Selecting Accurate Classifier Models for a MERS-CoV Dataset

Afnan AlMoammar(✉), Lubna AlHenaki, and Heba Kurdi

Computer Science Department, KSU, KSA Riyadh, Saudi Arabia
{437203909, 437204268}@student.ksu.edu.sa,
hkurdi@ksu.edu.sa

Abstract. The Middle East Respiratory Syndrome Coronavirus (MERS-CoV) is a viral respiratory disease that is spreading worldwide necessitating to have an accurate diagnosis system that accurately predicts infections. As data mining classifiers can greatly assist in enhancing the prediction accuracy of diseases in general. In this paper, classifier model performance for two classification types: (1) binary and (2) multi-class were tested on a MERS-CoV dataset that consists of all reported cases in Saudi Arabia between 2013 and 2017. A cross-validation model was applied to measure the accuracy of the Support Vector Machine (SVM), Decision Tree, and k-Nearest Neighbor (k-NN) classifiers. Experimental results demonstrate that SVM and Decision Tree classifiers achieved the highest accuracy of 86.44% for binary classification based on healthcare personnel class. On the other hand, for multiclass classification based on city class, the decision tree classifier had the highest accuracy among the remaining classifiers; although it did not reach a satisfactory accuracy level (42.80%). This work is intended to be a part of a MERS-CoV prediction system to enhance the diagnosis of MERS-CoV disease.

Keywords: Data mining · Medical data · Classification · Classifier model MERS-CoV · Accuracy measurement · Cross-validation model

1 Introduction

Middle East respiratory syndrome (MERS) is a viral respiratory disease that spread over 27 countries around the world. The disease was caused by a novel coronavirus called the Middle East respiratory syndrome coronavirus (MERS-CoV). Moreover, coronaviruses are a large family of viruses responsible for causing many diseases, from mild colds to Severe Acute Respiratory Syndrome (SARS). MERS-CoV is one of the most common major causes for the increase in mortality among children and adults in the world [1]. The first identification of MERS-CoV was in Saudi Arabia in 2012. It spread rapidly in Saudi Arabia and many other countries and caused a large number of deaths [2]. Therefore, early diagnosis of MERS-CoV infection may help to control the outbreak of the virus and reduce human suffering. Computer and data mining techniques can provide great help in analyzing, diagnosing, and predicting diseases, and they can assist in controlling virus infection [3].

K. Arai et al. (Eds.): IntelliSys 2018, AISC 868, pp. 1070–1084, 2019.
https://doi.org/10.1007/978-3-030-01054-6_74

Using data mining techniques in diagnosis and prediction of diseases has been developing fast over the last few decades. Data mining is the process of analyzing a large amount of complex data to find useful patterns and extract hidden information by applying machine learning algorithms [4]. In healthcare, the generated data is vast and too complex to be analyzed and processed by traditional methods. Due to this, the need for data mining in healthcare is becoming essential. Accordingly, data mining has been widely used in healthcare, including outcomes prediction, treatment effectiveness evaluation, infection control, and disease diagnosis [3]. Moreover, studies on using data mining in healthcare show that it succeeds in helping to improve diagnostic accuracy prediction and predicting health insurance fraud, which lightens the burden of increasing workloads, and reducing healthcare costs [5].

Recently, various types of data mining methods have been applied by a number of researchers [6, 7], using real MERS-CoV datasets based on several types of machine learning classifiers. MERS is a complex disease caused by MERS-CoV that spreads easily and has a high death rate; approximately 40% of patients diagnosed with MERS have died [1]. The challenge remains to provide prediction systems that accurately anticipate and diagnose MERS-CoV. Prediction systems are primarily motivated by the necessity of achieving maximum possible accuracy. Our motivation for this study is to utilize data mining techniques in order to control the spreading of MERS-CoV and to save people's lives. Motivated by the above needs, we make the following contribution in the application of classification algorithms to a MERS-CoV dataset for identifying the accurate classifier.

The main contribution of this study is to apply a support vector machine classifier beside two other classifiers to assess the classification accuracy on MERS-CoV dataset. Whilst the previous studies used datasets consisting of information about MERS-CoV cases only up to 2015, our dataset covers all affected cases in Saudi Arabia from 2013 to 2017.

The remaining parts of this paper are organized as follows: The literature review is introduced in Section 2. Then, the system design and implementation are presented in Section 3. The methodology is then described in Section 4. After that, the results and discussion are detailed in Section 5. Finally, the conclusions and directions for future work are discussed.

2 Literature Review

One of the early applications of data mining techniques was in medical areas where it could help in predicting and diagnosing diseases and support medical decision making. Several researchers have been working in data mining application and experimental use of medical datasets. This review will go through some of the related work in healthcare, but it is not meant to be exhaustive. The first part of the literature review introduces some applications of classification algorithms on different medical datasets. The second part is a review of the related works of the MERS-CoV diagnosis and prediction using data mining techniques.

For instance, the researchers in [8], apply data mining on historical health records to improve the prediction of chronic disease. In this study, two datasets from UC Irvine (UCI) repository are considered: heart disease and diabetes. Many data mining algorithms are applied, including: Naïve Bayes, Decision Tree, Support Vector Machine (SVM), and Artificial Neural Networks (ANN). From the experiment, SVM performs better than the other classifiers on the heart disease dataset, while Naïve Bayes classifier achieves the highest accuracy on the diabetes dataset.

A recent study [9] uses data mining to increase the diagnosis of neonatal jaundice in newborns. The dataset consists of records of healthy newborn infants with 35 or more weeks of gestation collected from the Obstetrics Department of the Centro Hospital. Several data mining algorithms are applied to the dataset: Decision Tree, CART, Naïve Bayes, Artificial Neural Networks, SVM, and Easy Logistic algorithms. The results of this study show that the most effective predictive models are Naïve Bayes, Neural Networks, and Easy Logistic algorithms.

The researchers in [10] compare different data mining algorithms to find the most efficient and effective algorithm in terms of accuracy, sensitivity, and precision. An experiment is conducted using an original Wisconsin Breast Cancer dataset from the UCI machine learning repository with four classifiers: SVM, Naïve Bayes, Decision Tree, and k-Nearest Neighbor (k-NN). The effectiveness of all classifiers is evaluated in terms of time to build the model, correctly classified instances, incorrectly classified instances, and accuracy. The results show that SVM is the most efficient classifier in Breast Cancer prediction and diagnosis with high precision and low error rate.

Another study [11], applies different machine learning algorithms on artificial lung cancer datasets systematically collected by the Hospital Information System in order to explore the advantages and disadvantages of each algorithm. Many experiments are conducted on the dataset using the following machine learning algorithms: Decision Tree, Bagging, Adaboost, SVM, k-NN, and Neural Network. The results show that, according to the high accuracy of these algorithms, Adaboost and Neural Network are suitable for this type of cancer analysis.

The researchers in [12] compare two classification algorithms: Decision Trees and Random Forest with Self-Organizing Map (SOM) to build a predictive model for diabetic patients. The dataset uses in this study is collected from the Hospital Information System of the Ministry of National Guard Health Affairs (MNGHA), Saudi Arabia, between 2013 and 2015. The authors found that the Random Forest algorithm achieves the highest recall and precision.

The authors in [13] introduce a MobDBTest Android mobile application. MobDBTest uses machine learning techniques to predict diabetes levels for the users. The propos Android mobile application is tested on real dataset collected from a reputed hospital in the Chhattisgarh state of India. Four machine learning algorithms such as J48, Naïve Bayes, SVM and Multilayer Perceptron are used to classify the collected data. The results show that J48 algorithm outperformed other methods in terms of sensitivity, specificity and ROC areas.

During the past six years, more information about the MERS-CoV disease has become available to the public. MERS-CoV is a well-known virus that is still rapidly growing. Finding the accurate classifier can help to improve the prediction accuracy of MERS-CoV infection. The study in [7] applies data mining techniques to a

MERS-CoV dataset to identify the accurate classifier models of binary, multi-class, and multi-label classification. The dataset includes all MERS-CoV cases in Saudi Arabia from the Saudi Ministry of Health from 2013 to the second half of 2016. Three classifier models are built using k-NN, Decision Tree, and Naïve Bayes algorithms. The outcome of this research is that the Decision Tree is the most accurate algorithm for the binary-class classification, whereas k-NN is the most accurate algorithm for the multi-class classification. Additionally, for the multi-label classification the Naïve Bayes is the most accurate algorithm.

Another related study [6], involves experimental data mining to build prediction models for MERS-CoV. The experiments are conducted on a dataset collected from the Saudi Ministry of Health. It consists of MERS-CoV cases between 2013 and 2015. The Naive Bayes and Decision Tree algorithms are used to develop recovery and stability predictive models based on the MERS-CoV dataset. The results of recovery models indicate that healthcare workers are more likely to survive. Moreover, symptoms and age are important attributes for predicting stability in stability models. In general, Decision Tree has better accuracy over all models.

The researchers in [14] propose a molecular approach to analyze DNA sequences of MERS-CoV to draw the route of transmission of MERS-CoV from Saudi Arabia to the world. Full DNA sequences that are collected from 15 different regions from the National Center for Biotechnology Information (NCBI) are converted into amino acid sequences to be used in the analysis process. Moreover, the proposed approach uses Apriori and Decision Tree algorithms to find the similarities and differences between different amino acid sequences. Relevance between several sequences is found using Decision Tree algorithm.

The study described in [15] proposes a cloud-based MERS-CoV prediction system to predict and prevent MERS-CoV infection spread between citizens and regions. The dataset consists of patients, medicines, and reports of each user. It is stored in multiple clouds known as a medical record (M.R.) database. In addition, this system is based on a statistical classifier in data mining, which is a Bayesian classification algorithm for initial classification of the patient base on predicting class membership probabilities. The outcome of this study is a prediction of MERS-CoV-infected regions on Google Maps with high accuracy in the classification.

A study [16] applies three data mining algorithms to compare two viruses with similar symptoms: severe acute respiratory syndrome (SARS) and MERS coronaviruses. Apriori, Decision Tree, and SVM data mining algorithms are used on data of the spike in glycoprotein from the NCBI to distinguish between the two viruses. From the experiment, it is clear that distinguishes between MERS and SARS spike glycoproteins with a high accuracy.

Table 1 presents a comparison of literature review that applied data mining techniques on medical data over the different categories. These categories are reference number, used data mining algorithm, used dataset, the objective of the research, used tool, and finally the outcome of the research. From the Table 1, it can be seen that several algorithms and techniques have been applied to medical datasets and that the most common methods for classification are Decision Tree, SVM, and k-NN algorithms.

Table 1. Comparison of relevant literature review

Ref. no.	Data mining techniques	Dataset	Objective	Tool	Outcomes
[8]	Naïve Bayes, Decision Tree, SVM, and ANN	Heart disease, and diabetes datasets	Predicting Chronic disease by mining the data containing historical health records	WEKA	SVM gives highest accuracy rate of 95.55% and Naïve Bayes classifier gives highest accuracy of 73.58% for the heart disease, diabetes respectively
[9]	Decision Tree, CART, Trusting Bayes classifier, neural networks SMO, and easy logistic	Records of Healthy newborn infants with 35 or more weeks of gestation	Improving the diagnosis of neonatal jaundice in newborns	WEKA	The most effective predictive models are Trusting Bayes with 88% accuracy, neural networks with 87% accuracy, and easy logistic with 89% accuracy
[10]	SVM, C4.5, Naive Bayes, and *k*-NN	Wisconsin Breast Cancer (original) dataset	Finding the most efficient algorithm for Breast Cancer prediction and diagnosis	WEKA	The SVM has proven its efficiency in Breast Cancer prediction and diagnosis with 97.13% accuracy
[11]	Decision Tree, Bagging, Adaboost, SVM, *k*-NN, and Neural Network	Artificial lung cancer dataset	Comparing different classification algorithms in order to explore the advantages and disadvantages of each one	RStudio	Adaboost algorithm and neural network algorithm have relative high accuracy with 97.5% accuracy
[12]	Self-Organizing Map (SOM), Decision Tees	Adult population data	Constructing intelligent predictive model for diabetic	RStudio and WEKA	The RandomForest model could assist health care

(*continued*)

Table 1. (*continued*)

Ref. no.	Data mining techniques	Dataset	Objective	Tool	Outcomes
	C4.5, and Random Forest		disease by using real healthcare data		providers with 90% accuracy to make better clinical decisions in identifying diabetic patients
[13]	J48, Naïve Bayes, SVM and Multilayer Perceptron	Reputed hospital in the Chhattisgarh state of India	Predict diabetes levels for the users uses machine learning techniques	Android mobile application	The results show that J48 algorithm outperformed other methods in terms of sensitivity, specificity and ROC areas
[7]	*k*-NN, Decision Tree, and Naïve Bayes algorithms	MERS-CoV cases in Saudi Arabia noted between 2013 and second half of 2016	Identifying accurate classifier modes for binary, multiclass, and multi-label classification of a text-based MERS-CoV dataset	RapidMiner Studio	The accurate algorithm for the Binary-Class classification is Decision Tree with 90% accuracy, for the Multi-Class classification is *k*-NN with 51.60% accuracy, and for the Multi-Label classification is Naïve Bayes with 77% accuracy
[6]	Naive Bayes, and Decision Tree algorithms	1082 records of MERS-CoV cases noted between 2013 and 2015	Building predictive models for MERS-CoV infection to understand which factors contribute to complications of this infection	WEKA	The results show that, Decision Tree classifier has better accuracy of 55.69%, and 68% for the stability and the recovery models respectively

(*continued*)

Table 1. (*continued*)

Ref. no.	Data mining techniques	Dataset	Objective	Tool	Outcomes
[14]	Decision Tree, and Apriori Algorithms	DNA sequences of MERS-CoV outbreak from different regions in the world where the viruses speared	Finding the similarities between different MERS-CoV amino acid sequences to know transmission route of MERS-CoV	Mathematical model	The results show that Riyadh, Makkah, and Buridah regions of MERS-CoV transmission in Saudi Arabia
[15]	BBN classification	Multiple attributes: 1-personal (static), 2-MERS (changes over time)	Identifying an intelligent system for predicting and preventing MERS-CoV infection	R Studio, WEKA, and Amazon EC2	The BBN achieve an accuracy of 83.1% on synthetic data
[16]	Decision Tree, and Apriori Algorithms	DNA sequences of MERS-CoV of outbreak	Finding the similarities between different MERS-CoV amino acid sequences to know transmission route of MERS-CoV	Mathematical model	The results show that Riyadh, Makkah, and Buridah regions of MERS-CoV transmission in Saudi Arabia

In conclusion, in related studies data mining is widely used for the prognoses and diagnoses of many diseases. However, the datasets used in [6, 7] are limited and include the MERS-CoV cases in Saudi Arabia from 2013–2015 only. It is important to increase the size of the dataset to cover new cases. Therefore, this study applied Data mining techniques using Decision Tree, SVM, and k-NN classification algorithms to a real dataset of MERS-CoV cases in the Kingdom of Saudi Arabia that was collected during 2013–2017.

3 System Design and Implementation

The system overview is illustrated in Fig. 1. It shows high-level components of the classification framework. The classification framework is composed of three subsystems, which are the MERS-CoV dataset, supervised learning, and data scientist. The MERS-CoV dataset subsystem aims to collect MERS-CoV data from different sources

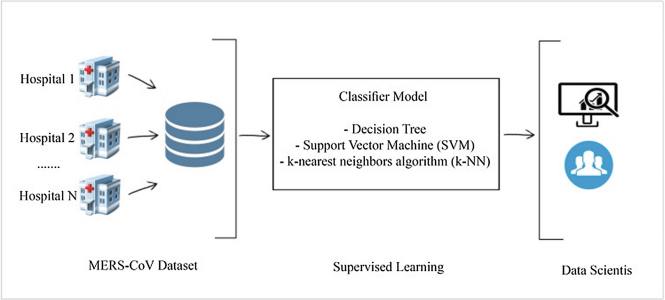


Fig. 1. System overview.

and integrate them into one database. The purpose of the supervised learning subsystem, which is the core of this study, is applying data mining techniques to build three different classifier models. Finally, the third subsystem consists of data scientists who analyze data and evaluate results.

Figure 2 shows the overall workflow of the classification framework, which is divided into two main phases. The first phase aims to collect data of patients who are affected by MERS-CoV from different cities in Saudi Arabia between January 2013 and October 2017. The second phase is the most important phase. Its purpose is to identify the classifier model and evaluate the classification accuracy using cross validation test mode.

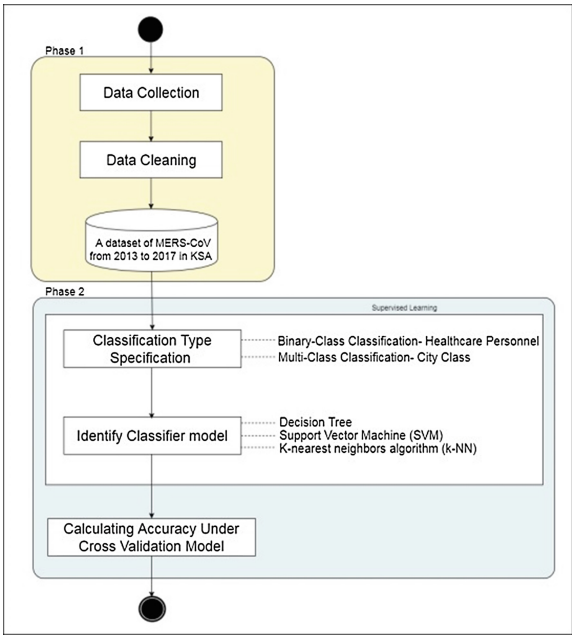


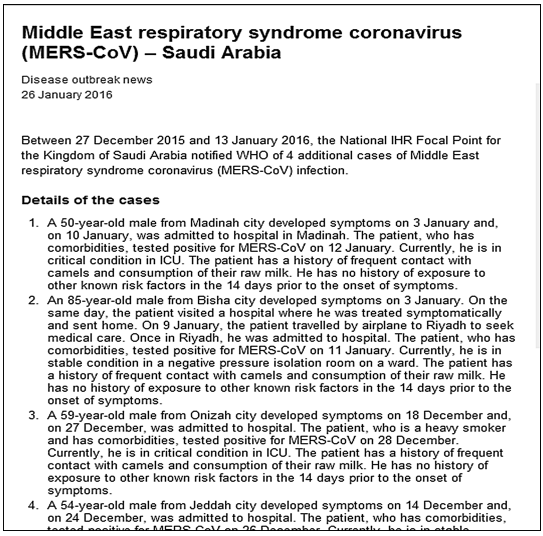
Fig. 2. System workflow.

4 Methodology

4.1 Dataset Description and Pre-processing

As mentioned, the dataset used in this study covers all MERS-CoV cases in Saudi Arabia, including 1,186 alive records and 224 death records, which were reported between 2013 and 2017. The dataset of MERS-CoV cases from 2013–2015 was obtained by a request from [2], the 2016–2017 dataset was collected from the website of the World Health Organization [2]. Moreover, The MERS-CoV dataset consists of the following information about MERS-CoV patients: gender, age, exposure to camels, comorbidities, exposure to MERS-CoV cases, city, and whether the patient is employed in healthcare or not. In addition, the dataset contains information about status to detect whether the patient is alive or dead.

The challenge of building the dataset was that data were published on the website as text description of details of the MERS-CoV cases, represented in Fig. 3, was not promptly usable by any data mining tool. This compelled us to construct the dataset from scratch. Furthermore, all records were prepared in Comma Separated Value (CSV) format, which is appropriate for a data mining tool.



Middle East respiratory syndrome coronavirus (MERS-CoV) – Saudi Arabia

Disease outbreak news
26 January 2016

Between 27 December 2015 and 13 January 2016, the National IHR Focal Point for the Kingdom of Saudi Arabia notified WHO of 4 additional cases of Middle East respiratory syndrome coronavirus (MERS-CoV) infection.

Details of the cases

1. A 50-year-old male from Madinah city developed symptoms on 3 January and, on 10 January, was admitted to hospital in Madinah. The patient, who has comorbidities, tested positive for MERS-CoV on 12 January. Currently, he is in critical condition in ICU. The patient has a history of frequent contact with camels and consumption of their raw milk. He has no history of exposure to other known risk factors in the 14 days prior to the onset of symptoms.
2. An 85-year-old male from Bisha city developed symptoms on 3 January. On the same day, the patient visited a hospital where he was treated symptomatically and sent home. On 9 January, the patient travelled by airplane to Riyadh to seek medical care. Once in Riyadh, he was admitted to hospital. The patient, who has comorbidities, tested positive for MERS-CoV on 11 January. Currently, he is in stable condition in a negative pressure isolation room on a ward. The patient has a history of frequent contact with camels and consumption of their raw milk. He has no history of exposure to other known risk factors in the 14 days prior to the onset of symptoms.
3. A 59-year-old male from Onizah city developed symptoms on 18 December and, on 27 December, was admitted to hospital. The patient, who is a heavy smoker and has comorbidities, tested positive for MERS-CoV on 28 December. Currently, he is in critical condition in ICU. The patient has a history of frequent contact with camels and consumption of their raw milk. He has no history of exposure to other known risk factors in the 14 days prior to the onset of symptoms.
4. A 54-year-old male from Jeddah city developed symptoms on 14 December and, on 24 December, was admitted to hospital. The patient, who has comorbidities,

Fig. 3. A sample of text description of MERS-CoV cases.

In order to enhance the quality of the classification framework, different preprocessing techniques were applied to the MERS-CoV dataset, including replacing missing values and reducing noise values. To handle the missing values, each was replaced with the mean of the attribute that includes missing values. Additionally, the

noise in the dataset appeared due to the existence of inconsistent data in the dataset. For instance, the gender attribute is represented in some instances using the full word "female" or "male," while in other instances it is represented using the abbreviations "F" or "M." So, the inconsistent values were integrated into a standard value which is "F or M." Furthermore, the data were converted from categorical to numerical data because the SVM algorithm deals only with numerical data.

4.2 Data Mining

In a similar approach to [7], in this study, the classifier model performance was examined for two classification types: binary classification based on the healthcare personnel class, and multi-class classification based on the city class. The SVM, Decision Tree, and k-NN classification algorithms were chosen because they are commonly used for medical mining as presented in the Literature Review section of this paper. In addition, they outperform other applied algorithms, as shown in [6, 12, 13]. Furthermore, SVM was used in this study because it was not applied to MERS-CoV dataset in recent studies [6, 7]. Also, the Decision Tree and k-NN classifiers in [7] achieved the highest accuracy on binary-class and multi-class classification respectively. Additionally, based on the structure of the dataset, each class was chosen to represent a different classification type.

The software used in this study was RapidMiner Studio version 7.6. RapidMiner is an open-source data mining software tool written in Java programing language. It is issued under the Affero General Public License that provides an integrated environment for data mining and predictive analytics. Moreover, RapidMiner is used to perform machine learning algorithms for data mining tasks [17]. SVM, Decision Tree, and k-NN algorithms were applied on the MERS-CoV dataset using a RapidMiner tool. In sum, this experimental study was recreated six times.

The essential parameters of the k-NN algorithm are k, which is determined by the numbers of nearest neighbors. The value of k was set to 5 because an odd value is recommended to prevent a tie, when two or more classes have the same number of votes. Additionally, the Euclidean distance function was used as a similarity measure between testing and training data [4, 18]:

$$Euclidean\ distance\ (x, x_i) = \sqrt{\sum_{i=0}^{m}} (x - x_i)^2 \tag{1}$$

Where x is the testing point, and x_i is the training point.

For the decision tree, gain ratio was used as the attribute selection method for splitting, because it measures the information that gained by each attribute easily and quickly. The information gain ratio calculated using the following formula [4, 18]:

$$Gain(A) = Info(D) - InfoA(D) \tag{2}$$

where the Info (D) is the average amount of information needed to identify the class label of a tuple in D also, known as the entropy of D, and it is calculated by:

$$Info(D) = \sum_{i=1}^{m} \mathrm{P_i log_2(p_i)} \tag{3}$$

Where $\mathrm{P_i}$ is the probability that an arbitrary tuple in D belongs to class C, where $Info(D)$ is the expected information required to classify a tuple from D based on the partitioning by attribute A, it is calculated by:

$$Info_A(D) = \sum_{\mathrm{j}=1}^{\mathrm{v}} \frac{|\mathrm{D_j}|}{|\mathrm{D}|} \times \mathrm{Info}(\mathrm{D_j}) \tag{4}$$

The term $\frac{|\mathrm{D_j}|}{|\mathrm{D}|}$ is the weight of the $\mathrm{j^{th}}$ partition [4].

The maximum depth of the Decision Tree was set to 20. Also, the Decision Tree was generated with a pruning function, which allows for reducing the size of the tree by removing low-power sub-trees.

The essential parameter of SVM classifier is the kernel function. The most common kernel function that used with SVM classifier is linear kernel function [6] it defined as:

$$F(x) = W.(x_i, y_i,) + b \tag{5}$$

Where $X = (x_i, y_i,)$ is the dataset, x_i, is the instances, y_i, is the class label and $i = 1, 2, \ldots\ldots n$ and W is the weight vector or the coefficient, and b is a scalar value called bias [4, 17]. Other important parameters are the value of complexity constant (C), and the tolerance parameter. In this study, the kernel function that used was the linear function because the data is linearly separable, the parameter C was set to 0, and the tolerance parameter was set 0.001.

Since the classification type of classes, classifier algorithm, and their parameters were specified, a model is needed for assessing the classification performance. Therefore, a cross-validation model is used to assess the classification performance. In k-fold cross validation technique, the dataset is randomly split into k equal-sized subsets. Each time, one of the k subsets is used as the test set and the other k-1 subsets are put together and used as training set (train on nine datasets and test on one). Then, the process is repeated ten times. In this empirical study, all models were built using 10-fold cross validation. The advantage of this method is that it data division into training and testing sets is irrelevant [19].

The most significant part of many studies is discovered during evaluation, and the value of the study can be assessed. To compare all results of the applied algorithms to the MERS-CoV dataset in this project, their performances were quantitatively measured using accuracy, which is the most widely used evaluation metric to reflect the percentage of correctly-classified records in the testing phase [21]. Therefore, the accurate classifier will be useful for building a MERS-CoV prediction system.

A confusion matrix is an important way for analyzing the performance of a binary-class classification. Moreover, in this matrix, each row contains information about actual

Table 2. Confusion matrix

	Negative (predicted)	Positive (predicted)
Negative (actual)	TP	FN
Positive (actual)	FT	TN

class while each column contains information about predicted class. Accordingly, the confusion matrix aims to analyze how well a classifier can recognize tuples of different classes. Table 2 illustrates the confusion matrix for a two-class classifier [20].

For evaluating the classification framework based on the confusion matrix, the accuracy formula of each classification type was used for the binary-class classifications; the accuracy was calculated based on the following formula:

$$Accuracy = \frac{100 \times (\mathrm{TP} + \mathrm{TN})}{\mathrm{TP} + \mathrm{FN} + \mathrm{TN} + \mathrm{FP}} \tag{6}$$

On the other side, for the multi-class classifications, the accuracy was calculated based on the following formula:

$$Accuracy = \frac{\sum_{i=1}^{l} \frac{100 \times (TP_i + TN_i)}{TP_i + FN_i + TN_i + FP_i}}{l} \tag{7}$$

Where, TP (True Positives) is the correctly classified positive cases, TN (True Negative) is the correctly classified negative cases, FP (False Positives) is the incorrectly classified negative cases, and FN (False Negative) is the incorrectly classified positive cases [21].

5 Results and Discussion

Based on the essential parameters of the classifier models, which are presented in the methodology section, the accuracies obtained on the MERS-CoV dataset with each classifier model for each classification type are shown in Table 3. The best accuracy is for the binary-class classification based on healthcare personnel class with 86.44%, which was produced by SVM and Decision Tree algorithms. Figure 4 illustrates the result of the binary-class classification; when applying SVM with the healthcare personnel class, the margin width was maximizing, making the prediction faster and more accurate. On other hand, as healthcare personnel class became a root of the decision tree the depth of the tree was minimized and the tree is not complex that generates accurate predictions.

Table 3. Classifier model accuracy for each classification type

Classification type	SVM classifier	Decision tree classifier	*k*-NN classifier
Binary-class classification	86.44%	86.44%	85.31%
Multi-class classification	18.24%	42.80%	30.80%

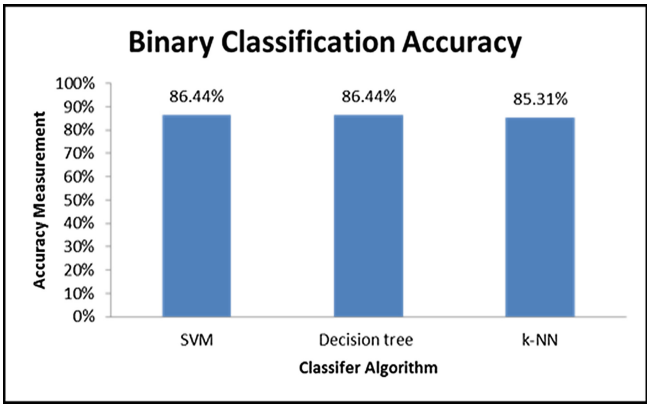


Fig. 4. Binary-class classification accuracy.

Another important finding is that, for multi-class classification the Decision Tree obtains the highest accuracy with 42.80% based on city class. Whereas the accuracy of *k*-NN was 30.80% and SVM classifier was under 20% as shown in Fig. 5. Moreover, for evaluating the effectiveness of the results of our method, we have to compare the experimental results with the results of a recent study [7]. A recent study [7] reported highest accuracy of 51.60% for multi-class classification based on *k*-NN classifier. This could be due to the value of parameter *k* that is set to 5 in this study while it is set to 3 in [7].

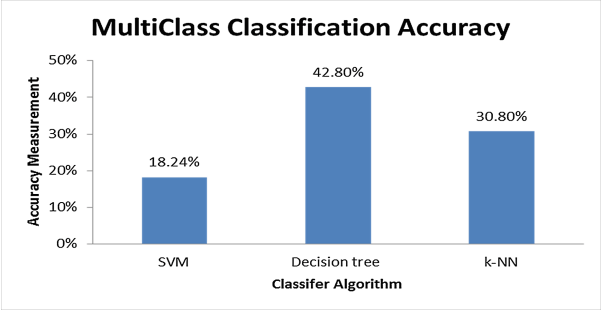


Fig. 5. Multi-class classification accuracy.

On another side, the researchers in [7] reported a higher accuracy on binary-class classification of 90.00%, when using the Decision Tree algorithm based on gender attribute, while our method achieved a lower accuracy of 86.44% when using SVM and Decision Tree classifiers based on healthcare personnel attribute. Therefore, using healthcare personnel attribute as binary-class may not be appropriate for MERS-CoV dataset classification.

The experimental results demonstrate that the accurate classifier models for binary-class and multi-class classification types are built by using Decision Tree and *k*-NN algorithms respectively. Additionally, the results of this study indicate that using SVM classifier is not suitable for classification of MERS-CoV dataset. In general, the main explanation of our results is based on the essential parameter settings.

6 Conclusions and Future Work

The classifier model performance of several classification types can greatly assist to enhance the prediction accuracy of MERS-CoV infection. In this study, we have identified a classifier model performance that applied binary and multiclass classification on real a MERS-CoV dataset. Three algorithms were used to build classifier models, which were SVM, Decision Tree, and *k*-NN. The algorithms were applied using RapidMiner, a data mining tool. The performance of classifier models was measured using the accuracy evaluation metric; in addition, cross-validation was used as a model for assessing classification performance.

The experimental results have shown that both SVM and Decision Tree classifiers achieved the highest accuracy of 86.44% on binary-class classification based on healthcare personnel class. On the other hand, the Decision Tree classifier had the highest accuracy of 42.80% among the remaining classifiers for multiclass classification based on city class, although it did not reach a satisfactory accuracy level. In general, the comparison of the experimental results and the results of a recent study indicate that Decision Tree and *k*-NN classifiers are the accurate classifiers for binary-class and multi-class classification types respectively. Additionally, using an SVM classifier is not suitable for classification of a MERS-CoV dataset. For future work, it is intended that this experiment will be applied to the universal MERS dataset. Furthermore, other preprocessing technique such as remove missing value can be used to measure its effect on the classifier models' performance. Additionally, other classification methods, such as ensemble learning, can be used. Also, another similarity metric, such as cosine similarity, may be used with the *k*-NN algorithm. Finally, for the multiclass classification, we suggest recreating the empirical study with different parameters to determine a classifier that gives accuracy greater than 50%.

References

1. Coronavirus website - Ministry of Health. http://www.moh.gov.sa/en/CCC/. Accessed 29 Oct 2017
2. WHO: Middle East respiratory syndrome coronavirus (MERS-CoV). http://www.who.int/emergencies/mers-cov/en/. Accessed 23 Oct 2017
3. Koh, H.C., Tan, G.: Data mining applications in healthcare. J. Healthc. Inf. Manag. **19**(2), 64–72 (2005)
4. Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Elsevier, Haryana, India, Burlington (2012)
5. Yoo, et al.: Data mining in healthcare and biomedicine: a survey of the literature. J. Med. Syst. **36**(4), 2431–2448 (2012)

6. Al-Turaiki, M., Alshahrani, M., Almutairi, T.: Building predictive models for MERS-CoV infections using data mining techniques. J. Infect. Public Health **9**(6), 744–748 (2016)
7. AlMansour, N., Kurdi, H.: Identifying accurate classifier models for a text - based MERS-CoV dataset. Presented at the Intelligent Systems Conference 2017, London, UK (2017)
8. Deepika, K., Seema, S.: Predictive analytics to prevent and control chronic diseases, pp. 381–386 (2016)
9. Ferreira, D., Oliveira, A., Freitas, A.: Applying data mining techniques to improve diagnosis in neonatal jaundice. BMC Med. Inform. Decis. Mak. **12**(1), December 2012
10. Asri, H., Mousannif, H., Moatassime, H.A., Noel, T.: Using machine learning algorithms for breast cancer risk prediction and diagnosis. Procedia Comput. Sci. **83**, 1064–1069 (2016)
11. Li, J., Zhao, Z., Liu, Y., Cheng, Z., Wang, X.: A comparative study on machine classification model in lung cancer cases analysis. In: Yen, N.Y., Hung, J.C. (eds.) Frontier Computing, vol. 422, pp. 343–357. Springer Singapore, Singapore (2018)
12. Daghistani, T., Alshammari, R.: Diagnosis of diabetes by applying data mining classification techniques. Int. J. Adv. Comput. Sci. Appl. **7**(7) (2016)
13. Sowjanya, K., Singhal, A., Choudhary, C.: MobDBTest: a machine learning based system for predicting diabetes risk using mobile devices, pp. 397–402 (2015)
14. Kim, D., Hong, S., Choi, S., Yoon, T.: Analysis of transmission route of MERS coronavirus using decision tree and apriori algorithm, pp. 559–565 (2016)
15. Sandhu, R., Sood, S.K., Kaur, G.: An intelligent system for predicting and preventing MERS-CoV infection outbreak. J. Supercomput. **72**(8), 3033–3056 (2016)
16. Jang, S., Lee, S., Choi, S.-M., Seo, J., Choi, H., Yoon, T.: Comparison between SARS CoV and MERS CoV using Apriori Algorithm, Decision Tree, SVM. In: MATEC Web of Conferences, vol. 49, p. 08001 (2016)
17. RapidMiner Studio - RapidMiner Documentation. http://docs.rapidminer.com/studio/. Accessed 11 Jan 2017
18. Witten, H., Frank, E., Hall, M.A.: Data Mining: Practical Machine Learning Tools and Techniques, 3rd edn. Morgan Kaufmann, Burlington (2011)
19. Kohavi, R.: A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence, vol. 2, pp. 1137–1143 (1995)
20. Stehman, S.V.: Selecting and interpreting measures of thematic classification accuracy. Remote Sens. Environ. **62**(1), 77–89 (1997)
21. Sokolova, M., Lapalme, G.: A systematic analysis of performance measures for classification tasks. Inf. Process. Manag. **45**(4), 427–437 (2009)