Journal of Information & Knowledge Management
Vol. 19, No. 1 (2020) 2040015 (17 pages)
© World Scientific Publishing Co.
DOI: 10.1142/S0219649220400158

Classification and Association Rule Mining Technique for Predicting Chronic Kidney Disease

Ahmad Alaiad*,‡, Hassan Najadat*,§, Belal Mohsen†,¶
and Khaled Balhaf†,||

**Department of Computer Information Systems*
Jordan University of Science and Technology
Irbid 22110, Jordan

†*Department of Computer Engineering*
Jordan University of Science and Technology, Jordan
‡*aiaiad@just.edu.jo*
§*najadat@just.edu.jo*
¶*blmohsen14@cit.just.edu.jo*
||*khbalhalf14@cit.just.edu.jo*

Published 11 March 2020

Abstract. Background and objective: Chronic kidney disease (CKD) is one of the deadly diseases that can affect a lot of vital organs in the human body such as heart, liver, and lungs. Many individuals might be at early stage of kidney disease and not have any signs, which might lead to a sudden death. Previous research showed that early prediction of CKD is very important in the medical field for physicians' decision-making and patients' health and life. To this end, constructing an efficient prediction system for CKD, which is the goal of this paper, often reduces medical errors and overall healthcare cost. **Methods:** Classification and association rule mining techniques were integrated and utilised to construct an efficient system for predicting and diagnosing CKD and its causes using weka and SPSS as platform environments. In particular, five classification algorithms, namely, naive Bayes, decision tree, support vector machine, K-nearest neighbour, and JRip were used to achieve the research goal. In addition, Apriori algorithm was used to discover strong relationship rules between attributes. The experiments were conducted on real medical dataset collected from hospitals and patient monitoring systems. **Results:** The experiments achieved high accuracy of 98.50% for K-nearest neighbour (KNN) classifier and achieved 96.00% when using classier based on association rule (JRip). **Conclusions:** We conclude by showing that applying integrative approach by combining classification algorithms and association rule mining can significantly improve the classification accuracy and be more useful for CKD prediction. This research has also several theoretical and practical implications for the medical field and healthcare industry.

Keywords: Chronic kidney disease; classification; association rule mining; KNN; predictive model.

1. Introduction

Chronic kidney disease (CKD) is one of the diseases that might lead to death (Basar and Akan, 2018). It is considered as one of the most common diseases in many countries in the world (Arasu and Thirumalaiselvi, 2017). According to National Kidney Foundation (n.d.), 10% of the population in the world have CKD, and millions die annually because of it. In middle-income countries, treatment of CKD creates a huge financial burden for the patients who need it (Arasu and Thirumalaiselvi, 2017). In other 112 countries, many patients cannot afford its treatment, resulting in death of over 1 million people annually from untreated kidney failure. In the US, treatment of CKD is likely to exceed $48 billion each year (Aqlan *et al.*, 2017). Treatment for kidney failure consumes 6.7% of the total Medicare budget to care for less than 1% of the covered population (Aqlan *et al.*, 2017). Many individuals might be at early stage of kidney disease and not have any signs, which might lead to a sudden death. In CKD, the kidneys do not work well as they should, thus wastes accumulate gradually in the blood. The harmful chemicals of the body such as phosphorus and potassium may grow to abnormal levels that may cause heart failure, bone problems, and anaemia. All these health problems can lead to kidney failure that makes patient life at danger (National Institute of Diabetes and Digestive and Kidney Diseases, NIDDK). However, CKD can be treated. With an early diagnosis and treatment, it is possible to slow or stop the progression of kidney disease. Previous research showed that early prediction of CKD is very significant in the medical field for physicians' decision-making, patients' health and life, and reduction of the overall healthcare costs (Arasu and Thirumalaiselvi, 2017; Aqlan *et al.*, 2017; Basar and Akan, 2018; Zeynu and Patil, 2018).

Ten years ago, researchers have attempted to utilise information system approaches in health care for minimising the cost and impact of the diseases on people and thus saving their life (Brossette *et al.*, 1998; Akiyama and Fujita, 2013; Maeda *et al.*, 2016). One of those approaches is Knowledge Discovery and Data Mining (KDDM) that helps to predict and discover signs of diseases (Boukenze *et al.*, 2017). The medical knowledge discovery and the findings of its research help governments, hospitals, and patients to reduce cost of treatments, reduce risks of human's errors, and save patients' life (Kaur and Sharma, 2017). KDDM is an important process for extracting knowledge from a huge amount of data. KDDM uses several methods and techniques to extract interesting knowledge that can be used to support decision-making. These methods and techniques consist of association rules, classification, and clustering (Zeynu and Patil, 2018). KDDM techniques involve many iterative steps to extract the significant knowledge, which is used to make a right decision in an efficient manner (Arasu and Thirumalaiselvi, 2017).

Previous studies used such techniques for extracting knowledge, predicting pathological cases, and discovering useful patterns of treatments for many patients (Pendyala *et al.*, n.d.; Piatetsky-Shapiro, 1991; Li *et al.*, 2001; Lee *et al.*, 2007; Duan *et al.*, 2011; Gopika and Vanitha, 2017). For instance, Maeda *et al.* (2016) used text mining techniques to study the health of Japanese people who claim having high health problems. In addition, Tsanas *et al.* (2012) used classification methods as data mining technique to diagnose Parkinson's diseases. In the study conducted by Kaur and Sharma (2018), Multibank algorithm was implemented for the prediction of asthma disease. Many data mining techniques were used for this purpose. In their results, 80% accuracy has been obtained. On the contrary, Arasu and Thirumalaiselvi (2017) and Aqlan *et al.* (2017) used association rule mining techniques for electronic health record (EHR). Recently, several studies such as Kaur and Sharma (2018), Otunaiya and Muhammad (2019), and Zeynu and Patil (2018) used data mining techniques to extract hidden information from electronic medical record (EMR). However, limited research used an integrative approach by combining several techniques together for the purpose of extracting knowledge from medical data. In particular and related to the domain of this research, none of the previous studies used an integrative approach for predicting and diagnosing CKD.

In order to fill this knowledge gap, classification and association rule mining algorithms were integrated and used in this research to design and develop a classification system for predicting CKD using weka and spss as platform environments (Weka 3, n.d.). The data mining approach used in Kumbhare and Chobe (2014) was followed and applied to achieve this goal. In particular, five classification algorithms, namely, naive Bayes (NB), decision tree (J48), support vector machine (SVM), K-nearest neighbour (KNN), and classier based on association rule (JRip), were used to predict and diagnose CKD. In addition, Apriori algorithm was used to discover strong association rules between attributes. The chronic disease dataset was taken from hospitals and patient monitoring systems. The data were first preprocessed properly in a careful manner and solid approach to resolve missing data issues. Weka and spss platform environments were used to analyse the data and apply the five algorithms as well as the Apriori algorithm. The findings were impressive and useful for patients, physicians, governments, and decision-makers in the medical and health informatics field.

This research makes several contributions to the literature. First, it shows that applying an integrative approach by combining classification algorithms with association rule mining improves the accuracy of prediction, especially for medical data. Secondly, it shows that KNN classifier has the highest accuracy among others for chronic disease data. Thirdly, it demonstrates that applying Apriori algorithm for CKD dataset to extract strong rules can significantly improve classification accuracy and gain strong rules from CKD dataset.

2. Related Work and Limitations

Boukenze *et al.* (2016) used SVM and NB learning algorithms to extract information from database. They classified the patients as chronic kidney disease patients (CKD) and not chronic kidney disease patients (notCKD). Kaur and Sharma (2018) also used SVM and KNN to predict CKD in Hadoop applying Matlab. Their results showed that SVM had a higher accuracy. In Kunwar *et al.* (2016), NB and artificial neural network (ANN) algorithms have been used to diagnose CKD using rapidminer tool. In addition, Otunaiya and Muhammad (2019) evaluated the performance of multilayer perceptron, NB, and J48 decision tree in the prediction of CKD using Weka platform and the results showed that J48 decision tree gave the best result. However, in our research, we used different datasets, different classifiers, and association rule techniques. We also followed an integrative approach by combining association rule mining and classification for predicting and diagnosing CKD.

On the contrary, several algorithms such as NB, majority (TestLearners), KNN, classification tree, SVM, and random forest classification methods have been used to classify the kidney stones data (Kaladha *et al.*, 2012). The results showed that SVM gave 91.98% of accuracy. In addition, KNN, J48, ANN, NB, and SVM classification techniques were used to diagnose CKD (Sharma and Kaur, 2018). Two important models were built; namely, feature selection method and ensemble model. It has been observed that ensemble model achieved better accuracy (99%). Furthermore, Garg and Mongia (2018) implemented four data mining algorithms (Decision Tree, Bayesian, SVM, and lazy learner) on CKD dataset. The accuracy of NB classification algorithm was being measured highest. Moreover, Rubini and Eswaran (2015) used three classification methods: RBF network, multilayer perceptron MLP, and logistic regression for classifying their CKD dataset, which was classified into: CKD and notCKD. As finding, experiments showed that MLP classifier provided highest accuracy of 99.75%. However, in our research, we used association rule techniques to generate strong rules from CKD dataset for making a better decision. We also applied four different classification methods, in addition to classifier based on association rule (JRip) that achieved a high accuracy for our problem.

In contrast, association rule is a technique used to discover patterns or relations between attributes (Piatetsky-Shapiro, 1991). Once the association has finished processing, we typically get a decision tree as if–then statements. Association rules were primarily introduced to extract multi-correlated items in transactions (Agrawal *et al.*, 1993a, 1993b; Moreno *et al.*, 2005). In several previous studies such as Agrawal *et al.* (1993a, 1993b) and Chandrakar and Kirthima (2013), the authors proposed the main concept of association between items and introduced the Apriori algorithm to discover the relation between attributes. After that, the efficient method has been studied for extraction association rules from databases. Gautam and Pardasani (2010) introduced new version of Apriori algorithm, called SC-BF multilevel, for mining multi-level association rule that is used with massive databases to extract maximum frequent item-set at lower level of abstraction. The SC-BF algorithm is fast and effective version that needs only single access of database for extracting frequent item-sets. Three association rule mining methods, namely, Apriori, FP-Growth, and tertius algorithm, have been applied on supermarket data and compared with each other (Arora *et al.*, 2013). The results showed that the FP-growth is better than the others.

Sunita and Lobo (2012) applied three data mining techniques: clustering (K-means algorithm), classification (alternating decision tree algorithm), and association rule (Apriori algorithm) on a dataset that contains data about courses learnt by students. These techniques aim to find the best set of courses. The results showed that the use of integrated approach is better than applying only association rule mining for this task. However, limited research in the medical field applied such an integrative approach on medical dataset, especially related to CKD.

In the medical field, Nahar *et al.* (2009) used association rule mining techniques for their experiment to predict heart diseases from heart disease data. They applied three association rule mining algorithms: Apriori, predictive Apriori, and tertius on heart disease dataset and compared their results. The results showed that the Apriori algorithm has high-quality performance for this kind of task. Therefore, we adopted this technique in our research. In addition, Ganda (2013) integrated clustering and association rule mining on kidney dataset that contains 7 attributes and 157 instances to form strong rules in each cluster by weka. However, in this paper, we applied Apriori algorithm for CKD dataset to extract a strong rule, improve classification accuracy, and gain strong rule from CKD dataset. Finally, Afhami (2018) employed J48, NB, Bayesian Network, SVM, SMO, Random Forest, Bagging, and MLP to predict the disease progression in 249 patients with diabetic chronic kidney. Random Forest had better performance in terms of precision, Recall, and F-measure. Afhami employed rough set theory to extract rules for prediction of the disease progression to the last stage of renal disease or mortality.

3. Algorithms Used

3.1. *Classification algorithms*

Naive Bayes (NB)

An NB classifier is a probabilistic machine learning model that is based on the Bayes theorem (Basar and Akan, 2018). The basic assumption of NB classifier is that the presence or absence of a particular feature is not related to the presence or absence of any other feature (Otunaiya and Muhammad, 2019). In a supervised learning environment, an NB classifier can be trained in an efficient manner. It only needs a limited size of training data to estimate the classification (Arasu and Thirumalaiselvi, 2017; Zeynu and Patil, 2018; Otunaiya and Muhammad, 2019). NB classifier calculates the variances of the variables for each class, not the entire covariance matrix because variables in NB are assumed to be independent (Bay and Le, 2008; Arasu and Thirumalaiselvi, 2017; Zeynu and Patil, 2018).

Using Bayes theorem, we can calculate posterior probability in the following way:

$$P(c|x) = [p(x|c) * p(c)]/p(x),$$

- $P(c|x)$ is the posterior probability of class c, given x.
- $P(c)$ is the prior probability of class.
- $P(x|c)$ is the likelihood which is the probability of predictor given class.
- $P(x)$ is the prior probability of predictor.

Support vector machine (SVM)

SVM is a supervised machine learning technique used for both classification (mostly used) and regression problems. In this algorithm, each data item is represented as a point in n-dimensional space (Garg and Mongia, 2018). An optimal hyper-plane that differentiates the two classes is found to conduct classification task on the dataset. SVM is a frontier which best segregates the two classes. SVM finds an optimal hyperplane applying the margin concept (Zeynu and Patil, 2018).

The margin is the distance between the hyperplane and the closest points on either side, which we want to maximise for better generalisation (Kaur and Sharma, 2018).

Decision Tree

Decision Tree is one of the most commonly used classifiers in machine learning (Basar and Akan, 2018). Some characteristics of decision tree algorithm include: it is a non-parametric method, it is used for both classification and regression, it employs the divide-and-conquer technique, and it can be converted easily to simple if–then rules (Zeynu and Patil, 2018). Decision tree represents a simple graphic structure in which each internal node represents a test, each branch represents an outcome, and leaf nodes represent classes (Otunaiya and Muhammad, 2019). In this algorithm, to classify an unknown instance, the values of attribute are tested against the decision tree. In particular, each instance attribute is assessed using a statistical method to determine how it classifies the training samples. Then, a path is traced from the root node to a leaf node that carries the class for that instance (Garg and Mongia, 2018).

K-nearest neighbour (KNN)

KNN is an instance-based learning commonly used for classification task (Arasu and Thirumalaiselvi, 2017). The dataset records are represented in a *d*-dimensional space. The attributes of the records are coordinated in the space (Zeynu and Patil, 2018). Given a new instance, KNN uses some similarity measures to calculate the similarity of the new point to the data points of the model in order to classify it (Kaur and Sharma, 2018). To this end, KNN states the class of the new point by first finding the K closest points to the new one, then KNN uses the majority vote to find the most common class among them to be the class of the new point. For example, if $K = 3$, then the class of the new point will be the most common class among these nearest neighbours (Kaur and Sharma, 2018).

In practice, the algorithm finds a specific class to the unseen example according to its similar neighbours (Arasu and Thirumalaiselvi, 2017; Zeynu and Patil, 2018). The similarity between two items is defined by a distance function. The most common distance function that is used to measure similarity is the Euclidian distance and it is defined by:

$$D(x, y) = \sqrt{\sum_i (xi - yi)^2},$$

where $x = x1, \ldots, xm$ and $y = y1, \ldots, ym$ and m is the attribute value of two points x and y. The more shorter the distance between x and y, the more similar x and y are.

3.2. *Classier based on association rule (JRip)*

JRip is one of the most popular classification rule algorithms. It implements a propositional rule technique and basically extracts the rules directly from the data (Kaur and Singh, 2017). JRip which is a bottom-up method learns rules by treating particular judgement of the examples in the training data as a class and finding the set of rules covering all the members of the class (Parsania *et al.*, 2014). Classes are assessed incrementally and the initial rules of the class are produced using incremental reduced error (Parsania *et al.*, 2014). JRip treats all the examples in the training data as a class, discovering the rules that cover all the members of that class. Thereafter, it proceeds to the next class and does the same, repeating this until all classes have been covered (Kaur and Singh, 2017). The JRip algorithm progresses through four phases: (i) growth, (ii) pruning, (iii) optimisation, and (iv) selection (Parsania *et al.*, 2014; Kaur and Singh, 2017).

3.3. *Apriori algorithm*

The problem of finding association rules from data is called the "market-basket problem" (Agrawal *et al.*, 1993a, 1993b). Basically, a set of items and a collection of transactions (baskets) are given. The goal is to find relationships between the various items within those baskets. The Apriori algorithm is the most commonly used Association Rule algorithm (Agrawal *et al.*, 1993a, 1993b). The algorithm utilises a level-wise search, in which k-itemsets are used to find $(k+1)$ itemsets (Győrödi *et al.*, 2017). In this algorithm, frequent itemsets are expanded one item a time (candidate generation process). Those candidates are tested against the data we have. Apriori uses breadth-first search method and a hash tree structure. It defines the frequent items in the dataset and expands them to larger itemsets (Kumbhare and Chobe, 2014).

4. Research Method

To achieve the goal of this research, we followed the data mining approach used in Rubini and Eswaran (2015). However, other algorithms to classify CKD dataset were used. In addition, rule information was extracted using association rule mining as data mining technique. The overall framework of research method is illustrated in Fig. 1.

Next, our method is presented in four subsections starting with the CKD dataset characteristics section. In the second subsection, some functions as preprocessing step for raw data to produce relevant data were applied. After that, five classification

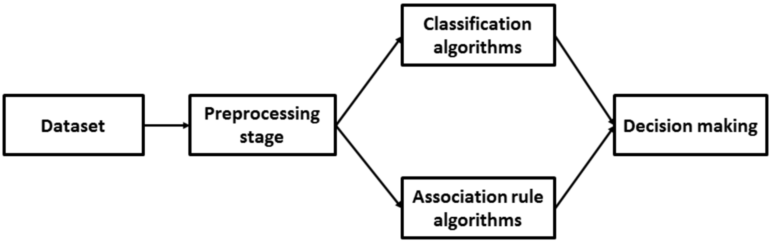

Fig. 1. An overall framework of research method.

algorithms were used, namely, NB, J48, SVM, KNN, and JRip to predict and diagnose CKD. Finally, Apriori algorithm was applied to discover strong association rule between attributes in the last subsection.

4.1. *Chronic kidney disease dataset*

In this research, online dataset from UCI website (UCI, n.d.; Rubini and Eswaran, 2015) was used. This dataset was taken for patients from hospital in India nearly over 2 months. In addition, this dataset was real data and contains 25 attributes for each patient but unfortunately has missing values. Table 1 shows the dataset characteristics. The dataset is used to classify the patients into infected with CKD or not.

CKD dataset is imbalanced data because it contains 250 patients infected with CKD and 150 healthy patients of CKD. Therefore, the class values are CKD for 250 instances and notCKD for 150 ones. In addition, the dataset contains a lot of missing

Table 1. Dataset characteristics.

Attribute characteristics	Real	Data tasks	Classification
# of Attributes	25	# of Instances	400
# of Numeric attributes	11	# of Nominal attributes	14

Table 2. Missing values for each attribute.

Attribute name	Attribute type	# Missing values	Percentage (%)
Age	Numerical	9	02
Blood pressure	Numerical	12	03
Specific gravity	Nominal	46	12
Albumin	Nominal	47	12
Sugar	Nominal	46	12
Red blood cells	Nominal	152	38
Pus cell	Nominal	152	38
Pus cell clumps	Nominal	4	01
Bacteria	Nominal	4	01
Blood glucose random	Numerical	44	11
Blood UREA	Numerical	19	05
Serum creatinine	Numerical	17	04
Sodium	Numerical	87	22
Potassium	Numerical	88	22
Haemoglobin	Numerical	52	13
Packed cell volume	Numerical	71	18
White blood cell count	Numerical	106	27
Red blood cell count	Numerical	131	33
Hypertension	Nominal	2	01
Diabetes mellitus	Nominal	2	01
Coronary artery disease	Nominal	2	01
Appetite	Nominal	1	00
Pedal Oedema	Nominal	1	00
Anaemia	Nominal	1	00

values in each attribute. Table 2 shows missing values for each attribute. We used two techniques to fix missing value that are explained in preprocessing subsection.

4.2. *Data preprocessing*

Data preprocessing is the process in which we convert raw data into relevant data. It is one of the most important steps in data mining process because it affects directly the data mining results. Data preprocessing techniques include: data integration, data cleaning, data reduction, data transformation, and data discretisation (Agrawal *et al.*, 1993a, 1993b).

Chronic Kidney dataset from UCI repository suffers from missing values, as shown in Table 2. In this work, data cleaning was applied to fill all missing values, remove noise, and correct inconsistencies on chronic kidney dataset by Weka (Weka 3).

The data cleaning in Weka can be conducted by using the attribute filter which replaces missing values. This filter replaces all missing values for nominal attribute by putting mode value. Further, for continuous values replacement filter used the mean value for each numeric attribute.

Association rule data mining technique only performed on categorical data; therefore, it requires applying discretisation on numeric attributes. There are 11 such attributes in CKD dataset (age, blood pressure, blood glucose random, blood urea, serum creatinine, sodium, potassium, haemoglobin, packed cell volume, white blood cell count, and red blood cell count). Weka was used to apply discretisation on these numeric attributes. Each of these attributes was divided into bins or intervals using discretisation filter. Furthermore, discretisation filter can do so by using various statistical techniques to automatically decide the best method of binning the data. In this case, simple binning was performed.

4.3. *Classifiers used*

Five classification algorithms were applied to classify CKD dataset into two classes: ckd and notckd. Weka framework was used to perform this task. Weka's classifiers have been classified into different groups such as Bayes, Functions, Lazy, etc. We used KNN from Lazy group, NB from Bayes group, SVM from functions, Decision Tree from trees group, and JRip from rule group. These different machine learning classifiers were used because these algorithms have highest accuracy and well perform on medical data (Bala and Kuma, 2014), and these machine learning classifiers have high performance compared with other classifiers (Rubini and Eswaran, 2015).

CKD dataset was divided into training data and testing data utilising cross-validation approach (Weka). The 10-fold cross-validation was used in order to get accurate test results for all data. Cross-validation is a statistical method of assessing learning algorithms by dividing data into two segments: model training data and model validating data. In 10-fold cross-validation, the data is first partitioned into 10 equally sized parts. Subsequently, 10 iterations of training and validation are conducted such that within each iteration a different part or fold of the data is kept for validation while the remaining nine folds are used for learning (Weka). The following five measures were applied: accuracy, sensitivity, specificity, precision, and mean absolute error (MAE). Equations (1)–(5) were used to calculate these measures values:

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{FP} + \text{TN} + \text{FN}), \tag{1}$$

$$\text{Sensitivity} = \text{TP}/(\text{TP} + \text{FN}), \tag{2}$$

$$\text{Specificity} = \text{TN}/(\text{FP} + \text{TN}), \tag{3}$$

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP}), \tag{4}$$

$$\text{MAE} = (\text{FP} + \text{FN})/(\text{TP} + \text{FP} + \text{TN} + \text{FN}). \tag{5}$$

Confusion Matrix was generated as shown in Table 3. True positive (TP) is number of correctly predicted for positive class (ckd class). False positive (FP) is number of wrongly predicted for positive class (ckd class). True negative (TN) is number of correctly predicted (notckd class). False negative (FN) is number of wrongly predicted (notckd class).

4.4. *Association rule algorithm used*

Apriori algorithm on CKD dataset was applied to generate rules. Before rule generation, the correlation between attributes using Weka tool by applying

Table 3. Confusion matrix.

	Confusion matrix	
Predicted positive (CKD)	Actual positive (CKD) TP	Actual negative (NotCKD) FP
Predicted positive (CKD)	TP	FP
Predicted negative (NotCKD)	FN	TN

Table 4. Rank of correlation between each attribute and class.

Attribute	Ranked	Attribute	Ranked
Hypertension	0.5904	Pus cell clumps	0.2653
Diabetes mellitus	0.5591	Red blood cell count	0.2440
Albumin	0.4770	Coronary artery disease	0.2361
Appetite	0.3933	Serum creatinine	0.2116
Pus cell	0.3752	Blood pressure	0.1966
Pedal Oedema	0.3752	Bacteria	0.1869
Specific gravity	0.3505	Blood glucose random	0.1799
Packed cell volume	0.3359	Blood urea	0.1737
Anaemia	0.3254	White blood cell count	0.1349
Haemoglobin	0.3249	Age	0.1145
Sugar	0.3009	Sodium	0.0912
Red blood cells	0.2826	Potassium	0.0671

"Correlation AttributeEval" was investigated. Table 4 shows the rank of correlation between each attribute and class.

Twelve attributes were selected that have the highest rank to generate strong association rules. The main goal of attributes selection is to extract useful information from CKD dataset and transform it into rules that can be helpful to predict the class value of unclassified examples easily and accurately. There are different metrics that can be used to measure the strength of the association rule. The following are some of these metrics:

- Support: The ratio of transactions which contains a certain itemset.
- Confidence: The percentage of the transactions that contains items X and Y.
- Lift: The proportion of the support to the expected if X and Y were independent.
- Conviction: The percentage of the expected frequency that X occurs without Y divided by the observed frequency of wrong predictions.
- Leverage: The number of counting obtained from the co-occurrence of the antecedent and consequent of the rule from the expected value.

The following are the corresponding equations for the above metrics:

$$\text{Support} = (P(X,Y))/(|T|), \tag{6}$$

$$\text{Confidence} = (P(X,Y))/(P(X)), \tag{7}$$

$$\text{Lift} = (P(X,Y))/(P(X) \cdot P(Y)), \tag{8}$$

$$\text{Conviction} = (P(X)P(\text{not}Y))/(P(X,Y)), \tag{9}$$

$$\text{Leverage} = P(X,Y) - P(X)P(Y). \tag{10}$$

5. Results

5.1. *Classification results*

In the experiment, five different classifiers were applied, namely, NB, J48, SVM, KNN, and JRip on CKD dataset. The dataset was divided into training and testing data using 10-fold cross-validation. Finally, five measures were calculated for each classification method based on Eqs. (1)–(5).

Table 5 shows the measures result for each classification algorithm. As shown in the table, the experiments show that KNN classifier algorithm achieved the highest classification accuracy (about 98.00%) when we assume K value 1.

Table 5. Measure matrix accuracy.

Classifier	Sensitivity (%)	Specificity (%)	Precision (%)	MAE (%)	Accuracy (%)
NB	99.56	87.65	91.60	5.50	94.50
DT(J48)	97.21	95.97	97.60	3.25	96.75
SVM	100.00	94.34	96.40	2.25	97.75
KNN	99.59	96.75	98.00	1.50	98.50
JRip	96.42	95.27	97.20	4.00	96.00

Table 6. KNN accuracy with different K values.

K-values	1	3	5	7	9
Accuracy	98.5%	97.75%	97.5%	97.25%	96.75%

Table 6 shows the accuracy of KNN algorithm with different K-values. As K value increases, the accuracy decreases accordingly. In addition, by applying JRip classifier, we gained two strong rules with 96.00% accuracy. In case of equality, we should care of the last part of the condition whether al $= 0$ or not. Only one rule should be fired based on the value of al.

Rule 1: $\{\text{hemo} >= 13.1, \text{bgr} <= 140, \text{al} = 0\} => \{\text{class} = \text{notckd}\}$.
Rule 2: $\{\text{hemo} <= 13.1, \text{bgr} >= 140, \text{al}! = 0\} => \{\text{class} = \text{tckd}\}$.

5.2. *Association rule result*

Twelve attributes were selected and Apriori algorithm was applied to extract strong rules based on Lift matrix. Lift as the measurement with minimum value equal to 1.5 was used. Furthermore, the rule x determines y was defined. X contains attribute set: {hypertension(htn), diabetes mellitus(dm), albumin(al), appetite(appet), pus cell(pc), pedal Oedema(pe), specific gravity(sg), packed cell volume(pcv), anaemia (ane), haemoglobin(hemo), sugar(su), and red blood cells(rbc)}. Y contains the class values: ckd and notckd.

Five strong rules generated using Apriori algorithm:

Rule 1: {al:0, su:0, pc:normal, htn:no, dm:no, appet:good, pe:no}=={class:notckd}.
Rule 2: {al:0, pc:normal, htn:no, dm:no, appet:good, pe:no, ane:no}=={class: notckd}.
Rule 3: {al:0, su:0, rbc:normal, pc:normal, htn:no, dm:no, appet:good}=={class: notckd}.
Rule 4: {htn:yes, dm:yes}=={class:ckd}.
Rule 5: {htn:no, dm:no}=={class=notckd}.

Table 7 shows the measure of the strength of the association rule.

Confidence represents the percentage of transactions that include both antecedent and consequent to the transactions that include only antecedent. Its range falls

Table 7. Measures of strength of each rule.

Rule #	Confidence (%)	Lift	Conviction	Leverage
Rule 1	90	2.4	5.8	0.22
Rule 2	89	2.38	5.53	0.22
Rule 3	89	2.37	5.28	0.22
Rule 4	100	1.6	39.75	0.10
Rule 5	68	1.8	1.9	0.17

in [0, 1]. If confidence = 1 then most interesting, if confidence = 0 then least interesting. Lift predicts the performance of an association rule to enhance response. Its range is [0, ∞]. If $0 <$ lift < 1 then $X \to Y$ is interdependent negatively, if lift=1 then interdependent, if $\infty >$ lift > 1 then $X \to Y$ is interdependent positively. Conviction addresses the limitation of confidence and lift. It evaluates the degree of implication of a rule. Its range is [0.5, ∞]. If conviction = 1 then rule is independent, if conviction > 1 then interesting rules. The range of leverage falls in $[-0.25, 0.25]$. If leverage = 0.25 then most interesting, If $leverage = -0.25$ then least interesting (Deora *et al.*, 2013).

6. Discussion

According to the World Health Organisation, there were approximately 58 million deaths in the world in 2015, with 35 million because of chronic disease. CKD is one of the deadly diseases that can affect a lot of vital organs in the human body such as heart, liver, and lungs. Previous research showed that early prediction of CKD is very significant in the medical field for physicians' decision-making, patients' health and life, and reducing the overall healthcare cost. Limited research had focused on constructing an efficient prediction system for CDK using an integrative approach by combining several techniques together for predicting and diagnosing CKD. Thus, this research aims to propose an efficient system for predicting and diagnosing CKD utilising an integrative approach. In particular, classification and association rule mining techniques were integrated and utilised to construct an efficient system for predicting and diagnosing CKD and its causes.

An integrative approach helps in identifying large data sets. The presented experiments shows that integration of association rule mining and classification technique yields more accurate results than simple classification method. In addition, it is very useful in generating rules when we have missing values in the data set. This integrated technique provides a promising classification result with utmost accuracy rate and robustness for CKD prediction.

The results showed that KNN achieved the highest accuracy followed by SVM. These findings suggest that KNN algorithm is the best for predicting CKD. NB was the least prediction accuracy as shown in our results. The results showed that the accuracy of KNN decreases gradually as K value increases. In terms of precision and specificity, the results showed that KNN is better compared with other classifiers.

Apriori algorithm was applied on 12 attributes to extract strong rules based on lift matrix. Two attributes, namely, hypertension (htn) and diabetes mellitus (dm), have highest ranks of correlation with class attributes. Thus, two rules from them were generated. When hypertension value equals yes and diabetes mellitus value equals yes the class value is ckd. In contrast, if hypertension and diabetes mellitus values are no then class value is nockd.

Comparing to existing research's results (Piatetsky-Shapiro, 1991; Jena and Kamila, 2015), our results outperformed their results and showed a higher accuracy applying our proposed integrative approach. For instance, Jena and Kamila (2015) used weka to classify CKD using six classification algorithms, namely, NB, MLP, SVM, J48, decision tree, conjunctive rule, and decision table classifier. Their research concluded by showing that all algorithms gave more than 90% classification accuracy (e.g. MLP gave 99.75%), except SVM which gave 62% accuracy. In our approach, SVM gave more than 95% accuracy, and KNN gave the highest accuracy of 98%. In addition, NB and ANN algorithms have been used to diagnose CKD using rapidminer tool (Kalaiselvi and Nasira, 2015). However, our integrative approach achieved a higher prediction accuracy as mentioned earlier.

Confirming our major contribution, the integrative approach achieved a higher accuracy than that of a single approach; several previous researches applied the integrative approach but in different domains or using different algorithms. Sunita and Lobo (2012) applied three data mining techniques: clustering (K-means algorithm), classification (alternating decision tree algorithm), and association rule (Apriori algorithm) on a dataset that contains data about courses learnt by students. The results showed that the use of integrated approach is better than applying only association rule mining. Moreover, clustering and association rule mining (Ganda, 2013) have been integrated on kidney dataset that contains 7 attributes and 157 instances to form strong rules in each cluster by weka. However, in our approach, we integrated classification algorithms with Apriori algorithm for CKD dataset. Overall, our findings confirmed the findings of previous research and showed that such integrative approach improves the prediction accuracy of CKD.

Based on the results of our approach, we conclude that the proposed integrative approach gives satisfactory results that are superior to single classifier in CKD prediction. Due to the complexity and high mortality of CKD, a diagnosis in timely and accurate manner is crucial. Thus, improving the prediction accuracy by applying integrative techniques is of great help to CKD treatment. Such a prediction system helps doctors to diagnose and suggest the treatment at an early stage of CKD. It also enables the patients at early stage of CKD to have necessary knowledge about their health status and follow appropriate diet and prescriptions.

7. Conclusion and Future Work

CKD is one of the diseases that may lead to death. This paper proposes building a strong prediction system to discover kidney diseases, before the kidneys are damaged totally, that helps to reduce cost of treatments, reduce risks of human's errors, and save patients' life.

CKD dataset was used to predict and diagnose patient case. Preprocessing for CKD dataset was conducted to produce relevant data. Furthermore, five algorithms were applied: NB, J48, KNN, SVM, and JRip for classification task. Experiments showed that KNN gives 98.50% as highest accuracy. Apriori algorithm was used to extract strong rules based on Lift matrix. Five different rules from 12 attributes that have high ranks of correlation with class attribute were produced.

This research has several implications for practice. It builds an efficient prediction system with a high accuracy for CKD. With early diagnosis and treatment, it is possible to slow the CKD progression. Early prediction of CKD is very significant in the medical field for physicians' decision-making, patients' health and life, and reducing the overall healthcare cost. By using data mining techniques, researchers have the scope to predict the CKD. This helps physicians to diagnose and suggest the appropriate treatment at an early stage. In addition, it enables the patients at early time to have knowledge about their health status and follow appropriate diet and prescriptions.

The research has some limitations; the method does not generalise to other classification methods, it has trouble dealing with continuous variables, and it cannot completely solve the challenge of working with large datasets. In future, we plan to extend this research using other classifier methods or apply FP-growth algorithm to generate association rule mining.

References

Afhami, N (2018). Prediction of diabetic chronic kidney disease progression using data mining techniques. *International Journal of Computer Science Engineering*, 7(2), 35–40.

Agrawal, R, T Imielinski and A Swami (1993a). Database mining: A performance perspective. *IEEE Transactions on Knowledge and Data Engineering*, 5(6), 914–925. doi: 10.1109/69.250074.

Agrawal, R, T Imielinski and A Swami (1993b). Mining association rules between sets of items in large databases. In *Proceedings of ACM SIGMOD International Conference on Management of Data*, Washington, D.C.: ACM, pp. 207–216.

Akiyama, M and K Fujita (2013). How to improve patient safety by text mining with medical incident reports: Innovative technologies using e-Health and Health Technology Assessment. In *Proceedings of PICMET '13 Technology Management in the IT-Driven Services*, San Jose, CA: IEEE.

Aqlan, F, R Markle and A Shamsan (2017). Data mining for chronic kidney disease prediction. In *Proceedings of the 2017 Industrial and Systems Engineering Conference*, Pittsburgh, PA: Institute of Industrial and Systems Engineers, pp. 1789–1794.

Arasu, D and R Thirumalaiselvi (2017). Review of chronic kidney disease based on data mining techniques. *International Journal of Applied Engineering Research*, 12(23), 13498–13505.

Arora, J, Shelza and S Rao (2013). An efficient ARM technique for information retrieval in data mining. *International Journal of Engineering Research and Technology*, 2(10), 1079–1084.

Bala, S and Kuma (2014). A literature review on kidney disease prediction using data mining classification technique. *International Journal of Computer Science and Mobile Computing*, 3(7), 960–967.

Basar, MD and A Akan (2018). Chronic kidney disease prediction with reduced individual classifiers. *Electrica*, 18(2), 249–255. doi: 10.26650/electrica.2018.99255.

Bay, V and B Le (2008). A novel classification algorithm based on association rule mining. In *The 2008 Pacific Rim Knowledge Acquisition Workshop (Held with PRICAI08), LNAI*, Ha Noi, Viet Nam.

Boukenze, B, A Haqiq and H Mousannif (2017). Predicting chronic kidney failure disease using data mining techniques. In *Advances in Ubiquitous Networking: Lecture Notes in Electrical Engineering*, Vol. 397. Singapore: Springer.

Boukenze, B, H Mousannif and A Haqiq (2016). Performance of data mining techniques to predict in Healthcare Case Study: Chronic kidney failure disease. *International Journal of Database Management Systems*, 8(3), 1–9. doi: 10.5121/ijdms.2016.8301.

Brossette, SE, AP Sprague, JM Hardin, KB Waites, WT Jones and SA Moser (1998). Association rules and data mining in hospital infection control and public health surveillance. *Journal of the American Medical Informatics Association*, 5(4), 373–381. doi: 10.1136/jamia.1998.0050373.

Chandrakar, I and A Kirthima (2013). A survey on association rule mining algorithms. *International Journal of Mathematics and Computer Research*, 1(10), 270–272.

Deora, C, S Arora and Z Makani (2013). Comparison of interestingness measures: Support-confidence framework versus Lift-Irule framework. *International Journal of Engineering Research and Applications*, 3(2), 208–215.

Duan, L, WN Street and E Xu (2011). Healthcare information systems: Data mining methods in the creation of a clinical recommender system. *Enterprise Information Systems*, 5(2), 169–181. doi: 10.1080/17517575.2010.541287.

Ganda, R (2013). Knowledge discovery from database using an integration of clustering and association rule mining. *International Journal of Advanced Research in Computer Science and Software Engineering*, 3(9), 13–18.

Garg, R and V Mongia (2018). A comparative study of different classification algorithms on kidney disease prediction. *International Journal for Research in Applied Science and Engineering Technology*, 6(2), 741–746. doi: 10.22214/ijraset.2018.2132.

Gautam, P and K Pardasani (2010). Algorithm for efficient multilevel association rule mining. *International Journal on Computer Science and Engineering*, 2(5), 1700–1704.

Gopika, S and M Vanitha (2017). Efficiency of data mining techniques for predicting kidney disease. *International Journal of Engineering and Technology*, 9(5), 3586–3591. doi: 10.21817/ijet/2017/v9i5/170905314.

Győrödi, C, R Győrödi and S Holban (2017). A comparative study of association rules mining algorithms. *SACI*, pp. 1–11.

Jena, L and N Kamila (2015). Distributed data mining classification algorithms for prediction of chronic kidney disease. *International Journal of Emerging Research in Management and Technology*, 4(11), 110–118.

Kaladha, D, K Rayavarapu and V Vadlapudi (2012). Statistical and data mining aspects on kidney stones: A systematic review and meta-analysis. *Open Access Scientific Reports*, 1(12), 1–542. doi: 10.4172/scientificreports.543.

Kalaiselvi, C and GM Nasira (2015). Prediction of heart diseases and cancer in diabetic patients using data mining techniques. *Indian Journal of Science and Technology*, 8(14). doi: 10.17485/ijst/2015/v8i14/72688.

Kaur, G and A Sharma (2017). Predict chronic kidney disease using data mining algorithms in Hadoop. In *2017 International Conference on Inventive Computing and Informatics (ICICI)*, Coimbatore: IEEE, pp. 973–979. doi: 10.1109/ICICI.2017.8365283.

Kaur, G and A Sharma (2018). Predict chronic kidney disease using data mining algorithms in Hadoop. *International Journal of Advances in Electronics and Computer Science*, 5(4), 6–13.

Kaur, G and J Singh (2017). Classification of malicious Urls for web using Ripper algorithm. *International Journal of Advanced Research in Computer Science*, 8(7), 137–139. doi: 10.26483/ijarcs.v8i7.4141.

Kumbhare, T and S Chobe (2014). An overview of association rule mining algorithms. *International Journal of Computer Science and Information Technologies*, 5(1), 927–930.

Kunwar, V, K Chandel, AS Sabitha and A Bansal (2016). Chronic kidney disease analysis using data mining classification techniques. In *2016 6th International Conference — Cloud System and Big Data Engineering (Confluence)*, pp. 300–305.

Lee, C, C Wu and H Yang (2007). Text mining of clinical records for cancer diagnosis. In *Second International Conference on Innovative Computing, Information and Control*, Kumamoto, Japan: IEEE, p. 172. doi: 10.1109/ICICIC.2007.556.

Li, W, J Han and J Pei (2001). CMAR: Accurate and efficient classification based on multiple class association rules. In *Proceedings of IEEE International Conference on Data Mining*, San Jose, CA, pp. 369–376.

Maeda, T, Y Fukushige and M Yajima (2016). Text mining analysis for E Health Information System. In *2015 17th International Conference on E-health Networking, Application & Services (HealthCom)*. Boston, MA: IEEE. doi: 10.1109/HealthCom.2015.7454472.

Moreno, M, S Segrera and V López (2005). Association rules: Problems, solutions and new applications. *ActasdelIIITallerNacionaldeMineríadeDatosyAprendizaje*, pp. 317–323.

Nahar, J, K Tickle, S Ali and Y Chen (2009). Diagnosis heart disease using an association rule discovery approach. In *The IASTED International Conference Computational Intelligence (CI 2009)*, Honolulu, Hawaii.

National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) (n.d.). Available at https://www.niddk.nih.gov/. Accessed on 5 February 2019.

National Kidney Foundation (n.d.). Available at http://www.kidney.org/. Accessed on 20 April 2019.

Otunaiya, K and G Muhammad (2019). Performance of datamining techniques in the prediction of chronic kidney disease. *Computer Science and Information Technology*, 7(2), 48–53. doi: 10.13189/csit.2019.070203.

Parsania, V, N Jani and N Bhalodiya (2014). Applying Naïve Bayes, BayesNet, PART, JRip and OneR algorithms on hypothyroid database for comparative analysis. *International Journal of Darshan Institute on Engineering Research and Emerging Technology*, 3(1), 60–64.

Pendyala, S, Y Fang, J Holliday and A Zalzala (n.d.). A text mining approach to automated healthcare for the masses. In *IEEE Global Humanitarian Technology Conference* (*GHTC* 2014). San Jose, CA: IEEE. doi:10.1109/GHTC.2014.6970257.

Piatetsky-Shapiro, G (1991). Discovery, analysis, and presentation of strong rules. *Knowledge Discovery in Databases*.

Refaeilzadeh, P, L Tang and H Liu (2009). Cross-Validation. In *Encyclopedia of Database Systems*. L Liu and MT Özsu (eds.). Boston, MA: Springer.

Rubini, L and P Eswaran (2015). Generating comparative analysis of early stage prediction of Chronic Kidney Disease. *International Journal of Modern Engineering Research*, 5(7), 49–55.

Sharma, P and G Kaur (2018). Review on data mining techniques for prediction of chronic kidney disease. *International Journal of Engineering Trends and Technology*, 63(1), 58–60. doi: 10.14445/22315381/ijett-v63p209.

Sunita, B and L Lobo (2012). Combination of clustering, classification & association rule based approach for course recommender system in e-learning. *International Journal of Computer Applications*, 39(7), 8–15. doi: 10.5120/4830-7087.

Tsanas, A, MA Little, PE Mcsharry, J Spielman and LO Ramig (2012). Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease. *IEEE Transactions on Biomedical Engineering*, 59(5), 1264–1271. doi: 10.1109/tbme.2012.2183367.

UCI Machine Learning Repository: Chronic Kidney Disease Data Set (n.d.). Available at https://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease. Accessed on November/December 2018.

Weka 3: Machine Learning Software in Java (n.d.). Available at https://www.cs.waikato.ac.nz/ml/weka/. Accessed on 8 July 2018.

Zeynu, S and S Patil (2018). Survey on prediction of chronic kidney disease using data mining classification techniques and feature selection. *International Journal of Pure and Applied Mathematics*, 118(18), 149–156.