Feature Selection in Machine Learning Models for Road Accident Severity

Isra Al-Turaiki

College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia

Summary
Traffic accidents are a major cause of serious injuries and deaths around the world. Building predictive models from traffic data can give insights that help authorities improve road safety. Feature selection is an important step in building effective machine learning models. Feature selection methods are used to determine features that are relevant to classification task. The chosen feature selection method can affect the performance of machine learning models. In this paper, a real dataset of traffic accidents in Saudi Arabia is used to model accident severity. Classification models are built using single and ensemble classification algorithms. In addition, we evaluate the performance of developed models to which feature selection is applied. Two feature selection methods are used in this study: information gain, which is a filter-based feature selection method, and a genetic algorithm, which is a wrapper-based method. Experimental results show that better classification performance is obtained with genetic algorithm feature selection. In particular, ID3 and naïve Bayes classifiers have improved results with genetic algorithm feature selection.

Key words:
Machine learning, Road, Traffic, Accidents, Severity, Classification Models, Ensemble

1. Introduction

Road accidents continue to be a major issue around the world. In 2018, the number of annual road traffic deaths reached 1.35 million [1]. The World Health Organization report on road traffic injuries indicates that injuries from road accidents are the eighth leading cause of death. The majority of traffic accident victims are males between the age of 15 and 44 years old [2]. Saudi Arabia is among the world's highest income countries; however, the number of deaths in Saudi Arabia due to road accidents is the highest in the region. In recent years, government authorities put much effort toward improving road safety. Many road safety laws have been enforced, such as the seat belt and the ban of mobile phone while driving [3]. In addition, several awareness campaigns were launched on both traditional and social media platforms [4]. Research suggests that victims of road accidents occupy 20% of hospital beds in Saudi Arabia. In addition, 81% of deaths in hospitals are caused by road accidents [5]. The social and economic impacts of road accidents are huge. Thus, it is of crucial importance to investigate the factors leading to the severity of driver injuries and to develop effective accident severity prediction models. Traffic accident data has been analyzed using various statistical methods [6][7][8]; however, it has been shown that statistical methods fail when applied to complex and nonlinear traffic data [9].

Machine learning has been applied with success in many fields, such as finance [10], education [11], and healthcare [12]. Machine learning algorithms overcome the limitations of statistical methods, as they have the advantage of being able to deal with large amounts of multidimensional data. Thus, they are suitable for the analysis of traffic safety data [9]. To be effective, machine learning techniques require a set of input features, called attributes, that are relevant to the prediction task. But, with the variety of features available for traffic accidents, determining the best subset that are useful in building a prediction model is not easy. A good selection of features can lead to better machine leaning prediction models.

Feature selection refers to the selection of attributes that are most representative of a given dataset. There are two types of feature selection methods: filter-based and wrapper-based methods [13]. In filter-based feature selection methods, the features are first ranked according to some measure, and then the top raking features are selected. In wrapper-based feature selection methods, a specified classifier is used to evaluate the quality of the selected features.

In this study, we investigated the effect of feature selection methods on the performance of single and ensemble accident severity classification models. Although this question has been addressed in the literature [10], it has not been investigated for traffic data. We compared the performance of a filter-based and a wrapper-based feature selection method on a real dataset of traffic accidents occurring in Saudi Arabia. We formulated the problem of traffic accident severity prediction as a binary classification problem and built single and ensemble classification accident severity models. Four machine learning techniques were used for this purpose: decision trees (ID3), naïve Bayes (NB), support vector machines (SVM), and logistic regression (LR). In addition, two well-known ensemble algorithms are used: bagging and boosting.

Manuscript received March 5, 2020
Manuscript revised March 20, 2020

The rest of the paper is organized as follows: In Section 2, some previous work on machine learning for traffic safety problems is reviewed. In Section 3, we describe the dataset and the research methodology. In Section 4, we present our experimental setup and discuss the obtained results. Finally, Section 5,concludes the paper with suggested future work.

2. Related Work

Analyzing road accident data using machine learning has been widely studied throughout the literature.

Kumeda et al. [14] modeled traffic accidents using six machine learning algorithms: Fuzzy-FARCHD, Random Forest, Hierarchal LVQ, Radial Basis Function Network, Multilayer Perceptron, and Naïve Bayes. An evaluation of a real dataset of traffic accidents in the UK showed that the Fuzzy-FARCHD algorithm was effective in classifying the dataset.

Singh et al. [15] utilized Naïve Bayes, k-Nearest Neighbours, Decision trees, and Support Vector Machines for the evaluation of road accidents occurring in India. A real dataset of 38,604 road accident records were used in their study. The research showed the superiority of the Decision Tree classifier, since it achieved the highest accuracy among the four tested classifiers.

In Ghana, Wahab, and Jiang [16] modeled the severity of injury from motorcycle crashes using the following algorithms: J48 decision tree, Random Forest, and instance-based learning with parameter k. Models were trained using 8,516 records for crashes reported during 2011–2015. The machine learning models were evaluated and compared to the results of a multinomial logit statistical model. It was found that machine learning algorithms outperformed the statistical model in terms of accuracy and effectiveness. The random forest model exhibited the best performance of the three machine learning algorithms due to its global optimization and extrapolation ability.

In Malaysia, Pradhan and Sameen [17] compared deep neural network and SVM models for the prediction of traffic accident severity. The experimental results of a dataset of 1138 records showed that the SVM model with a linear kernel and optimized penalty parameter outperformed the NN models. The best accuracy achieved by the DNN models is slightly higher than the best accuracy achieved by the shallow NN models.

Lee et al. [18] used machine learning algorithms in order to predict traffic accident severity in rainy seasons. The prediction models were built using random forest, ANN, and decision tree algorithms. The models were evaluated using a real dataset of 518 accidents in Korea, and performance was measured using the out-of-bag estimate of error rate, mean square error, and root mean square error. The results showed that the most accurate predictions were obtained by the random forest model.

Ensemble methods have also been applied to traffic problems. Zhang et al. [19] used two averaging methods, random forest and extremely randomized trees, and two boosting methods, adaptive boosting and gradient tree boosting, in order to model traffic crash frequency. All models were tested using a real dataset of crash records obtained from California, USA. The dataset consisted of 1.5 million valid crash records. The experimental results demonstrated that the two averaging models achieved better performance in crash frequency analysis than the two boosting models in terms of predictive accuracy, generalization ability, and stability.

3. Methodology

3.1 Dataset and Preprocessing:

In this study, we used a real dataset of road accidents that was provided by the Saudi General Department of Traffic. We used a sample of 2000 records for accidents occuring in Riaydh city between 2013 to 2015. Information on the accident details, including the involved vehicles and parties, is provided. The raw information was separated into three tables. In the accident table, there were 28 attributes describing the accidents, including location, time of accident, weather conditions, road surface conditions, illumination conditions, injury severity, and accident reasons. All attributes related to the vehicles involved in the accidents are stored in the vehicle table. The eighteen attributes include vehicle type, model, color, percentage of error, and hit side. The last table is the parties table, which gives information about the people involved in the accident. It has 14 attributes, including victim (passenger, driver, etc.), gender, nationality, license type, and health status.

For this study, we were interested in the severity attribute, which could be divided into two categories: serious injury and minor injury. This attribute is included in the accident table.

Before building our models, the three separate tables were integrated into one table. Then, missing values were replaced with the mode. All the irrelevant attributes were discarded, resulting in a total of 17 attributes. The final set of attributes are accident point, accident type, accident angle, description, direction, zone, district, nationality, land surface, vehicle make, vehicle model, vehicle status, vehicle color, victim, license type, registration, and severity. Our dataset was balanced to include 1000 records of severe injuries and 1000 of minor injuries.

3.2 Feature Selection

We used information gain, which is a well-known filter-based feature selection method. It is based on Claude Shannon's information theory. Information gain measures the amount of information gained by knowing the value of a feature. As for the wrapper-based method, we used genetic algorithms (GA) [20]. GA is a meta-heuristic that is inspired by natural evolution. It is a type of evolutionary algorithms that is used to solve optimization problems. In feature selection, GA is used to optimize the subset of relevant features.

3.3 Machine Learning Techniques

For building the single classification models, we used: logistic regression, naïve Bayes, ID3 decision tree, and support vector machines [21]. Then, ensemble models are built using bagging [22] and boosting [23]. In this section, we briefly describe each of these techniques.

Naïve Bayes: The naïve Bayes model is a probabilistic classifier that was built using the Bayes theorem. The main assumption of naïve Bayes is class conditional independence, in which features are assumed to be independent of each other. Naïve Bayes has demonstrated success in many applications. It also showed comparable performance to the other machine learning algorithms.

ID3: The ID3 is a decision tree model that constructs a tree for the classification task. The decision tree is generated by recursively splitting the training dataset based on a selected attuite. An attribute selection measure is used to determine the best choice for splitting, called the split attribute. A tree can be converted to a list of IF-THEN rules. Decision tree models are widely used since they easy to understand.

Support vector machines: SVM is a machine learning algorithm that transforms the training dataset into a higher dimension. It then finds the optimal hyperplane that separates data points of one class from another. The optimal hyperplane is the one that has the maximum margin, which is the distance between the data points of each class and the hyperplane.

Logistic regression: Logistic regression models the probability of a given data points belonging to some class based on the value of independent features. It then uses the model to predict the probability that a given data point belongs to a certain class. The sigmoid function is used in building the regression model. It is assumed that the data points follow a linear function.

Bagging: Bootstrap aggregation is an ensemble machine learning algorithm. Ensemble classifiers are based on combining the predictions of many classifiers in order to improve the prediction performance. In bagging, the training dataset is sampled with a replacement for a number of k times. In each iteration, a classifier is built. To classify an unknown data point, each individual classifier returns its prediction. The bagging classifier counts the votes of each individual classifier, and the new data point is assigned the class with the most votes.

Adaboost: Adaptive boosting is also an ensemble machine learning algorithm. Similar to bagging, boosting learns k classifiers from sampling the training dataset with a replacement. However, in boosting, weight is assigned to each data point, which reflects the difficulty of its classification. Initially, all data points are assigned an equal weight. In each iteration, if a data point is misclassified, its weight is increased. If a data point is correctly classified, its weight is decreased. Data point weights are used in each iteration to sample the dataset.

4. Experimental Results

4.1 Experimental setup

Our machine learning models are implemented using *RapidMiner Studio 9.6* [24]. We ran the experiments using the MacBook Pro operating system macOS Catalina version 10.15.3 with a 2.3 GHz 8-Core Intel Core i9 and 16 GB RAM.

The performance of the machine learning models was evaluated in terms of accuracy, precision, and recall. Accuracy is the percentage of accidents records that are correctly classified, and it is calculated as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

Precision refers to the percentage of records correctly classified as severe injuries out of all the records predicted as severe injuries, and it is calculated as follows:

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

Finally, recall is the percentage of records correctly classified as severe injuries out of the total number of severe injuries, and it is calculated as follows:

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

where TP (true positives) is the total number of accident records that are correctly classified as severe injuries, FP (false positives) is the total number of records incorrectly classified as severe injuries, FN (false negatives) is the total number of accidents that are incorrectly classified as minor injuries, and TN

(true negatives) represents the total number of accidents that are correctly classified as minor injuries.

4.2 Experimental Results

4.2.1 Single Classifiers

Figures 1 shows the accuracy, precision, and recall for the four single prediction models without applying feature selection. All models show comparable performance in all measures, except the SVM model, which shows the least accuracy and least recall but the best precision.

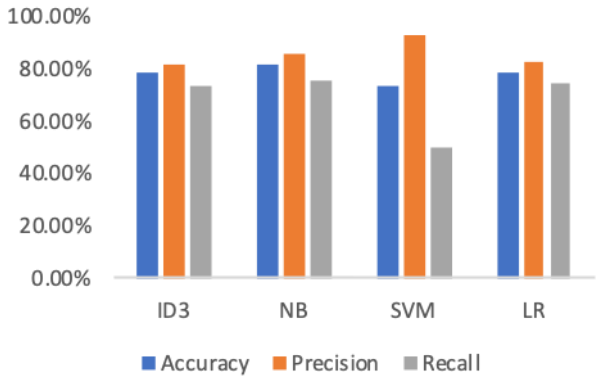

Fig. 1 Performance of single classifiers before applying feature selection

For single classifiers with feature selection, Table 1 shows the performance measures for the single classifiers when feature selection is applied before model construction. The table shows the performance measures with information gain and GA feature selection. Overall, the combination of ID3 and GA (ID3+GA) achieved the highest accuracy value (81.40%).

For the ID3, the prediction improved when using GA. For NB, no improvement was observed with feature selection. However, the prediction performance when GA was applied (NB+GA) outperformed that of applying information gain (NB+ InfoGain). We also noticed that GA slightly improved the recall of the SVM model. For LR, feature selection using information gain improved the precision of accident severity classification. Our results are consistent with the literature that wrapper-based methods exhibit have better performance than filter-based methods [25][10].

Table 1: The performance measures for the single classifiers with information gain and GA feature selection

	Accuracy	Precision	Recall
ID3	78.75%	82.10%	73.70%
InfoGain +ID3	76.90%	81.62%	69.40%
GA +ID3	**81.40%**	**86.53%**	**74.60%**
NB	**81.15%**	**85.43%**	**75.20%**
InfoGain+NB	78.25%	85.04%	68.70%
GA+NB	80.10%	85.24%	73.00%
SVM	**73.15%**	**92.85%**	50.20%
InfoGain+SVM	69.40%	89.85%	43.70%
GA +SVM	72.80%	90.53%	**51.00%**
LR	**79.05%**	82.29%	**74.10%**
InfoGain+LR	69.40%	**89.85%**	43.70%
GA +LR	76.25%	82.45%	66.80%

4.2.2. Ensemble Classifiers

Figures 2 and 3 show the performance of ensemble classifiers without feature selection using bagging and Adaboost, respectively. For bagging, the prediction models performed similarly; however, the best precision was obtained by SVM. ID3 with bagging had the best recall. For Adaboost, the best precision was obtained by SVM.

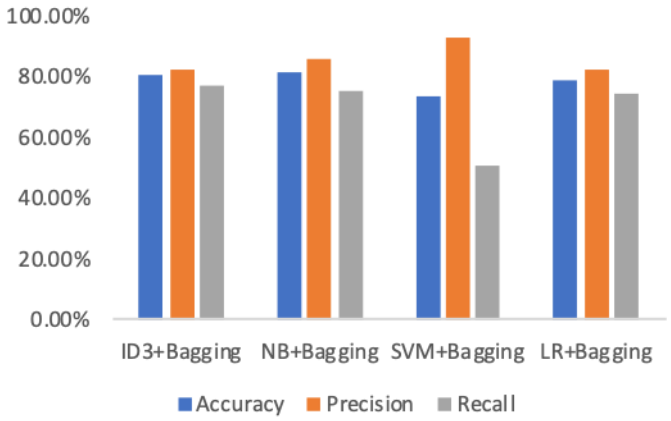

Fig. 2 Performance of bagging classifiers before applying feature selection

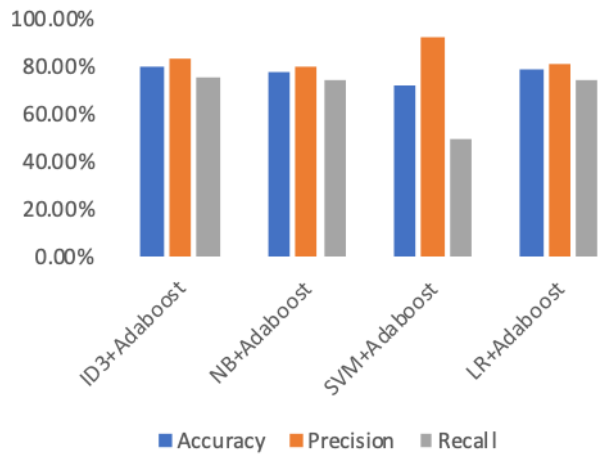

Fig. 3 Performance of Adaboost classifiers before applying feature selection

Table 2 and 3 show the results for the ensemble classifiers with feature selection using bagging and Adaboost, respectively.

In terms of bagging, ID3 with GA (GA+ID3+bagging) improved the accuracy and precision of the severity classification. It also outperformed the ID3 model in terms of bagging and information gain (InfoGain +ID3+bagging). For NB and SVM, bagging without feature selection performed better than bagging with feature selection. But, in both cases, performance with GA was better than with information gain.

Table 2: The performance measures for the bagging classifiers with information gain and GA feature selection

Algorithm	Accuracy	Precision	Recall
ID3+Bagging	80.35%	82.67%	**77.00%**
InfoGain+ID3+Bagging	77.65%	82.78%	70.00%
GA +ID3+Bagging	**81.55%**	**86.34%**	75.10%
NB +Bagging	**81.35%**	**86.05%**	**75.00%**
InfoGain +NB+ Bagging	77.90%	84.93%	68.00%
GA +NB +Bagging	79.85%	84.55%	73.30%
SVM +Bagging	**73.40%**	**92.96%**	**50.70%**
InfoGain+SVM+Bagging	69.40%	90.01%	43.70%
GA +SVM+ Bagging	72.55%	90.72%	50.30%
LR+ Bagging	**79.05%**	82.29%	**74.10%**
InfoGain+LR+Bagging	69.40%	90.01%	43.70%
GA +LR+ Bagging	76.40%	82.78%	66.60%

Table 3: The performance measures for the Adaboost classifiers with information gain and GA feature selection

Algorithm	Accuracy	Precision	Recall
ID3+ Adaboost	79.95%	83.28%	**75.00%**
InfoGain+ID3+Adaboost	77.55%	82.79%	69.60%
GA +ID3+Adaboost	**81.40%**	**86.53%**	74.60%
NB+Adaboost	77.85%	79.76%	**74.60%**
InfoGain +NB +Adaboost	76.15%	80.95%	68.60%
GA +NB+Adaboost	**78.25%**	**81.18%**	73.90%
SVM+ Adaboost	72.50%	**92.37%**	49.10%
InfoGain+SVM+Adaboost	69.40%	90.44%	43.70%
GA +SVM+Adaboost	**72.95%**	90.01%	**51.70%**
LR+ Adaboost	**78.60%**	81.06%	**74.70%**
InfoGain +LR +Adaboost	69.40%	**90.44%**	43.70%
GA+LR+Adaboost	76.00%	81.43%	67.40%

For Adaboost, we also observed better performance for ID3 and NB when GA is used. Overall, the best accuracy values were obtained for GA +ID3, GA +ID3+bagging, GA +ID3+Adaboost, GA +NB, and NB+bagging. In terms of precision, VM, SVM+ bagging, and SVM+ Adaboost produced the best results. All variations of the SVM had the least recall.

5. Conclusion

Traffic accidents worldwide are considered one of the leading causes of serious injuries and deaths. Thus, understanding the factors leading to car accidents can help decision makers better improve road safety.

In this research, we built several single and ensemble classification models for road accident severity. We studied the effect of feature selection methods on prediction performance. Two methods were applied in the dataset processing step. We chose a wrapper-based feature selection method: GA. In addition, a filter-based method, information gain, was tested. Experimental results from a real dataset of traffic accidents indicated that GA exhbits better performance than information gain. Among all models, ID3 and NB performed well when GA was applied. In the future, this experiment can be further investigated using other types of filter and wrapper-based methods. In addition, more datasets can be integrated.

Acknowledgment

The author would like to express her sincere thanks to the General Department of Traffic of Saudi Arabia for facilitating the data access. Special thanks to Major-general Abdullah Al-Zahrani and Colonel Sanad Al-Sanad.

References

[1] "Global status report on road safety 2018. [Online]. Available: https://www.who.int/publications-detail/global-status-report-on-road-safety-2018. [Accessed: 14-Mar-2020].

[2] A. A. Mohammed, K. Ambak, A. M. Mosa, and D. Syamsunur, "A review of traffic accidents and related practices worldwide, *Open Transp. J.*, vol. 13, no. 1, Jun. 2019.

[3] "Violations & Penalties, Ministry of Interior, General Department of Traffic, Saudi Arabia,." [Online]. Available: https://www.moi.gov.sa/. [Accessed: 16-Mar-2020].

[4] "@eMoroor witter," *Twitter*. [Online]. Available: https://twitter.com/emoroor. [Accessed: 16-Mar-2020].

[5] F. A. Mansuri, A. H. Al-Zalabani, M. M. Zalat, and R. I. Qabshawi, "Road safety and road traffic accidents in Saudi Arabia. A systematic review of existing evidence," *Saudi Med. J.*, vol. 36, no. 4, pp. 418–424, Apr. 2015.

[6] C. Dong, D. B. Clarke, X. Yan, A. Khattak, and B. Huang, "Multivariate random-parameters zero-inflated negative binomial regression model: An application to estimate crash frequencies at intersections," *Accid. Anal. Prev.*, vol. 70, pp. 320–329, Sep. 2014.

[7] J. C. Milton, V. N. Shankar, and F. L. Mannering, "Highway accident severities and the mixed logit model: An exploratory empirical analysis," *Accid. Anal. Prev.*, vol. 40, no. 1, pp. 260–266, Jan. 2008.

[8] C. Dong, D. B. Clarke, S. S. Nambisan, and B. Huang, "Analyzing injury crashes using random-parameter bivariate regression models," *Transp. Transp. Sci.*, vol. 12, no. 9, pp. 794–810, Oct. 2016.

[9] C. Dong, C. Shao, J. Li, and Z. Xiong, "An improved deep learning model for traffic crash prediction," *Journal of Advanced Transportation*, 2018.

[10] W.-C. Lin, Y.-H. Lu, and C.-F. Tsai, "Feature selection in single and ensemble learning-based bankruptcy prediction models," *Expert Syst.*, vol. 36, no. 1, p. e12335, 2019.

[11] C. Pierrakeas, G. Koutsonikos, A.-D. Lipitakis, S. Kotsiantis, M. Xenos, and G. A. Gravvanis, "The variability of the reasons for student dropout in distance learning and the prediction of dropout-prone students," in *Machine Learning Paradigms: Advances in Learning Analytics*, M. Virvou, E. Alepis, G. A. Tsihrintzis, and L. C. Jain, eds. Cham: Springer International Publishing, pp.91-111, 2020.

[12] M. Kashif, K. R. Malik, S. Jabbar, and J. Chaudhry, "Chapter 6: Application of machine learning and image processing for detection of breast cancer," in *Innovation in Health Informatics*, M. D. Lytras and A. Sarirete, eds. Academic Press, pp. 145–162, 2020.

[13] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Comput. Electr. Eng.*, vol. 40, no. 1, pp. 16–28, Jan. 2014.

[14] B. Kumeda, F. Zhang, F. Zhou, S. Hussain, A. Almasri, and M. Assefa, "Classification of road traffic accident data using machine learning algorithms," in *2019 IEEE 11th International Conference on Communication Software and Networks (ICCSN)*, pp. 682–687, 2019.

[15] J. Singh, G. Singh, P. Singh, and M. Kaur, "Evaluation and classification of road accidents using machine learning techniques," in *Emerging Research in Computing, Information, Communication and Applications*, Singapore, pp. 193–204, 2019.

[16] L. Wahab and H. Jiang, "A comparative study on machine learning based algorithms for prediction of motorcycle crash severity," *PLOS ONE*, vol. 14, no. 4, p. e0214966, Apr. 2019.

[17] B. Pradhan and M. Ibrahim Sameen, "Modeling traffic accident severity using neural networks and support vector machines," in *Laser Scanning Systems in Highway and Safety Assessment: Analysis of Highway Geometry and Safety Using LiDAR*, B. Pradhan and M. Ibrahim Sameen, eds. Cham: Springer International Publishing, pp. 111–117, 2020.

[18] J. Lee, T. Yoon, S. Kwon, and J. Lee, "Model evaluation for forecasting traffic accident severity in rainy seasons using machine learning algorithms: Seoul City study," *Appl. Sci.*, vol. 10, no. 1, p. 129, Jan. 2020.

[19] X. Zhang, S. T. Waller, and P. Jiang, "An ensemble machine learning-based modeling framework for analysis of traffic crash frequency," *Comput.-Aided Civ. Infrastruct. Eng.*, vol. 35, no. 3, pp. 258–276, 2020.

[20] D. E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*, 1st ed. USA: Addison-Wesley Longman Publishing Co., Inc., 1989.

[21] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995, doi: 10.1007/BF00994018.

[22] L. Breiman, "Bagging Predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, Aug. 1996, doi: 10.1023/A:1018054314350.

[23] Y. Freund and R. E. Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," *J. Comput. Syst. Sci.*, vol. 55, no. 1, pp. 119–139, Aug. 1997, doi: 10.1006/jcss.1997.1504.

[24] "RapidMiner | Best data science & machine learning platform," *RapidMiner*. [Online]. Available: https://rapidminer.com/. [Accessed: 15-Mar-2020].

[25] J. Hua, W. D. Tembe, and E. R. Dougherty, "Performance of feature-selection methods in the classification of high-dimension data," *Pattern Recognit.*, vol. 42, no. 3, pp. 409–424, Mar. 2009.

Isra Al-Turaiki is an assistant professor of Computer Science at King Saud University. She received her PhD degree in 2014 from the College of Computer Sciences at King Saud University. Her research interests include data mining, machine learning, and bioinformatics.