IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received March 29, 2021, accepted April 13, 2021, date of publication April 20, 2021, date of current version April 27, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3074559

From Big Data to Deep Data to Support People Analytics for Employee Attrition Prediction

NESRINE BEN YAHIA[1], JIHEN HLEL[1],
AND RICARDO COLOMO-PALACIOS[2], (Senior Member, IEEE)
[1]RIADI Laboratory, National School of Computer Sciences, University of Manouba, Manouba 2010, Tunisia
[2]Computer Science Department, Østfold University College, 1783 Halden, Norway

Corresponding author: Ricardo Colomo-Palacios (ricardo.colomo-palacios@hiof.no)

This work was supported in part by the ERASMUS + KA2 Projects "Information Technology Governance for Tunisian Universities" under Grant 561614-EPP-1-2015-1-ES-EPPKA2-CBHE-JP.

ABSTRACT In the era of data science and big data analytics, people analytics help organizations and their human resources (HR) managers to reduce attrition by changing the way of attracting and retaining talent. In this context, employee attrition presents a critical problem and a big risk for organizations as it affects not only their productivity but also their planning continuity. In this context, the salient contributions of this research are as follows. Firstly, we propose a people analytics approach to predict employee attrition that shifts from a big data to a deep data context by focusing on data quality instead of its quantity. In fact, this deep data-driven approach is based on a mixed method to construct a relevant employee attrition model in order to identify key employee features influencing his/her attrition. In this method, we started thinking 'big' by collecting most of the common features from the literature (an exploratory research) then we tried thinking 'deep' by filtering and selecting the most important features using survey and feature selection algorithms (a quantitative method). Secondly, this attrition prediction approach is based on machine, deep and ensemble learning models and is experimented on a large-sized and a medium-sized simulated human resources datasets and then a real small-sized dataset from a total of 450 responses. Our approach achieves higher accuracy (0.96, 0.98 and 0.99 respectively) for the three datasets when compared previous solutions. Finally, while rewards and payments are generally considered as the most important keys to retention, our findings indicate that 'business travel', which is less common in the literature, is the leading motivator for employees and must be considered within HR policies to retention.

INDEX TERMS Deep people analytics, employee attrition, retention, prediction, interpretation, policies recommendation.

I. INTRODUCTION

Employee attrition or voluntary turnover presents a key issue for organizations as it affects not only their productivity and work sustainability but also their long term growth strategies [1]. On this path, employee retention is a major challenge for recruiters and employers alike, since employee attrition means not only the loss of skills, experiences and personnel but also the loss of business opportunities [2]. In the era of Big Data, people analytics help organizations and their human resources (HR) managers to reduce attrition by changing the way of attracting and retaining talent [3]. In this context, HR analytics is considered as a 'must have' capability for the HR management and profession and "a tool for creating value from people and a pathway to broadening the strategic influence of the HR function" [4]. So, it represents the quantification and the systematic identification of the people drivers of the business outcomes with the purpose of making better decisions. There are interchangeable terms used for HR analytics that are talent analytics, people analytics, and workforce analytics [5]. Thanks to people analytics, HR managers gain the ability to understand their departments and their employees, by providing more accessible and interpretable data about employee attributes, performance and behaviours [6]. Thus, HR analytics plays a significant role in every aspect of the HR function in organizations including recruiting, training and development, retention, engagement and compensation. In the context of HR Analytics, employee attrition analysis has caught more and more attention in the business world. In fact, how to use analytic methods to predict

The associate editor coordinating the review of this manuscript and approving it for publication was Weiping Ding.

whether employees will leave or not can help the organization improve the HR management and save the cost on it [4].

Therefore for the HR managers, it is crucial to have a better idea of what kind of employees will tend to leave and what kind of features will influence them to leave [6]. Most commonly, organizations desire to make sure the right employees are in the right place at the right time and identifying employees' intention to leave by means of analytics [7]. Descriptive analytics are used to summarize or turn data into relevant information so investigate what has occurred. In other words, descriptive analytics have some meaningful impact by explaining what has already happened however, they are not much helpful in predicting what will happen or may happen in the future. On the contrary, predictive analytics have been proposed and used to forecast what will happen in the future. In the field of HR, predictive analytics lead to achievement of organizational benefits and help surely in better decision-making in the organization without any biasness, especially with the most prosperous trend of the big data era and data science basing on machine and deep learning techniques [8]. In fact, data is considered as one of the mandatory ingredients that a people analytics team requires to be effective [9]. Otherwise, HR is set to fail in handling Big Data challenges since Big Data focuses on capturing every piece of available information and collecting every suitable and unsuitable data. But, in HR analytics context, the issue must move from the size of the data to its smartness and making better use of data to create and capture value, being a necessary prerequisite to the more advanced forms of big data analysis [4]. Additionally, [10] highlighted the limits of the application of Big Data within a contextual HR case study, whilst also noting the need to shift the focus from a quantitative to a qualitative analysis of HR data. In this context, the concept of deep data was born to deal with collecting only relevant and specific information and excluding information that might be unusable or otherwise redundant [11].

Thus, in this paper, we mainly focus on two dimensions: a functional dimension and a data dimension. From a functional dimension, we aim to test, compare and select the best accurate predictive model that can early detect employee attrition. We also aim to interpret the positive attrition to find reasons behind it and so to support HR managers to build retention plan. From a data dimension, the key property of the proposed approach, we aim to shift from big data to deep data to address data issues that organizations may face when implementing HR analytics.

Big data is a label commonly used to identify large volumes of (structured or unstructured) data that can generally be defined with the help of the 3Vs volume, velocity, and variety. Volume refers to the quantity of data that are produced by various sources such as sensors, social media, business transactions, etc. Velocity represents the speed at which data are produced, and variety refers to the different formats of data. Over the last decade, the exploitation of big data has become very popular among organizations and these ones tend to adopt new data-driven strategic decision-making models and especially big data analytics across different HR functions [12]. One of the main challenges of using analytics in HR is the deficiency of empirical data. In fact, lack of enough empirical data can be in terms of both the number of candidates or samples, as well as the number of features and this fails to adequately train a reliable model based on such a small dataset. Hence, organizations that plan to use HR analytics first have to face the data availability challenge and they must be able to produce very large volumes of data [13]. Consequently, organizations need large-scale storage solutions that tend to be cloud-based and which require high costs. Moreover, small organizations may not have high-quality HR data and may lack the analytical capabilities to adapt techniques designed for big data to areas where the volume of data is quite small (big data). In this context, the main challenge is the quality of data where organizations must know exactly the data, they need to support their HR analytics functions as HR managers may not have need to all the data they collect. From this point of view, the volume of data is not very important, as what matters in this context is the value of data. Importantly, the identification of deep data, a high-quality data that focus on specific predict trends, is a major barrier to the use of HR analytics for some organizations. So, the main objective of our approach is to shift from big data to deep data perspective and to section down the massive amount of data by excluding useless or duplicate information.

Thereby, in this paper, we aim to propose a deep data-driven predictive approach that can early detect and predict employee intention to leave. Comparing with the related works, this approach focuses on small information-rich HR data within big data. In fact, recent related works such as [14]–[25] and [26] are commonly focusing on finding the best predictive models with high performances to predict employee attrition using generally benchmarks and simulated open data such as HR IBM[1] and HR Kaggle[2] datasets. But, in this paper, we argue that apart from models performances, the HR data must be well constructed and filtered to give relevant and rapid prediction without biases.

Thanks to this deep-data driven approach, which is based on small data providing the greatest business value at a lower cost than vast volumes of big data with regards to the real impactful factors on employee attrition. Thus, the main goals of this research are to: 1) create an effective employee attrition model that contains the necessary and sufficient factors for early detection of attrition intent by deploying a mixed method based on exploratory as well as quantitative analyses, 2) build decision models to predict attrition using Machine, Ensemble and Deep Learning techniques (ML, EL and DL),3) make interpretations to explain and identify the exact reasons behind employee attrition, and 4) make

[1] https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset

[2] https://github.com/ryankarlos/Human-Resource-Analytics-Kaggle-Dataset

recommendations to fight this possible attrition and to take necessary HR management policies.

The outline of this paper is as follows: In the second section, we will present an overview of related works. The research methodology conducted in this research to collect data for our study and to design our final employee attrition model will be presented in the third section. In the ford section, we will present our approach and the various intelligent and predictive models proposed in order to predict employee attrition as soon as possible. The fifth section will show the experimental results as well as the findings of this research i.e. interpretation of the results to understand what makes an employee quit. Finally, we conclude and present an outlook on future works.

II. RELATED WORKS

Literature reports several employee attrition and voluntary turnover predictive models. In this study, we particularly consider recent works that are based on machine and deep learning models applied to the simulated HR datasets of IBM and Kaggle e.g. [14]– [25] and [26]. This choice is motivated by the existence of experiments results of predictive models' accuracy for these open datasets so we can compare them with our proposed models.

IBM HR simulated dataset is a medium sized-dataset provided by IBM and it contains 1470 samples with 34 input features (Age, Business Travel, Daily Rate, Department, Distance From Home, Education, Education Field, Employee Count, Employee Number, Environment Satisfaction, Gender, Hourly Rate, Job Involvement, Job Level, Job Role, Job Satisfaction, Marital Status, Monthly Income, Monthly Rate, Num Companies Worked, Over18, Over Time, Percent Salary Hike, Performance Rating, Relationship Satisfaction, Standard Hours, Stock Option Level, Total Working Years, Training Times Last Year, Work Life Balance, Years At Company, Years In Current Role, Years Since Last Promotion, Years With Current Manager) and its target variable is attrition that is represented as ''No'' (employee did not left) or ''Yes'' (employee left).

Kaggle HR dataset is a large sized-dataset supplied by Kaggle that contains 15000 samples where its target variable is ''left'' and its 9 features are satisfaction level; last evaluation; number project; average monthly hours; time spend company; Work accident; promotion last 5 years; sales and Salary.

In Table 1 authors present an overview of recent solutions to predict employee turnover. For each solution, used datasets and proposed models are presented such as Support vector Machine (SVM), Decision Tree (DT), Logistic Regression (LR), Random Forest (RF), XGBoost (XGB) and K Nearest Neighbors (KNN).

TABLE 1. **Recent related works.**

References	Used simulated HR dataset	Proposed models
[14]	IBM	SVM, RF, KNN
[15]	IBM	LR, KNN, RF, XGB
[16]	IBM	DT, LR, KNN, RF, XGB, SVM, KNN
[17]	IBM	RF, LR
[18]	IBM	SVM
[19]	IBM	SVM, LR, RF
[20]	Kaggle	KNN, SVM, NB, DT, RF
[21]	Kaggle	DT, RF, SVM, MLP, KNN
[22]	IBM	LR, DT, SVM, Voting Classifier
[23]	IBM	LR, DT, KNN, RF
[24]	IBM	LR
[25]	IBM	DT, XGB
[26]	Kaggle	SVM, LR, RF. DT, XGB

While these solutions proposed accurate predictive models to predict employee attrition, they suffer from two major critics: 1) there are no deep studies of employee features selected and used to predict the attrition that justifies the choice of the features. 2) They generally focus only on the employee attrition prediction however for a HR manager it is important to not only predict as soon as possible an employee's intention to leave but also to interpret and explain why the employee has this intention to leave.

III. A MIXED METHOD FOR EMPLOYEE ATTRITION MODELING

As employee attrition or voluntary turnover is a non-avoidable phenomenon, modelling it is a key issue for the process of attrition prediction. In addition, as we aim to adopt a deep data-driven approach, a research methodology that allows us to match theoretical models and experiments must be adopted. That's why we propose to conduct a mixed research method based on the combination of an exploratory research and a quantitative method where the aim is to understand and explain employee attrition phenomena. These two combined methods are used sequentially (e.g., findings from one method inform the other). Thus, such a combined method can leverage the strengths and weaknesses of exploratory and quantitative methods and offer greater insights on a phenomenon that each of these methods individually cannot offer.

In fact, in order to gain a deeper understanding of the phenomenon of high attrition and identifying the factors behind it, an exploratory study based on reviewing available literature is firstly established in detail using studies, papers and open datasets provided by HR experts and researchers. Secondly, these collected features are compared with causal factors for attrition identified through a questionnaire and feature selection techniques (a quantitative research method).

The architecture of the conducted research methodology in this study is depicted in Fig. 1. We will explain in the following sections the different steps of the proposed mixed method.

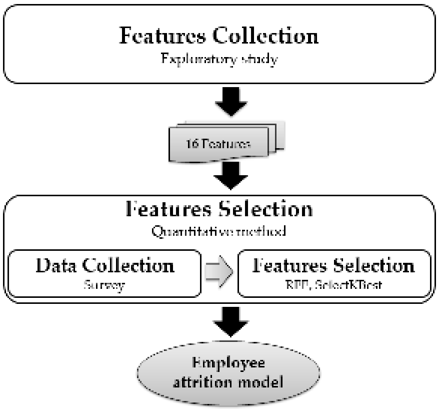


FIGURE 1. Our mixed method for employee attrition modelling.

A. FEATURES COLLECTION: EXPLORATORY STUDY

The first step in this study was to identify and collect employee features that are suitable for our analysis. This step is carried out using an exploratory research of factors responsible for employee turnover that relies on secondary researches by reviewing available literature. Thus, the exploratory research method, conducted in the first step, helped us identifying and collecting adequate and impactful features for our problematic that are most commonly used in different related works and researches in the available literature. In fact, through this exploratory study, based on reviewing of many researches, experiments in HR management and open simulated HR datasets (the fictional data set created by IBM data scientists, and the simulated HR dataset supported by Kaggle) referenced in Table 1, we found that the strongest consensual predictors for employee voluntary turnover are Age, Education, Gender, Job involvement (implication of employee in decision making), Job satisfaction (Career satisfaction), Marital status, Job performance (skills adequacy), Tenure, Promotability (promotions in work), Business Travel, Grade, Rewards (Pay, organization based-rewards, Motivation factors, Salary), Relationship Satisfaction (Hostile organization culture), Environment Satisfaction (favourable or unfavourable working conditions), Training (Training time number, Uncongenial Work environment), Work life/balance. In table 2, we summarize these most cited 16 features that are commonly and frequently used in the available literature.

B. FEATURE SELECTION: QUANTITATIVE METHOD

Following the exploratory study conducted to collect most common factors used in the literature that influence employee attrition, a survey research method is adopted to gather necessary data for the study. Then, some feature selection techniques are also adopted to better filter the chosen features and to end up with a final employee attrition model.

1) DATA COLLECTION: SURVEY

In order to collect employee real data and to tap the factors responsible for attrition in our study, an online questionnaire is prepared and used as a data gathering instrument from respondents (presented in the appendix and accessible through this link). Features collected through the exploratory method have been divided into three parts. Part 1 comprises demographic variables including: Gender, Age, Education, Marital status and Tenure. Part 2 is about their overall level of satisfaction, motivation, involvement and life interest (Job satisfaction, Job involvement, Job performance, Promotability, Environment satisfaction, Rewards, Relationship satisfaction, Business travel, Grade, Training, Work life/balance). Finally, part 3 aims to know the most impactful factors according to respondents and to collect their suggestions (if they are other features that can cause a turnover and so can be integrated into our study). From the designed survey we received450 responses. Respondents were university people from different countries (Tunisia, Norway, France, United States, China, Italy, Pakistan, India, England and Germany). The questionnaire is anonymous. 44,5% of respondents are female and 55,5% are male. Age of the respondents varies from 27 to 62. Out of the total participants, 47,3% want to leave their jobs and 52,7% don't have the intention to quit.

TABLE 2. Features collected after the exploratory study.

Features	References
Age	[2],[1], [14],[15], [16], [17],[18],[19], [22]
Education	[1], [14], [15], [16],[17],[18], [19],[20], [21], [22], [24]
Gender	[2], [1],[14],[15], [16],[17],[18],[19], [22]
Marital status	[1],[14], [15],[17],[18], [19], [22], [24]
Business Travel	[14], [15],[18],[19], [22], [24]
Job satisfaction	[1], [14],[15],[17],[18], [19],[20],[21], [22]
Job involvement	[2],[14],[15],[18],[19], [22], [24]
Job performance	[14],[15], [15],[16],[18],[19], [22], [24]
Relationship Satisfaction	[2],[1], [14],[15],[17],[18],[19], [22], [24]
Environment Satisfaction	[2],[1],[14],[15],[17],[18], [19], [22], [24]
Tenure	[2],[1],[14],[15],[16],[17],[18], [19],[20],[21], [22], [24]
Promotability	[14], [15],[17],[18], [19],[20],[21], [22]
Grade	[2],[14],[15],[18],[19], [22], [24]
Training	[2],[14],[15],[17], [18], [19], [22]
Rewards	[2],[1],[14],[15],[16],[17],[18], [19],[20],[21], [22], [24]
Work life/balance	[2],[14],[15],[17],[18], [19], [22], [24]

2) FEATURE SELECTION METHODS

To improve our first proposition of the employee attrition model, a feature selection procedure will be now followed to better filter features using collected real data from our survey. Feature selection method is the automatic selection of attributes in the data that are most relevant to the predictive modelling problem we are working on. They aim to create an accurate predictive model by choosing relevant features that will contribute to improve the accuracy and removing irrelevant and redundant attributes. In this study, we will

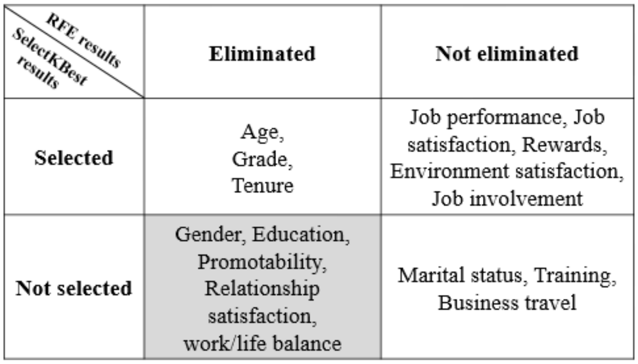
TABLE 3. Vote for keeping/eliminating collected features.

Features	XGB	RF	DT	LR	SVM	Vote results
Age	False	True	True	False	False	Eliminate
Education	False	False	False	False	False	Eliminate
Gender	False	False	False	True	True	Eliminate
Marital status	True	False	False	True	True	Keep
Business Travel	True	True	True	False	False	Keep
Job satisfaction	True	True	True	True	True	Keep
Job involvement	True	False	False	True	True	Keep
Job performance	True	False	False	True	True	Keep
Relationship Satisfaction	False	False	True	False	False	Eliminate
Environment Satisfaction	True	False	True	True	True	Keep
Tenure	False	True	True	False	False	Eliminate
Promotability	False	True	False	False	False	Eliminate
Grade	True	True	False	False	False	Eliminate
Training	True	True	True	True	True	Keep
Rewards	True	True	True	False	False	Keep
Work life/balance	False	False	False	True	True	Eliminate

jointly benefit from two popular feature selection methods namely, Recursive Feature Elimination (a wrapper method) and SelectKBest (a filter method).

Recursive Feature Elimination (RFE) is a feature selection method that fits a model using an external estimator that assigns weights to the features (e.g., the coefficients of a linear model) removes the weakest feature(s). Features are ranked by the model's coefficients or features importance attributes and by recursively eliminating a small number of features per loop RFE attempts to eliminate dependencies and collinearity that may exist in the model. When a predictive model or algorithm assigns the value False to an attribute meaning that the attribute has to be eliminated from the data columns and when the model assigns the value True to an attribute, which should be retained. In this step, we used 5 famous and accurate classifiers (XGB, RF, DT, LR and SVM). As employee attrition prediction is considered here as a classification problem, these classifiers have been chosen because they are the best representatives of the different classification approaches and at the same time they often be have well when dealing with statistical data [27]. Table 3 shows the results of RFE method applied by the 5 classifiers or predictive models. Then, from these results, a majority vote was made to select candidate features of the RFE algorithm, so the selected ones to be eliminated by RFE (that have False values more than True values) are: Gender, Age, Grade, Education, Tenure, Promotability, Relationship satisfaction, and Work/life balance.

SelectKBest is a feature selection algorithm that scores the features of a dataset using a score function and then removes all but the k-highest scored features. It then simply retains the first k features of training set with the highest scores. It is helpful to mention here that we used GridSearch technique to identify the value of k. In fact, we used cross-validation to divide data into three sets (10% for the validation set which is used by GridSearch to find the best hyperparameter k, 70% for the training data and 20% for the test set). The best k recommended by GridSearch here is 8 features for both IBM and Kaggle HR analytics datasets. After applying SelectKBest method to the collected data, the 5 algorithms select the 8 same features which are: Age, Grade, Tenure, Job performance, Job satisfaction, Rewards, Environment satisfaction and Job involvement.



RFE results / SelectKBest results	Eliminated	Not eliminated
Selected	Age, Grade, Tenure	Job performance, Job satisfaction, Rewards, Environment satisfaction, Job involvement
Not selected	Gender, Education, Promotability, Relationship satisfaction, work/life balance	Marital status, Training, Business travel

FIGURE 2. Combination matrix of SelectKBest and RFE results.

Fig. 2 presents a combination matrix of the results of the two feature selection techniques. Then, we propose to retain features that are selected by SelectKBest even though they are eliminated by RFE. Additionally, features that are not selected by SelectKBest and not eliminated by RFE are equally retained. Finally, features that are not selected by SelectKBest and eliminated by RFE are then removed. So, we ended up eliminating the following attributes: Gender, Education, Promotability, Relationship satisfaction and Work/life balance.

In conclusion, according to the combination of the two feature selection techniques (RFE and SelectKBest) and the collected data, the 11 main attritionary features necessary for the employee attrition prediction are: Age, Marital status, Tenure, Grade, Rewards, Job involvement, Training, Business Travel, Job satisfaction, Job performance, and Environment satisfaction.

IV. THE PROPOSED ATTRITION PREDICTION APPROACH

The second part of the study deals with proposing a solution for employee attrition prediction. To do so, we will start this section by an overview of the related works with regards to attrition prediction solutions based on predictive models. Then, we will focus on our proposed predictive approach and its steps details.

With the help of our previous research studies and collected data from the employees' survey, we found the main impactful features on employee attrition which will help us effectively predicting this attrition. The collected and selected data will be considered as an input to our predictive approach that is based on three steps. Fig. 3 presents the architecture of

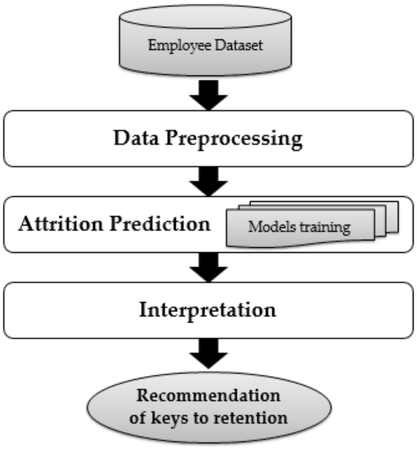


FIGURE 3. Architecture of the proposed approach.

our proposed approach. The first step is data pre-processing. The second one deals with attrition prediction based on machine, ensemble and deep learning models. And, the third one deals with interpretation to explain to HR managers the why of this employee attrition.

A. DATA PREPROCESSING

To better train predictive models, data pre-processing is one of the key steps. To do so, data provided by respondents is transformed and encoded to make them proper for processing and training using the library functions provided and implemented in Python's library scikit-learn [27]. For instance, categorical features were One-Hot Encoded, by which each of the distinct values in the categorical fields was converted to numerical values, and then scaling technique is used to put all the features on a similar scale by normalizing data to ranges from −1 to 1 which avoids outliers to affect the predictions.

B. ATTRITION PREDICTION MODELS

Employee attrition prediction is tackled as a supervised learning problem, and in particular, as a binary classification one. In other words, we are interested in detecting and confirming the existence or not of the employee's intention to leave. To do so, we have put to the test different supervised machine and deep learning techniques, using also the implementations provided in Python's library scikit-learn [29]. In particular, we have adhered to the following classifiers: Decision Tree, Logistic Regression and Support Vector Machine (as machine learning models), Random Forest, XGBoost and Vote Classifier (as ensemble learning models) and three deep learning models (DNN, LSTM and CNN). A grid-search algorithm was performed for each classifier over tuning hyperparameters and the dataset was split 10:70:20 into validation, training and test sets. Then, the different models were trained using their best configuration on the training dataset.

1) MACHINE LEARNING BASED PREDICTIVE MODELS

1. Decision Tree is built through a recursive partitioning process where paths from root to leaf represent classification rules [28]. Each internal node represents a "test" on an attribute, each branch represents the partitioned outcome of the test, and each leaf represents a class label in classification case or a numerical value in regression case.

2. Support Vector Machine is a supervised learning algorithm that is used for linear as well as nonlinear classification problems. To achieve class separation, it uses a hyper-plane or a set of hyper-planes in higher dimensional space. The intuition in this statistical learning based algorithm is that a good separation is achieved by the hyper-plane that has the largest distance to the nearest training data points of any class [29].

3. Logistic Regression is a simple statistical technique and one of the basic linear models for classification that uses the logistic function to model categorical or binary dependent variables. It's often used with regularization in the form of penalties based on L1-norm or L2-norm to avoid over-fitting [30].

2) DEEP LEARNING BASED PREDICTIVE MODELS

1. Deep Neural Networks (DNN), [31], are deep Artificial Neural Networks (ANN) with multiple (at least two) hidden layers where the "deep" refers to the number of hidden layers through which the data is transformed from the input to the output layers. In classical DNN, each layer is composed of a set of neurons and an activation function and is fully connected. A set of weights is affected to each neuron where each weight is multiplied by one input into the neuron. They are then summed to form the output from the neuron after it has been fed through the activation function.

2. Long Short-Term Memory Networks (LSTM) are an amelioration of recurrent neural networks (RNN) that are able to model sequential and temporal data and to predict times series [32]. More specifically, a cell state is added in LSTM to store long-term states and to build more stable RNN for time series prediction by detecting and memorizing the long-term dependencies existing in the time series.

3. Convolutional Neural Networks (CNN) [33], contain generally four types of layers in their structure: an input layer, convolutional layers, pooling layers, and fully connected layer (output). In the convolutional layer, which represents the most important CNN part, the input will be convoluted with different filters where each filter is considered as a smaller matrix. Then, corresponding feature maps will be generated after the convolution operation. The pooling operation consists in reducing the size, while preserving the important features. The efficiency of the network is thus improved, and over-fitting is avoided. So, the main role of convolutional and pooling layers is generally to extract features, and the main goal of fully connected layers is usually to output the information from feature maps together, and then provide them to latter layers.

3) ENSEMBLE LEARNING BASED PREDICTIVE MODELS

The main goal of ensemble learning (EL) is to combine several models in order to find a better solution that gives better results [34]. So, EL is used here to combine the classifiers and their predictions in order to improve robustness over a single classifier. In this study, we will test three ensemble learning models:

1. Random Forest is a popular tree-based ensemble learning technique and a bagging algorithm where successive trees are constructed using a different bootstrap sample of the dataset. By the end, a simple majority vote is taken for prediction. Random forests are different from standard trees as each node is split using the best among a subset of predictors randomly chosen at that node which makes it robust against over-fitting [35].

2. XGBoost is a gradient boosted tree algorithm that involves fitting a set of weak learners and in which final prediction is produced by the combination of predictions from all of them through a weighted majority vote (or sum). This boosting algorithm is based on the use of a regularized-model formalization to control over-fitting, which makes it highly robust and gives it better performance [36].

3. Voting Classifier is an ensemble learning model that trains on an ensemble of classifiers and then predicts the output class basing on a majority vote according to two different strategies. The first one is the Hard Voting where the predicted output class is the class which had the highest probability of being predicted by each of the classifiers. The second one is the Soft Voting where the output class is the prediction based on the average of probability given to that class. In our case, we use a Voting classifier that combines our chosen ML models and that is based on the majority vote strategy (Hard vote) to predict the output class. Such a classifier can be useful for a set of equally well performing model in order to balance out their individual weaknesses.

4. Stacked ANN-based model where outputs of the three chosen deep learners (DNN, LSTM and CNN) are collected to create a new dataset encompassing also for each row the real expected value that will be used to train a new DNN learning model, called meta-learner.

It is helpful to recall here that we used GridSeach for 10% of dataset as validation set to identify the best hyperparameters for each model (such as decision criterion and max-depth for DT, the hidden layers number and units or neurons number in each layer for DNN, LSTM and CNN).

C. INTERPRETATION OF THE EMPLOYEE ATTRITION PHENOMENON

Employee retention refers to organizations' practices and policies that are used to prevent valuable and skilled employees from leaving their jobs [37].

Thus, retention is totally opposite of attrition, it means the ability of organizations to keep their employees, in particular, productive ones, and stop them from going to work somewhere else. In fact, organizations retention policies and all other internal policies governance play a significant role in improving workplace productivity, engaging employees emotionally and, hence controlling attrition. How to retain productive employees and their valued skills is one of the biggest problem that plague organizations, so we aim in this study not only to help HR managers in early detection of employee intention to leave but also to enable them to be aware of the facts leading to employees' attrition, thus they can take few measures and effective management strategies to retain their employees. Indeed, it is equally very important for HR managers to not only have an accurate, but also an interpretable and an explicative predictive model that indicate which features triggering employee attrition and what makes an employee quit.

Thus, in this step of our approach, we will show how we can use our proposed models for attrition interpretation as well as attrition prediction using features importance. It is a statistical method that allows us to evaluate and quantify the participation of each feature in the prediction of the classification task. So, we will use it here to identify attritionary features and to understand these features' influence on employee attrition. Generally, features importance provides a score for each attribute that indicates either how much an attribute contributes to the improvement of the performance, or how much does the model depends on each of its features in the prediction.

So, our aim here is to search for real reasons behind the phenomenon of attrition, so interpretation has to focus only on attritional employees and those who have intention to leave, i.e. only taking into account samples where the value of Attrition = 1 (and we ignore samples where the value of Attrition = 0). Then, we consider "Job satisfaction" feature as our new target because employee job satisfaction is a key ingredient of employee retention. In fact, evidence suggests that employee attrition is triggered by job dissatisfaction and many researchers have shown that the employee satisfaction with job is significantly correlated to the intention to leave [38]. We then proceed to the following steps:

1. Remove rows that present employees who did not leave their jobs or don't have intention to leave (with Attrition = 0).
2. Delete the "Attrition" column and consider "Job satisfaction" as the new target.
3. Convert values of job satisfaction column 1, 2 and 3, 4 into respectively 0 and 1 as satisfied and not satisfied.
4. Apply features importance using the Random Forest (RF) classifier to identify the most impactful features on employee job satisfaction (we choose RF because it is the most performing predictor whereas ensemble method cannot be used here as its inputs are classifiers and not data).

Results of applying features importance on our RF classifier are depicted in Fig. 4.

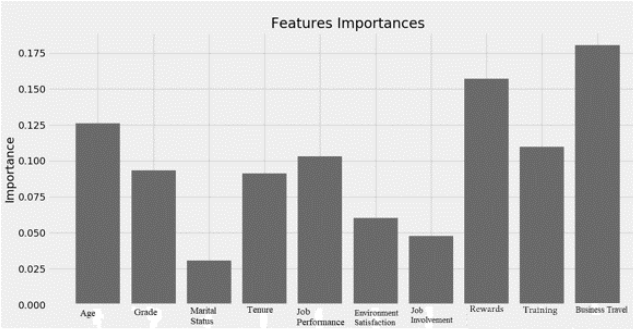


FIGURE 4. Features importance of the random forest model.

V. EXPERIMENTATION RESULTS

After conducting an exploratory and deep data analysis and then identifying all models settings (parameters and hyper-parameters), we are now ready to proceed onto building our models and to assess their performance. Indeed, we will present in this section the experimental results of machine, ensemble and deep learning predictive models. To best assess the performance of these prediction models in a variety of scenarios, the large-sized Kaggle HR simulated dataset (15000 samples), the medium-sized IBM HR simulated dataset (1470 samples) and our small-sized HR real dataset (450 samples) are used. Finally, the salient contribution of these models will be presented towards the end of this experimentation to enable the HR manager not only to predict attrition but also to understand why and so to identify keys to retention. Evaluation criteria for these models and the comparison of their results are explained in following sections.

A. RESULTS OF PREDICTIVE MODELS FOR TWO SIMULATED HR DATASETS

In this section, the two simulated human resources datasets are used to assess the performance of our predictive models. The first one is the large sized-dataset supplied by Kaggle that contains 15000 samples where its target variable is ”left” and its 9 features are satisfaction level; last evaluation; number project; average monthly hours; time spend company; Work accident; promotion last 5 years; sales and Salary. The second simulated human resources analytics dataset is a medium sized-dataset provided by IBM and it contains 1470 samples with 34 features and its target variable is attrition that is represented as ”No” (employee did not left) or ”Yes” (employee left). In this second simulated dataset, we find our 11 selected features as part of its 34 features, so we will check the performance of our predictors using the entire dataset of IBM with its 34 features. Then, we will assess their performance using the same dataset but we will keep only the 11 selected features of our employee attrition model (Marital status, Age, Tenure, Grade, Rewards, Job involvement, Training, Business Travel, Job satisfaction, Job performance, and Environment satisfaction). Table 4 shows the results in terms of accuracy (that is defined as the percentage of the correctly classified data by the model and it represents the ratio of the predictions total number that is correct) and F1-score using the two simulated datasets.

TABLE 4. Performance evaluation of models using the two simulated datasets.

Models	IBM dataset		11 features from IBM dataset		Kaggle dataset	
	Acc.	F1	Acc.	F1	Acc.	F1
DT	0.777	0.318	0.952	0.766	0.972	0.945
LR	0.83	0.368	0.849	0.15	0.782	0.37
SVM	0.85	0.5	0.837	0.42	0.782	0.495
DNN	0.80	0.42	0.84	0.4	0.89	0.18
LSTM	0.71	0.487	0.75	0.57	0.79	0.605
CNN	0.84	0.672	0.89	0.649	0.91	0.7
RF	0.858	0.169	0.953	0.828	0.978	0.967
XGB	0.853	0.434	0.956	0.728	0.976	0.946
VC	**0.93**	0.58	**0.96**	0.62	**0.98**	0.88
Stacked	0.88	0.5	0.9	0.631	0.96	0.67

TABLE 5. Performance evaluation of models using our real dataset.

Models	Data Before feature selection		After feature selection	
	Acc.	F1	Acc.	F1
DT	0.97	0.963	0.98	0.966
LR	0.73	0.582	0.778	0.574
SVM	0.757	0.596	0.787	0.475
DNN	0.81	0.43	0.95	0.891
LSTM	0.69	0.567	0.75	0.8
CNN	0.86	0.72	0.91	0.799
RF	0.98	0.966	0.983	0.972
XGB	0.92	0.93	0.935	0.89
VC	0.96	0.87	**0.99**	0.91
Stacked	0.86	0.79	0.92	0.9

B. RESULTS OF PREDICTIVE MODELS FOR OUR REAL DATASET

In this section, we compare our classification predictors for understanding which predictor is more benefiting to classify churners and non-churners using our real dataset. Models accuracies are measured before and after feature selection algorithms which means that for the first time we use the entire real dataset with its 16 features. Next, models are evaluated using only the 11 features selected after applying the feature selection process by combining RFE and SelectKbest. Results are shown in Table 5.

VI. FINDINGS AND DISCUSSION

In this section, we aim to discuss our experiment results and to put the light on the novelties of this research.

Firstly, regarding the quantitative assessment of our predictors' performance, results depicted in Tables 4 and 5 show that the ensemble learning model Voting Classifier VC performs better than the other models for the simulated as well as

TABLE 6. Comparison of accuracy models for IBM/Kaggle HR datasets with existing works.

Models	DT	LR	SVM	DNN	RF	XGB	VC
IBM HR Dataset							
[14]			0.74		0.71		
[15]		0.89			0.87	0.87	
[16]	0.82	0.87	0.86	0.86	0.86	0.86	
[17]		0.9			0.92		
[18]			0.85				
[19]		0.81	0.77		0.82		0.83
[22]	0.69	0.86	0.87				0.88
[23]	0.79	0.86			0.85	0.9	
[24]	0.85						
[25]	0.83						
Ours	0.77	0.83	0.85	0.8	0.858	0.853	**0.93**
Kaggle HR Dataset							
[20]	0.97		0.78	0.92	0.98		
[21]	0.97		0.94		0.97		
[26]	0.97	0.78	0.94		0.97	0.97	
Ours	0.97	0.78	0.78	0.89	0.978	0.976	**0.98**

real data. In fact, VC outperforms all the other classifiers in terms of accuracy especially when using our real dataset compared to the simulated ones. With regards to the different used machine learning classifiers, the use of ensemble learning VC gives better results in terms of accuracy for both simulated and real dataset regardless of the application of feature selection. In particular, for our final dataset, ensemble learning VC gives the best results with an accuracy of 0.99. This can be explained by the fact that the ensemble learning aims to combine (weak) learners in one method by taking advantage of their complementarity to output best accurate results. In addition, with regards to deep learning predictors, our ensemble learning VC also outperforms them in both simulated and real data. This result may be explained by the quantity of the provided data. In fact, deep learning algorithms require ‘‘relatively’’ large datasets to work well and to give better results, and it also needs the infrastructure to train them in reasonable time. Also, deep learning algorithms require many more experiences and they are more beneficial when we deal with complex problems and real big data with a greater number of features.

Moreover, in order to compare accuracy of our proposed models with regards to recent works that reused the simulated HR datasets, we show in Table 6 different results. We note here that for IBM HR simulated dataset, our ensemble learning VC gives the best results with an accuracy of 0.93. For Kaggle HR simulated dataset, ensemble learning VC equally gives the best results with an accuracy of 0.98.

Apart from proposed predictive models and their combination to get more accurate employee attrition predictions, the salient contributions in this paper basically deal with two points. The first one concerns the proposals of a deep data-driven predictive approach. In fact, our approach focuses on the use of relevant data and the selection of impactful features instead of using all the collected data. It is helpful mentioning here that feature selection gives an effective way to reduce the complexity of classification problems by removing irrelevant and redundant data, which can reduce computation time, improve learning accuracy, and facilitate a better understanding for the learning model. According to the results, those substantiations were experimentally proved here as shown in tables 4 and 5. In fact, an improvement of accuracy measures for most of the classifiers is marked when feature selection is used. We also note an improvement of the F1-score after features selection. This confirms the effectiveness of our chosen employee attrition model in this study and the good results from multiple classifiers after feature selection justify that the selected features are effectively contribute to voluntary attrition. Even for the human resources IBM simulated dataset, predictors' performance has been improved by reducing the number of existing features and keeping only our 11 selected features, and in particular, ensemble method VC accuracy has been slightly increased from 0.93 before feature selection to 0.96 after feature selection. Moreover, ensemble learning VC applied to our final dataset after feature selection gives the best results with an accuracy of 0.99. This also confirms that the choice of SelectKBest and RFE as the two feature selection algorithms is a good one to improve and validate our employee attrition model. So, this deep study also complements previous findings reported in the literature regarding the impactful features on employee attrition and confirms only the need of the 11 selected features.

The second salient contribution in this paper concerns the interpretation and the explanation of the attrition phenomena and so the recommendations for effective retention. According to [37], retention policies fall into three levels of HR management: High, medium and low levels. Each level considers a different perspective and requires a different kind of strategies that can help to combat the problem of attrition arising at that level. In the lower managerial level, understanding and money are keys to retention, whereas, for the medium managerial level managers' appreciation, training and business travel programs act as major keys. Finally, for the high-level management, retention policies include freedom of decision making and creation of a trustworthy environment. Thus, generally, organizations should create an environment that fosters work appreciation and a friendly collaborative atmosphere that makes an employee feel involved and connected to the organization. Especially for our real case study, results of feature importance applied to our RF classifier and plotted in figure 5 show that, for the 450 respondents, the highest importance is assigned firstly to the ''Business Travel'' feature and secondly to 'Rewards'. Meaning that Business travel presents the most motivational attribute and the key to employee retention with regards to the studied dataset. Thus, HR manager should adopt a retention strategy in the medium managerial level and try to organize some business travels for the employees. While rewards, pay and effort–reward imbalance are generally considered as the most impactful variables on employee attrition as in [2] and [16], findings here indicate, however, that one of the leading features identified is less common in the literature: business travel. In fact and as reported in the literature (e.g., [39]), business travel, whether domestic or international, undoubtedly brings benefits for employees and is shown to have a significant effect up and beyond technology transfer through innovation and inspiration from other environments. Indeed, it has been suggested that the experience of visiting clients, other companies, cities and countries broaden employees' understanding of different cultures and make them more open-minded.

At this stage, we assume that there might be some validity threats of our research findings, and we have self-assessed them here in order to denote the trustworthiness of our results, to what extent they are true and not biased by our subjective point of view. In addition, these potential threats are addressed according to the classification proposed in [40]. Regarding the construct validity, we assume that the provided measures could be biased regarding the researchers' expected results. However, we have used in this research, to validate and evaluate the performance of the adopted classifiers, accuracy which is considered as a standard metric often used for measuring performance by reducing biases. They are also robust, particularly for balanced data, which is almost our case here as for our real dataset 47,3% of respondents want to leave their jobs and 52,7% don't have the intention to quit. Regarding the external validity, there might be some issues regarding generalization of our predictive approach as collected data through the employee survey were small data (450 samples) which might indicate a low relevance of the obtained results. To overcome this issue, this approach and its learnt models are assessed on the large-sized Kaggle HR simulated dataset (15000 samples) and the medium-sized IBM HR simulated dataset (1470 samples) which will provide more consistent feedback about the relevance of our results. Finally, regarding reliability, there might be a potential threat that concerns the dependency of data and analysis on the specific researchers. However, we are doing an effort towards trying to minimize this threat by collecting data from different countries with different cultures.

VII. CONCLUSION AND FUTURE WORKS

The main goal of this research is to help HR managers to detect as soon as possible an employee's intention to leave using predictive analytics methods and so to fight this attrition. The contributions can be summarized into three points: i) The proposal of a new employee attrition model that contains only 11 features necessary and sufficient to detect intention to leave and to predict positive attrition using a mixed research methodology. ii) The proposal of machine, deep and ensemble learning predictive models and their experimentation in a variety of different settings (large-sized simulated dataset, medium sized simulated dataset and small-sized real dataset) to best assess their performance. iii)The interpretation and the explication that enables HR managers to understand what makes an employee want to leave and to help them in adopting key policies to retention.

In terms of study limitations, considering dynamic features that deal with employees' behaviour and their emotional states will be promising to study their impact on employee attrition. In this case, the predictive models training must be on-line as data will be dynamic and new data can be added whenever required. We acknowledge also that our questionnaire respondents have equally suggested other features to be considered and that can cause voluntary turnover and so can be integrated into our future study. In fact, they have proposed to consider health issues, job security and the use of new technologies in the company. Finally, in future research, considering unbalanced data is a real challenge especially for organizations and companies with high turnover rate because the adopted predictive models are experimentally not suitable for unbalanced data.

APPENDIX

QUANTITATIVE QUESTIONNAIRE

1. Country:
2. Gender: Female/Male
3. Grade:
4. Age:
5. Education: 1: 'Below College' 2: 'College' 3: 'Bachelor', 4: 'Master', 5: 'Doctor', 6: Other
6. Specialty (Computer Science, Electronics, Mechanics, Business, Medicine, Education, etc.):
7. Marital status: 1: Single, 2: Married, 3: Divorced
8. Organization tenure (number of years at your organization):
9. Years since last promotion in the organization:
10. Rate the degree of your job satisfaction (motivational work, spirit of challenge, contentment with career progress, personal development): 1: Low, 2: Medium, 3: High, 4: Very high
11. Rate the degree of job performance (productivity, skills adequacy) : 1: Low, 2: Medium, 3: High, 4: Very high
12. Rate the degree of environment satisfaction (simple tasks, clear roles, no stressors): 1: Low, 2: Medium, 3: High, 4: Very high
13. Do you feel you are well rewarded for your dedication and commitment towards the work (rewards, Pay)? Yes/No
14. How easy was it for you to get involved in your job (participation in decision making, opinions): 1: Slightly easy, 2: Moderately easy, 3: Very easy, 4: Extremely easy
15. Are you satisfied with your relationships at work (relationship with colleagues and manager)? ∗1: Slightly satisfied, 2: Moderately satisfied, 3: Very satisfied, 4: Extremely satisfied
16. Reward/Salary:
17. Trainings number offered by the organization:
18. How easy was it to balance your work life and personal life while working? 1: Low, 2: Medium, 3: Easy, 4: Very easy

19. How often did you travel for business at that organization? 1: Non-travel, 2: Travel rarely, 3 : Travel frequently
20. Intention to quit the organization Yes/No
21. Any other factors which you feel are responsible for Employee Attrition?

REFERENCES

[1] R. Punnoose and P. Ajit, "Prediction of employee turnover in organizations using machine learning algorithms," *Int. J. Adv. Res. Artif. Intell.*, vol. 5, no. 9, p. 5, 2016, doi: 10.14569/IJARAI.2016.050904.

[2] R. Colomo-Palacios, C. Casado-Lumbreras, S. Misra, and P. Soto-Acosta, "Career abandonment intentions among software workers," *Hum. Factors Ergonom. Manuf. Service Industries*, vol. 24, no. 6, pp. 641–655, Nov. 2014, doi: 10.1002/hfm.20509.

[3] *Amazon.fr—People Analytics in the era of big Data: Changing the way you Attract, Acquire, Develop, and Retain Talent—Jean Paul Isson—Livres*. Accessed: Dec. 15, 2019. [Online]. Available: https://www.amazon.fr/People-Analytics-Era-Big-Data/dp/1119050782

[4] D. Angrave, A. Charlwood, I. Kirkpatrick, M. Lawrence, and M. Stuart, "HR and analytics: Why HR is set to fail the big data challenge," *Hum. Resource Manage. J.*, vol. 26, no. 1, pp. 1–11, Jan. 2016, doi: 10.1111/1748-8583.12090.

[5] A. Tursunbayeva, S. D. Lauro, and C. Pagliari, "People analytics—A scoping review of conceptual boundaries and value propositions," *Int. J. Inf. Manage.*, vol. 43, pp. 224–247, Dec. 2018.

[6] T. Pape, "Prioritising data items for business analytics: Framework and application to human resources," *Eur. J. Oper. Res.*, vol. 252, no. 2, pp. 687–698, Jul. 2016.

[7] S. N. Mishra, D. R. Lama, and Y. Pal, "Human resource predictive analytics (HRPA) for HR management in organizations," *Int. J. Sci. Technol. Res.*, vol. 5, no. 5, pp. 33–35, 2016.

[8] P. Likhitkar and P. Verma, "HR value proposition using predictive analytics: An overview," in *New Paradigm in Decision Science and Management*. Singapore: Springer, 2020, pp. 165–171, doi: 10.1007/978-981-13-9330-3_15.

[9] T. Peeters, J. Paauwe, and K. Van De Voorde, "People analytics effectiveness: Developing a framework," *J. Organizational Effectiveness, People Perform.*, vol. 7, no. 2, pp. 203–219, Jul. 2020, doi: 10.1108/JOEPP-04-2020-0071.

[10] N. Shah, Z. Irani, and A. M. Sharif, "Big data in an HR context: Exploring organizational change readiness, employee attitudes and behaviors," *J. Bus. Res.*, vol. 70, pp. 366–378, Jan. 2017, doi: 10.1016/j.jbusres.2016.08.010.

[11] S. V. Kalinin, B. G. Sumpter, and R. K. Archibald, "Big-deep-smart data in imaging for guiding materials design," *Nature Mater.*, vol. 14, no. 10, pp. 973–980, Oct. 2015, doi: 10.1038/nmat4395.

[12] M. Nocker and V. Sena, "Big data and human resources management: The rise of talent analytics," *Social Sci.*, vol. 8, no. 10, p. 273, Sep. 2019, doi: 10.3390/socsci8100273.

[13] D. Pessach, G. Singer, D. Avrahami, H. C. Ben-Gal, E. Shmueli, and I. Ben-Gal, "Employees recruitment: A prescriptive analytics approach via machine learning and mathematical programming," *Decis. Support Syst.*, vol. 134, Jul. 2020, Art. no. 113290, doi: 10.1016/j.dss.2020.113290.

[14] S. S. Alduayj and K. Rajpoot, "Predicting employee attrition using machine learning," in *Proc. Int. Conf. Innov. Inf. Technol. (IIT)*, Al Ain, UAE, Nov. 2018, pp. 93–98, doi: 10.1109/INNOVATIONS.2018.8605976.

[15] M. Ganesh V. Aishwaryalakshmi, S. Aksshaya, and K. Abinaya, "Predicting employee attrition using machine learning," *Int. J. Sci. Res. Comput. Sci., Eng. Inf. Technol.*, vol. 3, no. 3, pp. 145–149, 2018.

[16] Y. Zhao, M. K. Hryniewicki, F. Cheng, B. Fu, and X. Zhu, "Employee turnover prediction with machine learning: A reliable approach," in *Intelligent Systems and Applications*, vol. 869, K. Arai, S. Kapoor, and R. Bhatia, Eds. Cham, Switzerland: Springer, 2019, pp. 737–758.

[17] X. Gao, J. Wen, and C. Zhang, "An improved random forest algorithm for predicting employee turnover," *Math. Problems Eng.*, vol. 2019, pp. 1–12, Apr. 2019, doi: 10.1155/2019/4140707.

[18] S. N. Khera and Divya, "Predictive modelling of employee turnover in Indian IT industry using machine learning techniques," *Vis., J. Bus. Perspective*, vol. 23, no. 1, pp. 12–21, Mar. 2019, doi: 10.1177/0972262918821221.

[19] S. Karande and L. Shyamala, "Prediction of employee turnover using ensemble learning," in *Ambient Communications and Computer Systems*, vol. 904, Y.-C. Hu, S. Tiwari, K. K. Mishra, and M. C. Trivedi, Eds. Singapore: Springer, 2019, pp. 319–327.

[20] D. S. Sisodia, S. Vishwakarma, and A. Pujahari, "Evaluation of machine learning models for employee churn prediction," in *Proc. Int. Conf. Inventive Comput. Informat. (ICICI)*, Coimbatore, India, Nov. 2017, pp. 1016–1020, doi: 10.1109/ICICI.2017.8365293.

[21] M. M. Alam, K. Mohiuddin, K. M. Islam, M. Hassan, A.-U. M. Hoque, and S. M. Allayear, "A machine learning approach to analyze and reduce features to a significant number for employee's turn over prediction model," in *Intelligent Computing*, vol. 857, K. Arai, S. Kapoor, and R. Bhatia, Eds. Cham, Switzerland: Springer, 2019, pp. 142–159.

[22] S. Shah, S. Alatekar, Y. Bhangare, B. Kasar, and R. Patil, "Analysis of employee attrition and implementing a decision support system providing personalized feedback and observations," *J. Crit. Rev.*, vol. 7, no. 19, pp. 2372–2380, 2020.

[23] F. Fallucchi, M. Coladangelo, R. Giuliano, and E. W. De Luca, "Predicting employee attrition using machine learning techniques," *Computers*, vol. 9, no. 4, p. 86, Nov. 2020, doi: 10.3390/computers9040086.

[24] S. R. Ponnuru, "Employee attrition prediction using logistic regression," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 8, no. 5, pp. 2871–2875, May 2020, doi: 10.22214/ijraset.2020.5481.

[25] S. Kakad, R. Kadam, P. Deshpande, S. Karde, and R. Lalwani, "Employee attrition prediction system," *Int. J. Innov. Sci., Eng. Technol.*, vol. 7, no. 9, p. 7, 2020.

[26] N. Jain, A. Tomar, and P. K. Jana, "A novel scheme for employee churn problem using multi-attribute decision making approach and machine learning," *J. Intell. Inf. Syst.*, vol. 56, no. 2, pp. 279–302, Apr. 2021, doi: 10.1007/s10844-020-00614-9.

[27] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, and J. Van der Plas, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011.

[28] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," *IEEE Trans. Syst., Man, Cybern.*, vol. 21, no. 3, pp. 660–674, May 1991.

[29] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statist. Comput.*, vol. 14, no. 3, pp. 199–222, Aug. 2004, doi: 10.1023/B:STCO.0000035301.49549.88.

[30] G. King and L. Zeng, "Logistic regression in rare events data," *Political Anal.*, vol. 9, no. 2, pp. 137–163, 2001.

[31] G. E. Hinton, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, Jul. 2006, doi: 10.1126/science.1127647.

[32] R. Pascanu, C. Gulcehre, K. Cho, and Y. Bengio, "How to construct deep recurrent neural networks," 2013, *arXiv:1312.6026*. [Online]. Available: https://arxiv.org/abs/1312.6026

[33] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, and T. Chen, "Recent advances in convolutional neural networks," *Pattern Recognit.*, vol. 77, pp. 354–377, May 2018, doi: 10.1016/j.patcog.2017.10.013.

[34] G. Brown, "Ensemble learning," in *Encyclopedia of Machine Learning*, vol. 312. 2010, pp. 15–19.

[35] A. Liaw and M. Wiener, "Classification and regression by randomForest," *R News*, vol. 2, pp. 18–22, Dec. 2002.

[36] J. H. Friedman, "Greedy function approximation: A gradient boosting machine.," *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001.

[37] B. K. Goswami and S. Jha, "Attrition issues and retention challenges of employees," *Int. J. Sci. Eng. Res.*, vol. 3, no. 4, pp. 1–6, Apr. 2012.

[38] A. H. Khan and M. Aleem, "Impact of job satisfaction on employee turnover: An empirical study of autonomous medical institutions of Pakistan," *J. Int. Stud.*, vol. 7, no. 1, pp. 122–132, May 2014, doi: 10.14254/2071-8330.2014/7-1/11.

[39] J. V. Beaverstock, B. Derudder, J. R. Faulconbridge, and F. Witlox, "International business travel: Some explorations," *Geografiska Annaler, B, Hum. Geogr.*, vol. 91, no. 3, pp. 193–202, Sep. 2009, doi: 10.1111/j.1468-0467.2009.00314.x.

[40] P. Runeson and M. Höst, "Guidelines for conducting and reporting case study research in software engineering," *Empirical Softw. Eng.*, vol. 14, no. 2, pp. 131–164, Apr. 2009, doi: 10.1007/s10664-008-9102-8.

NESRINE BEN YAHIA is currently a Doctor-Engineer in computer sciences. She is also an Associate Professor with the National School of Computer Science (ENSI), where she is a Coordinator of the master's degree in smart systems and the Chief of the Information and Decision Systems Department. She participated in scientific national and international projects. Her current research interests include artificial intelligence, cooperative systems, CSCW, knowledge engineering, social networks analysis, and intelligent decision support systems. Her teaching interests include machine and deep leaning, software engineering, UML, software design, and software reengineering.

JIHEN HLEL received the master's degree in computer sciences from the National School of Computer Sciences, in 2019. She is currently pursuing the Ph.D. degree. Her current research interests include artificial intelligence, machine and deep leaning, and intelligent decision support systems.

RICARDO COLOMO-PALACIOS (Senior Member, IEEE) received the bachelor's, master's, and M.B.A. degrees from the Instituto de Empresa, in 1994, 1997, and 2002, respectively, and the Ph.D. degree in computer science from the Universidad Politécnica of Madrid, in 2005. He is currently a Full Professor with Østfold University College, Norway. He has been working as a software engineer, a project manager, and a software engineering consultant with several companies, including Spanish IT Leader INDRA. His research interests include applied research in information systems, software project management, people in software projects, business software, and software and services process improvement. He is also an Associate Editor in journals, like *IEEE Software*, IEEE Access, and *Computer Standards & Interfaces*. He has edited several special issues in journals, like *Journal of Software: Evolution and Process*, *Software Quality Journal*, *Science of Computer Programming*, and *Future Generation Computer Systems*.