ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 2 - 2026

Development and Application of a Deep Belief Network-Based Performance Evaluation Model for University Faculty

Na Jiang[1], *Guoqing Chen*[2], *Tianwen Zhao*[3] *and Piyapatr Busababodhin*[4,*]

[1] Teacher Education Development Center, Chengdu Jincheng College, Chengdu, 611731, Sichuan, China
[2] Mathematical Modeling Research Center, Chengdu Jincheng College, Chengdu, 611731, Sichuan, China
[3] Department of Trade and Logistics, Daegu Catholic University, Gyeongsan, 38430, Daegu, Republic of Korea
[4] Department of Mathematics, Faculty of Science, Mahasarakham University, Kantarawichai, 44150, Maha Sarakham, Thailand

SUMMARY: *For solving the problems of dispersive data, non-uniform metrics, and not enough explanation of outcomes in yearly performance assessment work of university teachers, a deep belief network model that has missing-value-conscious inputs and double-task outputs was built. According to name-hidden materials from four same-grade undergraduate universities that cover the years 2021 to 2024, this research has collected 1,248 yearly teacher assessment samples. The raw data was systematically organized into five primary dimensions-teaching contributions, research output, student development support, public service, and professional growth-comprising 18 secondary indicators and 64 computable variables. Test results show that the model achieves a Mean Absolute Error (MAE) of 3.31, a Root Mean Square Error (RMSE) of 4.29, an R^2 of 0.895, an Accuracy of 0.861, and a Macro-F1 of 0.832, outperforming Linear Regression, Random Forest, XGBoost, Backpropagation Neural Networks, and the standard DBN.Compared to DBN-base, the RMSE decreased by 12.1%, and the Macro-F1 score increased by 3.4 percentage points. Robustness experimental outcomes show that when the missing data rate is increasing from 0% to 15%, the value of RMSE is risen only from 3.92 to 4.36; the descending of performance is more obvious when the two items, which are research output and student support variables, all encounter high degree of interference at the same time. The outcome shows that this model can give comparatively steady quantification support and a distinct explanation interface for university teacher performance assessment.*

KEYWORDS: *Deep Belief Network; the performance assessment of teachers in universities; data from many sources; study of two tasks together; interpretable assessment*

1 Introduction

The performance assessment of university teachers has direct connection with permanent employment, title promotion, performance-linked salary payment, teaching quality enhancement, and support for teachers' own development. To institutions of higher learning, the working achievement of teachers cannot be comprehended only by means of yearly work

*piyapatr.b@msu.ac.th
https://doi.org/10.65102/is20261010
Published: 30 April, 2026

load calculation figures. It is distributed among teaching work, curriculum making, research outputs, student guiding, department works, public services, and professional improvement. At present, the evaluation systems which exist now normally put together annual report forms, student feedback information, peer checking work, and management side examination. Although this structure already obtains very widespread use, it still exists continuous difficult problems on the aspects of data integration and feedback provision. The work tasks of faculty members are not same in different colleges, and the contribution modes also have differences among teaching-focused, research-focused, practice-based and service positions. Under these situations, evaluation models which are built mainly on even weights and manual gathering tend to squeeze different work into an excessively simple outcome.

Student course appraisals are still broadly applied in teaching evaluation, however their validity and fairness are not steady. Cook and other colleagues demonstrate that such grading methods only catch a portion of the teaching experience, and cannot separate teaching effect from course hardness, student anticipations, or teacher working input [1]. Quansah and other persons we further point out that scoring action, appraisal environment, and measuring mistake reduce their dependability in teacher evaluation [2]. These distorted situations become more influential when evaluation scores are connected to promotion, rewards, or work post reassignment. Daskalopoulou and other researchers also report the influences of gender, self-identity, curriculum type, and student's perception on evaluation results, faculty's pressure, and career developmentt [3]. These restrictions have impelled faculty appraisal toward data-based models that can seize nonlinear connections among multi-source indexes and support more fine-divided evaluation of instruction, research, and development possibility. Alakoum et al. noted regarding the application of AI in university faculty performance evaluation that intelligent evaluation systems can enhance evaluation automation, personalized feedback, and the ability to identify multidimensional contributions [4]. Almufarreh et al. constructed a teaching quality framework and employed machine learning methods to support faculty performance evaluation, demonstrating that educational evaluation data can be used to train models capable of generating computable quality judgments [5]. Almubarak et al. further applied deep learning to identify classroom interactions, proving that evaluation criteria-such as faculty behavior and student engagement-which traditionally relied on manual observation, can be transformed into trainable and verifiable model inputs [6].

Existing research has driven the transition of teacher evaluation from manual, experience-based methods to intelligent assessment; however, several key shortcomings remain in the context of university faculty performance evaluation. First, many models focus primarily on teaching quality or classroom behavior, with insufficient comprehensive modeling of research output, student mentoring, public service, and professional development.Second, some studies rely on fuzzy algorithms, regression models, or single classifiers, which can produce evaluation results but struggle to capture the hierarchical and implicit relationships among indicators. Yang et al. used adaptive fuzzy algorithms to construct a faculty performance evaluation method in the context of higher education reform, enhancing rule expression and fuzzy reasoning capabilities [7]; Qi et al. optimized the design of higher education performance evaluation indicators from the perspective of multi-objective feature regression, improving the modeling accuracy of multi-objective evaluations [8].While these studies provide a valuable model foundation for faculty evaluation, shallow models remain susceptible to limitations imposed by manual weighting and feature selection when performance indicators exhibit characteristics such as multi-source heterogeneity, weak labeling, missing records, and nonlinear interactions.

The problem of whether people can understand the meaning still is a core weak point in the evaluation of college teaching staff members. The higher education appraisal must indicate which quotas formed the outcome, in which places great differences appeared, and how

teaching staff in different positions can make improvement. Ben the Zion et al. the research indicates that AI-supported teaching assessment can make supplement to traditional evaluation and match what students think [9]. However, in the management of faculty members, merely consistency is not sufficient. The results also must give support to cross-department comparison, individual feedback, anomaly review, and year-to-year tracking work. When the attribution of results is not clear, the acceptance of faculty becomes weaker, and the follow-up work of administration thus loses its direction.

DBNs are fit for this situation because they can study hierarchical expression for score forecast and grade sort under restricted samples, high-dimension targets, lost data, and complex feature relation. The materials of teacher side have continuous and classification variables put together, and many connections are not straight line shapes. Teaching working burden, for example, cannot directly correspond to teaching quality, while the connection between research achievement and student guidance is moreover influenced by departmental resources and post type.

Therefore, this research carries out the development of a model which is based on DBN for the evaluation of faculty members. Yearly materials are transformed into multi-dimensional vectors that cover teaching, research, service and professional development, hence the model together carries out prediction of scores and grades. This work gives a united index framework, one missing-considering two-task DBN, and experiment verification through model comparison, stability checks, ablation, and case analysis.

2 Methods

2.1 Data Organization and Indicator Construction

This dataset was constructed by using information that removes personal identifiers from faculty files which come from four undergraduate universities of the same grade, in the time period of 2021 to 2024, it includes academic work, human resource, research administration, student score evaluations, colleague evaluation, and public service work. After we carry out ID and year matching, remove repeated samples, and carry out rule-standardized rearrangement work, 1,248 faculty-year samples are retained by us. Figure 1 shows the sample structure and its linkage to supervisory tasks.

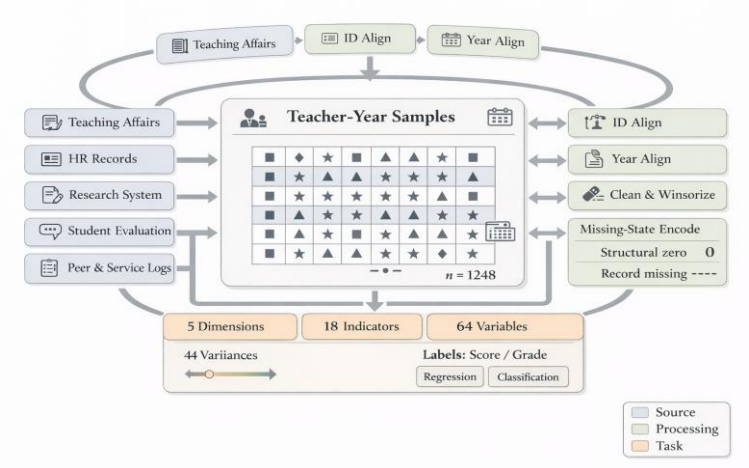


Figure 1: Organization of University Faculty Performance Data and Sample Construction Mechanism

If performance evaluations rely solely on teaching hours, the number of papers, or the average student evaluation score, they tend to reduce teachers with significantly different job responsibilities to the same scale. Based on this, this paper organizes the raw data into five primary dimensions: teaching contributions, research output, support for student development, public service, and professional growth. Under these, 18 secondary indicators are established, which are further expanded into 64 computable variables.This indicator system retains common items from annual university evaluations while incorporating elements-such as course development, competition coaching, academic services, and faculty development-that are often underrepresented in aggregate scores. The organizational structure of the core indicators is shown in Table 1.

Table 1: Teacher Performance Indicator System and Variable Processing Rules

Primary dimension	Secondary indicator	Variable count	Typical variables	Processing rule
Teaching contribution	Teaching load, teaching quality, course construction, teaching reform output	16	Annual teaching hours, student evaluation mean, peer review score, course construction output, teaching award count	Log compression for counts; score mapping to 0-100; weighted conversion for graded awards
Research output	Projects, papers and monographs, patents and transfer, funding	18	Project count, indexed papers, monograph contribution, patent output, research funding	Log compression for counts; comparable-value standardization for funding; weighted conversion by output level
Student support	Academic advising, thesis supervision, competition mentoring, student feedback	10	Thesis supervision load, competition mentoring results, advising frequency, student support score	Standardization for counts; unified scoring for feedback and outcomes
Public service	Internal governance service, academic community service, social service	8	Committee service, peer-review service, outreach activity, industry collaboration	Joint encoding of frequency and service level; duplicate-event removal
Professional growth	Training, promotion, honors, international and peer development	12	Training hours, promotion progress, honor count, visiting activity, academic exchange	Standardization for continuous items; intensity coding for stage-based events

Variables of different types are processed hierarchically. Count-based metrics-such as teaching hours, number of projects, number of papers, and number of students supervised-retain their original scale and undergo logarithmic compression; ratio and score-based metrics-such as average student evaluation scores, average peer review scores, and course objective

achievement rates-are uniformly mapped to the 0-100 range;For indicators showing significant scale differences across different schools, robust centralization is first performed at the "university-school-year" three-tier level, followed by a unified standardization process. Outliers are not directly removed but are truncated at the 1st and 99th percentiles to mitigate the influence of a few exceptionally high values on model weights.

Missing values are handled separately based on their business attributes. The first category is structural missing data, such as when some faculty members do not supervise graduate students, engage in international collaboration, or participate in industry-sponsored projects. Two types of missing-data states we have carried out the distinction. Activities that do not fit the institutional rules were encoded as effective zero values. Recording-connected missing items-like non-returned colleague reviews, not-complete business registration, or not-synchronized past records-were kept as missed, and one covering vector was created in the same time. This processing method, at the input stage, separates the condition of "no task is assigned" from the condition of "task has not been recorded". The supervised labels are obtained from the annual institutional evaluation outcomes. Because the rules of scoring were not same among every school, the original results were first projected onto a unified 0-100 scale, and then were divided into four grades: A ($\geq$85), B (75-84.99), C (60-74.99), and D ($<$60). The final sample distribution is 18.3% for Grade A, 41.8% for Grade B, 30.1% for Grade C, and 9.8% for Grade D. This distribution preserves the fundamental characteristic of the annual evaluation-a concentration in the middle range with relatively fewer schools at the extremes-and provides a supervised foundation for subsequently conducting score prediction and grade classification in parallel.

2.2 DBN-based Performance Representation and Evaluation Model

There are clear hierarchical relationships among faculty performance variables. Teaching workload and teaching quality do not always change in tandem, and the contribution of research output to overall performance is further influenced by position type, departmental tasks, and service workload. To extract stable representations from multi-source, heterogeneous variables, this paper adopts a Deep Belief Network (DBN) as the core model and introduces missing-value-aware inputs, feature group balancing constraints, and a dual-task output structure based on the standard DBN.The DBN architecture used in this paper, along with its missing-value-aware inputs and dual-task output relationships, is shown in Figure 2.

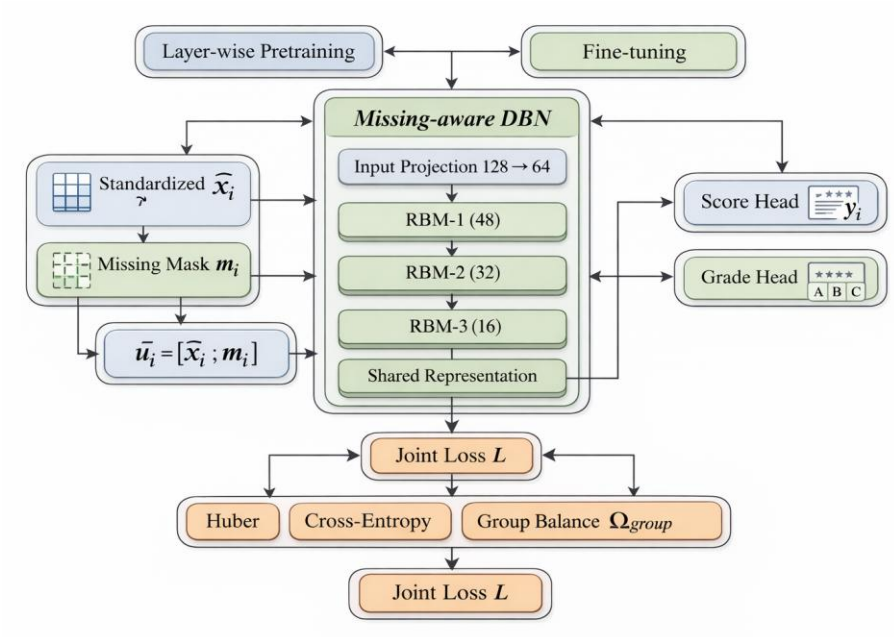


Figure 2: DBN Performance Representation and Dual-Task Output Mechanism

The sample input consists of a standardized feature vector and a missingness mask. Let the original features of the annual sample for the th faculty member i be x_{ij} , and let the mean and standard deviation of the th variable j be μ_j and σ_j , respectively. Then, the standardized input is given by Equation (1).

$$\tilde{x}_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j + \varepsilon}, \qquad u_i = [\tilde{x}_i; m_i] \tag{1}$$

In the equation, $\tilde{x}_{ij}$ represents the standardized j th variable, m_i is the missing mask vector for sample i, ε is a small constant to prevent the denominator from becoming zero, and u_i is the final input representation.The original performance variable in this paper is 64-dimensional, and the mask vector is also 64-dimensional; when concatenated, they form a 128-dimensional initial input. To avoid training instability caused by high-dimensional sparse inputs directly entering the RBM, this paper first compresses them to 64 dimensions using a single layer of linear mapping, and then feeds them into a three-layer DBN backbone network.

The DBN consists of stacked RBMs. The first layer uses a Gaussian-Bernoulli RBM to process continuous, standardized inputs, while the upper layers use a Bernoulli-Bernoulli RBM to learn latent representations. For any given RBM layer, its energy function is shown in Equation (2).

$$E(v, h) = -a^{\top} v - b^{\top} h - v^{\top} W h \tag{2}$$

Here, v denotes the visible layer variables, h denotes the hidden layer variables, W denotes the inter-layer weight matrix, and a and b denote the visible and hidden layer biases, respectively.This paper employs layer-wise unsupervised pre-training to initialize network parameters, followed by a supervised fine-tuning stage, to mitigate the issue of unstable updates that arises during direct end-to-end training under moderate sample sizes [10-14]. Given the input to the visible layer, the hidden layer activation probability and the visible layer reconstruction probability of the RBM are expressed in Equations (3) and (4), respectively.

$$p(h_j = 1|v) = \sigma\left(b_j + \sum_i W_{ij}\, v_i\right) \tag{3}$$

$$p(v_i = 1|h) = \sigma\left(a_i + \sum_j W_{ij}\, h_j\right) \tag{4}$$

Here, $\sigma(\cdot)$ denotes the Sigmoid function, and W_{ij} represents the connection weight between the i th visible unit and the j th hidden unit [15-17]. In this paper, the sizes of the three hidden layers are set to 48, 32, and 16, respectively. After layer-wise pre-training, the 16-dimensional representation from the top layer is fed into a shared discriminator layer, which then splits into a performance score regression head and a rating classification head.The former outputs continuous performance scores $\hat{y}_i$, while the latter outputs a probability vector indicating the probability of a sample belonging to one of four grades $\hat{p}_i$. A single regression objective tends to bias the model toward the middle range, whereas a single classification objective loses the distinction between continuous scores. To balance both types of information, this paper employs a joint loss to train the shared representation layer. The overall objective function is shown in Equation (5).

$$\mathcal{L} = \lambda_r \frac{1}{N}\sum_{i=1}^{N} \ell_{\text{Huber}}\left(\hat{y}_i, y_i\right) + \lambda_c \frac{1}{N}\sum_{i=1}^{N} \text{CE}\left(\hat{p}_i, c_i\right) + \lambda_g \Omega_{\text{group}} + \lambda_w \|\Theta\|_2^2 \tag{5}$$

where N is the number of training samples,y_i is the true performance score, c_i is the true grade label,$\ell_{\text{Huber}}(\cdot)$ denotes the Huber loss for the regression head, $\text{CE}(\cdot)$ denotes the cross-entropy loss for the classification head, Ω_{group} is the feature set balancing term, Θ is the total number of trainable parameters, and λ_r, λ_c, λ_g, and λ_w are the loss weights.In this paper, we set $\lambda_r = 1.0$, $\lambda_c = 0.6$, $\lambda_g = 0.2$, $\lambda_w = 10^{-4}$. The feature group balancing term is used to prevent the contributions of the five feature categories-teaching, research, student support, services, and development-from overly concentrating on a few high-variance metrics in the hidden layer representations, thereby reducing the dominance of a single dimension on the overall score.

The training work of the model possesses two stages. Each layer of RBM is first carried out pre-training for 60 epochs, with batch size 32, learning rate 0.01, and one-step contrastive divergence method being used. Then, this network we fine-tune by Adam at 1×10^{-3}, reaching a maximum of 120 epochs, whose batch size is 32, and we perform early stopping after 12 epochs that have no improvement. For the purpose of decreasing the influences brought by class imbalance, the classification loss is given weights through the inverse frequency of each class. After the shared layer, a dropout rate which is 0.2 is applied.

Compared with the direct use of multi-layer perceptron, the differences of this model are mainly manifested in three aspects. First, the input keeps the information of missing states, therefore it enables the model to make a distinction between the two kinds of missing values: "has not occurred" and "has not been recorded". Secondly, layer-by-layer pre-training can give more stable initial expression vectors, hence making it suitable for such tasks as teacher work performance assessment, which include strong relevance among variables and middle-sized sample quantities. Third, the sharing expression layer between score return and order classification keeps continuous performance differences while it increases the identifiability of order boundaries.

2.3 Experimental Protocol and Evaluation Metrics

To evaluate the model's applicability in predicting performance scores and classifying grades, this study employed stratified sampling based on performance grades and college types, dividing the entire dataset into training, validation, and test sets in the proportions of 70%, 15%, and 15%, respectively. The test set contains 187 annual faculty performance samples and is used solely for reporting final results; the remaining samples are used for parameter selection, model training, and validation. In addition to the fixed partitioning, this study conducts five sets of experiments with different random seeds to mitigate random fluctuations caused by a single partitioning. The corresponding relation among the experimental work flow, contrast models and assessment norms is displayed in Figure 3.

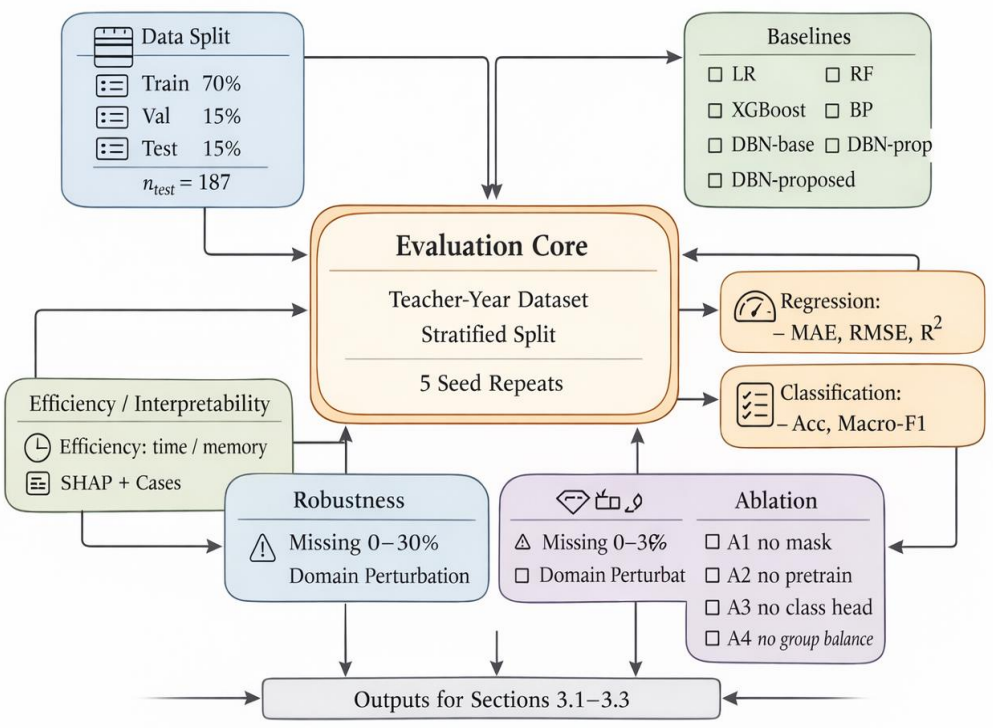


Figure 3: Experimental Setup, Comparison Strategy, and Evaluation Protocol

The comparison models cover three categories of methods: statistical learning, ensemble learning, and neural networks. Specifically, they include Linear Regression, Random Forest, XGBoost, BP Neural Network, DBN-base, and the DBN-proposed model introduced in this paper. Among these, DBN-base retains only the standard DBN backbone and a single regression output, without using missing value masking, feature group balancing, or the dual-task head structure, to verify the actual contributions of each improvement.The number of trees in the Random Forest is set to 300; the maximum depth of XGBoost is set to 6, with a learning rate of 0.05; the BP network adopts a three-layer fully connected structure with hidden layer dimensions of 128, 64, and 32; for the remaining models, optimal hyperparameters are selected on the validation set.

Prediction performance was evaluated using MAE, RMSE, and R^2, while classification performance was assessed using Accuracy and Macro-F1. Robustness testing was divided into two categories: one involved repeatedly applying random masks at missing rate levels of 0%, 5%, 10%, 15%, 20%, and 30% to observe the rate of model degradation;the other involves creating targeted missing values in specific indicator domains to compare performance changes

when key dimensions-such as teaching contributions, research output, and student support-are missing. Efficiency analysis records training time, per-sample inference time, and video memory usage.

To identify the roles of each module in the model, this paper conducts four sets of ablation experiments. A1 removes the missing value masking and retains only standardized performance variables; A2 removes RBM pre-training and replaces it with direct supervised fine-tuning after random initialization; A3 removes the binary classification head and retains only the score regression task; A4 removes the feature group balancing term to observe whether the hidden layer representations are more susceptible to being driven by high-intensity research metrics.All ablation experiments maintain the same number of training epochs, batch size, and optimizer settings.

SHAP was used for interpretability analysis to decompose metric contributions on representative samples from the test set, and error comparisons were conducted across three categories of teachers: teaching-oriented, research-oriented, and balanced-development-oriented. The experimental environment consisted of Python 3.11, PyTorch 2.2, and CUDA 12.1, with hardware configuration including an Intel i7 processor, 32 GB of RAM, and an NVIDIA RTX 4070 graphics card.

3 Results and Discussion

3.1 Overall Evaluation Performance and Baseline Comparison

Whether a model possesses practical usefulness is primarily determined by its simultaneous behaving on both continuous score giving and grade categorization. The whole outcomes of diverse methods upon the testing set are displayed in Table 2.

Table 2: Performance comparison of different models in score prediction and grade classification

Model	MAE	RMSE	R^2	Accuracy	Macro-F1
Linear Regression	4.82	6.21	0.781	0.748	0.701
Random Forest	4.36	5.67	0.817	0.781	0.742
XGBoost	4.08	5.31	0.839	0.802	0.766
BP Neural Network	3.98	5.19	0.846	0.812	0.775
DBN-base	3.73	4.88	0.864	0.832	0.798
DBN-proposed	3.31	4.29	0.895	0.861	0.832

In Table 2, DBN-proposed achieves the best results across all five core metrics, with an MAE of 3.31, an RMSE of 4.29, an R^2 of 0.895, an accuracy of 0.861, and a Macro-F1 of 0.832.Compared to the standard DBN, its MAE decreased by 11.3%, RMSE decreased by 12.1%, Accuracy improved by 2.9 percentage points, and Macro-F1 improved by 3.4 percentage points; compared to XGBoost, which performed most closely, RMSE further decreased by 19.2%, and Macro-F1 improved by 6.6 percentage points.These results indicate that the nonlinear relationships and missing values in the teacher performance data were not fully utilized by traditional ensemble models, whereas the DBN with missing value awareness and dual-task outputs can transform this information into more stable decision boundaries. The score fitting and calibration relationships are shown in Figure 4.

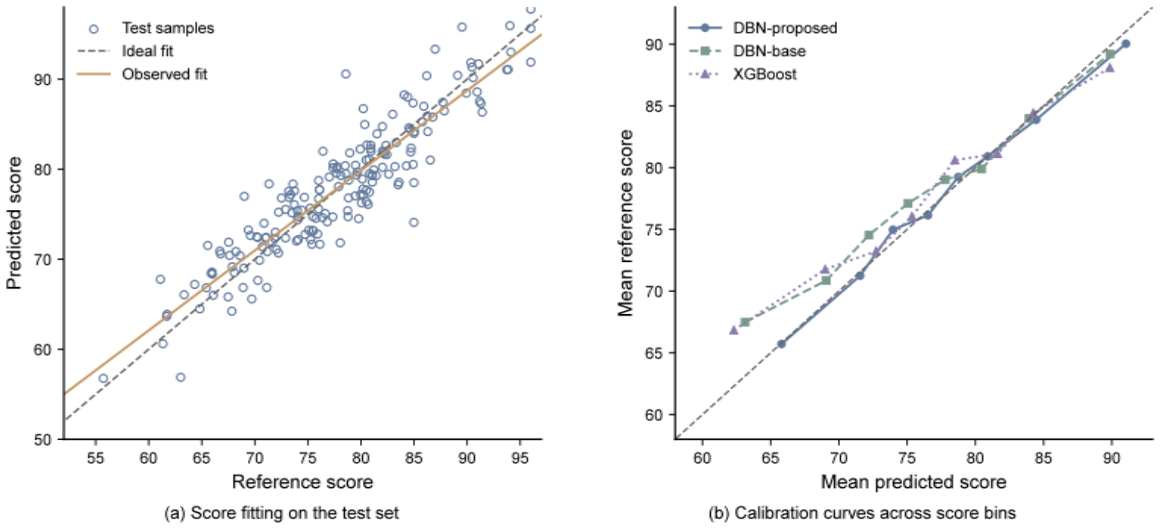


Figure 4: Score fitting and calibration performance

From Figure 4(a), we can know that most of the test samples are located beside the ideal diagonal line, hence the fitting line only possesses a tiny deviation from the reference line. The dispersion still remains under control in both the high-score and the mid-to-high-score areas, hence this indicates that the decreasing of mean error did not come at the expense of the fluctuation of scores. Figure 4(b) furthermore provide demonstration that our proposed DBN possesses better calibration effect than the baseline DBN and XGBoost in the interval of 65 to 90. This advantage is most obviously manifested in the 75–85 interval, where faculty samples are most concentrated together and the curve maintains stable without obvious systematic overestimation or underestimation. In the assessment of teachers, this calibration feature is extremely important hence outcomes are utilized for grade classification and arrangement; a continuous change of the central score range would directly affect excellent award and performance-related resource allocation. Figure 5 provides the error distributions which are present between different sorts of colleges.

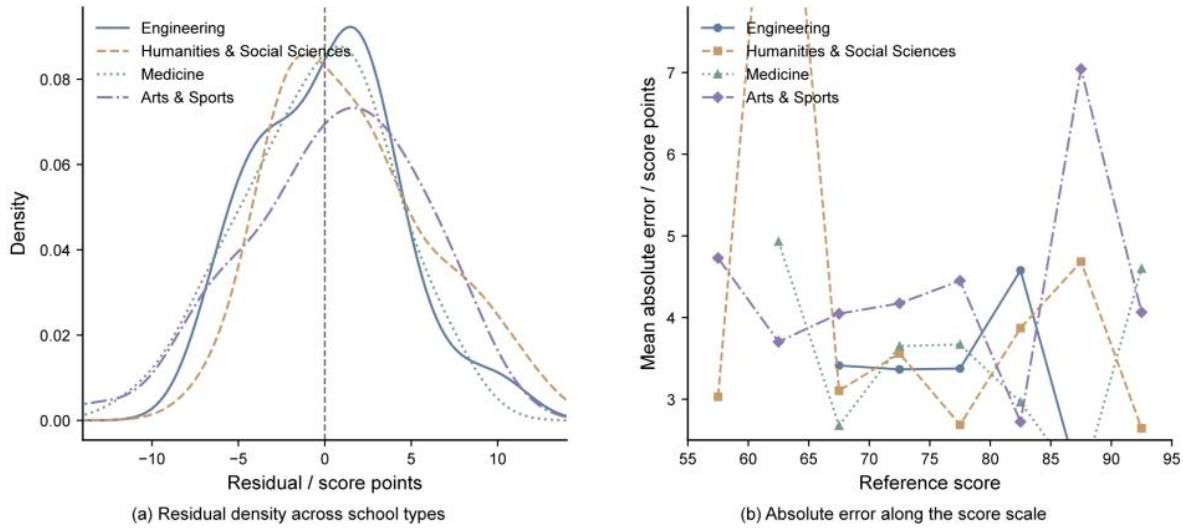


Figure 5: Residual distribution by school type

Figure 5(a) displays that residual concentration of Engineering and Medicine is tighter

around zero, Humanities & Social Sciences and Arts & Sports have shown wider distributions and a more clear right tail, this hence indicates that overestimation occurs more frequently. The average absolute errors are respectively 3.26, 3.38, 3.68 and 4.06. Figure 5(b) further displays that Arts & Sports have stronger fluctuation of middle-to-high scores, and Humanities & Social Sciences have a partial low-score error that is increased. This difference cannot be explained only by the size of the sample. The records of engineering and medicine related domains are usually more standardized, while the accomplishments in humanities, arts and sports are more heterogeneous and more difficult to uniformly quantify, especially when it comes to curriculum formulation, exhibitions, contests, public extension, and item services. Therefore, the model on the whole maintains high precision, but the deployment across schools still needs discipline-related calibration and artificial examination.

3.2 Scenario Robustness, Ablation, and Efficiency Analysis

After the overall accuracy has gotten been confirmed, hence it is necessary that we judge whether the model still keeps its stability under the situations of incomplete data and structural decreases. The connection among data missing situation, metric area changes, and model stability is shown in the Figure 6.

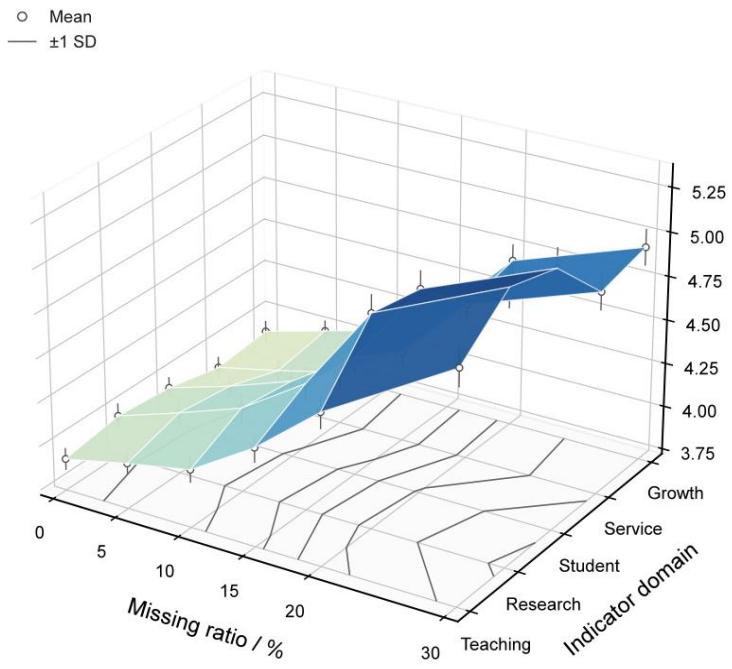


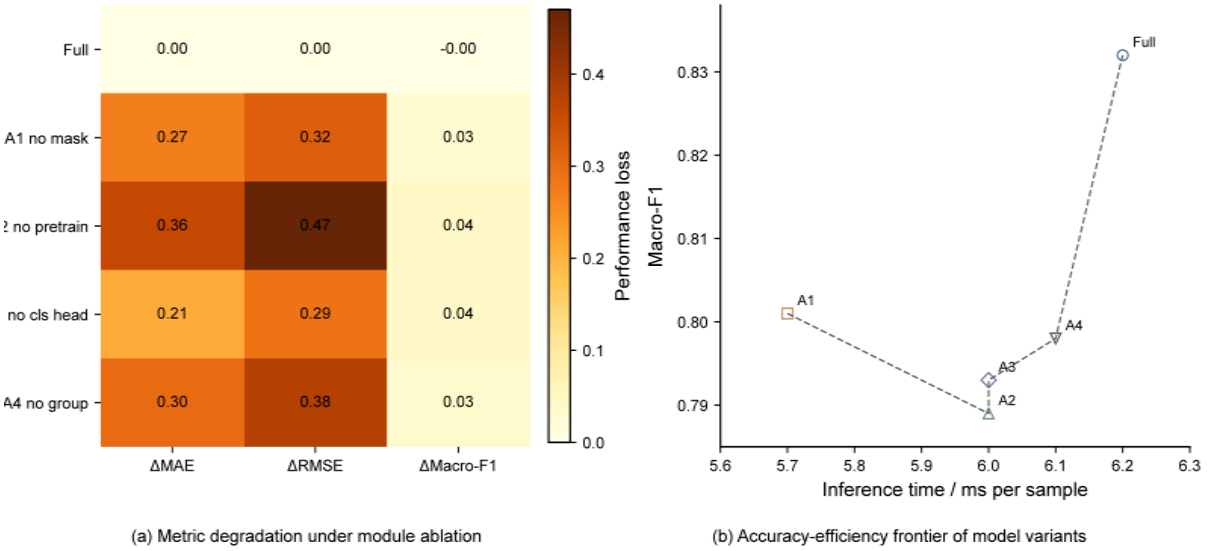
Figure 6: Three-dimensional respond surface of robustness under missing data and domain perturbation

Figure 6 shows the three-dimensional response curved surface of missing ratio, feature region, and RMSE, and Macro-F1 decrease contour lines are projected on the bottom. When the degree of missing data go up from 0% to 15%, the average value of RMSE increase from 3.92 to 4.36, meanwhile the value of Macro-F1 have a reduction of 2.18 percentage points. The surface still keeps comparatively level in this section, hence it shows that missing-value covering and layered expression can ease middle-degree data losing. When the missing degree arrives at 20% and 30%, RMSE further goes up to 4.72 and 5.05, hence the performance declination becomes clearly more quick.

The sensitivity has differences among different feature domains. When missingness is at 30 percent, the value of RMSE achieves 5.26 for Research output and 5.18 for Student support, therefore both of these are higher than Teaching contribution (5.02) and Public service (4.87). This kind of mode shows that indicators for research and student helping have stronger capability to distinguish near the grade dividing lines. When these variables do not exist, the model is still able to produce scores, but the stability of boundaries decreases, especially within the A/B and B/C transition intervals. Table 3 has reported the module ablation and efficiency results, and Figure 7 provides the corresponding visualization content.

Table 3: Ablation, efficiency, and resource consumption of model variants

Variant	MAE	RMSE	Macro-F1	Train time / min	Inference / ms	Memory / GB
DBN-proposed	3.31	4.29	0.832	8.6	6.2	1.48
A1: no missing mask	3.58	4.61	0.801	7.8	5.7	1.42
A2: no RBM pretraining	3.67	4.76	0.789	7.2	6.0	1.46
A3: no classification head	3.52	4.58	0.793	8.1	6.0	1.47
A4: no group balance	3.61	4.67	0.798	8.4	6.1	1.48



(a) Metric degradation under module ablation

(b) Accuracy-efficiency frontier of model variants

Figure 7: Ablation effects and efficiency frontier

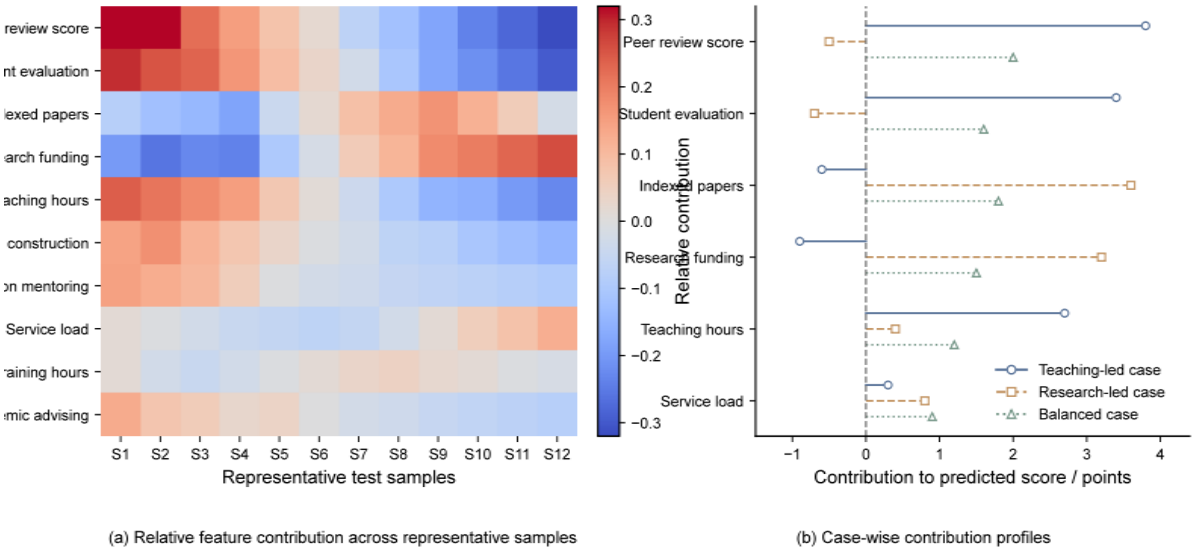
Figure 7(a) gives a summary of the loss of each ablation in one heatmap. Removing RBM pre-training produces the largest decline: RMSE rises from 4.29 to 4.76, up 11.0%, while Macro-F1 drops by 4.3 percentage points. This suggests that layer-wise pre-training still improves initialization under moderate sample size and strong variable coupling. Removing the missing mask raises RMSE to 4.61, indicating weaker distinction between true absence and missing record. After removal of the classification head, Macro-F1 falls to 0.793, which shows that the grade task still supports the shared representation. Removing the feature-group balancing term increases both MAE and RMSE, indicating that over-dominance of high-variance research features weakens separation of teaching and service contribution.

Figure 7(b) makes a comparison between performance and computational cost. The comprehensive model records 6.2 ms inference time for each single sample, 8.6 min training

time, and 1.48 GB GPU memory usage. Comparing with the ablation variant types, the inference time only has an increase of 0.1-0.5 ms, therefore error is lower and Macro-F1 is higher. To the offline batch evaluation which is like annual faculty assessment, this cost still keeps acceptable. The observed increment hence mainly originates from more forceful representation study instead of additional deployment load.

3.3 Error diagnosis, case interpretation, and application implications

Besides accuracy and robustness, the performance model must solve two additional practical problems: which indicators it mainly depends on to give judgments, and how these results are combined into the organization's evaluation flow. The contribution of each feature and the diagnostic outcome of each case are displayed in Figure 8.



(a) Relative feature contribution across representative samples

(b) Case-wise contribution profiles

Figure 8: Feature contribution and case diagnostics

Figure 8(a) displays a heatmap of relative contributions across a representative test sample. It can be seen that Peer review score, Student evaluation, Indexed papers, and Research funding are the most stable high-contribution features, though they do not vary in the same direction.Samples with a clear teaching advantage are typically supported by classroom evaluations, peer reviews, and teaching workload; samples with a research advantage are driven more by publications, funding, and project records. The model does not reduce all high-scoring samples to a single "research-led" or "teaching-led" category, but instead retains multiple performance combination pathways in the latent layers.

Figure 8(b) presents the contribution profiles of three typical faculty categories. In the Teaching-led case, Peer review score, Student evaluation, and Teaching hours contribute 3.8, 3.4, and 2.7 score points, respectively, while Indexed papers and Research funding act as negative offset terms; the Research-led case exhibits the opposite pattern, with papers and funding contributing 3.6 and 3.2 score points, respectively;In the Balanced case, the distribution of contributions is more even, with most features falling between 0.9 and 2.0 score points. This outcome shows that the model does not get limited to the linear weight rules that are buried in current organization patterns. It maintains the whole score steadiness while it separates the explanation roads which are connected to different performance structures. Large-error situations are gathered together among teachers having extremely uneven contribution situations and places close to grade dividing lines, especially those with powerful research work

but insufficient teaching support, or powerful teaching with few research and service materials. These samples, when compared with being targets of direct automated grading, are more suitably handled as outliers for manual review. Figure 9 is what displays the corresponding closed-loop application.

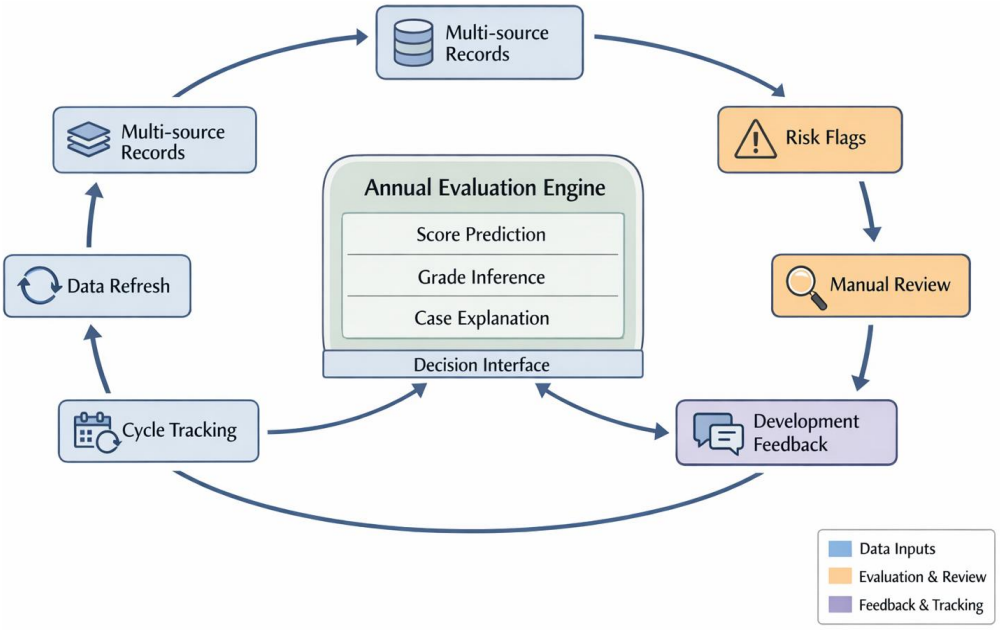


Figure 9: Closed-loop deployment pathway for annual faculty evaluation

Figure 9 puts the model into the annual evaluation work flow, and forms a closed cycle of data updating, model evaluation, abnormal checking, development feedback, and circle tracking. This model is not able to take the place of the institutional policy; it makes the separated performance records become a verifiable, interpretable, trackable analyzing interface that can be checked. For the management personnel, it marks out high-residual and boundary cases before the artificial check. As for faculty members, this method finds out specific deviations in teaching, research, and service, instead of only giving back a total score. Hence, evaluation has a shift which is from one-round confirmation to follow-up that is linked with feedback through all cycles.

4 Conclusion

This research carries out development and carries out verification for a DBN-founded model on yearly teacher work accomplishment assessment. By utilizing identity-removed documents from four equal-level undergraduate universities in the period of 2021–2024, it has constructed 1,248 samples of faculty per year and merged the affairs of teaching, personnel resources, management of research, appraisals by students, and service documents into one united evaluation space. The experiments concerning score forecasting and rank division prove that the steady distinguishing ability exists among different performance structures.

(1) Annual faculty data were reconstructed at the faculty member-year level. The records were reorganized into five dimensions-teaching, research, student support, public service, and professional growth-covering 18 secondary indicators and 64 computable variables. Structural missingness was separated from record-level missingness, which improves input stability and preserves role-specific contribution patterns.

(2) The model achieved the best overall results on the test set, with an MAE of 3.31, RMSE of 4.29, R^2 of 0.895, accuracy of 0.861, and Macro-F1 of 0.832, outperforming Linear Regression, Random Forest, XGBoost, BP Neural Network, and standard DBN. Relative to DBN-base, RMSE decreased by 12.1% and Macro-F1 increased by 3.4 percentage points. Ablation results show that removing RBM pre-training raises RMSE to 4.76, while removing the missing mask raises it to 4.61. Under a missing rate increase from 0% to 15%, RMSE changes from 3.92 to 4.36, although research-output and student-support variables produce larger boundary errors under heavier perturbation.

(3) The value that the model possesses also lies in the interpretation of obtained results. Peer review, student course evaluation, publication output, and research funding still keep relatively stable high-contribution variables, hence high scores are achieved via different contribution structures. Obvious distinctions exist between faculty who focus on teaching, faculty who focus on research, and faculty who pursue balanced development. The present research is still restricted to yearly de-identified data and cannot completely consider institutional difference, work-type rule alterations, or cross-period renewals. In the future, research work may extend this framework by adding longitudinal tracking, job-type adaptive correction, and more lightweight deployment schemes.

Funding

This research was financially supported by Mahasarakham University; Chengdu Jinjiang College Teacher Education Development Center; Sichuan Provincial Key Research Base for Social Sciences, Sichuan Provincial Education Development Research Center (No. CJF20011); 2025 Doctoral Special Support Program Project of Chengdu Jincheng College (NO. 2025JCKY(B)0018).

About the Author

Na Jiang is a lecturer in higher education resource allocation and smart education. Na Jiang obtained a master's degree in Education from China West Normal University and a bachelor's degree from Shandong Normal University. In 2021, she pursued a doctoral degree in Educational Management at the School of Management, Universiti Sains Malaysia. In 2007, she joined the General Education College of Chengdu Jincheng College to conduct research in higher education and served as a lecturer. She has authored numerous papers in the fields of higher education resource allocation, smart education technology and application, and educational management and policy. Her research interests also include the integration of general education and smart technologies, as well as sustainable development education.

Guoqing Chen is a lecturer in educational statistics and mathematical modeling. Guoqing Chen obtained a Ph.D. in Statistics from Mahasarakham University and a master's degree from Xihua University. In 2019, he joined the Mathematical Modeling Research Center of Chengdu Jincheng College to engage in educational statistics research and served as a lecturer. He has authored numerous papers in fields such as applied statistics, mathematical modeling, and smart education. His research interests also include educational economics and emotional labor, innovation in teaching models, digital teaching technology, smart education, etc.

Tianwen Zhao is currently a combined Master's and Ph.D. student in the Department of Trade and Logistics at Daegu Catholic University, South Korea. He holds a Bachelor's degree in Accounting from Chengdu Jincheng College, where he received honors such as the Chinese National Scholarship and the President's Special Scholarship. He has published multiple papers

as the first author or corresponding author in several internationally renowned SCIE/EI/ SCOPUS-indexed journals and conferences. Actively involved in academic service, he serves as a reviewer for several high-level SCI journals, including Renewable Energy (IF: 9.1, SCIE Q1 TOP), Expert Systems with Applications (IF: 7.8, SCIE Q1 TOP), and Applied Soft Computing (IF: 6.9, SCIE Q1 TOP), completing an impressive 195 manuscript reviews within just 1 year. His research interests include trade forecasting, artificial intelligence, applied science, econometrics, and other interdisciplinary studies.

Piyapatr Busababodhin is currently serving as Vice Dean of the Faculty of Science and holds the position of Associate Professor in the Department of Mathematics at Mahasarakham University. With a dedicated career spanning over 25 years since he joined the university in April 1998, he has established a distinguished record in both teaching and research. His scholarly work has garnered a total of 410 citations, with an h-index of 12 and an i10-index of 14, reflecting his significant academic influence. He completed his Bachelor's degree in Mathematics at Mahasarakham University, followed by a Master's degree in Applied Statistics at Thammasat University, and earned his Ph.D. in Applied Statistics from King Mongkut's University of Technology North Bangkok. To further deepen his expertise, he undertook postdoctoral research in Applied Statistics at Chonnam National University in South Korea. His research focuses on cutting-edge areas of applied statistics, including extreme value analysis, copula modeling, statistical modeling, hydrological analysis, and statistical process control (SPC).

References

[1] Cook, S., Watson, D., and Webb, R. (2024). Performance evaluation in teaching: Dissecting student evaluations in higher education. Studies in Educational Evaluation, 81, 101342.

[2] Quansah, F., Cobbinah, A., Asamoah-Gyimah, K., et al. (2024). Validity of student evaluation of teaching in higher education: A systematic review. Frontiers in Education, 9, 1329734.

[3] Daskalopoulou, A., Fox, A., Brookley, S., et al. (2024). Understanding the impact of biased student evaluations of teaching on academics' mental health and career progression. Studies in Higher Education.

[4] Alakoum, A., Nica, E., and Abiad, M. (2024). Revolutionizing faculty performance evaluation: The future role of AI in higher education. Journal of Self-Governance and Management Economics, 12(1), 25-49.

[5] Almufarreh, A., Noaman, K. M., and Saeed, M. N. (2023). Academic teaching quality framework and performance evaluation using machine learning. Applied Sciences, 13(5), 3121.

[6] Almubarak, A., Alhalabi, W., Albidewi, I., et al. (2025). An AI-powered framework for assessing teacher performance in classroom interactions: A deep learning approach. Frontiers in Artificial Intelligence, 8, 1553051.

[7] Yang, D., Malik, M. R. A., Abdullah, J. M. A., et al. (2025). Design of performance evaluation method for higher education reform based on adaptive fuzzy algorithm. PeerJ Computer Science, 11, e3090.

[8] Qi, F., Liu, Z., Zhang, W., et al. (2025). Optimization of multi-objective feature regression models for designing performance assessment methods in college and university educational reform. PeerJ Computer Science, 11, e2883.

[9] Ben Zion, Y., Yakov, S., Abramovitch, E., et al. (2025). AI-based teaching evaluations: How well do they reflect student perceptions? Computers and Education: Artificial Intelligence, 9, 100448.

[10] Hinton, G. E., Osindero, S., and Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. Neural Computation, 18(7), 1527-1554.

[11] Hinton, G. E., and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. Science, 313(5786), 504-507.

[12] Le Roux, N., and Bengio, Y. (2008). Representational power of restricted Boltzmann machines and deep belief networks. Neural Computation, 20(6), 1631-1649.

[13] Zhang, N., Ding, S., Zhang, J., et al. (2018). An overview on restricted Boltzmann machines. Neurocomputing, 275, 1186-1199.

[14] Zambra, M., Testolin, A., and Zorzi, M. (2023). A developmental approach for training deep belief networks. Cognitive Computation, 15, 103-120.

[15] Bai, A., and Hira, S. (2021). An intelligent hybrid deep belief network model for predicting students employability. Soft Computing, 25(14), 9241-9254.

[16] Kamakshamma, V., and Bharati, K. F. (2023). Adaptive-CSSA: Adaptive-chicken squirrel search algorithm driven deep belief network for student stress-level and drop out prediction with MapReduce framework. Social Network Analysis and Mining, 13, 90.

[17] Tang, B., Li, S., and Zhao, C. (2024). Predicting the performance of students using deep ensemble learning. Journal of Intelligence, 12(12), 124.