Online-Journals.org

iTDAF IETI Transactions on Data Analysis and Forecasting

iTDAF | eISSN: 2959-0442 | Vol. 4 No. 1 (2026) | OPEN ACCESS

https://doi.org/10.3991/itdaf.v4i1.59233

PAPER

Enhancing Retention Strategies through Deep Learning-Based Dropout Prediction

Ibrar Hussain[1], Sidra Tahir[1] (✉), Asif Nawaz[1], Kashif Mehmood[1], Ahthasham Sajid[2], Sabitha Banu[3]

[1]PMAS-Arid Agriculture University, Rawalpindi, Pakistan

[2]Air University, Islamabad, Pakistan

[3]PSGR Krishnammal College for Women, Coimbatore, India

stahir@uaar.edu.pk

ABSTRACT

Employees are an organization's most significant resource. Employee dropout may be costly for firms owing to the costs of recruiting, training, and lost productivity. By forecasting dropout, firms may take preventative actions such as developing retention programs, providing targeted assistance to at-risk employees, and addressing possible workplace concerns. The unpredictable dropout can assist in reducing dropout rates and saving money. Existing approaches to predicting employee dropout use machine learning (ML) techniques for employee dropout prediction, which do not present the correlation of various employee attributes that may have caused the dropout. Moreover, the imbalanced dataset affects the accuracy of prediction results. In this paper, Synthetic Minority Oversampling Technique (SMOTE) is applied to the dataset to solve the issue of imbalanced data. Following that, a deep learning technique, gated recurrent unit (GRU), is utilized to predict staff dropout effectively. It also aided in determining most of the relevant factors of employee results. For this purpose, the IBM employee dataset is utilized for training and assessing GRU using 10-fold test-train splitting. The ultimate objective is to effectively detect dropouts to assist any organization in improving various retention strategies. According to the results, the suggested technique achieves 95% accuracy, more significant than existing state-of-the-art approaches.

KEYWORDS

employee dropout, deep learning, synthetic minority oversampling technique (SMOTE), gated recurrent unit (GRU), prediction

1 INTRODUCTION

Employee dropout denotes the voluntary retreat of employees of a firm or a company. It is evaluated by various indicators such as return on equity, profitability, sales growth, and customer service quality to assess its impact on organizational success. Research has shown that high staff dropout rates hurt organizational efficiency, making it a significant concern [1, 2], Managing staff turnover is crucial for businesses as it leads to financial losses, a decline in the organization's knowledge base,

Hussain, I., Tahir, S., Nawaz, A., Mehmood, K., Sajid, A., Banu, S. (2026). Enhancing Retention Strategies through Deep Learning-Based Dropout Prediction. *IETI Transactions on Data Analysis and Forecasting (iTDAF)*, 4(1), pp. 35–55. https://doi.org/10.3991/itdaf.v4i1.59233

Article submitted 2025-10-16. Revision uploaded 2025-12-22. Final acceptance 2026-02-10.

and decreased employee engagement [3–5]. Dropout also has an impact on productivity and meeting organizational goals on time. It encompasses employee resignations, deaths, and retirements [6] and is an essential aspect of workforce planning to ensure the availability of suitable employees when needed.

Organizations and their management continuously seek ways to evaluate and improve human resource (HR) management operations. Traditional approaches like surveys and experimental studies have limitations, including reliance on self-reports, resource-intensive requirements, and the lack of real-time evaluations [7]. Deep learning, particularly predictive analytics, offers the potential to predict employee turnover by identifying patterns and forecasting future events [3, 8]. Deep learning employs artificial intelligence (AI) techniques to extract valuable information from data at various levels of representation and abstraction [9, 10]. By leveraging predictive analytics, companies can gain insights into real-time behaviors based on online data, enabling them to forecast and address employee dropout [11, 12]. Employee Dropout Prediction models assist HR professionals in estimating employee departures, allowing them to focus on retaining valuable resources.

The focus of this study work is on using and investigating a gated recurrent unit (GRU) for predicting employee dropout. The GRU model effectively utilizes relevant parameters to accelerate prediction [13, 14]. It leverages current neural networks and is a gating mechanism similar to long short-term memory (LSTM) models but with rare parameters [15–20]. The GRU model considers various input variables, including wage structure, work-life balance, job satisfaction, and relationships with supervisors, as indicators of potential dropout. By providing warnings of dropouts, this prediction framework aids management in making informed decisions and taking appropriate actions.

This study also analyzed HR data using deep learning and predictive analytic techniques. The data was obtained from an online repository and underwent cleaning, preparation, and preprocessing using various data exploration techniques. A data-balancing technique was applied to overcome the class imbalance in the class of a dataset. The GRU prediction model was applied to the preprocessed data, and its performance was enhanced through 10-fold cross-validation. The results demonstrated that the deep learning model performed exceptionally well on the HR dataset, surpassing previous prediction algorithms with an accuracy of 95%.

A) Research contributions

The contributions of the study can be summarized as follow:

- Conducting a comprehensive literature review to identify existing deep learning and machine learning (ML) techniques for employee dropout prediction, along with the analysis of relevant datasets.
- Using a deep learning-based GRU algorithm as a solution to problem.
- Performing experiments that compared the proposed model with state-of-the-art approaches using a 10-fold test-train split. The results showcased the superior performance of the GRU-based predictive analysis model, achieving an impressive accuracy of 95%.

Paper structured as follows: Section 2 presents an outline of the related studies. Section 3 describes the research methods employed. Section 4 explains the results and discussions, and finally, Section 5 concludes the study and future work.

2 RELATED WORK

This section reconnoiters the literature relevant to the proposed study investigation. The related literature is based on an examination of prior approaches and research findings for forecasting employee dropout. The most recently applied state-of-the-art techniques were chosen for the literature review.

Shawni [21] investigated the neural network cross-validation method for predicting employee turnover. To estimate employee dropout on a single platform, they proposed using a neural network based on feed-forward using a 10-fold validation strategy. The Purpose of their study was to look into the feasibility of employing relevant criteria and the possibility of being affected by the dropout process. Their proposed method was evaluated and compared to existing classifiers. The experimental analysis found that the proposed method significantly outperforms the existing classifier regarding performance measures.

In their work, [22] presented the ML method for determining employee turnover. Their work evaluated the performance of different supervised ML algorithms on several HR datasets. In the context of statistical analysis, this work established and applied several data mining techniques, such as data expanding, parameter surfing, and validation. Feature significance ranking and classifier visualization were used to improve the interpretability of the employee dropout prediction model.

Jinquan discussed graph neural networks of heterogeneous nature for employee turnover forecasts [23]. During their study, they developed a data-driven deep neural network to model proof of worker turnover from an internal and an external perspective. They combined BiLSTM and survival analysis to predict employee turnover from inside and outside viewpoints and exploited the attention mechanism. Finally, they ran extensive tests to assess the effectiveness of their approach using large-scale real-world personnel data.

In another state of the art, Salah argued for using deep neural networks to anticipate employee turnover [24]. They evaluated the employee's dataset to see what elements attract the individual to leave the company. There was also information about the relationships between various qualities. Their study found that the most compelling aspects affecting employee decisions were extra hours, work level, and monthly salary. Using the dataset provided by IBM Analytics was problematic due to its inconsistency. The dataset's prediction accuracy was 94%.

Pekel [25] created a convolutional neural network (CNN) model for predicting employee dropout. Employee dropout was expected in their study of retail sales personnel using deep learning algorithms. Initially, they used ML approaches to predict employee dropout. The CNN was also used to increase the accuracy ratio. The projected models were fruitfully executed for a dataset of 1186 salespeople at a retail organization based on the original and real data sets. According to the findings, the proposed ECDT framework is an accurate predictor of employee turnover. By roughly 11%, the ECDT-GRID model surpassed Decision Trees, which generated effective results from basic classification algorithms.

In another paper, [26] discussed using deep learning data to provision employee analytics for employee retention prediction. First, they offered a deep data-driven methodology dependent on a hybrid technique to build a model for employee dropout in order to discover essential employee characteristics impacting their dropout. Their dropout prediction technique was based on intelligent learning

models and was tested on simulated big and average-sized HR datasets, as well as a genuine small-sized dataset with 450 answers. When compared to earlier methods, their methodology obtained remarkable accuracy for the three datasets. Finally, while prizes and compensation were often considered important factors in retention, their data showed a factor, 'business travel,' was the primary motivation for workers.

Fahad [27] presented a forecasting mechanism for the dropout of employees using an ensemble using ML approaches. They developed an automated strategy for predicting employee dropout that used numerous predictive analytical methodologies. To select the best model, the systems were combined with various pipeline topologies. Furthermore, an auto-tuning technique was employed to find the finest arrangement of multiple parameters for generating the champion models. Finally, they provided an ensemble strategy for choosing the most efficient model by considering a variety of evaluation metrics. The consequences of the proposed model demonstrated that no prototype could be deemed optimal or ideal for a business environment up to this moment.

The research used ML approaches such as SVM, Naive Bayes, and random forest (RF) classifiers to predict employee dropout [28]. These methods were utilized to forecast which employees would depart the company over the next two years, using a few categorization models. As a result, in order to examine the trend, they divided their inquiry into two cases. In the first scenario, they analyzed all Focused Variable classes, but in the second, they removed those individuals who still raised concerns about leaving an organization soon. Compared to the other models, the RF classifier was the most efficient model in the study, with the highest accuracy and recall value.

An improved RF algorithm was used for predicting employee turnover in research [29]. They utilized their suggested weighted quadratic random forest (WQRF) method on employee dropout data having high-dimensional with imbalanced features. At first, the RF method was employed to prioritize features and minimize dimensions. Second, the selected characteristics were combined with RF, and the F-score for every decision tree (DT) was computed as a weights to develop the employee turnover likelihood framework. The primary determinants impacting employee turnover were discovered in an experiment utilizing an employee dataset of communications business: monthly income, years at the company, gender, distance from home, overtime, age, and percentage of wage rise. Monthly salary and overtime were the two most critical criteria among them.

Fallucchi [30] used ML to predict employee dropout. Following training, the model was tested on a dataset of IBM Analytics, having 35 attributes and around 1500 samples. The results were presented in traditional metrics, with the Gaussian Nave Bayes classifier producing the superlative results for the supplied dataset. It had the maximum recall rate because it looks at a classifier's ability to detect all positive occurrences and achieves an overall false negative rate of 4.5%.

Golande [22] demonstrated an ML technique for explaining and forecasting employee turnover in their work. The primary reason for conducting their study was developing a model that can predict whether an employee would quit or not. The primary objective was to experiment with the validity of employee assessment and satisfaction scores in the company, which can contribute to reducing employees' dropout. In their study, they have applied various ML algorithms in an HR data set. From their findings, RF is concluded to be better than the other ML classifiers.

Table 1. A summary of several articles on dropout prediction based on their methodologies and findings

Study	Methods	Dataset Size	Performance Measures	Variables	Limitations
[32]	NB, DT, RF	16,649	Accuracy (97.5%)	12	Unbalanced and biased data were disregarded. The inference of relationships between employee attributes was not investigated.
[33]	ML models	D1 = 1470, D2 = 14,999	Accuracy (96%)	35	Both datasets lack an obvious distinction between classes. Extensive preparation.
[30]	NB	1500	Accuracy (82%)	35	Extensive pretreatment Unbalanced data is disregarded.
[24]	DNN	1500	Accuracy (85%)	35	The significance of factors influencing employee turnover is unrelated. The harmonizing of datasets is inefficient.
[34]	GBM	577	Accuracy (89%)	17	Experimental evaluations are limited to particular real-world datasets.
[35]	DNN	720	Accuracy (89%)	20	The features of local employees are not discussed.
[36]	ADASYN, SVM, RF, KNN	1575t	F1 score (50%)	25	A prediction model with diminished precision and accuracy. Manual under sampling was performed on the dataset.
[27]	MLP	1500	Accuracy (93%)	35	Prediction model implying lower values

In another research, [31] presented an employee dropout prediction framework with an ML technique. Their study suggested a three-stage dropout prediction paradigm (pre-processing, processing, and post-processing). The dataset had various features. During the pre-processing step, dimension reduction was advised to employ the "max-out" feature selection strategy. That method was tested using the IBM HR dataset. Finally, the model's parameters were validated by training it on many bootstrap datasets. The model's confidence rating and stability were then determined by calculating the average and standard deviation parameters. Table 1 lists some of the other fundamental techniques for forecasting employee dropout.

A summary of several articles on dropout prediction based on their methodologies and findings, the literature review shows that most of the methods for predicting employee dropout are based on ML approaches. They are pretty remarkable in accuracy and predict employee dropout efficiently; they fail to capture the long-term dependency, handle variable-length sequences, and efficient training, which is done using the recurrent nature of GRU. It uses imputation to auto-handle missing data to improve the accuracy of the proposed model. Therefore, there exists a need to provide a deep learning GRU-based model that efficiently overcomes the said limitations. The displayed model will provide an employee dropout predicted result.

Which leadership of companies easily identifies an employee leaving and takes necessary actions to stop them.

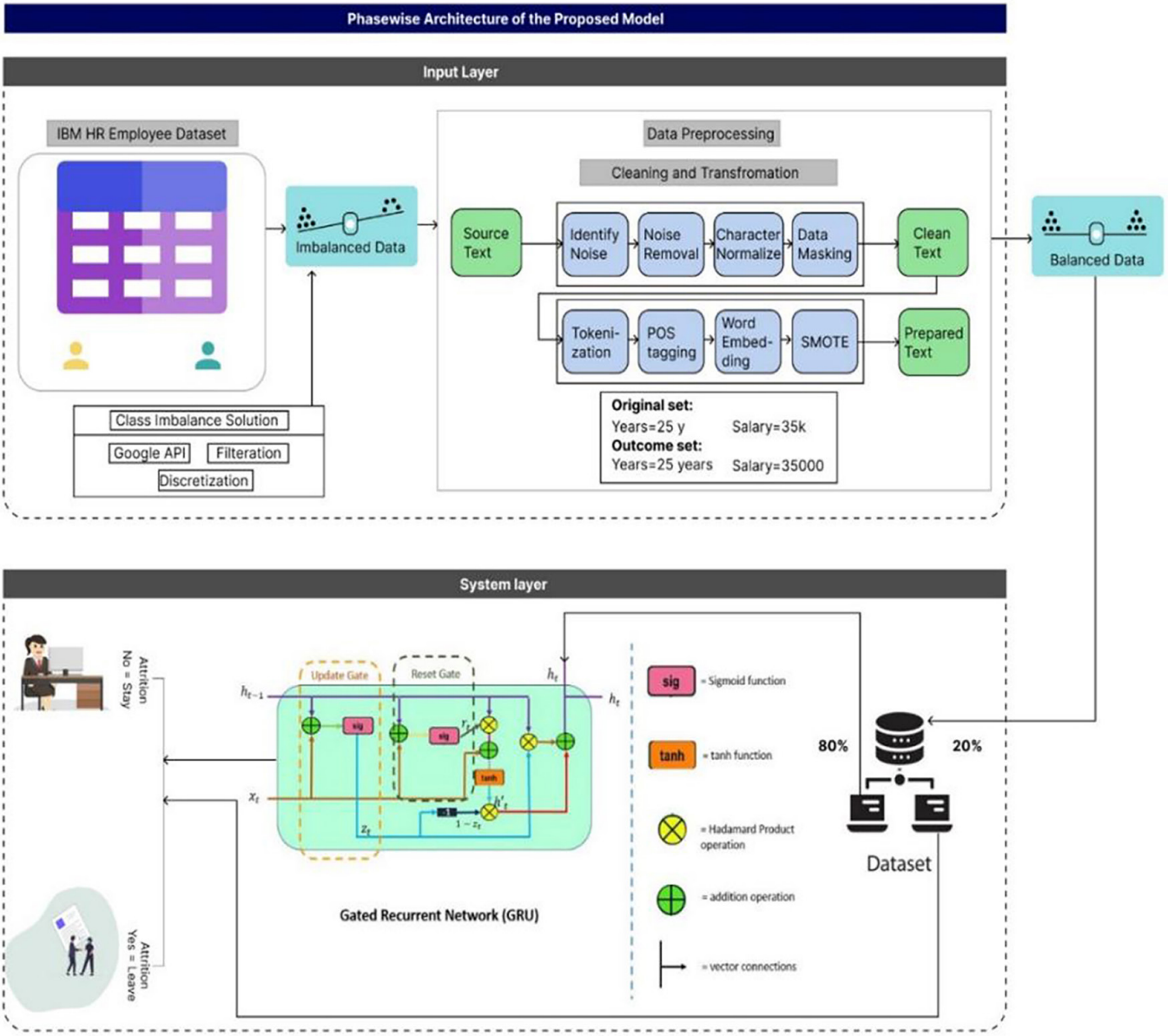

Fig. 1. GRU-model for employee dropout rate

3 PROPOSED METHODOLOGY

This study proposed an Synthetic Minority Oversampling Technique (SMOTE)-based hybrid deep learning GRU-based model to predict employee dropout in an organization. The critical steps of the proposed model have been depicted in Figure 1, as it is apparent that the vital stages of the proposed model comprise three layers: the input layer, the system layer, and the prediction layer. The input layer is primarily responsible for data collection, filtration, and preprocessing, which also handles the imbalanced nature of the data through the SMOTE [37, 38]. Whereas the system layer's work uses the GRU algorithm for analyzing and classifying test and training datasets, providing results to the prediction layer. The detailed description of each layer is explained in the below subsections.

3.1 Input layer

This is the first and most crucial layer of the proposed model. This layer is responsible for data collection, data filtration, and data preprocessing. Data preprocessing also has a sub-phase that is cleansing and transformation, where the collected data is pre-processed and balanced for better results.

Dataset collection. The IBM HR Analytics employee performance and dropout data (https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-dropout-dataset), which was published in Kaggle's Competitions, was used in this study. Figure 2 shows a snapshot of the initial details of the dataset. Figure 3 depicts the distribution of dropout factors in positive and negative scenarios. The IBM HR dataset has 1470 records and 35 factors, one of which is dependent (employee turnover) and the rest of which are independent. Personal qualities factors, job-related variables, and career-related variables are the three primary groups of independent variables. Because of the dataset's complexity, a feature reduction technique is necessary. This procedure is based on the filtration approach [39, 40], and the candidate feature list is given in Figure 3. The currently presented research uses the open IBM HR Analytics dataset, which already serves for many employee churn-related studies as a benchmark. Although this choice allows for a fair comparison with previously published results, relying on a single dataset may restrict the generalizability of findings. Organizational culture, HR policies, and workforce dynamics differ across industries and company sizes. Future research efforts, therefore, need to further this work by validating the proposed framework on multi-source enterprise datasets from diverse industrial domains.

Data balancing. The potential features, which had been cut down to 26 features, were used to examine the data for pre-processing. Preprocessing is often used in studies that try to predict employee turnover because datasets often have missing records, different noise levels, and significant scale changes per feature [41, 42]. Other data preparation methods were used to get the most valuable results.

Education	Education Field	EnvironmentSatisfaction	HourlyRate	JobRole	JobSatisfaction	MaritalStatus	MonthlyIncome	MonthlyRate	OverTime	PercentSalaryHike	PerformanceRating	RelationshipSatisfaction	StandardHours	TotalWorkingYears	TrainingTimesLastYear	WorkLifeBalance	YearsAtCompany	YearsInCurrentRole	YearsSinceLastPromotion	YearsWithCurrManager
2	Life Scien	2	94	Sales Exe	4	Single	5993	19479	Yes	11	3	1	80	8	0	1	6	4	0	5
1	Life Scien	3	61	Research	2	Married	5130	24907	No	23	4	4	80	10	3	3	10	7	1	7
2	Other	4	92	Laborator	3	Single	2090	2396	Yes	15	3	2	80	7	3	3	0	0	0	0
4	Life Scien	4	56	Research	3	Married	2909	23159	Yes	11	3	3	80	8	3	3	8	7	3	0
1	Medical	1	40	Laborator	2	Married	3468	16632	No	12	3	4	80	6	3	3	2	2	2	2
2	Life Scien	4	79	Laborator	4	Single	3068	11864	No	13	3	3	80	8	2	2	7	7	3	6
3	Medical	3	81	Laborator	1	Married	2670	9964	Yes	20	4	1	80	12	3	2	1	0	0	0
1	Life Scien	4	67	Laborator	3	Divorced	2693	13335	No	22	4	2	80	1	2	3	1	0	0	0
3	Life Scien	4	44	Manufact	3	Single	9526	8787	No	21	4	2	80	10	2	3	9	7	1	8
3	Medical	3	94	Healthcar	3	Married	5237	16577	No	13	3	2	80	17	3	2	7	7	7	7
3	Medical	1	84	Laborator	2	Married	2426	16479	No	13	3	3	80	6	5	3	5	4	0	3
2	Life Scien	4	49	Laborator	3	Single	4193	12682	Yes	12	3	4	80	10	3	3	9	5	0	8
1	Life Scien	1	31	Research	3	Divorced	2911	15170	No	17	3	4	80	5	1	2	5	2	4	3
2	Medical	2	93	Laborator	4	Divorced	2661	8758	No	11	3	3	80	3	2	3	2	2	1	2
3	Life Scien	3	50	Laborator	3	Single	2028	12947	Yes	14	3	2	80	6	4	3	4	2	0	3
4	Life Scien	2	51	Manufact	1	Divorced	9980	10195	No	11	3	3	80	10	1	3	10	9	8	8
2	Life Scien	1	80	Research	2	Divorced	3298	15053	Yes	12	3	4	80	7	5	2	6	2	0	5
2	Medical	4	96	Laborator	4	Divorced	2935	7324	Yes	13	3	2	80	1	2	2	1	0	0	0
4	Life Scien	1	78	Manager	4	Married	15427	22021	No	16	3	3	80	31	3	3	25	8	3	7
3	Life Scien	4	45	Research	4	Single	3944	4306	Yes	11	3	3	80	6	3	3	3	2	1	2
2	Other	1	96	Manufact	3	Divorced	4011	8232	No	18	3	4	80	5	5	2	4	2	1	3
4	Life Scien	3	82	Sales Rep	1	Single	3407	6986	No	23	4	2	80	10	4	3	5	3	0	3
4	Life Scien	1	53	Research	2	Single	11994	21293	No	11	3	3	80	13	4	3	12	6	2	11
2	Life Scien	3	96	Research	4	Single	1232	19281	No	14	3	4	80	0	6	3	0	0	0	0
1	Medical	2	83	Research	1	Single	2960	17102	No	11	3	3	80	8	2	3	4	2	1	3
3	Other	3	58	Manager	3	Divorced	19094	10735	No	11	3	4	80	26	3	2	14	13	4	8
1	Life Scien	2	72	Research	1	Single	3919	4681	Yes	22	4	2	80	10	5	3	10	2	6	7
4	Marketing	3	48	Sales Exe	2	Married	6825	21173	No	11	3	4	80	10	2	3	9	7	4	2
4	Medical	1	42	Healthcar	4	Married	10248	2094	No	14	3	4	80	24	4	3	22	6	5	17
4	Marketing	2	83	Manager	1	Single	18947	22822	No	12	3	4	80	22	2	2	2	2	2	1
3	Medical	3	78	Laborator	4	Single	2496	6670	No	11	3	4	80	7	3	3	1	1	0	0

Fig. 2. Dataset after filtration

Exploratory data analysis was employed to determine how the data was put together and its features. The measured numbers for "Standard Time," "Over 18,"

and “Employee Count” were all the same. “Employee Number,” an employee ID number, was taken out of the model variables. One-hot encoding and label encoding were used to code categorical values.

Education	Education Field	EnvironmentSatisfaction	HourlyRate	JobRole	JobSatisfaction	MaritalStatus	MonthlyIncome	MonthlyRate	OverTime	PercentSalaryHike	PerformanceRating	RelationshipSatisfaction	StandardHours	TotalWorkingYears	TrainingTimesLastYear	WorkLifeBalance	YearsAtCompany	YearsInCurrentRole	YearsSinceLastPromotion	YearsWithCurrManager
2	Life Scien	2	94	Sales Exe	4	Single	5993	19479	Yes	11	3	1	80	8	0	1	6	4	0	5
1	Life Scien	3	61	Research	2	Married	5130	24907	No	23	4	4	80	10	3	3	10	7	1	7
2	Other	4	92	Laborator	3	Single	2090	2396	Yes	15	3	2	80	7	3	3	0	0	0	0
4	Life Scien	4	56	Research	3	Married	2909	23159	Yes	11	3	3	80	8	3	3	8	7	3	0
1	Medical	1	40	Laborator	2	Married	3468	16632	No	12	3	4	80	6	3	3	2	2	2	2
2	Life Scien	4	79	Laborator	4	Single	3068	11864	No	13	3	3	80	8	2	2	7	7	3	6
3	Medical	3	81	Laborator	1	Married	2670	9964	Yes	20	4	1	80	12	3	2	1	0	0	0
1	Life Scien	4	67	Laborator	3	Divorced	2693	13335	No	22	4	2	80	1	2	3	1	0	0	0
3	Life Scien	4	44	Manufact	3	Single	9526	8787	No	21	4	2	80	10	2	3	9	7	1	8
3	Medical	3	94	Healthcar	3	Married	5237	16577	No	13	3	2	80	17	3	2	7	7	7	7
3	Medical	1	84	Laborator	2	Married	2426	16479	No	13	3	3	80	6	5	3	5	4	0	3
2	Life Scien	4	49	Laborator	3	Single	4193	12682	Yes	12	3	4	80	10	3	3	9	5	0	8
1	Life Scien	1	31	Research	3	Divorced	2911	15170	No	17	3	4	80	5	1	2	5	2	4	3
2	Medical	2	93	Laborator	4	Divorced	2661	8758	No	11	3	3	80	3	2	3	2	2	1	2
3	Life Scien	3	50	Laborator	3	Single	2028	12947	Yes	14	3	2	80	6	4	3	4	2	0	3
4	Life Scien	2	51	Manufact	1	Divorced	9980	10195	No	11	3	3	80	10	1	3	10	9	8	8
2	Life Scien	1	80	Research	2	Divorced	3298	15053	Yes	12	3	4	80	7	5	2	6	2	0	5
2	Medical	4	96	Laborator	4	Divorced	2935	7324	Yes	13	3	2	80	1	2	2	1	0	0	0
4	Life Scien	1	78	Manager	4	Married	15427	22021	No	16	3	3	80	31	3	3	25	8	3	7
3	Life Scien	4	45	Research	4	Single	3944	4306	Yes	11	3	3	80	6	3	3	3	2	1	2
2	Other	1	96	Manufact	3	Divorced	4011	8232	No	18	3	4	80	5	5	2	4	2	1	3
4	Life Scien	3	82	Sales Rep	1	Single	3407	6986	No	23	4	2	80	10	4	3	5	3	0	3
4	Life Scien	1	53	Research	2	Single	11994	21293	No	11	3	3	80	13	4	3	12	6	2	11
2	Life Scien	3	96	Research	4	Single	1232	19281	No	14	3	4	80	0	6	3	0	0	0	0
1	Medical	2	83	Research	1	Single	2960	17102	No	11	3	3	80	8	2	3	4	2	1	3
3	Other	3	58	Manager	3	Divorced	19094	10735	No	11	3	4	80	26	3	2	14	13	4	8
1	Life Scien	2	72	Research	1	Single	3919	4681	Yes	22	4	2	80	10	5	3	10	2	6	7
4	Marketing	3	48	Sales Exe	2	Married	6825	21173	No	11	3	4	80	10	2	3	9	7	4	2
4	Medical	1	42	Healthcar	4	Married	10248	2094	No	14	3	4	80	24	4	3	22	6	5	17
4	Marketing	2	83	Manager	1	Single	18947	22822	No	12	3	4	80	22	2	2	2	2	2	1
3	Medical	3	78	Laborator	4	Single	2496	6670	No	11	3	4	80	7	3	3	1	1	0	0

Fig. 3. Dataset after filtration

There were more employees who left compared to those who remained, so the SMOTE was employed to create new cases. SMOTE is a method used to create new instances by modifying existing data and employing a synthetic minority oversampling technique. The method of SMOTE identifies the nearby neighbor of the up-sampled data and then calculates the distance between them. Subsequently, the distance is added by a random number is k, which is added to the initial sample for creating additional synthetic examples. Algorithm 1 presents the SMOTE, where the features created synthetically are well-proportioned and can be applied in the system layer.

Algorithm 1: Working of SMOTE	
Variable	Majority Factors F^+, Minority Factor F^-, threshold $\hat{h}$, ratio $\widehat{R}$, Euclidean distance ∂, Generated Samples $\mathbb{S}$
Input:	Total number of majority factors F^+ and minority Factors F^-
Start 1. Set threshold $\hat{h}$ th-> max (degrees (class imbalance)) 2. For every minority factor f, calculate Euclidean distance ∂ 3. $\widehat{R}_i = \Delta\frac{i}{k}, k = 10$ 4. $\widehat{R}_f < \widehat{R}_i / \Sigma\widehat{R}_i$ 5. $\beta = \frac{F^+}{F^-}$ 6. $\mathbb{S} = (F^+ - F^-) * \beta$ Output: No of $\mathbb{S}$ End	

3.2 System layer

In particular, given the turnover profile elements of the employee *E*, the GRU circle is utilized to get these properties first. Next, evaluate the contributions of numerous turnover elements to the ultimate conclusion. When learning long-term dependencies, GRU networks outperform typical recurrent neural networks (RNNs) by overcoming the challenges of gradient vanishing and explosion. The GRU network is used in the system layer, and its enhanced prediction variant remedies this deficiency. Finally, this work applies an improved loss function to produce reliable dropout predictions for the unbalanced problem of the turnover dataset.

The purpose of our suggested model in this study is to forecast employee turnover. The model forecasts the label *r* for the *i*th personnel sample based on employee attributes and information related to the profile *Ei*. *Ei's* profile data includes number_project, average_monthly_hours, salary, increment, age, gender, evaluation, manager, satisfaction_level, time_spend_company, and so on. The classification label *r* denotes the projected outcome, such as turnover or non-turnover. This paper treats the turnover problem as a binary classification case with general loss, $r \in \{0, 1\}$, Let $r = 1$ if the *i*th men will quit the job. Otherwise, $r = 0$. As the objective of this study is to estimate the turnover category $T(ri, Ei)$, the following formal definition of personnel turnover prediction is provided in this study as shown in Equation 1:

$$r = \begin{cases} 1, & T(r\hat{}i = 1 | Ei) > 0.5 \\ 0, & otherwise \end{cases} \tag{1}$$

The properties of the input turnover rely on them, so a GRU neural network is used to train sample profile feature representations. As a type of recurrent neural network, GRU is efficient at modeling information about patterns. GRU can also determine how different parts of an individual's profile affect each other and stop gradients from spiking and dropping off. At the first layer, input coding is conducted. The GRU model used numerical or textual inputs to code the HR data that has already been analyzed and balanced. It usually involves changing categorical variables into one-hot encoded matrices using normalization or scaling numerical variables. Next, Sequence Processing starts.

GRU can handle the collapsed HR data in order, taking into account how different data points depend on each other in time. The GRU cell gets each input in order, and the model keeps a secret internal state to store the information from the previous inputs. Afterward, Hidden State Update gets activated. The GRU model's that hide state is updated every time step by adding current input to the earlier hidden state. The filtering system of GRU decides the amount of information from the prior hidden state that must be kept while determining how much new information will be added. After the hidden state update, Information Propagation is activated. The new hidden state is sent to the next time step. This lets the GRU spread information and track long-term relationships across the sequential data. The way of spreading information helps the GRU model figure out trends and connections in the HR data. Once the HR data has been processed, the GRU model generates outputs based on what it has learned. For instance, based on the input, as the goal is to predict employee turnover, the GRU gives a binary classification that shows how likely an employee will leave or stay. Finally, the output layer is where the HR data is paired with goal labels (such as "dropout labels") that describe whether an employee is going to stay or leave. The GRU model learns to use optimization methods like backpropagation and gradient descent to change its internal parameters so that the gap

between its predicted outputs and the actual labels is as small as possible. The working of the system layer is explained in Figure 4.

As an initial vector sequence, this study signifies all of the turnover characteristics in each sample.

$$E_{i=1}^{n} = \left\{ e_1^{turnover}, e_2^{turnover}, \ldots, e_n^{turnover} \right\}$$

n refers the number of profiles. Further, Eq. 2 to 5 are used in GRU.

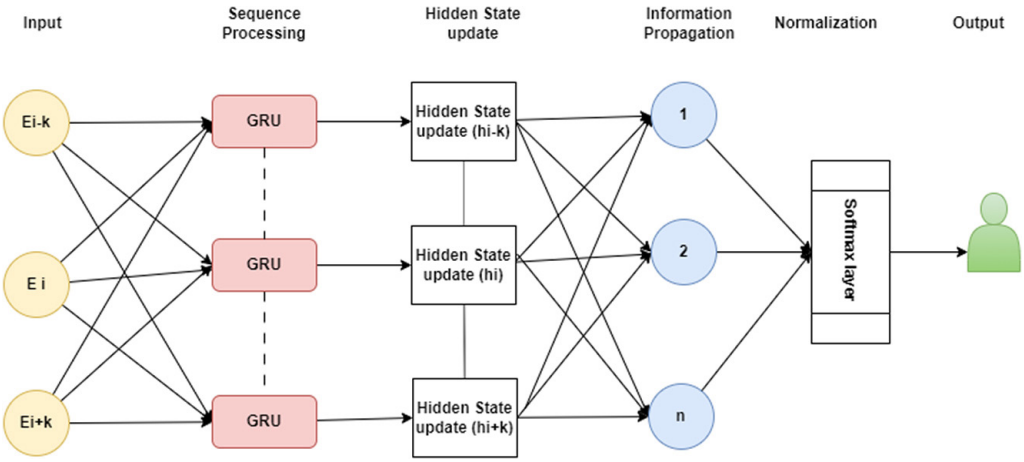

Fig. 4. GRU steps in the system layer

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r) \tag{2}$$

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z) \tag{3}$$

$$h\tilde{}t = \tanh(W x_t + U(r_t \odot h_{t-1}) + b_h) \tag{4}$$

$$h_t = z_t \odot h_{t-1} + (1 - zt) \odot h\tilde{}t \tag{5}$$

Where σ is the sigmoid function, an element-wise product is denoted as $\odot$, rt is used to control the mixture of new input with the current memory, z_t holds previous memories that are mixed with the current state, and $h\tilde{}t$ is the candidate state of h_t.

After obtaining the profile features of the input samples, attention is focused on calculating the weight coefficient, which evaluates the comparative significance of various attributes and characteristics of the data set. For example, for any given employee instance of the turnover sample, this study has N sets of output from N GRU cells; for the illustrative example, attention is computed via Eqs. (6)–(8). The vector c_i is obtained by weighted sum. The calculation can be written as

$$C_i = \sum_{j=1}^{N} \propto_{ij} \mathcal{V}_j \tag{6}$$

Where c_i represented as the formation of the sequence as the weighted sum of hidden representation, α_{ij} indicates the normalized importance. Each E_j can have weight coefficient $\propto_{ij}$ which is calculated

$$\propto_{ij} = softmax(E_{ij}) = \frac{e^{Eij}}{\sum_{k=1}^{N} e^{E\,ik}} \tag{7}$$

E_{ij} denote the result about degree of dependent between E_i and E_j.

$$E_{ij} = F\ score\ (E_i, E_j) = uT\ tanh(W_1 E_i + W_2 E_j) \tag{8}$$

F score is a function to retrieve the score in terms of *E_i* and *E_j*, where u, *W_1*, and *W_1* are learned attention parameters. *E_ij* obtained from function *F score* is normalized by *softmax* by Equation (7). The output of attention is a weighted sum that is routed through the prediction layer.

3.3 Prediction layer

This is the final layer that gives the end classification by adapting the proposed model. In this layer, it is only decided whether an employee will stay or leave an organization. The final classification labels for leave and stay are 0 and 1, respectively.

4 EXPERIMENTAL RESULTS AND EVALUATION

In order to validate the outcome of the proposed model, an evaluation was involved in two different aspects. First, the accuracy, MSE, and AUC of the proposed GRU-based approach were studied, and results were inferred. Afterward, the proposed GRU was compared with baseline methods to compare performance measures. The architecture of the model is a single-layer GRU network and a fully connected dense layer used for binary classification. The hidden units in the GRU layer are 64, and the dropout rate to curb overfitting from the model is 0.2. The activation function used is ReLU in the hidden layers, and the output layer is activated by the sigmoid function.

The loss function used in training the model is binary cross-entropy. Training went for 100 epochs using a learning rate of 0.001 and an Adam optimizer using a batch size of 32. Early stopping was used based on the validation loss to avoid overfitting. We conducted all the experiments using 10-folds cross-validation so that the results produced are robust and reproducible. This section explains the performance measures and experimental results and compares the solution with baseline methods to validate the proposed solution.

4.1 Performance measures

To ensure comprehensive coverage and unbiased evaluation of the dataset, various assessment measures were employed in this study. These measures serve as a basis for justifying the performance of any model. The performance evaluation measures used are as follows in Equations 9–12:

$$Precision\ (P) = \frac{Number\ of\ correctly\ Predicted\ Leaving\ Employee}{Number\ of\ all\ Predicted\ Leaving\ Employee} \tag{9}$$

$$Recall\ (R) = \frac{Number\ of\ retrieved\ Leaving\ Employee}{Number\ of\ Leaving\ Employee} \tag{10}$$

Accuracy: This statistic measures the proportion of right predictions to the overall number of instances evaluated. Calculated using Equation 11

$$Accurcay = \frac{No.\,of\ correct\ predictions}{Total\,No.\,of\ predictions} \tag{11}$$

Mathew correlation coefficient (MCC): Typically, the MCC is employed in cases of unbalanced datasets. As with the loss and accuracy patterns, the MCC also diminishes for the test set (see Equation 12); a value of +1 indicates flawless prediction.

$$\text{MCC} = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{12}$$

This metric is predominantly intended for evaluating binary classification problems. MCC is a singular number derived from the confusion matrix's parameter values.

4.2 Experimental results

Assessing the prediction model with merely train-test sets is not always sufficient. To achieve a more realistic measurement, cross-validation with 10-folds is performed. The proposed GRU-based model, employing 10-fold cross validations, achieved the highest performances among all specified ensemble classifiers. Figure 5 demonstrates the overall performance of the proposed GRU model across 10 folds of iteration, while Figure 6 presents the results of MSE over the same 10 folds. The results indicate that as the number of folds increased, the training and testing MSE decreased, demonstrating improved performance. Concurrently, accuracy increased across the folds.

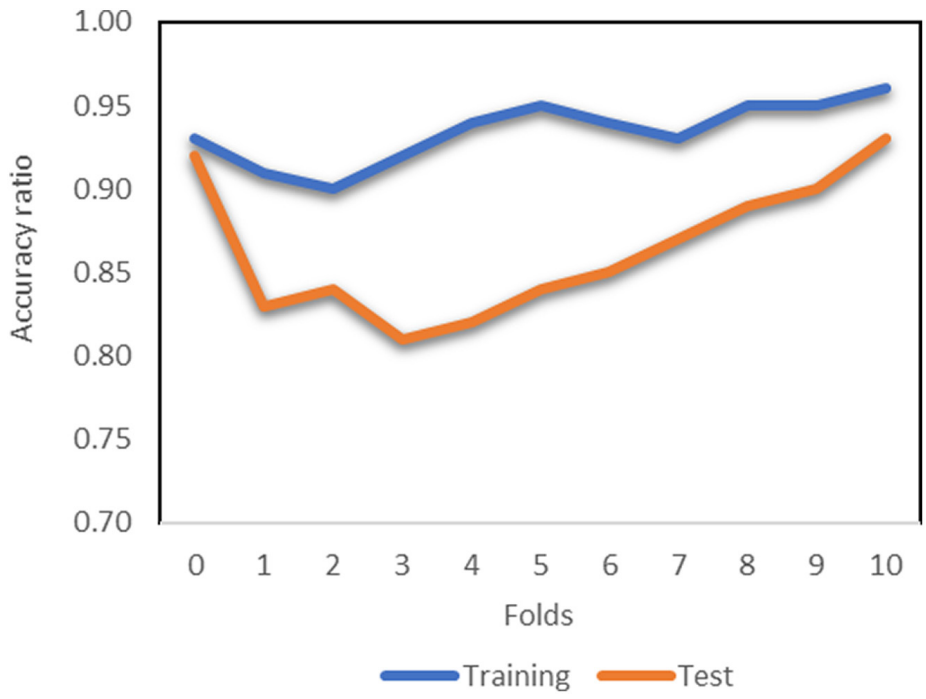

Fig. 5. Training and testing accuracy in each iteration of cross-validation

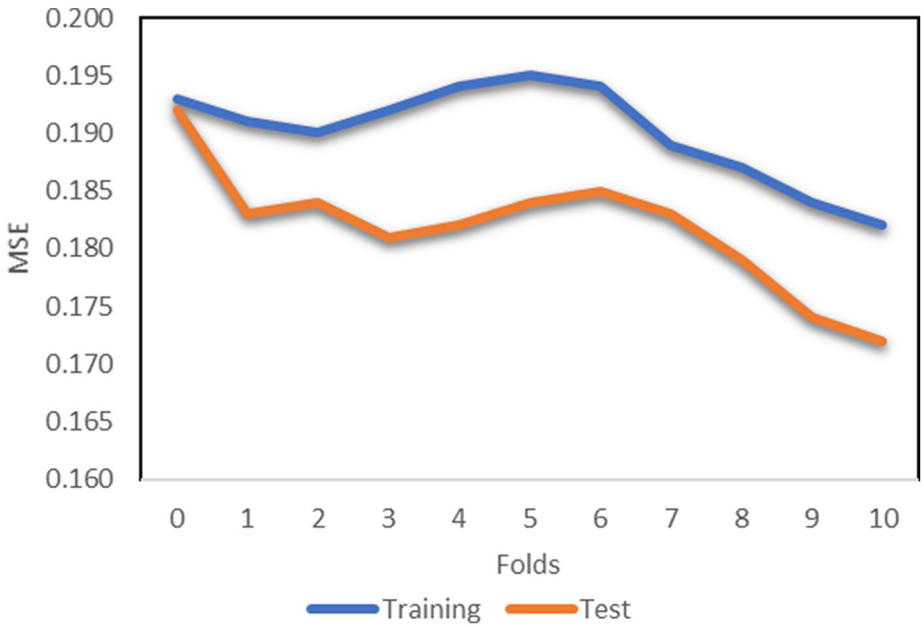

Fig. 6. Training and testing MSE in iteration of cross validation

The ROC curve is a visual display of a binary classifier's performance. It demonstrates the True Positive Rate versus the False Positive Rate for various classification limits. The curve depicts the balance between the classifier's capacity to accurately identify positive examples (True Positive Rate) and its likelihood to mistake negative cases for positive ones (False Positive Rate). The AUC is a summary metric derived from the ROC curve. It quantifies the classifier performance by measuring the area under the curve. The AUC ranges from 0 to 1, with a higher value indicating better classifier performance. An AUC of 0.5 suggests a classifier that performs no better than random guessing, while an AUC of 1 indicates a perfect classifier with perfect discrimination between positive and negative instances. Figure 7 shows the ROC of the selected model.

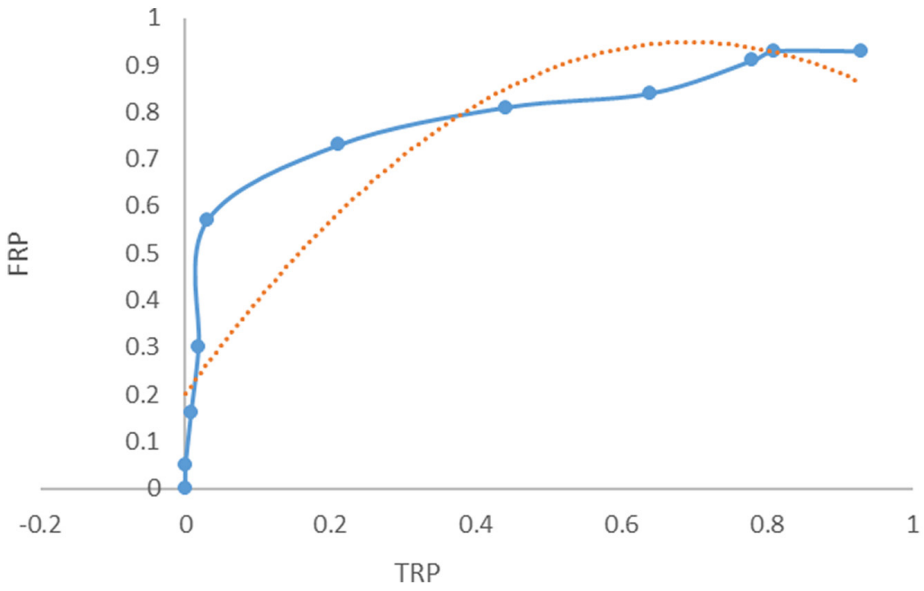

Fig. 7. AUC and ROC of the proposed framework

The MCC for the test set also achieved a value of 0.93, indicating that the proposed framework has a strong capacity for prediction. In the first scenario, the model's efficacy when classifying employees was severely diminished. This is because imbalanced values existed in the initial dataset. Later, after processing, these instances were rebalanced, and the framework was able to classify the departing class employees and the remaining class employees very accurately. Figure 8 presents the MCC of the selected dataset with the imbalance dataset. Whereas Figure 9 demonstrates MCC with a balanced dataset.

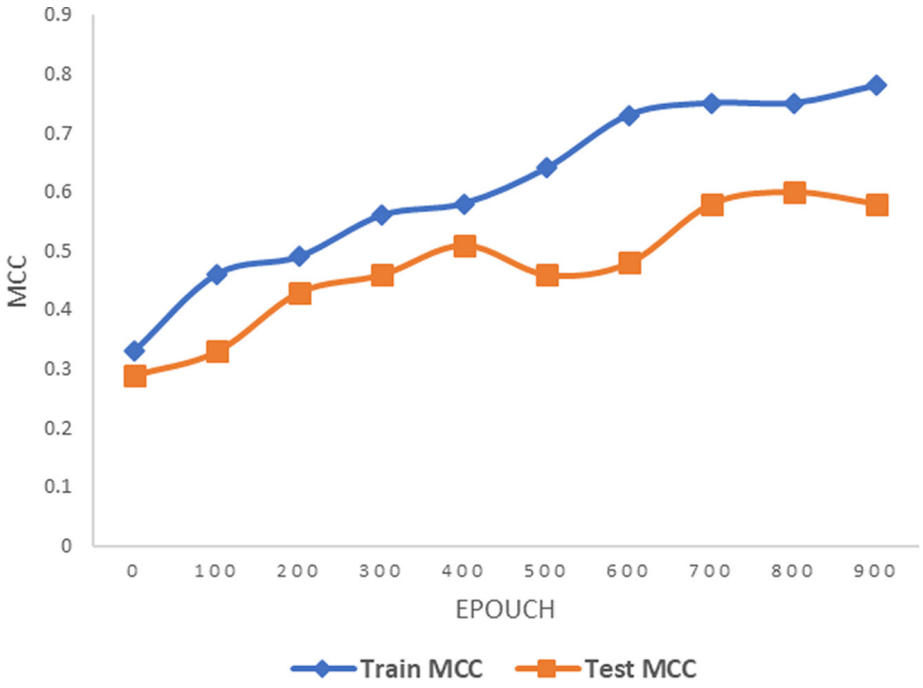

Fig. 8. MCC with imbalance dataset

Besides accuracy, the proposed framework was evaluated on precision, recall, F1-score, ROC-AUC, and MCC to ensure reliable assessment under imbalanced data conditions. The GRU-based model outperformed with high precision and recall values, reflecting its strong capability of correctly identifying the employee dropout cases while minimizing false alarms. The ROC-AUC score further confirms the discriminative power of the model, while the value of MCC as 0.93 reflects robust prediction performance across both the classes. Confusion matrix analysis demonstrates a great reduction in false negatives after SMOTE-based balancing, which is very critical for proactive employee retention strategies.

4.3 Performance comparison with baseline methods

The initial data set is utilized in a different study, which offers a problem because of the significant disparity in sample counts between targets 0 and 1. The suggested approach is to compare the existing state-of-the-art methods in Table 2. The results showed that the proposed model outperformed all competing systems regarding accuracy, owing to deep learning's classification strength and early balancing stages.

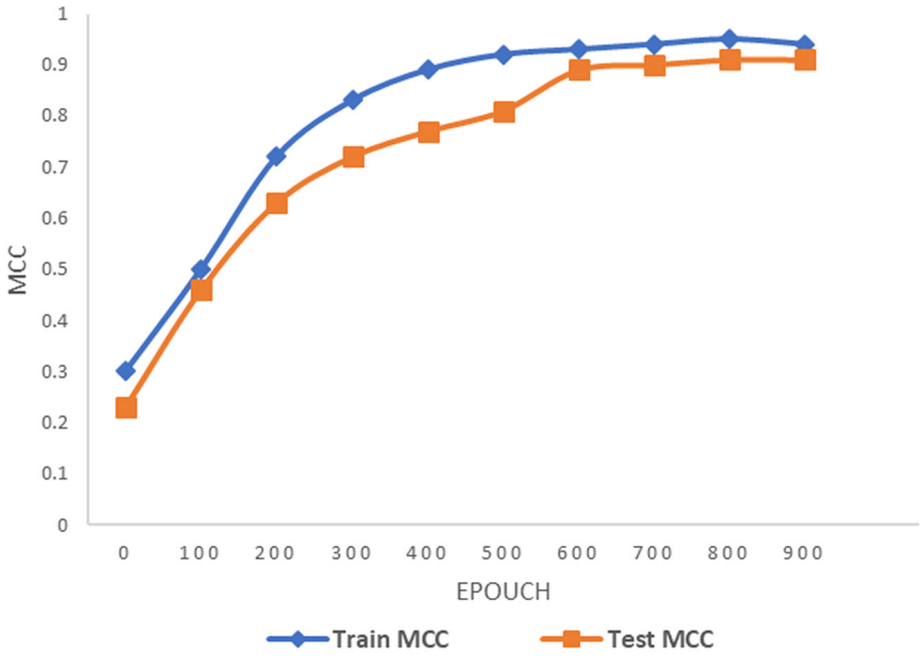

Fig. 9. MCC with balanced dataset

Table 2. Compare the proposed model with state-of-the-art contribution

Contribution	Accuracy (%)
ECDT [25]	82
DNN [24]	89.1
Proposed GRU	95

The GRU algorithm outperformed existing ML and deep neural network-based techniques in the prediction challenge. Furthermore, 10-fold cross-validation was used to get realistic performance using the balanced dataset. The performance measures of the proposed work are better than these methods using the same dataset. Once the dataset had been balanced, as described in previous sections, the proposed GRU-based technique outperformed all other state-of-the-art approaches in prediction accuracy. The adoption of the SMOTE approach before deep learning resulted in improved accuracy. The GRU-based deep learning strategy aided in selecting only the most compelling features. We compared the cross-validation approach with DNN [24] and ECDT [25]. As seen in Figure 10, the accuracy of our model is significantly higher than that of existing approaches.

5 DISCUSSION

This study emphasizes adding significant contribution and meaning to the research that has already been done, which uses the DL-based GRU method in system prediction. One thing that makes the proposed model stronger is that during the evaluation, we also tried to find the relationship between different factors that can impact the employee's decision to drop out. This is important because it can be used to measure precision and accuracy. Also, past studies do not emphasize identifying

any link between the different reasons that cause employees to quit or change jobs. We constructed a heatmap to further analyze the correlation among multiple factors of the employee IBM HR dataset. The primary purpose of a heatmap is to provide a visual summary of the data, highlighting patterns, trends, or relationships between variables. Heat maps can be particularly useful, as the HR dataset had multivariate data. Heatmaps can quickly identify areas of interest, spot outliers, and make comparisons across different variables or categories, as shown in Figure 11.

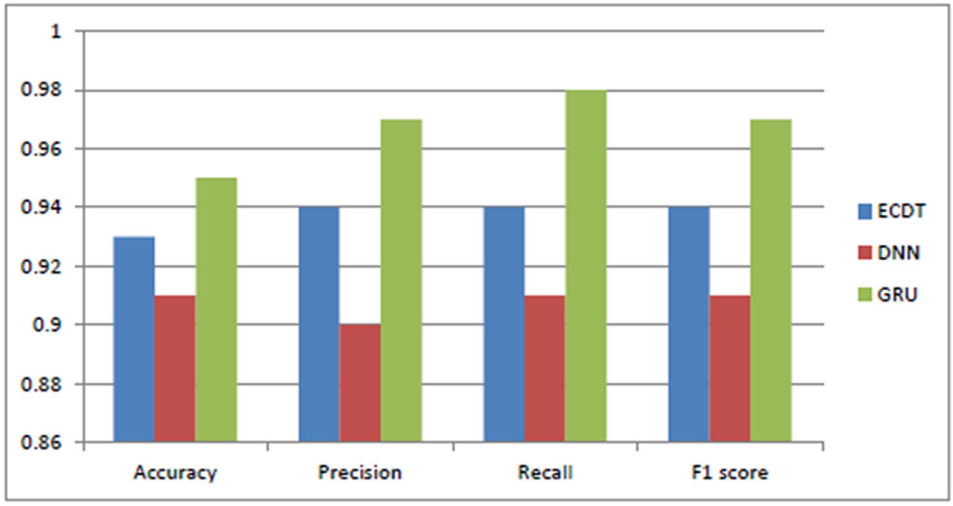

Fig. 10. Comparison of the proposed model with baseline methods

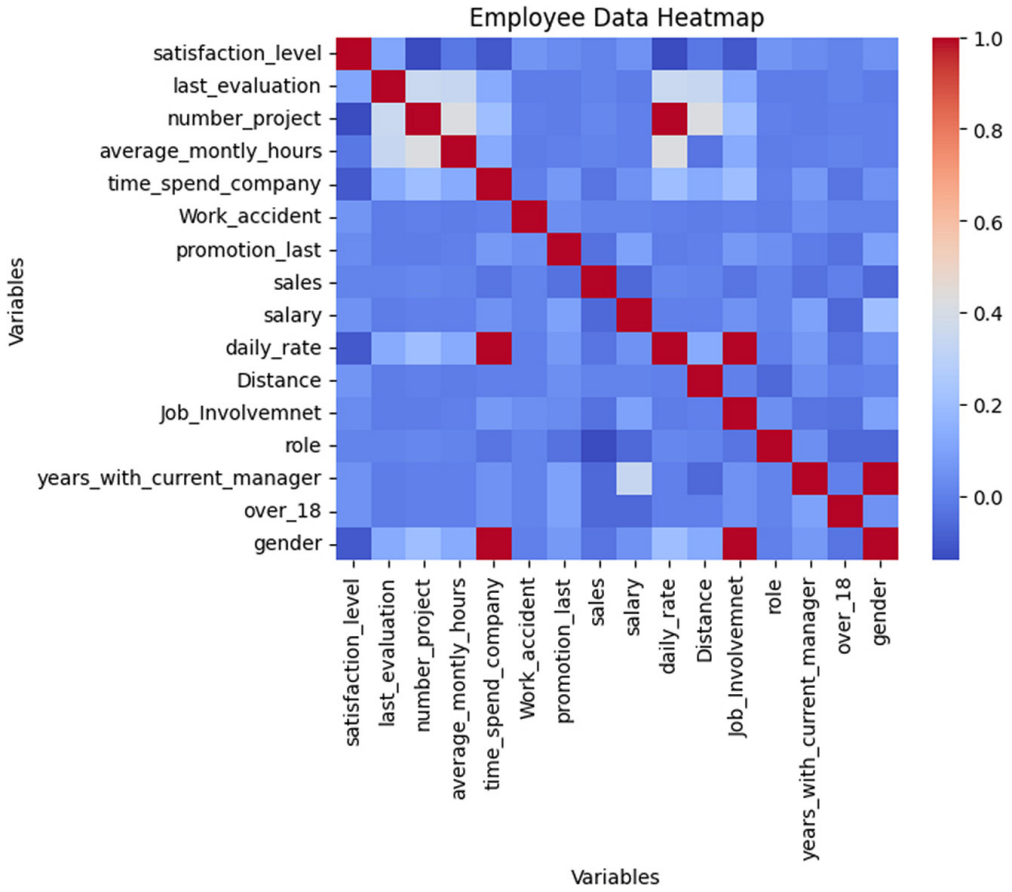

Fig. 11. Relationships of multiple employee factors affecting each other for turnout

The relationships of various factors of employee turnout are explained below. The heatmap can demonstrate correction between pairs of variables. A positive correlation indicates that if one element is increasing, it leads to an increase in another factor. For example, the employee's satisfaction level is increased with the number of projects, as it ensures that the employee is happy about his job. Similarly, role and job involvement are also directly proportional.

On the other hand, the number of projects does not impact distance from home, monthly hours, and job involvement. This relation is evident due to contracting shades of the heatmap. The key drivers of an employee include variables like salary, job involvement, role, and evaluation, as indicated in the figure. At the same time, gender appears to be an outlier that is not affecting the decision of employee turnout. We also observed some trade-offs between factors, like a negative correlation between the quantity of projects and monthly hours on average, indicating that employees with more ventures tend to work fewer hours per project.

The attention mechanism embedded within the GRU model enables identification of influential employee attributes contributing to turnover prediction. Experimental analysis indicates that monthly income, job involvement, work-life balance, overtime, and performance evaluation scores are among the most critical predictors. Correlation analysis further reveals interdependencies among these factors, offering actionable insights for HR managers. Unlike traditional black-box models, the proposed framework enhances interpretability by quantifying feature relevance, thereby supporting evidence-based retention strategies.

From this discussion, it is observed that multiple factors affect the decision of employee turnout in an organization, which plays an essential role in forming the carrier. Due to imbalanced datasets, traditional employee turnout prediction techniques may suffer from low relevance and precision. Furthermore, due to the complexity of many factors, determining the reason for employee turnover is exceptionally chaotic and time-consuming. Based on the balanced dataset and GRU-based prediction model, this study provides an effective predicting mechanism for employee dropout that identifies a particular employee from the dataset of an organization. The very most important thing of study is higher accuracy achieved due to GRU. This approach enables organizations to work on employee retention policies and offer better careers. Along with the prediction, the proposed model provides organizations with different factors that depend on each other with direct and indirect relations. Last but not least, the proposed model has certain limitations, including scalability.

Although the proposed SMOTE–GRU framework demonstrated strong predictive performance, several avenues for future research remain. Firstly, the model is to be validated with real-world human resources datasets collected from organizations across industries and organizational scales to enhance generalizability. Secondly, future studies might investigate the use of state-of-the-art hyperparameter optimization techniques, such as Bayesian optimization or evolutionary algorithms, to further fine-tune GRU. Thirdly, scalability issues can be explored by using distributed training frameworks and testing the model on large-scale enterprise HR systems. Finally, the integration of sentiment analysis from employee feedback, emails, or performance reviews may yield richer contextual insight and further improve dropout prediction accuracy.

6 CONCLUSION

In today's highly competitive business environment, employees influence a company's profitability. The higher the employee turnover rate, the greater the recruitment and training expenses, and the lower the customer satisfaction and level

of service. By anticipating employee behavior, employee attrition analysis seeks to reduce employee turnover. Classification and prediction algorithms like GRU offer significant opportunities for predicting employee turnover. This work proposed a GRU-based model for turnover prediction in this employee HR dataset comprising 1470 records against 35 factors. Precisely, we first balanced the IBM employee dataset using SMOTE, then used GRU to learn significant turnover variables, and last employed an attractive method of the model for profit information. Finally, we ran trials using turnover data to validate our model's effectiveness and accuracy. The numerical results indicate that the proposed GRU-based model was a practical approach for identifying employee dropout with an accuracy of 0.95. Furthermore, different related and interrelated factors are also studied to help the organization plan its retention approach. In the future, we intend to apply sentimental analysis to discover turnover elements to forecast employee turnover behavior.

7 ACKNOWLEDGMENT

Conflict of interests: The author declared no potential conflicts of interest.
Research involving Human Participants and/or Animals: No
Data Availability Statement: The dataset is publicly available at https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset

8 REFERENCES

[1] H. Han *et al.*, "A theoretical framework development for hotel employee turnover: Linking trust in supports, emotional exhaustion, depersonalization, and reduced personal accomplishment at workplace," *Sustainability*, vol. 12, no. 19, p. 8065, 2020. https://doi.org/10.3390/su12198065

[2] A. R. Skelton, D. Nattress, and R. J. Dwyer, "Predicting manufacturing employee turnover intentions," *J. Econ. Financ. Adm. Sci.*, vol. 25, no. 49, pp. 101–117, 2020. https://doi.org/10.1108/JEFAS-07-2018-0069

[3] S. Garg, S. Sinha, A. K. Kar, and M. Mani, "A review of machine learning applications in human resource management," *Int. J. Product. Perform. Manag.*, vol. 71, no. 5, pp. 1590–1610, 2022. https://doi.org/10.1108/IJPPM-08-2020-0427

[4] L. Sun *et al.*, "Employee engagement: A literature review," *Int. J. Hum. Resour. Stud.*, vol. 9, no. 1, pp. 63–80, 2019. https://doi.org/10.5296/ijhrs.v9i1.14167

[5] W. Y. Degbey, P. Rodgers, M. D. Kromah, and Y. Weber, "The impact of psychological ownership on employee retention in mergers and acquisitions," *Hum. Resour. Manag. Rev.*, vol. 31, no. 3, p. 100745, 2021. https://doi.org/10.1016/j.hrmr.2020.100745

[6] R. Jain and A. Nayyar, "Predicting employee attrition using xgboost machine learning approach," in *2018 Int. Conf. Syst. Model. & Adv. Res. Trends*, 2018, pp. 113–120. https://doi.org/10.1109/SYSMART.2018.8746940

[7] P. R. Srivastava and P. Eachempati, "Intelligent employee retention system for attrition rate analysis and churn prediction: An ensemble machine learning and multi-criteria decision-making approach," *J. Glob. Inf. Manag.*, vol. 29, no. 6, pp. 1–29, 2021. https://doi.org/10.4018/JGIM.20211101.oa23

[8] D. McCarthy, P. Alexander, and Y. Jung, "Enhancing the organisational commitment of public sector accounting staff through the pursuit of CSR objectives," *J. Account. & Organ. Chang.*, vol. 18, no. 2, pp. 304–324, 2022. https://doi.org/10.1108/JAOC-09-2020-0139

[9] B. Hmoud *et al.*, "Will artificial intelligence take over human resources recruitment and selection," *Netw. Intell. Stud.*, vol. 7, no. 13, pp. 21–30, 2019.

[10] R. Yasin, "Responsible leadership and employees' turnover intention. Explore the mediating roles of ethical climate and corporate image," *J. Knowl. Manag.*, vol. 25, no. 7, pp. 1760–1781, 2021. https://doi.org/10.1108/JKM-07-2020-0583

[11] N. Jain, A. Tomar, and P. K. Jana, "A novel scheme for employee churn problem using multi-attribute decision making approach and machine learning," *J. Intell. Inf. Syst.*, vol. 56, pp. 279–302, 2021. https://doi.org/10.1007/s10844-020-00614-9

[12] P. K. Jain, M. Jain, and R. Pamula, "Explaining and predicting employees' attrition: A machine learning approach," *SN Appl. Sci.*, vol. 2, pp. 1–11, 2020. https://doi.org/10.1007/s42452-020-2519-4

[13] W. Zhang, H. Li, L. Tang, X. Gu, L. Wang, and L. Wang, "Displacement prediction of Jiuxianping landslide using gated recurrent unit (GRU) networks," *Acta Geotech*, vol. 17, no. 4, pp. 1367–1382, 2022. https://doi.org/10.1007/s11440-022-01495-8

[14] C. Li, G. Tang, X. Xue, A. Saeed, and X. Hu, "Short-term wind speed interval prediction based on ensemble GRU model," *IEEE Trans. Sustain. Energy*, vol. 11, no. 3, pp. 1370–1380, 2019. https://doi.org/10.1109/TSTE.2019.2926147

[15] T.-Y. Kim and S.-B. Cho, "Predicting residential energy consumption using CNN-LSTM neural networks," *Energy*, vol. 182, pp. 72–81, 2019. https://doi.org/10.1016/j.energy.2019.05.230

[16] T. S. Kumar, "Data mining based marketing decision support system using hybrid machine learning algorithm," *J. Artif. Intell.*, vol. 2, no. 3, pp. 185–193, 2020. https://doi.org/10.36548//jaicn.2020.3.006

[17] S. Tahir, Y. Hafeez, M. A. Abbas, A. Nawaz, and B. Hamid, "Smart learning objects retrieval for e-learning with contextual recommendation based on collaborative filtering," *Educ. Inf. Technol.*, pp. 1–38, 2022. https://doi.org/10.1007/s10639-022-10966-0

[18] Z. E. Rasjid and R. Setiawan, "Performance comparison and optimization of text document classification using k-NN and naive bayes," *Classification Techniques, Procedia Comput. Sci.*, vol. 116, pp. 107–112, 2017. https://doi.org/10.1016/j.procs.2017.10.017

[19] C. V. G. Zelaya, "Towards explaining the effects of data preprocessing on machine learning," in *2019 IEEE 35th Int. Conf. data Eng.*, 2019, pp. 2086–2090.

[20] S. Ali, Y. Hafeez, M. Humayun, N. S. M. Jamail, M. Aqib, and A. Nawaz, "Enabling recommendation system architecture in virtualized environment for e-learning," *Egypt. Informatics J.*, vol. 23, no. 1, pp. 33–45, 2022. https://doi.org/10.1016/j.eij.2021.05.003

[21] S. Dutta, S. K. Bandyopadhyay, and S. Kumar Bandyopadhyay, "Employee attrition prediction using neural network cross validation method," *Int. J. Commer. Manag. Res.*, vol. 6, no. 3, pp. 80–85, 2020. https://doi.org/10.20944/preprints202006.0333.v1

[22] Y. Zhao, M. K. Hryniewicki, F. Cheng, B. Fu, and X. Zhu, "Employee turnover prediction with machine learning: A reliable approach," in *Intell. Syst. Appl. IntelliSys. 2018*, *Advances in Intelligent Systems and Computing*, K. Arai, S. Kapoor, and R. Bhatia, Eds., vol. 869, 2019. https://doi.org/10.1007/978-3-030-01057-7_56

[23] J. Hang, Z. Dong, H. Zhao, X. Song, P. Wang, and H. Zhu, "Outside in: Market-aware heterogeneous graph neural network for employee turnover prediction," in *Proc. Fifteenth ACM Int. Conf. Web Search Data Min.*, 2022, pp. 353–362. https://doi.org/10.1145/3488560.3498483

[24] S. Al-Darraji, D. G. Honi, F. Fallucchi, A. I. Abdulsada, R. Giuliano, and H. A. Abdulmalik, "Employee attrition prediction using deep neural networks," *Computers*, vol. 10, no. 11, p. 141, 2021. https://doi.org/10.3390/computers10110141

[25] E. Pekel Ozmen and T. Ozcan, "A novel deep learning model based on convolutional neural networks for employee churn prediction," *J. Forecast.*, vol. 41, no. 3, pp. 539–550, 2022. https://doi.org/10.1002/for.2827

[26] N. B. Yahia, J. Hlel, and R. Colomo-Palacios, "From big data to deep data to support people analytics for employee attrition prediction," *IEEE Access*, vol. 9, pp. 60447–60458, 2021. https://doi.org/10.1109/ACCESS.2021.3074559

[27] F. K. Alsheref, I. E. Fattoh, and W. M. Ead, "Automated prediction of employee attrition using ensemble model based on machine learning algorithms," *Comput. Intell. Neurosci.*, vol. 2022, p. 7728668, 2022. https://doi.org/10.1155/2022/7728668

[28] A. Jadhav *et al.*, "Churn prediction of employees using machine learning techniques," *Teh. Glas.*, vol. 15, no. 1, pp. 51–59, 2021. https://doi.org/10.31803/tg-20210204181812

[29] X. Gao, J. Wen, and C. Zhang, "An improved random forest algorithm for predicting employee turnover," *Math. Probl. Eng.*, vol. 2019, no. 1, pp. 1–12, 2019. https://doi.org/10.1155/2019/4140707

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

[31] S. Najafi-Zangeneh, N. Shams-Gharneh, A. Arjomandi-Nezhad, and S. Hashemkhani Zolfani, "An improved machine learning-based employees attrition prediction framework with emphasis on feature selection," *Mathematics*, vol. 9, no. 11, p. 1226, 2021. https://doi.org/10.3390/math9111226

[32] A. Alamsyah and N. Salma, "A comparative study of employee churn prediction model," in *2018 4th Int. Conf. Sci. Technol.*, Yogyakarta, Indonesia, 2018, pp. 1–4. https://doi.org/10.1109/ICSTC.2018.8528586

[33] X. Wang and J. Zhi, "A machine learning-based analytical framework for employee turnover prediction," *J. Manag. Anal.*, vol. 8, no. 3, pp. 351–370, 2021. https://doi.org/10.1080/23270012.2021.1961318

[34] F. Mozaffari, M. Rahimi, H. Yazdani, and B. Sohrabi, "Employee attrition prediction in a pharmaceutical company using both machine learning approach and qualitative data," *Benchmarking: An Int. J.*, vol. 30, no. 10, pp. 4140–4173, 2022. https://doi.org/10.1108/BIJ-11-2021-0664

[35] S. A. Ali Shah, I. Uddin, F. Aziz, S. Ahmad, M. A. Al-Khasawneh, and M. Sharaf, "An enhanced deep neural network for predicting workplace absenteeism," *Complexity*, vol. 2020, no. 1, pp. 1–12, 2020. https://doi.org/10.1155/2020/5843932

[36] S. S. Alduayj and K. Rajpoot, "Predicting employee attrition using machine learning," in *2018 Int. Conf. Innov. Inf. Technol.*, 2018, pp. 93–98. https://doi.org/10.1109/INNOVATIONS.2018.8605976

[37] J. Chen, H. Huang, A. G. Cohn, D. Zhang, and M. Zhou, "Machine learning-based classification of rock discontinuity trace: SMOTE oversampling integrated with GBT ensemble learning," *Int. J. Min. Sci. Technol.*, vol. 32, no. 2, pp. 309–322, 2022. https://doi.org/10.1016/j.ijmst.2021.08.004

[38] A. Fernández, S. Garcia, F. Herrera, and N. V. Chawla, "SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary," *J. Artif. Intell. Res.*, vol. 61, pp. 863–905, 2018. https://doi.org/10.1613/jair.1.11192

[39] J. A. Richards, "Feature reduction," in *Remote Sens. Digit. Image Anal.*, Springer, Cham, 2022, pp. 403–446. https://doi.org/10.1007/978-3-030-82327-6_10

[40] R. Abdulhammed, H. Musafer, A. Alessa, M. Faezipour, and A. Abuzneid, "Features dimensionality reduction approaches for machine learning based network intrusion detection," *Electronics*, vol. 8, no. 3, p. 322, 2019. https://doi.org/10.3390/electronics8030322

[41] S. A. Alasadi and W. S. Bhaya, "Review of data preprocessing techniques in data mining," *J. Eng. Appl. Sci.*, vol. 12, no. 16, pp. 4102–4107, 2017.

[42] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," *Knowl. Inf. Syst.*, vol. 33, no. 1, pp. 1–33, 2012. https://doi.org/10.1007/s10115-011-0463-8

9 AUTHORS

Ibrar Hussain is with the PMAS-Arid Agriculture University, Rawalpindi, Pakistan.

Sidra Tahir is with the PMAS-Arid Agriculture University, Rawalpindi, Pakistan (E-mail: stahir@uaar.edu.pk).

Asif Nawaz is with the PMAS-Arid Agriculture University, Rawalpindi, Pakistan.

Kashif Mehmood is with the PMAS-Arid Agriculture University, Rawalpindi, Pakistan.

Ahthasham Sajid is with the Department of Computer Science, Fazaia Bilquis College of Education for Women's PAF Nur Khan Base, Air University, Islamabad, Pakistan.

Sabitha Banu is with the Department of Computer Science and Cyber, PSGR Krishnammal College for Women, Coimbatore, Tamilnadu, India.