EMPLOYEE ATTRITION PREDICTION USING MACHINE LEARNING MODELS: A REVIEW PAPER

Haya Alqahtani, Hana Almagrabi and Amal Alharbi

Department of Information Systems, Faculty of Computing and Information Technology
King Abdulaziz University, Jeddah, Saudi Arabia

ABSTRACT

Employee attrition refers to the decrease in staff numbers within an organization due to various reasons. As it has a negative impact on long-term growth objectives and workplace productivity, firms have recognized it as a significant concern. To address this issue, organizations are increasingly turning to machine-learning approaches to forecast employee attrition rates. This topic has gained significant attention from researchers, especially in recent times. Several studies have applied various machine-learning methods to predict employee attrition, producing different resultsdepending on the employed methods, factors, and datasets. However, there has been no comprehensive comparative review of multiple studies applying machine-learning models to predict employee attrition to date. Therefore, this study aims to fill this gap by providing an overview of research conducted on applying machine learning to predict employee attrition from 2019 to February 2024. A literature review of relevant studies was conducted, summarized, and classified. Most studies agree on conducting comparative experiments with multiple predictive models to determine the most effective one.From this literature survey, the RF algorithm and XGB ensemble method are repeatedly the best-performing, outperforming many other algorithms. Additionally, the application of deep learning to employee attrition prediction issues also shows promise. While there are discrepancies in the datasets used in previous studies, it is notable that the dataset provided by IBM is the most widely utilized. This study serves as a concise review for new researchers, facilitating their understanding of the primary techniques employed in predicting employee attrition and highlighting recent research trends in this field. Furthermore, it provides organizations with insight into the prominent factors affecting employee attrition, as identified by studies, enabling them to implement solutions aimed at reducing attrition rates.

KEYWORDS

Employee Attrition, Prediction, Machine Learning, Ensemble Methods, Feature Selection

1. INTRODUCTION

The workforce is the most valuable asset in any company or organization [1][2]. It plays a fundamental role in achieving the organization's vision and mission[3]. Executives spend significant time selecting qualified employees and investing in employee training[2]. However, one of the most prominent issues organizations face is employee attrition[4].

Employee attrition is the term that refers to a decrease in the number of employees in a company due to employees' voluntary or involuntary departures from their jobs[4][5]. Employees choose to leave their positions for various reasons, such as work pressure, an unsatisfactory work environment, or inadequate compensation[6]. According to a McKinsey survey, nearly half (48%) of workers are leaving their current roles to transition into a new industry[7].

DOI:10.5121/ijaia.2024.1520223

The issue of employee attrition has garnered attention from many organizations due to its negative impact[8]. Employee turnover rates have reached their highest levels in the past decade, forcing many organizations to deal with this challenge [9]. This urgency prompts leaders to proactively address turnover rates to mitigate their adverse effects on business activities and operations[9].

The impact of this issue can be evaluated using various metrics, such as sales growth, return on equity, customer service quality, and profitability[4]. It has been shown that a high attrition rate presents significant challenges for businesses, as companies invest significant resources—financial, temporal, and material—in training employees for specific roles, incurring substantial losses when these employees leave[10]. On the other hand, organizations are forced to allocate additional resources to human capital, recruitment, orientation, and the development of new hires[8][11]. Importantly, the most significant expenses for organizations are investments in their human resources, including employee benefits, training, salaries, bonuses, and other forms of compensation[12].

Research indicates that voluntary employee turnover can cost an organization more than one and a half times an employee's annual salary, taking into account total expenses such as reassignment, recruitment, and training of replacements [13].Therefore, the significant impact of staff attrition on firms highlights the increased interest in this topic [4].

In such a context, the importance of using analytics in the human resources sector becomes increasingly clear, as it enables HR professionals to gain comprehensive insights into their personnel and make informed decisions [11]. Employee analytics provides human resources managers with easily accessible and understandable data about their workforce's characteristics, performance, and behavior, enhancing their understanding of their departments and employees[14]. Descriptive analytics are used to analyze data or translate it into relevant information, shedding light on past occurrences but offering limited insight into future events [14]. On the other hand, predictive analytics have emerged as a valuable tool for forecasting future events, facilitating improved and unbiased decision-making within organizations[14][15]. By leveraging historical data trends, machine-learning classifiers, such as data mining algorithms, help predict future employee behaviors, such as performance and attrition rates. As a result, through the mining and analysis of historical employee data, HR can anticipate future organizational behaviors [16].

The use of machine learning for predicting employee attrition has attracted considerable research attention, especially in recent years. The growing amount of data in this field has resulted in numerous studies investigating this topic [6].

Previous studies have used different machine-learning methods to forecast employee attrition, each helping to improve solutions for this issue. However, the results of these studies have differed based on the tools and methodologies used. Despite the substantial amount of research, as far as we know, there is no comprehensive review that compares the different research efforts that have employed machine-learning models to predict employee attrition.

Therefore, this study aims to provide an overview of the research conducted in the field of using machine learning to predict employee attrition in recent years, from 2019 to February 2024. A thorough literature review was conducted, gathering relevant studies published in electronic journals, which were then summarized and categorized. The goal of this study is to offer a concise review for new researchers, helping them understand the main approaches used in this field.

The rest of the paper is organized as follows. The second section discusses the issue of employee attrition. The third section reviews the factors that influence employee attrition, as explored in previous studies. The fourth section focuses on predicting employee attrition, including a classification of previous work in this field, the approaches used, and a review of the resources and techniques employed. The fifth section provides a comprehensive discussion of the previous sections, while the paper concludes with Section 6.

2. EMPLOYEE ATTRITION AS A PROBLEM

Employee attrition refers to the decline in the workforce of an organization whether employees leave the organization voluntarily or are retired. Employee turnover is the number of current employees who are replaced by new hires for a set amount of time. High attrition leads to high employee turnover in any organization[17]."Turnover" and "attrition" are two trading terms that are constantly at odds with one another. The term "attrition" describes a decline in the workforce. While "Turnover" occurs when a company searches for a replacement for a departing employee. These terms can be used interchangeably for workforce data analysis and other metrics required for workforce planning[18].

There are two types of employee attrition: voluntary attrition and involuntary attrition. Workers dismissed by their employers for a variety of reasons, such as poor performance or increased workload, are said to have undergone involuntary attrition. Voluntary attrition is when high-performing workers who choose to quit the company voluntarily do so despite the company's efforts to keep them[19].

The main objectives of human resource management are to attract, inspire, and keep talent within the company by creating management processes that guarantee that human talent is utilized efficiently [20].Although companies invest in their workforce by offering exceptional training and a pleasant working environment because they understand the value of their employees, they also experience voluntary attrition and the loss of brilliant individuals. Finding replacements is another issue that comes with substantial expenditures for the business, such as those connected to employment, training, and recruiting[19].

Employee attrition is one of the main issues facing businesses. According to statistics, 68% of workers quit their employment voluntarily because of a variety of problems that can be controlled at work, such as poor management, a lack of benefits, and a lack of recognition[21].High attrition rates prove to be costly for the company as the company invests time, money, and assets to train employees to make them job-ready in a particular company. If workers depart, the company not only loses their valuable employees, but also the money invested in finding, choosing, and preparing them for their positions. On the other side, the company must continually increase its investment in new hire recruitment, training, and development[10][22]. In addition, new hires will also go through a similar learning curve as seasoned internal staff members to reach equivalent proficiency levels in technical or business expertise[23].In addition, this issue may cause significant problems for organizations, such as lower earnings brought on by the departure of productive workers[21].

When highly effective individuals leave the company in search of better opportunities, the losses incurred when a competent employee quits are not limited to advanced product beliefs, great project management, or connections with customers. This can hurt companies as their productivity drops significantly, hampering organizational morale [10].

Additionally, employee attrition results in productivity and performance that are below optimal levels[3], and loss of effective project management, which leads to the dissatisfaction of

customers and other stakeholders[4][10]. It also risks leaking basic knowledge and technologies to competitors[15]. Studies have also revealed that organizations with higher voluntary turnover rates perform worse than their competitors, endangering their future market opportunities[13]. More significantly, a company is more likely to experience reputational harm and harm to its brand, from which it will be challenging to regain [9][21].Consequently, it makes sense that the reason behind the great interest in employee attrition is that it has serious consequences for organizations[4].

It has become more important than ever for companies to do their best to reduce employee turnover, given the proven detrimental effects of such turnover on their operations, and because of these effects, by implementing regulations and creating policies that are more satisfactory to employees and thus retain them.Companies can take advantage of the large amounts of employee data they have to help address this problem.

3. LITERATURE REVIEW: FACTORS INFLUENCING EMPLOYEE ATTRITION

While human resources are indeed valuable assets crucial to a company's success, one of the most significant challenges in human resource management is employee attrition [24]. Addressing this issue primarily involves identifying the underlying causes and researching associated factors. In this section, we review studies that have identified or examined some of these factors and their impact on employee attrition.

For instance, the authors in [25]identified five primary causes of voluntary employee turnover. These include lack of recognition, insufficiently competitive compensation systems, recruitment practices, managerial style, and a toxic work environment.

According to the study in [26], employee attrition occurs due to various reasons, including a lack of support for employees in lower positions within the organization, inadequate communication channels between these workers and their superiors, monotonous tasks that result in a loss of creativity, limited career prospects, and a lack of opportunities for advancement.

The authors in [27]emphasized two types of factors that contribute to employee turnover. These include job-related factors such as job stress, discordance regarding job tasks and duties, and job dissatisfaction, as well as organizational factors such as organizational instability, deficient communication systems, and limited employee involvement in decision-making processes. Additionally, the authors in[28]confirmed that employees leave their positions due to feelings of being underpaid, disagreements with management, issues with working hours, a lack of career growth opportunities, or family-related concerns. Furthermore, in addition to the mismatch between job roles and employee skills, other contributing factors include inadequate training, limited opportunities for professional growth, lack of recognition or appreciation, poor work-life balance resulting in excessive work-related stress, and diminished confidence in leadership[22].

Another study [29], identified administrative factors such as unstable management, inadequate work environments, lack of essential facilities, low salaries, and insufficient financial incentives as significant contributors to turnover. Additionally, the study highlighted further influences on the turnover rate, including limited opportunities for career advancement, influence from coworkers, alignment between the candidate and the job, and the availability of alternative job prospects.

According to the authors in[30], insufficient training and inadequate training duration can leave workers feeling unprepared to meet the requirements of their roles. The lack of opportunities for development and advancement, where organizations fail to provide career paths and training

opportunities for skilled and ambitious professionals, leads to the loss of talented employees. Incorrect job profiles also contribute to high turnover rates. Additionally, ineffective management and limited decision-making opportunities contribute to attrition [30].

Similarly, the author in [31]identified several causes of attrition, including a perceived lack of career growth prospects, ineffective appraisal systems that fail to adequately recognize individual efforts and contributions, management's indifferent attitude towards employees, limited long-term growth opportunities, inadequate provisions for ongoing learning, poor quality of HR practices, ineffective leadership, uninspiring leadership leading to reduced enthusiasm at work, inflexible work arrangements and inadequate compensation, excessive pressure with insufficient recognition, failure of the organization to adhere to ethical practices and guidelines, and unsafe work environments.

Another study [32]which focused on employees in private financial institutions, identified several factors that contribute to attrition. The study concluded that the strained relationship between management and employees, employee dissatisfaction, lack of job security, and job pressure were significant contributors to employee attrition [32].

Furthermore, research conducted by several scholars has highlighted various factors that influence turnover rates. These factors can be categorized into organizational factors, job-related factors, work environment factors, and personal factors [33].

4. EMPLOYEE ATTRITION PREDICTION

Several factors, such as personal, job-related, and organizational characteristics, have been proven to influence employee attrition, making them valuable predictors of attrition rates within an organization [16].

The process of machine-learning prediction involves training an algorithm using a historical dataset. Afterwards, for each record in new data, the algorithm generates potential values for unknown variables [34].

managers and HR departments to assess an employee's intention to leave the organization, considering factors such as length of employment and performance. As a result, prediction models can be created using historical data to evaluate employee performance, engagement, and behavior within the organization [16].

Machine-learning models are trained on preprocessed data, allowing them to make informed decisions when presented with new data. The primary objective of these models is to identify patterns within the data and apply the knowledge gained to make predictions [1].There are two main types of models: supervised learning and unsupervised learning. In supervised learning, the training set is labeled by humans, which enables the model to classify new data and predict outcomes once it learns the relationship between input and output data. On the other hand, unsupervised learning processes large amounts of data without human supervision, as there are no labeled instances[35].

To validate machine-learning classification models developed using current employee data, accuracy tests, and validations can be performed. This may involve using a different dataset or testing the trained prediction model using 20% of the main dataset that was not used during the training [36].

Efficiency and scalability are vital for machine-learning algorithms to effectively extract information from large datasets or dynamic data streams. Metrics such as classification accuracy and training time are crucial for evaluating the performance of data mining algorithms and selecting the best algorithms for classification or prediction tasks [37].

Applying machine learning and data mining with various optimization algorithms has consistently yielded valuable insights from different datasets [38].Predictive analysis uses these algorithms to forecast future trends, helping identify potential employees who may leave the company. This proactive approach enables quick decision-making and reduces the loss of human resources[38][39], promoting continuous growth and maintaining a higher business[16].

The following subsections will review the types of datasets used in previous studies and the approaches employed to predict employee attrition. Additionally, the focus will be on studies that utilized ensemble methods to enhance forecasting performance and discussions on feature selection techniques that improved prediction accuracy.

4.1.Datasets

Researchers have applied different machine-learning algorithms to different datasets in studies predicting employee attrition. Some studies relied on previous ready-made and available databases that were collected either for the study itself or other purposes, and some studies collected the study database from real employee data. In this sub-section, previous work is classified according to the database used:

4.1.1.Available Datasets

IBM HR Analytics Employee Attrition Dataset: The fictional dataset was created by IBM data scientists in 2016[21].It consists of 1470 observations and includes 34 features that are relevant to employees' work and personal data. The dataset also includes a binary target variable called "attrition," where "yes" indicates an employee who left the company and "no" indicates an employee who stayed[1]. However, the dataset is highly unbalanced, with 237 positive samples (former employees) and 1233 negative samples (current employees), which significantly skews the data[6]. Various researchers have conducted experiments using this dataset to predict employee attrition, and these studies are referenced in [15], [22], [23], [40], and others.

Datasets Available on Kaggle: Some researchers have used pre-existing datasets available online, such as those provided by the Kaggle website. For instance, the dataset utilized in [11]for human resource management is easily accessible on kaggle.com and comprises approximately 14,000 records with ten attributes pertaining to employee attrition problems.

In[14], the performance of predictive models was assessed using two simulated human resources datasets. The first dataset, obtained from Kaggle, consists of 15,000 samples with the target variable "left" and nine features, including time spent in the company, work accidents, promotions in the last five years, sales, satisfaction level, number of projects, average monthly hours, and salary. The second dataset is a medium-sized dataset obtained from IBM.

Dataset from SAS: In [8]], researchers utilized a real dataset from SAS, which consisted of 1.5k rows and 35 variables/columns. The dataset included categorical and interval variables and was sourced from the SAS library (www.sas.com).

4.1.2.Collected Employees Dataset

Many previous studies have collected datasets to analyze employee data and predict attrition. These datasets were either sourced from specific institutions and organizations or gathered through surveys that covered factors that could potentially contribute to workers leaving their jobs.

In[10], the authors proposed a method for predicting the rate of employee attrition and assessing employees' emotional well-being within the company. They collected data through a questionnaire that included attrition-related questions and the Goldberg Depression Questionnaire. The dataset considered various elements such as workplace equality, job satisfaction, overtime, working hours, travel, personal life, and others. Additionally, the authors analyzed the depression questionnaire using Goldberg's Depression Questionnaire [10].

In[41], the authors describe and examine a Europe-wide study on employee turnover intention. This study was based on the Global Employee Engagement Index (GEEI) survey conducted by Effectory in 2018. The survey included 18,322 employees from 56 European countries and consisted of 123 questions that covered socio-demographic, labor, and industry aspects, as well as an objective question [41].

The study conducted by[4]. identified the most significant factors contributing to employee attrition using data from the human resources department of an Iranian pharmaceutical company. The dataset included 577 employee observations and comprised 31 characteristics, including nine numeric-type variables, three logical variables, and 19 character-type variables [4].

In a separate study [42], the researchers investigated the correlation between employee performance and organizational growth. For this research, data on employees was collected from a well-established MNC business in Chennai. The company employs around 5000 individuals in different departments, age groups, and locations [42].

In an Indian study[16], employee information was obtained from the HR datasets of three IT companies in India. The dataset included 22 input features, and the employees' job status at the time of data capture was considered as the response variable [16].

In the study[14], an online questionnaire was created and used to collect information from participants. The goal was to gather authentic employee data and identify factors contributing to attrition. The collected features were divided into three sections using an exploratory approach: the first section covered demographic variables, the second addressed overall levels of satisfaction, motivation, engagement, and interest in life, and the third aimed to identify the elements participants considered most significant in causing turnover. The database consisted of 450 responses from individuals across various universities worldwide, including Tunisia, Norway, France, and the United States. Around 53% of participants expressed no intention to leave their current positions, while 47% indicated a desire to do so [14].

To predict employee attrition, researchers in the study [43]used an authentic HR dataset from employees of a Kazakhstani higher education institution. The dataset contained actual 2023 data on employee demographics, job characteristics, gender, and other relevant factors. The dataset included approximately 1500 observations, with instances of employee attrition accounting for 39% of the total[43].

4.2. Approaches

During the literature review of studies on predicting employee attrition, it is noteworthy that most studies relied on either traditional machine-learning approaches or deep-learning methods. In this

subsection, we categorize previous studies based on the approach used to predict employee attrition.

4.2.1.Machine Learning

In a study by [44], four machine-learning algorithms—logistic regression (LR), random forest (RF), naïve Bayes (NB), and k-nearest neighbors (KNNs)—were compared to predict employee attrition. The study used the IBM HR Analytics Employee Attrition dataset for model training. To address class imbalance, the synthetic minority oversampling technique (SMOTE) was applied. The LR algorithm performed the best with SMOTE and weighted class techniques, achieving an accuracy of 90% [44].

In another study[21], researchers used the IBM HR Analytics Employee Attrition dataset to assess the effectiveness of various machine-learning methods in predicting staff attrition. The algorithms were selected following a comprehensive evaluation of all machine-learning algorithms offered by MATLAB. The study aimed to compare the performance of five selected algorithms—LR, linear support vector machines (LSVMs), quadratic support vector machines, support vector machines kernel, and boosted trees—using default MATLAB parameters. The findings revealed that LR achieved the highest accuracy of 87.78%, making it the best algorithm for forecasting this issue [21].

In[1], an initial exploratory investigation was proposed to use machine-learning techniques to predict employee attrition. Five machine-learning models—LR, NB classifier, RF, classification trees, and a simple neural network (NN) architecture—were used. LR emerged as the most effective technique, achieving an accuracy of 88%[1].

In[45],the IBM HR Analytics Employee Attrition dataset was used to analyze how objective factors influence employee attrition and predict employee departure from the company. Various machine-learning algorithms, including decision tree (DT), Gaussian naïve Bayes (GNB), KNN, support vector machines (SVMs), RF, LR, and LSVMs, were applied. The GNB algorithm performed the best, achieving the highest recall rate of 0.54. Additionally, it was noted that the “Research and Development” department had the highest percentage of resigned workers (133 out of 237 employees), with an attrition rate of 13.8%, while the “Sales” and “Human Resources” departments had attrition rates of 20.6% and 19%, respectively[45].

In[20], the objective was to develop a model using supervised classification and machine-learning techniques to predict voluntary job turnover. The study described various professional and personal variables that could influence employee attrition rates. The correlation matrix analysis revealed a positive association between years at the company, years with the current manager, total working years, and monthly income, while job level, tenure in the current role, total years worked, and monthly salary were negatively correlated with attrition. Real data from IBM was used to implement Multinomial NB, GNB, Bernoulli NB, DTs, RFs, and LR algorithms for prediction. The GNB algorithm, known for its ability to minimize false negatives, achieved the highest recall rate of 70.76% [20].

In[10], the authors proposed predicting employee attrition rates and evaluating employee emotions within the company. They included employee attrition-related questions and the Goldberg Depression Questionnaire in their survey. Various graphs were used to establish connections between different variables, and machine-learning algorithms such as KNN, DT, SVM, LR, RF, and GNB were employed. The RF model emerged as the top performer, achieving an accuracy of 86% [10].

Similarly, in a study [40]that utilized the IBM HR Analytics Employee Attrition dataset, researchers compared the performance of various classification models, such as DT, RF, XGBoost (XGB), LR, and SVM, in predicting employee attrition. The RF classifier-based model showed superior predictive capability, achieving an accuracy score of 97%. Furthermore, the Chi-square test identified several factors that significantly impact an organization's profit levels, including overtime, marital status, monthly income, job role, age, total working years, distance from home, stock options level, job satisfaction, and years at the company [40].

In[43], researchers used a real HR dataset from Kazakhstani higher education institutions. The dataset included demographic, job characteristic, gender, and other relevant data from 2023. It consisted of approximately 1500 observations, with 39% of them representing cases of employee attrition. Machine-learning models, including LR, KNN, Multinomial NB, GNB, SVC, and DT, were applied. However, the initial model predictions were considered insufficiently accurate, leading to the use of optimization techniques such as Lasso regression. After optimization, DT emerged as the most accurate model, achieving a score of 95% [43].

In another study[46], the authors focused on systematically predicting attrition using Machine Learning and Data Analysis methods. They utilized the IBM HR Employee Analytics Attrition and Performance dataset. To address the imbalance in the dataset, SMOTE was employed to generate entries for the minority class. Various machine-learning techniques, including DTs, SVMs, KNN, RF, and NB, were applied. The RF classifier, with feature reduction, achieved an accuracy score of 83.3%, while the DT classifier reached 81%. This indicates the superiority of these classifiers over the others[46].

Additionally, a framework for predicting employee attrition using an LR model was presented in[2].By utilizing the IBM HR Analytics Employee Attrition dataset, the model's accuracy improved from 78% to 81% with feature selection, which highlights its effectiveness [2].Table 1 shows a summary of previous studies that applied the machine learning approach, highlighting the results, strengths, and limitations of each study.

Table 1. Summary of previous studies that applied the machine learning approach.

Ref	**Date**	**Dataset**	**Used Models**	**Result**	**Strengths**	**Limitations**
[44]	2023	IBM HR Analytics Employee Attrition dataset	LR, RF,NB, and KNNs	LR performed the best, accuracy = 90%	Results were improved by addressing the class imbalance (weighted classes and SMOTE)	The dataset used is limited and does not account for every psychological and subjective factor.
[21]	2022		LR,LSVMs, SVM, boosted tree	LRachieved the highest accuracy =87.78%.	MATLAB Classification Learner was applied to test 19 ML algorithms in 6 iterations.	Limited dataset.All factors were Taken into consideration without filtering them based on the

						most important.
[1]	2022		LR, NB, RF, classification trees, and NN	LR performed the best achieving an accuracy=88%	The main employee attrition reasons were analyzed.	Limited dataset. And the data imbalance was not handled.
[20]	2021		NB, GNB, Bernoulli NB, DTs, RFs, and LR	The GNB algorithm achieved the highest recall rate of 70.76%	The correlation between features was Analyzed.	
[45]	2020		DT, GNB, KNN,SVMs, RF, LR, LSVMs,	GNBperformed the best, achieving the highest recall = 0.54	Several basic indicators of employee attrition were Identified.	Limited dataset. Feature engineering was not applied.
[10]	2021	The dataset collected by aquestionnaire	KNN, DT, SVM, LR, RF, and GNB	RF emerged as the top, achieving an accuracy=86%	Real dataset, including the Goldberg Depression Questionnaire.	Improved methods can be applied to make models more accurate.
[43]	2023	real HR dataset from Kazakhstani higher education institutions	LR, KNN, Multinomial NB, GNB, SVC, and DT	DT emerged as the most accurate model, achieving a score of 95%.	A real HR dataset. Various optimization techniques were employed to enhance the performance of the model like Lasso .regression	The features of employee attrition were not analyzed.
[46]	2019	IBM HR Analytics Employee Attrition dataset	DTs, SVMs, KNN, RF, and NB	RF, with feature reduction, achieved an accuracy = 83.3%, while the DTreached 81%. This indicates the superiority of these classifiers over the others	SMOTE was applied to handle data imbalance. Feature Selection techniques were performed to improve prediction performance.	The dataset used is limited and does not account for every psychological and subjective factor.
[2]	2021		LR	the model's accuracy improved from	The max-out feature selection	

				78% to 81% with feature selection	technique was applied. Analysis of the most important features of employee attrition.	
[40]	2022		DT, RF, XGBoost (XGB), LR, and SVM	RF showed superior predictive capability, achieving anaccuracy=97%	Variables that had a significant effect on capture were recorded after the Chi-square test.	Limited dataset. Improved methods can be applied to make models more accurate.

4.2.2. Deep learning

The dataset used for analysis in [30]was the IBM HR Analytics Employee Attrition dataset. Machine-learning techniques and NNs were employed to analyze the data. Correlation analysis was used to identify the features that affect employee attrition. Different machine-learning models such as LR, RF, DT, NB, SVM, multilayer perceptron (MLP), and DNN were compared. RF performed the best with an accuracy of 97%. Although RFs proved to be the most effective, the article also demonstrated the promising accuracy of NNs, with an accuracy rate close to 95%. Initially, the NN model provided 90% accuracy, which improved to approximately 95% [30]. Factor analysis revealed that employees who lived far from their workplace had a higher attrition rate. In addition, lower-paid workers exhibited higher employee attrition compared to their higher-paid counterparts. Employee turnover was higher for those involved in more projects compared to those involved in fewer projects. Workers with more than five years of experience typically stayed in their jobs longer, and higher-level personnel had lower attrition rates. Workers with higher education tended to leave the organization more quickly. However, the natural environment was not found to be a significant factor in attrition analysis [30].

The researchers used the IBM HR Analytics Employee Attrition dataset in [22]to identify individual traits and structural factors that contribute to employee attrition. They compared various categorization models, including NB, LR, KNNs, and NNs (MLP). Despite the limited available data, the NN produced satisfactory results. However, it was noted that the NN model might be overfitted, which could affect generalization. They explored several hyperparameter configurations, with the best accuracy reaching 88.43%[22].

In another study[6],because the IBM HR Analytics Employee Attrition dataset was imbalanced and biased towards employed individuals, the authors used a balanced version obtained through the adaptive synthetic sampling method. They used a correlation matrix to evaluate the relationships between the features of the dataset. Deep neural networks (DNNs) were used as the model and achieved an accuracy of 94% [6].

In another study[8], researchers presented a model for predicting employee attrition using various predictive analytical techniques, including the artificial neural network (ANN) algorithm, gradient boosting, RF, and ensemble models. The dataset was sourced from the SAS library (www.sas.com). Key evaluation metrics, such as average squared error, misclassification rate,

and the Gini coefficient, were used to select the best model, which was determined to be the ANN algorithm[8].

In[14], an approach using people-analytics to predict employee attrition was introduced, giving priority to data quality rather than quantity. The study employed three datasets: the IBM HR Analytics Employee Attrition dataset, a simulated HR dataset from Kaggle, and data gathered through a questionnaire on the causes of employee attrition. Deep learning algorithms, such as DNNs, long short-term memory networks, and convolutional neural networks, were utilized, alongside various machine-learning algorithms and ensemble methods. The study's findings indicated that the voting classifier (VC), an ensemble method, outperformed other models across all datasets[14].Table 2 shows a summary of previous studies that applied the deep learning approach, highlighting the results, strengths, and limitations of each study.

Table 2. Summary of previous studies that applied the deep learning approach.

Ref	**Date**	**Dataset**	**Used Models**	**Result**	**Strengths**	**Limitations**
[30]	2023	IBM HR Analytics Employee Attrition dataset	LR, RF, DT, NB, SVM,MLP, and DNN	RF performed the best with an accuracy = 97%	Analysis of common causes of employee turnover rates.	The dataset used in these works is limited. Dataaugmentation techniques are not applied. feature selection is not used to achieve more accurate results.
[22]	2022		NB, LR, KNNs, and NNs MLP.	The best accuracy reached 88.43% for NN model	Highlighting the main employee attrition causes.	
[6]	2021		Deep neural networks (DNNs)	achieved an accuracy of 94%	ADASYN is used to transform the original dataset into a synthetic one. Analysis of employee attritionfeatures.	
[8]	2022	from the SAS library	ANN, gradient boosting, RF, and ensemble models, LSVMs,	The best model was ANN.	A real dataset from SAS was used.	Employee attrition factors were not analyzed. feature selection is not used to achieve more accurate results.

[14]	2021	Three datasets: an IBM dataset, a dataset from Kaggle, and a dataset gathered by a questionnaire.	DNNs, LSTM, and CNN, alongside various ML algorithms and ensemble methods.	The voting classifier (VC), an ensemble method, outperformed other models across all datasets	Feature selection methods were applied. Three datasets were used, and the prediction models' performance in various scenarios was evaluated.	Ignoring the class-unbalanced issue.

4.3. Ensemble Methods to Enhance Prediction

The process of systematically creating and merging multiple models, such as classifiers or experts, to address specific computational intelligence issues is known as ensemble learning. The main objectives of ensemble learning are to improve a model's performance (classification, prediction, function approximation, etc.) or reduce the risk of randomly selecting a poor model [38].Ensemble techniques strive to build a set of diverse learning algorithms and combine them to achieve better prediction performance compared to a single learning algorithm [47].

There are various types of ensemble models depending on how they were trained. Bagging involves training samples in parallel, using a bootstrap sampling approach with replacement, and then aggregating the results using a voting method. Boosting, similar to bagging, uses sampling with replacement, but instead of training in parallel, it learns sequentially. It assigns low weights to accurate predictions and high weights to incorrect ones, thereby giving more attention to correcting inaccurate predictions. Lastly, stacking involves generating additional predictions by using the outputs of each model as input variables. A meta-learner retrains on the data trained by individual algorithms to make predictions, giving stacking a dual-training characteristic [15].

Most previous research papers compare the accuracy of different models (such as DTs, RFs, SVMs, etc.) to determine which machine-learning model is best at handling staff attrition. However, a single model may not achieve the highest level of accuracy due to its development in a specific setting, which limits its generalizability [15].Therefore, researchers are interested in ensemble models that combine multiple models to enhance performance[47]. Many efforts have been made to improve pattern recognition and prediction on complex and atypical data using machine learning, which has led some employee attrition prediction studies to focus on applying ensemble models to improve prediction accuracy. This subsection reviews these studies.

In the study[15], researchers developed a predictive model based on 30 characteristics from the IBM HR Analytics Employee Attrition dataset that are known to influence employee attrition. They created eight predictive models, including RF, XGB, SVM, LR, ANN mode, and three different types of ensemble models that integrate individual models. The goal of the ensemble models was to combine multiple models into one to outperform any single model alone. By leveraging multiple models, the ensemble approach aimed to improve forecast accuracy and reduce the overall error rate. Specifically, the ensemble models included ESM1, which used LR as a meta-learner and RF and ANN as base algorithms; ESM2, which incorporated LR, ANN, XGB, and SVM as base algorithms; and ESM3, which combined LR, ANN, XGB, and SVM. Evaluation of the models' performance in predicting employee attrition revealed that ESM1 achieved the highest accuracy score of 97% [15].

In another study[23], HR analytics data from the Kaggle website was used to develop a model for predicting employee turnover rates. Predictive modeling was conducted using the AdaBoost classifier and the RF classifier. The findings showed that the RF classifier achieved an accuracy of 100%. This achievement can be attributed to the bagging strategy applied by the RF classifier, which involves dividing the dataset into separate random subsets and randomly selecting features during each stage of tree splitting. This approach effectively addresses the problem of overfitting[23].

In the study[42], the researchers investigated the correlation between employee performance and organizational growth. They collected real-time employee data from a reputable multinational corporation in Chennai, which consisted of around 5000 employees from different departments, age groups, and locations. The researchers used the extreme gradientboosting ensemble classifier, a boosting algorithm, to predict employee performance and then compared it with the gradient boosting ensemble classifier. The results indicated that both algorithms achieved an accuracy rate of over 90%, with the extreme gradientboosting method outperforming the traditional gradientboosting ensemble learning approach[42].

To accurately identify employee attrition using machine learning, the researchers utilized the IBM HR Analytics Employee Attrition dataset in their study [47]. They created six different training sets, each with a different number of features based on the associations between variables. They applied several machine-learning algorithms, such as DTs, AdaBoost, LR, RF, gradient boosting, and ensemble methods. The best-performing ensemble consisted of DT and LR, achieving an accuracy rate of 86%[47].

In[48], machine-learning algorithms were used to analyze the factors that contribute to employee attrition and predict attrition itself. The IBM HR Analytics Employee Attrition dataset was used in this study. An optimized approach, using the extra trees classifier (ETC) algorithm, was applied. Employee exploratory data analysis was conducted, and in order to balance the dataset, SMOTE dataset resampling was used. The ETC algorithm achieved the highest accuracy score of 93%, surpassing SVM, LR, and DT in this experiment [48].

The researchers in [49]used the IBM HR Analytics Employee Attrition dataset to predict the employee attrition rate using machine-learning algorithms. The study's findings showed that XGB performed the best, achieving an accuracy of 88%, when compared to the other algorithms trained in this study: DT, RF, KNN, NNs, and AdaBoosting [49].

In the study [50], a tree-based ensemble machine-learning model is used with the IBM HR Analytics Employee Attrition Performance dataset to conduct a comprehensive analysis of employee attrition. The study aims to optimize the effectiveness of the currently available tree approaches by evaluating the tree-based ensemble. The machine-learning models employed are RF and gradient boosting. The kernel density estimate plot is utilized to illustrate the distribution of data across the features of the dataset and to determine the density of data across different features. The gradient boost method, an ensemble method, outperforms the RF algorithm as it achieves a balance between accuracy and recall scores while also improving accuracy. This ensemble methodology achieves an accuracy of approximately 95%, which is significantly higher than the RF approach[50].

The model for predicting employee attrition was developed, and the study by [4]identified the most significant factors that contribute to employee attrition. The dataset used included data from the human resources department of an Iranian pharmaceutical company. A gradientboosting algorithm was applied, resulting in an accuracy rate of 89%[4].

In[51],the authors developed a method to assess the importance of each factor that impacts an employee's likelihood of leaving the organization, using the IBM HR Analytics Employee Attrition dataset. The following machine-learning algorithms were employed: AdaBoost, XGB, gradient boosting, KNN, DT, RF, and LR. After testing, evaluation, and validation, it was determined that the XGB classifier yielded the most favorable outcomes, achieving an accuracy of 87% [51].

A people-analytics approach for predicting employee attrition was proposed in [14]. The study focused on data quality rather than quantity. Three datasets were used: the IBM HR Analytics Employee Attrition dataset, a simulated HR dataset supported by Kaggle data, and a dataset collected through a questionnaire about the causes of employee attrition. Machine-learning algorithms like DT, SVM, LR, and RF, along with ensemble methods like XGB, VC, and stacked ANN-based, were applied, as well as a set of deep learning algorithms. The study's results reveal that VC outperforms the other models on all datasets [14].

The ensemble model showed superiority over the rest of the models, in both studies [52] and [53]. The attrition rate of employees is discussed by the authors in [52]. To benchmark employee attrition accuracy, three distinct supervised learning algorithms AdaBoost, SVM, and Random Forest are used. The models were trained using the IBM dataset. Most of the time, AdaBoost has outperformed the other classifiers in terms of accuracy. The accuracy gained from several algorithm executions on the random splitting of training and test data was in the range of 0.84 to 0.88 when there were 1000 estimators and a learning rate of 0.1 [52].

The authors in [53] suggest a technique for creating a turnover intention prediction model based on machine learning. An analysis was done in this study to confirm the influence of predictors using data from the Job Movement Path Survey for college graduates conducted by the Korea Employment Information Service. Extreme gradient boosting (XGB), logistic regression (LR), and k-nearest neighbor (KNN) classifiers were used for model learning and classification. Looking at the average values of the four sets of cross-validation, the turnover intention prediction model analysis revealed that XGB scored the highest accuracy (78.5%), followed by LR (78.3) and KNN (76%). One of the model benefits in this studyis calculating the significance of independent variable input to the prediction model. The significance ranking of the independent variables was examined in this investigation. The research revealed that contentment with work security was the most significant predictor in predicting the desire to leave one's employment. Following that, contentment with the organization and the work's alignment with the employee study major emerged sequentially[53].

The authors in[41]describe and analyze a special study on employee turnover intention conducted across Europe. The study is based on the GEEI survey carried out by Effectory in 2018. This labor market survey included a sample of 18,322 employees from 56 European countries. The survey consisted of 123 questions covering contextual information such as socio-demographic, labor, and industry aspects, as well as an objective question. A subset of the raw data from the GEEI survey, containing 9296 rows, was selected and limited to respondents from 30 European countries. The predictive abilities of various machine-learning classifiers, including KNN, LR, DT, RFs, LightGBM (LGBM), XGB, and TabNet, were compared. The top-performing models were found to be LGBM and LR. To determine which features were responsible for the performance of both LR and LGBM, the authors conducted an analysis of the relevance of each feature by performing 100 separate trials and compiling importance ratings from all trials. The absolute value of the coefficient for a feature in the LR model was used as a measure of its significance, while the relevance of a feature in the LGBM model was determined by how frequently it was used in the model tree split[41].Table 3 shows a summary of previous studies that applied ensemble methods, highlighting the results, strengths, and limitations of each study.

Table 3. Summary of previous studies that applied ensemble methods.

Ref	Date	Dataset	Used Models	Result	Strengths	Limitations
[15]	2022	IBM HR Analytics Employee Attrition dataset	RF, XGB, SVM, LR, ANN mode, and three different types of ensemble models that integrate individual models	ESM1 achieved the highest accuracy score of 97%	Multiple approaches were applied, including ML, deep learning, and a combination of ensemble models. SMOTE technique was performed.	The dataset used is limited and does not account for every psychological and subjective factor.
[49]	2022		XGB and DT, RF, KNN, NNs, and AdaBoosting	XGB performed the best, achieving an accuracy of 88%	Several feature selection methods were used.	
[48]	2022		extra trees classifier (ETC) algorithm , SVM, LR, and DT	The ETC algorithm achieved the highest accuracy score of 93%	The dataset of employee attrition was analyzed by the EEDA. SMOTE technique was applied.	

[47]	2021		DTs, AdaBoost, LR, RF, gradient boosting, and ensemble methods	The best-performing ensemble consisted of DT and LR, achieving an accuracy rate of 86%.	Multiple models in several scenarios were applied to determine the best model.	Limited dataset. Improved methods can be applied to increase the accuracy rate and reduce generalization errors.
[50]	2021	IBM HR Analytics Employee Attrition dataset	RF and gradient boosting	gradient boosting achieves an accuracy of approximately 95%, which is significantly higher than the RF approach.	Exploratory Data Analysis was Applied to Feature Engineering.	The dataset used is limited and does not account for every psychological and subjective factor.
[51]	2022		AdaBoost, XGB, gradient boosting, KNN, DT, RF, and LR	XGB yielded the most favorable outcomes, achieving an accuracy of 87%	Analysis and recommendations about employee attrition were provide.	
[52]	2024		AdaBoost, SVM, and Random Forest	AdaBoost has outperformed the others in terms of accuracy.	Accuracy was tested in three different random iterations. Analysis of features correlation was applied.	
[23]	2022	HR analytics data from the Kaggle.	AdaBoost classifier and the RF classifier.	RF classifier achieved an accuracy of 100%.	Feature engineering was applied to select the relevant data.	Ignore the dataset imbalance problem.
[42]	2022	They collected real-time employee data from a reputable multinational corporation in Chennai.	extreme gradientboosting ensemble classifier, a boosting algorithm.	both algorithms achieved an accuracy rate of over 90%, with the XGB method outperforming the traditional gradientboosting ensemble.	The importance of features was analyzed, and then the most prominent features were selected using chi-square and mutual information.	It is possible to compare the performance of the model with other ensemble learning algorithms.
[4]	2021	data from the human resources	A gradientboosting algorithm	A gradientboosting algorithm was	Analysis to identify employee	The valuation of experiments is limited to

		department of an Iranian pharmaceutical company		applied, resulting in an accuracy rate of 89%	attrition factors was performed.	the pharmaceutical sector.
[53]	2024	Dataset from the Job Movement Path Survey for college graduates conducted by the Korea	XGB, LR, and KNN	XGB had the highest accuracy (78.5%),	The reasons for choosing a job were examined, and recommendations to organizations for managing new talent were provided.	The limited application of classification techniques, and the data imbalance in turnover intention.
[41]	2022	The study is based on the GEEI survey carried out by Effectory in 2018	KNN, LR, DT, RFs, LGBM, XGB, and TabNet,	The top-performing models were found to be LGBM and LR	The study covered a dataset covering 30 European countries. The study combined analytical methodologies from the perspective of predictive analysis of human resources with a causal approach.	The limitation is the weighing of datasets to combat selection bias. which is restricted to the workforce of a single nation.

4.4. Feature Selection Techniques

Feature selection techniques in machine learning aim to find the optimal set of features for constructing optimized prediction models [3]. These techniques eliminate unnecessary and redundant attributes and select relevant features to improve accuracy[14].

Several feature selection techniques have been developed for this purpose, including filter, wrapper, embedded, and hybrid approaches. Filtering techniques select relevant features based on the correlation between feature values. Wrapper approaches choose features based on their performance with a selected subset of features. Embedded approaches incorporate feature selection directly into the learning process. Hybrid methods combine elements from these approaches to achieve optimal feature selection [2].

The use of feature selection algorithms in datasets is crucial for improving accuracy. Recent research has extensively investigated the selection of significant features from data to enhance the speed and accuracy of models [2].

The feature selection process is important because it improves the accuracy of classification The feature selection process is significant as it improves the accuracy of classification results for the underlying problem. Real-world datasets typically contain a large number of features, many of which may not be relevant to classification or prediction models. This abundance of unnecessary and redundant features can result in inconsistent and unreliable results when all features are used in classification or prediction tasks [37]. Therefore, feature selection plays a critical role in identifying attributes that genuinely contribute to classification accuracy, especially those closely related to the class variables[38].

In this subsection, employee attrition prediction studies are reviewed regarding their reliance on feature selection algorithms.

Since employee datasets contain numerous features, some of which may be irrelevant, researchers often choose to select features that are relevant to the issue of employee attrition. For example, in studies like [47] and[50], principal component analysis (PCA) was used to reduce the dimensionality of the feature space. However, PCA may struggle to handle binary variables that are common in attrition databases. Additionally, PCA relies on linear relationships between features for dimension reduction and may overlook potential connections between the output and filter features[2].

Researchers aim to identify the features that have the greatest impact on employee attrition by employing feature selection techniques that reduce or eliminate unnecessary features. In many experiments, researchers applied feature selection techniques and compared the results. Notably, the RF classifier technique for feature selection has proven effective in improving RF metrics. For example, the accuracy of the RF classifier increased from 85.22% to 86.10%[23].

In[3], the IBM HR Analytics Employee Attrition dataset was used to conduct RF classification modeling to predict employee attrition. The study focuses on comparing feature selection techniques to eliminate or reduce redundant characteristics, to identify features that have the greatest impact on employee turnover. Three feature selection algorithms were used: SelectKBest, information gain, and recursive feature elimination. According to the test scenario in this study, information gain achieved the highest accuracy value of 89.2% when utilizing 25 features [3].

In[38], a model was presented for predicting employee attrition using an enhanced weight-based forest optimization algorithm. Among all the sets of features, this algorithm chooses the best one. A new classifier, called improved random forest (IRF), was proposed, which selects n samples from a total of k samples to build a random number of DTs. The final prediction is assessed based on the performance of the classifier rather than a majority vote. The analysis was performed using the IBM HR Analytics Employee Attrition dataset. The study's findings show a model accuracy of 91% [38].

In[14],the feature selection techniques RFE and SelectKBest were used. Features selected by SelectKBest are kept even if they are eliminated by RFE. Similarly, features not eliminated by RFE but selected by SelectKBest are also kept. On the other hand, features rejected by RFE and not selected by SelectKBest are discarded. As a result, age, marital status, rewards, job involvement, training, business travel, tenure, grade, job performance, job satisfaction, and environment satisfaction are identified as the 11 main features related to employee attrition prediction, based on the combination of the RFE and SelectKBest feature selection techniques [14].

In[49], after applying the chi-square, recursive method, mutual information method, and tree selector methods to each attrition-related feature, only the top 15 elements from each approach were combined into a set, resulting in a consideration of 23 features. Additionally, eight PCA characteristics were included, resulting in a total of 31 features being considered [49].

In another experiment[42], mutual information and chi-square were two alternative methods used for feature selection. Features showing the highest mutual information and chi-square values with the class label were selected [42].

In several experiments, various techniques for selecting features were implemented, contributing to the improvement of prediction results. The Boruta method was used to identify features that influence employee attrition and rank them based on their average importance [4]. Additionally, the "max-out" feature method was applied to minimize the difference between features[2]. Exploratory data analysis was conducted in[51]to gain a better understanding of the dataset, using a variety of graphs to visualize employee attrition more comprehensively. Furthermore, the Shapley additive explanations index was used to determine the most significant factors contributing to the attrition[51].

5. DISCUSSION

General discussion: The study included a literature survey of previous studies that presented an attempt to predict employee attrition from 2019 to February 2024. Most previous studies took a machine learning approach to predict employee attrition, although a group of them relied on the machine learning approach. Most studies agree to conduct comparative experiments for several predictive models and then determine the model that performs best in predicting employee attrition.
Discussion related to algorithms: The most widely used machine learning algorithm in employee attrition prediction is linear regression.In most cases, the linear regression algorithm is used to compare with the study's suggested algorithms and verify the algorithm with the best performance, as in studies [10], [20], and others. However, its appearance in many studies does not mean that it is the best-performing model for predicting employee attrition. Through this literature survey, the Random Forest algorithm can be considered the best-performing algorithm, and its superiority over many algorithms has been repeated, as in [10], [40], and others. The model (RF) considers multiple factors and has an accuracy of 86.0% in predicting employee attrition, which is greater than the performance reported in [10]. According to the chosen performance metrics (Accuracy, Precision, Recall, and F1-score), the RF classifier-based model is considered to be the best. The model achieved an accuracy of 97.8% [40]. The RF model is the most promising in terms of performance metrics considered, compared to other models in the researchers' experiment [1], including the Neural Network model. The RF model achieved an accuracy of 87.96%, while the NN model achieved an accuracy of 84.76%. As a result, applying RF not only enhances data-driven decision-making but also yields plausible justifications for choices made. The last point is particularly significant in real-world circumstances where the outcomes' justice and dependability are paramount [1]. In addition, the turnover intention prediction analyses revealed that the XGB ensemble model outperforms many other models.

A recent use of deep learning (DL), a subfield of machine learning, for employee attrition prediction issues shows great promise. We also found some deep learning-based studies in this investigation. DL approaches have various potential qualities, like better performance and automated feature extraction. We anticipate that further studies on the application of DL techniques in employee attrition prediction will be carried out in the future.

Discussion related to datasets:A variety of datasets were utilized to study the issue of employee attrition and develop predictive models using them. Some studies relied on ready-made datasets available or archival employee datasets for specific institutions, and some researchers collected data using questionnaires that included questions about factors related to employee attrition. However, we can say that the available dataset (IBM HR Analytics Employee Attrition dataset) produced by IBM, which includes 1470 observations, is the most widely used in previous studies, as is the case in the studies of[15],[22], [23], [40], and many others.

This dataset is highly unbalanced, with 237 positive samples (from former employees) and 1,233 negative samples (from current employees), which leads to a significant data bias [6]. So we can conclude that without addressing the class imbalance in the data, most models will not perform well [44]. Addressing dataset imbalance using the Synthetic Minority Oversampling (SMOTE) technique and weighted classes has improved prediction results as in [44]. When paired with weighted class and SMOTE approaches, Logistic Regression yields the maximum recall score of 0.9014 and can accurately predict 64 out of 71 minority class situations. As a result, this study uses SMOTE and category weights to get the best results from the logistic regression model. Also, SMOTE is a strategy for oversampling the minority class; it was chosen over undersampling because the latter could lead to the omission of important data [46]. Also, this dataset is limited. Therefore, it is recommended to apply data augmentation techniques in the future.

On the other hand, we can conclude that the reliance of many research studies on the same dataset is an indication of the lack of ready and available datasets that include many advantages that contribute to predicting employee attrition with greater accuracy. It is recommended to create real datasets in the future, which in turn will help organizations examine the factors that lead to employee attrition, especially if they focus on a specific field or sector.

Discussion related to features: Some studies have taken all features of the database into account without analyzing and filtering them based on their importance. While other studies provided analyses to determine the most important factors that cause employee attrition.

To name a few, the outcome obtained in[45] shows that the main factors of attrition are monthly income, age, overtime, and distance from home. The results of the analysis in [20] revealed that there is a positive relationship between employee attrition and the following characteristics: years working in the company, years with the current manager, total years of work, and monthly income. There is also a negative relationship between attrition and job level, years in a current role, total years of work, and monthly income[20]. In[41] can be noted that time spent with the company, satisfactory development, and willingness to recommend the employer to friends and family are among the top five traits in the study. It's interesting to note that in the two data sets included in the study, gender—a widely acknowledged driver of turnover intention—is not ranked among the most significant characteristics. It also does not show any country-specific effect[41].

Therefore, the most prominent features whose importance in prediction was praised by previous studies were summarized.

The most important features affecting employee attrition that were extracted are shown in Table4.The number of times these features are mentioned is displayed, along with references. As shown in Table4, the most influential features are monthly income, age, years of experience, overtime, job role, environmental satisfaction, and job satisfaction. Depression level, bonuses, training, business travel, hourly wage, gender, department, last contract length, childcare

allowance, and working on multiple projects were among the features that were not significant in most studies.

Table 4. Summary of the features that studies have proven to affect employee attrition.

Factors	**Times**	**References**
Monthly Income	11	[1], [4], [6], [10], [15], [23], [30],[40],[45], [48],[51].
Age	10	[1], [4], [14], [15],[23],[38], [40],[45],[48],[51].
Years of Experience	8	[2], [4], [23], [30], [38],[40], [45], [48].
Overtime	7	[2], [6],[15],[38], [40],[45],[51].
Job Role	6	[2], [4], [6], [14],[40], [51].
Environment Satisfaction	6	[1], [14], [15], [38], [45], [51].
Job Satisfaction	6	[10], [14], [23], [38], [40], [51].
Job Involvement	5	[1], [2], [14],[30],[38].
Distance from Home	5	[1], [30], [40], [45], [51].
Marital Status	4	[14],[23],[38], [40].
Number companies worked	4	[2], [15], [23], [51].
Work-Life Balance	3	[1], [23], [38].
Stock Option Level	3	[40], [45], [51].
Years with current manager	3	[2], [23], [38].
Years in the company	3	[1], [40], [45].
Job Performance	2	[14], [15].
Relationship Satisfaction	2	[15], [51].
Direct manager	2	[4], [48].
Grade	2	[4], [14].
Years after the last promotion	1	[2].
Depression Level	1	[10].
Rewards	1	[14].
Training	1	[14].
Business Travel	1	[14].
Hourly rate	1	[48].
Gender	1	[15].
Department	1	[4].
Last Contract Duration	1	[4].
Allowance for Childcare	1	[4].
Work on multiple projects	1	[30].

As for the practical implications based on the results of this study, human resources management can use machine learning models to forecast and analyze employee attrition behavior and, in turn, develop plans to increase job satisfaction. Additionally, managers can lessen the causes of attrition behaviors within their company and offer the best ways to eliminate them.

The survey found several significant reasons why workers quit their jobs. It is possible to design effective policies and initiatives that will support employee retention by taking these aspects into account. The characteristics that have a major impact on employee attrition are monthly salary, age, years of experience, extra time, job role, environmental satisfaction, and job satisfaction, as has been repeatedly stated in prior research.Therefore, it can be suggested to include the following policies in the strategic plans:

- Competitive base pay.

- Reduction in the "number of projects/tasks assigned" to workers who are working beyond their capacity.
- Regular promotions, rewards, and incentives
- Motivate staff members and encourage them to work only during business hours.
- Recognizing talents and offering flexibility.
- providing a suitable work environment, providing the required opportunities and resources.
- Ensuring the satisfaction of employees to continue in the organization and thus reducing the “Employee Attrition” [16][38].

6. CONCLUSIONS

This study showed that the selected publications used various machine learning models along with multiple feature selection methods, depending on data availability. Each study looks at predicting employee attrition using machine learning, the various features used, and the type and size of the database. Some studies relied on ready datasets available on the Internet, while in other studies, researchers collected data for their studies.

Many models have been used in different studies. However, it cannot be considered that any specific model is the best model ever to predict employee attrition. However, the Random Forest model has shown superior performance in several studies. The most used models in previous studies were Linear Regression, Decision Trees, Random Forests, and Support Vector Machines. Most studies relied on comparing a group of different machine learning models to determine the best-performing model among them. It has been observed that the ANN and DNN algorithms are the most preferred deep learning algorithms. This article can be considered that pave the way for further research on developing the problem of crop yield forecasting.

In future work, we aim to build on the results of this study and focus on developing a DL-based employee attrition prediction model. It is also recommended to focus on improving the performance of models by implementing modern feature selection methods, which contribute to identifying the factors that most affect prediction accuracy and thus excluding unnecessary features.

Numerous research papers' reliance on the same dataset suggests that there aren't enough readily available datasets with a variety of factors that help forecast employee attrition more accurately. It is recommended to create real data sets in the future, that are collected based on the analysis of factors that lead to employee attrition, especially if they focus on a specific area or sector of society.
Since the process of predicting the departure of an employee is a critical decision that requires trust in the machine learning model, and at the same time, decision-makers in the human resources department are keen to know the reasons behind the departure of employees, it is recommended to apply explainable artificial intelligence.

REFERENCES

[1] F. Guerranti and G. M. Dimitri, “A Comparison of Machine Learning Approaches for Predicting Employee Attrition,”Applied Sciences, vol. 13, no. 1, p. 267, Dec. 2022, doi: 10.3390/app13010267.
[2] S. Najafi-Zangeneh, N. Shams-Gharneh, A. Arjomandi-Nezhad, and S. Hashemkhani Zolfani, “An Improved Machine Learning-Based Employees Attrition Prediction Framework with Emphasis on Feature Selection,”Mathematics, vol. 9, no. 11, p. 1226, May 2021, doi: 10.3390/math9111226.

[3] S. F. Sari and K. M. Lhaksmana, "Employee Attrition Prediction Using Feature Selection with Information Gain and Random Forest Classification,"JoSYC, vol. 3, no. 4, pp. 410–419, Sep. 2022, doi: 10.47065/josyc.v3i4.2099.

[4] F. Mozaffari, M. Rahimi, H. Yazdani, and B. Sohrabi, "Employee attrition prediction in a pharmaceutical company using both machine learning approach and qualitative data,"BIJ, vol. 30, no. 10, pp. 4140–4173, Dec. 2023, doi: 10.1108/BIJ-11-2021-0664.

[5] N. Alshanbri, M. Khalfan, M. A. Noor, D. Dutta, K. Zhang, and T. Maqsood, "Employees' Turnover, Knowledge Management and Human Recourse Management: A Case of Nitaqat Program,"IJSSH, vol. 5, no. 8, pp. 701–706, 2015, doi: 10.7763/IJSSH.2015.V5.543.

[6] S. Al-Darraji, D. G. Honi, F. Fallucchi, A. I. Abdulsada, R. Giuliano, and H. A. Abdulmalik, "Employee Attrition Prediction Using Deep Neural Networks,"Computers, vol. 10, no. 11, p. 141, Nov. 2021, doi: 10.3390/computers10110141.

[7] J. Farrugia, "7 statistics on employee turnover every HR manager should be aware of | Workforce.com." Accessed: Feb. 19, 2024. [Online]. Available: https://workforce.com/news/7-statistics-on-employee-turnover-in-2022-every-hr-manager-should-be-aware-of

[8] F. K. Alsheref, I. E. Fattoh, and W. M.Ead, "Automated Prediction of Employee Attrition Using Ensemble Model Based on Machine Learning Algorithms,"Computational Intelligence and Neuroscience, vol. 2022, p. 7728668, Jun. 2022, doi: 10.1155/2022/7728668.

[9] Abdelmohsen A. Nassani, Shaykah A. Al-Hassan, and Anhar A. Al-Dakhil, "The Nexus between Work Engagement, Job Stress, and Employee Turnover Intention in Riyadh Private Sector," 2021, [Online]. Available: https://doi.org/10.15520/jassh.v7i4.611

[10] R. Joseph, S. Udupa, S. Jangale, K. Kotkar, and P. Pawar, "Employee Attrition Using Machine Learning And Depression Analysis," in 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India: IEEE, May 2021, pp. 1000–1005. doi: 10.1109/ICICCS51141.2021.9432259.

[11] P. K. Jain, M. Jain, and R. Pamula, "Explaining and predicting employees' attrition: a machine learning approach,"SN Applied Sciences, vol. 2, no. 4, p. 757, Mar. 2020, doi: 10.1007/s42452-020-2519-4.

[12] G. Gabrani and A. Kwatra, "Machine Learning Based Predictive Model for Risk Assessment of Employee Attrition," in Computational Science and Its Applications – ICCSA 2018, vol. 10963, O. Gervasi, B. Murgante, S. Misra, E. Stankova, C. M. Torre, A. M. A. C. Rocha, D. Taniar, B. O. Apduhan, E. Tarantino, and Y. Ryu, Eds., in Lecture Notes in Computer Science, vol. 10963. , Cham: Springer International Publishing, 2018, pp. 189–201. doi: 10.1007/978-3-319-95171-3_16.

[13] E. Rombaut and M.-A. Guerry, "Predicting voluntary turnover through human resources database analysis,"MRR, vol. 41, no. 1, pp. 96–112, Jan. 2018, doi: 10.1108/MRR-04-2017-0098.

[14] N. B. Yahia, J. Hlel, and R. Colomo-Palacios, "From Big Data to Deep Data to Support People Analytics for Employee Attrition Prediction,"IEEE Access, vol. 9, pp. 60447–60458, 2021, doi: 10.1109/ACCESS.2021.3074559.

[15] D. Chung, J. Yun, J. Lee, and Y. Jeon, "Predictive model of employee attrition based on stacking ensemble learning,"Expert Systems with Applications, vol. 215, p. 119364, Apr. 2023, doi: 10.1016/j.eswa.2022.119364.

[16] S. N. Khera and Divya, "Predictive Modelling of Employee Turnover in Indian IT Industry Using Machine Learning Techniques,"Vision, vol. 23, no. 1, pp. 12–21, Mar. 2019, doi: 10.1177/0972262918821221.

[17] R. Punnoose and P. Ajit, "Prediction of Employee Turnover in Organizations using Machine Learning Algorithms,"ijarai, vol. 5, no. 9, 2016, doi: 10.14569/IJARAI.2016.050904.

[18] D. S. Sisodia, S. Vishwakarma, and A. Pujahari, "Evaluation of machine learning models for employee churn prediction," in 2017 International Conference on Inventive Computing and Informatics (ICICI), Coimbatore: IEEE, Nov. 2017, pp. 1016–1020. doi: 10.1109/ICICI.2017.8365293.

[19] S. S. Alduayj and K. Rajpoot, "Predicting Employee Attrition using Machine Learning," in 2018 International Conference on Innovations in Information Technology (IIT), Nov. 2018, pp. 93–98. doi: 10.1109/INNOVATIONS.2018.8605976.

[20] A. Habous, E. H. Nfaoui, and Y. Oubenaalla, "Predicting Employee Attrition using Supervised Learning Classification Models," in 2021 Fifth International Conference On Intelligent Computing in Data Sciences (ICDS), Oct. 2021, pp. 1–5. doi: 10.1109/ICDS53782.2021.9626761.

[21] N. Khalifa, M. Alnasheet, and H. Kadhem, "Evaluating Machine Learning Algorithms to Detect Employees' Attrition," in 2022 3rd International Conference on Artificial Intelligence, Robotics and Control (AIRC), Cairo, Egypt: IEEE, May 2022, pp. 93–97. doi: 10.1109/AIRC56195.2022.9836981.

[22] G. Raja Rajeswari., R. Murugesan, R. Aruna., B. Jayakrishnan, and K. Nilavathy., "Predicting Employee Attrition through Machine Learning," in 2022 3rd International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India: IEEE, Oct. 2022, pp. 1370–1379. doi: 10.1109/ICOSEC54921.2022.9952020.

[23] S. Krishna and S. Sidharth, "Analyzing Employee Attrition Using Machine Learning: the New AI Approach," in 2022 IEEE 7th International conference for Convergence in Technology (I2CT), Mumbai, India: IEEE, Apr. 2022, pp. 1–14. doi: 10.1109/I2CT54291.2022.9825342.

[24] F. H. Wardhani and K. M. Lhaksmana, "Predicting Employee Attrition Using Logistic Regression With Feature Selection,"SinkrOn, vol. 7, no. 4, pp. 2214–2222, Oct. 2022, doi: 10.33395/sinkron.v7i4.11783.

[25] Sami M Abbasi and Kenneth W Hollman, "Turnover: The Real Bottom Line,"Public Personnel Management, vol. 29, no. 3, pp. 333–342, Sep. 2000, doi: 10.1177/009102600002900303.

[26] A. Mhatre, A. Mahalingam, M. Narayanan, A. Nair, and S. Jaju, "Predicting Employee Attrition along with Identifying High Risk Employees using Big Data and Machine Learning," in 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), Greater Noida, India: IEEE, Dec. 2020, pp. 269–276. doi: 10.1109/ICACCCN51052.2020.9362933.

[27] H. Ongori, "A review of the literature on employee turnover," 2007, [Online]. Available: Available online http://www.academicjournals.org/ajbm

[28] s. R. Basariya and R. Ahmed, "A STUDY ON ATTRITION – TURNOVER INTENTIONS OF EMPLOYEES,"International Journal of Civil Engineering and Technology, vol. 10, pp. 1954–19601, May 2019.

[29] C. Abdullah Al Mamun and Md. Nazmul Hasan, "Factors affecting employee turnover and sound retention strategies in business organization: a conceptual view,"Problems and Perspectives in Management, vol. 15, no. 1, pp. 63–71, Mar. 2017, doi: 10.21511/ppm.15(1).2017.06.

[30] S. R. Pulari, A. Punitha, S. Raja Varshni Meenachi, and S. Vasudevan, "A Comparative Study of Employee Attrition Analysis Using Machine Learning and Deep Learning Techniques," in Inventive Communication and Computational Technologies, G. Ranganathan, X. Fernando, and Á. Rocha, Eds., Singapore: Springer Nature Singapore, 2023, pp. 1–12.

[31] S. Chatterjee, "Understanding the Ergonomics of Attrition,"European Journal of Studies in Management and Business, vol. 22, pp. 11–18, Sep. 2022, doi: 10.32038/mbrq.2022.22.02.

[32] K. Ashokkumar, S. Jacob, and A. Joseph, "A Study on Attrition Management in Private Sector Financial Institutions - A Special Reference to Kottayam District in Kerala,"commerce, vol. 8, no. 2, pp. 35–44, Apr. 2020, doi: 10.34293/commerce.v8i2.2319.

[33] P. Ghosh, R. Satyawadi, J. Prasad Joshi, and Mohd. Shadman, "Who stays with you? Factors predicting employees' intention to stay,"Int J of Org Analysis, vol. 21, no. 3, pp. 288–312, Jul. 2013, doi: 10.1108/IJOA-Sep-2011-0511.

[34] M. M. Taye, "Understanding of Machine Learning with Deep Learning: Architectures, Workflow, Applications and Future Directions,"Computers, vol. 12, no. 5, 2023, doi: 10.3390/computers12050091.

[35] A. Singh, N. Thakur, and A. Sharma, "A review of supervised machine learning algorithms," in 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), Mar. 2016, pp. 1310–1315.

[36] V. V. Saradhi and G. K. Palshikar, "Employee churn prediction,"Expert Systems with Applications, vol. 38, no. 3, pp. 1999–2006, Mar. 2011, doi: 10.1016/j.eswa.2010.07.134.

[37] U. Moorthy and U. D. Gandhi, "Forest optimization algorithm-based feature selection using classifier ensemble,"Computational Intelligence, vol. 36, no. 4, pp. 1445–1462, Nov. 2020, doi: 10.1111/coin.12265.

[38] S. Porkodi, S. Srihari, and N. Vijayakumar, "Talent management by predicting employee attrition using enhanced weighted forest optimization algorithm with improved random forest classifier,"IJATEE, vol. 9, no. 90, May 2022, doi: 10.19101/IJATEE.2021.875340.

[39] Dr. T. S. Poornappriya and Dr. R. Gopinath, "Employee Attrition In Human Resource Using Machine Learning Techniques," 2021.

[40] S. Aggarwal, M. Singh, S. Chauhan, M. Sharma, and D. Jain, "Employee Attrition Prediction Using Machine Learning Comparative Study," in Intelligent Manufacturing and Energy Sustainability, vol. 265, A. N. R. Reddy, D. Marla, M. N. Favorskaya, and S. C. Satapathy, Eds., in Smart Innovation, Systems and Technologies, vol. 265. , Singapore: Springer Singapore, 2022, pp. 453–466. doi: 10.1007/978-981-16-6482-3_45.

[41] M. Lazzari, J. M. Alvarez, and S. Ruggieri, "Predicting and explaining employee turnover intention,"Int J Data Sci Anal, vol. 14, no. 3, pp. 279–292, Sep. 2022, doi: 10.1007/s41060-022-00329-w.

[42] P. Sujatha and R. S. Dhivya, "Ensemble Learning Framework to Predict the Employee Performance," in 2022 Second International Conference on Power, Control and Computing Technologies (ICPC2T), Raipur, India: IEEE, Mar. 2022, pp. 1–7. doi: 10.1109/ICPC2T53885.2022.9777078.

[43] B. Meraliyev, A. Karabayeva, T. Altynbekova, and Y. Nematov, "Attrition Rate Measuring In Human Resource Analytics Using Machine Learning," in 2023 17th International Conference on Electronics Computer and Computation (ICECCO), Kaskelen, Kazakhstan: IEEE, Jun. 2023, pp. 1–6. doi: 10.1109/ICECCO58239.2023.10146602.

[44] K. Sheth, J. Patel, and J. Verma, "Machine Learning-Based Investigation of Employee Attrition Prediction and Analysis," in Emerging Technology Trends in Electronics, Communication and Networking, vol. 952, R. Dhavse, V. Kumar, and S. Monteleone, Eds., in Lecture Notes in Electrical Engineering, vol. 952. , Singapore: Springer Nature Singapore, 2023, pp. 221–238. doi: 10.1007/978-981-19-6737-5_19.

[45] F. Fallucchi, M. Coladangelo, R. Giuliano, and E. William De Luca, "Predicting Employee Attrition Using Machine Learning Techniques,"Computers, vol. 9, no. 4, p. 86, Nov. 2020, doi: 10.3390/computers9040086.

[46] N. Bhartiya, S. Jannu, P. Shukla, and R. Chapaneri, "Employee Attrition Prediction Using Classification Models," in 2019 IEEE 5th International Conference for Convergence in Technology (I2CT), Mar. 2019, pp. 1–6. doi: 10.1109/I2CT45611.2019.9033784.

[47] A. Qutub, A. Al-Mehmadi, M. Al-Hssan, R. Aljohani, and H. S. Alghamdi, "Prediction of Employee Attrition Using Machine Learning and Ensemble Methods,"IJMLC, vol. 11, no. 2, pp. 110–114, Mar. 2021, doi: 10.18178/ijmlc.2021.11.2.1022.

[48] Raza, K. Munir, M. Almutairi, F. Younas, and M. M. S. Fareed, "Predicting Employee Attrition Using Machine Learning Approaches,"Applied Sciences, vol. 12, no. 13, p. 6424, Jun. 2022, doi: 10.3390/app12136424.

[49] L. S. Ganthi, Y. Nallapaneni, D. Perumalsamy, and K. Mahalingam, "Employee Attrition Prediction Using Machine Learning Algorithms," in Proceedings of International Conference on Data Science and Applications, vol. 288, M. Saraswat, S. Roy, C. Chowdhury, and A. H. Gandomi, Eds., in Lecture Notes in Networks and Systems, vol. 288. , Singapore: Springer Singapore, 2022, pp. 577–596. doi: 10.1007/978-981-16-5120-5_44.

[50] V. Mehta and S. Modi, "Employee Attrition System Using Tree Based Ensemble Method," in 2021 2nd International Conference on Communication, Computing and Industry 4.0 (C2I4), Dec. 2021, pp. 1–4. doi: 10.1109/C2I454156.2021.9689398.

[51] K. M. Mitravinda and S. Shetty, "Employee Attrition: Prediction, Analysis Of Contributory Factors And Recommendations For Employee Retention," in 2022 IEEE International Conference for Women in Innovation, Technology & Entrepreneurship (ICWITE), Bangalore, India: IEEE, Dec. 2022, pp. 1–6. doi: 10.1109/ICWITE57052.2022.10176235.

[52] M. Karimi and K. S. Viliyani, "Employee Turnover Analysis Using Machine Learning Algorithms".

[53] J. Park, Y. Feng, and S.-P. Jeong, "Developing an advanced prediction model for new employee turnover intention utilizing machine learning techniques,"Scientific Reports, vol. 14, no. 1, p. 1221, 2024.

AUTHORS

Haya Alqahtani is a student in the master's program in Computer Information Systems at King Abdulaziz University. Her research interests are machine learning, data science, and natural language processing.

Hana Almagrabi is an assistant professor in the Department of Information Systems at King Abdulaziz University. Her research interests are in Text Mining and Natural Language Processing, Machine Learning,

and Data Science. Topics she has worked on include Information Extraction, Prediction Models, Text Mining, and Content Analysis.

Amal Alharbi is an assistant professor in the Department of Information Systems at King Abdulaziz University. Her research interests are in Natural Language Processing, Information Retrieval, Machine Learning, and Data Science. Topics she has worked on include Information Extraction, Data Mining, and Systematic Reviews.