A Semi-Supervised Learning Framework for Employee Performance Evaluation using Group-based Features

1st Seungchan Jeon
dept. of AI & DS
Sejong University
Seoul, South Korea
shermansc310@sju.ac.kr

2nd Jangkyum Kim
dept. of AI & DS
Sejong University
Seoul, South Korea
jk.kim@sejong.ac.kr

3rd Youngdae Ko
dept. of Hotel and Tourism Mgmt
Sejong University
Seoul, South Korea
youngdae.ko@sejong.ac.kr

***Abstract*—In the industry, accurate employee performance evaluations are essential for compensation and workforce planning. However, employee performance datasets generally lack reliable labels. Therefore, we present a semi-supervised learning framework that combines a small set of expert-labeled records with a large pool of unlabeled data. In addition, group-based features are incorporated to ensure fair workload allocation among working groups. Also, statistical smoothing is used to stabilize small groups and weights deviations according to task complexity. Lastly, an iterative high-confidence pseudo-labeling procedure is employed to expand the labeled set, while distributional stability is monitored to prevent imbalance. In numerical studies, the framework enlarged the labeled portion from 4% to 92% without destabilizing the core feature distributions. After semi-supervised learning, the main features had stable means and lower standard deviations. This shows that the framework reduces the burden of expert labeling, enables fairer comparisons through group-based features and statistical smoothing, and can be smoothly integrated into employee performance analytics pipelines.**

***Index Terms*—Employee performance evaluation, Group-based features, Pseudo-labeling, Semi-supervised learning, Workload analysis**

I. Introduction

Performance evaluation is a key element of human resource (HR) management, guiding major organizational decisions such as compensation and promotion. However, in efforts to improve efficiency, the evaluation process is often undermined by systemic issues. Conventional assessments typically rely on narrow metrics (e.g., task counts or attendance) [1]. Moreover, dependence on a single manager's judgment introduces evaluator-induced bias, and such subjectivity often leads to inconsistent and potentially unfair outcomes. Overall, these approaches provide only a partial view of employee performance.

Therefore, data-driven performance evaluation models aim to reduce such subjectivity [2]. However, they face significant challenges in HR contexts due to the scarcity of reliable ground-truth labels for employee performance data [3]. This scarcity hinders model performance and often causes supervised models to overfit. Moreover, many approaches lack feature engineering tailored to organizational structures, resulting in inconsistent predictions across groups.

To overcome these limitations, semi-supervised learning (SSL) that trains models using a small set of labeled data together with a large amount of unlabeled data offers a promising solution. SSL leverages both labeled and unlabeled employee performance data and demonstrated strong results in domain with rich data structures. Methods such as FixMatch [4] and Noisy Student [5] have demonstrated strong performance, and both are highlighted in the comprehensive survey by Van Engelen and Hoos [6]. However, leveraging unlabeled structured data poses several challenges. For instance, generative models like Generative Adversarial Networks (GANs) struggle to learn the complex distributions of tabular data [7]. In this context, while SSL emerges as a promising alternative, it also remains underexplored for structured tabular data (e.g., employee performance records). This can be attributed to the lack of clear and consistent patterns in tabular data, which are essential for common SSL techniques (i.e., augmentation or stable pseudo-labeling).

To overcome these challenges, we propose an SSL framework that combines workload-based group feature engineering with an iterative pseudo-labeling strategy. Our approach adjusts for departmental differences and incorporates work-complexity ratings. The model then expands the labeled set by adding high-confidence predictions. In this study, we propose a tailored SSL methodology designed to build accurate and generalizable models from complex and label-scarce employee performance data.

The main contributions of this works are summarized as follows.

- Group-based feature engineering by designing statistical features are proposed to normalize workload differences across departments, roles, and work complexities. With the proposed feature engineering method, we prove that it is enable to evaluate fairer performance comparisons.
- Through the iterative pseudo-labeling, we show that it is possible to control label noise and mitigate overfitting.

- By analyzing the stability of feature distributions, we track their means and standard deviations over self-training iterations to validate the consistency and reliability of the pseudo-labeling process.

The rest of this paper is organized as follows. Section III presents the proposed performance evaluation framework, including data preparation, group-based feature construction, and the SSL pipeline. Section IV analyzes feature means and standard deviations over iterations to verify the consistency of the pseudo-labeling process. Finally, Section V summarizes the key findings and outlines directions for future research.

II. Nomenclature

For reader's convenience, we represent the symbols and variables in Table I.

TABLE I
NOMENCLATURE FOR SYMBOLS AND VARIABLES

Symbol	Description
w	Workload of the employee
w_t	Global workload-tier classification
d	Department identifier
d_t	Department-specific workload-tier classification
g	Generic group index
$g^{(w)}$	Workload-tier–complexity grouping (w_t, c_w)
$g^{(d)}$	Department workload-tier–complexity grouping (d_t, c_w)
c_w	Work complexity rating
m_l	Manager label (expert-provided performance label)
n_g	Number of employees in group g
μ_g	Mean workload within group g
σ_g	Standard deviation of workload within group g
μ	Global mean workload across all employees
σ	Global standard deviation of workload
λ	Shrinkage intensity parameter (set to 50.0 in this study)
ε	Small constant to avoid division by zero (10^{-8})
μ_g^*	Smoothed mean workload of group g after shrinkage
σ_g^*	Smoothed standard deviation of workload of group g after shrinkage

The above symbols and variables are used throughout the subsequent sections to describe the dataset, group-based features, and the proposed SSL framework.

III. System Model

This section introduces the proposed system model for employee performance evaluation using SSL. The framework addresses the scarcity of labeled employee performance data by combining a small set of expert-labeled records with a large pool of unlabeled records. To ensure fair and consistent evaluation across departments and roles, the system applies workload-based group feature engineering with statistical smoothing, which reduces the influence of group size and role-specific differences.

As illustrated in Figure 1, the process consists of four stages. Section III-A introduces group-based statistical smoothing to stabilize department–role statistics. Section III-B then defines relative performance features that normalize workload differences across organizational groups. Section III-C describes the iterative semi-supervised learning pipeline, which progressively enlarges the labeled dataset through high-confidence pseudo-labeling. Finally, Section IV presents the numerical results, evaluating both the expansion of labeled data and the stability of feature distributions across iterations. Together, these components establish a scalable and robust approach to employee performance evaluation in label-scarce organizational settings.

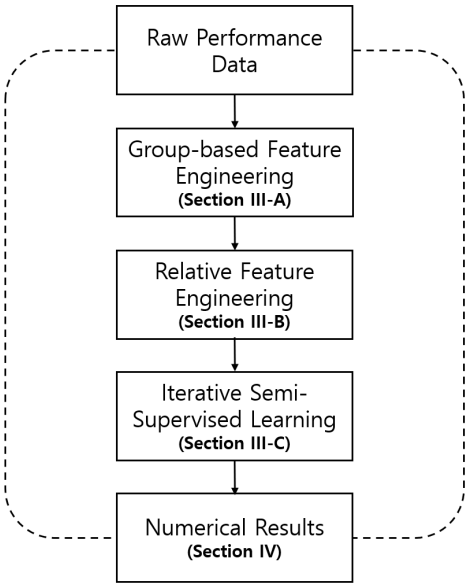

Fig. 1. Overall structure of the proposed system model.

A. Group-based Feature Engineering

Employee performance datasets commonly contain contextual information (e.g., department, role, and job level). While such attributes provide valuable background, this study primarily focuses on the relationship between workload and manager evaluation. To ensure fair comparisons across organizational groups, we construct features derived from workload as an objective measure. Rather than relying on absolute task counts, workload is normalized within department–role groups to capture relative performance. These group-based features establish a consistent foundation for comparison across different organizational contexts.

1) Necessity of Group-Based Smoothing: In employee performance evaluation, workload distributions can vary substantially across departments and roles. For example, employees in Department A with high-intensity roles might regularly handle over 500 tasks per evaluation period. On the other hand, in Department B with more specialized functions might complete fewer than 200 tasks in the same period. Such differences make direct comparison of raw workload values misleading. Moreover, some department–role groups may have very few employees. Small sample sizes yield unstable estimates of the mean and standard deviation, which can introduce noise and drive overfitting to atypical values.

To address this, we apply a smoothing approach that blends each group's statistics with the global statistics via a shrinkage parameter λ. When a group has few samples, its smoothed

value is pulled toward the global value, improving stability; when a group has many samples, the smoothed value preserves the group's unique characteristics. This yields comparable, stable, and meaningful relative measures across organizational groups.

2) Calculation of Smoothed Group Statistics: Let g denote a generic group index (Table I for the two grouping schemes). Given the group size n_g, group mean μ_g, group standard deviation σ_g, global mean μ, and global standard deviation σ, the smoothed estimates are:

$$\mu_g^* = \frac{n_g \mu_g + \lambda \mu}{n_g + \lambda} \tag{1}$$

$$\sigma_g^* = \frac{n_g \sigma_g + \lambda \sigma}{n_g + \lambda} \tag{2}$$

Here, λ controls how much we pull the group statistics toward the global values. In this study, $\lambda = 50.0$ provides a good balance between stability for small groups and specificity for large groups. From these stabilized group statistics, we derive new features that better capture each employee's relative performance, as described in the following subsection III-B.

B. Relative Feature Engineering

Using μ_g^* and σ_g^*, we construct six derived features:

(a) **Relative Workload to Group Expectation**

$$\mathrm{w_{ratio}} = \frac{w}{\mu_g^* + \varepsilon} \tag{3}$$

In equation (3), we shows the employee's workload relative to the group's expected value, removing scale differences between departments/roles and clearly showing whether the workload is above or below expectations.

(b) **Log-Transformed Workload Difference**

$$\mathrm{w_{logdiff}} = \log(1 + w) - \log(1 + \mu_g^*) \tag{4}$$

In equation (4), we show the workload difference from the group mean in log scale, which converts multiplicative gaps into additive ones. The constant 1 prevents undefined values when $w = 0$ and reduces the disproportionate influence of extremely large workloads.

(c) **Workload Difference from Group Mean**

$$\mathrm{w_{diff}} = w - \mu_g^* \tag{5}$$

In equation (5), we show the employee's workload as a difference from the group mean in the original scale. It takes positive values when above the group mean and negative values when below.

(d) **Standardized Workload Difference**

$$\mathrm{w_z} = \frac{w - \mu_g^*}{\sigma_g^*} \tag{6}$$

In equation (6), we show the employee's workload difference standardized by the group's variability, enabling fair comparisons across groups and identifying unusual workload patterns.

(e) **Complexity-Weighted Workload Difference**

$$\mathrm{w_{cdiff}} = \mathrm{w_{diff}} \times c_w \tag{7}$$

In equation (7), we show the employee's workload difference from the group mean weighted by the work-complexity rating, so that the same difference receives greater weight in higher-complexity tasks.

(f) **Complexity-Weighted Standardized Difference**

$$\mathrm{w_{cz}} = \mathrm{w_z} \times c_w \tag{8}$$

In equation (8), we show the employee's standardized workload difference weighted by the work-complexity rating, ensuring that standardized differences are emphasized more strongly for higher-complexity tasks.

TABLE II
SUMMARY OF DERIVED FEATURES

Feature	Purpose
w_{ratio}	Relative workload compared to the group mean; removes department/role scale effects.
$w_{logdiff}$	Log-transformed workload difference; converts multiplicative gaps into additive differences and reduces the impact of extremely large values.
w_{diff}	Workload difference from the group mean in the original scale.
w_z	Standardized workload difference that accounts for the group's variability and enables fair comparisons.
w_{cdiff}	Complexity-weighted workload difference; emphasizes differences more strongly for higher-complexity tasks.
w_{cz}	Complexity-weighted standardized difference; emphasizes standardized differences more strongly for higher-complexity tasks.

The derived features are used as inputs to the SSL framework (Section III-C) with selected base covariates from Table I.

C. Semi-Supervised Learning Framework

1) Problem Setting: The dataset contains only a limited proportion of labeled samples, with about 4% of employee records having a m_l. Such scarcity of labels makes it difficult to train a reliable supervised model without overfitting. To address this issue, we propose a semi-supervised learning (SSL) framework that leverages a small labeled set as a seed and iteratively expands it by assigning pseudo-labels to unlabeled samples. The framework builds on the group-based features described in Section III-A. These features normalize differences across departments and roles, as well as workload tiers and work complexity levels, to provide consistent predictive signals.

2) Input Features: For model training and optimization, we employ a total of 15 input variables. The base covariates are c_w, w , and $\log(w)$. The remaining twelve inputs are group-level means computed under $g^{(w)}$ and $g^{(d)}$ (in Table I and Section III-B). These are calculated for six per-employee features:

$$\{\, \mathrm{w_{ratio}},\ \mathrm{w_{logdiff}},\ \mathrm{w_{diff}},\ \mathrm{w_z},\ \mathrm{w_{cdiff}},\ \mathrm{w_{cz}} \,\}$$

3) Framework Overview: The SSL pipeline consists of three main stages:

(1) *Initial Model Training.* A base classifier is trained exclusively on the labeled subset to establish a reliable baseline. We employ a gradient boosting decision tree model (XGBoost), chosen for its robustness to mixed data types and strong performance on tabular datasets. Hyperparameters are tuned using cross-validation on the initial 440 labeled samples.

(2) *Pseudo-Label Assignment.* The trained model is applied to the unlabeled portion of the dataset to generate probability scores. Samples with prediction confidence above a threshold (initially 0.9, progressively lowered in subsequent iterations, e.g., 0.900, 0.895, 0.890) are assigned pseudo-labels to expand the labeled pool. The threshold is not decreased below 0.87 to preserve label quality and prevent error propagation.

(3) *Iterative Self-Training.* The model is retrained on the augmented labeled set, and the pseudo-labeling process is repeated. Iterations continue until the increase in labeled samples per iteration falls below 0.2% or until validation performance stabilizes.

IV. Numerical Results

We conduct numerical experiments to examine the framework's effectiveness in enlarging limited labeled datasets, ensuring feature distribution stability, and detecting reliable predictions.

A. Evolution of Labeled Samples

Figure 2 presents the progression of labeled and unlabeled employee records across 10 self-training iterations. The process began with 440 manager-labeled records, representing approximately 4% of the dataset. During the initial iterations, the labeled set expanded rapidly as the model incorporated high-confidence pseudo-labels, meeting an initial threshold of 0.900 and gradually relaxing it to 0.870. By the 10 times of iteration, the framework had generated 9,293 pseudo-labeled records from the original 440 labels, yielding a total of 9,733 labeled samples, or about 92% of the dataset. The growth curve shows a clear saturation pattern after the fifth iteration, indicating that early stages capture the most easily classifiable employees, while later stages primarily contain records with greater uncertainty.

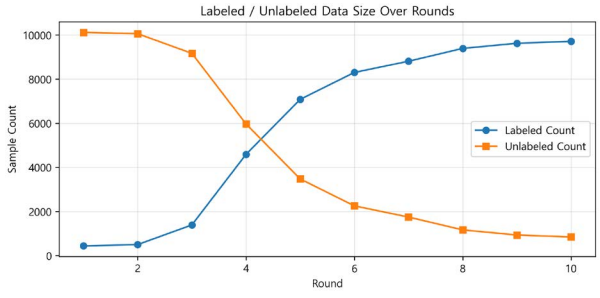

Fig. 2. Changes in the number of labeled and unlabeled employee records over 10 self-training iterations.

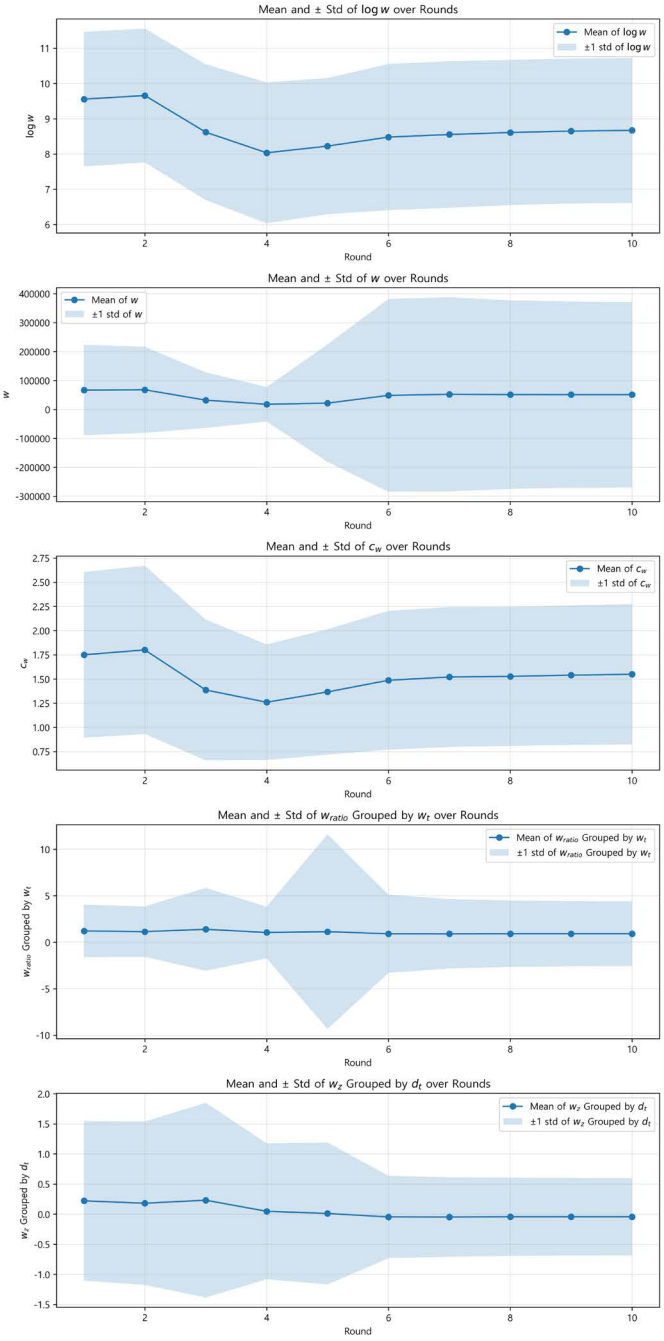

Fig. 3. Mean (dots) and standard deviation (shaded area) of five key features over 10 iterations.

B. Feature Distribution Stability

Figure 3 tracks the mean (dots) and the standard deviation (shaded bands) of five features over ten self-training rounds. This subsection summarizes each subplot and explains how the observed patterns indicate a well-behaved pseudo-labeling process. Each subplot is denoted as (a)–(e) from top to bottom, and their characteristics are described in sequence below.

Figure 3 (a) illustrates the trajectory of $\log w$, where the mean decreases from approximately 9.5 in the first round

to about 8.0 by round 4, then recovers and stabilizes near 8.6 from round 7 onward. The standard deviation remains stable across all rounds, showing a consistent pattern as high-confidence pseudo-labels are incorporated and demonstrating the robustness of the SSL process. Figure 3 (b) presents the raw workload w, whose mean follows a V-shaped path—initially declining and later rebounding—while the standard deviation contracts during rounds 1–4 but expands after round 5. This pattern implies that later pseudo-labeling cycles reintroduced records with larger raw workloads, reflecting controlled variability rather than instability. Figure 3 (c) shows the complexity-weighted workload c_w, with the mean declining from 1.75 to 1.25 before returning and stabilizing near 1.5. The early narrowing and slight widening of the standard deviation after round 5 suggest that the complexity mix stabilized quickly yet preserved some heterogeneity across tasks. Figure 3 (d) depicts the group-normalized ratio w_{ratio}, which remains centered near 1 throughout the rounds. Although a short increase in standard deviation appears at round 5, it decreases in the following round, showing that temporary changes in group ratios are quickly resolved within the SSL framework. Finally, Figure 3 (e) highlights the standardized workload w_z grouped by departments, where the mean consistently stays close to 0, while the standard deviation steadily decreases to about 0.6 by round 10. This pronounced reduction in variability reflects increasing homogeneity in standardized workloads across departments as self-training progresses.

Overall, these results show that the feature centers remain close to their targets ($w_{ratio} \approx 1$, $w_z \approx 0$), early fluctuations are dampened, and no persistent bias appears in raw-scale statistics. These patterns confirm that pseudo-labeling effectively expanded coverage without distorting the core distributions of the most predictive features.

C. Key Takeaways

The labeled set grew from 440 records (4%) to 9,733 records ($\approx$ 92%) by round 10, and the per-iteration gain fell below 0.2% after round 6, which matches the stopping rule. Normalized centers remained at their targets. w_{ratio} stayed near 1 across rounds, w_z hovered around 0 throughout, and $\log w$ converged near 8.6 after round 7. The standard deviation of w_z declined from about 1.4 to about 0.6 by round 10, and the standard deviation of $\log w$ shrank steadily across rounds 1–10. Raw-scale w narrowed early then widened after round 5, which suggests that later cycles reintroduced larger workloads while avoiding sustained bias. w_{ratio} showed one spike in standard deviation at round 5 that disappeared in the next round, and c_w dipped to about 1.25 then returned to about 1.5. Taken together, these results show that pseudo-labeling expanded coverage without distorting core feature distributions, and point to targeted manual labeling or active learning as next steps for role- or department-specific pockets.

V. Conclusion

This paper addressed the challenge of label scarcity and cross-unit heterogeneity in organizational performance evaluation. We introduced a semi-supervised framework that blends workload-tier–aware group feature engineering with iterative high-confidence pseudo-labeling. The approach uses smoothed group statistics to normalize scale across departments and roles, and applies work-complexity weighting to reflect task difficulty. In addition, we monitored feature-distribution stability during self-training to prevent bias and to keep inputs consistent over iterations. The framework is intended to reduce dependence on expert labels while supporting fairer comparisons across organizational units. It is practical for integration into employee performance analytics pipelines on tabular data.

Future work will extend the framework in four directions. First, adaptive confidence schedules and uncertainty-aware selection can improve pseudo-label quality. Second, active learning can be adopted for department- or role-specific performance evaluation, tailored to specific tasks.

Acknowledgment

This work was supported by Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency (KOCCA) grant funded by the Ministry of Culture, Sports and Tourism (MCST) in 2025 (Project Name: Cultivating masters and doctoral experts to lead digital-tech tourism, Project Number: RS-2024-00442006, Contribution Rate: 100%)

References

[1] A. S. DeNisi and K. R. Murphy, "Performance appraisal and performance management: 100 years of progress?" *Journal of Applied Psychology*, vol. 102, no. 3, pp. 421–433, 2017.

[2] P. Budhwar, A. Malik, M. T. De Silva, and H. Thees, "Artificial intelligence-challenges and opportunities for international hrm: a review and research agenda," *The International Journal of Human Resource Management*, vol. 33, no. 6, pp. 1165–1197, 2022.

[3] Z.-H. Zhou, "A brief introduction to weakly supervised learning," *National Science Review*, vol. 5, no. 1, pp. 44–53, 2018.

[4] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 596–608.

[5] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, "Self-training with noisy student improves imagenet classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 687–10 698.

[6] J. E. Van Engelen and H. H. Hoos, "A survey on semi-supervised learning," *Machine Learning*, vol. 109, no. 2, pp. 373–440, 2020.

[7] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling tabular data using conditional gan," *Advances in neural information processing systems*, vol. 32, 2019.