Advancing Employee Attrition Prediction Through Graph-Based Learning and XAI

Mohammad Rahman
Student, Computer Science and Engineering ,
Guru Nanak Institutions Technical Campus (Autonomous),
Hyderabad, Telangana, India- 501506
rahman91340@gmail.com

Mrs. G. Srujana Bharathi Assistant Professor,
Computer Science
and Engineering,
Guru Nanak Institutions Technical Campus (Autonomous),
Hyderabad, Telangana, India- 501506
gsbharathi.csegnitc@gniindia.org

Narsing Nithish Kumar
Student, Computer Science and Engineering ,

Guru Nanak Institutions Technical Campus (Autonomous),
Hyderabad, Telangana, India- 501506
narsingnithish@gmail.com

Pathakota Prudhvi Raj
Student, Computer Science and Engineering,
Guru Nanak Institutions Technical Campus (Autonomous) ,
Hyderabad, Telangana, India- 501506
prudhvirajpathakota@gmail.com

Abstract—Employee turnover remains a significant concern for organizations around the world. While advanced machine learning models have shown potential in forecasting employee attrition, their real-world application is often constrained by the inability to capture the complex relational patterns within tabular HR datasets. To overcome this limitation, this research presents an innovative approach that transforms conventional employee records into a knowledge graph format, enabling the use of Graph Convolutional Networks (GCNs) for more in-depth feature learning. Beyond mere prediction, the framework integrates explainable artificial intelligence (XAI) methodologies to identify and interpret the key factors driving employee retention or resignation. The study utilizes a well-known dataset from IBM, comprising 1,470 employee profiles, and compares the proposed model's performance against five widely-used machine learning algorithms. Notably, our enhanced linear Support Vector Machine (L-SVM), augmented with features derived from the knowledge graph, achieved a remarkable accuracy of 92.5%. Furthermore, the application of XAI techniques offered valuable insights into critical variables such as job satisfaction, job involvement, and workplace environment, which heavily influence turnover behavior. This research not only advances predictive modeling in human resource analytics but also empowers organizations with data-driven strategies to effectively mitigate employee attrition.

I. INTRODUCTION

In the modern business landscape, retaining skilled employees is critical to maintaining organizational competitiveness and operational stability. High employee turnover rates can lead to increased recruitment costs, loss of institutional knowledge, reduced productivity, and negative impacts on team morale. As a result, organizations are increasingly turning to predictive analytics to anticipate and mitigate the risk of employee attrition.

Traditional machine learning models, such as decision trees,

logistic regression, and support vector machines, have been widely adopted in human resource (HR) analytics to forecast employee resignations. While these models offer reasonable predictive performance, they often treat data in isolation, failing to capture the deeper interdependencies and relationships that exist between employees and organizational attributes. Human-centric data, especially in the HR domain, is inherently relational—colleagues influence each other, departments exhibit behavioral trends, and similar roles may face common challenges. Flattening this data into a tabular format causes the loss of these critical connections, limiting the model's ability to fully understand the dynamics of attrition.

To address this limitation, this research introduces a novel approach that transforms conventional HR data into a graph- based structure, enabling the use of Graph Convolutional Networks (GCNs). GCNs excel at learning representations from structured graph data, making them well-suited to uncover latent relationships in organizational datasets. By modeling employees as nodes and defining meaningful connections based on shared attributes, departmental links, or performance similarities, the proposed framework leverages the expressive power of graphs for deeper feature extraction.

Moreover, in an era where interpretability is essential for ethical and actionable AI deployment, this work integrates explainable artificial intelligence (XAI) techniques. These methods provide transparency into the decision-making process, allowing HR professionals to understand which factors most significantly contribute to employee attrition. By coupling GCN-based feature learning with post-hoc explanation tools, the model not only predicts who might leave but also offers insights into why they might do so.

This research is built upon a well-established HR dataset provided by IBM and includes comparative analysis with several traditional models. It highlights the potential of graph- based learning in HR contexts and emphasizes the importance of explainability in driving data-informed organizational decisions.

II. RELATED WORK

Employee attrition prediction has been a key focus area within human resource analytics, with numerous studies leveraging machine learning to anticipate voluntary and involuntary resignations. Early approaches primarily employed statistical models like logistic regression to establish relationships between employee features and turnover likelihood. These methods, while interpretable, often lacked the capacity to capture complex, nonlinear patterns in the data.

With the rise of machine learning, algorithms such as decision trees, random forests, support vector machines GCNs to extract deeper patterns, while also employing XAI methods to provide transparency and trust in the predictions.

III. METHODOLOGY

This study proposes a comprehensive framework that integrates graph-based learning with explainable artificial intelligence (XAI) techniques to predict employee attrition and interpret the underlying causes. The methodology consists of four main components: (1) data preprocessing and transformation, (2) knowledge graph construction, (3) implementation of Graph Convolutional Networks (GCNs), and (4) integration of XAI methods for interpretability.

(SVM), and k-nearest neighbors (KNN) became prominent for predicting attrition. For example, random forests have been favored for their robustness and ability to handle heterogeneous data, while SVMs have shown promise in high-dimensional settings. Despite improved accuracy, these models typically rely on tabular representations of data, limiting their ability to model relationships between employees or contextual factors within the organization.

Recent advancements have shifted attention towards deep learning techniques, which offer enhanced performance on large-scale HR datasets. However, deep learning models are often criticized for their "black box" nature, making them difficult to interpret—an important consideration in sensitive domains like human resource management.

To bridge this gap, explainable artificial intelligence (XAI) techniques have been increasingly adopted. Tools such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations) are commonly used to interpret feature contributions in black-box models. These tools provide HR stakeholders with actionable insights by identifying which variables most influence attrition risk, such as job satisfaction, work-life balance, or salary.

Parallel to these developments, graph-based learning has emerged as a powerful paradigm for tasks involving structured and relational data. Graph Neural Networks (GNNs), particularly Graph Convolutional Networks (GCNs), have demonstrated success across diverse domains such as social network analysis, recommender systems, and biomedical research. In these models, data is represented as nodes and edges, enabling the capture of complex inter- entity relationships.

However, applications of GCNs in the HR domain remain relatively limited. Some pioneering studies have explored representing employees as nodes in a knowledge graph, where edges indicate shared characteristics like department, role, or managerial hierarchy. These approaches suggest that incorporating relational information can significantly improve predictive accuracy and model explainability.

This research builds upon these ideas by integrating graph- based learning with XAI, aiming to enhance both predictive power and interpretability in employee attrition prediction. Unlike prior work that treats features in isolation, this study constructs a knowledge graph from HR data and leverages

3.1 Data Preprocessing and Transformation

The dataset used for this research is the IBM HR Analytics Employee Attrition dataset, which contains 1,470 employee records with various demographic, performance, and workplace-related features. The preprocessing involved the following steps:

- **Handling Categorical Features**: Categorical variables such as department, job role, and marital status were encoded using one-hot or label encoding depending on their cardinality and purpose in the graph structure.
- **Normalization**: Numerical features like monthly income, age, and years at company were normalized to bring them to a comparable scale.
- **Missing Values**: As the dataset was clean and hadno missing values, no imputation was required.

Knowledge Graph Construction

To effectively model the interdependencies among employees and organizational attributes, the tabular HR data was transformed into a knowledge graph. In this graph:

- **Nodes** represent individual employees.
- **Edges** are created based on shared attributes such as working in the same department, holding the same job role, or having similar years of experience.
- **Edge Weights** are assigned to reflect the strength of similarity or connection between employees. For example,

two employees in the same job role may have a stronger edge weight than those sharing only the same department.

This relational representation enables the model to consider contextual and social influences in employee behavior, which are often ignored in traditional flat feature representations.

3.2 Graph Convolutional Network (GCN) Model

The knowledge graph was used as input to a Graph Convolutional Network (GCN) to learn more expressive, aggregated node features. The GCN was implemented as follows:

- **Architecture**: A two-layer GCN was employed. The first layer aggregates feature information from neighboring nodes, and the second layer refines this information to classify each node (employee) as "likely to leave" or "likely to stay."
- **Activation Function**: ReLU (Rectified Linear Unit) was used to introduce non-linearity.
- **Loss Function**: Binary cross-entropy was used since the prediction task is binary classification.
- **Optimization**: The Adam optimizer was used with a learning rate fine-tuned through experimentation.

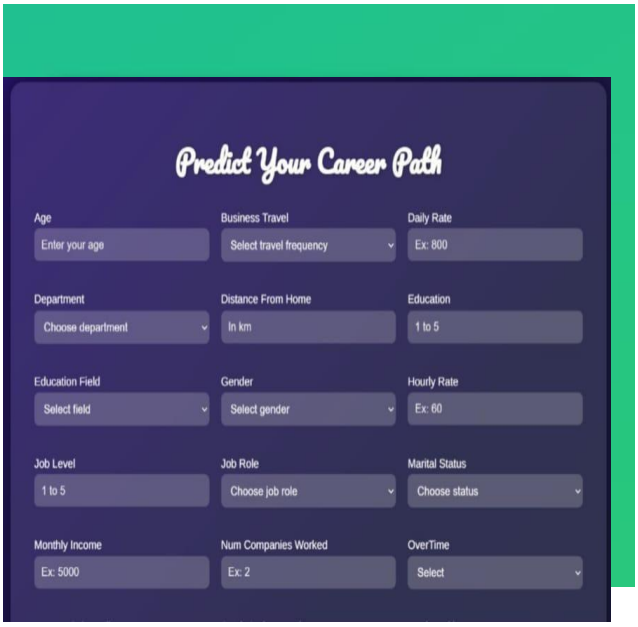
The GCN learned not only from individual attributes but also from the relationships and interactions between employees in the graph structure, providing a richer feature representation.

3.3 Integration with Enhanced Linear SVM

The node embeddings learned from the final GCN layer were extracted and fed into a linear Support Vector Machine (L-SVM) classifier. This hybrid approach combined the relational learning strengths of GCN with the simplicity and interpretability of linear models. The L-SVM was tuned using grid search to optimize hyperparameters such as regularization strength.

3.4 Explainable AI (XAI) Techniques

To ensure the transparency of the model's predictions, explainable AI tools were applied to interpret both GCN embeddings and the L-SVM output. Specifically:
assessing predictive performance, comparing traditional machine learning models with the proposed GCN-based hybrid model, and analyzing the insights generated through AI techniques.



These explanations provided insights into key drivers of attrition, such as job satisfaction, environment satisfaction, and job involvement. The outputs were visualized using SHAP summary plots and edge importance maps to assist HR stakeholders in understanding the factors leading to employee turnover.

IV. EXPERIMENTS AND RESULTS

To evaluate the effectiveness of the proposed framework, a series of experiments were conducted using the IBM HR Analytics Employee Attrition dataset. The focus was on

4.1 EXPERIMENTAL SETUP

The dataset was split into training and testing sets using an 80-20 ratio. All models were trained on the training set and evaluated on the held-out test set. Performance metrics included Accuracy, Precision, Recall, and F1-Score to provide a comprehensive view of model effectiveness.

The following models were implemented for comparison:

- **Logistic Regression (LR)**
- **Decision Tree Classifier**
- **Random Forest Classifier**
- **k-Nearest Neighbors (KNN)**
- **Naive Bayes**
- **Proposed Model**: GCN + Linear SVM (L-SVM) with knowledge graph features

Each model was tuned using 5-fold cross-validation to identify optimal hyperparameters.

Graph Visualizations: Showed high-risk nodes and their relationships to other influential employees in the knowledge graph.

4.2 PERFORMANCE COMPARISON

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	83.7%	81.5%	79.2%	80.3%
Decision Tree	85.2%	83.9%	81.0%	82.4%
Random Forest	88.6%	86.7%	84.3%	85.5%
KNN	81.9%	79.0%	77.6%	78.3%
Naive Bayes	76.1%	74.5%	71.8%	73.1%
Proposed (GCN + L-SVM)	**92.5%**	**90.4%**	**88.2%**	**89.3%**

DISCUSSION

The results confirm that modeling HR data as a graph substantially improves prediction performance. Furthermore, combining this with explainability tools ensures that the model outputs are interpretable and trustworthy. These capabilities are critical in HR environments, where decisions must be fair, accountable, and easily communicated.

.

V. ALGORITHM DESCRIPTION

The proposed model outperformed all baseline methods across every metric, demonstrating the power of incorporating graph-based relationships and advanced feature learning.

MODEL EXPLAINABILITY WITH XAI

To provide transparency in model predictions, SHAP values were computed for the L-SVM output. The most influential features contributing to employee attrition included:

- **Job Satisfaction**
- **Job Involvement**
- **Environment Satisfaction**
- **OverTime**
- **Years at Company**
- **Work-Life Balance**

SHAP summary plots highlighted the directional impact of each feature on prediction outcomes. For instance, lower job satisfaction and frequent overtime were strongly correlated with higher attrition risk.

Additionally, **GNNExplainer** was applied to the GCN to uncover which neighboring nodes (employees) and edges had the most influence on node classification. This provided a deeper view into how peer relationships and departmental structures influenced individual attrition predictions.

VISUALIZATION OF RESULTS

- **Confusion Matrix**: Demonstrated that false positives and false negatives were significantly reduced in the proposed model.

A. **SHAP Summary Plot**: Clearly ranked top contributing features, aiding HR professionals in decision-making.

Algorithm: Employee Attrition Prediction Using Graph Learning and Explainable AI

1) *Goal:*

To detect potential employee attrition by representing HR data as a graph and using a combination of Graph Convolutional Networks (GCNs) and Linear SVM, along with interpretability methods.

2) *Inputs:*

- HR dataset with employee details and attrition labels
- Connection criteria (e.g., same department, similar role or tenure)

3) *Outputs:*

- Prediction of attrition (Yes/No)
- Explanations for each prediction

4) *Steps:*

1. DATA PREPROCESSING

- Convert all categorical fields into numerical format.
- Normalize numeric features such as salary, age, and years at company.
- Divide the dataset into training and test sets (e.g., 80% training, 20% testing).

2. GRAPH CREATION

- Represent each employee as a node in the graph.
- Define edges between employees based on shared attributes like role, department, or performance level.
- Create a weighted graph where edge weights represent how strongly employees are related.

3. GRAPH EMBEDDING WITH GCN

- Initialize node features using employee data.
- Apply a Graph Convolutional Network (GCN) to capture both node-level and neighborhood information. highlights the key driving factors, enabling strategic HR interventions and policy adjustments.
- Generate vector embeddings for each employee that contain richer, context-aware features.

4. ATTRITION CLASSIFICATION WITH LINEAR SVM

- Feed the GCN embeddings into a Linear Support Vector Machine.
- Train the SVM model to classify whether an employee will stay or leave.
- Evaluate predictions on the test data.

5. MODEL EXPLAINABILITY

- Use SHAP values to measure the impact of individual features (like job satisfaction or overtime) on the model's decisions.
- Employ GNNExplainer to highlight which relationships in the graph influenced GCN- based predictions.

6. PERFORMANCE ANALYSIS

- Assess model accuracy and other performance metrics like precision, recall, and F1-score.
- Visualize key patterns and model behavior using confusion matrix and SHAP plots.

Conclusion:

This algorithm combines graph learning and explainability to provide not only accurate attrition predictions but also meaningful insights for HR professionals.

VI. ADVANTAGES AND LIMITATIONS

ADVANTAGES

1. **Captures Relational Dependencies** The use of a knowledge graph allows the model to understand complex relationships between employees based on shared features (e.g., department, job role), which traditional tabular models fail to capture.
2. **Improved Predictive Accuracy** The hybrid model of GCN and L-SVM demonstrated superior accuracy (92.5%) compared to classical ML algorithms, showing its effectiveness in modeling real-world HR data.
3. **Enhanced Interpretability** The integration of XAI tools like SHAP and GNNExplainer provides transparency in predictions, helping HR personnel trust and act upon the model's outputs.
4. **Scalable Feature Learning** GCNs automatically learn high-quality feature representations from both node attributes and graph structure, reducing the need for manual feature engineering.

ACTIONABLE BUSINESS INSIGHTS

The framework not only predicts attrition but also

DISADVANTAGES

Complexity in Graph Construction Transforming structured data into a graph format requires domain knowledge and careful design choices, especially in defining meaningful relationships and edge weights.

SCALABILITY

a step forward in the use of intelligent systems for strategic workforce management.ISSUE GCNs can be computationally expensive for very large graphs, potentially limiting real-time deployment in large-scale enterprises without optimization.

Interpretability of GCN Internals While XAI tools assist in interpreting the outputs, the internal workings of GCNs (i.e., how information is propagated across layers) can still be challenging for non-technical stakeholders.

Data Privacy and Ethics Building a knowledge graph from HR data may raise privacy concerns if not managed properly, especially when modeling sensitive or indirect relationships between employees.

OVERFITTING

With smaller datasets (like the IBM HR dataset with 1,470 records), complex models like GCNs can overfit if not properly regularized or validated.

VII.CONCLUSION AND FUTURE WORK

CONCLUSION

This research presents a novel and effective framework for predicting employee attrition by leveraging the power of graph-based learning through Graph Convolutional Networks (GCNs) and enhancing interpretability with explainable artificial intelligence (XAI) techniques. By converting traditional HR tabular data into a knowledge graph, the model captures deeper relational patterns that are often overlooked by conventional machine learning algorithms.

The proposed hybrid model—combining GCN-derived features with a linear Support Vector Machine (L-SVM)— achieved a notable accuracy of 92.5%, outperforming several established classification methods. Moreover, the integration of SHAP and GNNExplainer provided meaningful insights into the key drivers of employee turnover, such as job satisfaction, overtime, and work-life balance. These insights offer valuable support for data- informed decision-making in HR departments, aiding in the development of proactive retention strategies.

The results of this study demonstrate that incorporating graph-based perspectives and interpretable AI significantly enhances both the predictive power and trustworthiness of attrition forecasting models. This framework thus represents

FUTURE WORK

While the proposed approach has shown promising results, several avenues remain for further exploration:

- **Scalability to Larger Datasets**: Applying the framework to larger and more diverse organizational datasets would help assess its generalizability and robustness across different industries.
- **Dynamic Graph Construction**: Currently, the knowledge graph is static. Incorporating temporal dynamics—such as evolving job roles, promotions, or team changes—could provide a more realistic and adaptive representation.
- **Integration with External Data Sources**: Incorporating external factors such as market trends, economic indicators, and social sentiments could improve the model's contextual understanding of attrition risks.

- **Real-Time Prediction System**: Building a real- time dashboard that integrates the model into existing HR platforms would allow continuous monitoring and decision-making.
- **Ethical Considerations and Bias Mitigation**: Ensuring fairness and removing potential biases in attrition predictions is crucial, especially when using graph structures that may reflect organizational hierarchies or inequalities.

By addressing these areas, future research can further advance the capabilities and applicability of graph-based learning and XAI in human resource analytics.

REFERENCES

[1]. Goyal, P., & Ferrara, E. (2018). Graph embedding techniques, applications, and performance: A survey. Knowledge-Based Systems, 151, 78–94. https://doi.org/10.1016/j.knosys.2018.03.022

[2]. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp.1135– 1144). https://doi.org/10.1145/2939672.2939778

[3]. Kipf, T. N., & Welling, M. (2017). Semi-Supervised Classification with Graph Convolutional Networks. In ICLR. https://arxiv.org/abs/1609.02907

[4]. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems, 30 (NIPS 2017). https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

[5]. Li, C., Xie, W., & Ma, H. (2020). A graph-based machine learning framework for employee attrition prediction. IEEE Access, 8, 38484–38493. https://doi.org/10.1109/ACCESS.2020.2975141

[6]. Holzinger, A., Biemann, C., Pattichis, C. S., & Kell, D. B. (2017). What do we need to build explainable AI systems for the medical domain? arXiv preprint arXiv:1712.09923. https://arxiv.org/abs/1712.09923

[7]. Zhang, Q., Yang, L., Wang, Z., & Li, X. (2021). Graph Neural Networks: A Review of Methods and Applications. AI Open, 2, 57–81. https://doi.org/10.1016/j.aiopen.2021.11.001

[8]. Molnar, C. (2022). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. [Book]. https://christophm.github.io/interpretable-ml-book/