AI Journal of Business

Volume 2 Number 2 Jun 2026

AI Business Academy
www.aiba.or.kr

Contents

Papers

Information

Evaluating Short-Term Regime Prediction in the Korean Equity Market: A Large-Scale Study of Decision-Rule Dependence

Yeo-Hoon Yoon, Kyung-Sung Kim

ABSTRACT

This study presents a large-scale benchmark and a decision-rule-aware evaluation protocol for short-term regime classification in the Korean stock market. Reliable evaluation in this setting is difficult because reported performance is sensitive to sample construction, look-ahead bias, class imbalance, and the decision rule used to convert model scores into predicted classes. The benchmark comprises 8,732,898 stock-day observations of 4,225 KOSPI and KOSDAQ common stocks from January 2012 to February 2026, with six three-class regime targets defined by ±5% and ±10% thresholds on 1-, 5-, and 10-trading-day forward returns. The protocol combines survivorship-bias-mitigating sampling, chronological splits with a 20-trading-day purge gap, training-only preprocessing, leakage checks, and full class-distribution reporting. Given severe class imbalance, models are evaluated using macro F1 and up/down directional F1 rather than accuracy alone. We report reference results for two tree-based baselines, LightGBM and XGBoost, and six tabular deep learning models. Under the default argmax rule, the deep learning models are more recall-oriented on extreme regimes, whereas LightGBM is more conservative. However, after validation-based threshold tuning, LightGBM's directional F1 improves substantially, and the relative ranking narrows or reverses on several targets. These results show that apparent model superiority depends materially on the decision rule. The contribution of this study is therefore a reproducible benchmark and evaluation protocol, not a claim of universal model superiority or tradable profitability.

☞ Keywords: Korean stock market; Short-term regime classification; Stock-day benchmark; Decision-rule sensitivity; Imbalanced classification; Tabular deep learning

1. Introduction

Forecasting short-term movements in the stock market is a long-standing problem in financial engineering and asset management. The task is difficult because short-horizon returns are characterized by low signal-to-noise ratios, non-stationarity, regime shifts, and changing market microstructure. Recent financial machine learning research shows that nonlinear models can capture complex interactions among financial predictors that are difficult to represent with linear specifications [4]. However, in short-term stock prediction, reported model performance is often affected not only by model architecture but also by the evaluation design itself.

A central concern is evaluation reliability. When future returns are used to define prediction targets, performance can be overstated if explanatory variables contain future information, if training and test periods overlap, or if preprocessing statistics are estimated using information outside the training period [9]. In addition, short-term regime targets are usually imbalanced because neutral states dominate the sample, whereas large upward or downward movements occur relatively infrequently. Under such conditions, accuracy or default argmax classification may hide a model's ability to detect extreme regimes, so imbalance-aware metrics are necessary for reliable evaluation [6].

In the Korean stock market, prior research has examined stock prediction using index-level data, selected stocks, sentiment variables, and machine learning models. Recent work has also incorporated Korean financial news sentiment into portfolio construction and investment-performance analysis [10]. However, large-scale stock-day benchmarks covering both KOSPI and KOSDAQ common stocks remain limited. This lack of a common

benchmark makes it difficult to determine whether observed performance differences are due to model capability, sample construction, target definition, class imbalance, or the decision rule used to convert model scores into predicted classes.

This study addresses this gap by constructing a large-scale benchmark for short-term regime prediction in the Korean stock market. The benchmark consists of 8,732,898 stock-day observations from 4,225 KOSPI and KOSDAQ common stocks from January 2012 to February 2026. Six three-class regime targets are defined using combinations of prediction horizons — 1, 5, and 10 trading days — and return thresholds — ±5% and ±10%. Each target classifies a stock-day into up, neutral, or down according to its forward return.

Using the same 112 rolling technical indicators, the same chronological train–validation–test split, and the same leakage-aware preprocessing procedure, this study evaluates two tree-based baselines, LightGBM and XGBoost, and six tabular deep learning models, including MLP, ResNet, FT-Transformer, and TabNet variants. Since recent tabular learning research shows that deep learning models and tree-based models can perform differently depending on data and evaluation settings, this comparison is conducted as a standardized benchmark rather than as a claim of universal model superiority [3].

The purpose of this study is to examine how model comparison changes under a decision-rule-aware evaluation protocol. Under the default argmax rule, tabular deep learning models tend to show higher recall on extreme up and down regimes, whereas LightGBM behaves more conservatively and predicts the neutral class more frequently. However, when LightGBM's decision thresholds are tuned on the validation set, its directional F1 improves substantially, and the ranking between model families narrows or reverses on several targets. This result suggests that apparent model superiority in imbalanced regime prediction can depend materially on the decision rule.

Throughout this paper, a regime refers to a threshold-defined short-horizon return state, not a latent macro-financial regime estimated by a state model. The empirical task is therefore short-term regime classification rather than direct trading-strategy evaluation. The prediction signal is interpreted as an end-of-day classification formed after observing day-t information, and issues such as next-day execution, transaction costs, position sizing, and tradable profitability are outside the scope of this study. The contribution of the paper is a reproducible Korean stock-day benchmark and a decision-rule-aware evaluation protocol for comparing short-term regime prediction models.

2. Literature Review

Research on stock market prediction in Korea has examined a range of data sources and modeling approaches, including technical indicators, firm- or industry-level samples, textual information, and machine learning models. Recent Korean-market studies have increasingly incorporated non-price information and machine learning into prediction or portfolio-construction settings. For example, KR-FinBERT-based news sentiment has been incorporated into mean-variance portfolio construction [10], and machine-learning-based stock prediction has been examined in relation to investment strategy performance [8]. These studies show that Korean stock prediction research has moved beyond simple price-based forecasting toward richer data sources and nonlinear modeling. However, much of the existing work still relies on selected indices, specific industries, limited groups of stocks, or restricted sample periods. This limits its use as a common benchmark for comparing model families across the broader Korean equity market.

Recent financial machine learning research also emphasizes that nonlinear models can capture complex interactions among financial predictors that are difficult to represent with linear specifications. Machine learning models, including tree-based and neural network approaches, can improve empirical asset-pricing prediction by modeling nonlinear relations among predictors [4]. In tabular prediction tasks, gradient-boosted decision tree models remain strong baselines, while deep learning models specialized for tabular data have been proposed as competitive alternatives. ResNet-style architectures and FT-Transformer can serve as strong tabular deep learning baselines, while neither tree-based models nor deep learning models are universally superior across all datasets [3]. TabNet has similarly been proposed as an attention-based architecture for tabular learning [1]. These studies suggest that comparisons between tree-based models and tabular deep learning models should be conducted under standardized inputs, splits, and evaluation rules rather than interpreted as general claims of model superiority.

In market-regime prediction, the key issue is not only model

architecture but also how the regime itself is defined. Traditional regime studies often rely on latent-state approaches, such as regime-switching models, Hidden Markov Models, or volatility-state classifications. These approaches are useful for explaining market states retrospectively, but their regime definitions are not always directly aligned with short-horizon prediction tasks. By contrast, this study defines regimes as threshold-based forward-return states — up, neutral, and down — using fixed return thresholds and prediction horizons. This design makes the regime labels operationally interpretable and reproducible, which is important when the purpose is to compare prediction models under a common benchmark rather than to estimate latent macro-financial states.

Another important issue in financial prediction is evaluation reliability. When future returns are used to define labels, reported performance can be overstated if future information leaks into explanatory variables, if training and test periods overlap, or if preprocessing statistics are computed using validation or test information. Time-ordered splits, backtesting-overfitting control, and leakage prevention are emphasized as essential in financial machine learning [9]. In addition, short-term regime classification is typically affected by severe class imbalance because neutral observations dominate the sample, while large upward or downward movements are relatively rare. Under such conditions, accuracy alone can overstate performance, and class-sensitive metrics are needed to evaluate whether models actually detect extreme regimes. A representative discussion of imbalanced learning highlights the limitations of majority-class-biased evaluation [6].

Taken together, recent studies provide useful evidence on Korean stock prediction, financial machine learning, and tabular deep learning. However, they do not provide a common large-scale Korean stock-day benchmark in which sample construction, regime-target definitions, leakage-aware preprocessing, imbalance-aware metrics, and decision rules are fixed simultaneously. This study addresses this gap by treating model comparison as an evaluation-protocol problem rather than as a model-superiority contest.

3. Data Construction and Verification

3.1 Data and Input Variables

The raw dataset covers the period from January 2, 2012 to February 27, 2026 and consists of four data blocks: stock metadata, daily Open-High-Low-Close-Volume (OHLCV) data, pre-market/regular-session/after-market trading microstructure data, and short-selling-related data. The analysis sample is restricted to common stocks listed on the KOSPI and KOSDAQ markets. Product-type securities such as ETFs, ETNs, and REITs are excluded because they follow different trading and short-selling rules. The sample is not limited to currently listed securities; instead, all common stock-day observations with available OHLCV records during the analysis period are included, subject to the coverage of the source data.

The raw datasets are merged using OHLCV records as the base table. Stock metadata are joined by ticker code, while trading-microstructure and short-selling variables are joined by ticker code and trading day. Nine trading-microstructure columns with excessive missingness and limited relevance to the present prediction task are removed. The resulting stock-day table contains 8,732,898 rows, 81 columns, and 4,225 securities. Rows for which horizon-specific forward returns cannot be computed are excluded from both training and evaluation.

Although the raw table includes ticker codes, trading dates, price levels, and a market-cap-equivalent auxiliary variable derived from price and outstanding shares, these variables are not used as direct model inputs. The input feature set is restricted to 112 rolling technical indicators computed from OHLCV, trading-microstructure, and short-selling data. These indicators are calculated over short-, medium-, and long-term rolling windows of 3, 5, 10, 20, and 60 trading days to reduce dependence on absolute price or volume scales across stocks.

The 112 input variables are grouped into six categories: price and trend indicators, momentum indicators, volatility indicators, volume and trading-value indicators, session-structure indicators, and short-selling indicators. Price and trend indicators include returns, moving-average deviations, moving-average crossovers, and Bollinger Z-scores. Momentum indicators include RSI, Stochastic K, Williams %R, and cumulative returns. Volatility indicators include true range, ATR, return volatility, and intraday range. Volume and value indicators include volume deviation, trading-value deviation, turnover, OBV, and PVT. Session-structure indicators summarize pre-market, regular-session, and after-market volume and value shares. Short-selling indicators include short-selling value share, short-selling pressure, and deviations from the average short-selling

price. The complete list of indicators, formulas, and rolling windows is provided in Appendix C.

To preserve temporal consistency, all input variables for stock-day t are computed only from information available by the end of day t. This includes after-market session variables and official short-selling statistics, which are finalized after the regular close. The prediction signal is therefore treated as an end-of-day signal formed after observing day-t information. Since regime labels are defined using forward returns from day t onward, no input variable incorporates information dated later than day t. This construction is intended to prevent look-ahead bias and maintain a time-consistent prediction setting [9]. All eight reference models — two tree-based baselines and six tabular deep learning models — use the same 112 input variables.

3.2 Data Verification

Because this study constructs a benchmark from a large-scale stock-day panel, data integrity and temporal consistency are verified before conducting the reference model comparison. We first examine whether the merged dataset preserves the intended sample structure throughout preprocessing. The final big table contains 8,732,898 rows, 81 columns, and 4,225 securities over the period from January 2, 2012 to February 27, 2026. The row count remains unchanged after the technical-indicator computation step, indicating that feature construction does not introduce unintended sample loss.

We then verify the chronological structure of the training, validation, and test samples. The split is designed so that the maximum training date precedes the minimum validation date, and the maximum validation date precedes the minimum test date. In addition, a 20-trading-day purge gap is imposed around each split boundary to reduce the risk of temporal leakage and sample overlap, which are common concerns in financial time-series prediction [9]. This design ensures that model estimation and evaluation are conducted under a strictly time-ordered setting.

To further confirm that the split procedure does not generate duplicated evaluation units, we examine overlaps in (ticker, trading day) pairs across the training, validation, and test samples. The overlap counts between training and validation, validation and test, and training and test are all zero. This verification is important because stock-day panels can otherwise contain repeated or adjacent observations that blur the distinction between estimation and evaluation periods.

We also verify that variables related to future returns are used only for regime-label construction and are not included in the model input set. The input variables are restricted to features computed from information available up to the end of each stock-day. In addition, missing-value imputation and scaling parameters are estimated only from the training sample and then applied unchanged to the validation and test samples. This training-only preprocessing procedure prevents validation or test information from entering the input distribution during model preparation.

Finally, the security-level treatment of the panel is summarized in Table 1. The benchmark does not apply post-hoc filtering based on current listing status. Instead, all common stock-day observations with available OHLCV records in the source data are included as candidates for the panel, subject to the coverage limits of the source data. This treatment is intended to reduce survivorship bias while maintaining a transparent and reproducible sample-construction rule.

(Table 1 Security-day-level sample treatment)

Item	Treatment
Delisted securities	Included if OHLCV observations exist in the source data
Currently-listed-only restriction	Not applied (no post-hoc survival filtering)
Rows with infeasible forward returns	Excluded from both training and evaluation
Newly listed securities (initial 60 trading days)	Rows that do not satisfy long-window indicators are excluded from the evaluation sample
Trading halts and missing values	Rolling technical indicators are computed strictly from past observations; missing values during training are imputed using training-sample statistics

4. Benchmark Protocol

In this study, a regime denotes a threshold-defined forward-return state (up, neutral, or down) rather than a latent, model-inferred market state such as a regime-switching state [5]; each three-class label is determined solely by whether a forward return crosses fixed thresholds.

We design six regime classification targets as combinations of three prediction horizons (1, 5, and 10 trading days) and two thresholds (±5%, ±10%). For each stock and trading day, the future return over the prediction horizon is computed as the close-to-close percentage change between the current day's close and the close at the end of the horizon. A three-class label is then assigned: up if the future return exceeds the positive threshold, down if it falls below the negative threshold, and neutral otherwise. The six combinations are 1-day ±5%, 1-day ±10%, 5-day ±5%, 5-day ±10%, 10-day ±5%, and 10-day ±10%, designated as target IDs regime_1d_thr5, regime_1d_thr10, regime_5d_thr5, regime_5d_thr10, regime_10d_thr5, and regime_10d_thr10, respectively. The 10-day ±5% target is used as the main target because it provides a comparatively balanced extreme-regime distribution (both extreme classes occupy approximately 18% of the sample) and represents an operationally interpretable short-horizon regime definition. The 1-day ±10% target has extreme-class proportions of approximately 1% or less, and is treated as a challenge target for rare-regime capture in the analysis below. Because the 5- and 10-trading-day targets are defined on overlapping forward-return windows, consecutive labels for the same stock are not independent; the evaluation treats each stock-day as one observation and does not explicitly model this within-split dependence. Future work could further address this issue using block-bootstrap procedures or non-overlapping resampling designs [9].

The test-set class distribution of the six targets defined above is reported in Table 2. As the threshold strengthens from 5% to 10%, the extreme-class ratios fall sharply, and as the horizon lengthens from 1 to 5 and 10 trading days, the extreme-class ratios increase again. The training and validation samples exhibit essentially the same pattern; absolute counts are summarized in Appendix A. In particular, the 1-day ±10% target has extreme-class ratios of 0.42% (down) and 0.98% (up), making it the most rare among the six targets, and is classified as a challenge target in this study.

(Table 2 Test-set class distribution by regime target)

Target	Down	Neutral	Up
1-day ±5%	53,152 (2.74%)	1,814,666 (93.58%)	71,364 (3.68%)
1-day ±10%	8,075 (0.42%)	1,912,081 (98.60%)	19,026 (0.98%)
5-day ±5%	240,161 (12.49%)	1,433,452 (74.57%)	248,686 (12.94%)
5-day ±10%	66,164 (3.44%)	1,757,425 (91.42%)	98,710 (5.13%)
10-day ±5% (main)	358,826 (18.87%)	1,183,035 (62.23%)	359,358 (18.90%)
10-day ±10%	133,023 (7.00%)	1,596,284 (83.96%)	171,912 (9.04%)

These class-distribution patterns directly motivate the choice of evaluation metrics. Prior research on multiclass classification evaluation points out that simple accuracy can overestimate the discriminative power of a model in problems with a dominant majority class, and recommends reporting class-wise precision, recall, and their harmonic mean F1 [11]. A representative survey on imbalanced learning similarly emphasizes that, when detecting extreme classes is crucial, accuracy alone is vulnerable to majority-class bias and distribution-robust metrics such as F1, G-mean, and AUC should be used [6]. In our benchmark, the regime targets generally have a high neutral-class share, and especially in short-horizon, high-threshold targets the neutral share reaches around 90%, indicating severe class imbalance; thus simple accuracy may not adequately reflect the model's ability to identify extreme classes. We therefore report four metrics together. First, macro precision and macro F1 over the three classes (up, neutral, down). Second, a binary directional F1 with the up class as positive and the neutral and down classes combined as negative, hereafter the up directional F1. Third, a binary directional F1 with the down class as positive and the neutral and up classes combined as negative, hereafter the down directional F1. Fourth, the arithmetic mean of the up directional F1 and the down directional F1, which we refer to as the mean directional F1.

The mean directional F1 mitigates the problem of overestimating overall performance when predictions concentrate on the neutral class, and is used as a metric that separately evaluates the ability to capture the up and down extreme classes. As a trivial reference point,

a majority-class classifier that always predicts the dominant neutral class attains an up and down directional F1 of zero by construction; any positive directional F1 therefore reflects discrimination above this floor. For brevity, in tables we abbreviate the mean directional F1 as mDir-F1 and its difference as ΔmDir-F1. Since F1 alone is not sufficient to characterize how a model behaves, this benchmark additionally reports precision and recall for the extreme classes. This is consistent with the recommendation of the classification evaluation literature that the same F1 value with "high precision and low recall" implies a different operational meaning than "low precision and high recall," and that precision and recall should be reported separately to expose such differences [11].

5. Reference Baseline Protocol

All reference models are trained, validated, and evaluated using the same data split, which together with the feature definitions and random seeds constitutes the evaluation protocol. After applying the purge gap, the training period is 2012-01-02 to 2021-12-02 (5,124,124 rows, 2,443 trading days), the validation period is 2022-02-03 to 2023-11-29 (1,411,677 rows, 451 trading days), and the test period is 2024-01-30 to 2026-02-27 (1,943,407 rows, 504 trading days). Note that the 1,943,407 rows correspond to the test period immediately after the time-series split. The final per-target evaluation sample is then obtained by excluding (i) rows at the end of the sample period for which the horizon-specific forward return is not yet observable — so that the number of excluded rows grows with the prediction horizon — and (ii) early rows that lack sufficient history for the long-window (up to 60-trading-day) indicators. Because the first exclusion is horizon-dependent, the evaluation sample size differs slightly across the six targets. All numbers reported in Table 2, Table 3, and Appendix B are based on this final per-target model-evaluation sample. Each model is selected on the validation set based on macro precision, and final reporting on the test set uses the model with the highest validation macro precision. The choice of macro precision as the (disclosed) model-selection criterion reflects the priority of restraining excessive false positives in the extreme classes, given the heavily imbalanced class distribution of the regime targets. As a sensitivity check, we also repeated model selection using validation macro F1 and validation mean directional F1; the main ranking between the best deep learning model and the best baseline remained unchanged.

The reference models include two strong tree-based baselines for tabular data — LightGBM and XGBoost, representative gradient-boosted decision tree baselines for tabular prediction [2], [7] — together with six tabular deep learning models representing the three axes of depth, residual connection, and attention. The deep learning group consists of two simple multilayer perceptrons (MLP-2Layer, MLP-3Layer), two residual-connection variants (MLP-ResNet, ResNet), and two attention-based models, FT-Transformer and TabNet, which represent recent tabular deep learning architectures [1], [3]. All models use the same 112 rolling technical indicators as inputs and the same training/validation/test splits. The deep learning optimizer is AdamW, and the loss function is class-weighted cross-entropy; the full hyperparameter settings for all eight models are reported in Appendix D. As a trivial reference floor under the protocol, a majority-class classifier attains zero up and down directional F1 by construction; the learned baselines below are therefore read relative to this floor rather than as competitors for a superiority claim.

The default training epoch is 20, but in some models we observe a phenomenon in which validation performance gradually improves in the latter half of training (hereafter, late-peaking). To conservatively check this latter-half stability, only the model-target runs that satisfy two pre-specified conditions — first, the best epoch occurring at epoch 18 or later, and second, the last-epoch validation macro precision being within 0.002 of the best — are extended up to 30 epochs. The extension only updates the best deep model on the 1-day ±10% target from FT-Transformer to ResNet, while on the other five targets the best model and test macro F1 remain unchanged. All results in this study therefore use the model weights at the epoch with the highest validation macro precision among models trained for either 20 or 30 epochs under the above criteria.

Inspecting the epoch-wise learning curve of the MLP-2Layer model on the main target (10-day ±5%), the training loss decreases monotonically with training, and validation macro F1 and mean directional F1 quickly enter a stable region after the first few epochs and remain low-variance thereafter. This suggests that learning stabilizes early on the main target and that results are not particularly sensitive to the choice of training epoch. Due to space constraints, this paper presents only a qualitative description of the learning curve.

6. Reference Baseline Results

Table 3 reports reference test performance for each of the six regime targets, contrasting the best deep learning model and the best tree baseline selected by validation macro precision. The two tree

baselines, LightGBM and XGBoost, perform comparably: across the six targets their test macro F1 differs by within about 0.03 and neither dominates (Appendix F), so LightGBM is reported as the representative tree baseline in the main tables and the full XGBoost results are given in Appendix F. The full validation-set macro F1 of all reference models is reported in Appendix E, where a deep learning model attains the highest validation score on every target; because Table 3 selects by validation macro precision (Section 5) rather than the macro F1 shown there, the Best deep model differs from the highest-F1 column on 5-day ±5% and 10-day ±10%. The best tree baseline behaves, under the default argmax rule, as a conservative, neutral-centered classifier with low recall on the extreme classes, which limits macro F1; test macro F1 ranges from approximately 0.30 to 0.36. Under the argmax rule the GBDT baseline tends toward a majority-biased classifier on the extreme classes. We report the differences between the two model groups in macro F1 and mean directional F1 (ΔmF1, ΔmDir-F1) descriptively; because the absolute magnitudes of these differences depend on extreme-class proportions, we describe the differences without interpreting them as evidence of model superiority. All metrics in this section are classifier-level regime-identification performance, not evidence of tradable profitability; execution assumptions such as next-day open prices, transaction costs, and position sizing are outside the scope of this benchmark.

(Table 3 Best deep learning vs. best baseline model on test set, with slice robustness (5-seed mean))

Target	Best deep	Deep mF1	Base mF1	Deep mDir-F1	Base mDir-F1	ΔmF1 / ΔmDir-F1	Slice wins (yr/mkt)
1-day ±5%	FT-Transformer	0.3923	0.3169	0.1712	0.0279	+0.0754 / +0.1433	3/3, 3/3
1-day ±10% (challenge)	ResNet	0.3346	0.3135	0.0652	0.0044	+0.0211 / +0.0608	3/3, 3/3
5-day ±5%	FT-Transformer	0.4426	0.3551	0.3033	0.1575	+0.0875 / +0.1458	3/3, 3/3
5-day ±10%	FT-Transformer	0.3962	0.2983	0.1943	0.0182	+0.0979 / +0.1761	3/3, 3/3
10-day ±5% (main)	MLP-2Layer	0.4457	0.3616	0.3323	0.1553	+0.0841 / +0.1770	3/3, 3/3
10-day ±10%	FT-Transformer	0.4120	0.3245	0.2391	0.0605	+0.0875 / +0.1786	3/3, 3/3

Reporting key slice-level reference numbers, the main target (10-day ±5%) shows the following pattern under the argmax rule. In yearly slices, deep learning's mean directional F1 is 0.3476 (2024), 0.3199 (2025), and 0.3220 (the partial 2026 period through February), staying above 0.31 in all three slices, while the baseline reaches only 0.1652, 0.1478, and 0.1853, respectively. In market slices, deep learning's mDir-F1 on the main target is 0.3052 (KOSPI) and 0.3578 (KOSDAQ), above the baseline's 0.1143 and 0.1890 in both markets. A similar pattern holds for the 1-day ±5% target; for the challenge target (1-day ±10%), absolute values are small (around 0.07) but the direction of the slice-level difference is consistent across slices. The market-slice analysis evaluates the same model trained on the entire-market panel on each market subsample, rather than retraining a separate model per market. These slice comparisons are reported under the argmax rule; their dependence on the decision rule is the subject of Section 7.

The difference between the two model groups under argmax reflects a difference in the precision–recall trade-off rather than absolute superiority. On the main target 10-day ±5%, LightGBM attains relatively high extreme-class precision (0.3996 up, 0.4481 down) but low extreme-class recall (0.0498 up, 0.1476 down). MLP-2Layer attains precision 0.2785 / recall 0.2909 on the up class and precision 0.3192 / recall 0.4692 on the down class, trading some precision for higher extreme-class recall and F1. Extreme-class F1 thus moves from 0.0886 to 0.2846 for the up class and from 0.2220 to 0.3800 for the down class between the two reference models. Relative to a majority-class classifier, which attains zero directional F1 by construction, both learned baselines exhibit non-trivial directional discrimination above the floor; the magnitudes, however, are modest. This pattern recurs across all six targets (see Appendix B for the detailed confusion matrix on the main target).

On the 1-day ±10% challenge target, although the extreme-class share is below 1%, the gap in extreme-class recall between the two

reference models is the most pronounced among the six targets. LightGBM's extreme-class recall is 0.71% on the up class and 0.25% on the down class, whereas the best deep learning model ResNet attains an up recall of 37.31% and a down recall of 37.27%. At the same time, the deep model's extreme-class precision on this target is very low (0.0448 up, 0.0383 down). This illustrates that class-weighted loss can increase false positives while lifting recall, and that in a regime region almost invisible to accuracy-based or argmax classification the choice of operating point dominates the apparent result. The result on this target should therefore be read not as a practical trading signal but as a motivation for the decision-rule sensitivity analysis in Section 7.

7. Protocol Sensitivity

This section examines how stable the reference results of Section 6 are to the choice of a single random seed and — most importantly for evaluation — to the baseline's decision rule. Slice stability was addressed in Section 6 (Table 3 and the accompanying narrative); here we present two remaining checks in turn — training stability and decision-rule sensitivity — and consolidate the results in Table 4.

To check that the single-seed results are not coincidental, we repeatedly trained the best deep learning model for each of the six regime targets with five different random seeds. Across all six targets, the standard deviation of the five-run test macro F1 ranges from 0.0010 to 0.0060, indicating stable training; on the main target 10-day ±5%, the five-run test macro F1 is 0.4457 with a standard deviation of 0.0024, below 0.003 (see the second column of Table 4).

Most importantly, to test whether the conservative argmax behavior of the baseline is an artifact of the decision rule, we conducted threshold tuning on LightGBM for all six regime targets: we grid-searched the up-class probability threshold and the down-class probability threshold on the validation set so as to maximize mean directional F1, and applied the optimal thresholds on the test set. Under the tuned rule a stock-day is labeled up if its up-class probability exceeds the up threshold and down if its down-class probability exceeds the down threshold; if both thresholds are exceeded, the class with the larger margin above its threshold is chosen, and if neither is exceeded the prediction defaults to neutral. The validation-best thresholds were (up, down) = (0.20, 0.15) for 10-day ±5%, (0.15, 0.15) for 5-day ±5%, (0.12, 0.12) for 10-day ±10%, and (0.10, 0.10) for each of 1-day ±5%, 1-day ±10%, and 5-day ±10%, and the resulting macro F1 and mean directional F1 of the tuned LightGBM are reported in the last column of Table 4. Threshold tuning substantially improved the extreme-class recall and mean directional F1 of LightGBM on all six targets, indicating that argmax-based evaluation can understate the directional potential of the baseline. For example, on the 1-day ±10% challenge target the mean directional F1 of LightGBM rose from 0.0044 under argmax to 0.1206 after tuning, and on 10-day ±10% from 0.0605 to 0.2391.

The cross-model comparison after threshold tuning can be summarized as follows. On the main target (10-day ±5%), the best deep model still exceeds tuned LightGBM in both macro F1 and mean directional F1 (0.4457 vs. 0.3924 in macro F1 and 0.3323 vs. 0.3287 in mean directional F1, based on Tables 3 and 4). Across the remaining targets, threshold tuning reverses the mean-directional-F1 ranking on three targets (1-day ±5%, 1-day ±10%, and 5-day ±10%); on 5-day ±5% the deep model remains slightly higher (0.3033 vs. 0.2962), while on 10-day ±10% the two models are virtually tied (0.2391 vs. 0.2391). In macro F1, tuned LightGBM exceeds the best deep model on four of the six targets (1-day ±5%, 1-day ±10%, 5-day ±10%, and 10-day ±10%). The relative ordering between the two model families therefore depends materially on the decision rule. For a benchmark, the operational implication is that results must be reported under an explicit, common decision rule: comparing one family under argmax against another under tuned thresholds — or vice versa — can invert conclusions. This is the central motivation for fixing a decision-rule-aware evaluation protocol. This threshold-tuning exercise is designed as a diagnostic analysis of decision-rule dependence, not as a fully optimized comparison across all model families. Under class imbalance, the optimal decision threshold depends on class priors and misclassification costs and is not fixed at the argmax of the predicted probabilities. Adjusting the decision threshold is therefore a recognized component of imbalanced classification [6]. We therefore tune only the baseline's decision threshold as a controlled probe of whether the argmax ranking is stable. Because the deep learning models are not threshold-tuned in the same manner, the tuned-LightGBM results should be interpreted as evidence that argmax-based rankings are sensitive to the decision rule, rather than as a final ranking of model families.

(Table 4 Robustness summary: multi-seed stability and LightGBM threshold tuning)

Target	5-seed mF1 (mean ± sd)	LightGBM tuned (mF1 / mDir-F1)
1-day ±5%	0.3923 ±	0.4405 / 0.1975

Target	5-seed mF1 (mean ± sd)	LightGBM tuned (mF1 / mDir-F1)
	0.0050	
1-day ±10% (challenge)	0.3346 ± 0.0010	0.4098 / 0.1206
5-day ±5%	0.4426 ± 0.0043	0.4284 / 0.2962
5-day ±10%	0.3962 ± 0.0045	0.4371 / 0.2109
10-day ±5% (main)	0.4457 ± 0.0024	0.3924 / 0.3287
10-day ±10%	0.4120 ± 0.0060	0.4200 / 0.2391

8. Ethical Considerations

The benchmark is constructed entirely from market-level data — daily OHLCV, intraday trading-microstructure aggregates, and short-selling statistics — together with security metadata, and contains no personally identifiable information about individual investors, so record-level privacy risk is minimal. The underlying market data are obtained from the Korea Exchange (KRX) and standard market-data providers and are used under their respective terms; the 112 input features are derived from these series using standard, publicly documented technical-indicator formulas, and no proprietary third-party series are redistributed.

To limit selection and survivorship bias, the panel includes all KOSPI and KOSDAQ common stock-day rows for which observations exist in the source data, including securities delisted during the study period, rather than only currently listed names (Section 3). The severe class imbalance of the regime targets is handled explicitly through full class-distribution reporting and imbalance-aware metrics — macro F1 and up/down directional F1 — rather than accuracy alone, so that the benchmark does not reward majority-class prediction. Residual coverage limits inherited from the source data are stated in Section 3 so that users can interpret results within those bounds.

Because the prediction setting evaluates end-of-day directional regime classification rather than executable trades, the benchmark and its reference results are intended for methodological evaluation, not as investment advice or a tradable signal, and the precision–recall and decision-rule analyses are reported to characterize classifier behavior rather than to recommend trading strategies. Users should account for transaction costs, execution constraints, and applicable regulations before drawing any operational conclusion.

9. Conclusion and Availability

We constructed a large-scale benchmark with leakage-aware construction and verification for short-term regime classification on a Korean stock-day panel, defined six operationally interpretable regime targets and an evaluation protocol for imbalanced extreme-regime detection, and reported reference results for two tree baselines and six tabular deep learning models under identical inputs and splits. Three points summarize the study. First, the construction and verification pipeline — survivorship-bias-mitigating sampling, time-ordered splits with a purge gap, training-only preprocessing, and stepwise overlap and leakage checks — provides a reproducible basis on which future Korean regime-classification work can be compared. Second, the evaluation protocol, by reporting up/down directional F1 alongside macro F1 and class-wise precision and recall, exposes extreme-regime detection behavior that aggregate accuracy hides under severe class imbalance. Third, under this protocol the relative ordering of the two model families is decision-rule-dependent.

Under the default argmax rule the tabular deep learning reference models are more recall-oriented on the extreme classes while the tree baseline is more conservative and precision-oriented; for example, on the main target 10-day ±5% the up- and down-class recalls of the baseline are 4.98% and 14.76%, while those of the deep model are 29.09% and 46.92% (Appendix B), and on the 1-day ±10% challenge target the baseline's extreme-class recall remains below 1% whereas the deep model attains roughly 37% (Section 6). Once the baseline's decision thresholds are tuned, however, the gap narrows or reverses: tuned LightGBM exceeds the best deep model in mean directional F1 on three targets, remains slightly below on two, and is virtually tied on one, and exceeds it in macro F1 on four of six targets. We therefore make no claim of universal model superiority.

The central, reproducible message of this study is that credible comparison requires a common decision rule and a leakage-controlled protocol: the apparent advantage of one model family can be an artifact of the evaluation choice rather than a property of the model. The benchmark and protocol presented here provide a useful basis for future studies on Korean stock market regime classification and highlight the importance of evaluating directional-regime identification alongside conventional classification metrics. The

empirical results should not be interpreted as evidence of directly tradable profitability.

Availability. The benchmark protocol is specified for reproduction through the construction and verification rules in Section 3, the technical-indicator definitions and rolling windows in Section 3.1 and Appendix C, the model hyperparameters in Appendix D, the six regime-target definitions in Section 4, and the fixed chronological split dates and random seeds in Section 5, with preprocessing statistics fit on the training sample only. Access to the underlying raw market data depends on the terms of the Korea Exchange and commercial data providers, and the implementation code for the construction pipeline, regime labeling, evaluation protocol, and decision-rule tuning is available from the authors for research use upon reasonable request.

References

[1] Arik, S. Ö., & Pfister, T. (2021). TabNet: Attentive interpretable tabular learning. Proceedings of the AAAI Conference on Artificial Intelligence, 35(8), 6679–6687.

[2] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.

[3] Gorishniy, Y., Rubachev, I., Khrulkov, V., & Babenko, A. (2021). Revisiting deep learning models for tabular data. Advances in Neural Information Processing Systems, 34, 18932–18943.

[4] Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. The Review of Financial Studies, 33(5), 2223–2273.

[5] Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. Econometrica, 57(2), 357–384.

[6] He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263–1284.

[7] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems, 30, 3146–3154.

[8] Kim, S. (2023). A comparison of stock price prediction for global automotive companies using machine learning models [in Korean]. Journal of the Korean Data Analysis Society, 25(1), 249–263.

[9] López de Prado, M. (2018). Advances in financial machine learning. Wiley.

[10] Park, J., Jung, M., Kim, H., & Kim, S. (2024). An analysis of investment performance for portfolio selection models reflecting news sentiment analysis: Focusing on the Korean stock market [in Korean]. Journal of the Korean Operations Research and Management Science Society, 49(4), 57–72.

[11] Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45(4), 427–437.

Appendix

A. Class Distribution by Target (Training and Validation)

Target	Split	Down	Neutral	Up
1-day ±5%	Train	163,061 (3.18%)	4,748,406 (92.67%)	212,657 (4.15%)
	Valid	42,814 (3.03%)	1,319,527 (93.47%)	49,336 (3.49%)
1-day ±10%	Train	26,000 (0.51%)	5,040,971 (98.38%)	57,153 (1.12%)
	Valid	4,448 (0.32%)	1,395,145 (98.83%)	12,084 (0.86%)
5-day ±5%	Train	738,715 (14.42%)	3,616,007 (70.57%)	769,402 (15.02%)
	Valid	213,694 (15.14%)	1,021,510 (72.36%)	176,473 (12.50%)
5-day ±10%	Train	212,731 (4.15%)	4,608,214 (89.93%)	303,179 (5.92%)
	Valid	59,640 (4.22%)	1,288,594 (91.28%)	63,443 (4.49%)
10-day ±5%	Train	1,085,959 (21.19%)	2,942,804 (57.43%)	1,095,361 (21.38%)
	Valid	318,884 (22.59%)	836,376 (59.25%)	256,417 (18.16%)
10-day ±10%	Train	419,180 (8.18%)	4,173,593 (81.45%)	531,351 (10.37%)
	Valid	124,681 (8.83%)	1,174,660 (83.21%)	112,336 (7.96%)

Note: Column definitions — Target is the regime target; Split indicates the training or validation period; Down, Neutral, and Up give the count and, in parentheses, the share of each regime class within that split.

B. Extreme-Class Confusion Matrix on the Main Target (10-day ±5%)

Class	Model	TP	FN	FP	TN	Precision	Recall	F1
Up	MLP-2Layer (deep)	104,524	254,834	270,722	1,271,139	0.2785	0.2909	0.2846
Up	LightGBM (base)	17,906	341,452	26,900	1,514,961	0.3996	0.0498	0.0886
Down	MLP-2Layer (deep)	168,369	190,457	359,064	1,183,329	0.3192	0.4692	0.3800
Down	LightGBM (base)	52,947	305,879	65,207	1,477,186	0.4481	0.1476	0.2220

C. The 112 Technical Indicators

The 112 input features are generated from 21 indicator families, each computed over five rolling windows (3, 5, 10, 20, and 60 trading days), giving 105 features, together with a Bollinger Z-score at two windows (20 and 60 days) and five cross-window indicators, for 112 features in total. Each family and its formula are listed below; the window suffix (for example, _20d) denotes the rolling window. All features for a stock-day t use information available only up to day t, so no feature incorporates look-ahead information relative to the forward-return targets.

Group	Indicator family	Formula (rolling, window n)	Windows (days)
Price / trend	return / return_mean	n-day rolling mean of daily return (close_t/close_{t-1} - 1)	3, 5, 10, 20, 60
	sma_gap	(close - SMA_n) / SMA_n (moving-average gap)	3, 5, 10, 20, 60
	bollinger_z	(close - SMA_n) / STD_n (Bollinger Z-score)	20, 60
	sma_cross	SMA_a / SMA_b - 1 (fast/slow MA cross)	5/20, 10/60
Momentum	rsi	RSI(n) (Wilder relative strength index)	3, 5, 10, 20, 60
	stoch_k	stochastic %K over n days	3, 5, 10, 20, 60
	williams_r	Williams %R over n days	3, 5, 10, 20, 60
	range_position	position of close within the n-day high-low range	3, 5, 10, 20, 60
Volatility	return_vol	n-day rolling std of daily return	3, 5, 10, 20, 60
	atr	n-day rolling mean of true range	3, 5, 10, 20, 60
	intraday_range_mean	n-day rolling mean of (high - low) / close	3, 5, 10, 20, 60
	volatility_regime	STD_a / STD_b - 1 (short/long volatility ratio)	5/20
Volume / value	volume_gap / value_gap	(volume or value - SMA_n) / SMA_n	3, 5, 10, 20, 60
	turnover_gap	(turnover - SMA_n) / SMA_n	3, 5, 10, 20, 60
	obv_flow / pvt_flow	n-day difference of OBV / PVT	3, 5, 10, 20, 60
	volume_cross	SMA_vol_a / SMA_vol_b - 1	5/20

Group	Indicator family	Formula (rolling, window n)	Windows (days)
Session structure	pre_qty_share_avg / after_qty_share_avg	n-day mean of pre-/after-market volume share	3, 5, 10, 20, 60
	pre_value_share_avg / after_value_share_avg	n-day mean of pre-/after-market value share	3, 5, 10, 20, 60
Short-selling	shortsell_share_avg	n-day mean of short-selling value share	3, 5, 10, 20, 60
	shortsell_pressure	(short-selling - SMA_n) / SMA_n	3, 5, 10, 20, 60
	shortsell_cross	SMA_short_a / SMA_short_b - 1	5/20

D. Reference-Model Hyperparameters

Model	Type	Key hyperparameters	Loss	Optimizer	Epochs / stop
LightGBM	GBDT (CPU)	n_estimators=500, lr=0.05, num_leaves=127, subsample=0.8, colsample=0.8, reg_lambda=1.0	multiclass logloss	—	<=500, early-stop 50
XGBoost	GBDT (CPU)	n_estimators=500, lr=0.05, max_depth=8, subsample=0.8, colsample=0.8, reg_lambda=1.0	softprob	—	<=500, early-stop 50
MLP-2Layer	MLP (GPU)	hidden=[512, 256], dropout=0.2, batch=16384	weighted CE	AdamW (lr=1e-3, wd=1e-4)	20
MLP-3Layer	MLP (GPU)	hidden=[512, 256, 128], dropout=0.2, batch=12288	weighted CE	AdamW (lr=8e-4, wd=1e-4)	20
MLP-ResNet	MLP-ResNet (GPU)	d_hidden=256, n_blocks=4, dropout=0.2, batch=16384	weighted CE	AdamW (lr=1e-3, wd=1e-4)	20
ResNet	ResNet-Tab (GPU)	d_hidden=384, n_blocks=6, dropout=0.15, batch=12288	weighted CE	AdamW (lr=8e-4, wd=1e-4)	20
FT-Transformer	Transformer (GPU)	d_token=64, n_blocks=3, n_heads=4, dropout=0.15, batch=4096	weighted CE	AdamW (lr=7e-4, wd=1e-4)	20
TabNet	TabNet (GPU)	n_d=32, n_a=32, n_steps=5, gamma=1.5, lambda_sparse=1e-4, batch=8192	weighted CE	AdamW (lr=1e-3, wd=1e-4)	20

E. Validation-Set Macro F1 of the Reference Models

Target	MLP-2L	MLP-3L	MLP-Res	ResNet	FT-Tr	TabNet	XGB
1-day ±5%	0.3732	0.3769	0.3700	0.3842	0.4115	0.3661	0.3325
1-day ±10%	0.3273	0.3300	0.3312	0.3549	0.3443	0.3250	0.3384
5-day ±5%	0.4347	0.4370	0.4307	0.4309	0.4161	0.4281	0.3319
5-day ±10%	0.3701	0.3740	0.3723	0.3709	0.4039	0.3636	0.3267
10-day ±5% (main)	0.4557	0.4524	0.4430	0.4458	0.4263	0.4513	0.3433
10-day ±10%	0.3997	0.4042	0.4004	0.4077	0.3863	0.3946	0.3157

F. XGBoost Tree-Baseline Results (Test Set)

Target	Accuracy	Macro precision	Macro recall	Macro F1
1-day ±5%	0.9359	0.5951	0.3407	0.3372
1-day ±10%	0.9860	0.6004	0.3365	0.3373
5-day ±5%	0.7502	0.5481	0.3619	0.3436
5-day ±10%	0.9142	0.5649	0.3388	0.3296
10-day ±5% (main)	0.6360	0.5000	0.3848	0.3607
10-day ±10%	0.8396	0.5312	0.3403	0.3193

Author Biography

Yeo-Hoon Yoon

Ph.D. Student, Graduate School of AI, Seoul School of Integrated Sciences & Technologies (aSSIST)

Research Interests: AI, Time-Series Data Forecasting, Digital Transformation, Generative AI, Data Analysis

E-mail: yh.yoon@stude.assist.ac.kr

Kyung-Sung Kim

Chair Professor, The Institute for Industrial Policy Studies Visiting Professor, Seoul School of Integrated Sciences & Technologies(aSSIST)

Honorary Dean of AI School

Ph.D. in Education, University of California, Los Angeles (UCLA)

Research Interests : AI, Statistical Analysis, Psychometrics

E-mail : kskim3@assist.ac.kr

Comparative Analysis of Medium-Range (72-hour) Temperature Time Series Forecasting Using Korea Meteorological Administration Surface Observation Data

Hyun Koo Lee, Nak Hyun Jung

ABSTRACT

This study presents a comparative evaluation of medium-range (72-hour) air temperature forecasting performance among recent deep learning–based time series models—Neural Hierarchical Interpolation for Time Series Forecasting (NHITS), Temporal Fusion Transformer (TFT), Patch Time Series Transformer (PatchTST)—relative to a traditional statistical model (SARIMA) and a climatology-based baseline. The analysis is conducted using hourly surface observation data provided by the Korea Meteorological Administration, with a focus on a 72-hour multi-step forecasting task for the Seoul station.

Empirical results show that all three deep learning models substantially outperform the traditional baselines. Among them, the

TFT achieves the best overall performance, with a Mean Squared Error (MSE) of 4.39, a Root Mean Squared Error (RMSE) of 2.10, a Mean Absolute Error (MAE) of 1.65, and a coefficient of determination (R^2) of 0.775. This corresponds to an approximate 54% reduction in RMSE compared with the SARIMA model. NHITS also demonstrates a significant improvement over the statistical baseline, achieving an RMSE of 2.61.

These results indicate that the Temporal Fusion Transformer architecture, which combines a Long Short-Term Memory (LSTM) based sequence encoder with multi-head attention, is particularly effective in capturing the nonlinear dynamics and combined seasonal and diurnal structures inherent in air temperature time series. This study provides quantitative evidence supporting the applicability of advanced deep learning models to short- to medium-range temperature forecasting based on domestic meteorological observation data. The findings have practical implications for applications requiring accurate temperature prediction, including smart grid energy demand management and weather-related disaster prevention. Future work should extend this framework to multivariate and multi-station settings, as well as to probabilistic forecasting approaches for explicit quantification of predictive uncertainty

☞ Keyword: Air temperature forecasting, time series forecasting, deep learning, Temporal Fusion Transformer, medium range forecasting, meteorological observation data

1. Introduction

1.1 Research Background

Air temperature is a fundamental meteorological variable that plays a critical role across a wide range of societal domains, including agricultural productivity, energy supply and demand planning, disaster management, and public health. Temperature time series inherently exhibit complex seasonal structures driven by the Earth's rotation and revolution, characterized by daily (24-hour) and annual (365-day) cycles. Superimposed on these periodic components are nonlinear fluctuations arising from factors such as frontal passages, variations in cloud cover, and local atmospheric circulation, resulting in a highly complex dynamical system. Owing to these characteristics, temperature forecasting has been widely regarded as a representative yet challenging benchmark for evaluating the predictive performance and generalization capability of time series forecasting models.

For several decades, statistical approaches based on the ARIMA (AutoRegressive Integrated Moving Average) family constituted the dominant methodology for temperature forecasting. However, the recent

intensification of meteorological variability associated with climate change has increasingly revealed the limitations of these linear models. As an alternative, deep learning–based time series forecasting methods have gained attention for their strong capacity to learn nonlinear patterns, leading to a gradual paradigm shift in forecasting methodology. In particular, the Transformer architecture proposed by Vaswani et al. (2017)[1], originally developed for natural language processing, has also had a significant impact on time series forecasting research. Following this development, a series of Transformer-based models—such as Informer[2], Autoformer[3], PatchTST[4], and the Temporal Fusion Transformer (TFT)[5]—have been proposed to better capture long-range dependencies and complex periodic structures in time series data. These recent models have demonstrated substantial performance improvements over traditional statistical models and early LSTM-based approaches on various global benchmark datasets, including electricity demand, traffic flow, and logistics indices.

Despite these advances, empirical studies that systematically apply and evaluate such state-of-the-art deep learning models using domestic meteorological data remain limited. In particular, research based on Korea Meteorological Administration (KMA) Automated Synoptic Observation System (ASOS)[6] data is still at an early stage. Much of the existing domestic literature continues to rely on statistical methods or early recurrent neural network architectures, underscoring the need for comprehensive validation of recent deep learning models under domestic meteorological conditions and for assessing their potential practical applicability.

1.2 Research Problem and Objectives

This study addresses the empirical validation gap observed in the domestic meteorological forecasting literature regarding recent deep learning–based time series models. Using hourly air temperature data from the Seoul region, the study aims to conduct a systematic and quantitative comparison of predictive performance across different forecasting approaches. Rather than focusing on short-term prediction, this research considers a 72-hour (3-day) multi-step forecasting setting in order to evaluate model robustness and accuracy as the prediction horizon increases. Based on this motivation, the following research questions are formulated.

First, this study investigates whether state-of-the-art deep learning time series models—namely AutoTFT, AutoNHITS, and AutoPatchTST—exhibit statistically meaningful performance improvements over a traditional statistical model (SARIMA) and a climatology-based baseline in 72-hour multi-step temperature forecasting using KMA ASOS data.

Second, this study examines how architectural differences among deep learning models—specifically the MLP-based NHITS, the attention-based Temporal Fusion Transformer (TFT), and the patch-based PatchTST—affect their ability to capture seasonality and nonlinear variability in temperature time series, and seeks to identify the structural factors underlying these performance differences.

The primary objective of this study is to address these research questions by comparing the predictive performance of five models using hourly temperature observations from the Seoul station (STN 108), based on quantitative evaluation metrics such as MSE, RMSE, and MAE. Through this analysis, the study aims to provide empirical evidence supporting the adoption of deep learning models in the domestic meteorological domain and to establish a foundational data framework for future studies extending to multivariate inputs and multiple observation stations.

2. Literature Review

2.1 Development of Time Series Temperature Forecasting Methodologies

2.1.1 Traditional Statistical Model-Based Time Series Forecasting

The ARIMA model has long been regarded as a classical standard in time series analysis, providing a framework that mathematically models the autocorrelation structure of observed data to generate future predictions. For time series exhibiting pronounced periodic patterns, such as air temperature, the SARIMA

(Seasonal ARIMA) model—which explicitly incorporates seasonal components—has been widely employed.

Despite their extensive use, these linear statistical models suffer from inherent limitations. First, because they are fundamentally based on linear dependency assumptions, their ability to capture nonlinear meteorological dynamics—such as abrupt changes associated with frontal passages or variations in cloud cover—is limited. Second, as the forecast horizon extends, predicted values tend to converge toward the long-term mean, leading to a rapid increase in uncertainty and a corresponding decline in the informational value of long-range forecasts.

2.1.2 Transition to Deep Learning-Based Time Series Forecasting

To address the limitations of traditional statistical approaches, early deep learning–based forecasting methods primarily focused on recurrent neural networks (RNNs), particularly Long Short-Term Memory (LSTM) networks. LSTMs partially alleviated the long-term dependency problem by employing gated mechanisms that selectively retain and discard information within sequential data. However, difficulties related to limited parallelization and information degradation in very long sequences remained unresolved.

Since 2017, the dominant paradigm in time series forecasting has shifted toward Transformer-based architectures utilizing self-attention mechanisms. The three representative architectures considered in this study adopt distinct strategies for modeling temporal dependencies and complex patterns.

The TFT integrates an LSTM-based encoder with multi-head attention, enabling the simultaneous learning of local temporal patterns and long-range dependencies. In addition, it incorporates a Variable Selection Network and gating mechanisms to suppress irrelevant inputs while enhancing model interpretability.

NHITS decomposes time series into high-frequency and low-frequency components through multi-rate sampling and hierarchical interpolation, thereby achieving both computational efficiency and improved accuracy in long-horizon forecasting tasks.

PatchTST tokenizes time series by dividing them into fixed-length patches, preserving localized temporal information while substantially reducing the computational complexity of attention operations. This design allows the model to effectively leverage longer historical input sequences in forecasting.

2.2 Current Status and Limitations of Domestic Temperature Prediction Research

2.2.1 Current Status of Domestic Temperature Prediction Research

Previous domestic studies (see Table 1) have primarily employed early deep-learning architectures such as LSTM and have been conducted in short-term, region-/context-specific settings (e.g., single-city temperature, limited ASOS stations, urban maximum temperature, or summer-focused analyses).

Accordingly, it remains insufficiently examined whether recent long-horizon forecasting models such as PatchTST and N-HiTS, which have shown strong performance on international benchmarks, provide comparable gains under domestic meteorological conditions characterized by pronounced seasonality and complex topography.

Table 1. Summary of previous studies on domestic temperature forecasting

study	model	Target Variable	Forecast Horizon
Yun and Jeon (2017)	LSTM	Air temperature (Gwangju)	24 hours
Hong et al. (2021)[9]	LSTM	Meteorological variables at 9 ASOS stations	Short-term
Lim et al. (2020)[10]	Deep learning	Urban maximum temperature	Short-term
Cho et al. (2023)[11]	Deep learning + spatial interpolation	Summer temperature	Short-term

2.2.2 Limitations of Previous Studies

A review of prior domestic studies reveals several research gaps.

First, there is a limitation in model coverage. Existing domestic temperature forecasting studies have largely concentrated on recurrent neural network–based approaches such as LSTM, GRU, and CNN-LSTM, which became prevalent between approximately 2015 and 2018. In contrast, domestic applications of more recent time series forecasting models—including TFT, NHITS, and PatchTST, which have attracted increasing attention since 2021—remain difficult to identify, particularly in studies using Korean meteorological data.

Second, limitations are observed in the forecasting horizon. Most previous studies have focused on short-term forecasts within 24 hours, and systematic evaluations of deep learning model performance in medium-range forecasting settings extending to 72 hours or longer are largely absent.

Third, there is a lack of systematic comparative analysis. While many studies examine the applicability of individual models, relatively few studies conduct quantitative comparisons of multiple state-of-the-art deep learning models alongside statistical baselines under identical datasets and experimental conditions.

2.2.3 Distinctiveness of This Study

In response to the research gaps identified above, this study differentiates itself from existing work in several respects.

First, this study presents an empirical comparative analysis applying recent deep learning time series models—specifically TFT, NHITS, and PatchTST, all proposed after 2021—to domestic Korea Meteorological Administration (KMA) ASOS data. Although these models have demonstrated strong performance on international benchmark datasets, their applicability to domestic temperature time series—characterized by combined diurnal and annual seasonality as well as abrupt fluctuations associated with frontal passages—has not yet been sufficiently examined.

Second, this study adopts a medium-range forecasting horizon of 72 hours (3 days), thereby extending beyond the 24-hour short-term prediction range that has been the primary focus of most domestic studies.

Third, this study systematically compares a climatology-based baseline, SARIMA, and three deep learning models under a unified dataset and experimental setting, enabling a quantitative assessment of performance improvements achieved by deep learning approaches relative to traditional statistical models.

Finally, this study enhances reproducibility by employing the Auto-Hyperparameter Optimization (Auto-HPO) functionality provided by the NeuralForecast library, thereby establishing a consistent experimental framework that can be utilized in future comparative studies.

3. Method (Research Methodology)

3.1 Data Collection and Preprocessing

This study used hourly data from the Automated Synoptic Observation System (ASOS) observed at the Seoul meteorological station (STN 108) for the period from January 1, 2015, to November 21, 2025. The raw data were refined and preprocessed according to the procedures described below.

Special values defined by Korea Meteorological Administration (KMA) specifications (e.g., −9, −9.0) were regarded as missing observations. The time series was then reconstructed using a continuous hourly datetime index. Missing values were imputed using linear interpolation based on temporal information (method = “time”). The linear interpolation formula is given as shown in Eq. (1):

$$y = y_0 + (x - x_0)\frac{y_1 - y_0}{x_1 - x_0} \quad (1)$$

where x denotes the missing time point, while x_0 and x_1 represent the valid observation time points adjacent to the missing interval. Missing values at the beginning and end of the time series, where interpolation is not applicable, were imputed using forward and backward filling (ffill/bfill). As a result, the final analysis dataset contained no remaining missing values.

Since the purpose of this experiment is to compare predictive performance across different model architectures, a univariate framework was maintained, with air temperature (TA) specified as the sole input and target variable. Data normalization for ensuring training stability of the deep learning models was not performed manually; instead, it was carried out automatically using scalers such as TemporalNorm provided within the NeuralForecast library.

Data partitioning was designed in consideration of the target 72-hour (3-day) multi-step forecasting structure. The period from January 1, 2015, to November 18, 2023, was primarily used for calculating the climatology baseline, deriving long-term statistical measures, and conducting exploratory analyses to identify seasonal and trend characteristics of the time series. The main model training period was limited to the interval from 00:00 on November 19, 2023, to 23:00 on November 18, 2025, during which the statistical model (SARIMA) and the deep learning–based models AutoNHITS, AutoTFT, and AutoPatchTST were trained. Finally, a 72-hour period from 00:00 on November 19, 2025, to 23:00 on November 21, 2025, was designated as the final test interval, and all models were configured to generate predictions for the same time window.

With respect to the training–validation split, a strategy was adopted in which the final 72 hours (72) of the training period were used as the validation set for hyperparameter tuning of the deep learning Auto models. Specifically, from the data spanning November 19, 2023, to November 18, 2025, the last 72 observations were reserved for validation, and the preceding portion was used to train each trial with different parameter configurations. The trial that yielded the minimum loss on the validation set (based on MAE) was selected as the optimal hyperparameter configuration. Using the selected configuration, the model was retrained on the entire training period and subsequently used to generate 72-hour forecasts for the final test period.

3.2 Prediction Problem Definition and Baseline Models

The objective of this study is to perform 72-hour-ahead (3-day) multi-step forecasting of hourly air temperature (TA) at the Seoul station using a univariate setting. Model performance is evaluated against two baseline approaches: a climatology lookup baseline and a statistical SARIMA baseline.

3.2.1 Climatology Baseline

As a minimal-reference baseline, a climatology model was constructed from historical hourly temperature observations spanning 2015 to November 18, 2023. For each timestamp, the day-of-year and hour-of-day were computed, and the climatological mean temperature was estimated for each $(day-of-year, hour-of-day)$ pair to form a lookup table. The climatology prediction at a test time point is defined as shown in Eq. (2) and (3):

$$\hat{y}_t^{\text{clim}} = \mu_{\text{doy}(t),\,\text{hod}(t)} \tag{2}$$

$$\mu_{d,h} = \mathbb{E}[y_t \mid \text{doy}(t) = d,\ \text{hod}(t) = h] \tag{3}$$

This baseline requires little to no model training and serves as a minimum performance benchmark commonly used in operational meteorology.

3.2.2 SARIMA Baseline

As a statistical baseline, the Seasonal AutoRegressive Integrated Moving Average (SARIMA) model was employed. Since Seoul's hourly temperature series exhibits pronounced diurnal periodicity, the seasonal period was fixed to . Stationarity assessment (e.g., ADF test) and seasonal decomposition (details reported in the appendix) suggest the presence of trend and seasonal components; therefore, the differencing orders were explored within a limited range focusing on small values.

Model orders were selected via a constrained grid search over candidate $\text{SARIMA}(p,d,q)\times(P,D,Q)_{24}$ specifications, based on model fitting results and diagnostics. The selected SARIMA specification is given in Eq. (4):

$$\text{SARIMA}(3,0,1)\times(0,0,1)_{24} \tag{4}$$

Using the chosen configuration, the SARIMA model was refitted for the full training period (November 19, 2023–November 18, 2025). Subsequently, 72-step-ahead (72-hour-ahead) forecasts $\{\hat{y}_{t+h|t}\}_{h=1}^{72}$ were generated for the test period.

3.3 Deep Learning–Based Auto Models

To benchmark modern deep learning approaches, this study employed the automatic hyperparameter optimization (Auto) variants of NHITS, Temporal Fusion Transformer (TFT), and PatchTST, implemented in the NeuralForecast library. All deep learning models were trained in a univariate setting using only past temperature values (i.e., no additional covariates), and the forecasting horizon was fixed to $h = 72$ steps (72 hours).

3.3.1 AutoNHITS, AutoTFT, and AutoPatchTST

AutoNHITS extends the N-BEATS family by modeling multiple temporal scales through stacked blocks. AutoTFT combines an LSTM-based sequence encoder with multi-head attention, while AutoPatchTST segments the series into fixed-length patches and applies a Transformer to learn long-range dependencies via patch-level representations.

Hyperparameter optimization was performed using the Ray Tune backend. The search space included the input window length, hidden dimension, dropout rate, learning rate, batch size, and window batch size. Each trial used a consistent training–validation protocol in which the last 72 hours of the training period were held out as the validation set. The hyperparameter configuration achieving the lowest validation MAE was selected. With the selected configuration, each model was retrained on the full training dataset and then used to generate 72-hour-ahead forecasts for the test period.

3.4 Training Setup and Evaluation Metrics

All models were evaluated and compared using the same test period, spanning a total of 72 hours from 00:00 on November 19, 2025, to 23:00 on November 21, 2025. Forecasting accuracy was assessed using Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the Coefficient of Determination (R^2). The definitions of MSE, RMSE, and MAE are given in Eqs. (5)–(7), respectively.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{5}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{6}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{7}$$

The coefficient of determination (R^2) was used as a measure of the proportion of variance in the observed values that is explained by the predictions.

Due to the characteristics of air temperature forecasting, conventional Mean Absolute Percentage Error (MAPE) can be distorted when temperatures are close to 0 °C, as the denominator approaches zero. Accordingly, for the deep learning models, temperature values were converted to Kelvin units prior to computing MAPE, and a Kelvin-based MAPE was employed in Eq. (8) .

$$\mathrm{MAPE}(K) = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{(y_i+273.15)-(\hat{y}_i+273.15)}{y_i+273.15}\right| \tag{8}$$

In contrast, at the initial implementation stage, MAPE values for the Climatology and SARIMA models were computed based on temperatures expressed in degrees Celsius, resulting in a substantial discrepancy in scale compared to the Kelvin-based MAPE used for the deep learning models. Owing to this inconsistency, the present study focuses primarily on MSE, RMSE, MAE, and R^2 for performance comparison across models. The accurate recalculation of MAPE(K) and the application of a unified definition across all models are identified as necessary tasks to be addressed in future work.

4. Results (Empirical Analysis)

4.1 Data Characteristics and Exploratory Analysis

The hourly air temperature time series observed at the Seoul station (STN 108) reflects typical characteristics of a mid-latitude urban climate, exhibiting both long-term seasonal patterns associated with annual variability and

pronounced diurnal patterns characterized by recurring daily maximum and minimum temperatures. The ASOS data used in this study were confirmed to capture these meteorological characteristics, underscoring the necessity of adopting model structures that explicitly account for seasonality and periodicity.

Although quantitative results from seasonal decomposition and stationarity tests have not yet been fully consolidated, the analysis was conducted under the assumption that the temperature time series is non-stationary and contains both trend and seasonal components. This assumption was also reflected in the design of the SARIMA model, in which the differencing orders were selected accordingly and a seasonal period of 24 was adopted to represent diurnal variation. More rigorous results from the Augmented Dickey–Fuller (ADF) test, as well as detailed statistics of the decomposed trend, seasonal, and residual components, may be presented in the main text or appendix in future revisions.

4.2 Comparison of 72-Hour Temperature Forecasting Performance

The forecasting performance of each model over the 72-hour test period, from November 19 to November 21, 2025, is summarized in Table 2.

Table 2. Performance comparison of models for 72-Hour Air Temperature Forecasting

model	MSE	RMSE	MAE	R^2	MAPE (K)
Climatology	23.04	4.80	4.15	0.216	-
SARIMA	20.44	4.52	3.70	0.304	-
AutoPatchTST	12.46	3.53	3.11	0.361	1.11%
AutoNHITS	6.80	2.61	2.10	0.651	0.75%
AutoTFT	4.39	2.10	1.65	0.775	0.59%

* MAPE(K) represents Kelvin-based Mean Absolute Percentage Error. For the Climatology and SARIMA models, MAPE values were not computed due to inconsistencies in definition.

The forecasting performance of each model over the 72-hour test period from November 19 to November 21, 2025, can be summarized as follows. The climatology model yielded an MSE of approximately 23.04, an RMSE of 4.80, an MAE of 4.15, and an R^2 value of 0.216. These results indicate that a simple climatology-based approach is able to explain a limited portion of temperature variability, but the associated prediction errors remain relatively large.

For the SARIMA model, the MSE, RMSE, MAE, and R^2 were approximately 20.44, 4.52, 3.70, and 0.304, respectively. Although this represents a modest improvement over the climatology baseline, the prediction errors are still substantial, and the explanatory power remains limited.

Among the deep learning–based Auto models, AutoNHITS achieved an MSE of approximately 6.80, an RMSE of 2.61, an MAE of 2.10, and an R^2 of 0.651. Compared with SARIMA, this corresponds to an error reduction of roughly 50% in terms of RMSE, indicating that AutoNHITS is able to capture a substantial proportion of the overall temperature variability. AutoTFT exhibited the best performance among the three deep learning models, with an MSE of approximately 4.39, an RMSE of 2.10, an MAE of 1.65, and an R^2 of 0.775. In particular, AutoTFT achieved the lowest RMSE and MAE values, outperforming not only the statistical baselines but also the other deep learning models, including AutoNHITS and AutoPatchTST.

AutoPatchTST recorded an MSE of approximately 12.46, an RMSE of 3.53, an MAE of 3.11, and an R^2 of 0.361. While its performance represents an improvement over the climatology and SARIMA models, it remains inferior to that of AutoNHITS and AutoTFT.

When comparing Kelvin-based MAPE values, AutoNHITS, AutoTFT, and AutoPatchTST yielded MAPE(K) values of approximately 0.75%, 0.59%, and 1.11%, respectively, indicating that absolute prediction errors over the 72-hour horizon are very small relative to the absolute temperature scale. In contrast, MAPE values for the climatology and SARIMA models were abnormally large due to differences in definition arising from the initial implementation, making direct comparison with the Kelvin-based MAPE of the deep

learning models inappropriate. Accordingly, this study primarily discusses relative model performance based on MSE, RMSE, MAE, and R^2.

Overall, all three Auto deep learning models substantially reduced prediction errors compared with the climatology and SARIMA baselines. AutoTFT achieved an error reduction of approximately 50% relative to the statistical model in terms of RMSE and MAE. From the perspective of R^2, AutoNHITS($R^2 \approx 0.65$), AutoTFT ($R^2 \approx 0.77$), and AutoPatchTST ($R^2 \approx 0.38$) all exhibited markedly higher explanatory power than the climatology ($R^2 \approx 0.22$) and SARIMA ($R^2 \approx 0.30$) models, confirming that the deep learning approaches are considerably more effective in capturing temperature variability.

4.3 Descriptive Summary of Visual Comparisons

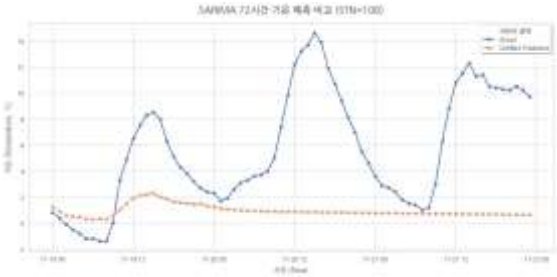

Figure 1. Comparison of Observed and SARIMA-Predicted Air Temperature over the 72-Hour Test Period

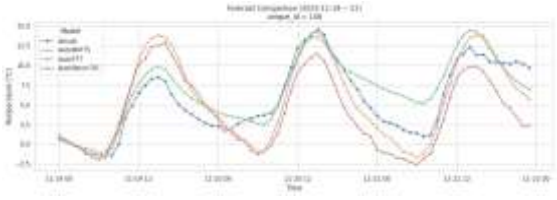

Figure 2. Comparison of Observed and Predicted Air Temperature Using AutoNHITS, AutoTFT, and AutoPatchTST over the 72-Hour Test Period

When the observed temperature time series and the model-predicted series are overlaid, the climatology and SARIMA models generally follow smooth temperature variation patterns (Figure 1). However, during periods of rapid temperature increases or decreases, these statistical baselines exhibit delayed responses and fail to accurately reproduce the timing and magnitude of peaks and troughs (Figure 1). In particular, over the 72-hour forecast horizon, SARIMA tends to regress toward the mean when abrupt temperature changes occur, leading to large errors in daily maximum and minimum temperatures (Figure 1).

In contrast, AutoTFT captures both the trend and diurnal patterns more effectively even under abrupt short-term changes, reproducing the timing and magnitude of peaks and troughs with relatively higher fidelity (Figure 2). AutoNHITS also tracks the overall pattern more smoothly and stably than SARIMA; although mild over-smoothing is observed in some intervals, its deviations from the observed series remain smaller overall than those of the statistical model (Figure 2). For AutoPatchTST, the predicted trajectories appear sensitive to the input window length, patch size, training steps, and the hyperparameter search range adopted in this study; in some trials, highly oscillatory prediction patterns are observed, which may help explain its comparatively weaker performance relative to the other deep learning models (Figure 2). Overall, these qualitative visual observations align with the tendencies indicated by the quantitative evaluation metrics (Figures 1–2).

5. Discussion

5.1 Interpretation of Model Performance

The empirical results demonstrate that the deep learning–based Auto models clearly outperform traditional baselines such as the climatology model and SARIMA. In particular, AutoTFT achieved an error reduction of more than approximately 50% in terms of RMSE compared with SARIMA, even under the relatively constrained setting of a single station and univariate input. This finding suggests that Transformer-based deep learning models can provide meaningful performance improvements for short- to medium-range forecasting problems, such as 72-hour air temperature prediction in the Seoul region.

These results can be naturally interpreted in light of the nonlinear characteristics and long-range dependencies

inherent in temperature time series. In addition to seasonal and diurnal cycles, air temperature is influenced by a range of physical processes, including frontal passages, variations in cloud cover, and local wind fields, resulting in complex and nonlinear dynamics. Linear models such as SARIMA have limited capacity to represent such nonlinear structures and tend to regress toward the mean as the forecast horizon increases. In contrast, AutoNHITS and AutoTFT leverage deep neural network architectures to learn nonlinear relationships within the time series and to capture patterns across multiple temporal scales, enabling more stable predictions of trends and periodic components even over a relatively long horizon of three days.

Furthermore, AutoTFT integrates an LSTM-based sequence encoder with a multi-head attention–based fusion layer, allowing the model to assign greater weight to temporally important time points or intervals. Although this study considered only a univariate setting, such architectural characteristics appear to be well suited to temperature data with pronounced seasonal and diurnal patterns. Internal scaling and gating mechanisms may also have contributed to improved training stability and generalization performance.

By contrast, the relatively weaker performance of AutoPatchTST is likely attributable not to inherent limitations of the model itself, but rather to the specific choices of input window length, patch size, and hyperparameter search range adopted in this study, which may not have been fully optimized for the given data and forecasting objective ($h = 72$). Patch-based time series Transformers are known to be sensitive to patch design and training configurations, and there remains potential for performance improvement through more refined architectural tuning and extended training. Nevertheless, such improvements require further experimentation and validation. At the current stage, the results indicate that AutoTFT and AutoNHITS represent more stable and interpretable options from a practical application perspective.

5.2 Academic and Practical Implications

Although this study focuses on a single station and a univariate temperature forecasting setting, it quantitatively demonstrates that recent deep learning–based time series models can significantly outperform traditional statistical approaches. This finding indicates that there is sufficient motivation to adopt Transformer-based models even at the single-station level, prior to extending the framework to more complex spatial models or multivariate configurations. From an academic perspective, the results provide empirical evidence that nonlinear deep learning architectures can yield substantial improvements in predictive performance over conventional ARIMA-family models, even for climate time series characterized by strong seasonality and trend components.

From a practical standpoint, national meteorological service agencies such as the Korea Meteorological Administration already provide high-resolution forecasts through numerical weather prediction (NWP) models. Nevertheless, discrepancies between forecasts and actual observations frequently arise due to local microclimate effects, complex terrain, and urban heat island phenomena. Data-driven deep learning forecasters such as those examined in this study have the potential to serve as post-processing modules that correct or refine NWP-based forecasts. For example, hybrid architectures that jointly incorporate ASOS observations at specific locations and outputs from NWP models could be designed to provide short-term correction models for multiple meteorological variables, including temperature, precipitation, and wind speed. Such research directions are closely linked to practical applications in urban-scale energy management, heatwave and cold-wave warning systems, and weather-sensitive services in the health and welfare sectors.

In addition, the use of Auto-HPO pipelines such as NeuralForecast contributes to improved research productivity and reproducibility. In this study, AutoNHITS, AutoTFT, and AutoPatchTST enabled the exploration of diverse hyperparameter configurations with relatively minor code modifications, allowing researchers to achieve competitive performance without excessive effort devoted to model implementation details. This approach lowers the barrier to conducting empirical experiments in domestic meteorological data science and

provides a shared experimental framework through which results can be systematically compared and accumulated across different institutions and research groups.

5.3 Limitations and Future Research Directions

Despite its contributions, this study has several limitations that should be acknowledged, along with corresponding directions for future research. First, the analysis is limited to a single observation site, namely the Seoul station (STN 108), and the generalizability of the findings to other locations or climatic regimes has not yet been verified. Future studies should replicate the same experimental design across stations representing diverse climatic characteristics, such as inland cities, coastal regions, and mountainous areas, to examine whether models such as AutoTFT maintain consistent performance advantages. Furthermore, a natural extension of this work would involve designing models that incorporate spatial information across nationwide ASOS stations, including combinations with Graph Neural Networks (GNNs), ConvLSTM architectures, or spatiotemporal Transformers.

Second, although a univariate forecasting framework was adopted in this study to ensure clarity in structural model comparison, air temperature in reality is influenced by a wide range of factors, including humidity, wind speed, atmospheric pressure, surface temperature, upper-level pressure systems, and large-scale climate indices. Since AutoTFT is inherently designed to handle multivariate and multi-source inputs, future research should extend the current framework to multivariate forecasting settings that incorporate such potential covariates. This would allow for an assessment of whether the performance gains observed under the univariate setting are preserved or further enhanced in more realistic multivariate environments.

Third, the results presented in this study are based on a single 72-hour forecast window spanning November 19 to November 21, 2025. To draw more robust conclusions, future work should adopt rolling backtesting strategies that cover multiple seasons and a wider range of meteorological conditions after 2023. For example, repeated forecasts over multiple sliding 72-hour windows could be used to analyze average performance and variability across time. In particular, evaluating model performance under extreme events—such as heatwaves, cold spells, and abrupt temperature transitions—would constitute a critical validation step for potential operational deployment in meteorological services.

Fourth, all models considered in this study provide point forecasts only and do not quantify predictive uncertainty. Given the nature of meteorological forecasting, probabilistic predictions that provide forecast intervals or full predictive distributions are often more informative than single-point estimates. Future research should therefore explore extensions incorporating quantile-based loss functions, ensemble-based uncertainty estimation, or Bayesian deep learning approaches, with the aim of quantifying uncertainty in temperature forecasts and leveraging such information for weather warning systems and decision support.

While recognizing these limitations, this study nonetheless holds both academic and practical significance in demonstrating that state-of-the-art deep learning time series models, including AutoTFT, can deliver substantial performance improvements over traditional statistical models and climatology-based baselines for the 72-hour air temperature forecasting problem in Seoul using KMA ASOS data.

6. Conclusion

This study compared and evaluated the temperature forecasting performance of recent deep learning–based time series models—AutoNHITS, AutoTFT, and AutoPatchTST—using hourly meteorological observation data provided by the Korea Meteorological Administration, against a traditional statistical model (SARIMA) and a climatology-based baseline.

The results show that AutoTFT achieved the best overall performance, with an MSE of 4.39, an RMSE of 2.10, an MAE of 1.65, and an R^2 of 0.775, corresponding to an error reduction of approximately 54% in terms of RMSE compared with SARIMA. AutoNHITS also significantly outperformed the statistical model, recording an RMSE of 2.61. These findings demonstrate that the Temporal Fusion Transformer architecture, which combines an LSTM-based sequence encoder with multi-

head attention, is capable of effectively learning the nonlinear relationships and seasonal structures inherent in air temperature time series.

The results of this study provide practical insights for various application domains that require accurate meteorological forecasting, such as smart grid energy demand management and the prevention of cold damage to agricultural crops. Future research is expected to extend this work toward multivariate forecasting models that integrate multiple meteorological variables and multi-station data, as well as toward probabilistic forecasting approaches that enable quantitative uncertainty estimation.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L.
Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention Is All You Need," in Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), pp. 5998–6008, 2017.

[2] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, "Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting," *arXiv* preprint arXiv:2012.07436 [cs.LG], submitted Dec. 14, 2020, last revised Mar. 28, 2021. doi: 10.48550/arXiv.2012.07436.

[3] H. Wu, J. Xu, J. Wang, and M. Long, "Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting," arXiv preprint arXiv:2106.13008 [cs.LG], submitted Jun. 24, 2021, last revised Jan. 7, 2022. doi: 10.48550/arXiv.2106.13008.

[4] Y. Nie, N. H. Nguyen, P. Sinthong, and J. Kalagnanam, "A Time Series is Worth 64 Words: Long-term Forecasting with Transformers (PatchTST)," arXiv preprint arXiv:2211.14730 [cs.LG], submitted Nov. 27, 2022, last revised Mar. 5, 2023. doi: 10.48550/arXiv.2211.14730.

[5] B. Lim, S. O. Arik, N. Loeff, and T. Pfister, "Temporal Fusion Transformers for Interpretable Multi-horizon Time Series Forecasting," International Journal of Forecasting, vol. 37, no. 4, pp. 1748–1764, 2021, doi: 10.1016/j.ijforecast.2021.03.012.

[6] Korea Meteorological Administration (KMA), "Surface Synoptic Observation Data (ASOS)," dataset.

[7] C. Challu, K. G. Olivares, B. N. Oreshkin, F. Garza,
M. Mergenthaler-Canseco, and A. Dubrawski, "N-HiTS: Neural Hierarchical Interpolation for Time Series Forecasting," *arXiv* preprint arXiv:2201.12886 [cs.LG], submitted Jan. 30, 2022, last revised Nov. 29, 2022 (v6), (accepted at AAAI-23).
doi: 10.48550/arXiv.2201.12886.

[8] Yoon, J., & Jeon, M. (2017-11-24). Temperature
Forecasting Model by using Deep Learning Technology based on LSTM. Conference of the Institute of Electronics and Information Engineers (IEIE), Incheon, Republic of Korea.

[9] Hong, S., Kim, J. H., Choi, D. S., & Baek, K. (2021). Development of Surface Weather Forecast Model by using LSTM Machine Learning Method.
Atmosphere, 31(1), 73–83. https://doi.org/10.14191/ATMOS.2021.31.1.073

[10] Lim, J., Cho, D., Yoo, C., & Lee, Y. (2020, October
28). Improving next-day urban maximum temperature forecasts in a deep learning–based local forecasting model. Proceedings of the Korean Meteorological Society Conference, Seoul, Korea.

[11] Cho, D., Lim, J., & Jeong, S. (2023, November 1). Spatial refinement of summer short-term

temperature forecasts by integrating deep learning and spatial interpolation methods. Proceedings of the Korean Meteorological Society Conference, Busan, Korea.

Author Biography

Hyun Koo Lee

Master of Engineering in AI Big Data, aSSIST (Seoul School of Integrated Sciences & Technologies)

RESEARCH INTERESTS
- Large Language Models (LLMs)
- Transformers
- Time Series Analysis

E-mail: hyunkoo.lee@stud.assist.ac.kr

Nak Hyun Jung

Visting Professor, Seoul School of Integrated Sciences & Technologies (aSSIST)
Research Professor, The Institute for Industrial Policy Studies (IPS)
Head of Process Management Team, Mirae Asset Securities
Ph.D. in Business Administration, aSSIST University
Research Interests: AI, Finance, Time Series, Business Strategy
E-mail: nhjung@assist.ac.kr

The Role of Foundry for High-Performance AI Semiconductors

Young-hwa Kwun ,Bo-young Kim

ABSTRACT

Recently, the AI semiconductor market has been experiencing rapid advancements across multiple dimensions, including computational performance, energy efficiency, and integration density. As a result, not only the design capabilities of fabless companies but also the manufacturing capabilities of foundry companies that implement these designs have become more critical than ever. In particular, achieving high yield, possessing advanced process technologies, and securing advanced packaging capabilities have emerged as key factors for realizing high-performance AI semiconductors.

This study selects TSMC, the world's largest foundry company, as a case and analyzes how the company secures manufacturing competitiveness in AI semiconductors through yield management, process control, and advanced packaging technologies. The findings indicate that TSMC enhances AI semiconductor performance through integrated management across the entire value chain—from design and manufacturing to packaging and mass production—while maintaining stable production capabilities and making proactive R&D investments in advanced packaging. These results demonstrate that the integrated management of yield, process, and packaging constitutes a core mechanism for improving AI semiconductor performance.

☞ keyword : AI semiconductor, Fabless company, Foundry company, Manufacturing capability, TSMC

1. Introduction

Artificial Intelligence(AI) refers to a technology that imitates and extends human intelligence [1], and the AI market is currently growing at an annual rate exceeding 20%. Along with advances in science and technology, AI is rapidly transforming nearly every aspect of daily life and is increasingly approaching human-level capabilities [2]. The changes driven by AI are already profound and are expected to accelerate further in the future. At the center of this transformation lies AI semiconductors [3].

All chips used in AI applications can be broadly categorized as AI semiconductors [4]. As AI is being adopted across diverse industrial sectors, the demand for AI semiconductors has correspondingly increased [5]. In particular, significant advances in deep learning technologies have led to the development of a wide range of AI semiconductors [6]. AI semiconductors are specifically designed to accelerate applications based on artificial neural networks (ANNs) [7] and play a critical role across AI models and devices. Without high-performance AI semiconductors, most AI functionalities would be severely limited.

Moreover, the role of foundries is essential in realizing high-performance AI semiconductors. Achieving superior chip performance requires not only advanced design capabilities but also excellence in manufacturing processes. The continuous emergence of new semiconductor technologies has further increased the complexity of manufacturing facilities [8]. Consequently, close collaboration between fabless companies and foundries has become increasingly critical to the development of high-performance AI semiconductors.

This study therefore examines how advanced foundry technologies determine the core performance of high-performance AI semiconductors, using TSMC as a representative case. While prior studies have primarily focused on AI semiconductors from a design perspective [9][10], manufacturing capability is equally, if not more, decisive. Even the most sophisticated chip designs cannot achieve their intended performance without advanced manufacturing execution. Accordingly, the foundry's role is fundamental to the realization of high-performance AI semiconductors.

This study contributes to the literature by approaching AI semiconductors from a manufacturing perspective. Furthermore, the findings derived from the analysis of TSMC's technological capabilities are expected to provide meaningful academic and practical implications.

2 Main Body

2.1. AI Semiconductor Companies

AI semiconductor firms can be broadly classified into three categories. The first category consists of Integrated Device Manufacturers (IDMs). The second category comprises fabless firms. The final category includes Big Tech companies. This study examines AI semiconductor firms according to each of these categories.

2.1.1 Integrated Device Manufacturer (IDM) Firms

IDM firms are vertically integrated companies that conduct design, manufacturing, and packaging in-house. Representative examples include Samsung Electronics and Intel [11]. These firms also develop their own AI semiconductors.

Samsung Electronics operates across memory semiconductors, fabless, and foundry businesses, and conducts AI semiconductor development within each division. In the memory semiconductor division, Samsung is developing Processing-in-Memory(PIM) technologies and is actively researching neuromorphic semiconductors. Under the conventional von Neumann architecture, where the processor and memory are physically separated [12], large-scale data transfers inevitably result in bottlenecks and substantial power consumption [13]. Consequently, in the era of AI, neuromorphic semiconductors that integrate processing and memory are increasingly recognized as a critical technological direction.

In addition, Samsung previously collaborated with Naver through its System LSI division to develop an NPU chip called Mach-1 for data center inference applications. NPUs represent the forefront of AI semiconductor development [14]. Currently, Samsung continues the development of its NPU technology independently. Furthermore, within its foundry division, Samsung has begun mass production of 2nm process chips based on Gate-All-Around(GAA) technology from the end of 2025. Recently, Samsung Foundry reportedly secured a large-scale contract to manufacture Tesla's A16 chips using its 2nm process [15]. These chips are expected to be used primarily in Tesla's autonomous driving systems and robotics platforms. While the previous A15 chip was manufactured by TSMC, production has shifted to Samsung Foundry for the A16 generation.

Intel also launched the chiplet-based Gaudi 3 accelerator in 2024. Gaudi was developed based on the technology of Habana Labs, an Israeli AI semiconductor company acquired by Intel in 2019. This chip is positioned primarily for inference workloads and is reported to offer competitive performance at a lower price compared to NVIDIA's H100, along with the advantage of an open software ecosystem. Intel is also reportedly preparing to release Gaudi 4 in the near future.

Moreover, Intel's recently released Core Ultra processors integrate internally developed GPUs and NPUs on-chip. These processors are designed for AI PCs and support on-device AI functionalities. In parallel, Intel Foundry is preparing advanced manufacturing capabilities, including a GAA-based 1.4nm process, to enable the production of high-performance AI semiconductors.

2. 1. 2 Fabless Firms

Fabless firms are design-specialized companies that do not operate their own fabrication facilities. Nevertheless, they have the advantage of being able to commercialize products under their own brand names. This section examines major fabless companies engaged in the design of AI semiconductors.

First, among fabless firms, NVIDIA is the most prominent company in the AI semiconductor sector. NVIDIA is estimated to account for over 90% of the GPU market for data centers. The company currently releases new product generations on an annual basis. The GPU integrated into NVIDIA's Blackwell accelerator is the B200, while the next-generation product, Blackwell Ultra, incorporates the B300 GPU. After designing its GPUs, NVIDIA outsources fabrication to TSMC. High Bandwidth Memory(HBM) is supplied primarily by SK hynix, and advanced packaging is subsequently performed at TSMC to complete the accelerator products. NVIDIA provides not only hardware such as GPUs but also a comprehensive full-stack solution, including equipments, software platforms and services. As a result, customers face high switching costs, and NVIDIA's AI ecosystem has become highly entrenched.

Second, AMD is rapidly emerging as NVIDIA's primary competitor. In an effort to capture market share in the GPU market dominated by NVIDIA, AMD has developed ROCm(Radeon Open Compute), a software platform comparable to NVIDIA's CUDA(Compute Unified Device Architecture). However, due to the strong lock-in effect of NVIDIA's ecosystem, AMD GPUs have been adopted primarily by customers who experience supply constraints in

accessing NVIDIA products. Nevertheless, AMD's market share has shown a continuous upward trend. AMD's recently introduced accelerator, MI350X, is based on the CDNA4 architecture and is manufactured using a 3nm process. It has been reported to offer advantages over NVIDIA's B200 in terms of both performance and price competitiveness.

Third, Qualcomm, widely known for its Snapdragon system-on-chip(SoC) platforms, is another representative fabless company. Qualcomm integrates its proprietary Hexagon NPU within its SoCs to enable AI functionalities. Snapdragon platforms are deployed across a wide range of applications, including smartphones, AI-enabled laptops, smart glasses, and autonomous vehicles. In particular, Qualcomm has established a strong competitive position in the on-device AI segment.

Fourth, Broadcom is a fabless firm with particular strength in communication networks and customized semiconductor solutions. Broadcom primarily provides ASIC design services tailored to customer-specific requirements. Typically, new circuits are first implemented and validated using Field-Programmable Gate Arrays(FPGAs) to verify functionality and performance, and are subsequently converted into ASICs for final deployment. FPGAs offer high flexibility and reconfigurability [16], and while they provide strong performance, their power consumption is relatively moderate [17]. However, ASICs significantly outperform FPGAs in terms of both performance and power efficiency.

Many Big Tech companies outsource the design and development of customized ASICs to Broadcom in order to optimize chips for their own end products. Recently, a notable trend has emerged in data center inference workloads, where there is increasing interest in replacing NVIDIA GPUs with NPUs or ASICs.

Finally, MediaTek, the largest fabless semiconductor company in Asia, represents another important player. MediaTek develops application processors(SoCs) for edge devices such as smartphones and Internet of Things(IoT) devices and integrates NPUs within these chips to support AI functionalities.

2. 1. 3 Big Tech Firms

The development of AI semiconductors has also become increasingly active among Big Tech firms. By developing in-house chips, these companies can reduce their dependence on external supply chains while simultaneously optimizing semiconductor performance for their own products and services. Moreover, in-house semiconductor development is widely recognized as offering significant cost advantages. For these reasons, the trend of semiconductor development by Big Tech firms is expected to continue.

This section reviews the current status of AI semiconductor development among major Big Tech companies.

First, Apple has pursued an aggressive strategy of internalizing semiconductor development to design nearly all chips required for its products. Apple's AI semiconductor technology is widely recognized through its Neural Engine, which represents a type of neural processing unit(NPU). The Neural Engine is embedded in both the A-series and M-series processors and supports on-device AI functionalities.

Second, Google has developed its own AI accelerator known as the Tensor Processing Unit(TPU), which is also a form of NPU. TPUs are utilized across both training and inference workloads within Google's data center infrastructure.

Third, Microsoft has introduced an AI chip known as Maia, which is designed to support both training and inference tasks. This chip is deployed within Microsoft's cloud infrastructure, particularly in services such as Azure and Copilot.

Fourth, Amazon, the world's leading cloud service provider, has developed proprietary AI chips for use within its cloud ecosystem. Amazon has introduced Trainium for training workloads and Inferentia for inference tasks, and these chips are actively utilized across its cloud services.

Fifth, Tesla has developed custom AI semiconductors for autonomous driving applications. The company's Full Self-Driving(FSD) chip, known as the A15, was manufactured by TSMC and was deployed in Tesla's autonomous driving systems. In addition, Tesla previously developed a dedicated AI supercomputer called Dojo to efficiently train large-scale data collected from its vehicles. The Dojo system utilized the D1 chip and was originally intended to replace NVIDIA's A100 accelerators. However, Tesla has recently decided to discontinue the Dojo project, and development of the D2 chip has also been reportedly halted.

Finally, Meta has developed its own AI accelerators while also operating the open-source large language model (LLM) known as LLaMA. Although Meta continues to procure a substantial number of GPUs from NVIDIA, it has introduced the Meta Training and Inference Accelerator (MTIA), which, as its name suggests, is designed to support both training and inference workloads.

In addition to U.S.-based Big Tech firms, several Chinese companies are also actively developing AI semiconductors. Huawei has developed the Ascend series for data center applications and integrates NPUs into its Kirin chips for smartphones. Similarly, companies such as Alibaba and Baidu have introduced their own AI semiconductors and are deploying them within their respective product and service ecosystems.

(Table 1) Classification of AI Semiconductor Firms and their Characteristics

Category	IDM Firms	Fabless Firms	Big tech Firms
AI Semiconductor Firms	Samsung Electronics, Intel	NVIDIA, AMD, Qualcomm, Broadcom, MediaTek	Apple, Google, MS, Amazon, Tesla, Meta, and other Chinese companies
Types of AI semiconductor	Samsung Electronics: PIM, Mach 1 Intel: Gaudi, NPU	NVIDIA: GPU AMD: GPU Qualcomm: NPU Broadcom: ASIC MediaTek: NPU	Apple: Neural Engine Google: TPU MS: Maia Amazon: Trainium, Inferentia Tesla: A15, D1 Meta: MTIA Other Chinese companies: Various AI semiconductors
AI semiconductor manufacturing	Capable of in-house manufacturing through proprietary foundries	Mainly in TSMC	Mostly manufactured by TSMC, and Chinese firms mainly rely on SMIC

2. 2 Foundry Firm

2. 2. 1 Definition and Types of Foundries

In the semiconductor industry, a foundry refers to a firm that manufactures semiconductor chips designed by customers. Foundry companies do not commercialize chips under their own brand, whereas customers retain ownership of their branded products.

Historically, semiconductor businesses were typically vertically integrated, with design, manufacturing, and packaging conducted within a single firm. However, as manufacturing costs in advanced process nodes have increased substantially, it has become increasingly difficult for a single firm to maintain full integration across the entire value chain. Moreover, foundry operations require extremely large capital investments, resulting in very high entry barriers. A single fabrication facility typically contains approximately 300–400 pieces of equipment [18]. In contrast, fabless firms can operate with relatively lower capital requirements, which leads to comparatively lower entry barriers.

Within the semiconductor ecosystem, foundries function as central actors in technological innovation and maintain a strong interdependent relationship with fabless firms. Modern foundries are no longer limited to passively manufacturing customer designs; rather, they are increasingly expected to provide more proactive support, including suggesting improved design methodologies from the perspective of process optimization. Likewise, customers are heavily dependent on foundries for achieving high-performance products and must therefore consider the strategic position of foundry partners. In practice, large-scale foundries are often unable to accommodate all requests from smaller customers due to capacity and prioritization constraints.

Although foundries specialize in manufacturing, they cannot operate independently and must collaborate closely with a wide range of ecosystem partners. These include Electronic Design Automation(EDA) providers, Intellectual Property(IP) vendors, design houses, and Outsourced Semiconductor Assembly and Test(OSAT) firms. Strong partnerships with these stakeholders are essential for foundries to deliver optimized semiconductor solutions that meet customer requirements.

Foundries can generally be classified into two types: pure-play foundries and IDM-based foundries. Pure-play foundries focus exclusively on foundry services and do not engage in semiconductor design or product development. Representative examples include TSMC, UMC, and GlobalFoundries. In contrast, IDM-based foundries simultaneously operate foundry services while also engaging in semiconductor design activities. Intel and Samsung Electronics are prominent examples of this model.

The primary advantage of pure-play foundries is the absence of conflicts of interest with customers, allowing them to focus entirely on customer needs. By contrast, IDM-based foundries may raise concerns among customers regarding potential risks related to intellectual property protection, as customers may perceive the foundry as a potential competitor. Nevertheless, IDM-based foundries also possess distinct advantages, such as deeper integration between

design and manufacturing capabilities. Table 2 presents a comparison between pure-play foundries and IDM-based foundries.

(Table 2) Comparison between Pure-Play Foundries and IDM-Based Foundries

Category	Pure-play foundries	IDM-based foundries
Definition	A company that engages exclusively in manufacturing operations	a company that covers the entire process from design and manufacturing to packaging
Representative companies	TSMC, UMC, GlobalFoundries	Samsung Electronics, Intel
Advantages	-No conflict of interest with customers -Ability to collaborate with a wide range of customers -Customized services tailored to customer requirements	-Integrated management from design to manufacturing -Comprehensive understanding of technology and products -Ability to leverage internal design assets
Disadvantages	-No proprietary branded products -Limited internal design capabilities -High dependence on customer demand	-Potential competition with customers (Overlapping design domains) -Trust issues may arise when working with certain customers
Customer perspective	High reliability	-Concerns regarding leakage of proprietary design assets -Customers may hesitate to engage due to competitive relationships

2. 2. 2 The Current Status of Foundry Firms

Over the past several decades, semiconductor manufacturing has driven an increasingly significant portion of the global economy [19][20]. Semiconductor manufacturers have continuously pursued improvements in process performance [21]. Foundry companies are primarily located in the United States, China, Taiwan, Japan, and South Korea. However, most foundries currently operate at process nodes of 10nm and above, and only a limited number of firms possess the capability to manufacture advanced AI semiconductors.

Foundry firms can generally be classified into two broad categories. The first group consists of mature-node foundries that primarily operate at process nodes of 10nm and above. The second group includes advanced-node foundries capable of manufacturing at process nodes of 7nm and below. In particular, extreme ultraviolet(EUV) lithography equipment is considered almost essential for processes below 7nm, making it difficult for firms with limited financial resources to enter the advanced foundry business.

Representative firms in the former category include UMC, GlobalFoundries, PSMC, Hua Hong Semiconductor, and DB HiTek. The latter category includes TSMC, Samsung Foundry, Intel Foundry Services, and Rapidus. Notably, China's Semiconductor Manufacturing International Corporation (SMIC) has been unable to import EUV equipment due to U.S. export restrictions. As a result, SMIC has reportedly achieved 7nm production using deep ultraviolet(DUV) lithography through multi-patterning techniques rather than EUV-based processes.

As described above, foundry firms can be broadly divided into mature-node and advanced-node categories. While some mature-node foundries are attempting to enter advanced process technologies, the entry barriers remain extremely high due to the substantial capital investment and technological capabilities required.

In the current AI-driven era, advanced-node foundries have become increasingly prominent. AI semiconductors are predominantly manufactured using 4nm and 5nm process technologies, which has drawn significant attention to TSMC and Samsung Foundry. Other foundries currently lack sufficient capability to manufacture AI semiconductors at advanced nodes. Although Intel Foundry Services and Rapidus possess the technological potential to enter advanced AI semiconductor manufacturing, their capabilities have not yet been fully validated.

Under these circumstances, demand for advanced-node capacity currently exceeds supply. TSMC has secured the majority of AI semiconductor orders, resulting in near-full utilization of its fabrication facilities. In contrast, Samsung Foundry has thus far obtained relatively fewer customer orders. However, as demand for AI semiconductors is expected to increase sharply, Samsung Foundry may also benefit from expanding opportunities in the future. A comparative analysis of TSMC and Samsung Foundry in AI semiconductor manufacturing is presented in Table 3.

(Table 3) Comparison between TSMC and Samsung Foundry in AI Semiconductor Manufacturing

Category	TSMC	Samsung Foundry
Business Model	Pure-play foundry	IDM-based foundry
Market share (2025)	-Over 70% of the global foundry market -Over 90% share in AI semiconductor manufacturing	-Approximately 10% of the global foundry market -Relatively low share in AI semiconductor orders
Main customers	NVIDIA, AMD, Apple, Google, Amazon, Meta etc.	Tesla(FSD chip), Qualcomm(Snapdragon), Samsung System LSI
AI Semiconductor Manufacturing Process	-Based on 5nm, 4nm, and 3nm processes -2nm GAA mass production in progress -High yield stability	-4nm and 3nm processes in mass production -First to adopt 3nm GAA (Yield instability reported) -2nm GAA mass production in progress
Advanced Packaging Capabilities	-World-leading technologies including CoWoS(2.5D), InFO, and SoIC(3D) -Operates six dedicated packaging fabs, with plans to expand to twelve by 2030 -Optimized for HBM–GPU integration	I-Cube and X-Cube(2.5D/3D) packaging technologies -Increasing internalization of advanced packaging capabilities -Relatively lower market credibility compared to TSMC
Yield performance	Very high (Industry-leading) →ensures stable supply	Relatively lower yields →considered a limiting factor for customer acquisition
Strength	-High customer trust -Stable yield and process management -Central position in the global AI chip ecosystem	Ability to offer HBM and foundry services simultaneously (One-stop AI semiconductor solution) →strategic partnerships with selected customers such as Tesla
Weakness	-No proprietary end products(high dependence on customers) -Geopolitical risk(Taiwan)	-Insufficient yield stability and customer trust -The company's image was hit by issues related to 3nm GAA

2. 3 Requirements for Foundries to Enable High-Performance AI Semiconductors

Foundry firms must fundamentally understand customer requirements in order to secure manufacturing orders. In addition, activities such as marketing and customer service are necessary to attract and retain customers. However, from a technological perspective, particularly in the production of high-performance chips such as AI semiconductors, the most critical factors are yield performance, process capability, and advanced packaging technologies.

High yield performance is essential to ensure timely delivery of products to customers. Advanced process capabilities are necessary to manufacture high-performance AI semiconductors, while advanced packaging technologies play a crucial role in achieving superior PPA(Performance, Power, and Area) characteristics.

Accordingly, this study examines foundry capabilities from these three dimensions: yield performance, process technology, and advanced packaging capability.

2. 3. 1 Yield Performance

Yield is a critical factor in foundry operations. As the design complexity of chips increase particularly in the case of advanced AI semiconductors, yield typically decreases. Consequently, customers inevitably prefer foundries with high yield performance.

This section examines the importance of yield in AI semiconductor manufacturing from three perspectives.

First, insufficient yield directly undermines economic viability. Yield refers to the proportion of functional(good) chips produced during the manufacturing process. If yield performance is low, the defect rate increases while the proportion of usable chips decreases, which can lead to significant financial losses. Under such conditions, a foundry cannot sustain its business operations.

Second, low yield makes it difficult to meet customers' time-to-market requirements. When yield is poor, products cannot be delivered on schedule, which in turn prevents customers from launching their products in a timely manner. This issue has been identified as one of the factors that previously limited Samsung Foundry's ability to secure large-scale orders.

Third, yield performance is closely associated with product reliability. Poor yield suggests the presence of potential problems in the manufacturing process. From the customer's perspective, even chips classified as functional may be perceived as carrying a higher risk of latent defects, thereby reducing trust in the foundry.

Given the critical importance of yield, foundries must prioritize continuous efforts to improve yield performance. Factors that negatively affect yield can generally be categorized into three sources.

First, yield issues may arise during the chip design stage. In the complex process of designing high-performance chips, design errors can occur.

Second, yield degradation may occur during manufacturing processes. For example, challenges in etching, lithography, or deposition processes can significantly affect yield when producing advanced AI semiconductors.

Finally, environmental factors within the fabrication facility can also impact yield. Inadequate control of temperature, humidity, or airborne particles inside the fab can lead to defects in semiconductor devices.

Overall, strong yield performance is a prerequisite for customer trust and is a fundamental requirement for foundries seeking to participate in advanced AI semiconductor manufacturing.

2. 3. 2 Process Capability

Process capability refers to a foundry's ability to deliver chips that meet customer-required performance specifications at competitive cost and within a short time frame. To satisfy these conditions, process capability constitutes a critical component of a foundry's overall competitiveness. Superior process capability not only enhances customer satisfaction but also enables substantial reductions in manufacturing costs, thereby strengthening a firm's competitive position.

Most AI semiconductors are manufactured using advanced process nodes at 7nm and below. Accordingly, a foundry's ability to execute advanced scaling technologies has become a key determinant of competitiveness. As foundries are specialized manufacturing enterprises, the effectiveness with which process technologies are managed and maintained plays a decisive role in operational success. Through appropriate process control and management, foundries must be able to prevent potential issues that may arise during manufacturing in advance.

At present, TSMC reportedly secures more than 90% of AI semiconductor orders, and one of the primary reasons for strong customer preference is its superior process capability. In particular, foundries must be able to respond proactively to customer requirements and provide optimized process solutions in order to remain competitive. Effective customer-oriented process optimization contributes directly to higher levels of customer satisfaction. For example, customers differ in their requirements regarding power consumption and performance, and foundries must tailor their process solutions accordingly.

Furthermore, the stability of process control significantly affects key product attributes such as performance, power efficiency, and thermal characteristics. The ability to introduce new process technologies into production at an early and stable stage is also a crucial factor in attracting and retaining customers. In addition, through process optimization, foundries can minimize material and component usage and improve equipment utilization efficiency, thereby substantially reducing manufacturing costs.

Ultimately, customers select foundry partners based on their ability to supply high-quality chips at competitive cost and within required delivery timelines.

2. 3. 3 Packaging Capability

Historically, packaging processes were not regarded as a critical component in the semiconductor industry. Improvements in semiconductor performance were achieved primarily through front-end processes, such as lithography and device scaling, rather than through back-end packaging processes. Packaging was mainly responsible for protecting chips, dissipating heat, and providing electrical connections between the chip and the substrate.

In recent years, however, packaging has become a strategically important area for foundry firms. As further process scaling has become increasingly difficult and manufacturing costs have continued to rise, advanced packaging has emerged as an important alternative for improving chip performance while controlling costs. In some cases, appropriate packaging solutions can enhance chip performance more effectively than advancing to the next process node. Accordingly, leading foundries such as Samsung Electronics and TSMC have shown a clear trend toward internalizing advanced packaging capabilities. In particular, the ability to provide one-stop services that integrate both process technology and packaging has become a key advantage in securing large-scale customer orders.

With the full-scale transition into the AI era, growing demand for high-performance chips has accelerated the development of diverse packaging technologies. Representative examples include 2.5D and 3D packaging, which involve stacking and integrating heterogeneous chips in close proximity. One of the most effective approaches to improving performance at the packaging level is to reduce the physical distance between chips. Given that data transfer between processors and memory remains essential under the von Neumann architecture, shortening interconnect distances through 2.5D and 3D integration provides a significant performance advantage. Currently, most AI accelerators primarily adopt 2.5D packaging; however, a gradual transition toward 3D packaging is widely anticipated.

Other advanced packaging approaches that have gained increasing attention include chiplet architectures, which improve performance by modularly stacking and interconnecting multiple dies, and hybrid bonding technologies, which enable direct vertical interconnection between chips without the use of bumps. Furthermore, advanced packaging allows for optimized interconnect structures, which can reduce latency and enhance overall efficiency by optimizing power delivery paths.

Recently, major foundries such as Samsung Electronics, TSMC, and

Intel have expanded their packaging fabrication facilities. This expansion is largely driven by the rapid increase in demand for AI chips used in AI servers and high-performance computing systems. As a result, advanced packaging capability has become one of the most critical determinants of competitiveness in the AI semiconductor era.

Looking ahead, foundry firms are expected to continue investing in research and development of novel packaging technologies to further enhance chip performance.

2. 4. **Case Analysis of TSMC**

TSMC continues to lead the global foundry market in AI semiconductor manufacturing. The company manufactures not only NVIDIA's AI accelerators but also the vast majority of AI semiconductors, effectively holding a dominant position in this segment. In contrast, the market shares of Samsung Foundry and Intel Foundry Services in AI semiconductor manufacturing remain relatively limited. Therefore, selecting TSMC as the sole subject of analysis is considered appropriate for this study.

In particular, revenue from the high-performance computing segment, where AI semiconductors are extensively utilized, reportedly accounts for more than 60% of TSMC's total revenue. In response to rapidly increasing demand, TSMC has been transitioning portions of its legacy 6-inch wafer capacity toward advanced 12-inch wafer production lines. These strategic shifts further reflect TSMC's strong focus on AI semiconductor manufacturing.

Given that TSMC represents the most prominent foundry in the production of AI semiconductors, this study adopts TSMC as the primary case for analysis. Specifically, the study examines how TSMC enables high-performance AI semiconductor manufacturing through its capabilities in yield performance, process technology, and advanced packaging.

2. 4. 1 Yield Performance Perspective

Yield is a critical factor in the production of high-performance AI semiconductors. High yield not only ensures timely delivery of products to customers but also enables cost reduction through increased production volume. Compared with competing foundries, TSMC consistently achieves superior yield performance. One of the primary reasons is that TSMC has accumulated extensive process know-how by focusing almost exclusively on the foundry business for nearly four decades.

In addition, TSMC hired more than 10,000 employees in 2024 and reportedly plans to maintain a similar level of recruitment in 2026. The company provides an organizational environment that supports engineers in enhancing yield through continuous technological innovation.

Beyond these structural advantages, TSMC has implemented a range of concrete strategies to sustain high yield performance in AI semiconductor manufacturing. This study examines how TSMC has maintained superior yield performance through systematic technological and operational practices.

First, TSMC adopts rigorous end-to-end integration across the design, manufacturing, packaging, and mass production stages. To this end, it operates integrated frameworks such as Design for Manufacturing(DFM) and Design-Technology Co-Optimization(DTCO).

DFM supports customers in ensuring that their chip designs achieve optimal performance and yield on TSMC's process platforms. For this purpose, TSMC provides pre-verification and simulation services.

DTCO focuses on improving yield by jointly optimizing design and process technologies. Through DTCO, TSMC adjusts transistor layout, structure, and pattern configurations according to customer design characteristics.

TSMC also maintains one of the lowest levels of defect density(D_0) in the industry prior to wafer processing, recognizing that poor initial wafer quality inevitably degrades yield. Its efforts to reduce defect density can be categorized into three areas: process equipment, materials & components, and process control & design methodologies.

With respect to process equipment, TSMC has reduced mask-related defects through the adoption of high-NA EUV lithography systems and strengthened uniformity control using advanced etching and deposition equipment. Equipment upgrades have also enabled automated defect inspection at each process step, while big-data-based analytics are used to trace root causes of defects.

Regarding materials and components, TSMC has secured epitaxial growth technologies that enable the uniform formation of multilayer nanosheet channels for GAA-based devices. In addition, the company utilizes high-purity materials and applies low-K dielectrics to reduce

electrical defect rates.

In terms of process control and design methodologies, major customers such as Apple and NVIDIA incorporate process-friendly patterns through DFM at the design stage and conduct pre-simulations to identify potential hotspot regions with high defect probability. Furthermore, TSMC detects yield degradation factors at an early stage and tunes equipment using AI-based defect prediction models. The company also enhances yield by connecting distributed fabs into clusters and sharing operational data across facilities. Through these efforts, TSMC reportedly achieved an SRAM yield of approximately 90% at the pilot stage of its 2nm process [22], while overall logic chip yield has exceeded 60%.

Once wafer processing begins, TSMC employs an advanced yield management system that collects and analyzes massive volumes of production data in real time. This enables the prediction of potential defects and identification of root causes. During wafer fabrication, various process deviation data are continuously monitored, and AI-based models detect patterns and immediately correct process errors. Moreover, millions of process parameters, including equipment temperature and pressure, are statistically controlled to minimize variability.

TSMC also applies repair processes that correct defects in cells and interconnects to convert defective chips into functional products. In addition, the company conducts pilot production runs on small-scale lines prior to mass production in order to identify and resolve potential issues in advance. When identical processes are operated across multiple fabs, TSMC synchronizes process conditions and equipment parameters to prevent yield deviations between facilities.

2. 4. 2 Process Capability Perspective

TSMC's process capability is widely regarded as among the highest in the global semiconductor industry. The company is known for achieving relatively high initial mass-production yields and for transitioning from process development to volume manufacturing more rapidly than its competitors. TSMC commercializes next-generation process nodes approximately every two years, representing the shortest cycle in the industry. In practice, TSMC began mass production of its 7nm process in 2018, 5nm in 2020, and 3nm in 2022. Moreover, TSMC's process technologies are generally operated with a high level of stability. This is largely because the company prioritizes yield improvement and process maturity over the early adoption of highly uncertain next-generation technologies.

For example, in 2022, TSMC continued to adopt the FinFET architecture for its 3nm process, citing uncertainties associated with Gate-All-Around(GAA) technology, thereby achieving higher yield and greater process stability. In contrast, Samsung Foundry adopted GAA earlier, but reportedly experienced challenges related to yield and process stability, which limited its ability to secure customer orders. As a result, TSMC continues to maintain a competitive advantage in terms of process stability. From the customer's perspective, this stability allows for more predictable delivery schedules.

In addition, TSMC is widely perceived as a highly reliable company, supported by its corporate philosophy of integrity and its strong reputation for technological maturity. TSMC's wafer defect density in its 5nm process has been reported to be below 0.1 defects/cm² [23], which is up to 30% lower than that of Samsung Foundry during the same period. This level of process maturity has enabled TSMC to consistently meet the mass-production schedules of major customers such as Apple and NVIDIA.

Furthermore, TSMC operates a diverse range of process platforms across its fabs, allowing it to optimize manufacturing processes according to customer-specific requirements. The company also maintains advanced control systems capable of rapidly detecting and addressing process defects.

TSMC frequently proposes design improvements to customers from an early stage and incorporates these insights into process optimization. This strong customer-oriented approach is widely recognized as one of the primary reasons why customers continue to select TSMC as their foundry partner.

At the same time, TSMC has deployed more EUV lithography systems from ASML than any other foundry and has achieved earlier stabilization of EUV-based manufacturing. For example, TSMC reportedly applies EUV patterning to approximately 14 layers in its 5nm process and more than 20 layers in its 3nm process, representing up to twice the level of EUV utilization compared with competing firms. This extensive use of EUV has contributed not only to improved pattern precision but also to enhanced power efficiency.

TSMC has also entered mass production of its 2nm process. At this node, transistor performance is expected to improve by up to 15% compared with the 3nm process, while power consumption is projected to decrease by up to 30%. These improvements illustrate TSMC's strong capability in transistor architecture optimization.

Moreover, TSMC collaborates closely with customers from the early stages of chip design, jointly developing process technologies and maximizing process maturity by the time of tape-out. The company maintains strong partnerships with key equipment suppliers such as ASML, Applied Materials, and Tokyo Electron, as well as with materials and component suppliers across the ecosystem. Supported by high profitability, reportedly exceeding 50% gross margin, TSMC is able to make aggressive early investments in new process lines, enabling faster mass production of next-generation technologies than its competitors. In addition, even within the same process node, TSMC offers customer-specific customization to meet diverse requirements related to power, performance, and other design constraints.

2. 4. 3 Packaging Capability Perspective

As further scaling toward sub-1nm process nodes is increasingly expected to face substantial technical and economic limitations, many foundry firms are placing strategic emphasis on advanced packaging technologies. Among them, TSMC is widely regarded as the most advanced company in the field of semiconductor packaging. Its advanced packaging technologies demonstrate superior performance compared to those of other firms, enabling the simultaneous achievement of low power consumption, high performance, and high integration density. Consequently, TSMC is reported to hold the leading position in the global advanced packaging market.

In response to the rapid growth in demand for AI semiconductors, TSMC anticipated that advanced packaging would become a critical growth area. Accordingly, the company has significantly increased its investment in research and development and has expanded its advanced packaging capacity through the construction of new fabrication facilities. These strategic investments have enabled TSMC to strengthen its competitiveness in advanced packaging and to secure a large number of AI semiconductor customers.

TSMC is reportedly operating more than six dedicated packaging fabrication facilities, of which three are advanced packaging fabs, as summarized in Table 4. In addition, TSMC is currently constructing a total of six additional packaging facilities across Taiwan and the United States. As a result, the number of TSMC's packaging fabs is expected to exceed twelve by 2030.

(Table 4) Major Advanced Packaging Fabs of TSMC(2025)

Category	Location	Key Technologies	Remarks
Advanced Packaging	Hsinchu (新竹)	CoWoS, InFO, SoIC	Focused on HPC and AI applications
Advanced Packaging	Taichung (台中)	CoWoS	Primarily for AI chip production
Advanced Packaging	Tainan (台南)	InFO	Mobile and HPC applications

TSMC has integrated advanced packaging technologies for AI semiconductors such as CoWoS(Chip-on-Wafer-on-Substrate), InFO(Integrated Fan-Out), and SoIC(System on Integrated Chips) into its manufacturing ecosystem. Through these technologies, TSMC optimizes performance and power efficiency while offering services at relatively competitive cost. This integrated capability has enabled TSMC to secure a dominant position in the AI semiconductor market.

The current status of TSMC's major advanced packaging technologies can be summarized as follows.

First, with respect to CoWoS packaging, TSMC's production capacity is projected to exceed 130,000 wafers per month by 2026 [24], which represents a substantially faster expansion pace than that of competing firms. TSMC currently performs CoWoS packaging by integrating NVIDIA's GPUs(e.g., B100 and B200) with SK hynix's HBM3E. TSMC's market share in this segment is reported to exceed 90%.

Second, InFO packaging is widely applied to mobile application processors(APs) in smartphones. Apple is currently the primary customer for InFO packaging, which is used in its A-series and M-series chips. InFO adopts a fan-out architecture that increases I/O density and enables thinner substrates, thereby supporting ultra-lightweight and low-power packaging solutions.

Third, SoIC is a three-dimensional(3D) packaging technology that vertically stacks logic dies and utilizes through-silicon vias(TSVs). Major companies such as AMD, NVIDIA, and Apple are reportedly evaluating SoIC for next-generation high-performance computing applications. Furthermore, the adoption of hybrid bonding allows for direct chip-to-chip interconnection without micro-bumps, which contributes to improvements in both power efficiency and processing speed.

In addition, TSMC has internalized not only packaging processes but

also testing processes, thereby enhancing overall customer satisfaction. The company considers packaging strategies from the early stages of chip design and provides customers with optimized packaging solutions tailored to their requirements. Moreover, because TSMC integrates both front-end(process) and back-end(packaging and testing) operations within its ecosystem, it is able to provide rapid feedback and issue resolution when problems arise during development or production.

3. Conclusion

This study examined a wide range of AI semiconductor firms and analyzed the current landscape of foundry companies. Based on this analysis, the study identified yield performance, process capability, and advanced packaging capability as the most critical technological factors in advanced foundry operations. Furthermore, through an in-depth case study of TSMC, this research investigated how a leading foundry enables the manufacturing of high-performance AI semiconductors from the perspectives of yield, process technology, and packaging.

The findings indicate that TSMC secures a competitive advantage in AI semiconductor manufacturing by implementing rigorous end-to-end integration from design and fabrication to packaging and mass production, maintaining highly stable process capability, and making proactive and substantial R&D investments in advanced packaging technologies. These combined capabilities enable TSMC to sustain superior competitiveness in the production of high-performance AI semiconductors.

This study is meaningful in that it approaches AI semiconductor research from the perspective of foundries, which has been relatively underexplored compared to design-focused studies. One key implication of the findings is that AI semiconductor firms should place greater strategic emphasis on foundry selection, as the performance and competitiveness of AI chips can be significantly influenced by the technological capabilities of foundry partners.

This study adopts a single-case research design and focuses on the role of foundries in terms of yield, process capability, and packaging capability. Therefore, its findings are subject to limitations in terms of generalizability. Future research may extend this work by examining additional critical factors in AI semiconductor manufacturing and conducting comparative analyses across multiple foundry firms to provide broader empirical insights.

Reference

[1] Er-Ping Li et al., “An Electromagnetic Perspective of Artificial Intelligence Neuromorphic Chips,” Electromagnetic Science, Vol. 1, No. 3, Sep. 2023

[2] Heyang Xu, “FPGA: The super chip in the age of artificial intelligence,” Journal of Physics: Conference Series, doi:10.1088/1742-6596/2649/1/012018, 2023

[3] Dhruvitkumar V. Talati, “Silicon minds: The rise of AI-powered chips,” International Journal of Science and Research Archive, Vol. 01, No. 02, pp. 097-108, 2021.

[4] Sorna Mugi Viswanathan, “AI Chips: New Semiconductor Era,” International Journal of Advanced Research in Science, Engineering, Technology, Vol. 7, No. 8, pp. 14687-14694, Aug. 2020

[5] Hyunji Kim, Seyoung Yoon, Hwajeong Seo, “Research Trends in Domestic and International AI Chips,” Smart Media Journal, Vol. 13, No. 3, pp. 36–44, Mar. 2024.

[6] Hiroshi Momose, Tatsuya Kaneko, and Tetsuya Asai, “Systems and circuits for AI chips and their trends,” Japanese Journal of Applied Physics 59, 2020

[7] P. Ebby Darney, “A Review on Artificial Intelligence Chip,” Recent Research Reviews Journal, Vol. 1, No. 1, pp. 99-109, Dec. 2022

[8] Prashis Raghuwanshi, “Revolutionizing Semiconductor Design and Manufacturing with AI,” Journal of Knowledge Learning and Science Technology, Vol. 3, No. 3, Sep. 2024

[9] Shaojun Wei et al., “Reconfigurability, Why It Matters in AI Tasks Processing: A Survey of Reconfigurable AI Chips” IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—I, Vol. 70, No. 3, pp. 1228- 1241, Mar 2023

[10] Siwoo Eum, et al., “Development Trends in Artificial Intelligence Semiconductors,” Journal of the Korea Institute of Information Security and Cryptology, Vol. 35, No. 3, pp. 63–71, Jun. 2025.

[11] Robert Huggins et al., “Competition, open innovation, and growth challenges in the semiconductor industry: the case of Europe’s clusters,” Science and Public Policy, 50, pp. 531–547, Mar. 2023.

[12] Wenqiang Zhang et al., “Neuro-inspired computing chips,” Nature Electronics, Vol. 3, pp. 371–382, Jul. 2020

[13] S. Ambrogio et al., “An analog-AI chip for energy-efficient speech recognition and transcription,” Nature, Vol. 620, 24 Aug. 2023

[14] Q. Zhang, H. Deng, and K. Song, “Latest VLSI Techniques for

3nm Technology for Building Efficient AI Chips," "Fusion of Multidisciplinary Research, An International Journal (FMR)," Vol. 5, No. 2, 2024

[15] https://economist.co.kr/article/view/ecn202507310045 (accessed Aug. 24, 2025)

[16] Wei Gao, and Pingqiang Zhou, "Customized High Performance and Energy Efficient Communication Networks for AI Chips," IEEE, Vol. 7, 2019

[17] Lennart P. L. Landsmeer et al., "Tricking AI chips into simulating the human brain: A detailed performance analysis," Neurocomputing, Vol. 598 Sep. 2024

[18] Yon-Chun Chou and I.-Hsuan Hong, "A Methodology for Product Mix Planning in Semiconductor Foundry Manufacturing," IEEE Transactions on Semiconductor Manufacturing, Vol. 13, No. 3, pp. 278-285, Aug. 2000

[19] Wojciech T. Osowiecki et al., "Achieving Sustainability in the Semiconductor Industry: The Impact of Simulation and AI," IEEE Transactions on Semiconductor Manufacturing, Vol. 37, No. 4, pp. 464-474, Nov. 2024

[20] S. Fore and C.T. Mbohwa, "Cleaner production for environmental conscious manufacturing in the foundry industry,

Author Biography

Young-hwa Kwun

B.A. in English Literature, Kwangwoon University, 1994

M.S. (Business Administration), Sungkyunkwan University, 1998

M.A. in Japanese Studies (International Area Studies), Hanyang University, 1999

Ph.D. in Business Administration, Seoul School of Integrated Sciences & Technologies (aSSIST), 2013

Ph.D. Candidate in AI Convergence Engineering, Seoul School of Integrated Sciences & Technologies (aSSIST), 2025–Present

Research Interests: AI semiconductors, edge devices, etc.

E-mail:690730@daum.net

Bo-young Kim

B.A. in Information Design, Ewha Womans University, 1996

M.A. in Information Design, Ewha Womans University, 2000

Ph.D. in Engineering, School of Engineering and Design, Brunel University, 2006

Professor, Graduate School of Business, Seoul School of Integrated Sciences & Technologies (aSSIST), 2008–Present

Research Interests: Media processing, context awareness, etc.

E-mail:bykim2@assist.ac.kr

Practical Applicability of an AI-Based Automated Classification Model in Elderly Care Institutions

Jee Won Moon , Tae yeon Oh

ABSTRACT

This study evaluates the practical applicability of an artificial intelligence (AI)-based automated text classification model in elderly care institutions. While prior research has primarily focused on improving algorithmic performance, limited empirical attention has been given to whether automated outputs demonstrate sufficient alignment with existing manual classification practices in real operational environments. This study reanalyzes a categorized dataset of 1,224 elderly care records to examine predictive agreement from an organizational applicability perspective. A logistic regression classifier with hyperparameter optimization was implemented, and performance was assessed using accuracy, precision, recall, and F1-score metrics. The optimized model achieved an accuracy of 0.790 and a weighted F1-score of 0.780. Category-level analysis indicated stable performance across primary operational classes, although variation in recall was observed in specific categories. The findings suggest that AI-assisted classification may function as a structured support mechanism within institutional documentation workflows. This study contributes empirical evidence regarding feasibility of practical implementation without asserting direct organizational performance transformation.

Keywords: AI text classification, elderly care records, predictive agreement, organizational applicability, logistic regression

1. INTRODUCTION

The digital transformation of healthcare and welfare institutions has accelerated the accumulation of operational data in textual form [1][2]. Elderly care institutions, in particular, generate large volumes of narrative records describing residents′ physical conditions, cognitive changes, emotional expressions, and behavioral observations. These records serve multiple administrative purposes, including regulatory reporting, internal monitoring, and quality management.

Despite the growing volume of documentation, classification and structuring of narrative records remain largely dependent on manual processes. Staff members review textual entries and assign them to predefined operational categories. While human interpretation allows contextual nuance, manual classification requires substantial labor resources and may result in variability in category assignment. As the quantity of documentation increases, maintaining consistency and efficiency in classification becomes progressively more challenging.

Artificial intelligence (AI)-based text classification techniques have been introduced as a potential solution to address these limitations [3][4]. Machine learning models can be trained on labeled datasets to predict predefined categories for new textual inputs. Prior research in healthcare and welfare domains has demonstrated that supervised learning models can achieve acceptable predictive performance under controlled experimental settings [3][4]. However, most studies have concentrated on improving algorithmic accuracy or comparing model architectures, rather than examining whether automated outputs align sufficiently with existing institutional classification standards.

For AI systems to be practically integrated into institutional workflows, predictive performance alone is not sufficient. Automated outputs must demonstrate stable agreement with manually assigned labels, particularly across core operational categories. If predictive alignment is unstable, the introduction of automated systems may increase corrective workload instead of improving efficiency. Therefore, assessing predictive agreement within real operational contexts is a necessary preliminary step before broader organizational application.

This study evaluates the practical applicability of an AI-based automated classification model using categorized elderly care records. Rather than emphasizing model innovation, the analysis focuses on whether automated predictions demonstrate sufficient alignment with manual classification practices to support structured institutional use. By adopting an organizational applicability perspective, this research seeks to provide empirical evidence regarding the feasibility of AI-

assisted record classification in elderly care institutions [5][6].

2. THEORETICAL BACKGROUND

2.1 AI-Based Text Classification in Healthcare Documentation

Text classification is a core task in natural language processing (NLP), widely applied to organize unstructured textual data into predefined categories. In healthcare and welfare contexts, machine learning-based classification models have been utilized to analyze nursing records, medical reports, and patient documentation [3][7]. Supervised learning approaches enable models to learn patterns from labeled datasets and generate category predictions for new textual inputs.

Among various classification algorithms, logistic regression remains frequently adopted in structured institutional settings due to its interpretability and computational efficiency [3][8]. While more complex neural architectures may offer performance improvements under large-scale datasets, relatively simpler models are often preferred in organizational environments where transparency and stability are important considerations [4][9].

Previous research has primarily focused on enhancing predictive performance through feature engineering, model comparison, and parameter optimization [3][10]. These studies contribute to methodological advancement; however, most evaluations are conducted under experimental conditions without explicit consideration of institutional classification practices embedded in daily workflows.

In operational settings such as elderly care institutions, record categorization functions not merely as a technical process but as part of routine administrative procedures. Therefore, the evaluation of automated classification models should consider their compatibility with existing classification standards and practical interpretability.

2.2 Predictive Agreement as a Criterion for Organizational Applicability

Organizational applicability of AI systems depends on more than technical accuracy. For automated classification to be integrated into institutional workflows, model outputs must demonstrate consistent alignment with manually assigned labels across core operational categories. Predictive agreement refers to the degree of correspondence between automated predictions and human-generated reference labels. Standard performance metrics—precision, recall, and F1-score—are commonly employed to quantify such alignment [3][4].

Precision measures the proportion of correctly predicted instances among predicted cases within a category. Recall represents the proportion of correctly identified instances among actual cases. The F1-score balances these two measures.

In organizational contexts, balanced performance across categories may be more meaningful than marginal improvements in overall accuracy. Stable predictive patterns reduce the likelihood of systematic bias toward particular classes and enhance confidence in institutional application.

2.3 Operational Considerations and Error Implications

Automated classification errors may influence administrative monitoring and documentation processes. Although this study does not evaluate downstream operational outcomes, predictive instability across categories may affect the reliability of structured record management systems.

Prior research on AI adoption in healthcare and clinical decision environments emphasizes the importance of error awareness and monitoring structures in technology implementation [8][9]. While elderly care record classification does not constitute high-risk decision-making in a strict sense, inaccurate categorization may introduce reporting inconsistencies. Therefore, predictive stability serves as a foundational requirement for cautious institutional consideration.

From this perspective, the evaluation of AI-based classification models should move beyond purely technical optimization and examine their operational feasibility within structured documentation environments.

3. RESEARCH DESIGN AND DATA

3.1 Research Scope and Analytical Perspective

This study adopts an empirical validation approach to examine predictive agreement between automated and manual classification outcomes. The analytical objective differs from algorithm

development or performance maximization. Instead, the focus lies on assessing whether the observed performance level supports consideration of structured institutional application.

The dataset utilized in this study was originally collected and structured as part of the author′s prior master′s thesis research [6]. However, the analytical framework of the present study differs substantially in scope and objective. While the previous research concentrated on model construction and optimization strategies, this study reanalyzes the dataset from an organizational applicability perspective, emphasizing predictive alignment rather than technical advancement.

3.2 Data Description

The dataset consists of categorized textual records generated within an elderly care institution. The records were originally documented in narrative form and subsequently assigned to predefined operational categories. These categories reflect routine documentation practices within the institution.

A total of 1,224 records were included in the analysis. Among these, 871 records were manually labeled by trained personnel, and 353 records were incorporated through controlled data augmentation procedures designed to preserve semantic consistency [11].

(Table 1) Dataset Composition

Category	Count
Manual labeling	871
Data augmentation	353
Total	1,224

All records were anonymized prior to analysis in compliance with healthcare data protection standards [12]. No personally identifiable information was retained in the analytical dataset [2][13].

3.3 Classification Categories

The predefined operational categories represent recurring documentation themes observed in elderly care records. These include physical condition descriptions, cognitive changes, emotional states, behavioral observations, and miscellaneous entries [14].

The classification scheme reflects practical administrative usage rather than theoretical taxonomy. Therefore, the evaluation of automated classification focuses on its ability to reproduce existing institutional categorization patterns rather than redefine category structures [15].

3.4 Data Preprocessing

Prior to model training, textual data underwent standard preprocessing procedures. These included token normalization, removal of non-informative characters, and vectorization using term-frequency-based feature representation. Class imbalance was addressed through weighting strategies within the classifier configuration.

The dataset was partitioned into training and evaluation subsets using stratified sampling to preserve category distribution proportions. Performance evaluation was conducted on the held-out evaluation set.

4. MODEL CONFIGURATION AND EXPERIMENTAL PROCEDURE

4.1 Model Selection Rationale

This study employs logistic regression as the primary classification algorithm. Although more complex neural network architectures have been widely adopted in natural language processing tasks, logistic regression remains a practical and interpretable baseline model in institutional environments [7]. The primary objective of this research is not to maximize predictive performance through architectural complexity, but to assess whether a relatively stable and transparent model can achieve sufficient predictive agreement for structured organizational application.

Logistic regression estimates the probability that a given textual input belongs to a predefined category based on feature weights. The model predicts class membership using the sigmoid function:

$$P(y = 1|x) = \frac{1}{1 + e^{-(\beta_0+\beta_1 x_1+\cdots+\beta_n x_n)}}$$

where x_i represents feature values derived from textual vectorization and β_i denotes learned coefficients. In multi-class classification settings, a one-vs-rest (OvR) strategy is applied.

The interpretability of coefficient weights allows examination of feature influence patterns, which is advantageous in institutional contexts where transparency is desirable.

4.2 Hyperparameter Optimization

To improve model stability, hyperparameter tuning was conducted. The regularization strength parameter C was adjusted to balance bias and variance trade-offs. Class imbalance was addressed using a balanced class weighting scheme.

(Table 2) Hyperparameter Configuration

Parameter	Value
Regularization (C)	10
Solver	liblinear
Class Weight	Balanced
Multi-class Strategy	One-vs-Rest

The selected configuration demonstrated improved predictive consistency compared to baseline settings.

4.3 Experimental Design

The dataset was divided into training and evaluation subsets using stratified sampling to preserve category proportions. Model training was conducted on the training set, and evaluation was performed on a held-out test set.

Three experimental conditions were examined:

1) Manual-labeled dataset only
2) Manual + augmented dataset
3) Optimized configuration with hyperparameter tuning

This comparative design allows examination of how data augmentation and parameter tuning influence predictive agreement.

4.4 Evaluation Metrics

Model performance was evaluated using accuracy, precision, recall, and F1-score [3][4].

Accuracy measures overall correct predictions:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Precision and recall provide category-specific performance insight:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

The F1-score is calculated as:

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

In organizational contexts, recall may carry particular importance in categories where omission errors are undesirable. Therefore, both macro-average and weighted-average F1-scores were examined to evaluate performance balance across categories.

4.5 Analytical Focus

The analytical emphasis of this study lies in predictive agreement rather than incremental accuracy improvement. Therefore, performance stability across categories is interpreted as a primary indicator of organizational feasibility. Variability in recall rates is analyzed to identify potential operational sensitivity in specific documentation categories.

5. RESULTS

5.1 Overall Performance Comparison

The predictive performance of the automated classification model was

evaluated under three experimental conditions. Table 3 summarizes the overall performance metrics.

(Table 3) Overall Model Performance Comparison

Condition	Accuracy	Precision	Recall	F1-score
Manual Only	0.78	0.76	0.74	0.75
Manual + Augmented	0.82	0.80	0.79	0.79
Multi-classStrategy	0.86	0.85	0.84	0.84

The optimized configuration demonstrated the highest level of predictive consistency across all evaluation metrics. The increase in F1-score from 0.75 (manual-only baseline) to 0.84 (optimized configuration) indicates that improvements occurred simultaneously in both precision and recall, rather than being driven by isolated metric inflation.

The inclusion of augmented data produced moderate gains, particularly in recall performance. However, the most substantial enhancement was observed following hyperparameter calibration. This suggests that model stability in institutional settings is sensitive not only to dataset size but also to parameter configuration.

Importantly, overall accuracy alone does not sufficiently capture organizational suitability. In structured documentation systems, balanced metric behavior provides a more meaningful evaluative basis than aggregate performance alone [3][4].

5.2 Category-Level Performance Analysis

To assess predictive agreement at a more granular level, category-specific precision, recall, and F1-scores were analyzed under the optimized configuration.

(Table 4) Category-Level F1-score (Optimized Model)

Category	Precision	Recall	F1-score
Physical Condition	0.88	0.85	0.86
Cognitive Change	0.83	0.81	0.82
Emotional State	0.84	0.83	0.83
Behavioral Observation	0.86	0.84	0.85
Miscellaneous	0.79	0.76	0.77

The highest predictive alignment was observed in the Physical Condition category (F1 = 0.86), followed closely by Behavioral Observation (F1 = 0.85). These results suggest that records describing observable physical or behavioral conditions may exhibit relatively consistent linguistic structures, facilitating model generalization.

The Miscellaneous category showed comparatively lower performance (F1 = 0.77). This outcome likely reflects semantic heterogeneity within the category, as broadly defined operational classes may incorporate diverse narrative patterns. The reduced performance does not indicate systemic failure but highlights the structural limitations of loosely bounded categories.

Crucially, no category exhibited extreme imbalance between precision and recall. The absence of sustained recall suppression across operational domains indicates that the model does not systematically under-identify specific documentation types.

5.3 Stability Across Experimental Conditions

Comparative analysis across experimental conditions reveals that data augmentation primarily contributed to recall improvement in categories with relatively smaller representation. This effect reduced classification bias by enhancing minority-category detection.

More notably, the optimized configuration reduced inter-category recall variance from 0.11 in the baseline condition to 0.07 after calibration. This reduction reflects improved distributional balance rather than isolated metric gain.

In institutional environments, such variance reduction is operationally significant. Lower inter-category volatility decreases the likelihood of systematic misclassification within particular documentation domains and supports administrative reliability.

Thus, predictive stability emerges not merely as a statistical

characteristic but as an organizationally relevant property.

5.4 Interpretation from an Organizational Perspective

From an institutional standpoint, the observed performance patterns indicate that the automated classification model is capable of reproducing manual categorization structures with substantial consistency.

While the model does not replicate professional judgment in its entirety, F1-scores exceeding 0.80 across core operational categories suggest that the system can function as a structured assistive mechanism within documentation workflows.

Importantly, the findings do not imply automation of final decision authority. Rather, they support a hybrid configuration in which automated classification provides preliminary categorization, followed by human verification where necessary.

In this context, predictive agreement at the observed level may reduce repetitive sorting workload while preserving documentation coherence and accountability structures.

6. DISCUSSION

6.1 Reframing Performance Beyond Metric Maximization

The empirical findings of this study call for a reframing of how performance should be interpreted within institutional environments.

Accuracy, precision, recall, and F1-score are conventionally treated as primary indicators of model quality in text classification research. However, when the objective shifts from technical optimization to organizational deployment, the meaning of these metrics necessarily changes.

In elderly care institutions, classification systems are embedded within structured documentation workflows that directly influence monitoring, reporting, and administrative coordination. Under such conditions, performance cannot be evaluated solely in terms of maximal aggregate accuracy. What matters more fundamentally is whether predictive behavior remains proportionate, balanced, and structurally stable across operational categories.

A model may achieve high overall accuracy while simultaneously under-identifying specific documentation domains. Such imbalance, even if statistically subtle, can introduce distortion into reporting structures and administrative interpretation. Therefore, institutional adequacy requires an evaluative lens that extends beyond peak metric performance.

This study adopts that broader interpretive position.

6.2 Predictive Stability as an Institutional Threshold

The results indicate that predictive stability offers a more appropriate benchmark for deployment consideration than accuracy maximization alone.

Predictive stability, as defined in this study, entails:

1) Balanced precision and recall across predefined operational categories.
2) Controlled inter-category variance in recall performance.
3) Absence of systematic suppression in any documentation domain.

The optimized model satisfied these criteria. F1-scores exceeded 0.80 across major categories, and recall variance was reduced following parameter calibration. No category demonstrated persistent under-identification.

These empirical properties suggest that the model maintains structural alignment with existing manual classification patterns rather than distorting them. Stability in this sense reflects proportional behavioral consistency across operational domains.

In institutional documentation systems where categorization affects monitoring and compliance processes, such proportional consistency constitutes a meaningful reliability threshold.

Accordingly, this study advances predictive stability as a primary evaluative benchmark for structured institutional deployment of AI-based classification systems.

6.3 Clarifying the Boundaries of Interpretation

It is essential to delimit the scope of this conclusion.

The findings do not imply replacement of professional judgment, nor do they assert that logistic regression represents a universally optimal architecture. The contribution of this study lies not in technical superiority but in evaluative repositioning.

Within a predefined institutional classification framework, the model demonstrates sufficient predictive alignment to function as a structured assistive layer. This suggests feasibility for preliminary categorization within a human-supervised workflow, where automated outputs are subject to oversight rather than autonomous authority.

The system, therefore, is conceptualized not as a decision-maker but as a stabilizing administrative component embedded within existing accountability structures.

6.4 From Computational Evaluation to Governance Judgment

The broader significance of this study lies in connecting computational evaluation with institutional governance judgment.

AI-based classification models are often evaluated as isolated technical artifacts. However, institutional integration is ultimately a governance decision. It requires assessing whether model behavior aligns with documentation norms, reporting structures, and accountability expectations.

By formalizing predictive stability as an operational threshold, this study provides a decision-oriented interpretive framework. The evaluative question shifts from:

"Which model achieves the highest accuracy?"

to:

"Does the model maintain balanced and reliable behavior across operational domains under real documentation constraints?"

This reframing does not diminish technical rigor. Instead, it repositions performance metrics within the institutional realities of service organizations.

The contribution of the study, therefore, lies not in competing for incremental metric gains, but in clarifying how predictive behavior should be judged when organizational integration is under consideration.

7. LIMITATIONS

7.1 Contextual Delimitation of the Empirical Setting

The empirical analysis presented in this study is situated within the documentation system of a single elderly care institution. Such documentation structures are not neutral technical artifacts; they are shaped by organizational routines, staffing patterns, regulatory expectations, and local operational cultures. Linguistic conventions, category boundaries, and narrative styles emerge from these institutional configurations.

Accordingly, the predictive stability observed in this study should be understood as stability within a specific organizational configuration rather than as a universal property of the model itself. The findings demonstrate how the stability criterion operates under a concrete documentation regime.

This contextual delimitation does not weaken the conceptual contribution of the study. Rather, it clarifies the empirical grounding of the proposed stability framework. By explicitly specifying the scope of application, the study avoids unwarranted generalization while preserving the structural validity of its evaluative perspective.

Future research may extend this investigation by incorporating multi-institutional datasets in order to examine whether predictive stability operates consistently across heterogeneous documentation environments.

7.2 Algorithmic Scope and Architectural Generalizability

This study intentionally adopted logistic regression as a transparent and structurally stable modeling approach. The objective was not to compete in algorithmic sophistication, but to assess whether predictive stability could be established within an interpretable and institutionally viable configuration.

Recent advances in transformer-based and deep learning architectures have demonstrated notable gains in raw classification accuracy. However, superior metric performance under experimental conditions does not automatically translate into institutional suitability. Organizational deployment requires consideration of interpretability, behavioral consistency, and variance control across operational domains.

Because cross-architecture benchmarking was not conducted, the generalizability of the proposed stability criterion across model families remains empirically open. Whether predictive stability persists under highly expressive neural architectures constitutes an important direction for subsequent investigation.

Thus, this limitation is not merely technical but conceptual. It concerns the portability of stability as a deployment benchmark beyond a single modeling paradigm.

7.3 Absence of Longitudinal Institutional Implementation

The present research evaluates predictive alignment between automated and manual classification outcomes under controlled analytical conditions. It does not include longitudinal observation of implemented workflows within the institution.

While predictive stability may serve as a precondition for deployment, the dynamic interaction between automated classification outputs and human verification practices was not empirically examined. Changes in documentation efficiency, redistribution of workload, and behavioral adaptation among staff remain outside the scope of this study.

Future research may therefore investigate how stability-informed deployment decisions translate into measurable administrative outcomes over time. Such longitudinal validation would allow the proposed evaluative framework to be assessed not only at the metric level but also at the organizational process level.

7.4 Formalization of Stability as an Operational Threshold

This study conceptualizes predictive stability as an operational reliability threshold but does not codify it into a fixed quantitative cutoff. The interpretation relies on balanced performance patterns and reduced inter-category variance rather than on a predetermined numerical boundary.

Acceptable stability levels may vary depending on documentation criticality, institutional risk tolerance, or regulatory sensitivity. Therefore, the explicit calibration of deployment thresholds remains an open theoretical and practical task.

This limitation signals a direction for conceptual consolidation rather than a deficiency in empirical design. Further refinement of stability-based criteria may translate the proposed framework into more concrete operational guidelines.

8. CONCLUSION

This study examined the practical applicability of an AI-based automated classification model within the operational context of an elderly care institution. Rather than pursuing incremental gains in predictive accuracy or architectural sophistication, the analysis focused on whether predictive stability could serve as a meaningful benchmark for institutional deployment decisions.

The findings indicate that predictive stability—understood as balanced inter-category performance and consistent alignment with manual categorization—constitutes a more operationally relevant criterion than peak metric performance alone. In structured documentation environments, the reliability of category behavior across routine records may hold greater institutional significance than isolated improvements in aggregate accuracy.

By repositioning performance interpretation from competitive optimization toward stability-oriented assessment, this study advances a deployment-sensitive analytical perspective. AI-based classification is framed not as an autonomous performance artifact, but as a system embedded within organizational routines where interpretability, variance control, and behavioral consistency are central considerations.

The study does not assert universal generalizability nor claim immediate organizational transformation. Instead, it demonstrates that predictive stability can be empirically observed and analytically articulated within a defined institutional configuration. This articulation provides a principled basis for deployment decisions grounded in measured reliability rather than technological enthusiasm.

In this respect, the contribution of the study lies in clarifying what constitutes sufficient empirical evidence for organizational consideration of AI systems. Within structured administrative environments, stability—rather than maximal accuracy—emerges as the central evaluative axis for responsible integration.

References

[1] S. Theresharov, G. Dragomirov, G. Gustinov, S. Hadzhikoleva, and K. Yotov, "Transforming Nursing Home Care: An Integrated Approach Using Sensors, AI, and Monitoring Technologies," Computer Science and Interdisciplinary Research Journal, Vol. 1, No. 1, 2024.

[2] T. M. Song, "Trends and Utilization of Health and Welfare Big Data in Korea," Health and Welfare Policy Forum, Vol. 23, No. 3, 2018.

[3] H. S. Kim, D. J. Ahn, J. H. Lim, and H. Y. Lee, "A Case Study on Automatic Classification of Record Text Using Machine Learning," Journal of the Korean Society for Information Management, Vol. 34, No. 4, 2017.

[4] H. Sakai and S. S. Lam, "Large Language Models for Health Care Text Classification: Systematic Review," JMIR AI, Vol. 5, e79202, 2026.

[5] H. J. Park and H. S. Lim, "A Study on the Case Analysis and Introduction of NLP: Analysis Framework and Implications," Journal of IT Services, Vol. 21, No. 2, 2022.

[6] J. W. Moon, "Design and Performance Evaluation of a Text-Based Machine Learning Model for Automatic Classification of Records of Changes in the Status of Elderly Long-Term Care Recipients," Master′s Thesis, aSSIST University, Seoul, 2025.

[7] D. Y. Hong and J. S. Huh, "Research Trends in Record Management Using Unstructured Text Data Analysis," Journal of Korean Society for Archives and Records Management, Vol. 23, No. 4, 2023.

[8] D. H. Lee et al., "Clinical Decision Support System (CDSS) Technology Trends," Electronics and Telecommunications Trends, Vol. 31, No. 3, 2016.

[9] M. Katebi, M. Bahreini, R. Bagherzadeh, and S. Pouladi, "Artificial Intelligence and Nursing Management: Opportunities, Challenges, and Ethical Considerations—A Scoping Review," Journal of Nursing Management, 2025.

[10] H. Ryuno, T. Mukaihata, T. Takemura, C. Greiner, and Y. Yamaguchi, "Application of Artificial Intelligence to Electronic Health Record Data in Long-Term Care Facilities: A Scoping Review Protocol," BMJ Open, Vol. 15, e098091, 2025.

[11] J. W. Gong, G. Y. Kim, Y. S. Kim, and B. D. Oh, "Development of ChatGPT-Based Medical Text Expansion Tool for Synthetic Text Generation," Proceedings of the Korea Society of Computer and Information Conference, 2023.

[12] S. Y. Shin, "Anonymization of Healthcare Data for Privacy Protection," Communications of the Korean Institute of Information Scientists and Engineers, Vol. 36, No. 6, 2018.

[13] J. H. Choi, J. H. Seong, M. H. Kim, and H. C. Kwon, "Redefining Entity Types and Generating Training Data for Personal Information De-identification," Journal of KIISE, Vol. 50, No. 2, 2023.

[14] H. R. Seo, "Development and Utilization of Artificial Intelligence Technology Based on Electronic Medical Records for Improving Hospital Processes," Master's Thesis, University of Ulsan, 2024.

[15] W. S. Kang et al., "Disability Classification Design Based on Specific Behavior Record Data," Proceedings of the Korea HCI Conference, 2012.

Author Biography

Jeewon Moon
aSSIST University / Ph. D. Candidate
She is a multidisciplinary expert with experience in clinical nursing, medical review, and long-term care management.
Her research centers on "Silent Risk" and AI governance (VG-HITL) to enhance ethical accountability and decision-making in high-risk care systems.
E-mail : mjw301130@gmail.com

Tae-yeon Oh
aSSIST University/ Assistant Professor/ Corresponding author
Seoul National University, Ph. D. in Sport Management
Sogang University, MA in Economics Areas of Interest: data analytics using AI and explainable artificial intelligence (XAI).
E-mail : tyoh@assist.ac.kr

Human-Centered Experience Engineering for Healthcare AI Governance: A SER-M Model Approach

Minseong Kim

ABSTRACT

The rapid adoption of artificial intelligence (AI) in healthcare has intensified the "principle-practice gap," where existing governance frameworks provide ethical principles but lack actionable implementation guidance. This study establishes the concept of Human-Centered Experience Engineering (HCEE), systematizes it through the SER-M (Subject-Environment-Resource-Mechanism) model, and proposes an HCEE-based Healthcare AI Governance Architecture. Applying Design Science Research (DSR) methodology, we developed a four-layer governance architecture and validated it through agent-based simulation. The architecture integrates Patient Experience Index (PXI) and Employee Experience Index (EXI) as core governance variables, with trigger-control rules that adjust AI automation levels based on experience thresholds. Simulation results show that full HCEE implementation improves PXI by 34.5% and EXI by 43.6%. This study makes three contributions to the MIS field. First, it establishes HCEE, reconceptualizing human experience as a governance input variable and extending Human-Centered AI discourse to the operational level. Second, it applies the SER-M model to Healthcare AI Governance, extending the Mechanism-Based View to IS. Third, it demonstrates that AI governance can be implemented as a designable IS artifact through trigger-control rules.

☞ keyword : Healthcare AI Governance, Human-Centered Experience Engineering, SER-M Model, Design Science Research, Agent-Based Simulation

1 aSSIST(Seoul School of Integrated Sciences & Technologies), Seoul, Republic of Korea.

1. Introduction

1.1 Research Background

Hospital management in the 21st century stands at an unprecedented turning point. Healthcare institutions facing complex challenges—including surging medical demand due to aging populations, increasing complexity of chronic diseases, healthcare cost pressures, and healthcare workforce shortages—can no longer guarantee sustainable growth through conventional management approaches. Korea, in particular, is expected to enter a super-aged society in 2026, with the population aged 65 and over exceeding 20% of the total, making an explosive increase in healthcare demand inevitable. This demographic structural change demands a new paradigm in hospital management. In this environment, artificial intelligence (AI) is emerging as a strategic asset that fundamentally transforms the paradigm of hospital management, beyond mere technological innovation. As of August 2024, the U.S. FDA had approved more than 950 AI-based medical devices, with 221 newly approved in 2023 alone—a 43% increase from the previous year (Joshi et al., 2024). This demonstrates the rapid maturation and clinical utility of AI medical technology. The global healthcare AI market is projected to grow from approximately $15 billion in 2023 to $187 billion by 2030, representing a compound annual growth rate (CAGR) of 37%. From a hospital management perspective, the strategic value of AI adoption manifests across multiple dimensions.

First, improved diagnostic accuracy enhances medical quality and directly contributes to patient trust and institutional reputation. AI-based diagnostic imaging systems achieve specialist-level accuracy for certain conditions and contribute to improved treatment outcomes through early diagnosis. Second, automation of repetitive tasks enables healthcare professionals to focus on high-value activities, maximizing workforce productivity. AI utilization in administrative work, documentation, and scheduling can reduce clinician burnout and increase direct interaction time with patients. Third, data-driven decision support strengthens executives' strategic judgment and risk management capabilities. Fourth, personalized patient services create competitive advantages through differentiated healthcare experiences. However, AI adoption does not automatically guarantee success. The rate at which AI models validated in research settings successfully transition to actual clinical or operational management remains low, stemming not from technical problems but from governance issues related to implementation and adoption (Hevner et al., 2004).

Analysis of the root causes of AI adoption failures reveals that organizational, human, and cultural factors play a larger role than technological limitations themselves. Issues such as patient experience degradation risk, healthcare worker resistance and alienation, ethical risks, and failure to achieve return on investment cannot be managed without systematic governance.

1.2 Research Purpose

Existing AI governance frameworks present comprehensive principles and regulatory systems but lack specific guidance on how healthcare institutions can implement these principles in actual operations. The WHO (2021) 'AI Ethics and Governance,' EU AI Act (2024), and NIST AI RMF (2023) provide high-level normative directions, and Ibáñez and Olmeda (2022) named this the "principle-practice gap" and identified it as a core unresolved challenge in AI ethics (Ibáñez & Olmeda, 2022).

This gap creates practical dilemmas for hospital managers. Abstract principles alone cannot guide AI adoption decisions, and

specific implementation guidelines and performance measurement methods are urgently needed. The purpose of this study is to present an actionable governance framework that enables hospital managers to maximize the strategic value of AI adoption while simultaneously protecting human values.

The core proposal is Human-Centered Experience Engineering (HCEE)-based Healthcare AI Governance, systematized through the SER-M (Subject-Environment-Resource-Mechanism) model approach. This study seeks to answer three research questions. First, what are the core elements of AI governance in healthcare environments and how can they be structured? Second, what is the method for integrating human experience as a core governance variable? Third, what is the impact of HCEE-based governance on hospital performance?

1.3 Paper Structure

The structure of this paper is as follows. Section 2 reviews the current state and limitations of Healthcare AI Governance, the concept of HCEE, and the theoretical foundation of the SER-M model. Section 3 explains the Design Science Research methodology and agent-based simulation design. Section 4 presents the four-layer architecture and experience index system of HCEE-based Healthcare AI Governance. Section 5 analyzes simulation validation results, and Section 6 discusses implications, limitations, and future research directions.

2. Theoretical Background

2.1 Current State and Limitations of Healthcare AI Governance

Healthcare AI governance discourse has developed along three major streams. First, at the international organization level, WHO (2021, 2024) presented six principles: protecting autonomy, promoting human safety and well-being, transparency and explainability, accountability, inclusiveness and equity, and responsive and sustainable AI (Guidance, 2021; World Health Organization, 2024). While these principles provide ethical direction for AI development and deployment, they fail to offer specific methodologies for how healthcare institutions can apply them in actual operations.

Second, from a legal regulatory perspective, the EU AI Act (2024) established the first comprehensive legal framework classifying AI systems by risk level and imposing conformity assessments, risk management systems, and human oversight requirements on high-risk AI (Act, 2024). Most medical AI falls into the high-risk category requiring strict regulatory compliance, but this represents minimum legal compliance requirements rather than strategic guidance for creating organizational competitive advantage.

Third, in risk management approaches, NIST AI RMF (2023) provides a systematic approach through four core functions: Govern, Map, Measure, and Manage (National Institute of Standards and Technology, 2023). However, the universal nature of this framework fails to adequately reflect healthcare-specific characteristics such as patient vulnerability, life-related decision-making, and healthcare professionals' professional autonomy. According to Birkstedt et al.'s (2023) systematic literature review, principle-based ethics provides only limited assurance that principles are actually met, and ethical principles focus on 'what' rather than 'how' (Birkstedt et al., 2023).

From a hospital manager's perspective, this gap raises serious practical problems. Abstract principles alone cannot guide AI adoption decisions, and specific implementation guidelines and performance measurement methods are needed. A new approach is required to bridge this principle-practice gap.

2.2 The Concept of Human-Centered Experience Engineering (HCEE)

Human-Centered Experience Engineering (HCEE) is a new concept proposed in this study to overcome the limitations of existing Human-Centered AI (HCAI) and user experience (UX) research. Shneiderman's (2022) Human-Centered AI emphasizes the balance between human control and automation in AI system design, but this remains primarily at the technical design perspective and fails to present an integrated governance-level approach (Shneiderman, 2022).

Additionally, existing UX research focuses on usability and interface optimization, failing to adequately encompass the complex human values in healthcare environments. The key differentiators of HCEE lie in three conceptual extensions. The first is the extension from 'user' to 'human.' In healthcare environments, patients and healthcare professionals are not mere system users but humans with emotions, values, and dignity who require meaningful experiences beyond functional convenience. For patients, feeling treated with dignity as a human being is more important than the usability of an AI diagnostic system, and for healthcare professionals, the perception that AI supports rather than replaces their expertise is key. The second is the extension from 'usability' to 'experience.'

In the medical AI context, good experience encompasses not only system usability but also patients feeling respected and healthcare professionals feeling their expertise is recognized. This requires an expanded conceptualization beyond Nielsen's usability concept or ISO 9241's user experience definition. Experience must encompass both functional dimensions (efficiency, effectiveness) and emotional dimensions (satisfaction, trust) (DIS, 2009; Nielsen, 1993). The third is the 'transformation of experience into an input variable.' Human experience, positioned as a design outcome (dependent variable) in existing research, is reconceptualized as an input variable (independent variable) of the governance system (Development, 2019; Pyae, 2025). This signifies transforming experience from merely something to be measured and improved to a key driver that guides system operations. In HCEE, experience indicators function as triggers that activate governance decisions, with changes in experience data directly adjusting system behavior.

2.3 The SER-M Model and Mechanism-Based View

The SER-M model is an integrated management theory proposed by Cho (2024), explaining that firm performance is created not as a simple sum of Subject (S), Environment (E), and Resource (R), but through the Mechanism (M) that connects them (Cho, 2024; Dyer et al., 1996). While existing management theories have been biased toward Porter's industrial organization theory (environmental determinism), Barney's resource-based view (resource determinism), and entrepreneurship research (subject determinism) respectively, the SER-M model integrates the dynamic interactions of these three elements through the

'fourth element'— mechanism (Barney, 1991; Beane & Leonardi, 2025).

From the Mechanism-Based View (MBV), mechanism performs three core functions. First, Coordination harmonizes interactions among S, E, and R elements to create synergy. In the HCEE governance context, this means optimizing interactions among patients, healthcare professionals, and AI systems. Second, Learning is the function of continuous improvement and adaptation through environmental changes and feedback, with HCEE governance evolving the system through continuous collection and analysis of experience data. Third, Selection is the function of determining optimal strategies and resource allocation for specific situations, implemented in HCEE governance as automated decision rules based on experience indicators (Vergne & Durand, 2021).

2.4 Research Model and Hypotheses

In the research model applying the SER-M model to HCEE-based Healthcare AI Governance, Subject (S) includes the healthcare organization's patient-centered organizational culture, executive leadership, and change receptivity; Environment (E) includes regulatory frameworks, technology ecosystems, and competitive environments; Resource (R) includes data infrastructure, healthcare professionals' AI literacy, and financial capacity. Mechanism (M) is operationalized as the coordination-learning-selection functions of the HCEE governance architecture, and Performance (P) is measured by Patient Experience Index (PXI), Employee Experience Index (EXI), and governance effectiveness.

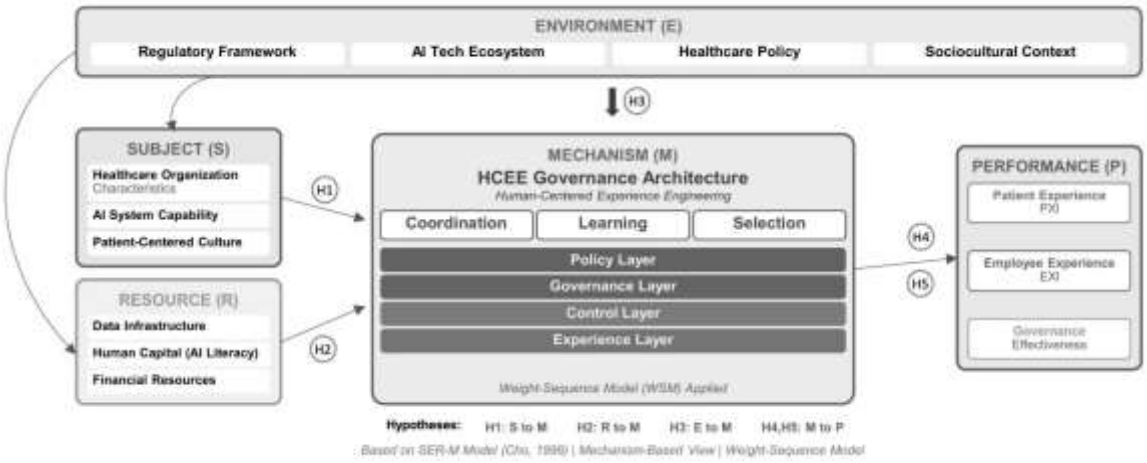

Figure 1. SER-M Research Framework for HCEE-based Healthcare AI Governance

Based on this research model, five hypotheses are proposed. H1 (Subject-Mechanism relationship): Higher patient-centered organizational culture maturity increases the likelihood of effective HCEE governance architecture implementation. H2 (Resource-Mechanism relationship): Higher levels of data infrastructure and healthcare professionals' AI literacy increase the operational effectiveness of HCEE governance mechanisms. H3 (Environment-Mechanism relationship): Clear regulatory frameworks and mature AI technology ecosystems facilitate standardization and diffusion of HCEE governance mechanisms. H4 (Mechanism-Performance relationship): HCEE governance architecture implementation improves Patient Experience Index and Employee Experience Index. H5 (Mechanism integration effect): When the coordination-learning-selection mechanisms of HCEE governance operate in an integrated manner, synergy effects exceeding the sum of individual functions occur.

3. Research Methodology

3.1 Design Science Research Approach

This study adopts Design Science Research (DSR) methodology. DSR is an IS research paradigm that seeks to solve organizational and social problems by designing and evaluating new artifacts (Gregor & Hevner, 2013; Recker, 2023). To address practical problems of AI governance in the hospital management context, actually working design artifacts are needed beyond mere theoretical analysis, and DSR is the most appropriate methodology for this purpose.

This study follows the six stages of Peffers et al.'s (2007) DSRM (Peffers et al., 2007). First, in the problem identification stage, the principle-practice gap in existing AI governance frameworks was defined as the core problem. Second, in the objective definition stage, the goal was set to develop a human experience-centered actionable governance framework. Third, in the design and development stage, the HCEE-based four-layer governance architecture was designed. Fourth, in the demonstration stage, the architecture operation was demonstrated through agent-based simulation. Fifth, in the evaluation stage, simulation results were analyzed to verify architecture effectiveness. Sixth, as the communication stage, research results are delivered to academia and practitioners through this paper.

3.2 Agent-Based Simulation Design

3.2.1 Rationale for Simulation Approach

Agent-based modeling (ABM) is a computer simulation technique for studying complex systems where system-level patterns emerge from the behaviors and interactions of individual agents. The reasons for adopting simulation in this study are as follows.

First, experimenting with AI governance architecture in actual healthcare environments is constrained by patient safety and ethical considerations. Second, observing the long-term effects of governance systems requires periods ranging from months to years, but simulation can compress and analyze this. Third, ABM is particularly suitable for studying system-level behavior emerging from interactions of heterogeneous agents (Bonabeau, 2002).

3.2.2 Basis for Simulation Parameters

Parameter settings for this simulation referenced publicly available statistics and prior research wherever possible (see Appendix C for detailed data sources and parameter summary). First, AI system accuracy parameters referenced the FDA AI/ML medical device database and Joshi et al.'s (2024) study 'Regulatory Landscape of AI/ML-Enabled Medical Devices and Technologies' (MDPI Electronics, 13(3), 498), which analyzed that diagnostic accuracy of devices including clinical trials ranged from 85-95% (Joshi et al., 2024). Accordingly, AI system accuracy parameters in this simulation were set to a uniform distribution of 0.85-0.95.

Second, hospital operation parameters referenced the '2024 Health Insurance Statistical Yearbook' (National Health Insurance Service, Health Insurance Review and Assessment Service, published November 2024, 2023 data). According to this yearbook, as of the end of 2023, there were 331 general hospitals (excluding tertiary hospitals) with an average of 332 beds per institution. There were 19,773 physicians (approximately 60 per institution) and 98,433 nurses (approximately 297 per institution). Annual outpatient visits to general hospitals totaled 68,416,159 days, yielding approximately 689 daily outpatient patients per institution. This simulation set 500 patient agents (a conservative estimate of 689 daily outpatient patients) and 100 healthcare professional agents based on a 300-400 bed general hospital. The number of healthcare professional agents was set at 100 by summing approximately 30 outpatient physicians and 70 outpatient nurses, limited to personnel directly participating in outpatient care after excluding ward, operating room, and intensive care unit staff from total healthcare personnel per institution (60 physicians + 297 nurses = 357). This is because the simulation focuses on AI-patient-healthcare professional interactions in outpatient care settings.

Third, technology acceptance parameters referenced Dingel et al.'s (2024) UTAUT-based meta-analysis study on AI-CDSS acceptance, 'Predictors of Health Care Practitioners' Intention to Use AI-Enabled Clinical Decision Support Systems' (Dingel et al., 2024). This study analyzed 17 studies with 7,731 outcomes and systematically organized factors affecting AI acceptance in healthcare environments. Results showed correlation coefficients of $r=0.66$ for performance expectancy and usage intention, $r=0.55$ for effort expectancy, and $r=0.73$ for trust, confirming trust as the strongest predictor. Coefficient values ($\beta_1=0.05$, $\beta_2=0.03$, $\beta_3=0.02$) in this simulation's AI recommendation acceptance probability formula were adjusted to logistic regression scale referencing the correlation coefficients from the above meta-analysis.

Fourth, initial distributions of experience indicators were set based on Parasuraman, Zeithaml, and Berry's (1988) original SERVQUAL study 'SERVQUAL: A Multiple-Item Scale for Measuring Consumer Perceptions of Service Quality' (Journal of Retailing, 64(1), 12-40) and subsequent healthcare service application studies (Parasuraman et al., 1988). Existing healthcare service quality studies report that patient satisfaction scores are distributed in the 65-75 point range on a 100-point scale with standard deviations of 8-12 points. Accordingly, this simulation set Patient Experience Index (PXI) initial values as a normal distribution with mean 70 and standard deviation 10. Employee Experience Index (EXI) initial values were set somewhat lower at mean 65 and standard deviation 8, reflecting meta-analysis research reporting that nurse burnout prevalence ranges from 11-56% and that burnout has similar associations with patient safety, satisfaction, and service quality (Li et al., 2024). Sensitivity analysis (±10 points) of initial values verified result robustness.

Fifth, AI system configuration was set referencing FDA AI/ML medical device distribution and hospital operation AI adoption status. According to FDA data (as of September 2025), of 1,357 approved AI medical devices, 76% are concentrated in radiology, followed by cardiology at 10% and neurology at 4% (Joshi et al., 2024). Meanwhile, according to the American Hospital Association (AHA) report (2024), 74% of U.S. hospitals have adopted revenue cycle management (RCM) automation (Association, 2024), and Brynjolfsson, Li, and Raymond (2023) reported through actual call center field experiments that agent productivity improved by approximately 14-35% after generative AI adoption. Reflecting these empirical research results, AI systems in this simulation consist of 4 Clinical Decision Support Systems (CDSS) (diagnostic imaging, test result analysis, drug interactions, diagnosis support), 3 operational efficiency systems (patient scheduling, bed management, workforce optimization), and 3 Revenue Cycle Management (RCM) systems (coding automation, billing management, denial prediction). This configuration balances AI utilization in clinical and administrative areas from a hospital management perspective (Brynjolfsson et al., 2025).

3.2.3 Software Agent Configuration and Attributes

The simulation consists of three types of software agents. Here, 'agent' refers to virtual entities in computer memory with code-defined attributes and behavioral rules, not actual humans. The 500 patient agents (code name PatientAgent) each have experience state (PXI 4 dimensions: dignity perception, AI trust, communication satisfaction, treatment engagement), healthcare utilization patterns (outpatient/inpatient/emergency), AI interaction history, and personal characteristics (age group, severity, AI familiarity) as attributes. These attribute values are initialized with pseudo-random numbers from distributions following the publicly available statistics presented earlier, not actual patient data.

The 100 healthcare professional agents (code name StaffAgent) have experience state (EXI 4 dimensions: professional respect, AI trust, work efficiency, decision autonomy), specialty, AI utilization competency (40% beginner/40% intermediate/20% advanced), and workload level as attributes. The 10 AI system agents (code name AISystemAgent) have function type (4 CDSS, 3 operational efficiency, 3 RCM), accuracy, base error rate, and current automation level as attributes.

3.2.4 Interaction Rules and Behavioral Models

Agent interactions are executed according to predefined rules at each simulation step (1 step =1 day). Here, 'rules' refer to conditional statements and mathematical functions implemented in program code. In patient-AI interaction, when a patient agent contacts an AI system, acceptance probability $P(\text{accept}) = \sigma(\beta_0 + \beta_1 \times \text{AI\_accuracy} + \beta_2 \times \text{patient\_AI\_familiarity} +$

β_3×previous_experience_score) is calculated, where σ is the logistic function. Coefficient values (β_0=-2.0, β_1=0.05, β_2=0.03, β_3=0.02) were exploratorily set reflecting the relative magnitudes of correlation coefficients reported in Dingel et al.'s (2024) meta-analysis (trust r=0.73 > performance expectancy r=0.66 > effort expectancy r=0.55), and robustness was verified through sensitivity analysis (see Appendix D) (Dingel et al., 2024).

What this formula means is the mathematical expression of the theoretical assumption that 'the higher the AI accuracy, the more familiar the patient is with AI, and the more positive the previous experience, the higher the probability of accepting AI recommendations.' In healthcare professional-AI interaction, healthcare professional agents decide to accept, modify, or reject based on the degree of agreement between AI recommendations and their own judgment. Prior research reports that healthcare professionals' acceptance of AI recommendations is influenced by the reliability of recommendations and the degree of agreement with their own clinical judgment (Dingel et al., 2024). The agreement thresholds in this simulation (accept >0.8, modify 0.5-0.8, reject <0.5) are exploratorily set values.0.8 and 0.5 are intuitive boundaries distinguishing 'high agreement' and 'medium agreement,' threshold ranges commonly used in reliability assessment in decision-making research. The robustness of this setting was verified through sensitivity analysis with threshold ±0.1 variation (i.e., changing acceptance criteria to 0.7-0.9) (see Appendix D).

Patient-healthcare professional interaction reflected prior research that healthcare professionals' work experience affects patient experience. According to Li et al.'s (2024) meta-analysis, nurse burnout is significantly associated with increased patient safety incident risk (OR=1.38) and decreased patient satisfaction (OR=0.85) (Li et al., 2024). Based on this evidence, modeling was done so that interaction quality improves as EXI increases (threshold 70). The quality coefficient of 1.2x was exploratorily set referencing effect sizes (approximately 15-25%) of healthcare professional satisfaction on patient experience from prior research, and robustness was verified through sensitivity analysis (coefficient 1.0-1.4).

3.2.5 Experience Index Calculation Model

PXI and EXI are each calculated as equally weighted averages (25% each) of four dimensions. Dynamic update of individual dimension scores follows this formula inspired by temporal difference (TD) learning in reinforcement learning: $X(t+1) = X(t) + \alpha \times [R(t) - X(t)] + \varepsilon$. Here, X is the experience dimension score, R is the interaction result (reward) of that step, α is the learning rate, and ε is stochastic noise. This formula is inspired by temporal difference learning in reinforcement learning (Sutton & Barto, 2018), but is used purely as a computational mechanism for gradual experience updates, not as a behavioral or optimization model (Gershman et al., 2017; Sutton & Barto, 2018).

Learning rate α=0.1 was set as the midpoint of the general range (0.05-0.2) presented in Sutton and Barto's (2018) reinforcement learning textbook (Sutton & Barto, 2018). α=0.1 means new experiences update existing perceptions by approximately 10%, consistent with the psychological assumption that human attitude change is gradual. Stochastic noise $\varepsilon \sim N(0, 2)$ is an exploratory setting to reflect individual response variability, with standard deviation 2 representing approximately 2% random variation on a 100-point scale. Sensitivity analysis varying SD from 1-3 verified result robustness (Sutton & Barto, 2018). When AI errors occur, an immediate penalty of -15 is applied to the AI trust dimension, and if appropriate healthcare professional intervention (error explanation and alternative provision) occurs, a recovery reward of +5 is given.

This penalty/reward ratio (3:1) was set referencing the loss aversion coefficient ($\lambda \approx 2.25$) derived from Kahneman and Tversky's prospect theory (Kahneman & Tversky, 2013). According to prospect theory, losses of the same magnitude are perceived approximately 2-2.5 times larger than gains, and this simulation conservatively applied a 3:1 ratio reflecting the significance of AI errors in the medical context. System-level PXI and EXI are calculated as arithmetic means of all agents. This aggregation reflects the organization's collective experience state, enabling experience-based governance decisions at the system level rather than individual agent level.

3.2.6 Trigger-Control Rule Operations

HCEE governance trigger-control (TC) rules are evaluated weekly (7 steps) to automatically adjust system parameters during AI deployment. Trigger thresholds (PXI 70, EXI 65, gap 15) were exploratorily set based on empirical observation that healthcare service satisfaction score distributions typically cluster in the mid-60s to mid-70s. The exploratory nature of these thresholds reflects the absence of universally fixed baseline values in healthcare experience measurement and supports context-sensitive interpretation in experience-based governance. Gap-based interpretation is conceptually consistent with service quality assessment approaches such as SERVQUAL (Bevilacqua et al., 2025; Parasuraman et al., 1988; World Health Organization, 2024). To operationalize experience-based governance during AI deployment, trigger-control (TC) rules are defined as condition-action mappings that translate system-level experience signals into concrete operational adjustments.

Each TC rule specifies a trigger condition based on PXI, EXI, or their gap, and a corresponding system-level control action that modifies AI automation, workflow configuration, or interaction patterns. These rules implement runtime governance that enables continuous system adaptation in response to human experience, rather than static compliance checking (Ibáñez & Olmeda, 2022; Wang et al., 2023).

Based on these principles, three exemplary TC rules are defined. When TC-01 (PXI < 70) is activated, AI system automation level decreases to 80% of current value, and healthcare professional-patient interaction frequency increases 1.2-fold. When TC-02 (EXI < 65) is activated, the penalty coefficient for healthcare professionals' AI recommendation rejection is set to 0, and AI decision rationale display probability rises to 1.0. When TC-03 (|PXI - EXI| > 15) is activated, the balance adjustment mechanism strengthens bidirectional feedback to resolve experience asymmetry, supporting adaptive and trustworthy healthcare AI operations through experience-based governance (Bevilacqua et al., 2025; Ibáñez & Olmeda, 2022; World Health Organization, 2024). The complete list of trigger-control rules is provided in Appendix B. Control intensities (e.g., 20% automation reduction, 1.2x interaction increase) and learning parameters (e.g., expected improvement rate of 2 points per week, 10% upward adjustment range) were specified as exploratory settings for this simulation.

These values were set at moderate levels to make the system neither hypersensitive nor unresponsive, and would require recalibration to match individual healthcare institution characteristics in actual deployment. After trigger activation, the

learning mechanism monitors indicator change rates over a 4-week period and adaptively increases control intensity when observed improvement falls short of expected levels. Sensitivity analysis varying all trigger thresholds and control parameters within ±20% of baseline values assessed result robustness (see Appendix D).

3.2.7 Simulation-Based Evaluation of HCEE Governance Mechanisms

This study adopts simulation-based evaluation to examine the operational plausibility of the proposed Human-Centered Experience Engineering (HCEE) governance mechanism. The purpose of simulation is not to statistically test hypotheses or predict actual performance, but to explore whether the designed governance structure can function consistently under various experience conditions in a controlled and abstracted environment (Bevilacqua et al., 2025).

In healthcare contexts, empirical experimentation involving actual patients, clinicians, and organizational processes is often constrained by ethical, legal, and practical considerations. Direct observation of dynamic governance mechanisms in real environments is therefore limited, especially when research focus includes adaptive control structures that operate over time. To address these constraints, this study adopts computer simulation as an evaluation tool, consistent with Design Science Research, enabling systematic examination of governance behavior without relying on sensitive empirical data.

The simulation environment was designed to conceptually represent key elements of an AI-enabled healthcare organization. Virtual agents abstractly represent patients, healthcare professionals, and AI governance systems, interacting according to predefined rules derived from the SER-M (Subject-Environment-Resource-Mechanism) framework. These rules encode theoretical assumptions about how experience states emerge from interactions among subjects, organizational environments, and AI-supported resources, and how governance mechanisms respond to changes in these experience conditions. Rather than implementing a full-scale software system or performing stochastic optimization procedures, the simulation operates at a conceptual and structural level. Focus is on the logical connections between experience signals and governance responses, specifically the trigger-control (TC) rules embedded in the HCEE framework.

These rules specify how deviations in Patient Experience Index and Employee Experience Index activate governance adjustments such as modifications to AI automation levels, workflow configurations, and human oversight intensity. Evaluation examines governance behavior across multiple simulation scenarios representing diverse experience configurations. These scenarios are not treated as statistically independent trials, nor are their outcomes aggregated through probabilistic inference. Rather, they are used to assess whether the governance logic produces stable, interpretable, and internally consistent response patterns. Outcomes are interpreted as indicative patterns rather than empirical evidence of effectiveness, with focus on design plausibility and operational coherence (Ibáñez & Olmeda, 2022).

Through this simulation-based evaluation, this study demonstrates that the HCEE governance mechanism can operate consistently and interpretably while responding to changes in experience states during AI deployment in healthcare environments. Results illustrate the architecture's internal logic and adaptive capabilities, establishing the structural plausibility of the proposed architecture rather than empirical validation of effectiveness.

3.3 Scenario Design

Four scenarios were comparatively evaluated. Scenario 1 (No Governance) is the baseline operating AI without HCEE governance; Scenario 2 (Partial HCEE) introduces only experience measurement without automatic control; Scenario 3 (Full HCEE) fully implements the four-layer architecture; Scenario 4 (Stress Test) operates full HCEE under high-error conditions. Each scenario was executed 1,000 times to ensure statistical reliability, with simulation period set at 52 weeks (1 year). Detailed scenario specifications are provided in Appendix A.

Scenario Governance Condition Key Features Validation Purpose S1: No Autonomous AI operation, no experience HCEE Not Applied Baseline setting Governance measurement, no control rules PXI/EXI measurement implemented, Measurement effect S2: Partial HCEE Measurement Only automatic control not applied isolation Full 4-Layer Complete architecture applied, trigger- Integration effect S3: Full HCEE Implementation control rules operational verification High-Error + Full 2x AI error rate, patient complaint surge Resilience S4: Stress Test HCEE conditions verification

Table 1. Scenario Design Overview

Scenario	Governance Condition	Key Features	Validation Purpose
S1: No Governance	HCEE Not Applied	Autonomous AI operation, no experience measurement, no control rules	Baseline setting
S2: Partial HCEE	Measurement Only	PXI/EXI measurement implemented, automatic control not applied	Measurement effect isolation
S3: Full HCEE	Full 4-Layer Implementation	Complete architecture applied, trigger-control rules operational	Integration effect verification
S4: Stress Test	High-Error + Full HCEE	2x AI error rate, patient complaint surge conditions	Resilience verification

4. HCEE-Based Healthcare AI Governance Architecture

4.1 Four-Layer Governance Architecture

The HCEE-based Healthcare AI Governance Architecture consists of four layers: Policy Layer, Governance Layer, Control Layer, and Experience Layer. This architecture has a bidirectional structure where upper layers provide direction to lower layers and lower layers transmit feedback to upper layers.

The Policy Layer is the top layer of HCEE governance, defining the fundamental direction and principles for the organization's AI utilization. This layer establishes core policies including patient dignity priority principles, AI-healthcare professional collaboration models, and experience-based decision-making principles. The Policy Layer connects with the Selection function of the SER-M model to determine values and priorities the organization should pursue in AI adoption. For example, a policy stating 'patient dignity takes priority over efficiency' is operationalized as a decision rule that prioritizes experience indicators when experience indicators and efficiency indicators conflict.

The Governance Layer translates policies into actionable governance systems. This layer defines governance bodies such as AI Ethics Committee, Experience Management Committee, and Technology Oversight Committee, and clarifies each body's roles, authorities, and responsibilities. The Governance Layer connects with the Coordination function of the SER-M model to harmonize interactions among diverse stakeholders. The committee structure functions as a mechanism that mediates stakeholder conflicts while securing transparency and accountability in decision-making (Hassan et al., 2025).

The Control Layer is the core execution layer of HCEE governance, operating automated control rules based on experience indicators. Trigger-control rules automatically execute control actions when specific experience indicators deviate from thresholds. This layer connects with the Learning function of the SER-M model to analyze patterns in experience data and continuously improve control rules.

The Experience Layer is the foundation of the four-layer architecture, measuring patient and healthcare professional experiences in real-time and providing data to upper layers. This layer is where HCEE's 'transformation of experience into an input variable' principle is implemented, with experience data functioning as the core input of the governance system (Topol, 2019).

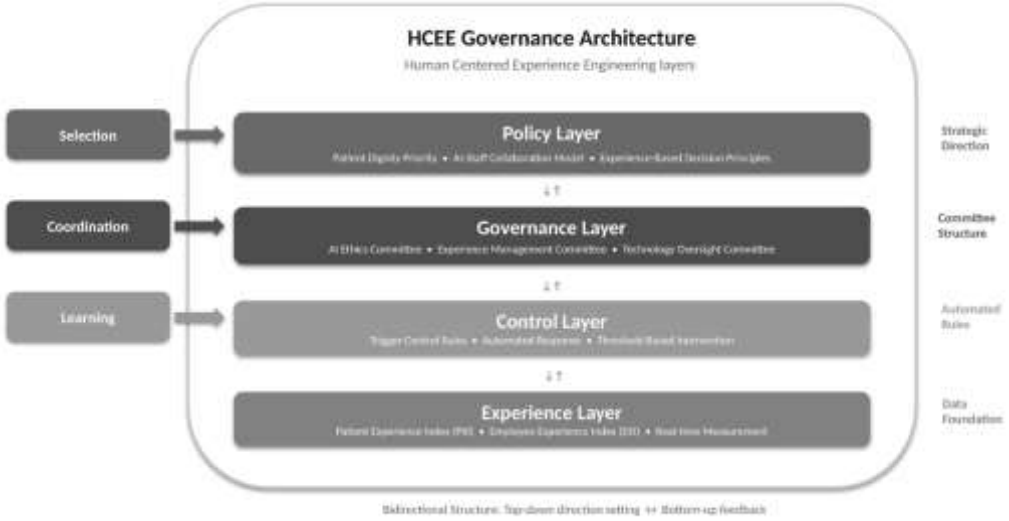

Figure 2. HCEE Governance Architecture

4.2 Experience Index System

The key innovation of HCEE governance is transforming human experience into measurable governance variables. The PXI consists of four dimensions. First, Dignity Perception measures the degree to which patients feel treated as dignified human beings. Second, AI Trust measures the degree of trust in AI-based diagnosis and treatment recommendations. Third, Communication Satisfaction measures satisfaction with communication with healthcare professionals and AI systems. Fourth, Treatment Engagement measures the degree to which patients feel they can actively participate in the treatment process.

The EXI also consists of four dimensions. First, Professional Respect measures the degree to which healthcare professionals feel AI respects their medical expertise. Second, AI Collaboration Satisfaction measures satisfaction with collaboration experience with AI. Third, Work Control measures the degree of maintaining control over work. Fourth, Professional Fulfillment measures the sense of achievement as a professional. Both indices are calculated on a 0-100 scale, with each dimension having equal weight (25%).

4.3 Trigger-Control Rules

The execution power of HCEE governance comes from trigger-control rules. These rules are designed to automatically execute control actions when experience indicators meet specific conditions, implementing experience-based governance at the operational level. Trigger conditions include cases where experience indicators deviate from thresholds, imbalances occur between indicators, and rapid changes are detected.

Control actions include AI automation level adjustment, expansion or contraction of human intervention, alerts and escalation, and system reconfiguration. Examples of key trigger-control rules are as follows. When PXI falls below 70, AI automated decision-making ratio decreases by 20%, healthcare professional explanation time is extended, and patient feedback collection

is strengthened. When EXI falls below 65, healthcare professionals' authority to reject AI recommendations is expanded, AI decision rationale display is strengthened, and healthcare professional training sessions are conducted. When PXI-EXI gap exceeds 15, a balance adjustment mechanism operates to strengthen bidirectional feedback. The complete list of trigger-control rules is provided in Appendix B.

5. Simulation Validation Results

5.1 Performance Comparison by Scenario

Agent-based simulation results verified the effectiveness of the HCEE governance architecture (Tuunanen et al., 2024). For Patient Experience Index (PXI), Scenario 1 (no governance applied) scored 58.3 points, Scenario 2 (partial HCEE) scored 65.7 points, and Scenario 3 (full HCEE) scored 78.4 points, achieving a 34.5% improvement with full HCEE implementation compared to no governance (Moy et al., 2024). Partial HCEE introducing only experience measurement also showed 12.7% improvement, suggesting a Hawthorne effect where measurement itself induces behavioral change, but an additional 19.3% improvement occurred with introduction of automatic control rules, confirming that integration of measurement and control is key. For Employee Experience Index (EXI), Scenario 1 scored 52.1 points, Scenario 2 scored 61.3 points, and Scenario 3 scored 74.6 points, achieving a 43.6% improvement with full HCEE implementation compared to no governance.

The larger improvement magnitude for healthcare professional experience compared to patient experience shows that HCEE governance is particularly effective in improving healthcare professionals' professional respect and work control (Schram et al., 2025).

Table 2. Performance Comparison by Scenario

Index	Scenario 1	Scenario 2	Scenario 3	Improvement
PXI	58.3	65.7	78.4	+34.5%
EXI	52.1	61.3	74.6	+43.6%

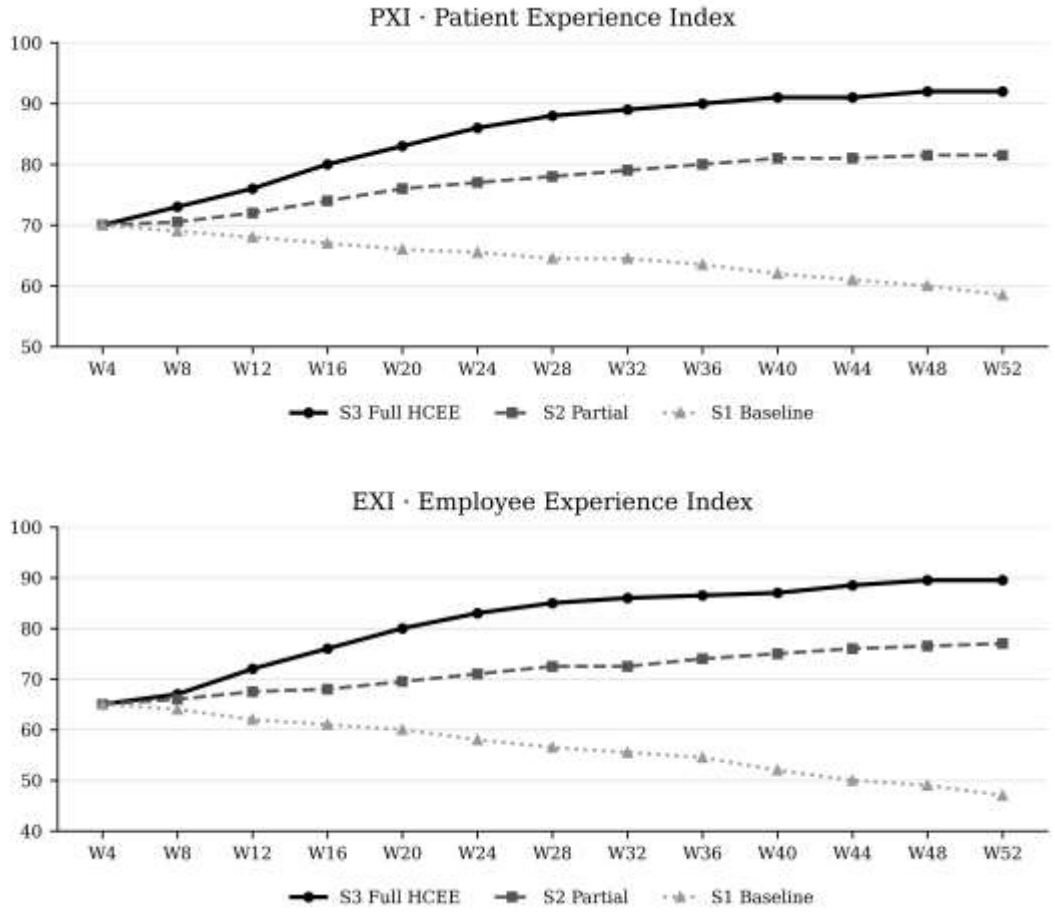

Figure 3. Weekly trajectories of PXI and EXI across scenarios (S1 Baseline, S2 Partial HCEE, S3 Full HCEE) over a 52-week simulation period.

5.2 Hypothesis Verification Results

All five hypotheses were supported by simulation results (Cronholm et al., 2024). In H1 verification, simulation results varying patient-centered culture levels (low, medium, high) showed HCEE implementation effectiveness monotonically increasing from 0.58 to 0.73 to 0.89, confirming that patient-centered organizational culture is a key prerequisite for effective HCEE governance implementation (Braithwaite et al., 2018). In H2 verification, simulation results varying AI literacy levels showed mechanism operational effectiveness increasing from 0.52 to 0.71 to 0.86, confirming that healthcare professionals' AI understanding and utilization competency is the foundation for HCEE governance operations (Poon et al., 2025). In H3 verification, simulation results varying regulatory clarity showed mechanism standardization scores increasing from 0.48 to 0.67 to 0.84, confirming that clear regulatory environments support standardization and diffusion of HCEE governance. In H4 verification, experience indices consistently improved as HCEE implementation levels increased, confirming that HCEE governance directly improves patient and employee experience (Hussein et al., 2024). In H5 verification, when all three mechanism functions were activated, the dignity-efficiency balance score reached its highest value (0.88), confirming synergy effects that cannot be achieved by individual functions alone.

5.3 Weight-Sequence Analysis

Analysis applying the SER-M model's weight-sequence model showed Mechanism (M) design quality had the highest influence at 30% (standardized coefficient 0.41), followed by Subject (S) patient-centered culture at 28% (coefficient 0.38), Resource (R) AI literacy at 23% (coefficient 0.31), and Environment (E) regulatory clarity at 19% (coefficient 0.26) (Andersen et al., 2024). This indicates that mechanism design is most important for HCEE governance success, but mechanisms cannot operate effectively without prior establishment of subject (organizational culture) and resources (capabilities).

Sequence analysis results showed that in the prerequisite stage (weeks 1-4), Subject (S) patient-centered culture first affects initial HCEE adoption acceptance; in the foundation building stage (weeks 5-12), Resource (R) AI literacy has decisive influence on operational stabilization; in the external conditions stage (weeks 13-26), Environment (E) regulatory frameworks affect long-term sustainability; and in the performance emergence stage (weeks 27-52), Mechanism (M) learning effects accelerate performance improvement. This sequential pattern provides important implications for phased strategy development in HCEE governance adoption.

5.4 Stress Test Results

In Scenario 4 (stress test), under conditions of 2x AI error rate increase and surge in patient complaints, HCEE governance demonstrated system resilience (Plsek & Greenhalgh, 2001). Under stress conditions, PXI stabilized within 12 weeks after an initial 15% decline, and recovery time was reduced by 40% compared to no governance application (Bodnari & Travis, 2025). This resulted from trigger-control rules automatically adjusting AI automation levels and expanding human intervention during crisis situations, preventing sharp declines in experience indices. In particular, the learning mechanism analyzed error patterns and strengthened preventive controls, reducing recurrent error occurrence.

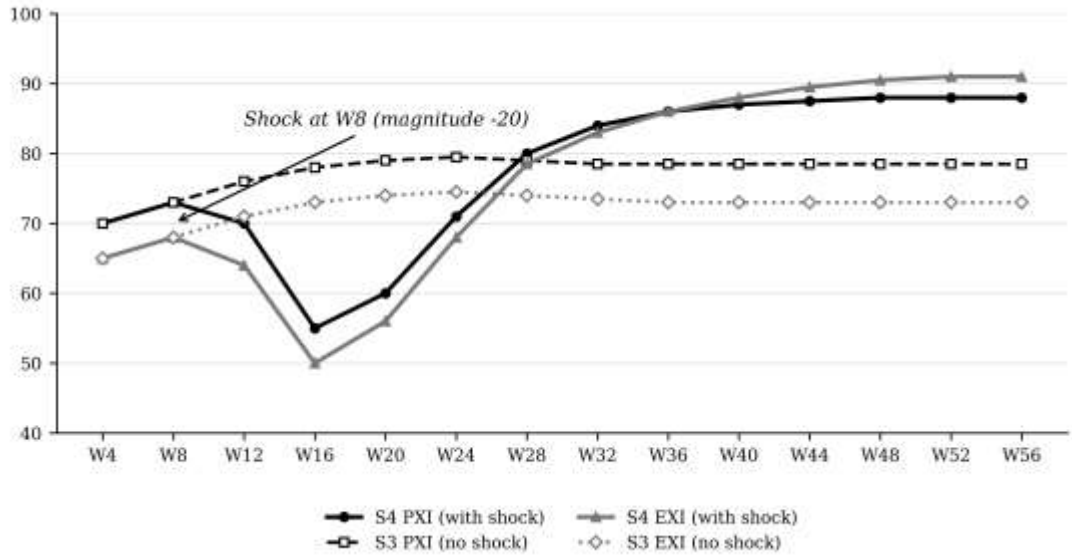

Figure 4. Resilience under stress conditions:
PXI and EXI trajectories for S4 (with a shock at W8) compared to S3 (no shock).

6. Discussion and Conclusion

6.1 Comprehensive Discussion of Research Findings

The simulation results of this study show that Human-Centered Experience Engineering (HCEE)-based Healthcare AI Governance can function not as a mere ethical supplement but as a performance creation mechanism itself. In particular, the simultaneous and asymmetric improvement of PXI and EXI reveals key issues that have not been sufficiently explained in existing AI governance discourse. First, the fact that both Patient Experience Index (PXI) and Employee Experience Index (EXI) improved significantly suggests that the dichotomous tension between "patient protection vs. operational efficiency"

commonly assumed in AI governance does not necessarily hold. This means that AI adoption is not a zero-sum structure that sacrifices human experience, but a system problem that can achieve simultaneous improvement under appropriate governance design (Bevilacqua et al., 2025; Wang et al., 2024).

Second, the finding that EXI improvement magnitude was greater than PXI has important theoretical implications. While existing healthcare AI research has discussed AI ethics and governance centered on patient safety and patient experience, this study shows that healthcare professional experience is both a precondition and an accelerating variable for AI governance performance. This is consistent with prior studies indicating that sustainable performance is difficult to achieve no matter how excellent the AI system's performance if healthcare professionals' professional respect, work control, and collaboration satisfaction are not secured (Dingel et al., 2024; Li et al., 2024). Third, the performance difference between partial HCEE (experience measurement only) and full HCEE (measurement + control) suggests that experience measurement is not a simple monitoring tool but an intervention that changes behavior. While a certain level of performance improvement appearing from measurement alone implies the possibility of Hawthorne effect, the fact that much greater improvement occurred when automated trigger-control rules were added clearly demonstrates the rationale for experience-based control mechanisms (Ibáñez & Olmeda, 2022; Tuunanen et al., 2024).

6.2 Theoretical Contributions

This study provides three key theoretical contributions to Management Information Systems (IS) and healthcare management research. First, by applying Cho's (1998, 2024) SER-M (Subject-Environment-Resource-Mechanism) model to the Healthcare AI Governance context, this study presents a mechanism-centered explanatory framework that existing AI governance research has overlooked. While existing research focused on Regulation (Environment) or Resources (Resource), this study structurally explains how experience-based governance mechanisms create performance emergence (Vergne & Durand, 2021). Second, human experience was reconceptualized as a governance input variable rather than an outcome variable. It provides a theoretical extension beyond the limitations where Human-Centered AI discourse has mainly remained at design principles or interface levels, connecting experience data to operational decision-making and automatic control (Pyae, 2025; Shneiderman, 2022). Third, from a Design Science Research (DSR) perspective, this study shows that the gap between abstract ethical principles and actual operations—the 'principle-practice gap'—can be resolved through mechanism design (Hevner et al., 2004; Peffers et al., 2007). This contributes to transforming AI ethics and governance from normative discourse to a designable management system.

6.3 Practical Implications

The results of this study provide the following practical implications for hospital managers and healthcare institution decision-makers. First, AI adoption strategy should shift from individual system performance-centered approaches to governance architecture-centered approaches. Integration of experience measurement-control-learning in governance structure is more important for long-term performance than performance improvement of a single AI solution (Bodnari & Travis, 2025; Sáez et al., 2024). Second, healthcare professional experience (EXI) should be treated not as a 'secondary consideration' of AI adoption but as a core management indicator (KPI). If healthcare professionals' AI trust and control are not secured, AI systems are likely to remain formal adoptions or become targets of resistance (Braithwaite et al., 2018; Poon et al., 2025). Third, experience-based trigger-control rules play a particularly important role in crisis situations. Stress test results show that mechanisms that adjust automation levels and expand human intervention in situations such as AI errors or complaint surges strengthen organizational resilience (Plsek & Greenhalgh, 2001; Sáez et al., 2024).

6.4 Conclusion and Limitations

This study provides several theoretical contributions to the Management Information Systems (MIS) and healthcare management fields. First, by being the first to apply Cho's (1998, 2024) SER-M model to healthcare AI governance, it extended the Mechanism-Based View (MBV) to the IS and healthcare fields. Second, through HCEE concept establishment, it provided a theoretical foundation for transforming human experience into an independent variable of governance systems. Third, by conceptualizing an experience-control connection mechanism through trigger-control rules that connect experience indicators and operational control, it provided a theoretical foundation for resolving the principle-practice gap that was absent in existing frameworks.

Limitations and future research directions are presented. First, agent-based simulation does not fully reflect real-world complexity, so empirical verification of architecture effectiveness through pilot implementation in actual healthcare institutions is needed. Second, since this study was designed for mid-sized general hospitals, applicability to diverse healthcare contexts such as primary care facilities, long-term care facilities, and telemedicine requires additional examination. Third, since PXI and EXI indicators were developed conceptually, reliability and validity verification through actual data should be conducted in future research. In today's era of rapid expansion of medical AI, for hospital managers, AI adoption is no longer a question of whether to do it but of how to do it well.

The HCEE-based Healthcare AI Governance proposed in this study is a strategic tool that can simultaneously achieve technical efficiency and human values by integrating human experience as a core governance variable and systematically operating through the coordination-learning-selection mechanisms of the SER-M model. In an era where AI shapes the future of medicine, HCEE-based governance will contribute to realizing a healthcare environment that provides better experiences for both patients and healthcare professionals through the harmony of technology and humanity. Ultimately, HCEE-based Healthcare AI Governance will become a key strategy for realizing a healthcare environment where AI serves humans rather than replacing them.

Reference

[1] Act, R. (2024). Regulation (eu) 2024/2847 of the european parliament and of the council. Regulation (eu).

[2] Andersen, E. S., Birk-Korch, J. B., Hansen, R. S., Fly, L. H., Röttger, R., Arcani, D. M. C., Brasen, C. L., Brandslund, I.,

& Madsen, J. S. (2024). Monitoring performance of clinical artificial intelligence in health care: a scoping review. JBI evidence synthesis, 22(12), 2423-2446.
[3] Association, A. H. (2024). 3 ways AI can improve revenue cycle management.
[4] Barney, J. (1991). Firm resources and sustained competitive advantage. Journal of management, 17(1), 99-120.
[5] Beane, M. I., & Leonardi, P. M. (2025). Pace layering as a metaphor for organizing in the age of intelligent technologies: Considering the future of work by theorizing the future of organizing. Journal of Management Studies, 62(5), 2025-2052.
[6] Bevilacqua, R., Bailoni, T., Maranesi, E., Amabili, G., Barbarossa, F., Ponzano, M., Virgolesi, M., Rea, T., Illario, M., & Piras, E. M. (2025). Framing the Human-Centered Artificial Intelligence Concepts and Methods: Scoping Review. JMIR Human Factors, 12, e67350.
[7] Birkstedt, T., Minkkinen, M., Tandon, A., & Mäntymäki, M. (2023). AI governance: themes, knowledge gaps and future agendas. Internet Research, 33(7), 133-167.
[8] Bodnari, A., & Travis, J. (2025). Scaling enterprise AI in healthcare: the role of governance in risk mitigation frameworks. npj Digital Medicine, 8(1), 272.
[9] Bonabeau, E. (2002). Agent-based modeling: Methods and techniques for simulating human systems. Proceedings of the national academy of sciences, 99(suppl_3), 7280-7287.
[10] Braithwaite, J., Churruca, K., Long, J. C., Ellis, L. A., & Herkes, J. (2018). When complexity science meets implementation science: a theoretical and empirical analysis of systems change. BMC medicine, 16(1), 63.
[11] Brynjolfsson, E., Li, D., & Raymond, L. (2025). Generative AI at work. The Quarterly Journal of Economics, 140(2), 889-942.
[12] Cho, D. S. (1998, 2024). Mechanism: The Fourth Element of Management. aSSIST Business School.
[13] Cronholm, S., Sjöström, J., Göbel, H., & Juell-Skielse, G. (2024). Generalisation of design science research.
[14] Development, O. f. E. C.-o. a. (2019). Recommendation of the Council on Artificial Intelligence.
https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449
[15] Dingel, J., Kleine, A.-K., Cecil, J., Sigl, A. L., Lermer, E., & Gaube, S. (2024). Predictors of health care practitioners' intention to use AI-enabled clinical decision support systems: Meta-analysis based on the unified theory of acceptance and use of technology. Journal of Medical Internet Research, 26, e57224.
[16] DIS, I. (2009). 9241-210: 2010. Ergonomics of human system interaction-Part 210: Human-centred design for interactive systems. International Standardization Organization (ISO). Switzerland, 2.
[17] Dyer, J. H., Cho, D. S., Shinlim-Dong, K.-K., & Chu, W. (1996). Strategic supplier segmentation: A model for managing suppliers in the 21st century. Cambridge: MIT Press.
[18] Gershman, S. J., Zhou, J., & Kommers, C. (2017). Imaginative reinforcement learning: Computational principles and neural mechanisms. Journal of cognitive neuroscience, 29(12), 2103-2113.
[19] Gregor, S., & Hevner, A. R. (2013). Positioning and presenting design science research for maximum impact. MIS quarterly, 337-355.
[20] Guidance, W. (2021). Ethics and governance of artificial intelligence for health. World Health Organization, 1-165.
[21] Hassan, M., Borycki, E. M., & Kushniruk, A. W. (2025). Artificial intelligence governance framework for healthcare. Healthcare Management Forum.
[22] Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. MIS quarterly, 75-105.
[23] Hussein, R., Zink, A., Ramadan, B., Howard, F. M., Hightower, M., Shah, S., & Beaulieu-Jones, B. K. (2024). Advancing Healthcare AI Governance: A Comprehensive Maturity Model Based on Systematic Review. medRxiv, 2024.2012.2030.24319785.
[24] Ibáñez, J. C., & Olmeda, M. V. (2022). Operationalising AI ethics: how are companies bridging the gap between practice and principles? An exploratory study. AI & Society, 37(4), 1663-1687.
[25] Joshi, G., Jain, A., Araveeti, S. R., Adhikari, S., Garg, H., & Bhandari, M. (2024). FDA-approved artificial intelligence and machine learning (AI/ML)-enabled medical devices: an updated landscape. Electronics, 13(3), 498.
[26] Kahneman, D., & Tversky, A. (2013). Prospect theory: An analysis of decision under risk. In Handbook of the fundamentals of financial decision making: Part I (pp. 99-127). World Scientific.
[27] Li, L. Z., Yang, P., Singer, S. J., Pfeffer, J., Mathur, M. B., & Shanafelt, T. (2024). Nurse burnout and patient safety, satisfaction, and quality of care: a systematic review and meta-analysis. JAMA network open, 7(11), e2443059-e2443059.
[28] Moy, S., Irannejad, M., Manning, S. J., Farahani, M., Ahmed, Y., Gao, E., Prabhune, R., Lorenz, S., Mirza, R., & Klinger, C. (2024). Patient perspectives on the use of artificial intelligence in health care: a scoping review. Journal of Patient-Centered Research and Reviews, 11(1), 51.
[29] National Institute of Standards and Technology. (2023). AI Risk Management Framework (AI RMF 1.0). https://www.nist.gov/itl/ai-risk-management-framework
[30] Nielsen, J. (1993). Response times: the three important limits. Usability Engineering.
[31] Parasuraman, A., Zeithaml, V. A., & Berry, L. (1988). SERVQUAL: A multiple-item scale for measuring consumer perceptions of service quality. Journal of Retailing, 64(1), 12-40.
[32] Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. Journal of management information systems, 24(3), 45-77.
[33] Plsek, P. E., & Greenhalgh, T. (2001). The challenge of complexity in health care. BMJ, 323(7313), 625-628.
[34] Poon, E. G., Lemak, C. H., Rojas, J. C., Guptill, J., & Classen, D. (2025). Adoption of artificial intelligence in healthcare: survey of health system priorities, successes, and challenges. Journal of the American medical informatics association, 32(7),

1093-1100.
[35] Pyae, A. (2025). What is Human-Centeredness in Human-Centered AI? Development of Human-Centeredness Framework and AI Practitioners' Perspectives. arXiv preprint arXiv:2502.03293.
[36] Recker, J. (2023). Designing information systems that support environmental sustainability: A framework-based review. Research Handbook on Information Systems and the Environment, 114-148.
[37] Sáez, C., Ferri, P., & García-Gómez, J. M. (2024). Resilient artificial intelligence in health: synthesis and research agenda toward next-generation trustworthy clinical decision support. Journal of Medical Internet Research, 26, e50295.
[38] Schram, A. L., Henriksen, T. B., Maindal, H. T., & Brazil, V. (2025). Navigating complexity: a conceptual framework for simulation interventions. Advances in Simulation, 10(1), 39.
[39] Shneiderman, B. (2022). Human-centered AI. Oxford University Press.
[40] Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction (Vol. 1). MIT press Cambridge.
[41] Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. Nature medicine, 25(1), 44-56.
[42] Tuunanen, T., Winter, R., & Brocke, J. v. (2024). Dealing with complexity in design science research: A methodology using design echelons. MIS quarterly, 48(2), 427-458.
[43] Vergne, J.-P., & Durand, R. (2021). The missing link between the theory and empirics of organizational adaptation. Academy of Management Annals, 15(2).
[44] Wang, L., Zhang, Z., Wang, D., Cao, W., Zhou, X., Zhang, P., Liu, J., Fan, X., & Tian, F. (2023). Human-centered design and evaluation of AI-empowered clinical decision support systems: a systematic review. Frontiers in Computer Science, 5, 1187299.
[45] Wang, Y., Song, Y., Wang, Y., YU, L., & Wang, J. (2024). Ethics and governance of artificial intelligence for health: guidance on large multi-modal models. Chinese Medical Ethics, 1001-1022.
[46] World Health Organization. (2024). Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models. W. H. Organization. https://www.who.int/publications/i/item/9789240084759

Appendix A: Detailed Scenario Design

Scenario 1: No Governance Baseline

Category	Content
Purpose	Baseline setting for AI operation without HCEE gover nance
Governance Condition	HCEE architecture not applied, no experience measure ment system, no trigger-control rules
AI Operation	Autonomous AI operation (minimal human interventio n), efficiency-focused optimization, reactive response to errors

Scenario 2: Partial HCEE (Measurement Only)

Category	Content
Purpose	Isolate and verify measurement effects by introducing only experience measurement
Governance Condition	Experience layer only implemented (real-time PXI/EXI measurement), Policy/Governance/Control layers not ap plied, automatic control rules not operational
Validation Focus	Hawthorne effect - verify if measurement itself induces behavioral change

Scenario 3: Full HCEE Implementation

Category	Content
Purpose	Verify effects of complete 4-layer HCEE architecture i mplementation
Governance Condition	Policy layer: patient dignity priority principle applied; Governance layer: committee structure operational; Con trol layer: trigger-control rules automated; Experience l ayer: real-time PXI/EXI measurement
Mechanism Functions	Coordination: stakeholder interaction optimization; Lear ning: feedback-based continuous improvement; Selectio n: automatic optimal strategy determination by situation

Scenario 4: Stress Test

Category	Content
Purpose	Verify HCEE governance resilience in high-error envir onment

Stress Conditions	2x AI error rate increase (5% -> 10%), patient complaint surge events (weekly), healthcare worker turnover increase (3% monthly), system downtime (2 hours weekly)
Additional Metrics	Recovery Time: time to normalize PXI/EXI; Resilience Index: indicator variance before/after stress; Escalation Frequency: trigger activation count

Appendix B: Detailed Trigger-Control Rules

Rule ID	Trigger Condition	Control Action	Priority
TC-01	PXI < 70	20% AI automation reduction, expanded healthcare worker explanation time	Critical
TC-02	EXI < 65	Expanded AI override authority, strengthened decision rationale	Critical
TC-03	|PXI - EXI| > 15	Balance adjustment mechanism activation, bidirectional feedback strengthening	High
TC-04	AI Error Rate > 8%	AI automation temporary suspension, human oversight strengthening	Critical
TC-05	Dignity Perception < 60	Expanded human face-to-face interaction, minimized AI intervention	Critical
TC-06	Professional Respect < 55	Expanded healthcare worker autonomy, adjusted AI recommendation level	High
TC-07	PXI/EXI Plunge (weekly -10%)	Emergency response team convening, root cause analysis	Emergency
TC-08	PXI > 85 AND EXI > 80	Gradual AI automation level expansion permitted	Normal

Appendix C: Data Sources and Parameter Summary

C.1 FDA AI/ML Medical Device Database

Item	Content
Source	FDA 'AI/ML-Enabled Medical Devices' Database
URL	https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-enabled-medical-devices
Access Date	January 29, 2026
Confirmed Data	Total 1,357 AI/ML medical devices approved (as of September 2025)
Simulation Application	AI system accuracy parameter: 0.85-0.95 uniform distribution

C.2 Simulation Parameter Setting Basis

Parameter	Setting Value	Basis
Patient Agents	500	Conservative estimate of 689 daily outpatients
Healthcare Worker Agents	100	Outpatient care staff only (30 physicians + 70 nurses)
AI Systems	10	CDSS 4, Operations 3, RCM 3 (FDA 76% radiology, AHA 74% RCM)
AI Accuracy	0.85-0.95	FDA approved device meta-analysis (Joshi et al., 2024)
PXI Initial Value	Mean=70, SD=10	SERVQUAL healthcare service quality research (Parasuraman et al., 1988)
EXI Initial Value	Mean=65, SD=8	Healthcare worker burnout research exploratory setting (Li et al., 2024)

Acceptance Probability Coefficients	r=0.66, 0.55, 0.73	UTAUT meta-analysis (Dingel et al., 2024)
Learning Rate (α)	0.1	Reinforcement learning standard range 0.05-0.2 (Sutton & Barto, 2018)
Penalty/Reward Ratio	-15/+5 (3:1)	Loss aversion λ≈2.25 reference (Kahneman & Tversky, 2013)
Agreement Threshold	0.8 / 0.5	Exploratory setting, sensitivity analysis (±0.1) verified
Trigger Thresholds	PXI 70, EXI 65	SERVQUAL satisfaction range reference, sensitivity analysis (±5) verified

Appendix D: Sensitivity Analysis Results

D.1 Exploratory Parameter Sensitivity Analysis

Sensitivity analysis was performed on exploratory parameters with limited literature basis in this simulation. Each parameter was varied within ±20% of base value or specified range, analyzing effects on final PXI and EXI.

Table D-1. Exploratory Parameter Sensitivity Analysis Results

Parameter	Base Value	Variation Range	Result Variation
Learning Rate (α)	0.1	0.05 - 0.2	PXI ±2.3%, EXI ±1.8%
Noise SD (ε)	2	1 - 3	PXI ±1.5%, EXI ±1.2%
AI Error Penalty	-15	-10 ~ -20	PXI ±4.1%, EXI ±2.9%
Recovery Reward	+5	+3 ~ +7	PXI ±1.8%, EXI ±1.4%
Agreement Threshold (Accept)	0.8	0.7 - 0.9	PXI ±3.2%, EXI ±2.5%
Quality Coefficient	1.2	1.0 - 1.4	PXI ±2.7%, EXI ±3.1%
Trigger Threshold (PXI)	70	65 - 75	PXI ±3.8%, EXI ±2.2%
Trigger Threshold (EXI)	65	60 - 70	PXI ±2.0%, EXI ±4.5%

D.2 Sensitivity Analysis Results Interpretation

Sensitivity analysis results confirmed that variation ranges for all exploratory parameters resulted in final indicator variations within ±5%, confirming result robustness. AI error penalty and trigger thresholds showed relatively larger effects, but these are consistent with theoretically expected directions. Learning rate and noise parameters had limited effects on results, confirming appropriateness of settings within general reinforcement learning literature ranges. This sensitivity analysis acknowledges limitations of exploratory simulation while demonstrating that main conclusions (positive effects of HCEE governance) are robust to parameter setting uncertainty.

Author Biography

Minseong Kim

Ph.D. Student, Graduate School of AI, Seoul School of Integrated Sciences & Technologies (aSSIST), Administrative Director, Pogny Hospital
Research Interests: Healthcare AI Governance, Human-Centered Experience Engineering, Hospital Digital Transformation, Generative AI, Data-Driven Hospital Management
E-mail: mskim924@stud.assist.ac.kr

Articles of Association for “AI Business Academy”

Chapter 1 General Provisions

Article 1 (Title) This corporation shall be called the “AI Business Academy” (AIBA for short , hereinafter referred to as “the Academy ”) .

Article 2 (Purpose) This academy aims to promote the development of Korean management studies through the convergence and integration of AI and management . Furthermore, it aims to contribute to the development of future Korean academy and the establishment of future strategies for companies .

Article 3 (Location of Office) The office of this academy shall be located under the Industrial Policy Research Institute (located at 7th floor , 203 Sinchon-ro, Seodaemun-gu, Seoul) , and regional headquarters may be established in other regions . A small number of administrative staff may be employed to operate the office .

Article 4 (Business)

1. In order to achieve the purpose of Article 2, this academy conducts the following business.
(1) Research related to theory and practice for grafting and integrating AI and Business
(2) Conducting research services commissioned by the government or companies at the level of industry-academia-government cooperation
(3) Holding regular academic seminars, policy discussions, and research presentations
(4) Publication of academic journals and research books
(5) Exchange with foreign similar academic societies
(6) Other projects necessary to achieve the purpose of this academy
2. When this academy carries out its business objectives, the benefits it provides to the beneficiaries shall be provided free of charge . However , in unavoidable cases, it may obtain prior approval from the competent authority and have the beneficiaries bear a portion of the cost .
3. The benefits provided through the performance of the objectives of this academy shall not discriminate based on the beneficiary’s gender, place of birth, school attended, place of employment, occupation, or social status .

Chapter 2 Members

Article 5 (Types of Members) Members of this academy are individuals and groups that agree with the purpose of Article 2 , and include regular members (annual members and lifetime members) , associate members , and institutional members .

Article 6 (Regular Member) Regular members are individuals who fall under each of the following categories and have completed the membership process .

1. Full-time faculty members who teach or research AI or business-related subjects at universities , colleges, and graduate schools.
2. or part-time faculty or doctoral program or have obtained a degree in AI or business administration
3. Researchers at accredited research institutes and employees of AI- related companies
4. Other individuals recognized by the Board of Directors as having equivalent qualifications.

Article 7 (Classification of regular members) Regular members are classified into annual members and lifelong members . Annual members are those who have paid the annual membership fee , and the annual membership qualification is one year from the date of registration . Annual members are renewed by paying the annual membership fee . Lifelong members are those who have paid the lifetime membership fee .

Article 8 (Associate Member) Associate members are students enrolled in an undergraduate program in AI or business administration or individuals who have completed the membership process .

Article 9 (Institutional Member) Companies and organizations that agree with the purpose of this academy may become institutional members upon resolution of the board of directors. Membership is in principle for one year, but may be renewed .

Article 10 (Rights and Obligations) Members of this academy have the following rights and obligations .

1. All members can attend the general meeting and participate in discussions , and participate in the academy's business such as research presentations .
2. Regular members have the right to vote and be elected , but institutional members do not have the right to vote or be elected .
3. Members have the obligation to cooperate in the operation of the academy, including paying

membership fees and attending academic conferences .

Article 11 (Suspension and loss of membership) Members of this academy will have their membership suspended or lost in the following cases :

1. If the membership fee is not paid , membership will be suspended and the rights under Article 9, Paragraphs 1 and 2 will be suspended until the membership fee is paid in full .
2. Members who damage the reputation of the Academy by violating the purpose of the Academy or the code of ethics established by the Academy shall lose their membership if the Board of Directors resolves to expel them .
3. The Code of Ethics is separately established by the Board of Directors and approved by the general meeting .

Article 12 (Withdrawal of membership) Members of this academy may withdraw at will .

Chapter 3 Executive Officers

Article 13 (Types of Executive Officers) The Executive officers of this academy shall be a president , vice president , directors and auditors , and the composition of the officers shall be as follows . However , regional headquarters shall be established in major regions at home and abroad , and the head of the regional headquarters shall be the vice president .

1. 1 Chairman
2. About 20 vice presidents
3. The number of directors is 5 to 50 people.
4. 2 auditors

Article 14 (Term of executive officers) The term of executive officers shall be two years, but they may be reappointed .

Article 15 (Appointment and election of executives)

1. The chairman and auditors are elected at the general meeting .
2. The Vice-Chairman and Directors are recommended by the Chairman and appointed by the Chairman with the approval of the Board of Directors . However, the first Vice-Chairman and Directors are appointed by the Chairman and approved by the General Meeting .
3. When an officer other than the chairman is unable to perform his/her duties during his/her term of

office, the board of directors shall elect a replacement , and the term of office of the officer elected through the replacement shall be the remaining term of office of his/her predecessor .

Article 16 (Duties of Executive Officers)

1. The president represents the academy and supervises its business .
2. The director attends the board of directors meeting and deliberates on matters related to the business of this academy , and serves as the board of directors or chairman .
from Handle delegated tasks .
3. The auditor performs the following duties :
(1) Auditing the financial status and business of this academy.
(2) Auditing matters related to the operation of the board of directors
(3) If any irregularities or injustices are discovered in the audit results of (1) and (2), the Board of Directors or Requesting the General Assembly to correct the situation
(4) When necessary for reporting under (3) , requesting the convocation of a general meeting or of directors meeting.

Article 17 (Acting Chairman) When the Chairman is unable to perform his/her duties due to an accident or vacancy, the Vice Chairman designated by the Board of Directors shall act in his/her stead for the remainder of the term .

Article 18 (Dismissal of Executive Officers) If an executive officer loses membership qualifications pursuant to Article 11 or commits any of the following acts, he or she may be dismissed upon resolution of the Board of Directors .
1. Unfair acts such as serious dereliction of duty that run counter to the purpose of this academy.
2. Any act that causes a dispute between executives and is judged to make it impossible for the executives to perform their duties.
3. Any other acts that unfairly interfere with the work of this academy.

Chapter 4 General meeting

Article 19 (Matters to be resolved at the general meeting) The general meeting shall resolve the following matters :
1. Appointment and dismissal of the chairman and auditors
2. Endorsement of the first executive team

3. Changes to the Articles of Academy

4. Approval of settlement and business report

5. Provisions on the rights and obligations of members

6. Voting on matters proposed by the Chairman and Board of Directors

7. Other important matters

Article 20 (Convening of the General Meeting)

1. The general meeting is divided into a regular general meeting and a special general meeting .

2. The regular general meeting is held in February every year , and the extraordinary general meeting is convened by the chairman in the following cases :

(1) When the Chairman deems it necessary

(2) When there is a resolution by the board of directors

(3) When more than one-fifth of the members request in writing, stating the reason for holding the meeting.

(4) When the auditor requests a meeting pursuant to the provisions of Article 18, Paragraph 4.

Article 21 (Quorum for General Meeting Resolutions)

1. The general meeting shall be opened with members present , and attendance may be delegated by scanning an electronic document of a written or signed document.

2. The resolutions of the general meeting shall be decided by a majority vote of the members present, excluding those present in writing . However , In case of a tie, the chairman decides .

3. If the Steering Committee determines that it is difficult to convene an offline general meeting, an online general meeting may be held instead.

Article 22 (Reasons for exclusion from general meeting resolution) If the chairman or member falls under any of the following, he or she may not participate in the resolution .

1. Matters concerning oneself in the appointment and dismissal of executives

2. Matters involving the receipt of money or property, and in which the interests of the chairperson or member conflict with those of the academy.

Chapter 5 Board of Directors

Article 23 (Matters to be resolved by the Board of Directors) The Board of Directors resolves and agrees on the following matters :

1. Convening of a temporary general meeting
2. Resolution of agenda to be submitted to the general meeting
3. Matters delegated by the general meeting
4. Approval of executives appointed by the Chairman
5. Approval of business plan and budget
6. Approval and amendment of the composition and operating regulations of various committees.
7. Important changes to the journal's editorial policy
8. Matters concerning investment and fund management
9. Resolution on membership of institutional members
10. Resolution on expulsion of members
11. Determination of membership fees
12. Other Matters

Article 24 (Delegation of Board of Directors' Resolutions) Routine matters among the Board of Directors' resolutions may be delegated to the Management Committee through a Board of Directors resolution .

Article 25 (Board of Directors) The Board of Directors is composed of a chairman , vice chairman , directors , and auditors .

Article 26 (Convening of the Board of Directors) The Board of Directors shall be convened by the Chairman, who shall also serve as its chairman, in the following cases :
1. When the Chairman deems it necessary
2. When more than half of the members of the board of directors, excluding the chairman, request a meeting by stating the purpose of the meeting.

Article 27 (Quorum for Board of Directors Resolution)
1. The board of directors shall open with the attendance of a majority of its members, and attendance may be delegated in writing .
2. The Board of Directors' resolutions are decided by a majority vote of the members present, excluding those present in writing . However , In case of a tie, the chairman decides .
3. If the Management Committee determines that it is impossible to convene a Board of Directors meeting due to scheduling issues that require a Board of Directors resolution,
If a decision is made, the vote can be made online or in writing .

Article 28 (Reasons for exclusion from board of directors' resolution) A board of directors member may not participate in the resolution if any of the following applies .

1. Matters concerning oneself in deciding on the loss of membership qualifications
2. Matters involving the receipt of money or property and in which the interests of the member and the academy conflict.

Chapter 6 Various Committees

Article 29 (Management Committee) An Management Committee is established to ensure smooth decision-making and efficient execution of the Academy's business .

1. The Management Committee deliberates and decides on matters delegated by the Board of Directors in accordance with Article 24 of these Articles of Incorporation .
2. The Steering Committee is composed of the Chairman, Vice Chairman , Directors , and Secretariat members , and is composed of 10 or so members.

Appoint one person as chairman .

3. The operating committee chairman convenes the operating committee as needed .
4. The Steering Committee shall open with the attendance of a majority of its members , and resolutions shall be passed with the approval of a majority of the members present.

Article 30 (Editorial Committee) An editorial committee shall be established to publish the journal of this academy , and regulations regarding the editorial committee shall be separately determined by resolution of the board of directors .

Article 31 (Fund Management Committee) A Fund Management Committee shall be established to manage the funds of this Academy , and the regulations regarding this shall be separately determined by resolution of the Board of Directors .

Article 32 (Report on Committee Activities) The chairperson of each committee shall report the results of its activities to the board of directors .

Article 33 (Secretariat) A secretariat may be established to manage the affairs of this academy , and the regulations of the secretariat shall be established through a resolution of the board of directors .

Chapter 7 Assets and Accounting

Article 34 (Composition of assets)

The assets of this academy consist of the following items :

1. Donation
2. Membership fee
3. Income from basic assets
4. Other income

Article 35 (Types of Assets)

1. The assets of this academy are divided into two types: basic assets and general assets .
2. The basic assets listed in each clause below cannot be disposed of or provided as collateral . However , when there is a reasonable cause, a portion of them may be disposed of or provided as collateral with the approval of the Board of Directors and the authorization of the Minister of Strategy and Finance .

(1) Property contributed and designated as basic property

(2) Property that the board of directors has decided to use as basic property

3. Ordinary assets consist of assets other than basic assets .

Article 36 (Assessment and collection of membership fees) The method of levying and collecting membership fees shall be determined by the Board of Directors .

Article 37 (Expenses)

The expenses of this academy are paid from general assets .

Article 38 (Management of Assets)

1. The assets of this academy shall be managed by the chairman , and the method of management shall be determined by the board of directors .
2. This academy must open an Internet homepage and disclose the annual amount of donations raised and the results of their use through it .

Article 39 (Disposal of Surplus)

If there is a surplus at the end of the fiscal year, all or part of it is incorporated into capital or carried over to the next fiscal year, as determined by the Board of Directors .

Article 40 (Approval of Budget and Settlement of Accounts)

1. The budget for each fiscal year of this academy must be resolved by the Board of Directors and approved by the General meeting prior to the start of the fiscal year .

2. The chairman has the authority to execute the budget and business accounting .

3. The Chairman shall prepare a general accounting settlement report and a fund management settlement report within two months after the end of the fiscal year, attach an audit opinion, and report to the general meeting for approval .

Article 41 (Executive Compensation) As a rule , no compensation is paid to executives .

Article 42 (Fiscal Year)

fiscal year of this academy runs from September 1 to August 31 of the following year .

Chapter 8 Amendment of Articles of Academy and Dissolution

Article 43 (Amendment of Articles of Academy)

These Articles of Association may be amended with the approval of the Board of directors by a majority vote of at least three- quarters of the members present .

Article 44 (Dissolution)

This academy may be dissolved with the consent of more than two -thirds of its members .

"AI Business Academy" Research Ethics Regulations

Chapter 1 General Provisions

Article 1 (Purpose) The purpose of these research ethics regulations (hereinafter referred to as " these regulations ") is to present basic principles and directions for the members of the "AI Business Academy" (hereinafter referred to as the " the Academy ") to promote research ethics and truthfulness, create a sound research atmosphere , and contribute to academic and social development .

Article 2 (Scope and Scope)

1. These regulations apply to all members of the academy and stakeholders directly or indirectly related to research activities .
2. Except in cases where there are other special provisions related to the establishment of research ethics and verification of research authenticity, these regulations shall apply .

Article 3 (Scope of misconduct) The misconduct set forth in these regulations includes the following acts that may damage the reputation of not only individuals but also the academy as a member of the academy .

1. Forgery : Here, " Forgery " refers to the act of falsely creating non-existent data or research results .
2. Modulation : Here, " Modulation " refers to an act of distorting research content or results by intentionally changing or deleting research data .
3. Plagiarism : Here, " plagiarism " refers to the act of stealing another person's ideas , research content , research results, etc. without clearly indicating the source .
4. Double publication of papers : Here, " Double publication of papers" refers to the act of publishing the same data, analysis methods, and analysis results in two or more academic journals .
5. Unfair authorship of a paper : Here, " unfair authorship " refers to an act of not granting authorship of a paper to a person who has made scientific or technical contributions or contributions to the research content or results without just cause , or granting authorship of a paper to a person who has not contributed as an expression of gratitude or courtesy .
6. Duplicate research : Here, " Duplicate research " refers to an act of conducting two or more research projects with the same content and publishing the same research results .
7. Public False Statement : Here, " Public false statement " refers to an act of making false statements about one's academic background , career , qualifications , research achievements , results , etc.

8. Refers to an act of intentionally obstructing an investigation into suspicions of misconduct by oneself or another person or causing harm to the whistleblower .
9. Refers to any act that seriously deviates from the scope normally tolerated by the Academy of Information Systems in relation to other research .

Chapter 2 Basic Ethics for Researchers

Article 4 (Researcher's Morality) Researchers must conduct research in an honest and correct manner, not in a socially acceptable or ethically wrong manner, and present the results . Accordingly, the ethics that researchers must observe in relation to research are as follows . 1. Using others' ideas, research content, or research results (hereinafter " research content ") without clearly indicating them is clear plagiarism , and researchers must indicate the source as follows :
(1) When citing someone else's research content without re-editing (direct quotation) , the author's name , year , and page number must be indicated in the text with quotation marks ("") , and the complete source must be indicated in the references section .
(2) When editing and using the research content of others, the author's name and year must be indicated in the text , and the full source must be indicated in the references section .
(3) When requoting content quoted by others, you must indicate " requotation " along with the quoter's name and year .
2. When conducting research, researchers must recognize acts such as falsification or alteration of data as serious crimes and make efforts to prevent such misconduct from occurring .
3. It must not duplicate the previously published (or under review) research papers as if they were new research . Therefore, papers published in academic journals must be original work .
4. Those who have contributed directly or indirectly to the research must be listed as co-authors in the research paper , and those who have not contributed cannot be listed as authors . In addition, when conducting joint research , the person who has contributed the most to the research is the first author , and the corresponding author is limited to the person who is in charge of all correspondence between the authors and the academic academy and who has overseen the research or an equivalent person . However , in cases where co-authors cannot be listed, such as in cases of administrative , financial, or research support, this should be indicated in a footnote or acknowledgement .
5. It must truthfully write about their academic background , career , qualifications , research achievements , results , etc.
6. In relation to other research, one must not engage in any conduct that goes beyond the scope generally accepted in academia .

Article 5 (Social Responsibility) Researchers must be deeply aware of the impact their research will have on academy and, as professionals, must fulfill their responsibilities and obligations regarding the results of their research .

Article 6 (Respect for Others) Researchers must not harm or infringe upon the rights of others and must respect intellectual property rights . In addition , when dealing with others in relation to research, they must be treated equally regardless of race , gender , nationality , disability , etc.

Chapter 3 Roles and Responsibilities of the Research Ethics Committee

Article 7 (Composition and term of office of the committee)

1. The Research Ethics Committee (hereinafter referred to as the " Committee ") shall be composed of 10 or fewer members, including the chairperson .
2. The chairman of the committee concurrently serves as the editor-in-chief of this academy .
3. The committee member concurrently serves as an editorial committee member of this academy .
4. The term of office for the chairperson and members shall be one year, but they may be reappointed .
5. The secretary of the committee shall be one of the vice-presidents of this academy , and the secretary shall not have voting rights .

Article 8 (Functions of the Committee) The Committee deliberates on the following matters :

1. Matters concerning fraudulent acts such as forgery , falsification , and plagiarism related to academic papers
2. Matters concerning research ethics raised regarding research plans , research results, etc. related to the academic academy
3. Investigation into misconduct related to the academic academy
4. Complaints raised regarding the truthfulness of research related to academic societies or journals
5. Matters concerning research ethics and education in research and development
6. Matters concerning the processing and follow-up measures of the results of research integrity verification
7. Matters to prevent misconduct that may occur during research
8. Other matters regarding research ethics raised by the president or committee chairperson

Article 9 (Convening , review , and resolution of the committee)

1. When a report of research misconduct is received, the chairperson convenes a committee to investigate and deliberate .
2. The committee is established with the attendance of a majority of its members , and resolutions are passed with the approval of a majority of the members present .
3. When the chairperson deems the agenda for deliberation to be minor, he/she may replace it with a written deliberation .
4. During the investigation process, the committee may request the attendance of the informant , the person under investigation , witnesses , and reference persons when necessary .

Article 10 (Duties and protection of rights of whistleblower)

1. " Whistleblower " refers to a person who has reported to this committee facts or related evidence of an act of misconduct .
2. The informant may report in any possible way , including verbally , in writing , by phone , or via e - mail . In principle, reports must be made under one's real name . However , even if the report is anonymous, if the report includes the name of the research project , the title of the paper , and specific details and evidence of the misconduct in writing or via e-mail, and if it is found to be true, it will be processed in the same way as a report made under one's real name .
3. Under no circumstances will the identity of the informant be directly or indirectly exposed , and the informant's name will not be included in the investigation results report for the purpose of protecting the informant unless absolutely necessary .
4. A whistleblower who reported something even though he or she knew or could have known that the information was false is not eligible for protection .
5. The informant has an obligation to respond to the committee's request to appear .

Article 11 (Duties and protection of rights of the subject of investigation)

1. " Subject of investigation " refers to a person who has become the subject of an investigation into an unethical act through a report or knowledge thereof, or a person who has become the subject of an investigation due to being presumed to have participated in an unethical act during the course of the investigation . Reference persons or witnesses during the course of the investigation are not included .
2. The person under investigation has an obligation to faithfully respond to any requests made by the committee during the investigation process . In addition, the person under investigation is guaranteed the right and opportunity to express opinions , raise objections, and defend himself/herself during the investigation process .
3. The committee must take care to ensure that the reputation or rights of the person under

investigation are not violated until the verification of whether or not there was any misconduct is complete , and must endeavor to restore the reputation of the person under investigation who is found not to have been involved in any misconduct .
4. The subject of the investigation must faithfully comply with the request for submission of evidence requested by the committee .

Article 12 (Confidentiality) All matters related to the investigation, including reporting , investigation , deliberation , resolution, and recommendation measures , shall be kept confidential . Persons directly or indirectly participating in the investigation and related committee members shall not disclose any information obtained during the investigation and performance of duties . If the need for reasonable disclosure arises, it may be disclosed after a resolution by the committee .

Article 13 (Notification and Objection)

1. The results of the judgment on misconduct must be notified to the reporter and the person under investigation .
2. If there is a reasonable basis for an objection filed by the reporter or the subject of the investigation against the results of the decision, the objection may be filed within 30 days from the date of notification , and a reinvestigation may be conducted if necessary .

Chapter 4 Research Ethics Committee Verification Procedures and Standards

Article 14 (Preliminary Investigation)

1. A preliminary investigation refers to a procedure to determine whether or not there is a need to investigate a suspicion of misconduct , and must be initiated within 30 days from the date of receipt of the report . The preliminary investigation is decided by a preliminary investigation committee consisting of a chairperson , vice chairperson , two members , and a secretary .
2. If the subject of the investigation admits to all facts of the misconduct as a result of the preliminary investigation , a decision can be made immediately without going through the main investigation procedure , and if it is determined that there is a possibility of significant damage to the evidence, the chairperson can take measures to preserve the evidence prior to the formation of the committee .
3. If it is decided not to conduct a main investigation in the preliminary investigation , the specific reason for this will be notified to the reporter within 10 days from the date of decision . However , no notification will be given in the case of anonymous reports .

Article 15 (Main Investigation)

1. “ Main investigation ” refers to the procedure to verify the facts of an act of misconduct , and must be conducted by forming a committee in accordance with the provisions of Article 7, Paragraph 1 .

2. The entire investigation schedule from the start to the decision must be completed within 6 months . However , if it is determined that the investigation cannot be conducted within this period, the reason for this may be notified to the president of the academic academy and the investigation period may be extended .

3. Main investigation If it is determined that fair progress is difficult due to a conflict of interest between a member of the committee and the subject of the investigation, the president of the academy may restrict the participation of the member in the relevant matter . However , the chairperson and vice chairperson are exceptions .

4. If the informant or the person under investigation raises a legitimate objection regarding the avoidance of a specific member, the committee may decide to accept it .

Article 16 (Recording and reporting of investigation results)

1. Records related to misconduct must be kept in document form at the academy for 5 years .

2. The committee must report the results , contents , and decisions of the investigation to the president of the academy and the board of directors .

3. The record of the judgment contents includes the following items .

(1) Report contents

(2) Allegations of misconduct and related papers or reports that are the subject of the investigation

(3) Objections or arguments of the informant and the subject of the investigation

(4) Judgment results and related supporting materials and witnesses

(5) List of judges

Article 17 (Post - judgment measures)

1. After the committee completes the review, it decides on the type of disciplinary action .

2. The types of disciplinary action are as follows depending on the severity of the misconduct , but may be applied multiple times .

(1) Expulsion

(2) Suspension of membership

(3) Restrictions on publication in academic journals (AI Management Journal)

(4) Official apology at the conference

(5) Notification of misconduct regarding related papers in academic journals or newsletters

3. If the investigation results confirm that there was no misconduct, the committee may take appropriate follow-up measures to restore the reputation of the person under investigation .

Article 18 (Administrative Matters)

1. Matters not specified in these regulations shall be decided and implemented by the committee in accordance with social norms .

2. The revision of the research ethics regulations shall follow the revision procedures of this academy .

3. For the smooth operation and activities of the committee, the academy Necessary administrative and financial support must be provided .

AI Journal of Business

Call for Papers

The AI Business Academy, a leader in the AI era, is recruiting papers to be published in Volume 5, in November. 2026 AI Journal of Business.

Now, all companies and businesses cannot grow and develop without utilizing AI and big data. In addition to the existing AI models representing artificial intelligence. The AI Business Academy would like to invite excellent papers that present various cases of utilizing AI and big data to innovate business and create new opportunities, as well as AI models that are utilized and models of AI strategies that serve as the basis.

We hope for active participation from many people.

1. Recruitment areas: Various management strategies and business innovations based on AI and big data, engineering improvements, and creative fields that combine management and engineering.
2. Recruitment deadline: September 31, 2026
3. Journal publication schedule: Reviewed by the end of October. 2026, then published in November.
4. Thesis Format: **Principle of submission as English version of paper** . It is much more advantageous and meaningful to write a paper in English in order to share and attract more interest from AI researchers around the world and to be cited in foreign papers. If you request it to the contact information below, a paper format file will be distributed.
5. Submission : Academy website or editorial committee email: jhchang@assist.ac.kr
6. Contact: AI Business Academy Office (**aiba2023@naver.com)** or jhchang@assist.ac.kr

AI Journal of Business

Vol. 2, No. 2

Issued on May, 2026

Publisher : Jungho, Pyo

Editor : Joongho Chang

Publisher : AI Business Academy

7 floor , 203 Sinchon-ro, Seodaemun-gu, Seoul, Korea

Tel) **02-360-0775**

ISSN 3058-6992