Prediction of Employee Attrition Using Machine Learning and Ensemble Methods

Aseel Qutub, Asmaa Al-Mehmadi, Munirah Al-Hssan, Ruyan Aljohani, and Hanan S. Alghamdi

***Abstract*—Employees are the most valuable resources for any organization. The cost associated with professional training, the developed loyalty over the years and the sensitivity of some organizational positions, all make it very essential to identify who might leave the organization. Many reasons can lead to employee attrition. In this paper, several machine learning models are developed to automatically and accurately predict employee attrition. IBM attrition dataset is used in this work to train and evaluate machine learning models; namely Decision Tree, Random Forest Regressor, Logistic Regressor, Adaboost Model, and Gradient Boosting Classifier models. The ultimate goal is to accurately detect attrition to help any company to improve different retention strategies on crucial employees and boost those employee satisfactions.**

***Index Terms*—Employee attrition, ensemble learning, gradient boosting classifier, machine learning, random forest regressor, stochastic gradient decent.**

I. Introduction

Globally, more than half of organizations have several challenges in retaining the most marketable or high-performance employees [1]. Identifying employees with the highest retention risk or who might be highly targeted for poaching are of great importance for many decision makers within any organization. The assessment of retention risk and the estimating likelihood of leaving would support to establish or update retention strategies and thus to avoid the high cost associated with hiring and training new employees. To assess the employee attrition automatically, in this work we use IBM HR Analytics Employee Attrition Performance dataset available at Kaggle.org. This dataset has been analyzed by many analysts and they published their results on Kaggle's kernel. There are 295 published kernels for this dataset. The most significant number of publications analyzed the dataset by visualization graphs and illustrative curves, tables, and others. In this paper, several machine learning models were trained, optimized and evaluated to predict weather a certain employee will leave the company or not and according to this predication the company will improve different retention strategies on targeted employee.

Manuscript received August 19, 2019; revised September 8, 2020.

Aseel Qutub, Asmaa Al-Mehmadi, Munirah Al-Hssan, Ruyan Aljohani were with the King Abdulaziz University, Information Systems Department, Faculty of Computing and Information Technology Jeddah, Saudi Arabia (e-mail: aseel.qutub@gmail.com, Asmam3008@gmail.com, Munirah.h.hassan@gmail.com, Atheer.kh10@gmail.com, Ruyan.n191@gmail.com).

Hanan. S. Alghamdi Author is with the King Abdulaziz University, Information Systems Department, Faculty of Computing and Information Technology Jeddah, Saudi Arabia (e-mail: hsaalghamdi@kau.edu.sa).

This paper is structured as follows. Section II discusses some related works. Section III describes the machine learning models used. Section IV illustrates the conducted experiments. Section V presents and discusses the results and finally Section VI concludes the paper.

II. Related Works

To identify the factors of employee voluntary attrition or turnover, in [2], the author studied the information provided from 112 respondents between the ages of 18 and 40 in the Chilean labor market. The author concluded that the turnover is the consequence of a combination of factors which include pay, recognition and career development opportunities, among others. Other factors have been discussed in [3] such as the lack of career mobility and challenges and the lack of role clarity. However, identifying employees at risk by a manager using some indicators or factors is a challenging task and needs many years of experience. Moreover, the author also confirmed that these factors are related to people's preferences and expectations, which differ between generations, the type of work and the employees' life stage. Therefore, in this work, we aim to support Human Resource (HR) managerial decisions by automatically predicting employee attrition through using machine learning methods.

III. Methodology

Six different machine learning models have been trained and evaluated in this work; decision tree model, random forest model, gradient boosting model, adaboost, and logistic regression model. Furthermore, to improve the accuracy of our model, three ensemble models combining the most promising algorithms were developed. The following subsections provide general overviews of the theory behind each of these models.

A. Decision Tree Models

Decision trees (DT) are very powerful algorithms, capable of fitting complex datasets and have been applied to wide range of tasks such as medical diagnosis and credit risk of loan application. Decision tree learning approximates a target function which is represented as a tree of "if-then" rules to improve human readability [4]. Thus, it breaks down a dataset into smaller and smaller subsets starting from the topmost node called "root", and then an associated decision tree is incrementally developed. The final decision tree has two nodes; decision nodes and leaf nodes, which can handle both categorical and numerical data [5]. Fig. 1 shows an example of a decision tree generated by using the highest correlated features to the target, i.e. the attrition. As shown,

doi: 10.18178/ijmlc.2021.11.2.1022

we have two classes, "Leave" and "Stay" that can be determined by the examining the features values for each instance.

B. Adaboost Model

Linear regression Adaboost (AB), short for Adaptive Boosting, is an ensemble learning method based on the idea of weighted instances. A set of weak base learners is built consecutively, and each learner focuses on the misclassified instances by the previous learner. To build an AdaBoost classifier, a first base learner is trained and evaluated on the training set. Then, the relative weight of misclassified cases is increased. Then, each learner is trained using the updated weights, makes predictions and updates the weights [3].

C. Random Forest Model

Random forest (RF) is an ensemble learning method for classification and regression that combines many decision trees (weak learners) to form a stronger learner and get a more accurate and stable prediction [6]. Those decision trees vote on how to classify a given instance of input data and outputting the class that is the mode of the classes in case of classification tasks or the mean of predictions in the case of regression tasks. Thus, random forest reduces the problem of overfitting. The more trees in the forest, the better the result will be produced. Random forest learning algorithm is flexible and widely used. It is able to produce good results even without hyper-parameter tuning [7].

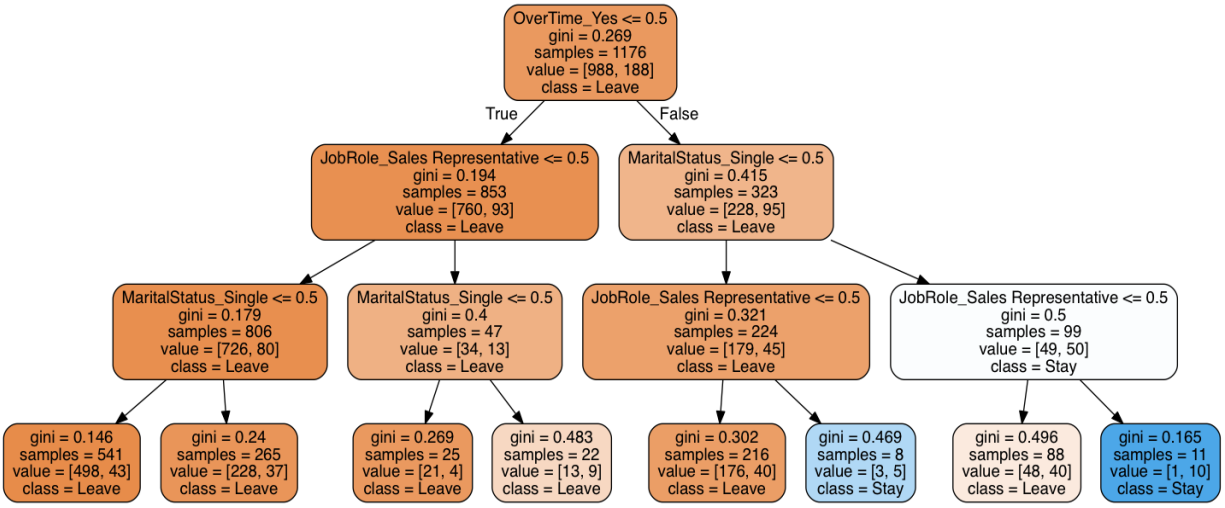


Fig. 1. A decision tree generated by using extracted features from IBM HR dataset.

D. Logistic Regression Model

Logistic regression (LR) estimates the parameters of a logistic (or logit) binomial regression model. Logistic regression is commonly used when the target variable is categorical and with problems of two class values such as pass/fail, win/lose or leave/stay as in the IBM attrition dataset [8]. A simple logistic function is defined by the formula:

$$Y = \frac{e^x}{1+e^x} = \frac{1}{1+e^{-x}} \quad (1)$$

where Y is the dependent variable or (the label) and x represents the independent variables or the features.

E. Adaboost Model

Adaboost (AB), short for Adaptive Boosting, is an ensemble learning method based on the idea of weighted instances. A set of weak base learners is built consecutively, and each learner focuses on the misclassified instances by the previous learner. To build an AdaBoost classifier, a first base learner is trained and evaluated on the training set. Then, the relative weight of misclassified cases is increased. Then, each learner is trained using the updated weights, makes predictions and updates the weights [3].

F. Random Forest Model

Random forest (RF) is an ensemble learning method for classification and regression that combines many decision trees (weak learners) to form a stronger learner and get a more accurate and stable prediction [6]. Those decision trees vote on how to classify a given instance of input data and outputting the class that is the mode of the classes in case of classification tasks or the mean of predictions in the case of regression tasks. Thus, random forest reduces the problem of overfitting. The more trees in the forest, the better the result will be produced. Random forest learning algorithm is flexible and widely used. It is able to produce good results even without hyper-parameter tuning [7].

G. Gradient Boosting Model

Gradient boosting (GB) is another ensemble technique similar to the RF where a combination of weak tree learners is brought together to form a relatively stronger learner and produces a prediction model in the form of an ensemble of weak predictors [9]. The main difference between RF and GB is in GB sequential adding of predictors until no further enhancement can be achieved. [8]. Gradient boosting supports different loss function. It is fast and memory-efficient, however, it is hard to visualize and interpret compared to simpler models such as LR or DT.

H. Ensemble Methods

Ensemble methods (EM) try to construct a set of multiple learning algorithms and combine them to have better predictive performance than what one learning algorithm could be obtained [10]. Most ensemble methods such as RF and GB use a single machine learning algorithm, i.e. DT to produce homogeneous ensembles, however, it is also helpful to produce heterogeneous ensembles, i.e., models of different

types. Ensemble methods can help to reduce both variance and bias [11].

IV. Experimental Evaluation

In this section, we first present our experimental environment and the dataset used in this work. Then, we discuss the data preprocessing steps followed by the statistical analysis and the evaluation metrics we have adopted to evaluate the proposed models. Finally, we explain the cross-validation procedures we employed to avoid overfitting the training set.

A. Dataset

In this paper, IBM HR Analytics Employee Attrition & Performance dataset is used. This dataset contains standard HR features such as age, education, gender and rate. It consists of a total of 1470 observations with 35 different attributes. IBM dataset, although it is fictional, has the characteristics which can represent real-world HR scenarios and its attributes can be readily available to the HR department in any organization. For example, the dataset includes attributes of the number of years since the last promotion, years spent in the company, number of companies the employee worked in and the training times in the last year. There was a total of 35 attributes, out of which two were the same for all data samples, i.e. standard hours and employee count.

B. Data Preprocessing

Data preprocessing is commonly performed before machine learning models training as datasets usually contain missing values, noise and significant differences in features' scale. For the IBM HR dataset, we have applied the following preprocessing steps:

- **Missing value imputation:** we found that the IBM HR dataset used in this paper contains no missing values. Thus, this step was skipped.
- **Feature selection:** feature selection and dimensionality reduction methods are commonly used to enhance machine learning model's performance. In this work, six training sets were created, each with different sets of features. The first set, D1 contains all features and the other sets contains subsets of this whole set. D2 was based on the features highly correlated to the target, i.e. the attrition, D3 contains features demonstrating high correlations to other features in the main set, D4 contains both of D2 and D3 subsets, D5 contains the features showing high positive correlation to the target and the last set D5 was created by using the Principle Component Analysis (PCA) algorithm.
- **Data type conversion:** some machine learning such as logistic regression, are not able to deal with categorical variables. Thus, it is essential to convert these variables into numerical format. In this research, data conversion for categorical attribute was performed using one hot encoding. There were 9 categorical attributes in the dataset and by using one hot encoding, 29 binary new features were added to the dataset.
- **Feature scaling:** In HR datasets, features generally have different scales, for example, employee ages in the IBM Attrition dataset range between 18 to 60 years old, whereas the monthly income range from $2,094 to $26,999. However, having a significant scale gap between features usually slow down the optimization algorithms such as the gradient descent. Indeed, for some machine learning algorithms, feature scaling may help to improve the classification performance and the learning efficiency. In this research, both normalization and standardization were performed on the original dataset after the data conversion step.

C. Evaluation Metrics

In IBM Attrition dataset analytics, the distribution of employees who left and those who stayed is imbalanced. For the "attrition" attribute, only 237 out of 1470 were positive, i.e. the employee who left. This imbalance should be taken into account when evaluating the proposed models. Therefore, in this work, five evaluation metrics were employed to provide a complete coverage and unbiased analysis of the results. This include the following:

- **Accuracy:** calculated as the percentage of the correctly classified data **samples** by the model

$$\text{Accuracy} = \frac{TP+TN}{N} \quad (2)$$

where N is the total number of data samples in the dataset, TP is the true positive and TN is the true negative.

- **Precision: calculated** as the number of TP divided by the sum of TP and false positives TP

$$\text{Precision} = \frac{TP}{TP+FP} \quad (3)$$

- **Recall:** calculated as the number of TN divided by the sum of TP and false positives FP

$$\text{Recall} = \frac{TP}{TP+FN} \quad (4)$$

- **F1 Score:** defined as the harmonic mean of precision and recall

$$F1\ \text{Score} = 2 * \frac{\text{Precision}*\text{Recall}}{\text{Precision}*\text{Recall}} \quad (5)$$

Receiver operating characteristic (ROC) curve: is the plot that displays the trade-off between the $Precision$ and $Recall$ across a series of cut-off points [12]. The closer the curve to the top left corner, the better the classifier. The Area under the ROC Curve (AUC) provides another important metric and is used in this paper to provide further insights of the classifiers' performance. For the perfect classifier the AUC is equal to 1.

V. Results and Discussions

Table I presents the average test results of the five supervised algorithms trained on the six different training sets and Fig. 2 – Fig. 7 show the corresponding ROC curves. As it can be seen from Table I, logistic regression resulted in the highest averaged accuracy, recall and AUC. However, LR performance on different features subsets vary as depicted by Fig. 3. This also applicable to the other models and the DT and the SGB in particular as shown in Fig. 2 and Fig. 6

respectively. As it can be noticed that most models perform the best using D5, thus D5 was evaluated on the ensembles consisting of two based model as shown in Table II. It is worth notice that the best performance ensemble were the ensemble consists of the DT and the LR, the least complex models evaluated in this work. In addition, the LR, the highest performance base model was also the only linear model. This could be due to the dataset size which is relatively small. Further investigation is needed to evaluate these models on larger real-world dataset. Also, although the performance of the ensembles is slightly lower than their best base models, these ensembles would generalize better for unseen examples and larger datasets.

TABLE I: SUPERVISIED MODELS' AVERAGE EVALUATION RESULTS

Model	Evaluation Metric				
	Accuracy	**Precision**	**Recall**	**F1 Score**	**AUC**
DT	82.31 %	0.43	0.06	0.11	0.7763
RF	85.03 %	0.60	0.28	0.39	0.7502
LR	88.43 %	0.74	0.46	0.57	0.8593
GB	84.01%	1.0	0.04	0.07	0.783
AB	86.7%	0.81	0.03	0.06	0.712

TABLE II: ENSMEBLES' MODELS EVALUATION RESULTS

Ensemble	Base Model 1	Base Model 2	Ensemble's Accuracy
DT + LR	85.03 %	88.44 %	**86.39 %**
AB + RF	86.73 %	85.03 %	86.05 %
SG + GB	83.33 %	82.31 %	81.23 %

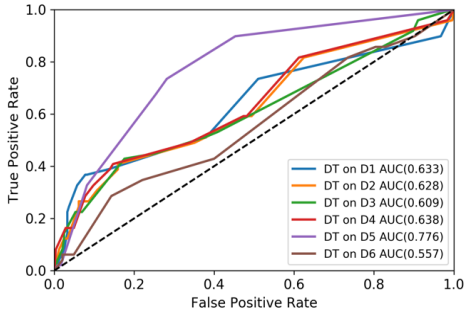


Fig. 2. Decsion tree models' results.

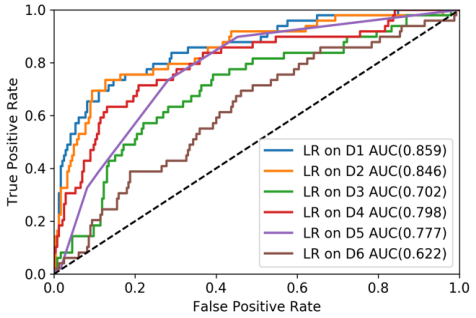


Fig. 3. Logistic regresssion models' results.

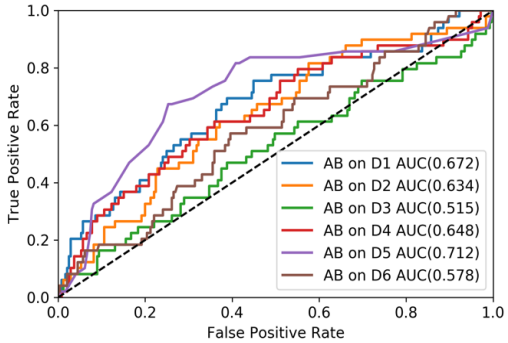


Fig. 4. Adaboost models' results.

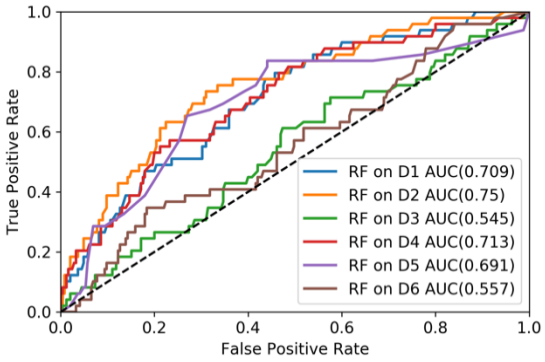


Fig. 5. Random forest models' results.

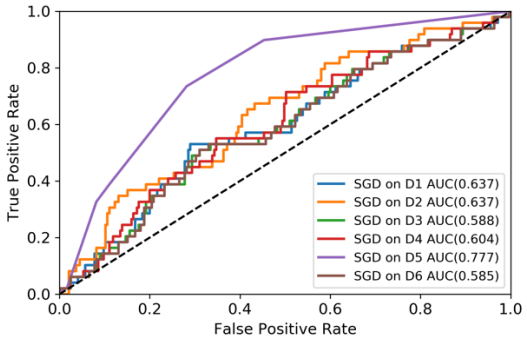


Fig. 6. Stochastic gradient decent models' results.

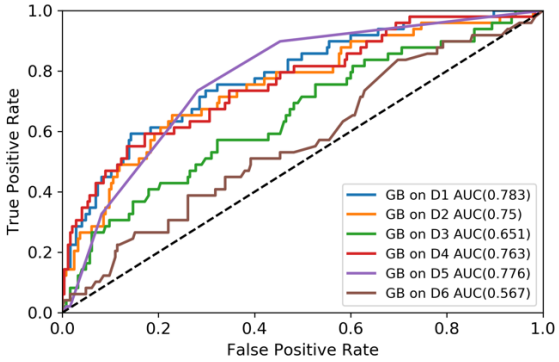


Fig. 7. Gradient boosting models' results.

VI. CONCLUSION AND FUTURE WORK

This paper presented the evaluation of multiple machine learning models on different subsets of features to predict employee attrition using IBM HR dataset. First, five base models were trained and evaluated. Then, three ensembles were constructed using multiple combinations of these five base models. The results show the superiority of the linear model in terms of accuracy, recall and AUC. Although the achieved accuracy demonstrates the capability of the LR model to capture the pattern in the dataset, the authors believe that a higher accuracy rate and a lower generalization error can be achieved by exploring further models' combinations and ensembles using larger and real-world dataset.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Aseel Qutub, Asmaa Al-Mehmadi, Munirah Al-Hssan and Ruyan Aljohani conducted the research; Aseel Qutub, Asmaa Al-Mehmadi, Munirah Al-Hssan and Ruyan Aljohani analyzed the data and evaluated the models; Aseel Qutub,

Asmaa Al-Mehmadi, Munirah Al-Hssan and Ruyan Aljohani and Hanan S. Alghamdi wrote the paper; all authors had approved the final version.

ACKNOWLEDGMENT

The authors would like to thank the IBM data scientists and Kaggle team for developing IBM HR Analytics Employee Attrition & Performance dataset and making it publicly available for developers and researchers. The authors also would like to thank Dr. Kawther Saeedi, Information Systems Department, Faculty of Computing and Information Technology, King Abdulaziz University, for her continues guidance and support.

REFERENCES

[1] M. P. Debono, "Are organisations doing enough to retain their talent?" *The Importance of Employee Retention*, 2018.
[2] J. L. E. E. Liu, "Main causes of voluntary employee turnover: A study of factors and their relationship with expectations and preferences," PhD thesis, Univ. Chile, 2014.
[3] G. V. Sridhar, "Employee attrition and employee retention-challenges & suggestions employee attrition and employee retention-challenges & suggestions," *Rajalakshmi Eng. Coll. Dep. Manag. Stud.*, January 2018.
[4] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, no. 1986, Kluwer Academic Publishers, Boston-Manufactured in The Netherland, pp. 81–106, 1978.
[5] C. Sehra. (2018). *Decision Trees Explained Easily*. [Online]. Available: https://medium.com/@chiragsehra42/decision-treesexplained-%0Deasily-28f23241248
[6] R. Y. Zou and M. Schonlau, *The Random Forest Algorithm for Statistical Learning with Applications in Stata The Random Forest algorithm*, pp. 1–20, 2016.
[7] G. Louppe, "Understanding random forests from theory to practice," PhD dissertation, University of Li ège, 2014.
[8] C. M. Dayton, *Logistic Regression Analysis*, vol. 1, pp. 1–9, 2001.
[9] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Front. Neurorobot.*, December 2013.
[10] Z.-H. Zhou, *Ensemble Methods Foundations and Algorithms*, CRC Press Taylor & Francis Group, 2012.
[11] D. Opitz and R. Maclin, *Popular Ensemble Methods: An Empirical Study*, vol. 11, pp. 169–198, 1999.
[12] R. K. A. A. Indrayan "Receiver operating characteristic (roc) curve for medical researchers RAJEEV," *Perspective*, vol. 48, pp. 277–287, 2011.

Aseel Qutub was born in Jeddah, KSA in 1996. She received the B.S. degree in computer information system from King Abdelaziz University in 2019 and CCNA certification from Cisco and Red Hat certification in Linux.

In 2018, she trained as a system analyst in the Deanship of Information Technology in King Abdulaziz University. Since 2019 she has been working as an IT service officer in National Commercial Bank, Jeddah, KSA.

Mrs. Qutub Award got the best lighting presentation in 11th Annual Undergraduate Research Conference from Zayed University, Dubai, UAE, in 2019.

Asmaa Al-Mehmadi was born in Jeddah, KSA in 1996. She received the B.S. degree in e-systems development, information systems from King Abdulaziz University, in 2019.

In 2018, she trained as a web applications developer and database administrator in the Faculty of Pharmacy, King Abdulaziz University.

In 2019, she has been a technical support and system administrator at Waad holding Company, Jeddah, Saudi Arabia.

Ms. Al-Mehmadi's became a member of Jeddah Data Geeks, in 2019.

Munirah Al-Hassan was born in Riyadh, Saudi Arabia in 1996. She obtained a bachelor's degree in computer information system (IS) from King Abdulaziz University (KAU) in Jeddah Saudi Arabia, in 2019.

In 2018, she had training as a system analyst and application developer in the Deanship of E-Learning and Distance Education in KAU.

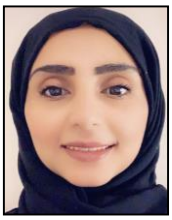
Miss Alhassan got two awards and honors for her senior project, including the second-best poster in the 16th Learning and Technology Conference in Effat University and the Best Lighting Presentation in undergraduate research conference in Zayed University.

Ruyan Aljohani was born in Jeddah, KSA in 1996. She received the B.S. degree in information system form King Abdulaziz University in 2019 and Nanodegree Program in data analysis from Udacity, Jeddah, in 2019.

In 2018, she trained as an application developer in the Deanship of E-Learning and Distance Education.

Since 2020 she has been a system analyst in Shahini Group, Jeddah, KSA.

Mrs. Aljohani got the Best Lighting Presentation in 11th Annual Undergraduate Research Conference from Zayed University, Dubai, UAE, in 2019.



Hanan Alghamdi was born in Jeddah, Saudi Arabia. She received the B.S and the master degrees in computer science from King Abdulaziz University, Jeddah, KSA in 2002 and 2008. She received the PhD from the University of Surrey, Guildford, United Kingdom in computer science 2018.

She worked as a strategic communication manager at the Savola Group in Saudi Arabia and as a demonstrator at the University of Surrey in the United Kingdom. Currently she is working as an assistant professor at King Abdulaziz University, Jeddah Saudi Arabia. Her research interests include machine learning, deep learning and medical image analysis.

Dr. Alghamdi has published several papers in the field of retinal images analysis; examples include "Measurement of Optical Cup to Disk Ratio in Fundus Images for Glaucoma screening", "Ensemble Learning Optimization for Diabetic Retinopathy Image Analysis" and "Automatic Optic Disc Abnormality Detection in Fundus Images: A Deep Learning Approach,". She is involved in several research activities and participated as a reviewer for International Conference on Computational Intelligence and Intelligent Systems, CIIS 2019 and in International Conference on Medical Image Computing and Computer Assisted Intervention, MICCAI 2019 and MICCAI 2020.