Research Article

International Journal of Advanced Computer Research, Vol 8(35)
ISSN (Print): 2249-7277 ISSN (Online): 2277-7970
http://dx.doi.org/10.19101/IJACR.2018.834003

A framework for chronic kidney disease diagnosis based on case based reasoning

Seham Abd Elkader[1*], Mohammed Elmogy[2], Shaker El-Sappagh[3] and Abdel Nasser H. Zaied[1]
Faculty of Computers and Informatics, Zagazig University, Egypt [1]
Faculty of Computers and Information, Mansoura University, Egypt[2]
Faculty of Computers and Informatics, Benha University, Egypt[3]

Received: 13-October-2017; Revised: 12-February-2018; Accepted: 16-February-2018

Abstract
Chronic kidney diseases are very critical. Case-based reasoning (CBR) is a reasoning technique suitable for problems that depend on experiences. The first step in building CBR system is preparing a comprehensive case base from patients' electronic health records (EHRs). EHR data need quality improvement steps, such as normalization, feature selection, feature weighting, and outlier detection. In the medical field, the representation of resulting case base using formalized concepts and terminologies is highly needed. There are many structures for representing case bases, but the most powerful method for representation is using ontologies. The manuscript proposes a methodology for diagnosing the chronic kidney disease based on an ontology reasoning mechanism. In this paper, we first prepare the chronic kidney dataset of 400 real cases with 25 features by utilizing a set of data mining algorithms. Next, we construct an ontology structure to represent this case base in the W3C web ontology language (OWL) ontology format and populate this ontology with the individual cases. The fuzzy rough set algorithm achieved the highest accuracy for selecting the most suitable feature set. The resulting OWL ontology is based on disease ontology (DO) semantics, which is the most common and standardized ontology in the medical field.

Keywords
Case-based reasoning, Preprocessing, Chronic kidney diagnosis, Ontology structure, Disease ontology (DO).

1.Introduction
Chronic kidney disease (CKD) is a disease that causes damage to both kidneys and continues for a long time. The problem is that the damage is usually severe, which may lead to ill health. Studies have proven that CKD has become more common than previously thought [1–4]. CKD has become more common in older people especially who have diabetes mellitus. This may result in the spread of the disease in a broader range of population [1].

In addition, the high-risk factors that may face CKD patients include having cardiovascular diseases. In this case, CKD should be diagnosed early to suitable preventative measures can be considered. Treatment of patients in their early stage of CKD is the right decision to slow progression of the disease. Otherwise, if this disease is not handled on time, patients will get infected with kidney failure where both kidneys fail to perform their functions sufficiently [5–7].

Otherwise, if this disease is not handled on time, patients will get infected with kidney failure where both kidneys fail to perform their functions sufficiently [5–7].

The only treatment for patients with kidney failure (i.e., end-stage of CKD) is the kidney transplantation or dialysis where they are considered as a necessary solution to maintain patient's life. To manage CKD properly, an earlier diagnosis of the disease is required. Physicians should adopt their experiences for better CKD diagnosis. The CKD diagnosis is a critical step that needs a decision support system for helping physicians takes accurate decisions. Among the clinical decision support systems (CDSS), which have effective roles in the medical field, is the case-based reasoning (CBR) systems [4, 8–10]. Today, CBR systems are given more attention by various fields like finance, manufacturing, business, and healthcare. Those fields apply CBR to solve complex and ambiguous problems. Regarding healthcare domain, CBR has a vital role where domain knowledge, as well as information about cases, helps

*Author for correspondence

significantly in addressing the current problems [10]. Among CBR system's advantages are that they are used in situations where some cases may not exist in the case base. In addition, CBR systems also are beneficial in treating missing input values efficiently.

In CBR applications, the same problems have the same solutions. In CBR systems, past experiences are retained as cases [10]. When the particular purpose of the system being developed is well known, this will facilitate identifying the case contents. Any case in CBR systems has two main components, which are the problem and solution parts. The description data of patients with CKD include symptoms, physical examinations, past medical history, and laboratory tests. While, the solution part is about diagnosis, medications, and outcome. Case base in CBR systems is about a set of stored cases. A new problem in CBR system is solved by applying the solution of the CBR cases, which have the same description part of a current problem. The solution of the similar cases is retrieved by retrieval techniques. After that, the solution part of those recovered case is regarded the solution to the new problem [10].

The step of preparing a case base for patients is the initial and the most critical step when constructing a CBR system. Nevertheless, the case base preparation is not an easy step. The reason is that EHR data are inconsistent, incomplete, and noisy. Therefore, low-quality EHR data needs to be prepared for producing high-quality case base knowledge [10–12]. The quality of case bases content determines the CBR's quality [13]. The CBR systems' accuracy will be enhanced when the medical dataset of EHR is pre-processed.

After applying preprocessing steps on EHR data of CKD and CBR case base is generated, it is considered as necessary to represent the kidney case-base knowledge using many structures like ontology. Case base representation is highly required in the medical field. The properties, clinical terms, and various association types can be represented using ontologies [10, 14]. The benefit of applying the ontological representation method is that the knowledge base can be reused and shared later by other users on the semantic web. Regarding our knowledge in the medical field, there are no ontologies have been previously made in the chronic kidney domain. In addition, due to the high importance of this domain, this motivates us to represent it with a set of standardized attributes, classes, and relationship types. These formalized terms and associations can be easily applied to the biomedical community. Disease ontology (DO) [https://bioportal.bioontology.org/ontologies/DOID] is a standard ontology that has a great importance in building our domain ontology. DO aims to integrate clinical terms and diseases and to map them according to SNOMED, MeSH, OMIM terminologies.

Our paper is organized as follows. Section 2 is the related work. Section 3 shows the description of the kidney dataset with a sample of cases before applying preprocessing steps. Section 4 discusses in detail our preprocessing work along with the results of comparing different techniques with each other. Section 5 highlights our work that is related ontology development. Section 6 contains the conclusion and future work.

2.Related work

Many studies stated that EHR is only a comprehensive record, which retains all historical and current low-quality dataset of patients [11]. Thus the quality of stored data is not enhanced by EHR technology [15, 16]. Many studies illustrated that EHR is a starting source for the construction of CBR case base for medical systems [17–20]. However, because of the inconsistency, incompleteness, and existence of noisy and outlier's data values, low-quality EHR data needs to be prepared for producing high-quality CBR knowledge. The quality of the case base affects significantly on the quality of CBR application [16]. The research done by Richter and Weber [9] illustrated that to have high-quality content of case base, they should be prepared using effective and dependable sources. There must be a regular or a uniform distribution of cases of the problems [9]. When cases are not appropriately distributed or fairly among problems, this results in the existence of any problems without solutions while others may have redundant and useless cases. The CBR systems' accuracy will be enhanced when the medical dataset of EHR is pre-processed. The step of data pre-processing is about CBR as well as applying the many techniques of artificial intelligence, such as genetic algorithm, k-nearest neighbor (KNN), Bayesian network, and fuzzy approach [21]. The steps of data preprocessing steps involve handling missing data, feature selection and weighting, data integration, discretization of data, normalization, and outliers' detection and removal [22]. These data preprocessing steps are applied on EHR for converting database structure of EHR to case base

structure and transforming EHR generic data to specified case base [11].

Among the studies conducted, which focused on the missing data problem, was by Jagannathan and Petrovic [23]. The authors of this study illustrated that case base, which contains missing data, is a big problem. It affects negatively on the CBR system's performance. So, these missing values must be treated using imputation approaches like mean/mode method, KNN imputation, etc. However, the missing data values of some attributes have been handled while others are not treated. The performance of CBR applications is enhanced when applying preparation algorithms on the case base, such as feature selection. Every CDSS system, which is based on knowledge, needs a step of preprocessing to produce a high-quality dataset.

For obtaining superior results during the retrieval process, cleaning, and normalizing data steps are done on retrieval algorithms like KNN algorithm [23]. Gu et al. [21] conducted a study to evaluate the performance of a dental CBR system after normalizing EHR data. This study illustrated the importance of data normalization in cases retrieval stage. Data normalization is useful in the matching process between the new case and CBR cases. In another meaning, the normalization results in fair and accurate comparison between cases.

Many efforts have been made to preprocess EHR data as data quality needs to be measured as a knowledge source. Among these studies, Weiskopf and Weng [24] stated that five factors could be used for measuring EHR data quality. Xie et al. [25] handled the unmatched feature and missing value problem in case retrieval algorithms. Guessouma et al. [26] proposed five techniques for dealing with the problem of missing data in CBR system. Han et al. [27] has suggested RapidMiner for processing of the diabetes mellitus dataset. Pla et al. [17] proposed an eXit*CBR system that includes necessary preprocessing steps, such as normalization, discretization, and feature reduction and selection approaches. Therefore, we can determine the required steps of preprocessing based on the nature of the dataset as well as the goal needed for CBR application [13]. Missing data values can be handled by many methods like KNN Algorithm [23].

In addition, coding process can be done using RapidMiner Studio 6.0 by applying discretize operator. The discretize operator is about mapping process where the chosen features are presented in fixed classes. The generated features of patients are massively in most cases, and they have different importance levels. Consequently, the feature selection methods take all features set as input and then produce the most critical and needed attributes, which support the CBR decision-making process. Two attractive and complementary methods are applied in selecting the most important features, namely filter and wrapper methods [28]. There are also some algorithms that can be used for instance naïve Bayes (NB), decision tree (DT), C4.5, and KNN. The elected features using these methods will be used in case representation. The Rough set approach is also considered a powerful analysis tool for generating the most important and relevant features of the whole set of features.

Weights vector, which is assigned to attributes, is very useful for the case retrieval. Consequently, essential characteristics will have higher weights. There are a variety of algorithms calculate feature weightings, such as genetic algorithms, rule induction, DT, and correlation techniques [29, 30]. It is well known that the performance of algorithms, especially retrieval algorithms in CBR will be affected adversely if there are an outlier and extreme values in the dataset. Outliers affect the normalization process. The RapidMiner detection outlier operator checks for outliers. Identifying n outliers in data depends on how much the distance to k-nearest neighbors. Thus, extreme values of the feature can be replaced by domain experts. On the other hand, many studies have been done on case base representation using ontologies for different diseases, such as breast cancer and liver diseases. These studies followed many methodologies, evaluation, and validation methods for representing knowledge using ontologies. Now, we will shed light on some of them. El-Sappagh and Elmogy [31] developed a fuzzy ontology for diabetes mellitus diagnosis based on CBR technique. This ontology handles the vague aspects in diabetes domain.

Moawad et al. [32] applied an ontology which depends on the reality biomedical framework for constructing ontology for viral hepatitis disease. The authors followed three stages in developing this ontology. These developmental stages are gathering stage, validation stage, and OWL ontology representation stage. Their design of the ontology was about bottom-up technique. In addition, they implement their ontology using protégé OWL editing community. Jusoh et al. [33] constructed ontology for

breast cancer using a hybrid technique. The authors followed three phases of developing their ontology. These phases are preparation phase, using a hybrid process, and ontology construction phase. At the initial stage, all relevant data that will be needed in the study of the user, software, and other sources are prepared. In addition, in this phase, it is important to prepare domain knowledge, either from domain experts, documentation, or available ontologies. In this research, data are gathered from domain experts, documentation, journals, articles, and websites. Finally, the phase of ontology development was established where breast cancer ontology was developed using a hybrid mechanism.

Regarding evaluation and validation of ontologies, many studies have been done in this area. Among the studies that are conducted on ontology evaluation, Salem and Katoua [34] applied some criteria in evaluating ontologies. These evaluation criteria are decidability, completeness, maintainability, correctness, efficiency, and minimum redundancy. The authors showed that ontology could be validated by ensuring its quality as well as ensuring whether the ontology is related to the field or not. They also assured that completeness, consistency, clarity, consistency, robustness, and generality are factors used for assuring the quality of ontologies.

3.Materials and methods

3.1Dataset description

Table 1 represents a sample of kidney 13 patient cases (C1 to C13) along with all attributes, which describe the kidney disease patients. Data of these cases are about raw data where they are not yet prepared or preprocessed. In our dataset, there are about 400 cases with 25 features that help in the diagnosis of kidney disease. Steps of preprocessing will be applied to these raw data sources to produce high-quality case base enhanced for CKD diagnosis.

Table 1 A set of 13 cases

CASE	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13
Age	48	7	62	48	51	60	68	24	52	53	50	63	68
Blood Pressure	80	50	80	70	80	90	70	?	100	90	60	70	70
Specific Gravity	1.02	1.02	1.01	1.005	1.01	1.015	1.01	1.015	1.015	1.02	1.01	1.01	1.015
Albumin	1	4	2	4	2	3	0	2	3	2	2	3	3
Sugar	0	0	3	0	0	0	0	4	0	0	4	0	1
Red Blood Cells	?	?	N	N	N	?	?	N	N	AN	?	AN	?
Pus Cell	N	N	N	AN	N	?	N	AN	AN	AN	AN	AN	N
Pus Cell Clumps	NP	NP	NP	P	NP	NP	NP	NP	P	P	P	P	P
Bacteria	NP	NP	NP	NP	NP	NP	NP	NP	NP	NP	NP	NP	NP
Blood Glucose Random	121	?	423	117	106	74	100	410	138	70	490	380	208
Serum Creatinine	1.2	0.8	1.8	3.8	1.4	1.1	24	1.1	1.9	7.2	4	2.7	2.1
Sodium	?	?	?	111	?	142	104	?	?	114	?	131	138
Potassium	?	?	?	2.5	?	3.2	4	?	?	3.7	?	4.2	5.8
Hemoglobin	15.4	11.3	9.6	11.2	11.6	12.2	12.4	12.4	10.8	9.5	9.4	10.8	9.7
Packed Cell Volume	44	38	31	32	35	39	36	44	33	29	28	32	28
White Blood Cell Count	7800	6000	7500	6700	7300	7800	?	6900	9600	12100	?	4500	12200
Red Blood Cell Count	5.2	?	?	3.9	4.6	4.4	?	5	4	3.7	?	3.8	3.4
Hypertension	Yes	No	No	Yes	No	Yes	No	No	Yes	Yes	Yes	Yes	Yes
Diabetes Mellitus	Yes	No	Yes	No	No	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes
Coronary Artery Disease	No	No	No	No	No	No	No	No	No	No	No	No	Yes
Appetite	Good	G	Poor	Poor	G	G	G	G	G	Poor	G	Poor	Poor
Pedal Edema	No	No	No	Yes	No	Yes	No	Yes	No	No	No	Yes	Yes
Anemia	No	No	Yes	Yes	No	No	No	No	Yes	Yes	Yes	No	No
Class	Ckd	Ckd	Ckd	Ckd	Ckd	Ckd	Ckd	Ckd	Ckd	Ckd	Ckd	Ckd	Ckd

N = Normal, AN = Abnormal, P = Present, NP = Not present, G= Good, ? = Missing value.

3.2The proposed case base building framework

The case base building is the most critical step to build an intelligent CDSS system based on CBR technique. This process must be handled carefully because failure in the medical domain has critical costs that may be cannot be paid. Building this knowledge base from EHR content is a challenge. In this section, we have proposed a general framework for this process, as shown in *Figure 1*. The framework has three sequential steps. Data preparation applies a set of data mining techniques to improve the quality of case base data. Ontology creation constructs the ontology structure based on DO. The ontology population instantiates the created ontology by real cases from the prepared case base.

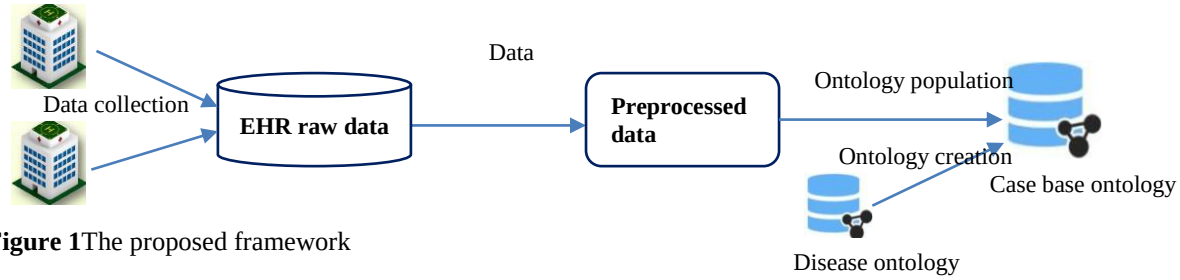

Figure 1 The proposed framework

4.Results and discussion

In this section, the results of each module of the proposed framework are detailed.

4.1Data normalization

This phase includes gathering dataset firstly and producing it in information table form. Then, data are normalized. It is critical to know that retrieval algorithms of CBR systems use similarity functions, which are based on distance comparison, such as Euclidean distance. Therefore, it is necessary to put all features of a given dataset in a normalized form. The advantage of data normalization is that it facilitates the comparison of features as they all will have the same scale. Thus, normalization step will be applied when we use features that have various scales. Regarding numerical features, normalization means that all features are in range, for example, a scale [0, 1]. In our dataset, we used Weka version 3.7.12 for normalizing numerical attributes in the scale [0.0, 1.0].

4.2Feature selection

In reality, the medical history of patients contains a large number of attributes. These description attributes do not have the same importance level. In addition, not all of these attributes will be used in the diagnosis of CKD. In addition, the full list of features has a negative impact on the performance of retrieval algorithms. This problem motivates us to apply algorithms for reducing the full list of features and selecting the most effective ones. The selected features will then be considered as input data for the classification algorithm. In addition, selected features will enhance the accuracy of classification. Two popular methods are used for feature selections, which are filter approach and wrapper approach. Regarding our kidney dataset, we adopted some machine learning algorithms for both filter and wrapper methods [28], see *Table 2*.

Table 2 Results of different feature selection techniques

Selection methods / **Classification algorithms**	**J48 subset evaluator method**	**SVM subset evaluator method**	**Filtered subset evaluator method**	**Wrapper subset evaluator method**	**Fuzzy rough feature selection method**
J48	Precision: 98.3% Recall:98.3% F-Measure:98.2% TP Rate:0.983 FP Rate: 0.027	Precision: 98% Recall:98% F-Measure:98% TP Rate:0.98 FP Rate: 0.025	Precision: 95.6% Recall:95.5% F-Measure:95.5% TP Rate:0.955 FP Rate: 0.04	Precision: 95.6% Recall:95.5% F-Measure:95.5% TP Rate:0.955 FP Rate: 0.04	Precision: 99% Recall:99% F-Measure:99% TP Rate:0.99 FP Rate: 0.014
SVM	Precision: 95.4% Recall:94.8% F-Measure:94.8% TP Rate:0.948 FP Rate: 0.032	Precision: 96% Recall:95.8% F-Measure:95.8% TP Rate:0.958 FP Rate: 0.034	Precision: 91.1% Recall:88.3% F-Measure:88.4% TP Rate:0.883 FP Rate: 0.071	Precision: 91.1% Recall:88.3% F-Measure:88.4% TP Rate:0.883 FP Rate: 0.071	Precision: 97.9% Recall:97.8% F-Measure:97.8% TP Rate:0.978 FP Rate: 0.013

Selection methods / Classification algorithms	J48 subset evaluator method	SVM subset evaluator method	Filtered subset evaluator method	Wrapper subset evaluator method	Fuzzy rough feature selection method
Fuzzy Rough NN	Precision: 39.1% Recall:62.5% F-Measure:48.1% TP Rate:0.625 FP Rate: 0.625	Precision: 77% Recall:63.5% F-Measure:50.3% TP Rate:0.635 FP Rate: 0.608	Precision: 77% Recall:63.5% F-Measure:50.3% TP Rate:0.635 FP Rate: 0.608	Precision: 77% Recall:63.5% F-Measure:50.3% TP Rate:0.635 FP Rate: 0.608	Precision: 96.4% Recall:96.3% F-Measure:96.3% TP Rate:0.963 FP Rate: 0.031
Naïve Bayes	Precision: 97.8% Recall:97.8% F-Measure:97.8% TP Rate:0.978 FP Rate: 0.016	Precision: 95.4% Recall:94.8% F-Measure:94.8% TP Rate:0.948 FP Rate: 0.032	Precision: 93.8% Recall:93% F-Measure:93.1% TP Rate:0.93 FP Rate: 0.047	Precision: 93.8% Recall:93% F-Measure:93.1% TP Rate:0.93 FP Rate: 0.047	Precision: 95.2% Recall:94.5% F-Measure:94.6% TP Rate:0.945 FP Rate: 0.033

To specify which group of features will be taken from the different algorithms of feature selection, we applied some classification algorithms, such as a J48 tree, SVM, and fuzzy- rough NN algorithm [35]. Next, we select the group of features, which have the highest accuracy level of classification. When comparing the results of classification accuracy against various feature selection methods, we noticed that fuzzy-rough features selection method achieves the highest classification accuracy, as shown in *Table 2*. As the number of features selected by the rough fuzzy algorithm achieves the highest accuracy results, we will take those attributes for handling their missing values and assigning weights to them.

4.3Handling missing values

It is worth mentioning that missing and incomplete data values have a bad impact on the performance of the retrieval algorithm [36]. Missing data can be handled by several methods [37–40]. Among these methods, removing instances, which include missing, attribute values and using the remaining instances for analysis. This method will decrease the size of data and thus produce some problems. Expectation-Maximization (EM) technique is also another method for treating missing values of attributes. This technique models the distribution of data input and includes two main steps, which are expectation and maximization. Its main idea is based on repeating those two steps many times until the maximum probability estimations produced. This method suffers from the complexity of both computation and determining the probability density function in advance. Another popular method for solving missing data problem is imputation. Imputation means filling missing values of attributes with values using techniques of machine learning. Techniques that can be used in imputation include mean, mode, and KNN techniques. Mean and mode imputation techniques include filling missing values using the average (i.e., mean) of all numerical features or mode of all nominal features. This imputation method is considered a powerful approach that may produce satisfactory results of accuracy. On the other hand, KNN imputation approach applies KNN technique for filling missing values. Its main idea is based on searching all instances to find the most similar instance for instance that has missing data. The barrier of this approach is that search becomes very hard when the size of the database is large.

Missing data can also be treated by using some algorithms, which handle missing data without imputation. Techniques that can handle missing values efficiently without imputation include NB, SMO, and J48. Regarding missing data in our dataset, we used Weka and applied a number of these methods along with comparing their results. From produced results, we noticed that replacing missing values with the mean values for numerical data or mode value for nominal data is better than other methods like to delete instances with missing data because it causes distortion and loss of some dataset instances. In addition, we noticed that fuzzy rough NN classification algorithm has the highest values of precision, recall, and F-measure compared to other algorithms like J48, NB, and SMO.

4.4Assigning weights to features

The performance of retrieval algorithms is improved when we rank features according to assigned weights by weighting algorithms. These weighting assigning algorithms can assign weights to cases as well as attributes. Cases with higher weights are considered critical for applications. On the other hand, weight can be calculated for attributes to determine the most important ones. Various algorithms for machine learning can be utilized in calculating weights for base case features. Regarding kidney dataset, we used RapidMiner and applied some algorithms for

attaching weights to the attributes. We already calculated selected features weights using evolutionary algorithms, particle swarm optimization (PSO), correlation algorithm, gain information algorithm, and rule induction algorithm, as shown in *Table 3*. However, features selected in features selection step should have a higher weight than other features. There are some features with weight equals zero. So, we will choose the maximum weight value for each attribute to form a weight vector for features in the CBR system. In *Table 3*, it is obvious that blood pressure and hemoglobin features have the highest value of weight. Consequently, those two attributes will be very efficient in diagnosing the health condition of patients. Features of pedal edema, specific gravity, diabetes mellitus, serum creatinine, hypertension, sodium, blood glucose random, blood urea, appetite, age, white blood cell count, and potassium come later.

Table 3 Features weights using different weigh assignment methods

	Evolutionary algorithm	**Particle swarm optimization**	**Correlation algorithm**	**Information gain algorithm**	**Rule induction algorithm**	**Chi-Squared statistics algorithm**	**Max. weight**
Age	0.139	0.479	0.227	0.008	0.142	0.199	0.479
Blood Pressure	1.0	0.318	0.327	0.189	0.0	0.272	1.0
Specific Gravity	0.593	0.873	0.066	0.734	0.550	0.692	0.873
Blood Glucose Random	0.121	0.596	0.497	0.362	0.133	0.274	0.596
Blood Urea	0.535	0.067	0.452	0.334	0.017	0.223	0.535
Serum Creatinine	0.498	0.472	0.333	0.804	0.008	0.060	0.804
Sodium	0.287	0.772	0.407	0.231	0.0	0.168	0.772
Potassium	0.380	0.266	0.0	0.0	0.0	0.0	0.380
Hemoglobin	0.864	0.215	1.0	1.0	1.0	1.0	1.0
White Blood Cell Count	0.0	0.423	0.197	0.027	0.0	0.126	0.423
Hypertension	0.285	0.227	0.787	0.511	0.392	0.489	0.787
Diabetes Mellitus	0.472	0.865	0.739	0.452	0.308	0.438	0.865
Appetite	0.024	0.244	0.485	0.181	0.0	0.213	0.485
Pedal Edema	0.766	0.974	0.457	0.155	0.0	0.194	0.974

Regards our kidney dataset, we tried to change the order of steps followed previously, which are normalization, feature selection, handling missing values, and feature weighting. In other words, we begin with feature selection and apply different methods, such as filtered method, wrapper method, information gain, and fuzzy rough methods. Then, we check the classification result for features selected. The result was the selected features by applying the filtered method to achieve the highest result of accuracy in classification. So, we choose those features for handling missing data. In dealing with missing data stage, we replace missing values with mean/ mode values. After that, we normalize all numerical attributes. Finally, we assign weights for selected features. We check the classification accuracy, using various algorithms. We noticed that accuracy is the same as steps applied previously.

5.The kidney disease ontology

At the beginning of our ontology development, we start the development by analyzing the terms and vocabularies of CKD field. This analysis step helps us determine the most known concepts, which describe the basics of this domain. The hierarchy of CKD ontology starts with a disease superclass. In addition, this class hierarchy is built based on the formalized concepts of DO [41]. The DO hierarchy for the kidney disease superclass is presented in *Figure 2*. The DO hierarchy introduces the most known terminologies in the kidney disease field. *Figure 3* shows a simple representation of this DO hierarchy using the protégé OWL tool. The kidney ontology includes five superclasses, which are *Disease*, *Patient*, *Medical_interference*, *Patient_country*, and *References*.

These superclasses include subclasses. In other words, we mean that Disease superclass has a subclass called *Disease_of_anatomical_entity*, then *Urinary_system_disease* subclass, and then *Kidney_disease* subclass. Secondly, the *Patient* superclass includes *Female* and *Male* subclasses. Thirdly, *Medical_interference* superclass has both *Diagnosis* and *Treatment* subclasses. After that, *Treatment* subclass are then composed of Treatment for *Early-stage-kidney-disease* and Treatment for *End-stage-kidney-disease*. While, *Patient_country* has three subclasses, which are *EgyptCountry*, *EuropeanUnionCountry*, and *UnitedStatesCountry*. Finally, the *References* superclass includes *Causes_of_kidney_disease*, *Kidney_complications*, *Kidney_Risk_Factors*, and *Symptoms_of_kidney_disease*.

Regarding relationships in our kidney ontology, as illustrated in *Table 4*. There are seven associations, which are *hasSymptoms*, *hasRiskFactors*, *hasComplications*, *isCausedBy*, *isLocatedIn*, *diagnosedBy*, and *treatedBy*, as shown in *Figure 4*. Each one of these relationships has both domain and range. The domain is a built-in property, which connects a property to a class description. While the range is a built-in property that relates a property to either a class description or a data range. *Table 5* describes class instances. Finally, the full classes of kidney ontology are shown in *Figure 5*.

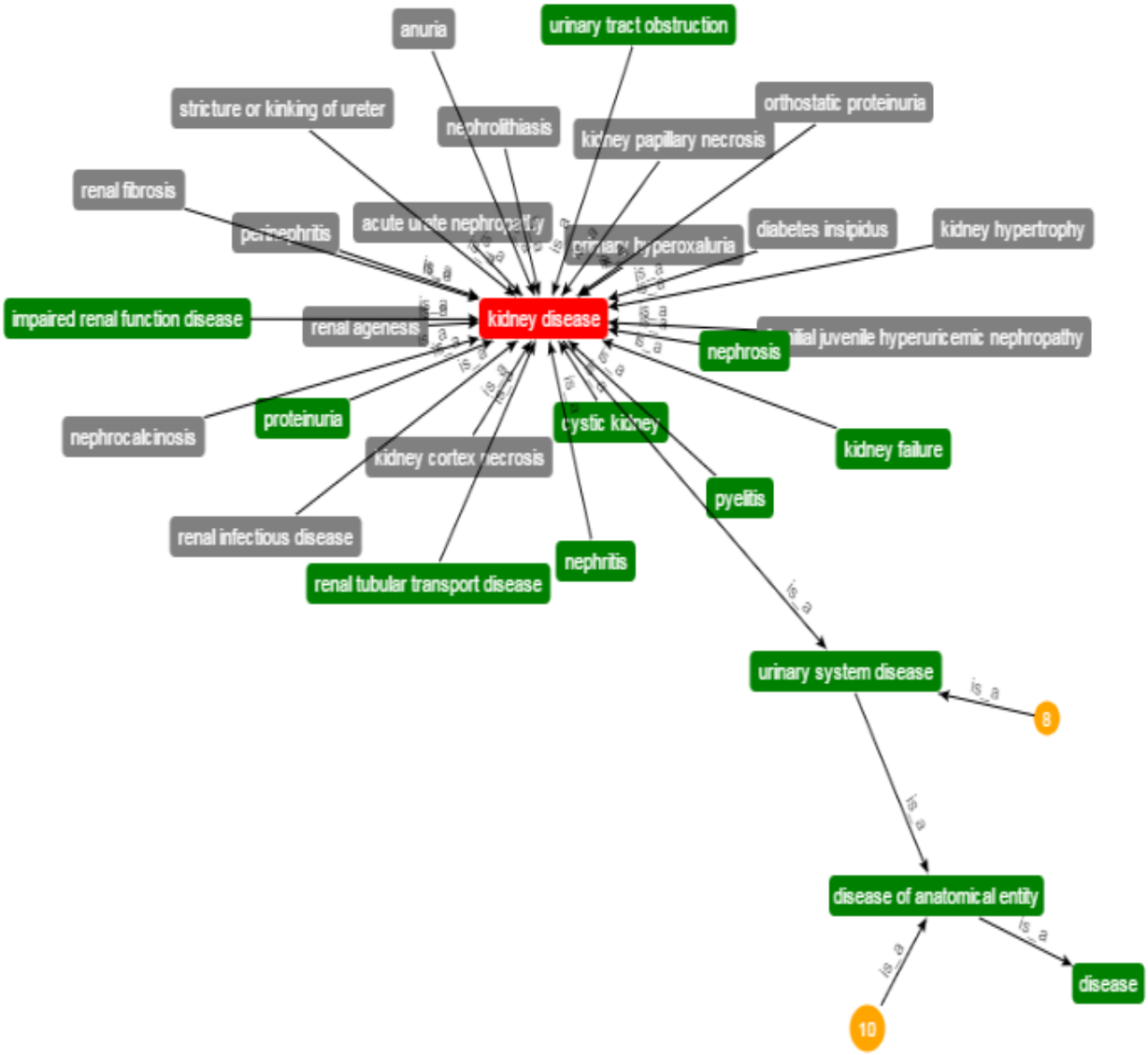

Figure 2 The kidney disease class visualization

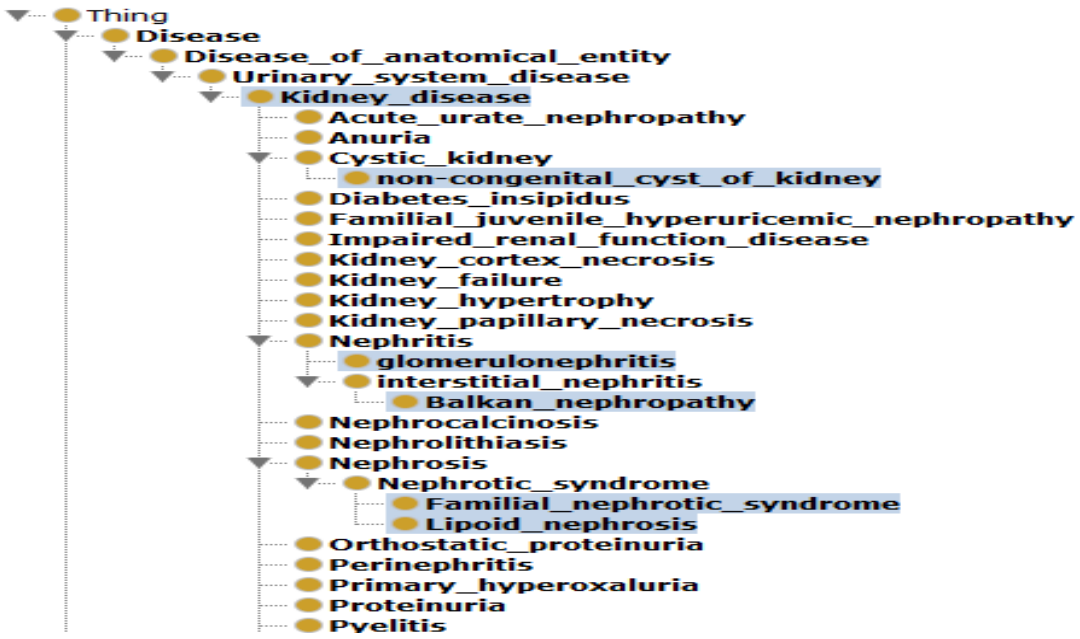

Figure 3 The kidney types in the ontology using the structure of DO

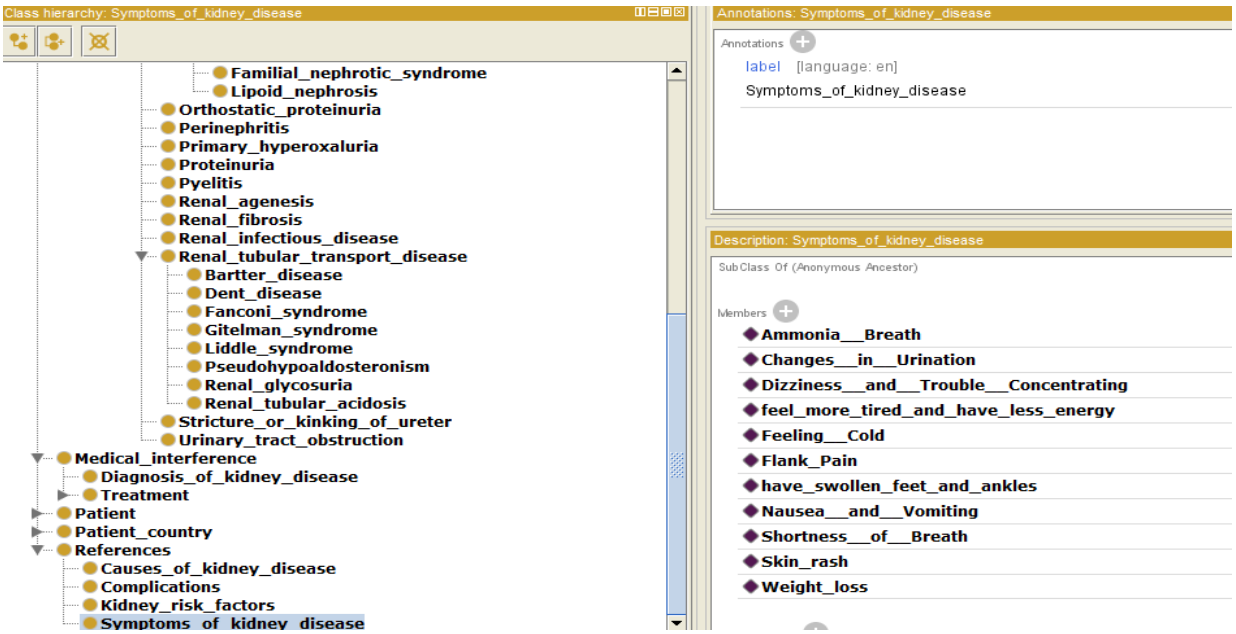

Figure 4 The representation of symptoms class and its individual members

Table 4 The object properties of kidney disease ontology

Property	Domain	Range
hasSymptoms	Kidney_Disease	Symptoms_of_Kidney_Disease
hasRiskFactors	Kidney_Disease	Kidney_Risk_Factors
hasComplications	Kidney_Disease	Kidney_Complications
isCausedBy	Kidney_Disease	Causes_of_Kidney_Disease
isLocatedIn	Patient	Patient_country
diagnosedBy	Kidney_Disease	Diagnosis_of_kidney_disease
treatedBy	Kidney_Disease	Treatment

Table 5 The instances or individuals of the classes of kidney disease ontology

Class	Instances
Kidney_Risk_Factors	Age_65_or_older, Atherosclerosis, Autoimmune disease, Being_AfricanAmerican_or_NativeAmerican_or_AsianAmerican,Bladder_cancer,Cigarette_smoking,Cirrhosis_and_liver_failure,Diabetes_both_type1_and_type2,Family_history_of_kidney_disease,High_cholersterol,Kidney_cancer,Kidney_infection,Kidney_stones,Lupus,Narrowing_of_the_artery_that_supplies_the_kidney,Obesity,Obstructive_kidney_disease,Scleroderma,Vasculitis, Vesicoureteral_reflux
Symptoms_of_Kidney_Disease	Dizziness_and_Trouble_Concentrating,feel_more_tired_and_have_less_energy,Feeling_Cold,Flank_Pain,have_swollen_feet_and_ankles,Nausea_and_Vomiting,Shortness_of_Breath,Skin_rash,Weight_loss
Kidney_complications	Decreased_immune_response,Decreased_sex_drive_or_impotence,Fluid_retention,Heart_and_blood_vessel_disease,Hyperkalemia,Irreversible_damage_to_kidneys,Pericarditis,Pregnancy_complications,Weak_bones_and_bone_fractures
Causes_of_Kidney_Disease	Diabetes,Glomerulonephritis,High_blood_pressure,Interstitial_nephritis,Lupus_and_other_diseases,Malformations,Polycystic_kidney_disease,Reccurent_kidney_infection
Diagnosis_of_kidney_disease	Blood_pressure_test,kidney_biopsy,Renal_ulterasound,Serum_creatinine,Urine_albumin_test
Treatment	Ask_doctor_about_medicines_for_protecting_kidneys,Control_blood_sugar_when_having_diabetes,Dialysis,Do_not_smoke_or_use_tobacco,Eat_a_heart_healthy_diet,Exercise_most_days_of_week,Keep_a_healthy_weight,Keep_healthy_blood_pressure,Kidney_transplant,Limit_alcohol

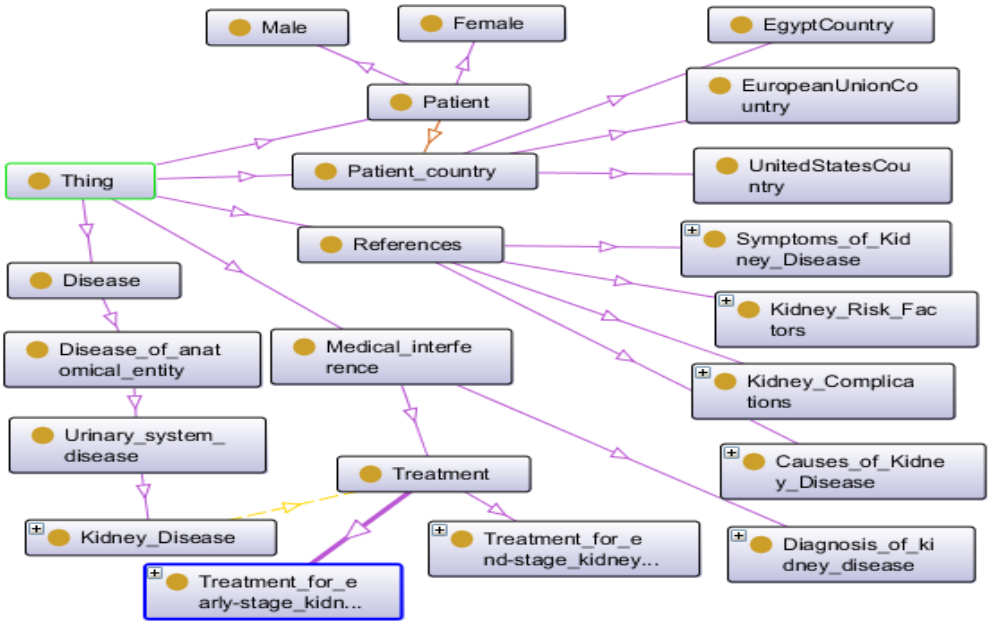

Figure 5 The kidney full ontology classes

After building the OWL ontology for the CBR system, in the next future work, we will utilize this knowledge base to build the full CDSS system for the kidney disease diagnosis. The resulting ontology is compatible with EHR semantics because it uses OWL format with rich semantics [41].

6.Conclusion

This paper concentrated on the preprocessing step of kidney dataset and representing the resulting data in the form of ontology. The followed steps were numerical dataset normalization, feature reduction and selection, treating missing data values, and features weights assignment. In each of these preprocessing steps, we adopted many algorithms. In addition, we record the results of classification accuracy of these techniques. These steps were done on a dataset of kidney patients. The resulted high-quality data can be considered as a case base knowledge, which can improve the retrieving process in CBR systems. We then focused on representation the produced dataset in ontology structure. We followed a mechanism for developing the ontology of

chronic kidney disease. We applied the most common terms and concepts of disease ontology, which are a well-known and standardized medical ontology. In our future work, we will concentrate on applying the remaining steps in case base preparation like handling fuzziness in the case base. Also, we will implement our CBR system for diagnosing patients with kidney disease. In addition, we seek to enlarge the size of kidney case base for improving the retrieval results of the CBR system.

Acknowledgment
None.

Conflicts of interest
The authors have no conflicts of interest to declare.

References
[1] Ali M, Han SC, Bilal HS, Lee S, Kang MJ, Kang BH, et al. iCBLS: An interactive case-based learning system for medical education. International Journal of Medical Informatics. 2018; 109:55-69.
[2] Relich M, Pawlewski P. A case-based reasoning approach to cost estimation of new product development. Neurocomputing. 2018; 272:40-5.
[3] Williams JK. Management strategies for patients with diabetic kidney disease and chronic kidney disease in diabetes. Nursing Clinics. 2017; 52(4):575-87.
[4] Chazara P, Negny S, Montastruc L. Flexible knowledge representation and new similarity measure: application on case based reasoning for waste treatment. Expert Systems with Applications. 2016; 58:143-54.
[5] Kidney Disease: Improving Global Outcomes (KDIGO) CKD-MBD Update Work Group. KDIGO 2017 Clinical Practice Guideline Update for the Diagnosis, Evaluation, Prevention, and Treatment of Chronic Kidney Disease–Mineral and Bone Disorder (CKD-MBD). Kidney International Supplements. 2017; 7: 1–59.
[6] Strippoli GF, Tong A, Johnson D, Schena FP, Craig JC. Catheter-related interventions to prevent peritonitis in peritoneal dialysis: a systematic review of randomized, controlled trials. Journal of the American Society of Nephrology. 2004; 15(10):2735-46.
[7] Renal association. Treatment of adults and children with renal failure: standards and audit measures, 2002. Royal College of Physicians of London and the Renal Association, London.
[8] Blanco X, Rodríguez S, Corchado JM, Zato C. Case-based reasoning applied to medical diagnosis and treatment. In distributed computing and artificial intelligence 2013 (pp. 137-46). Springer, Cham.
[9] Richter MM, Weber RO. Case-based reasoning. Springer-Verlag Berlin An; 2016.
[10] Yadav P. Case retrieval algorithm using similarity measure and adaptive fractional brain storm optimization for health in formaticians. Arabian Journal for Science and Engineering. 2016; 41(3):829-40.
[11] Abidi SS, Manickam S. Leveraging XML-based electronic medical records to extract experiential clinical knowledge: an automated approach to generate cases for medical case-based reasoning systems. International Journal of Medical Informatics. 2002; 68(1-3):187-203.
[12] Borges KC, De Barcelos Tronto IF, De Aquino Lopes R, Da Silva JD. A methodology for preprocessing data for application of case based reasoning. In informatica 2012 (pp. 1-8). IEEE.
[13] Andritsos P, Jurisica I, Glasgow JI. Case-based reasoning for biomedical informatics and medicine. In springer handbook of bio-/neuroinformatics 2014 (pp. 207-21). Springer Berlin Heidelberg.
[14] Chen SM, Huang YH, Chen RC, Yang SW, Sheu TW. Using fuzzy reasoning techniques and the domain ontology for anti-diabetic drugs recommendation. In Asian conference on intelligent information and database systems 2012 (pp. 125-35). Springer, Berlin, Heidelberg.
[15] El-Sappagh S, Elmogy M. An encoding methodology for medical knowledge using SNOMED CT ontology. Journal of King Saud University-Computer and Information Sciences. 2016; 28(3):311-29.
[16] Weiner MG, Embi PJ. Toward reuse of clinical data for research and quality improvement: the end of the beginning? Annals of Internal Medicine. 2009; 151(5):359-60.
[17] Pla A, LóPez B, Gay P, Pous C. eXiT* CBR. v2: Distributed case-based reasoning tool for medical prognosis. Decision Support Systems. 2013; 54(3):1499-510.
[18] O'Connor PJ, Sperl-Hillen JM, Rush WA, Johnson PE, Amundson GH, Asche SE, et al. Impact of electronic health record clinical decision support on diabetes care: a randomized trial. The Annals of Family Medicine. 2011; 9(1):12-21.
[19] Gu D, Liang C, Zhao H. A case-based reasoning system based on weighted heterogeneous value distance metric for breast cancer diagnosis. Artificial Intelligence in Medicine. 2017; 77:31-47.
[20] Gonzalez C, M.lopez D, Blobel B. Case-based reasoning in intelligent health decision support systems. In PHealth 2013: proceedings of the international conference on wearable micro and nano technologies for personalized health 2013 (pp. 44-9). IOS Press.
[21] Gu DX, Liang CY, Li XG, Yang SL, Zhang P. Intelligent technique for knowledge reuse of dental medical records based on case-based reasoning. Journal of Medical Systems. 2010; 34(2):213-22.
[22] Begum S, Ahmed MU, Funk P, Xiong N, Folke M. Case-based reasoning systems in the health sciences: a survey of recent trends and developments. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews). 2011; 41(4):421-34.
[23] Jagannathan R, Petrovic S. Dealing with missing values in a clinical case-based reasoning system. In

international conference on computer science and information technology 2009 (pp. 120-4). IEEE.
[24] Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. Journal of the American Medical Informatics Association. 2013; 20(1):144-51.
[25] Xie X, Lin L, Zhong S. Handling missing values and unmatched features in a CBR system for hydro-generator design. Computer-Aided Design. 2013; 45(6):963-76.
[26] Guessoum S, Laskri MT, Lieber J. RespiDiag: a case-based reasoning system for the diagnosis of chronic obstructive pulmonary disease. Expert Systems with Applications. 2014; 41(2):267-73.
[27] Han J, Rodriguez JC, Beheshti M. Diabetes data analysis and prediction model discovery using rapidminer. In international conference on future generation communication and networking 2008 (pp. 96-9). IEEE.
[28] Molina LC, Belanche L, Nebot À. Feature selection algorithms: a survey and experimental evaluation. In international conference on data mining 2002 (pp. 306-13). IEEE.
[29] Lausch A, Schmidt A, Tischendorf L. Data mining and linked open data-new perspectives for data analysis in environmental research. Ecological Modelling. 2015; 295:5-17.
[30] Kar D, Chakraborti S, Ravindran B. Feature weighting and confidence based prediction for case based reasoning systems. In international conference on case-based reasoning 2012 (pp. 211-25). Springer, Berlin, Heidelberg.
[31] El-Sappagh S, Elmogy M. A fuzzy ontology modeling for case base knowledge in diabetes mellitus domain. Engineering Science and Technology, an International Journal. 2017; 20(3):1025-40.
[32] Moawad IF, Marzoqi GA, Salem AB. Building OBR-based OWL ontology for viral hepatitis. Egyptian Computer Science Journal. 2012; 36(1):89-98.
[33] Jusoh F, Ibrahim R, Othman MS, Omar N. Development of breast cancer ontology based on hybrid approach. International Journal of Innovative Computing. 2013; 3(1):15-22.
[34] Salem AB, Katoua HS. Web-based ontology of knowledge engineering. Journal of Communication and Computer. 2012; 9: 254-9.
[35] Goswami S, Das AK, Chakrabarti A, Chakraborty B. A feature cluster taxonomy based feature selection technique. Expert Systems with Applications. 2017; 79:76-89.
[36] Ramos-González J, López-Sánchez D, Castellanos-Garzón JA, De Paz JF, Corchado JM. A CBR framework with gradient boosting based feature selection for lung cancer subtype classification. Computers in Biology and Medicine. 2017; 86:98-106.
[37] Saraiva R, Perkusich M, Silva L, Almeida H, Siebra C, Perkusich A. Early diagnosis of gastrointestinal cancer by using case-based and rule-based reasoning. Expert Systems with Applications. 2016; 61:192-202.
[38] Silva LO, Zárate LE. A brief review of the main approaches for treatment of missing data. Intelligent Data Analysis. 2014; 18(6):1177-98.
[39] Soley-Bori M. Dealing with missing data: key assumptions and methods for applied analysis. Boston University. 2013.
[40] Suthar B, Patel H, Goswami A. A survey: classification of imputation methods in data mining. International Journal of Emerging Technology and Advanced Engineering. 2012; 2(1):309-12.
[41] Schriml LM, Arze C, Nadendla S, Chang YW, Mazaitis M, Felix V, et al. Disease ontology: a backbone for disease semantic integration. Nucleic Acids Research. 2011; 40(D1):D940-6.

Seham Abd Elkader is a teaching assistant at Faculty of Computers and Informatics, Zagazig University, Egypt. She had received her B.Sc. and M.Sc. from Information Systems Department, Faculty of Computers and Informatics, Zagazig University, Egypt. She is currently studying a PhD degree at the department of computer science, University of New South Wales, Canberra, Australia. Her current research interests are Signal Processing, Machine learning, Pattern recognition, ontology engineering, distributed and hybrid clinical decision support systems, medical data encoding, medical terminology, semantic interoperability, cloud computing, and data integration.
Email: seham.abdo2@gmail.com

Dr. Mohammed Elmogy is an associate professor at Faculty of Computers and Information, Mansoura University, Egypt. He had received his B.Sc. and M.Sc. from Faculty of Engineering, Mansoura University, Mansoura, Egypt. He had received his Ph.D. from Informatics Department, MIN Faculty, Hamburg University, Germany in 2010. He is currently working as a post-doctoral research associate at the bioengineering department, University of Louisville, Louisville, USA. He is a member of IEEE and ACM. He authored and co-authored over 120 publications in recognized international journals and conferences. He advised approximately 20 master and doctoral graduates. His current research interests are Computer vision, Machine learning, Pattern recognition, and Biomedical Engineering.
Email: melmogy@mans.edu.eg

Dr. Shaker El-Sappagh was born in El-Behara, Egypt, in 1977. He received the bachelor degree in computer science from Information Systems Department, Faculty of Computers and Information, Cairo University, Cairo, Egypt, in 1997, and the master degree from the same university in 2007. He received the Ph.D. degrees in computer science from Information Systems Department, Faculty of Computers and Information, Mansura University, Mansura, Egypt in 2015. In 2003, he joined the Department of Information Systems, Faculty of Computers and Information, Benha University, Banha, Egypt as a teaching assistant. In 2009, he joined the Collage of Science, King Saud University as a lecturer. Since June 2016, he has been with the Department of Information Systems, Faculty of computers and Information, Benha University as a lecturer. He has publications in clinical decision support systems and semantic intelligence. His current research interests include medical informatics, ontology engineering, distributed and hybrid clinical decision support systems, semantic data modeling, distributed database systems, big data, semantic query languages, medical data encoding, medical terminology, semantic interoperability, description logic, fuzzy logic, fuzzy mathematics, fuzzy database, semantic database, cloud computing, data integration, semantic web, and fuzzy expert systems.
Email: shaker_elsapagh@yahoo.com

Dr. Abdel Nasser H. R. Zaied is a Professor of Systems Engineering (Information Systems), Dean, College of Computers and Informatics, Zagazig University, Egypt. Dr. Zaied is adviser to the Minister of Higher Education for Private and National Universities, Egypt. He received B.SC. and M.SC. in Mechanical Engineering from Department of Mechanical Engineering, College of Engineering, Ain Shams University, Egypt. He received Ph.D. from Department of Industrial Engineering, College of Engineering, Zagazig University, Egypt. He authored and co-authored over 50 publications in recognized international journals and conferences.
Email: nasserhr@zu.edu.eg