Journal of Informatics and Web Engineering

Vol. 5 No. 2 (June 2026) eISSN: 2821-370X

Machine Learning for Employee Attrition Prediction: A Comparative Study

Low Kai Jia[1], Lew Sook Ling[2*], Sri Winarno[3]**

[1]Faculty of Information Science and Technology (FIST), Multimedia University (MMU), Jalan Ayer Keroh Lama, Melaka, 75450 Bukit Beruang, Malaysia.

[2]Centre for Advanced Analytics, CoE for Artificial Intelligence, FIST, MMU, Jalan Ayer Keroh Lama, Melaka, 75450 Bukit Beruang, Malaysia.

[3]Research Center for Intelligent Distributed Surveillance and Security (IDSS), Faculty of Computer Science, Universitas Dian Nuswantoro, Jl. Imam Bonjol No.207 Gedung H, Pendrikan Kidul, Kec. Semarang Tengah, Kota Semarang, Jawa Tengah 50131, Indonesia.

**corresponding author: (sllew@mmu.edu.my; ORCiD: 0000-0003-4545-1163)*

***corresponding author: (sri.winarno@dsn.dinus.ac.id; ORCiD: 0000-0002-6801-2767)*

Abstract - Employee turnover remains a long-standing issue that can lead to significant economic losses and a decline in workplace productivity. Traditional methods for predicting natural attrition, such as surveys and rumour assessment, usually fail to capture the complexity and multi-factorial nature of natural attrition. To address this limitation, this paper proposes to develop an advanced prediction model using Machine Learning (ML) algorithms to enhance the accuracy and reliability of export prediction. These methods are applicable to a full set of data covering parameters such as demographics, workplace performance and survey scores. The objective of this model is to provide a data-driven model that can identify employees at risk of leaving the organization, thereby enabling intervention at the right time and place. The objectives in the paper are to identify employees who are more likely to leave due to work and social experience, second to understand the main variables that lead to employee turnover, and finally to make the model effective in reducing employee turnover and improving employee satisfaction. In this paper, a ML-based approach is introduced to enhance turnover prediction accuracy by integrating comprehensive features and addressing class imbalance demonstrating improved performance and interpretability as compared to previous studies. Through the application of these ML methods, this paper is beneficial to both academic research and practical practice in human resource management. The results of the paper emphasized the data analysis and ML opportunities, to help solve the problem of employee retention, and create a more close, more effective, more motivated staff. In this paper, the proposed ML model performs well in predicting employee turnover by effectively integrating key features as compared with traditional methods, this method has a higher accuracy rate.

Keywords—Employee Turnover, Machine Learning, Predicting, Random Forecast, Retention Strategies.

Received: 16 July 2025; Accepted: 21 August 2025; Published: 16 June 2026

1. INTRODUCTION

Human resource management is one of the fundamental operations of any organization and is of great significance in the selection, training and maintenance of talents. Good human resource management practices have the potential to enhance employees' engagement, enthusiasm and productivity, create a favourable working environment, prevent excessive employee turnover, and at the same time, they can artificially increase employees' satisfaction and loyalty [1]. In the current economic world, the knowledge base has become a key element of any type of business and employees are mostly considered as the main asset of the organization. Additionally, especially the loss of key personnel or high performance may lead to the high cost of recruitment and training, it may damage business continuity and operating efficiency.

In addition, job changes of employees usually refer to employee attrition, such as resignation, new hiring, internal transfer, etc. This is not only an important indicator for measuring the degree of satisfaction of the human resource management process with the organization and employees but also reveals the trends in the labour market [2]. High employee turnover rates can lead to low morale among employees, high operating costs, and have a negative impact on organizational performance. However, to a certain extent, employee turnover can bring about new paths and new ways of thinking, thereby enhancing the flexibility and adaptability of the organization [3].

Employee turnover has always been a challenge in Human Resource Management, especially in highly competitive modern enterprises. This is because many organizations lack effective and precise tools to predict the risk of employee turnover, and thus it is not easy for them to propose specific intervention and retention measures at the appropriate time. Traditional methods for analysing employee turnover largely rely on exit interviews, making it difficult to describe the underlying reasons for an employee's departure [4].

With the rapid development of AI and data analysis technologies, an increasing number of organizations are beginning to use Machine Learning (ML) algorithms to predict employee turnover. This prediction model, based on characteristic variables such as employee performance, job satisfaction, overtime situation, and job changes, reveals potential employees who may leave and provides relevant intervention suggestions for the human resources department, including training, job adjustment, and salary optimization. This information-based approach is more accurate and proactive than the one based on subjective judgment [5]. At present, deep learning algorithms such as Decision Tree (DT), Random Forest (RF) and Gradient Boosting (GB) have been widely applied in the field of employee turnover prediction. They will have the opportunity to fully identify the trends in the historical data of employees, thereby contributing to the scientific and intelligent level of Human Resource Management [6].

Therefore, this paper aims to construct an employee turnover prediction model that integrates multiple ML algorithms. Analysing the historical behaviours and performance characteristics of employees helps organizations identify employees with high turnover risks early and formulate targeted retention strategies to enhance the talent stability and overall competitiveness of the organization.

In addition, the employee attrition prediction model based on ML not only enhances the scientific and forward-looking nature of Human Resource Management but also provides enterprises with more effective and targeted decision support. The paper results are expected to provide data-driven strategic references for organizations in fierce competition for talent, optimize the allocation of human resources, and help enterprises achieve sustainable development.

2. LITERATURE REVIEW

2.1 Theory Turnover

Employee turnover, which refers to the frequency with which existing employees leave and are replaced by new employees, is a common management problem. High turnover increases the expenses associated with recruitment and training, causes the depletion of human capital, and significantly affects organizational performance [7]. However, current theories are insufficient to fully explain and predict employee mobility. This paper can focus on analysing this issue, which should become more comprehensive and apply some more advanced research methods. This paper aims to provide different perspectives on issues related to employee turnover and how organizations can manage employee turnover moderately by adopting human resource analysis methods.

2.1.1 Turnover and Commitment

The literature explores organizational commitment (OC) and its impact on employee turnover. According to most paper, high levels of OC are strongly associated with lower employee turnover, because dedicated employees are more likely to adhere to the organization's goals and values. Hence, by providing opportunities for career development and recognizing employee contributions, emotional attachment can be enhanced, which reduces turnover risk [8]. The second paper further highlighted competitive compensation packages, clear career development plans and regular performance reviews as key measures to increase employee commitment [9]. In addition, these studies also point to the significant role of predictive analytics in identifying employees at high risk of turnover, and with data-driven forward-looking interventions, Human Resource (HR) departments can develop personalized retention strategies, such as providing career development guidance or adjusting compensation packages, to effectively reduce employee turnover.

2.1.2 Turnover and Trust

Employee turnover is the term used to describe the process by which an organization naturally reduces its workforce owing to several unavoidable circumstances, which typically results in significant losses for the organization. Past study has shown that higher levels of trust are positively correlated with lower employee turnover [10]. When employees feel trusted by management and colleagues, they are less likely to leave the company and more likely to remain with it. Trust can be built by enhancing job satisfaction, promoting open communication, and supportive leadership, resulting in a stable work environment. In addition, the level of employee turnover also reflects the effectiveness of organizational culture and management practices. Therefore, by establishing a culture of trust, organizations can effectively reduce employee turnover and improve overall performance.

2.1.3 Turnover and Competencies

The concept of employee turnover and skill level argues that turnover rate indicates the frequency with which employees leave the company and are replaced by new employees. Therefore, the possibility of expecting employee turnover is crucial to the long-term development of an organization. Competence refers to the specific skills, knowledge and abilities of an employee. In the highly competitive market environment, organizations pay more attention to the matching of employees' abilities and needs [11]. When employees feel they have the necessary competencies and are challenged on the job, they are less likely to leave.

2.2. Support Vector Machine (SVM)

SVM is a supervised learning method frequently applied in classification and regression tasks. Its fundamental concept is to enhance category discrimination by effectively classifying through the recognition of hyperplanes. According to one study, SVM outperforms Logistic Regression (LR) in nonlinear classification using kernel approach, achieving 84.38% accuracy [12]. According to another study, SVM may achieve up to 91.4% accuracy in generalization when working with high-dimensional data [8]. By optimizing hyperparameters, their performance can be further improved, and more targeted interventions can be provided for HR departments to help effectively control employee turnover. In these papers is evaluated with other classifiers to compare its effectiveness in predicting employee turnover, demonstrating its competitive performance in dealing with complex and high-dimensional HR data.

2.2.1 LR

LR is a statistical method mainly applied in binary classification, namely the possibility of an employee leaving or staying. This model provides a probability value such that the probability of the observed value belonging to the specified category is the most likely. During training, the optimal model is sought by minimizing the cost function to best estimate the probability of the target variable. Simplicity is one of the main advantages of LR, especially in analysing the importance of features. For binary, it can be used to provide predictive probabilities and classify them by applying the thresholds provided by the user [6]. The baseline model used in this paper is LR to test its

interpretability and rationality in predicting employee turnover rates, serving as a baseline for comparing more complex ML algorithms.

2.2.2 Neural Network (NN)

The employee turnover prediction model uses NN technology and combines multi-dimensional information from large data sets for in-depth analysis. Through the training of the NN model, the model can accurately predict the turnover risk of employees based on the personal characteristics of employees, such as age, position, years of work, performance evaluation, salary level, job satisfaction and other data. In this paper, the NN model used consists of seven hidden layer nodes, each of which maps the input of data nonlinearly through activation functions to help the model identify key characteristics of potential departing employees. Through the optimization of the activation function and the cross-entropy loss function, the accuracy rate of this model reaches 97.0%, demonstrating considerable strength in predicting employee turnover [13]. In this paper, the NN model to the task of employee exit prediction and evaluated its performance, verifying its effectiveness in dealing with complex nonlinear relationships and improving prediction accuracy.

2.2.3 RF

RF is an ensemble learning method that makes predictions through the voting mechanism of multiple DTs. In the task of employee turnover prediction, RF is widely used to analyse many variables and improve the accuracy and robustness of prediction through voting mechanism. By constructing multiple DTs, each tree is trained with different data subsets and feature subsets, and finally the prediction results of all trees are summarized, thereby reducing the risk of overfitting and improving the generalization ability of the model. RF can automatically assess the importance of features, which makes it particularly good at feature selection, and can identify the factors that have the greatest impact on employee turnover, such as monthly income and years of service [14]. In addition, the RF model has strong processing power and can cope with complex data sets, including high-dimensional data and missing data. Random forecasts can increase prediction accuracy and prevent the bias issue that a single DT may face by combining the prediction outcomes of several DTs. As a result, RF is a strong tool for forecasting employee turnover in human resource management, having demonstrated great efficiency and consistent performance in this area. RF's accuracy in forecasting staff turnover was 86.4% [15]. Managing complicated data, lowering the chance of overfitting, managing huge data efficiently, and recognizing the significance of evaluating traits to assist in identifying critical elements influencing employee mobility are some of its advantages. In this paper, RF algorithm to model and analyse the employee turnover risk, verifying its advantages in multivariate processing and feature importance assessment, providing support for enterprises to formulate more targeted employee retention strategies.

2.2.4 DT

The DT model which performed efficient and intuitive forecasting tool that splits data by characteristics such as monthly income and years of service to generate a clear and regular path for predicting employee turnover. Its good interpretability and adaptability make it suitable for processing complex data. The DT's accuracy was 78%, and the recall rate was 57.35% [15]. The main affecting factors were age, monthly salary, and business tenure. With its modest requirements for data quality, ability to handle nonlinear interactions and missing values, and intuitiveness, DT rules are a dependable tool for effective management in HR applications. In this paper, the DT model to predict employee turnover not only identified the key influencing factors but also further verified its effectiveness in interpretability and practical application, providing HR with an intuitive and operational analytical tool.

In addition, the models applied in this study are RF, XGBoost, GB, NN, and DT. We selected these models for employee turnover prediction because each offers distinct strengths. DT serves as an interpretable baseline, while RF enhances robustness and reduces overfitting. GB and XGBoost were included for their strong predictive performance and ability to capture complex non-linear patterns in HR data, with XGBoost providing regularization and computational efficiency. NN were employed to explore the potential of deep learning in detecting subtle interactions between employee attributes, despite their lower interpretability. Overall, this combination of models balances accuracy, robustness, and interpretability, which are essential considerations in HR analytics.

2.3. Comparison Between ML Methods

This section presents findings from several related studies, highlighting the results, limitations, and values of each study. However, various studies predict the employee turnover rate has become an important research area for organizations aiming to enhance employee satisfaction and reduce the employee turnover rate. The traditional Human Resource Management methods are confronted with the problems of low efficiency and poor accuracy when dealing with the analysis of employee behaviours in the big data environment. Therefore, an increasing number of researchers are attempting to apply ML techniques in the development of models for predicting employee turnover. This paper collates and analyses the relevant research results on employee turnover prediction in recent years and systematically compares the current mainstream ML methods from aspects such as algorithm types, model performance, data feature processing, interpretability and application prospects.

Among different ML algorithms, ensemble learning technology is particularly remarkable. Empirical studies show that XGB, GB and RF models provide an impressive prediction performance. [16] used the DT model based on gradient enhancement to predict employee turnover. The Receiver Operating Characteristic - Area Under Curve (ROC AUC) value of this model is 0.9892 and the PR AUC value is 0.9795, which has greater superiority compared with the traditional methods. Social media behaviour is incorporated into the model, and the prediction accuracy and generalization ability are improved by combining techniques such as GB, LR, and k-Nearest Neighbors (KNN) [17]. Furthermore, [10] Utilized The Optimized (ETC) and verified the stability of the model through cross-validation based on achieving an accuracy rate of 93%, indicating that the integrated model has significant advantages in dealing with complex and imbalanced Human Resource data.

In addition to the integrated model, the traditional DT class method and the LR model are still widely adopted in practical applications due to their good interpretability. The study by [4] indicates that DT models not only have high predictive accuracy but also can assist human resource managers in identifying the key factors influencing employee turnover, thereby formulating targeted intervention measures. [18] achieved the highest R^2 value (0.94) by using the RF model. Meanwhile, it has short training time and high computational efficiency and is suitable for deployment in actual scenarios. According to the IBM HR dataset, the SVM and NN algorithms achieved accuracy rates of 91.4% and 90% respectively, and the generation feature variable was added [8]. They found that young, single and low-income employees aged between 18 and 35 had a higher turnover tendency. The explanatory ability of the model has been further enhanced.

Although deep learning has made significant progress in various prediction tasks, there are still specific challenges in using it for employee turnover prediction [9]. Several comparisons between ensemble models and deep NN models were studied. Their studies results show that deep learning technology has achieved the highest accuracy rate (94.52%), but it brings a large amount of training costs, a strong reliance on computing resources, and the limited interpretability of the model, which poses a challenge to small and medium-sized enterprises with limited resources. Meanwhile, some studies have also turned their attention to social factors and local data analysis. [19] conducted a questionnaire survey among manufacturing employees in Malaysia and found that salary levels, leadership styles, and interpersonal relationships significantly influence employees' intention to leave. Although the study did not use a complex model, its conclusion has strong practical reference value for regional enterprises. Table 1 shows the summary of the related studies.

According to the summary table, several key findings regarding the use of ML to predict employee turnover have emerged. Gradient-based models such as XGBoost and GB have consistently demonstrated very high prediction accuracy. The ROC AUC of the study by [16] was 0.9892, indicating its effectiveness in human resource analysis. Integrated and tree-based methods, including RFs and additional tree classifiers, have also demonstrated powerful performance in multiple studies, especially when combined with feature engineering or data balancing techniques. It is notable that some studies have introduced innovative perspectives. For instance, [17] utilized the Stimulus-Organic Response (S-O-R) theory to integrate social media features and expand the range of predicted features. The use of real-world datasets, such as the IBM HR Analytics dataset or primary surveys, can enhance the practical relevance of the model and allow for a deeper exploration of key variables such as age, tenure, and managerial engagement.

Table 1. Summary of the Related Studies

Ref.	Technique/ Algorithm	Dataset	Result	Value Added of the Current Study
[4]	DT, AdaBoost, Support Vector Machine	The dataset used in this study contained 50,000 employee records, which included the status of employees who were employed or had left.	T (AUC) = 0.78 AdaBoost (AUC =0.74) SVM (AUC = 0.50)	This study offers HR departments a predictive tool to reduce employee turnover through targeted incentives, helping lower training costs. It also contributes to an empirical case of ML in HR analytics, highlighting the value of data-driven decision-making in workforce management.
[8]	LR SVC NN	The IBM HR Analytics dataset used in this study was taken from Kaggle. This is a medium-sized dataset provided by IBM and it contains 1470 samples with 34 input features.	LR = 90 SVC = 91.40 NN=91.40	The value added to this study is to reveal the potential influencing factors of employee turnover through the generation of characteristics such as age group to generations and in-depth correlation analysis. By comparing the performance of various algorithms comprehensively, it provides a strong reference for the selection of employee retention prediction model in the future.
[9]	The study employed a combination of descriptive analysis, correlation analysis, and regression modelling. It also used ANOVA tests to evaluate model performance and identify significant predictors of employee attrition.	The dataset used in this analysis was sourced from Kaggle.com and comprises a total of 35 columns.	This model integrates multiple related variables and has the highest explanatory power with an R^2 value of 0.033, while its predictive power is average. However, the analysis confirms that younger employees and those with lower management engagement are more likely to leave. Although the R^2 value of a single model is relatively low, the combined method reveals meaningful wear prediction patterns.	This study provides actionable insights for HR professionals by identifying the key factors that influence employee turnover. It highlights YearsWithCurrManager as the most influential variable, followed by YearsInCurrentRole and YearsSinceLastPromotion, which can help organizations like IBM develop targeted retention strategies more effectively.
[10]	ETC	The IBM HR Employee Attrition dataset, created by IBM data scientists, was used for data analytics and building ML models to predict valuable employee attrition.	Accuracy = 93 Precision = 93 Recall = 93 F1 score = 93	This paper will use an integrated approach like ETC to make predictions with high accuracy, while using data balancing techniques to improve model performance.
[14]	LR DT SVM NN RF	Referring to the Price-Mueller model, the field information of data shared by IBM Watson Analytics platform, and the actual data collected in the survey, we have established the employee turnover prediction model.	LR = 94.5 DT = 94.63 SVM = 98.3 NN = 97.0 RF = 98.10	The paper will apply ML techniques to predict employee turnover with high accuracy, particularly using SVM, which achieved 98.3%.
[15]	Gaussian NB Bernoulli NB DT LR RF Multinomial NB	The dataset that we used in this paper is distributed by IBM Analytics.	Gaussian NB = 73.34 Bernoulli NB = 86 DT = 78 LR = 61.05 RF = 86.4 Multinomial NB = 50.5	This study presents a practical approach to predicting employee attrition using supervised ML models, leveraging real-world IBM data. By analysing key personal and professional factors, it identifies patterns that influence voluntary resignation decisions. Among the tested algorithms, Gaussian Naive Bayes proved most effective, particularly in minimizing false negatives, making it a valuable tool for HR departments aiming to proactively manage talent retention and reduce turnover-related costs.

[16]	Gradient-Based DTs	This study uses the dataset, which is HR Analytics dataset from Sisodia, Vishwakarma & Pujahari (2017) to develop our prediction model.	ROC AUC = 0.9892 PR AUC = 0.9795	This study offers a practical tool for predicting employee turnover using gradient-boosted trees, supporting better retention strategies and organizational stability. The approach is also applicable to other business data analysis tasks.
[17]	GB LR KNN	The dataset consists of 597 valid survey responses from IT personnel exposed to IoT, collected via an online questionnaire between 2021 and 2022. Respondents had an average age of 35 and 11 years of experience. The survey included 35 questions rated on a 5-point Likert scale.	GB = 88.4 LR = 56.5 KNN = 92.6	This study added the stimulus-organic-response theory, to incorporate social media characteristics into the prediction of employee turnover intention, highlighting the importance of social media in human resource analysis.
[18]	LR KNN Classifier Support Vector Machines Naive Bayes DTs RF	IBM HR Analytics Employee Attrition & Performance dataset from Kaggle	LR = 87.71 KNN = 59.22 SVM = 86.59 NB = 83.24 DT = 80.45 RF = 83.24	The value added is that it not only demonstrates the application of ML in employee turnover prediction but also the importance of data-driven methods in improving the efficiency of enterprise human resource management is emphasized, which provides a new perspective and methodological basis for future research in related fields.
[19]	Multiple Regression Analysis, Correlation Coefficient Analysis	The data for this paper came from 136 employees of a manufacturing company in Malaysia.	The results showed that 27.2% and 25.7% of respondents had worked for the company between five and twelve years. Respondents with less than RM3000 income level made up the highest percentage at 60.3%.	This study reveals the key factors that influence the turnover intention of Malaysian manufacturing employees. The results of this study provide empirical evidence for enterprises to optimize salary policy, improve leadership style and strengthen employer-employee relationship, so as to help formulate more effective employee retention strategies.

In conclusion, different ML methods have their own advantages in employee turnover prediction. Ensemble learning is suitable for processing large-scale and complex data, with high accuracy and stability. LR and DT models have strong interpretability and are convenient for practical applications. Deep learning is precise but requires many resources. In the future, local data and social factors should be combined to enhance the practicality of the model and assist enterprises in optimizing their talent retention strategies. In this paper, the ML model applied across the different datasets to predict the employee turnover, achieving high accuracy and demonstrating the practical value of ML in supporting HR retention strategies.

3. RESEARCH METHODOLOGY

3.1 Proposed Framework

The process of predicting employee turnover includes several basic stages. The initial stage involves creating a dataset that includes attributes such as job title, age, department, tenure, performance evaluation, and employment status (resignation or retention) as the basis for analysis. Next, data preprocessing is carried out to clean up the original data, including filling in missing values, converting categorical variables into numerical formats such as one-hot encoding or label encoding, standardizing or normalizing the numerical variables, and addressing class imbalance through methods such as oversampling or Synthetic Minority Over-sampling Technique (SMOTE). Ensemble models, such as XGBoost and GB, are applied to enhance predictive performance. After data cleaning, Exploratory Data Analysis (EDA) is conducted using visualization tools such as histograms, bar charts, box plots, and correlation heat maps to identify patterns, trends, key variables, and eliminate invalid features. Once the data is prepared, it is divided into a training set (70%) and a test set (30%) for model training. Multiple ML algorithms, such as LR, DT, RF, XGB, and NN, are applied for modelling. The model evaluation stage assesses predictive performance through metrics such as accuracy, precision, recall, F1 score ,and ROC AUC. Models that do not reach a minimum accuracy of 70% are optimized through hyperparameter tuning or replaced with better-performing models. After the best-performing model

is selected, it is used to predict employee turnover, helping organizations identify potential risks and implement preventive measures. Finally, presenting the model results and actionable insights supports the human resources department in formulating data-driven employee retention strategies, thereby enhancing overall organizational productivity. Additionally, Figure 1 illustrates the employee attrition purpose framework.

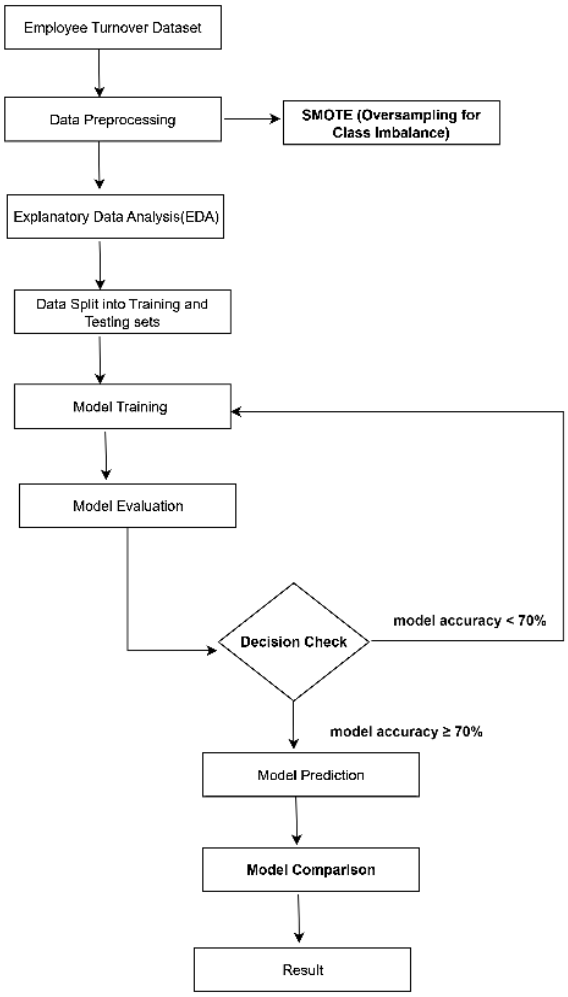

Figure 1. Purpose Framework

3.2 Dataset

This report utilizes the data obtained from Kaggle, which is a platform that allows users to access and share a wide range of datasets. This paper focuses on two datasets specifically related to employee mobility. The first dataset is named "Employee Performance and Productivity"[20], aiming to evaluate employee performance and productivity and provide valuable insights to support business decisions. It contains 100,000 entries, including 20 attributes: "Employee_ID" "Department", "Gender", "Age", "Job_Title", "Price", "Hire_Date", "Years_At_Company", "Education_Level","Performance_Score","Monthly_Salary","Work_Hours_Per_Week","Projects_Handled","Overtime_Hours",","Sick_Days","Remote_Work_Frequency","Team_Size","Training_Hours","Promotions_Employee"," Employee_Satisfaction_Score", and "Resigned". Table 2 shows the features of the Dataset 1.

Table 2. Features in Dataset 1 Without Pre-processing

No	Attributes	Explanation
1	Employee_ID	The unique identity of each employee member.
2	Department	The fields in which employees are employed (such as sales, Human Resources or Information Technology).
3	Gender	Sex of the Employee (male, female, other).
4	Age	The age of the employees between 22 and 60.
5	Job_Title	The positions held by the staff (manager, analyst, developer).
6	Hire_Date	The date on which the employee started working.
7	Years_At_Company	The length of time an employee has worked in the company.
8	Education_Level	Most advanced level of education attained (High School, Bachelor's, Master's).
9	Performance_Score	Performance evaluation of employees (scale of 1 to 5).
10	Monthly_Salary	The monthly salary of employees in USD is linked to their job title and performance rating.
11	Work_Hours_Per_Week	Weekly hours worked.
12	Projects_Handled	Overall count of projects managed by the employee.
13	Overtime_Hours	Total hours of overtime completed over the past year.
14	Sick_Days	Total sick days utilized by the employee.
15	Remote_Work_Frequency	Proportion of time spent working from home (0%, 25%, 50%, 75%, 100%)
16	Team_Size	Count of individuals in the employee's team.
17	Training_Hours	Total hours dedicated to training.
18	Promotions Employee	The number of promotions received during the period of employment.
19	Employee_Satisfaction_Score	Employee satisfaction score (on a scale from 1.0 to 5.0)
20	Resigned	A Boolean value that shows whether the employee has left the company.

However, the second dataset that consists of 1470 data and 34 columns, which are the following attributes:" Age","Attrition","BusinessTravel","DailyRate","Department","DistanceFromHome","Education","EducationField", "EmployeeCount","EmployeeNumber","EnvironmentSatisfaction","Gender","HourlyRate","JobInvolvement","JobLevel",,"JobSatisfaction","MaritalStatus","MonthlyIncome","MonthlyRate","NumCompaniesWorked","Over18","OverTime","PercentSalaryHike","PerformanceRating","RelationshipSatisfaction","StandardHours","StockOptionLevel","TotalWorkingYears","TrainingTimesLastYear","WorkLifeBalance","YearsAtCompany","YearsInCurrentRole"," YearsSinceLastPromotion ",and " YearsWithCurrManager ".Therefore, this dataset named as HR Employee Attrition [21]. Table 3 shows the features of the Dataset 2.

Table 3. Features in Dataset 2 Without Pre-processing

No	Attributes	Explanation
1	Age	Employee's age
2	Attrition	Show whether the employee has left the company.
3	BusinessTravel	Frequency of business trips
4	DailyRate	Daily wage rate
5	Department	Division where the employee is employed
6	DistanceFromHome	Distance from home to the workplace
7	Education	Education level (1 = Below College, 5 = Doctor)
8	EducationField	Area of study or educational background
9	EmployeeCount	Count of staff members (fixed amount)
10	EmployeeNumber	Distinct identification number given to every employee.
11	EnvironmentSatisfaction	Degree of contentment with the workplace atmosphere
12	Gender	Employee's gender
13	HourlyRate	Hourly pay rate
14	JobInvolvement	Degree of engagement in the work
15	JobLevel	Position or tier within the organization

16	JobSatisfaction	Contentment with the position
17	MaritalStatus	The employee's marital situation
18	MonthlyIncome	Monthly income
19	MonthlyRate	Monthly payment rate
20	NumCompaniesWorked	Count of organizations where the employee has held a position.
21	Over18	Specifies whether the employee is above 18 years old (fixed value).
22	OverTime	If the employee works additional hours.
23	PercentSalaryHike	Percentage rise in salary
24	PerformanceRating	Performance assessment scoring
25	RelationshipSatisfaction	Contentment with workplace connections
26	StandardHours	Regular working hours (fixed value)
27	StockOptionLevel	Amount of stock options awarded
28	TotalWorkingYears	Overall years of professional experience
29	TrainingTimesLastYear	Count of training sessions conducted over the last year.
30	WorkLifeBalance	Assessment of work-life balance
31	YearsAtCompany	Duration of employment at the organization
32	YearsInCurrentRole	Duration of time in the present position
33	YearsSinceLastPromotion	Time elapsed since the previous promotion
34	YearsWithCurrManager	Duration with the present manager

3.3 Data Preparation

This paper uses Python to handle missing values and duplicate columns. After loading the dataset into Jupyter Notebook, a thorough check was conducted on the missing values to confirm that there were no missing values, thereby eliminating the need to delete any rows. That is no missing value to handle in the dataset 1 as shown in Figure 2. Furthermore, the "Hire Date" column is converted to the date-time format to ensure the consistency of the datetime analysis. Therefore, in the dataset 1 it showed that unique values of the dataset 1 were also checked to better understand the distribution of data and identify any anomalies as shown in Figure 3.

Employee_ID	0
Department	0
Gender	0
Age	0
Job_Title	0
Hire_Date	0
Years_At_Company	0
Education_Level	0
Performance_Score	0
Monthly_Salary	0
Work_Hours_Per_Week	0
Projects_Handled	0
Overtime_Hours	0
Sick_Days	0
Remote_Work_Frequency	0
Team_Size	0
Training_Hours	0
Promotions	0
Employee_Satisfaction_Score	0
Resigned	0

Figure 2. Check Missing Values for Dataset 1

```
Unique values in Department: ['IT' 'Finance' 'Customer Support' 'Engineering' 'Marketing' 'HR'
 'Operations' 'Sales' 'Legal']
Unique values in Gender: ['Male' 'Female' 'Other']
Unique values in Job_Title: ['Specialist' 'Developer' 'Analyst' 'Manager' 'Technician' 'Engineer'
 'Consultant']
Unique values in Education_Level: ['High School' 'Bachelor' 'Master' 'PhD']
```

Figure 3. Check Unique Values for Dataset 1

In Dataset 2, the unique values of several key features—such as Age, DistanceFromHome, Education, JobLevel, and Work-Life Balance are presented in Figure 4. These values highlight the diversity and distribution within each variable. Additionally, the dataset contains no duplicate rows, as illustrated in Figure 5.

```
Unique values in Age: [41 49 37 33 27 32 59 30 38 36 35 29 31 34 28 22 53 24 21 42 44 46 39 43
 50 26 48 55 45 56 23 51 40 54 58 20 25 19 57 52 47 18 60]
Unique values in DistanceFromHome: [ 1  8  2  3 24 23 27 16 15 26 19 21  5 11  9  7  6 10  4 25 12 18 29 22
 14 20 28 17 13]
Unique values in Education: [2 1 4 3 5]
Unique values in JobLevel: [2 1 3 4 5]
Unique values in WorkLifeBalance: [1 3 2 4]
```

Figure 4. Check Unique Values for Dataset 2

```
Duplicate Rows:
Empty DataFrame
Columns: [Age, Attrition, BusinessTravel, DailyRate, Department, DistanceFromHome, Education, EducationField, EmployeeCount, EmployeeNumber, EnvironmentSatisfaction, Gender,
Index: []

[0 rows x 35 columns]
```

Figure 5. Dataset 2 Checking Duplicated Rows

3.4 Employee Attrition Prediction Models Classifications

In the prediction of employee's turnover, that used models such as DT, RF, NN, XG, and GB, and evaluated their performance through accuracy rate, precision, and F1 score. The F1 score combines accuracy and recall rates, providing a more balanced evaluation. The confusion matrix clearly illustrates the prediction results of the model, including classification scenarios such as true positives and false positives. The model's ability to distinguish categories can also be evaluated using the ROC curve and AUC value. The higher the AUC, the better the performance of the model. Furthermore, analysing the importance of characteristics helps identify the key factors influencing employee turnover. Comparing these indicators can help determine the model that is most suitable for predicting employee turnover.

3.4.1 RF

RF was chosen for employee turnover prediction due to its ability to capture complex relationships, strong resistance to noise, and improved predictive accuracy through the integration of multiple DTs. The model shows solid performance, especially in identifying employees who remain with the organization. However, the presence of many false negatives highlights its limitations in accurately detecting employees at risk of resignation—an essential aspect for developing effective retention strategies, as illustrated in Figure 6.

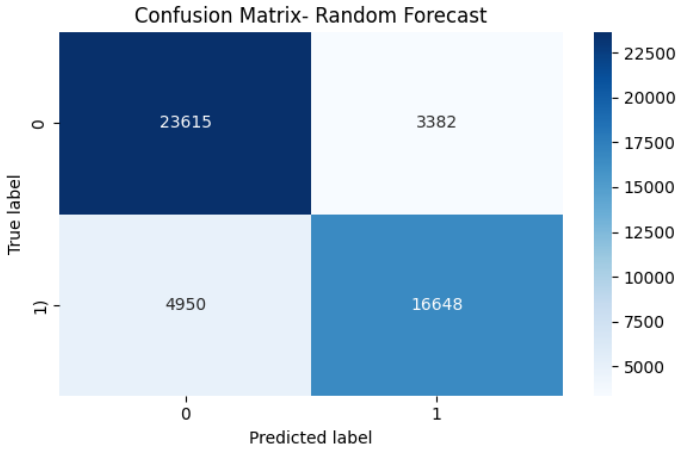

Figure 6. Confusion Matrix of RF Classifier in Dataset 1

According to the confusion matrix in this dataset 2 as shown in the below Figure 7, the model correctly detected 372 people who stayed, correctly predicted 372 of them, but predicted 8 people who were classified as wrongly leaving. This critical mass identified only 4 out of 61 employees who had but did not notice the accounts of 51 employees who had left and remained in the system.

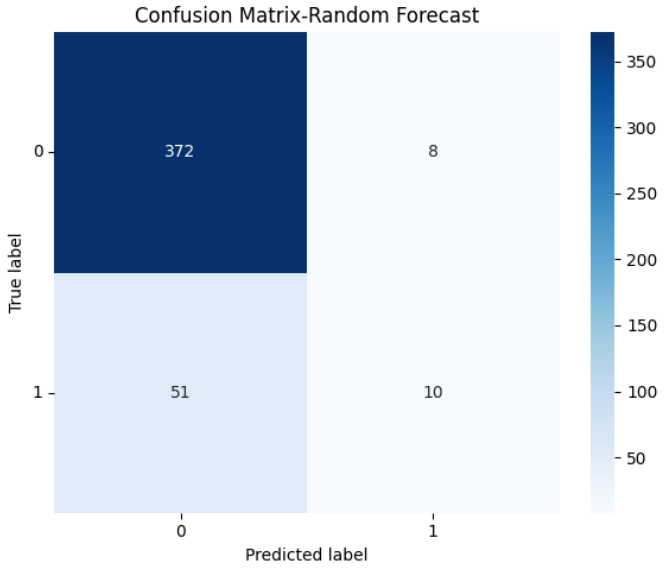

Figure 7. Confusion Matrix of RF Classifier in Dataset 2

3.4.2 GB

This method improves the prediction effect by gradually correcting the errors of the previous model and is applicable to problems of data imbalance such as employee turnover. GB can enhance the identification of a small number of departing employees and improve the sensitivity to rare events. This model can capture the complex nonlinear interactions among features and precisely locate the key factors influencing employee turnover, thereby enhancing the accuracy of predictions. This indicates that the model is highly effective in identifying employees who remain in the company and demonstrates a strong ability to predict departing employees, even if some at-risk employees are not accurately identified. Additionally, the results reflect a high level of classification accuracy with no false alarms and a relatively low number of missed turnover cases as shown in the Figure 8.

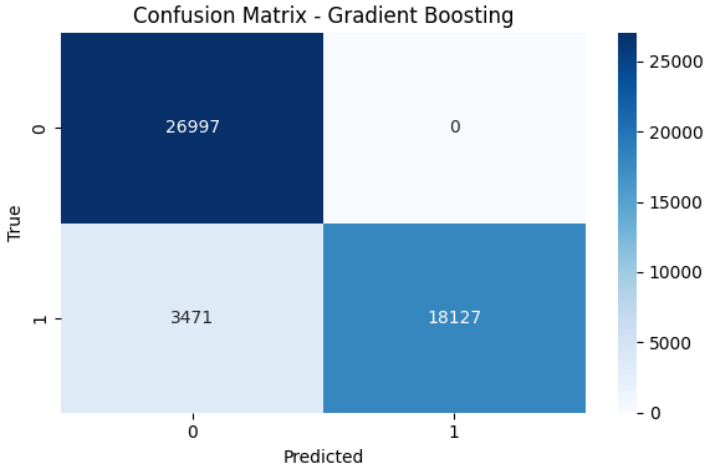

Figure 8. Confusion Matrix of GB Classifier in Dataset 1

The confusion matrix provides a detailed assessment of the model's classification capability. Out of the employees who remained, 369 were correctly classified as true negatives, whereas 11 were incorrectly predicted as false positives.

Regarding employees who resigned, the model successfully identified 18 true positives but failed to recognize 43 actual cases, resulting in false negatives. These errors suggest limitations in the model's ability to detect at-risk employees, which may weaken the effectiveness of retention-focused interventions, as illustrated in Figure 9.

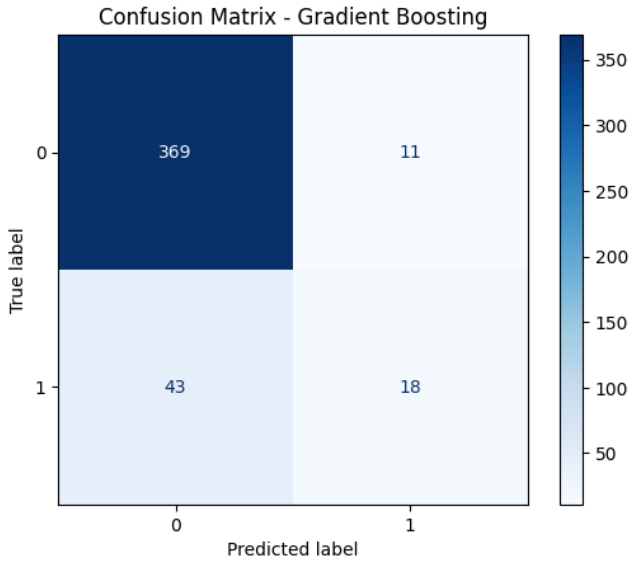

Figure 9. Confusion Matrix of GB Classifier in Dataset 2

3.4.3 DT

DTs are widely used in employee turnover prediction because of their simplicity, intuitiveness and ease of explanation. They can capture the most important factors affecting liquidity and address the correlation between missing values and complex features. To evaluate its predictive effect, a confusion matrix is needed when analysing "resignation" and "retention" to clarify the true positive, true negative, false positive and false negative of the model. In the event of data imbalance, integrating accuracy, recall rate and F1 score can conduct a comprehensive assessment of the model, helping the human resources department make more informed decisions, as shown in Figure 10 below.

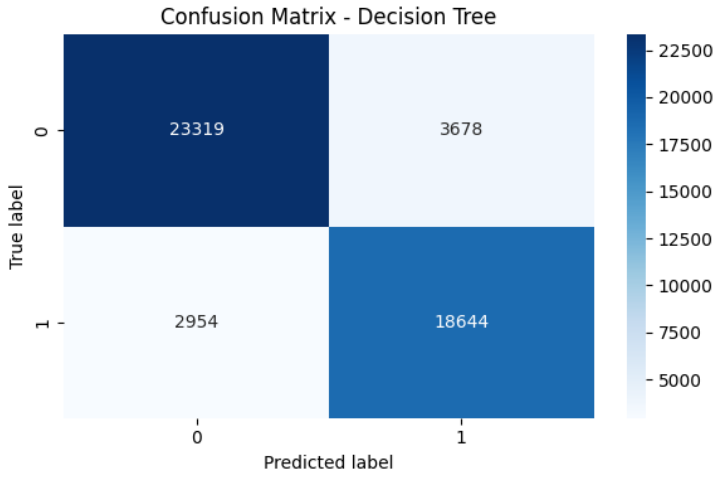

Figure 10. Confusion Matrix of DT Classifier in Dataset 1

The confusion matrix illustrates the performance of the DT model on a binary classification task. The model correctly classified 10 instances of category 1 and 362 instances of category 0. However, it misclassified 51 employees who belonged to category 1 as category 0 (false negatives) and 18 employees from category 0 as category 1 (false positives).

These results suggest that while the model performs well in identifying employees who are likely to stay, it struggles to accurately detect those who are at risk of leaving, as shown in Figure 11.

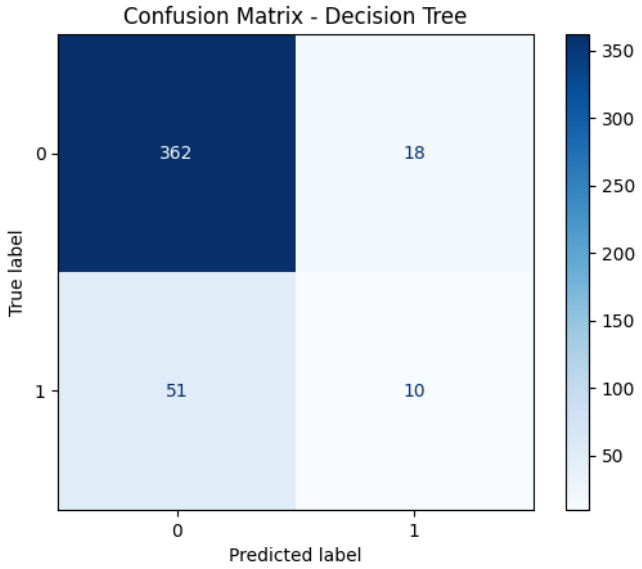

Figure 11. Confusion Matrix of DT Classifier in Dataset 2

3.4.4 XG Boost

This paper leverages the XGBoost technique due to its ability to efficiently manage complex nonlinear relationships and improve predictive performance by progressively improving weaker models. When addressing issues related to data imbalance, such as employee turnover, XGBoost demonstrates remarkable stability and is highly sensitive to complex details. The results derived from the weighted average indicate that the overall performance of the model is reliable. The confusion matrix results show that for the prediction of type 0 (employee retention), the model successfully identified 26,995 instances, with only 2 false positives. For the first category (employee turnover), it correctly predicted 18,448 instances, but there were 3,150 false negatives. This indicates that although the model shows a relatively high accuracy rate, there is still potential for improvement in the recall rate, as shown in Figure 12.

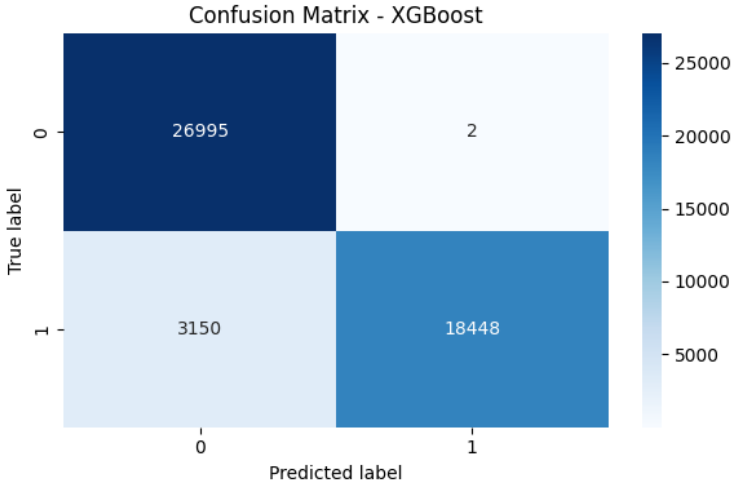

Figure 12. Confusion Matrix of XGB in Dataset 1

The confusion matrix illustrates the performance of the GB model. The model accurately classified 348 instances of class 0 as true negatives and 263 instances of class 1 as true positives. However, it incorrectly classified 22 instances of class 0 as class 1 (false positives) and 33 instances of class 1 as class 0 (false negatives), as shown in Figure 13.

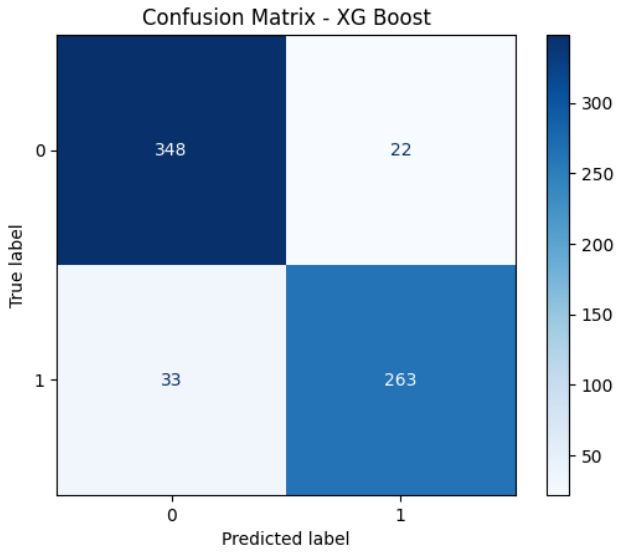

Figure 13. Confusion Matrix of XGB in Dataset 2

3.4.5 NN

NN learn by passing data sequentially through multiple layers of interconnected neurons. In this model, employee characteristics are fed into the input layer, transformed through nonlinear operations in the hidden layers, and finally used to generate predictions in the output layer. Due to their ability to capture complex, multi-dimensional relationships, NNs are well-suited for analysing factors that influence employee turnover. The model's performance in the binary classification task is evaluated using a confusion matrix. The results indicate that it correctly predicted 14,030 cases of category 0 (retention) and 10,779 cases of category 1 (resignation). However, it also produced 3,968 false positives (employees retained but misclassified as resigned) and 3,620 false negatives (employees who resigned but were misclassified as retained), as shown in Figure 14.

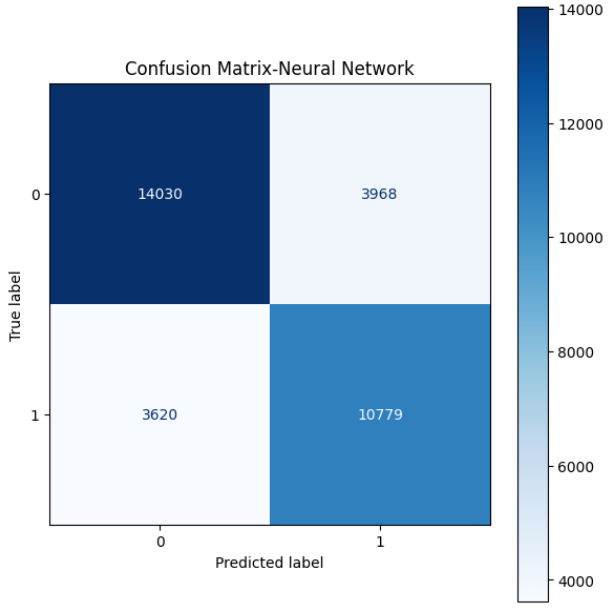

Figure 14. Confusion Matrix of NN Classifier in Dataset 1

The confusion matrix in Figure 15 summarizes the performance of the binary classification model. It correctly predicted 357 negative cases (true negatives) and 17 positive cases (true positives). However, the model also produced 23 false positives and failed to detect 44 actual positive cases (false negatives). This highlights that while the model performs well in identifying negative cases, it struggles with correctly detecting positive cases.

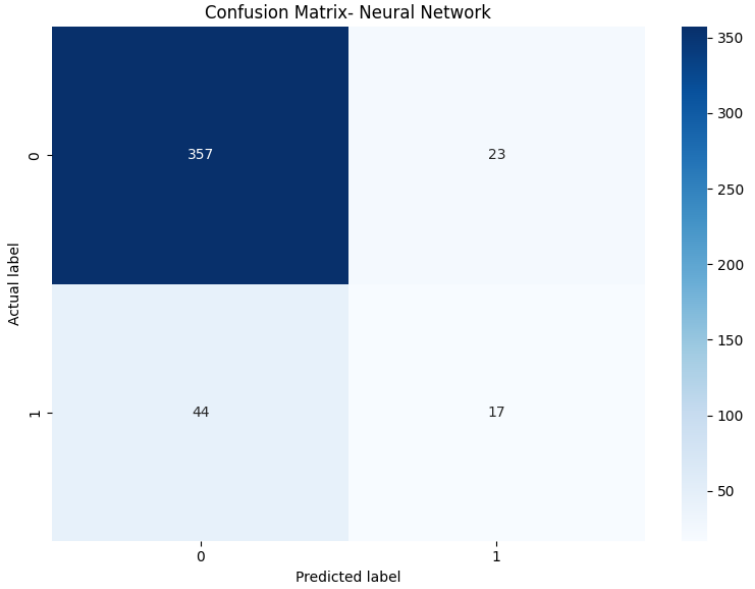

Figure 15. Confusion Matrix of NN Classifier in Dataset 2

4. RESULTS AND DICUSSION

4.1 Comparison of the Two Dataset

In this paper, two datasets were used to investigate five ML models aimed at predicting employee turnover. Since Dataset 1 are very large and most of the data belongs to one class, we use the SMOTE method to fix the imbalance. The company listed that approximately 90,000 people were still working, and 10,000 people had resigned. Dataset 2 is less than half the size of Dataset 1 because there are 1,200 remaining and 237 leaving. When calculating the accuracy rate, precision, recall rate and F1 score of the model, category 1 (resignation) is of the greatest importance. The comparison of model performance across both Dataset 1 and Dataset 2 reveals that XGBoost consistently outperforms other models, especially in predicting employee resignation (Class 1) in imbalanced datasets. In Dataset 1, which contains a large and heterogeneous sample, XGBoost achieved the highest accuracy of 0.94, with a precision of 1.00, recall of 0.85, and an F1 score of 0.92 for the resignation class. GB also performed well with an accuracy of 0.93 and an F1 score of 0.91, demonstrating the effectiveness of oversampling techniques combined with ensemble methods. In contrast, models like RFs, DTs, and NNs showed poor performance in identifying resignation cases, with low recall and F1 scores. A similar trend is observed in Dataset 2, where XGBoost again achieved the best results with an accuracy of 0.92, precision of 0.92, recall of 0.89, and an F1 score of 0.91. While GB showed a decent accuracy of 0.88, it struggled with resignation detection, with a recall of only 0.30 and F1 score of 0.40. Additionally, XGBoost consistently demonstrated superior performance in handling both large and small employee turnover datasets, outperforming other models including Linear Discriminant Analysis (LDA) in identifying resignation risks. Table 4 and 5 show the model performance comparison of Datasets 1 and 2.

Table 4. Model Performance Comparison – Dataset 1

Model	Accuracy	Precision (Class 1)	Recall (Class 1)	F1 score (Class 1)
RF	0.83	0.83	0.77	0.80
GD	0.93	1.00	0.84	0.91
DT	0.86	0.84	0.86	0.85
XG BOOST	0.94	1.00	0.85	0.92
NN	0.77	0.73	0.75	0.74

Table 5. Model Performance Comparison – Dataset 2

Model	Accuracy	Precision (Class 1)	Recall (Class 1)	F1 score (Class 1)
RF	0.87	0.56	0.16	0.25
GD	0.88	0.62	0.30	0.40
DT	0.84	0.36	0.16	0.22
XG BOOST	0.92	0.92	0.89	0.91
NN	0.85	0.42	0.28	0.34

4.2 K-fold Cross-validation

The 10-Fold Cross-Validation results for Dataset 1 and Dataset 2 show that ensemble methods generally outperform single-tree and NN models. In Dataset 1, RF achieved the highest accuracy (0.900 ± 0.000), followed by GB (0.863 ± 0.006), while XGBoost (0.653 ± 0.006) and NN (0.643 ± 0.026) performed worse, likely due to class imbalance, small dataset size, or unoptimized hyperparameters. For Dataset 2, GB (0.856 ± 0.017) and RF (0.852 ± 0.012) again demonstrated strong performance, with XGBoost (0.846 ± 0.017) now achieving higher accuracy, suggesting that Dataset 2 may have a more balanced distribution or features better suited to XGBoost. DT and NN models showed moderate performance in both datasets, indicating that single-tree models and default NN configurations may struggle with dataset complexity. The use of 10-Fold Cross-Validation provides a more reliable estimate of model performance by reducing variance and ensuring each data point is used for both training and validation, helping to identify models that generalize well to unseen data. Additionally, the results highlight the effectiveness of ensemble methods and the importance of tuning models like XGBoost and NNs for improved performance.

Table 6 and 7 show the 10-fold cross-validation of Datasets 1 and 2.

Table 6. 10-Fold Cross-Validation Accuracy of Dataset 1

Model	Accuracy (Mean ± Std)
RF	0.900 ± 0.000
DT	0.793 ± 0.002
GB	0.863 ± 0.006
XGB	0.653 ± 0.006
NN	0.643 ± 0.026

Table 7. 10-Fold Cross-Validation Accuracy of Dataset 2

Model	Accuracy (Mean ± Std)
RF	0.786 ± 0.024
DT	0.852 ± 0.012
GB	0.856 ± 0.017
XGB	0.846 ± 0.017
NN	0.767 ± 0.070

4.3 Paired T-test Results

Table 7 presents the paired t-test results comparing model accuracies across folds for Dataset 1 and Dataset 2. For Dataset 1, nearly all comparisons are statistically significant ($p < 0.05$), indicating that the differences in accuracy between models are unlikely to be due to random variation. Ensemble methods, such as RF and GB (GB), consistently outperform single-tree models, including DT, as well as NNs (NN), whereas XGBoost (XGB) exhibits performance comparable to that of the NN ($p = 0.2627$, not significant). In Dataset 2, the pattern differs slightly: DT demonstrates significant differences from RF, GB, and XGB, but not from the NN ($p = 0.4874$), while RF, GB, and XGB do not differ significantly among themselves, indicating comparable performance. Certain comparisons, such as RF versus NN, GB versus NN, and XGB versus NN, remain statistically significant, suggesting that NN accuracy is lower than that of specific ensemble models. In addition, these results indicate that ensemble methods generally achieve higher

and more consistent accuracies, with performance distinctions being more pronounced in Dataset 1 and less marked in Dataset 2. Table 8 shows the paired t-test results.

Table 8: Paired t-test Results (Accuracy Across Folds)

Comparison	Dataset 1 (p-value / Significance)	Dataset 2 (p-value / Significance)
DT vs RF	Significant	Significant
DT vs GB	Significant	Significant
DT vs XGBoost	Significant	Significant
DT vs NN	Significant	0.4874 (Not Significant)
RF vs GB	Significant	0.2967 (Not Significant)
RF vs XGBoost	Significant	0.3023 (Not Significant)
RF vs NN	Significant	Significant
GB vs XGBoost	Significant	0.0665 (Not Significant)
GB vs NN	Significant	Significant
XGBoost vs NN	0.2627 (Not Significant)	Significant

4.4 Comparison within Past Studies

However, among the five articles that have conducted previous studies in the existing literature review, as shown in Table 5, each article used different datasets and had different authors. By comparing the accuracy of employee turnover prediction in this paper with different models used by previous scholars, the model adopted in this paper has certain advantages in terms of accuracy. The models with the highest accuracy in predicted employee turnover rate used by the author are the LR and RF models. Additionally, this paper evaluates the performance of five ML models in employee departure prediction by using two datasets of different sizes and class distributions. It was found that XGBoost showed the best performance in both datasets, especially in identifying departing employees (category 1). The F1 scores reached 0.92 (dataset 1) and 0.91 (dataset 2), which significantly outperformed the models such as RF, DT, and NN. Compared to previous studies, [9] achieved 92.55% accuracy for HR dataset using RF model and [18] achieved 87.71% accuracy using LR model. In addition, [6] achieved 90.2% accuracy using the RF model. Table 9 shows the model performance comparison of the past studies.

Table 9. Model Performance Comparison of the Past Studies

Reference	Dataset	Technique/ Algorithm	Result
[9]	HR Employee Attrition	LGBM Classifier	91.25
		KNeighbors Classifier	84.77
		RF	92.55
		GB	89.63
[22]	HR dataset of employees of organization in Kazakhstan with information actual for 2023 was used for the prediction of Attrition.	LR	75
		DT	67
		SVM	73
		K-Nearest Neighbors	70
[18]	IBM HR Analytics Employee Attrition & Performance	LR	87.71
		KNN Classifier	59.22
		SVM	86.59
		NB	83.24
		DT	80.45
		RF	83.24
[23]	This study utilized data from the 2019 Graduates Occupation Mobility Survey (GOMS) conducted by the Korea employment information service, comprising 18,163 samples	XGB	78.5
		LR	78.3
		KNN	76.1
[6]	IBM Watson Analytics1	GB	87.5
		SVM	84.29
		LR	86.41
		KNN	84.23
		NB	80.70
		RF	90.20

4.5 Supporting Results for Key Contributions

In addition, compared with previous studies, this paper demonstrates significant innovations and improvements in several aspects. Unlike prior studies that mostly relied on a single dataset, limiting generalization, this paper employs two datasets with different scales and category distributions, each representing distinct practical scenarios, thereby enhancing the adaptability and stability of model results. The 10-Fold Cross-Validation results and paired t-test analyses provide strong support for the key contributions of this study. In Dataset 1, RF achieved the highest accuracy (0.900 ± 0.000), followed by GB (GB, 0.863 ± 0.006), while XGBoost (XGB, 0.653 ± 0.006) and NN (NN, 0.643 ± 0.026) performed less effectively. Paired t-test results indicate that almost all model comparisons in Dataset 1 are statistically significant, confirming that ensemble methods such as RF and GB significantly outperform single-tree and NN models.

For Dataset 2, GB (0.856 ± 0.017) and RF (0.852 ± 0.012) showed strong performance, with XGBoost (0.846 ± 0.017) achieving comparable accuracy, while DT (0.786 ± 0.024) and NN (0.767 ± 0.070) were moderately effective. Several paired t-test comparisons in Dataset 2 were not significant among top ensemble models, indicating comparable performance, whereas differences between ensemble models and NN remained significant. To address the common category imbalance in employee turnover prediction, this paper introduced the SMOTE data augmentation technique, which significantly improved the recall rate and F1 score for identifying the minority category, mitigating the insufficient minority identification observed in previous studies.

In terms of model evaluation, a clear performance threshold was set to replace models with accuracy below 70%, ensuring that only reliable models were retained for final comparisons, representing a dynamic optimization strategy superior to traditional fixed-model approaches. By integrating SMOTE with ensemble learning models, such as XGBoost and GB, the proposed approach achieved up to 94% accuracy and an F1 score of 0.92, surpassing most reported results in the literature. Overall, this study not only improves model accuracy but also emphasizes practical value: the results can serve as a predictive tool to assist HR departments in establishing early warning mechanisms and personalized intervention strategies to reduce organizational turnover, while providing a comprehensive analytical framework, including data preprocessing, model selection, performance evaluation, and cross-dataset validation, for subsequent related research.

5. PREDICTIVE ACCUCARY AND BUSINESS IMPLICATIONS

In conclusion, this study demonstrates that XGBoost and GB are highly effective in predicting employee turnover, particularly in identifying at-risk employees (Class 1). In Dataset 1, XGBoost achieved the highest performance with 94% accuracy and an F1 score of 0.92, while GB and RF also showed strong results, though with slightly lower recall for resignations. In Dataset 2, XGBoost again outperformed other models, achieving 92% accuracy and an F1 score of 0.91, indicating its robustness in handling class imbalance. Paired t-test analyses further confirm that ensemble methods generally outperform single-tree and NN models, although differences among top-performing ensembles are sometimes not statistically significant, reflecting comparable performance. From an organizational perspective, high recall rates for resignations enable HR departments to proactively implement retention strategies, such as performance discussions, career development support, or compensation adjustments, reducing turnover costs and maintaining workforce stability. These results align with previous research on ensemble learning, validating the use of XGBoost and GB as reliable decision-support tools in HR analytics. By integrating robust predictive models with techniques like SMOTE for minority class augmentation, this study not only improves predictive accuracy but also provides a practical, empirically validated framework for workforce planning and employee retention across datasets with varying distributions.

6. MODEL LIMITATION AND FUTURE WORK

In addition, several problems have still been exposed during the construction process. One of the main problems is that the category distribution of Dataset 1 is extremely uneven. Even with the adoption of data balancing techniques, this imbalance still had a certain impact on the generalization ability of the model. Furthermore, judging from the results of Dataset 2, the prediction accuracy of most models is not ideal, which indicates that in the case of limited information, it is quite challenging to identify employees who may resign. Since models mainly rely on historical data, they often fail to identify the latest developments within an organization or among employees. If variables such as employee engagement, performance indicators and real-time feedback can be further integrated, the prediction of

employee turnover risk in the future may be more accurate. Furthermore, the introduction of more complex analytical methods such as ensemble learning and interpretive tools such as SHapley Additive exPlanations (SHAP) or Local Interpretable Model-agnostic Explanation (LIME) will help enhance the understanding of the model's prediction mechanism, thereby more effectively supporting enterprises in their application in recruitment and human resource decision-making. In this paper, ML models have addressed the issues of class imbalance and data constraints, while also emphasizing the need for richer, real-time employee data and more interpretable methods to improve staff turnover prediction.

7. CONCLUSION

The use of ML to predict employee turnover has demonstrated significant impact on human resource management by enabling data-driven decision-making. This study employed multiple algorithms on two datasets with differing scales and category distributions, enhancing the adaptability and robustness of the models. Data preprocessing, including handling missing values, encoding categorical variables, and addressing class imbalance with SMOTE, ensured reliable model performance. The 10-Fold Cross-Validation results and paired t-test analyses show that ensemble methods, particularly XGBoost and GB, consistently achieve higher and more stable accuracies than single-tree models and NNs.

For Dataset 1, RF achieved the highest accuracy (0.900 ± 0.000), followed by GB (0.863 ± 0.006), while XGBoost (0.653 ± 0.006) and NNs (0.643 ± 0.026) were less effective due to class imbalance. In Dataset 2, XGBoost (0.846 ± 0.017) performed comparably to GB (0.856 ± 0.017) and RF (0.852 ± 0.012), highlighting the influence of dataset characteristics. XGBoost demonstrated strong predictive capabilities for identifying employees at risk of resignation, achieving high recall and F1 scores. This is crucial for HR managers to implement timely interventions such as job adjustments, performance discussions, or compensation reviews, thereby reducing turnover, maintaining workforce stability, and preserving organizational knowledge.

Traditional models like RF, DT, and NNs showed lower performance in detecting departures, emphasizing the advantages of ensemble learning. Overall, this study demonstrates the practical value of integrating ML with HR strategies, recommending XGBoost and GB as effective tools for employee turnover prediction, particularly when combined with SMOTE for class imbalance. Additionally, the study provides a methodological framework for future research encompassing data preprocessing, model selection, performance evaluation, and cross-dataset validation to support informed HR decision-making and retention planning.

ACKNOWLEDGEMENT

The authors would like to thank the anonymous reviewers for their suggestions to improve the paper.

FUNDING STATEMENT

The authors received no funding from any party for the research and publication of this article.

AUTHOR CONTRIBUTIONS

Low Kai Jia: Data Preparation, Modelling, Validation, Writing – Original Draft Preparation;
Lew Sook Ling: Conceptualization, Supervision, Review – Editing;
Sri Winarno: Conceptualization, Review – Editing.

CONFLICT OF INTERESTS

No conflicts of interest were disclosed.

ETHICS STATEMENTS

Our publication ethics follow The Committee of Publication Ethics (COPE) guideline. https://publicationethics.org/

DATA AVAILABILITY

The data that support the findings of this study are openly available in the Kaggle repository:

1. Employee Performance and Productivity Data. Available at: https://www.kaggle.com/datasets/mexwell/employee-performance-and-productivity-data Reference number: [20]
2. IBM HR Analytics Employee Attrition & Performance. Available at: https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset Reference number: [21]

REFERENCES

[1] Z. Taner, O. Hiziroglu, and A. Hiziroglu, "Leveraging machine learning methods for predicting employee turnover within the framework of human resources analytics", *Journal of Intelligent Systems Theory and Applications*, vol. 7, no. 2, pp. 145-158, 2024, doi: 10.38016/jista.1440879.

[2] W. A. Al-Suraihi, S. A. Samikon, A.-H. A. Al-Suraihi, and I. Ibrahim, "Employee turnover: causes, importance and retention strategies", *European Journal of Business Management and Research*, vol. 6, no. 3, pp. 1-10, Jun. 2021, doi: 10.24018/ejbmr.2021.6.3.893.

[3] M. A. Akasheh, E. F. Malik, O. Hujran, and N. Zaki, "A decade of research on machine learning techniques for predicting employee turnover: A systematic literature review", *Expert Systems with Applications*, vol. 238, pp. 121794, Mar. 2024, doi: 10.1016/j.eswa.2023.121794.

[4] R. Costa, and P. Carvalho, "Machine learning predictive models for preventing employee turnover costs", in *2022 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*, Maldives, Maldives: IEEE, pp. 1–6, Nov. 2022, doi: 10.1109/ICECCME55909.2022.9987976.

[5] E. M. T. A. Alsaadi, S. F. Khlebus, and A. Alabaichi, "Identification of human resource analytics using machine learning algorithms", *Telecommunication Computing Electronics and Control (Telkomnika)*, vol. 20, no. 5, pp. 1004, Oct. 2022, doi: 10.12928/telkomnika.v20i5.21818.

[6] R. Chakraborty, K. Mridha, R. N. Shaw, and A. Ghosh, "Study and prediction analysis of the employee turnover using machine learning approaches", in *2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON)*, Kuala Lumpur, Malaysia: IEEE, pp. 1–6, Sep. 2021, doi: 10.1109/GUCON50781.2021.9573759.

[7] D. Avrahami, D. Pessach, G. Singer, and H. Chalutz Ben-Gal, "A human resources analytics and machine-learning examination of turnover: Implications for theory and practice", *International Journal of Manpower*, vol. 43, no. 6, pp. 1405–1424, Aug. 2022, doi: 10.1108/IJM-12-2020-0548.

[8] F. Khan, "Data-driven strategies for predicting and preventing employee turnover", *Journal of Innovative Computing and Emerging Technologies*, vol. 4, no. 2, Oct. 2024, doi: 10.56536/jicet.v4i2.133.

[9] F. Maloku, "Analyzing IBM HR data: Employee attrition and performance insights", *Journal of Engineering and Applied Sciences Technology*, pp. 1-10, 2024, doi:10.47363/jeast/2024(6)268.

[10] A. 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, pp. 6424, Jun. 2022, doi: 10.3390/app12136424.

[11] K. Adeusi, P. Amajuoyi, and L. Benjami, "Utilizing machine learning to predict employee turnover in high-stress sectors", *International Journal of Management & Entrepreneurship Research*, vol. 6, no. 5, pp. 1702-1732, 2024, doi: 10.51594/ijmer.v6i5.1143.

[12] D. K. Sardar, S. Chourasiya, and V. Vijyalakshmi, "Employee turnover prediction by machine learning techniques", in *2023 International Conference on Circuit Power and Computing Technologies (ICCPCT)*, Kollam, India: IEEE, pp. 265–272, Aug. 2023, doi: 10.1109/ICCPCT58313.2023.10244896.

[13] R. Chakraborty, K. Mridha, R. N. Shaw, and A. Ghosh, "Study and prediction analysis of the employee turnover using machine learning approaches', in *2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON)*, Kuala Lumpur, Malaysia: IEEE, pp. 1–6, Sep. 2021, doi: 10.1109/GUCON50781.2021.9573759.

[14] J. Yuan, "Research on employee turnover prediction based on machine learning algorithms", in *2021 4th International Conference on Artificial Intelligence and Big Data (ICAIBD)*, Chengdu, China: IEEE, pp. 114–120, May 2021, doi: 10.1109/ICAIBD51990.2021.9459098.

[15] 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)*, Fez, Morocco: IEEE, pp. 1–5, Oct. 2021, doi: 10.1109/ICDS53782.2021.9626761.

[16] C. Zhang, and W. Han, "Ensembles of decision trees and gradient-based learning for employee turnover rate prediction", *PeerJ Computer Science*, vol. 10, p. e2387, Oct. 2024, doi: 10.7717/peerj-cs.2387.

[17] A. K. Biswas, R. Seethalakshmi, P. Mariappan, and D. Bhattacharjee, "An ensemble learning model for predicting the intention to quit among employees using classification algorithms", *Decision Analytics Journal*, vol. 9, pp. 100335, Dec. 2023, doi: 10.1016/j.dajour.2023.100335.

[18] P. Kumar, S. B. Gaikwad, S. T. Ramya, T. Tiwari, M. Tiwari, and B. Kumar, "Predicting employee turnover: A systematic machine learning approach for resource conservation and workforce stability", in *RAiSE-2023*, *MDPI*, Dec. 2023, p. 117. doi: 10.3390/engproc2023059117.

[19] N. R. Romaiha, R. Othman, N. E. Alias, S. A. N. Mizi, N. H. Mohamad Roseli, and Z. H. Abdul Karim, "Employees' turnover intention in Malaysian manufacturing company", *Information Management and Business Review*, vol. 15, no. 4(SI)I, pp. 258–263, Nov. 2023, doi: 10.22610/imbr.v15i4(SI)I.3599.

[20] "Employee performance and productivity data," *Kaggle*, 2024. [Online]. Available: https://www.kaggle.com/datasets/mexwell/employee-performance-and-productivity-data (accessed Jul. 5, 2025).

[21] "IBM HR analytics employee attrition & performance". [Online]. Available: https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset (accessed Jul. 5, 2025).

[22] 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, pp. 1–6, Jun. 2023, doi: 10.1109/ICECCO58239.2023.10146602.

[23] J. Park, Y. Feng, and S.-P. Jeong, "Developing an advanced prediction model for new employee turnover intention utilizing machine learning techniques", Sci Rep, vol. 14, no. 1, p. 1221, Jan. 2024, doi: 10.1038/s41598-023-50593-4. J. Park, Y. Feng, and S. Jeong, "Developing an advanced prediction model for new employee turnover intention utilizing machine learning techniques", *Scientific Reports*, vol. 14, no. 1, 2024, doi: 10.1038/s41598-023-50593-4.

BIOGRAPHIES OF AUTHORS

	**Low Kai Jia** received the bachelor's degree in information technology, specializing in business intelligence and analytics, from Multimedia University. Her research interests include employee turnover, machine learning method, and predictive analytics. She can be contacted at email: 1211200365@student.mmu.edu.my.
	**Lew Sook Ling** is an Associate Professor at the Faculty of Information Science and Technology, Multimedia University (MMU), Malaysia. She has been with MMU since 2001 and received her Ph.D. in 2013. She is also a Senior Member of IEEE. Her current research interests include educational technology, business analytics, image processing, and machine learning. She actively contributes to academic development through publications and collaborative research in emerging technologies. She can be contacted at email: sllew@mmu.edu.my.
	**Sri Winarno** is an Associate Professor at the Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang, Indonesia. He has been with Universitas Dian Nuswantoro since 1998 and received his Ph.D. in 2019. His current research interests include Educational Technology, Big Data analytics, Internet of Things (IoT), Image Processing, and Machine Learning. He is currently active in developing the expertise area of Intelligent Distributed Surveillance and Security through publications and collaboration research. He can be contacted at email: sri.winarno@dsn.dinus.ac.id.