Predicting Employee Turnover through Machine Learning and Data Analytics

Kiran Kumar Reddy Yanamala

Central Michigan University

Abstract

In today's rapidly evolving business environment, employee retention has become a critical concern for organizations striving to maintain operational efficiency and reduce turnover costs. This paper presents a data-driven analysis of employee engagement, job satisfaction, salary, training hours, and their impact on turnover risk. Using a simulated dataset of 1,000 employees, we explore the relationships between these key factors through statistical analysis and predictive modeling. A Random Forest model is employed to predict employee turnover and assess the relative importance of each variable. The results reveal that job satisfaction and salary are the strongest predictors of turnover, with engagement scores and training hours also playing significant but less influential roles. Our findings provide actionable insights for human resource professionals, enabling them to develop targeted retention strategies that address the most impactful factors driving employee attrition. The study contributes to the existing literature by integrating multiple metrics into a comprehensive predictive model, offering a holistic understanding of employee retention dynamics.

Keywords: *Employee Engagement, Job Satisfaction, Salary, Training Hours, Turnover Prediction, Data-Driven Analysis, Workforce Management.*

Introduction

Current competitive business landscape, organizations are increasingly recognizing the critical role of employee retention in maintaining operational efficiency, ensuring consistent service delivery, and reducing recruitment costs [1]. High employee turnover not only disrupts workflow but also incurs significant financial and intellectual losses, as experienced personnel leave with valuable institutional knowledge. Retaining top talent has therefore become a strategic priority for human resource management (HRM), prompting the need for deeper insights into the factors influencing employee retention and turnover [2]–[4].

Traditionally, employee turnover has been analyzed through discrete factors such as job satisfaction, compensation, and career progression opportunities. However, these conventional approaches often fail to account for the complex interplay between multiple influencing factors, limiting their predictive power. Recent advancements in data analytics and machine learning have opened up new opportunities for HR professionals to adopt a more data-driven, holistic approach to workforce management. By leveraging comprehensive employee data, organizations can better understand how a combination of factors—ranging from employee engagement and training hours to salary and job satisfaction—affect turnover risks [5]–[7].

This study aims to explore the relationships between these key factors in employee retention by analyzing a simulated dataset of 1,000 employees. We focus on four critical variables: engagement scores, training hours, salary, and job satisfaction, examining how each correlates with the likelihood of turnover. Utilizing statistical analysis and machine learning techniques such as Random Forest models, this paper seeks to identify the most influential predictors of employee turnover, offering actionable insights for organizations to improve retention strategies.

Through this analysis, we aim to fill the gap in HR literature where multiple employee metrics are rarely integrated into a single predictive model. While numerous studies have explored the impact of individual variables like salary or job satisfaction on turnover, few have examined how these variables interact to form a comprehensive picture of retention dynamics. Our findings will provide organizations with a clearer understanding of the key drivers of employee turnover and offer data-driven recommendations for enhancing workforce retention and engagement. The objective of this research is twofold: first, to identify the relationships and interactions between engagement, job satisfaction, salary, and training hours; and second, to determine the relative importance of each factor in predicting employee turnover. Ultimately, this paper aims to guide HR departments in designing more effective, data-driven retention strategies that address the underlying causes of employee attrition.

Literature RevieW

Employee engagement, job satisfaction, salary, and training are vital factors in understanding turnover and retention in modern organizations. Recent literature highlights the complex interplay of these variables and the impact of data-driven approaches on predicting and managing employee turnover. This review synthesizes findings from 20 studies to explore the relationships between these factors and their implications for retention strategies, with a focus on machine learning applications in HR management.

Employee engagement is often seen as a critical factor influencing turnover intention. Engaged employees tend to display higher levels of commitment, which reduces their likelihood of leaving an organization. Memon et al. (2020) explored the impact of HR practices on work engagement and found that satisfaction with training and performance appraisals significantly improved engagement, which in turn reduced turnover intentions [8]. Similarly, Brunetto et al. (2012) emphasized that emotional intelligence positively influenced job satisfaction and engagement, which led to lower turnover rates among police officers [9]. Karatepe & Ağa (2012) also demonstrated that work engagement acts as a mediator between personality traits and job outcomes, such as job satisfaction and turnover intentions. Their findings highlight the importance of fostering engagement to enhance retention [10].

Salary is frequently discussed in turnover models, but research suggests that job satisfaction might play a more nuanced role in predicting turnover. Alarcon & Edwards (2011) found that engagement was a stronger predictor of job satisfaction and turnover intentions than burnout, indicating that engagement has a direct and powerful effect on retention [11]. Moreover, Gevrek et al. (2017) discussed how employee perceptions of salary raises, relative to their peers, influenced turnover decisions, suggesting that perceptions of fairness are as critical as actual compensation levels in turnover prediction [12]. Piyush et al. (2014) examined the antecedents of job satisfaction and found that drivers such as work environment and personal characteristics significantly affected job satisfaction and, subsequently, turnover. These results suggest that organizations must consider individual employee profiles to address turnover effectively [13].

Several studies have emphasized the importance of training in enhancing job satisfaction and retention. Liu et al. (2018) applied machine learning models to predict turnover based on job skills, highlighting that employees with well-defined expertise are less likely to leave their organizations [14]. Additionally, Hall (2018) found that machine learning models that include training hours, among other variables, can accurately predict turnover risks, allowing organizations to intervene early and retain key talent [15]. Saraswati (2019) also demonstrated that psychological capital and organizational justice significantly impacted work engagement, which was further linked to turnover intention. Employees who felt fairly treated and invested in training were more likely to stay with their organizations [16].

Machine learning has emerged as a valuable tool for predicting employee turnover. Jones et al. (2010) tested the Unfolding Model of Voluntary Turnover (UMVT) and found that traditional models based on job satisfaction were insufficient to capture the complexity of turnover decisions. They proposed an extended UMVT model that integrates labor market forces and personal circumstances [17]. Porter et al. (2019) conducted a meta-analysis to examine how different network positions affect turnover. Their study showed that instrumental and expressive network centrality had distinct effects on turnover. Employees in central network positions were more likely to stay due to the social capital they accumulated, underscoring the importance of social networks in retention strategies [18], [19].

Methodology

The methodology section details the approach used to analyze the relationships between employee engagement, job satisfaction, salary, training hours, and their impact on turnover risk. This study adopts a combination of exploratory data analysis and machine learning modeling to uncover significant correlations and predict employee turnover. The significance of the stuided variables is shown in Figure 1.

A. Data Collection

For the purpose of this study, a simulated dataset of 1,000 employees was constructed to emulate real-world organizational contexts. The dataset includes multiple departments, job positions, and varying experience levels to capture a wide range of employee characteristics. The key variables selected for this analysis are commonly associated with employee engagement and retention in the HR literature. These variables include engagement score, training hours, salary, job satisfaction, and turnover status.

Engagement score is measured on a continuous scale from 1 to 10, reflecting the level of emotional commitment an employee has toward the organization. Training hours represent the total number of hours an employee has undergone training per quarter, ranging from 0 to 20 hours. Salary is captured as the employee's annual compensation in U.S. dollars, while job satisfaction is rated on a scale of 1 to 5. Turnover status, the dependent variable for the prediction model, is binary, indicating whether an employee has stayed with the organization or left. The constructed dataset, while simulated, is designed to mirror actual organizational data that would typically be collected by HR departments. This provides a rich foundation for both exploratory analysis and predictive modeling, allowing for a deep exploration of the factors influencing employee turnover.

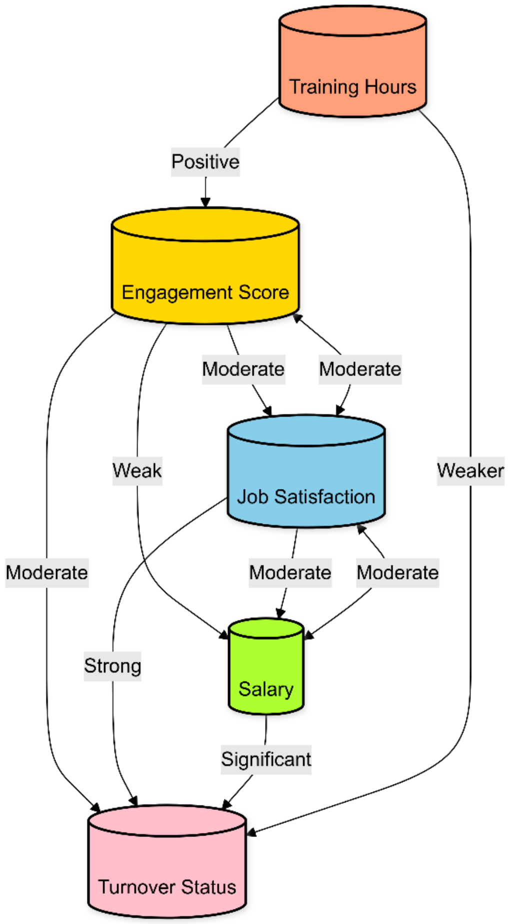

Figure 1 The relationships between the primary variables studied: Engagement Score, Job Satisfaction, Salary, Training Hours, and Turnover Status.

B. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was conducted to gain initial insights into the dataset’s structure and distributions. This step is crucial for identifying patterns, relationships, and any potential anomalies that might influence the predictive models. The EDA involved calculating descriptive statistics, such as mean, median, and standard deviation, for each variable to understand their central tendencies and dispersions.

Visualizations were employed to illustrate the distribution of key variables. Histograms were used to examine the frequency distributions of engagement scores and training hours, revealing that the majority of employees scored around 6 on the engagement scale and underwent approximately 10 hours of training per quarter. Scatterplots were also generated to observe pairwise relationships between the variables, with a focus on how engagement, job satisfaction, and salary correlate with turnover risk. These visualizations were instrumental in highlighting potential patterns that could be explored further through modeling. Additionally, a correlation matrix was computed to quantify the strength and direction of relationships between engagement scores, training hours, salary, and job satisfaction. This matrix provided a foundational understanding of how these variables interact and how they might collectively influence turnover risk.

C. Predictive Modeling

To predict employee turnover and assess the influence of various factors, a Random Forest model was employed. This machine learning technique was chosen for its ability to handle complex interactions between variables and to assess the importance of each predictor in determining turnover risk. Random Forest operates by constructing multiple decision trees during training and aggregating their results to improve prediction accuracy and reduce overfitting. The model was trained on 80% of the dataset, leaving 20% for testing. Turnover status was the target variable, with engagement score, training hours, salary, and job satisfaction as predictors. The Random Forest algorithm is well-suited for this study due to its robustness in handling continuous and categorical data, making it ideal for examining how these factors contribute to turnover prediction. Feature importance was calculated to identify the most influential variables in the prediction model. This analysis revealed that job satisfaction and salary were the strongest predictors of turnover, followed by training hours and engagement score. The results offer HR professionals a clearer understanding of which areas to focus on when developing retention strategies.

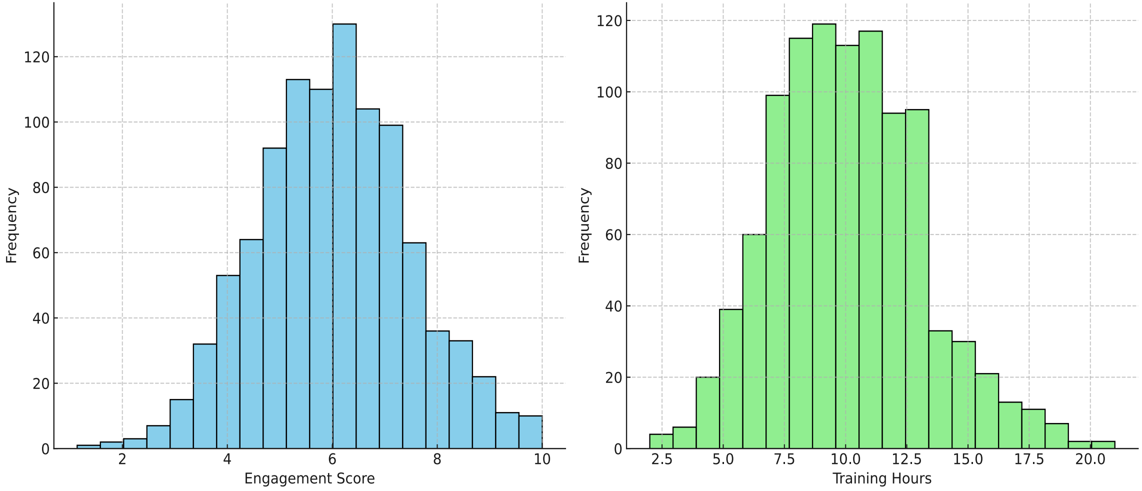

Figure 2 the distributions of engagement scores (left) and training hours (right) across employees in the dataset. The majority of employees have engagement scores around 6, and most employees undergo around 10 hours of training. Both distributions appear roughly normal.

II. RESULTS AND DISCUSSION

This section presents the key findings from the exploratory data analysis (EDA) and predictive modeling, focusing on the relationships between engagement scores, training hours, salary, job satisfaction, and turnover risk. Each result is accompanied by a discussion on its implications for employee retention strategies.

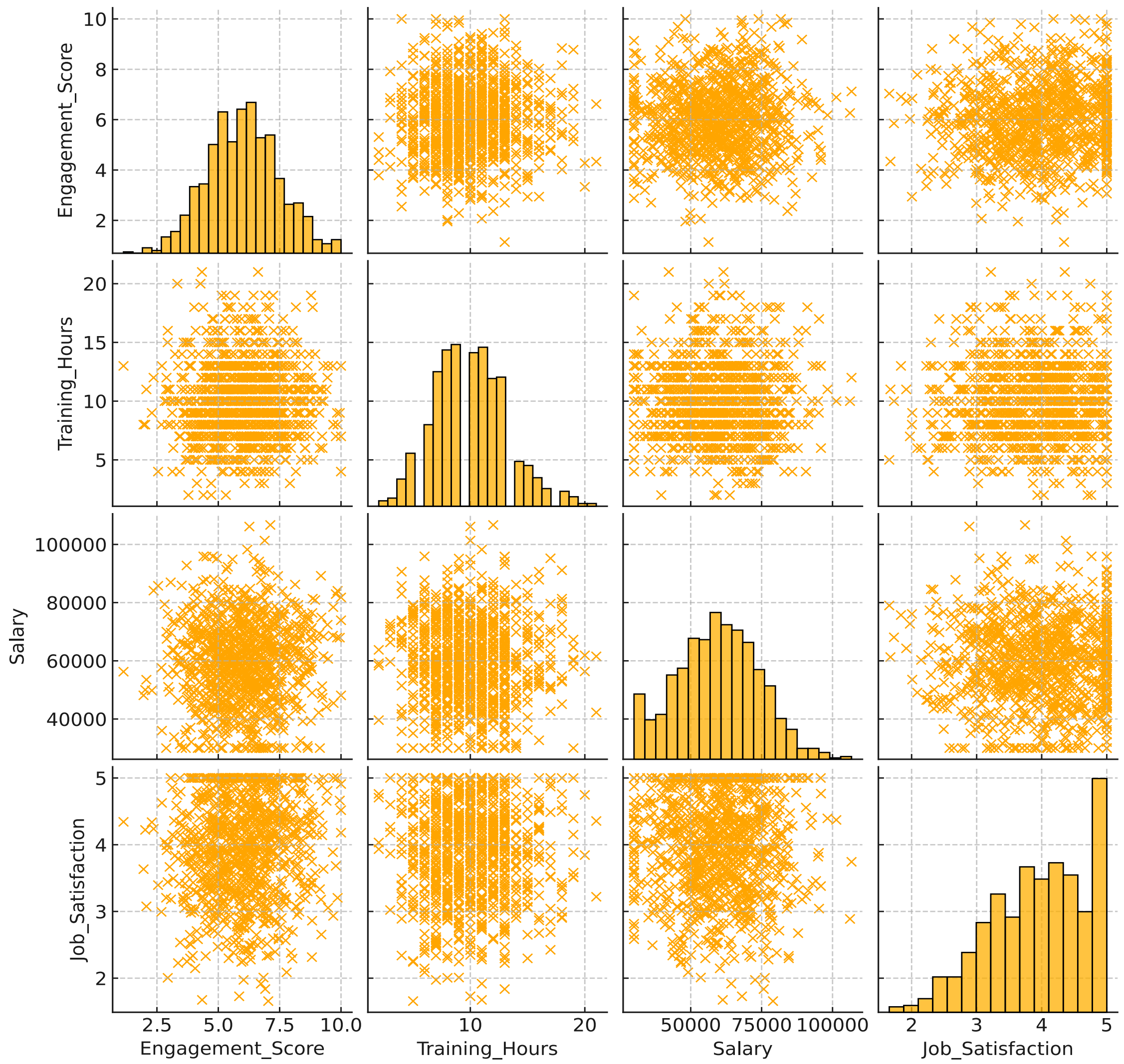

Figure 3 The relationships between engagement score, training hours, salary, and job satisfaction for the dataset. The diagonal elements display the distribution of each variable, while the off-diagonal elements provide scatter plots to visualize any potential correlations between the variables.

A. Distribution of Engagement Scores and Training Hours

The distribution of employee engagement scores and training hours is depicted in Figure 2. The histogram on the left shows that the majority of employees have engagement scores clustered around a score of 6, with a relatively symmetric distribution ranging from 2 to 10. The histogram on the right indicates that most employees undergo around 10 hours of training per quarter, with the distribution slightly skewed toward higher training hours. Both variables exhibit roughly normal distributions, which provides a reliable basis for further statistical analysis.

These results suggest that employee engagement, on average, is moderate, while most employees receive a substantial amount of training. The normal distribution of these variables indicates that there are no significant outliers, which simplifies the modeling process. The focus on employees with lower engagement scores and fewer training hours could be critical for designing interventions aimed at improving overall workforce retention.

B. Correlation Analysis

Figure 3 presents a matrix of scatter plots showing the relationships between engagement score, training hours, salary, and job satisfaction. The diagonal elements illustrate the distribution of each variable, while the off-diagonal scatter plots highlight the pairwise relationships between them. The correlation analysis reveals several notable patterns:

- **Engagement Score and Job Satisfaction**: A positive correlation is observed, suggesting that higher engagement scores are associated with higher levels of job satisfaction. Employees who are more engaged tend to report greater satisfaction with their jobs, underscoring the importance of engagement as a driver of employee morale and retention.

- **Training Hours and Salary**: The scatter plot shows a weak positive relationship between training hours and salary, indicating that employees who receive more training tend to earn higher salaries. However, this relationship is not particularly strong, suggesting that while training may contribute to skill development and potential salary increases, other factors such as experience and job role likely play a more significant role in determining salary levels.

- **Job Satisfaction and Salary**: A moderate positive correlation is observed between job satisfaction and salary. Employees who earn higher salaries tend to report greater job satisfaction, aligning with the well-established idea that compensation is a key factor in employee retention.

- **Engagement Score and Training Hours**: The relationship between engagement and training hours is relatively weak, indicating that while training may improve engagement for some employees, it is not a decisive factor in determining overall engagement levels.

These correlations provide insights into the interconnectedness of the variables, suggesting that job satisfaction and salary are more closely linked than engagement and training. The weak correlation between engagement and training hours highlights the need for organizations to explore other factors that might influence employee engagement, such as leadership, work environment, and career development opportunities.

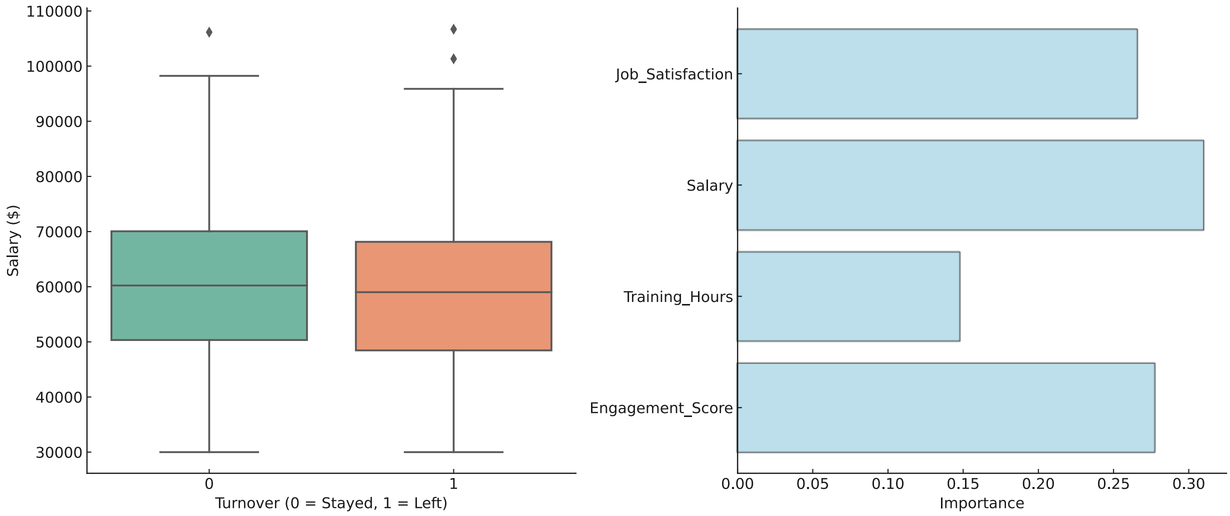

Figure 4 The left plot shows the salary distribution by employee turnover status, comparing employees who stayed (Turnover = 0) with those who left (Turnover = 1). The right plot illustrates the importance of different features (Job Satisfaction, Salary, Training Hours, and Engagement Score) in predicting employee turnover, based on a Random Forest model.

C. *Feature Importance in Turnover Prediction*

The predictive modeling results are summarized in Figure 4, which compares the salary distributions of employees who stayed with the company (Turnover = 0) versus those who left (Turnover = 1). The box plot on the left shows that employees who stayed tend to have slightly higher median salaries than those who left. However, there is significant overlap between the two groups, suggesting that salary alone is not a definitive predictor of turnover.The right side of Figure 4 presents the feature importance analysis based on the Random Forest model. The model identified Job Satisfaction as the most significant predictor of turnover, followed closely by Salary. Training Hours and Engagement Score were also found to be important, though their impact on turnover prediction was less pronounced compared to job satisfaction and salary.These findings suggest that organizations looking to reduce employee turnover should prioritize efforts to enhance job satisfaction and ensure competitive compensation packages. While training and engagement are also important, they appear to have a more indirect influence on turnover. Therefore, HR departments may benefit from focusing on improving the overall employee experience, particularly in terms of job satisfaction, to reduce attrition rates.

The salary distribution by turnover status, shown in Figure 4, provides additional context for understanding the role of compensation in employee retention. Employees who left the organization had a broader range of salaries compared to those who stayed, with some turnover occurring even among higher-paid employees. This finding underscores the complexity of turnover dynamics, where factors such as job satisfaction and career development opportunities may outweigh salary considerations in some cases

Conclusion

This study has provided a comprehensive analysis of the key factors influencing employee retention by examining the relationships between engagement scores, training hours, salary, job satisfaction, and turnover risk. The findings demonstrate that job satisfaction and salary are the most significant predictors of turnover, reinforcing the importance of addressing both financial and non-financial aspects of the employee experience. While engagement scores and training hours also contribute to turnover prediction, their influence is more indirect, suggesting that engagement strategies should focus on broader organizational factors such as work environment, career development opportunities, and employee recognition. The application of a Random Forest model allowed us to quantify the relative importance of these factors, offering HR professionals a data-driven tool for identifying at-risk employees and developing more targeted retention strategies. By emphasizing job satisfaction and salary improvements, organizations can mitigate turnover risks and enhance employee loyalty.

Despite its contributions, the study's reliance on a simulated dataset limits its generalizability. Future research should validate these findings using real-world data to confirm the robustness of the predictive model. Additionally, incorporating other variables such as leadership quality, work-life balance, and organizational culture could provide a more nuanced understanding of the factors driving employee retention.

Reference

[1] A. Lertxundi, “Transfer of HRM practices to subsidiaries: Importance of the efficiency of the HRM system,” *Manag. Res. J. Iberoam. Acad. Manag.*, vol. 6, no. 1, pp. 63–73, Apr. 2008.

[2] A. A. Mumin, A. S. Achanso, M. I. Mordzeh-Ekpampo, B. Boasu, and D. Dei, “Employee Turnover and Job Satisfaction: A synthesis of Factors influencing employee turnover in Institutions of Higher Learning in Ghana,” *Research Square*, 13-Sep-2018.

[3] D. N. Muhammad, D. A. Ihsan, and D. K. Hayat, “Effect of workload and job stress on employee turnover intention: A case study of higher education sector of Khyber Pakhtunkhwa,” *jbt*, vol. 7, no. 1, pp. 51–64, Jun. 2017.

[4] D. Bufquin, J.-Y. Park, R. M. Back, J. V. de Souza Meira, and S. K. Hight, “Employee work status, mental health, substance use, and career turnover intentions: An examination of restaurant employees during COVID-19,” *Int. J. Hosp. Manag.*, vol. 93, no. 102764, p. 102764, Feb. 2017.

[5] J. Yuan, "Research on employee turnover prediction based on machine learning algorithms," in *2018 4th International Conference on Artificial Intelligence and Big Data (ICAIBD)*, Chengdu, China, 2018.

[6] R. Valarmathi, M. Umadevi, and T. Sheela, "Employee turnover prediction using single voting model," in *Applied Learning Algorithms for Intelligent IoT*, Boca Raton: Auerbach Publications, 2018, pp. 153–174.

[7] R. Chakraborty, K. Mridha, R. N. Shaw, and A. Ghosh, "Study and prediction analysis of the employee turnover using machine learning approaches," in *2018 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON)*, Kuala Lumpur, Malaysia, 2018.

[8] M. A. Memon *et al.*, "Satisfaction matters: the relationships between HRM practices, work engagement and turnover intention," *Int. J. Manpow.*, vol. 42, no. 1, pp. 21–50, Mar. 2019.

[9] Y. Brunetto, S. T. T. Teo, K. Shacklock, and R. Farr-Wharton, "Emotional intelligence, job satisfaction, well-being and engagement: explaining organisational commitment and turnover intentions in policing," *Hum. Resour. Manag. J.*, vol. 22, no. 4, pp. 428–441, Nov. 2012.

[10] O. M. Karatepe and M. Aga, "Work engagement as a mediator of the effects of personality traits on job outcomes: A study of frontline employees," *Serv. Mark. Q.*, vol. 33, no. 4, pp. 343–362, Oct. 2012.

[11] G. M. Alarcon and J. M. Edwards, "The relationship of engagement, job satisfaction and turnover intentions," *Stress Health*, vol. 27, no. 3, pp. e294–e298, Aug. 2011.

[12] D. Gevrek, M. K. Spencer, D. Hudgins, and V. Chambers, "I can't get no satisfaction: The power of perceived differences in employee retention and turnover," *SSRN Electron. J.*, 2017.

[13] P. Kumar, M. Dass, and O. Topaloglu, "Understanding the drivers of job satisfaction of frontline service employees," *J. Serv. Res.*, vol. 17, no. 4, pp. 367–380, Nov. 2014.

[14] J. Liu, Y. Long, M. Fang, R. He, T. Wang, and G. Chen, "Analyzing employee turnover based on job skills," in *Proceedings of the International Conference on Data Processing and Applications*, Guangdong China, 2018.

[15] O. P. Hall, "Managing employee turnover: machine learning to the rescue," *Int. J. Data Sci.*, vol. 6, no. 1, p. 57, 2019.

[16] K. D. H. Saraswati PHD, "Work engagement: The impact of psychological capital and organizational justice and its influence on turnover intention," *J. Mgt. Mkt. Review*, vol. 4, no. 1, pp. 86–91, Mar. 2019.

[17] S. M. Jones, A. Ross, and B. Sertyesilisik, "Testing the unfolding model of voluntary turnover on construction professionals," *Constr. Manage. Econ.*, vol. 28, no. 3, pp. 271–285, Mar. 2010.

[18] C. M. Porter, S. E. Woo, D. G. Allen, and M. G. Keith, "How do instrumental and expressive network positions relate to turnover? A meta-analytic investigation," *J. Appl. Psychol.*, vol. 104, no. 4, pp. 511–536, Apr. 2019.

[19] "Supplemental material for how do instrumental and expressive network positions relate to turnover? A meta-analytic investigation," *J. Appl. Psychol.*, 2019.