www.ijecs.in
International Journal Of Engineering And Computer Science ISSN: 2319-7242
Volume 6 Issue 1 Jan. 2017, Page No. 19884-19889
Index Copernicus Value (2015): 58.10, DOI: 10.18535/ijecs/v6i1.07

Prediction of Breast Cancer using Random Forest, Support Vector Machines and Naïve Bayes

***Madeeh Nayer Elgedawy*[1]**
[1]Institute of Public Administration, Computer Department,
Prince Mohammed Bin Fahd Road, Dammam
M_nayer@hotmail.com

Abstract*: Machine learning techniques can be used to judge important predictor variables in medical datasets. This paper applies three machine learning techniques: Naïve Bayes, SVM and Random Forest to Wisconsin Breast Cancer Database. The three developed models predict whether the patients' trauma are benign or malignant. The paper aims at comparing the performance of these three algorithms through accuracy, precision, recall and f-measure. Results show that Random Forest yields the best accuracy of 99.42%, which is slightly better than both SVM and Naïve Bayes that have accuracies of 98.8% and 98.24% respectively. These results are very competitive and can be used for diagnosis, prognosis, and treatment.*

Keywords: Breast Cancer, random forest, support vector machine, naïve bayes, R.

1. Introduction

Breast cancer is the most common cancer in women worldwide, with 1.7 million new patients identified in 2012. This represents about 12% of all new cancer cases and 25% of all cancers in women[1]. The highest prevalence of breast cancer was in Northern America and the lowest prevalence in Asia and Africa [1]. In general, the disease targets women in the age of 30 in Arab countries, while affecting women above 45 years in European countries [2].

Cancer cells do differ in both shape and size. Consequently, uniformity of cell steers to a benign class. Similarly bare nuclei, bland chromatin and normal nucleoli are signs of benignity. Moreover, Normal cells stick to each other, but cancerous cells are not. Thus, loss of adhesion is a sign of malignancy and epithelial cells that are considerably puffy may be a malignant cell [3]. Malignancy is determined by taking a sample tissue from the patient's breast and performing a surgery. A benign diagnosis is assured either by surgery or by constant checkup, depending on the patient's choice. Any medical reports use all the previously mentioned factors to determine whether the lump is benign or not[2].

In this paper, three machine learning algorithms are used to automatically classify the patients' lumps: Random Forest, Support Vector Machine and Naïve Bayes. Random Forest depends on bagging and random subspace methods, SVM is a very solid classifier frequently reported as the best classifier possible if fine tuning of its parameters is wisely handled and Naïve Bayes is a simple classifier that assume that predictors are independent. These three algorithms will be compared using the same training and testing sets and performance will be measured by generating the confusion matrix of each and then calculating the accuracy, precision, recall and f-measure.

The Wisconsin Breast Cancer Database[3] consists of visually evaluated nuclear attributes of fine needle aspirates (FNAs) taken from patients' breasts. The three machine learning algorithms will focus on diagnosing cancer tissues or in other words predicting breast cancer susceptibility risks before occurrence, they will not be used to predict breast cancer reoccurrence or prognosis. All the experimentations are implemented in R, which is a language and environment for statistical computing and graphics, it has a huge collection of machine learning and data mining algorithms and it is a little faster than WEKA and new packages are coming up very regularly.

This paper is organized as follows, section 2 provides the brief of the related work to applying machine learning to Breast Cancer, section 3 gives a brief summary of the three classifiers used, section 4 declares the procedures steps followed to get the results, section 5 provides a detailed description of the database and some important data explorations, section 6 represents the experimentations and comments on the results. Finally section 7 concludes the result.

2. Related Work in Breast Cancer

[4] classified the Wisconsin Breast Cancer database by building a J48 decision tree using WEKA, their confusion matrix has 38 observations where are either false positive or false negative. A mix of Particle Swarm Optimization and Support Vector Machines was used in [5]; they got an excellent classification accuracy of 99.03%. [6] used a map reduce approach and proposed a boosting parallel algorithm, PCM. Moreover, Naive Bayes was used as a classifier in [7], and it yielded an accuracy of 96.6%. Some tools are giving impressive results as [8] when used RapidMiner to build an SVM classifier and achieved an accuracy of only 80%. [9] built a hybrid classifier of Support Vector Machines and Decision Trees in WEKA resulting in 91% accuracy. [10] reached an accuracy of 98.5% on The WDBC dataset which consists of 569 observations: 357 benign and 212 malignant and there are 30 variables. [2] made a comprehensive study among several algorithms and their fusion classifier of decision trees and MLP yielded the best accuracy while using PCA for feature reduction. [11] had applied the supervised fuzzy clustering

[1] http://www.wcrf.org/int/cancer-facts-figures/data-specific-cancers/breast-cancer-statistics
[2] http://www.cancer.org/
[3] https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Original)

technique and reported an accuracy of 95.57%. [12] introduced a system based on association Rules and neural networks of accuracy equals to 97.4%.

3. Classifiers

Three classifiers are compared in this research: Random Forest, Support Vector Machines and Naïve Bayes.

3.1 Random Forest

Random Forest is a bagging algorithm that successfully aims at the regularization point where the model quality is as high as possible and variance and bias problems are compromised [13]. To overcome the overfitting problem that encounter decision trees, Random Forests builds hundreds or may thousands of them. To make the trees different from each other, Random Forest uses random samples with replacement [14]. On the average, 37% of the rows will be left out of each sample [15]. Each tree classify its observations, and at the end majority votes [16], decisions are chosen. Random Forest can also be used in unsupervised mode for assessing proximities among data points [17].

3.2 Support Vector Machines

SVM is centered on the idea of defining a hyperplane that divides a dataset into two classes in the best possible way. SVM just takes the data points nearest to the hyperplane into consideration and names them as Support vectors [18]. The distance between the hyperplane and the nearest data point from either set is known as the margin. The goal is to choose a hyperplane with the greatest possible margin between the hyperplane and any point within the training set. As the number of features increases, the number of dimensions also increases, the hyperplane can no longer be a line. It must then be a plane. The idea is that the data will continue to be mapped into higher and higher dimensions until a hyperplane can be formed to segregate it. SVM is accurate and it can automatically mark the far data points as outliers but it is the best model to build for very large datasets [19].

In order to improve the performance of the support vector regression we will need to select the best parameters for the model. The most important parameters are Gamma that specifies number of support vectors and cost that can be adjusted to avoid overfitting [20]. The process of choosing these parameters is called hyper parameter optimization, or model selection using grid search, where different models are trained for the different couples of gamma and cost, and the best values are chosen.

3.3 Naïve Bayes

It is based on Bayesian theory, so the more data points you see, the more experience you gain, and the more accurate your decision will be [21]. It is naïve because it assumes that all features are independent from each other, which of course not the case in most real life scenarios, nevertheless, Naïve Bayes proves to be efficient for wide variety of machine learning problems.

In Naïve Bayes, two types of probabilities are distinguished from each other: the posterior probability of class given predictor and the prior probability of class which is simpler to compute and there is the likelihood, which is the probability of predictor given class [22]. In Naïve Bayes, likelihoods and prior probabilities are calculated first and then Bayesian theorem is used to calculate the posteriors. It is easy and fast, and performs well in case of categorical input variables. If categorical variable has a category (in test data set), which was not observed in training data set, then model will assign a zero probability and will be unable to make a prediction. To solve this, we can use the smoothing technique such as Laplace estimation [23].

4. Procedure

This is the followed steps while building the three models:

1. Acquire dataset from Wisconsin Breast Cancer datasets from the UCI Machine Learning
Repository.
2. Pre-process data for applying Random Forest, Support Vector Machines and Naive Bayes.
Remove Sample Code Number from attribute list
Set output variable as nominal.
3. Loading relevant packages.
4. Doing some exploratory analysis.
5. Diving the dataset into training and testing data.
6. Implementing the three algorithms where the class label is the output variable.
7. Diagnosis of testing set observations based on each of the three models.
8. Building the confusion matrix for each model.
9. Computing the accuracy, precision, recall and f-measure and compare them across all classifiers.

5. Data Description and Cleansing

The Wisconsin Breast Cancer Database has 699 observation and 11 variables: 10 predictors and one outcome variable. The output variable is either benign (458 observations) or malignant (241 observations). Each variable except for the first was converted into 11 primitive numerical attributes with values ranging from 0 through 10. With value one corresponding to a normal state and 10 to a most abnormal state [24].

Table 1: Database Columns Description

Column	*Description*
Id	Sample code number
Cl.thickness	Clump Thickness
Cell.size	Uniformity of Cell Size
Cell.shape	Uniformity of Cell Shape
Marg.adhesion	Marginal Adhesion
Epith.c.size	Single Epithelial Cell Size
Bare.nuclei	Bare Nuclei
Bl.cromatin	Bland Chromatin
Normal.nucleoli	Normal Nucleoli
Mitoses	Mitoses
Class	benign or malignant

There are 16 missing attribute values. All these missing values occur in the Bare.nuclei variable, and most of these cases are benign.

DOI: 10.18535/ijecs/v6i1.07

Table 2: Observations having Missing Values

Cl.thickness	*Cell.size*	*Cell.shape*	*Mitoses*	*Class*
8	4	5	1	malignant
6	6	6	1	benign
1	1	1	1	benign
1	1	3	1	benign
1	1	2	1	benign
5	1	1	1	benign
3	1	4	1	benign
3	1	1	1	benign
3	1	3	1	benign
8	8	8	1	malignant
1	1	1	1	benign
5	4	3	1	benign
4	6	5	1	benign
3	1	1	1	benign
1	1	1	1	benign
1	1	1	1	benign

Table 3: Impossible Imputations

Cl.thickness	*Cell.size*	*Cell.shape*	*Bare.nuclei*	*Class*
8	4	5	5.809919	malignant
6	6	6	9.400442	benign
1	1	1	-0.22687	benign
1	1	3	3.371236	benign
1	1	2	-2.42992	benign
5	1	1	4.245125	benign
3	1	4	0.094111	benign
3	1	1	1.568057	benign
3	1	3	6.292119	benign
8	8	8	6.084727	malignant
1	1	1	2.80573	benign
5	4	3	1.922176	benign
4	6	5	3.749635	benign
3	1	1	-0.80364	benign
1	1	1	3.408875	benign
1	1	1	2.306711	benign

For exploratory analysis, the correlations among the variables were plotted to check if there are any strong collinearity, then a quick decision tree was built on the dataset, and it was observed that not all of the variables were important. The most important variables are Cell Size, Cell Shape, Clump Thickness and Bare Nuclei.

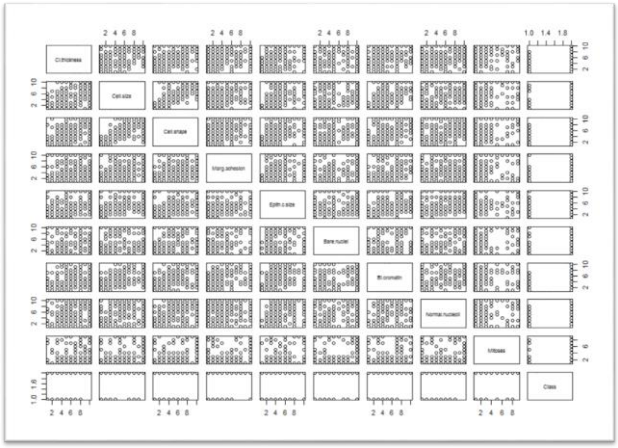

Figure 1: Correlations among Features

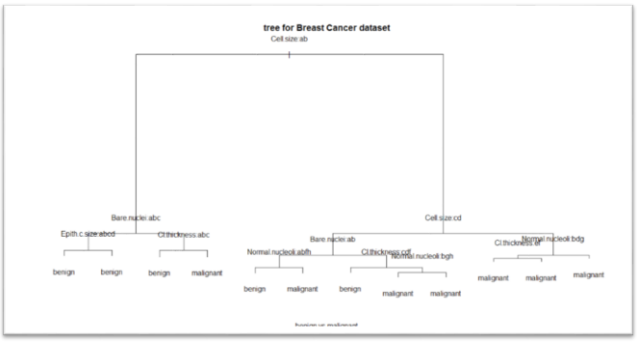

Figure 2: Exploratory Tree

Five imputations have been tested to replace the missing values in the Bare Nuclei variable, but all these imputations have impossible negative values.

As that is the case, it is most probable that the best solution is to delete these 16 cases altogether before proceeding. Nevertheless, two runs were tested for each classifier: one where the 16 observations are deleted and the other when they are replaced by the mean value.

6. Experimentation

For measuring performance, the following expressions are used [25]:

True positive (TP): hit.

True negative (TN): correct rejection.

False positive (FP): false alarm or Type I error.

False negative (FN): miss or Type II error.

Recall (or sensitivity) $=\frac{TP}{TP+FN}$

Precision$=\frac{TP}{TP+FP}$

Accuracy$=\frac{TP+TN}{TP+TN+FP+FN}$

F-measure$=\frac{2\times Precision\times Recall}{Precision+Recall}$

6.1. Random Forest Model

The dataset has been divided into training and testing sets, quarter of the dataset are allocated to testing phase. In the first run, there are 512 observations for training and 171 for testing, after removing the 16 observations that have missing values, which makes 683 observations. In the second run, the missing values will be replaced with the mean value, thus all 699 observations are included and divided into 524 observations for training and 175 for testing. Number of trees to grow is set to 2000. Therefore, it is ensured that every observation is predicted at least a few times. After training the model, it is possible to try to neglect each of the predictors one by one and see which of them negatively affected the accuracy and GINI Index by being removing.

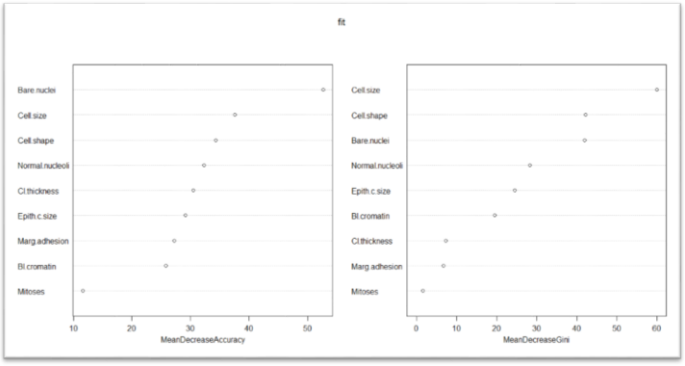

Figure 3: Features Importance

It is obvious that Cell Size, Bare Nuclei and Cell Shape are influential variables. Mitoses variable added little insight to the model. The other predictors are of moderate influence, but none of them can increase the model accuracy by being neglected.

The confusion matrix of Random Forest is promising. There is only one observation that was misclassified and the accuracy equals to 99.42%. By dividing the 103 true positives by true and false positives, then the precision is one, but the recall is 0.99, which can be calculated by dividing the true positives by both true positives and false negatives. According to precision and recall, f-measure is equal to 0.995. The misclassified observation is actually benign.

When missing values are taken into account and replaced by mean value, then there will be four misclassified observations and all the performance Metrix will clearly decrease; accuracy equals to 97.71%, precision is 0.992, recall is 0.975 and f-measure yields 0.983. The drop in accuracy and f-measure are 1.7% and 0.02 respectively.

Table 4: Random Forest Confusion Matrix (Missing Values Removed)

		Predicted	
		Benign	*Malignant*
Actual	*Benign*	103	1
	Malignant	0	67

Table 5: Random Forest Performance Matrix (Missing Values Removed)

Model Performance			
Accuracy	*Precision*	*Recall*	*F-measure*
99.42%	1	0.990	0.995

Table 6: Random Forest Confusion Matrix (Missing Values Replaced with the Mean)

		Predicted	
		Benign	*Malignant*
Actual	*Benign*	118	3
	Malignant	1	53

Table 7: Random Forest Performance Matrix (Missing Values Replaced with the Mean)

Model Performance			
Accuracy	*Precision*	*Recall*	*F-measure*
97.71%	0.992	0.975	0.983

6.2. Support Vector Machines Model

512 observations for training and 171 for testing, the missing observations are removed. For the Cost parameter, these values were tested: 0.001,0.01,0.1,1,4,5,8,10,16,50,100 and for the Gamma parameter, many different values were tested, the best obtained value was 0.123. RBF kernel function was implemented in the model. The best Gamma value is 0.5; best Cost is 4 and number of support vectors is 280: 102 malignant and 178 benign, respectively. The confusion matrix shows that the model accuracy is slighter worse than Random Forest.

Table 8: Support Vector Machines Confusion Matrix

		Predicted	
		Benign	*Malignant*
Actual	*Benign*	103	1
	Malignant	1	66

Table 9: Support Vector Machines Performance Matrix

Model Performance			
Accuracy	*Precision*	*Recall*	*F-measure*
98.8%	0.99	0.99	0.99

Table 10: Support Vector Machines Confusion Matrix (Missing Values Replaced with the Mean)

		Predicted	
		Benign	*Malignant*
Actual	*Benign*	116	5
	Malignant	0	54

Table 11: Support Vector Machines Performance Matrix (Missing Values Replaced with the Mean)

Model Performance			
Accuracy	*Precision*	*Recall*	*F-measure*
97.14%	1	0.959	0.979

There are only two misclassified observations: one of them is Benign and the other is malignant. Thus, accuracy equals to 98.8%, precision and recall both equals to 0.99 and f-measure yields 0.99. When missing values are taken into account and replaced by mean value, then there will be five misclassified observations, all of them are benign but predicted as malignant. Moreover, the recall will decrease, but the precision has a slightly better figures, but the overall effect will negatively affect the performance as accuracy equals to 97.14%, precision is 1, recall is 0.959 and f-measure yields 0.979. The drop in accuracy and f-measure are 1.66% and 0.011 respectively.

6.3. Naïve Bayes Model

The same 512 observations for training and 171 for testing, the missing observations are removed. The confusion matrix shows that the model accuracy is slighter better than Support Vector Machine.

Table 12: Naïve Bayes Confusion Matrix

		Predicted	
		Benign	*Malignant*
Actual	*Benign*	102	2
	Malignant	1	66

Table 13: Naïve Bayes Performance Matrix

Model Performance			
Accuracy	*Precision*	*Recall*	*F-measure*
98.24%	0.99	0.981	0.986

Table 14: Naïve Bayes Confusion Matrix (Missing Values Replaced with the Mean)

		Predicted	
		Benign	*Malignant*
Actual	*Benign*	118	3
	Malignant	0	54

Table 15: Naïve Bayes Performance Matrix (Missing Values Replaced with the Mean)

Model Performance			
Accuracy	*Precision*	*Recall*	*F-measure*
98.23%	1	0.975	0.987

There are three misclassified observations: two of them are Benign and the third is malignant. Thus, accuracy equals to 98.24%, precision and recall equals to 0.99 and 0.981 respectively, and f-measure yields 0.986. When missing values are taken into account and replaced by mean value, number of misclassified observations will be the same, but all of them will be benign, yet predicted as malignant. Moreover, the recall will decrease, but the precision has a slightly better figures, but the overall effect will positively affect the performance as accuracy equals to 98.24%, precision is 1, recall is 0.975 and f-measure yields 0.987. The rise in accuracy and f-measure are 0.01% and 0.001 respectively.

7. Conclusion

Random Forest yields the best accuracy and f-measure, it is slightly better than both Support Vector Machine and Naïve Bayes. Naïve Bayes was the only classifier that positively affected by added the missing values to the model.

By applying a simple voting scheme that include Random Forest, SVM and couple of other weak leaners, a slightly better overall accuracy may be achieved. This voting scheme should give a greater weight to Random Forest as it has the best accuracy. Moreover, applying other techniques, rather than imputation that clearly failed, on missing values may positively affect the models' accuracies. In addition, more algorithms need to be tested such as Fuzziers, Artificial Neural Networks and Logistic Regression. It is noteworthy, that there are several available breast cancer databases that have very huge number of observations and more features to consider, it will be important to check if Random Forest will do the same good job on these huge databases.

References

[1] Ferlay, J., Soerjomataram, I., Ervik, M., Dikshit, R., Eser, S., Mathers, C., & Bray, F. (2015). Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11 [Internet]. Lyon, France: International Agency for Research on Cancer. GLOBOCAN.

[2] Salama, G. I., Abdelhalim, M., & Zeid, M. A. E. (2012). Breast cancer diagnosis on three different datasets using multi-classifiers. Breast Cancer (WDBC), 32(569).

[3] Raju, R. (2012). Relative Importance of Fine Needle Aspiration Features for Breast Cancer Diagnosis: A Study Using Information Gain Evaluation and Machine Learning. Journal of the American Society of Cytopathology, 1(1), S11.

[4] Sumbaly, R., Vishnusri, N., & Jeyalatha, S. (2014). Diagnosis of Breast Cancer using Decision Tree Data Mining Technique. International Journal of Computer Applications, 98(10).

[5] Chen, H. L., Yang, B., Wang, G., Wang, S. J., Liu, J., & Liu, D. Y. (2012). Support vector machine based diagnostic system for breast cancer using swarm intelligence. Journal of medical systems, 36(4), 2505-2519.

[6] Umesh D. R. & B. Ramachandra (2016). Parallel Computing to Predict Breast Cancer Recurrence on SEER Dataset using Map-Reduce Approach. International Journal of Computer Applications, 149, 31-35.

[7] Gayathri, B. M., & Sumathi, C. P. (2016). An Automated Technique using Gaussian Naïve Bayes Classifier to Classify Breast Cancer. International Journal of Computer Applications, 148(6).

[8] Priyanka Jain & Santosh Kr. Vishwakarma (2016). Collaborative Analysis of Cancer Patient Data using Rapid Miner. International Journal of Computer Applications, 145, 8-13.

[9] Sivakami, K. (2015). Mining Big Data: Breast Cancer Prediction using DT-SVM Hybrid Model. International Journal of Scientific Engineering and Applied Science, 1(5).

[10] Animesh Hazra, Subrata Kumar Mandal & Amit Gupta (2016). Study and Analysis of Breast Cancer Cell Detection using Naïve Bayes, SVM and Ensemble Algorithms. International Journal of Computer Applications, 145, 39-45.

[11] Nauck, D., & Kruse, R. (1999). Obtaining interpretable fuzzy classification rules from medical data. Artificial intelligence in medicine, 16(2), 149-169.

[12] Karabatak, M., & Ince, M. C. (2009). An expert system for detection of breast cancer based on association rules and neural network. Expert systems with Applications, 36(2), 3465-3469.

[13] Ziegel, E. R. (2012). The Elements of Statistical Learning. Technometrics.

[14] Kotsiantis, S. B. (2013). Decision trees: a recent overview. Artificial Intelligence Review, 39(4), 261-283.

[15] Montano-Gutierrez, L. F., Ohta, S., Kustatscher, G., Earnshaw, W. C., & Rappsilber, J. (2016). Nano Random Forests to mine protein complexes and their relationships in quantitative proteomics data, 050302.

[16] Pudlo, P., Marin, J. M., Estoup, A., Cornuet, J. M., Gautier, M., & Robert, C. P. (2016). Reliable ABC model choice via random forests. Bioinformatics, 32(6), 859-866.

[17] Afanador, N. L., Smolinska, A., Tran, T. N., & Blanchet, L. (2016). Unsupervised random forest: a tutorial with case studies. Journal of Chemometrics, 30(5), 232-241.

[18] Dogan, U., Glasmachers, T., & Igel, C. (2016). A unified view on multi-class support vector classification. Journal of Machine Learning Research, 17, 1-32.

[19] Amer, M., Goldstein, M., & Abdennadher, S. (2013, August). Enhancing one-class support vector machines for unsupervised anomaly detection. In Proceedings of the ACM SIGKDD Workshop on Outlier Detection and Description (pp. 8-15).

[20] Subasi, A. (2015). A decision support system for diagnosis of neuromuscular disorders using DWT and evolutionary support vector machines. Signal, Image and Video Processing, 9(2), 399-408.

[21] Marsland, S. (2015). Machine learning: an algorithmic perspective. CRC press.

[22] Buntine, W. L. (2013). Decision tree induction systems: a Bayesian analysis. arXiv preprint arXiv:1304.2732.

[23] Marsland, S. (2015). Machine learning: an algorithmic perspective. CRC press.

[24] Fryback, D. G., Stout, N. K., Rosenberg, M. A., Trentham-Dietz, A., Kuruchittham, V., & Remington, P. L. (2006). The Wisconsin breast cancer epidemiology simulation model. MONOGRAPHS-NATIONAL CANCER INSTITUTE, 36, 37.

[25] Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. Journal of machine learning research, 3(Mar), 1289-1305.

Author Profile

Madeeh Nayer received the B.S. in Electrical Engineering from Cairo University in 2002; he worked in the Information and Decision Support Center – Egyptian Cabinet for seven years, first as a system analyst and developer, then as a data and text mining analyst. He received his M.S. in Information Systems in 2008 from Arab Academy for Science, Technology and Maritime Transport. He received his Ph.D. in Information Systems from Faculty of Computers and Information, Cairo University in 2013. He has been working in the Institute of Public Administration in Saudi Arabia as a lecturer and assistant professor for nearly five years.