

# Modesty of Interpretability and Arrogance of Predictive Power: A Comparative Study of Statistical Models versus Machine and Deep Learning in Trade Forecasting

OhSuk Yang\* and SeongHun Kim

Department of Business Administration, Kangwon National University  
Gangwondaehakgil 1, Chuncheon-si, Gangwon State, Republic of Korea

(Received November 22, 2025; accepted September 14, 2026)

**Keywords:** machine learning, deep learning, simple model, complex model, predictive power

We comparatively examined the relative superiority of traditional general statistical models, i.e., simple models, and machine learning (ML) and deep learning (DL) models, i.e., complex models, in the context of trade forecasting. Pooled ordinary least squares (OLS) was adopted as the general statistical model. The random forest (RF) and Transformer models were selected as the representative ML and DL models, respectively. Furthermore, a robustness check was performed using the panel regression and long short-term memory (LSTM) models. We utilized country-year panel data spanning from 2000 to 2023. The key findings derived from the analysis are as follows. First, the performance of the ML and DL models was found to be superior, suggesting improved nonlinear structure learning and the ability to capture interactions between variables. Second, the DL model exhibited higher metrics than the ML model, indicating that the learning of long- and short-term dependencies utilizing self-attention effectively captured the complex interconnectedness of trade and macroeconomic variables. Third, the out-of-sample  $R^2$  values for both the RF and Transformer models appeared negative. This result serves as evidence to raise skepticism regarding the claim of absolute superiority of complex models. Therefore, it is advisable for researchers to select an appropriate analysis model on the basis of the data's characteristics and the research objective, navigating the space between the humility of interpretability and the arrogance of predictive power.

## 1. Introduction

Recent social science research has expanded from a traditional focus on causal inference toward prediction and decision-making, leading to a rapid increase in the use of machine learning (ML) and deep learning (DL) techniques that capture nonlinear relationships. These techniques—often categorized as “complex models”<sup>(1,2)</sup>—efficiently handle high-dimensional data and nonlinearities, offering an alternative that overcomes the model misspecification problems inherent in conventional statistical models, which are conceptualized as “simple models”.<sup>(3)</sup> In particular, within economics, attempts to analyze complex interactions among

\*Corresponding author: e-mail: [osyang30@kangwon.ac.kr](mailto:osyang30@kangwon.ac.kr)  
<https://doi.org/10.18494/SAM6222>

macroeconomic indicators—interactions that linear models such as ordinary least squares (OLS) or ARIMA fail to capture—through models such as random forest (RF) and support vector machine (SVM) have become increasingly common.<sup>(4)</sup>

In business and finance research, where predictive accuracy is critical—such as customer churn prediction, asset price forecasting, and credit risk assessment—complex models have rapidly established themselves as standard analytical tools.<sup>(5–9)</sup> The same trend is evident in the trade research domain, in which the economic, policy, and institutional dimensions underlying cross-border flows of goods, services, capital, and technology are analyzed. Whereas trade studies have traditionally focused on identifying causal determinants of trade and interpreting coefficients—primarily using methods such as pooled OLS and panel regression—recent work has shifted toward maximizing predictive accuracy using ML and DL models, and toward extracting nonlinear patterns from complex data structures through data-driven forecasting.<sup>(10–13)</sup>

In modern global trade systems, macroeconomic trade indicators are not merely abstract statistical constructs but are increasingly derived from large-scale physical data collection infrastructures. Real-time trade-related data are generated through IoT sensors, RFID-based tracking systems, and automated port monitoring technologies. These physical sensing systems continuously capture logistics flows, cargo movements, and port activities, which are subsequently aggregated into higher-level macroeconomic indicators such as trade volumes, commodity indices, and exchange rate dynamics. From this perspective, trade forecasting models do not operate on purely economic variables but rather on transformed representations of sensor-generated data. Therefore, linking physical data acquisition mechanisms with high-level predictive models such as Transformer and long short-term memory (LSTM) is essential for understanding the full data-generating process underlying trade forecasting.

Against this backdrop, the key puzzle addressed in this article is “Does the predictive power of complex models consistently dominate the explanatory power of simple models? Specifically for trade research, do complex models offer theoretical and practical value as analytical tools?” The author’s answer to this question is partially negative. Although the RF and Transformer models—adopted as representative ML and DL substitutes—demonstrated substantially improved root mean square error (*RMSE*) and mean square error (*MSE*) performance characteristics in export forecasting compared with linear models, and although they reduced bias by automatically capturing nonlinear and interaction effects, their out-of-sample coefficient of determination (OOS  $R^2$ ) values turned out to be negative. This indicates that these models excessively increased variance, thereby elevating statistical risk. Nevertheless, the relative comparison revealed a clear hierarchy of performance: Transformer > RF > pooled OLS. This suggests that in forecasting trade and macroeconomic indicators—characterized by high volatility and strong nonlinearity—DL and ML are inherently more suitable tools than linear models.

The primary purpose of this article is to examine whether complex models such as ML and DL exhibit relative superiority over simple models on the basis of traditional statistical analysis in terms of explanatory and predictive power. A complementary purpose is to investigate the methodological advantages inherent to ML and DL as members of the complex-model family.

This article is structured as follows. Following the introduction in Sect. 1, an outline of the theoretical background and methodology is given in Sect. 2. Section 3 comprises a summary of

the empirical results comparing various analytical models and robustness checks. In Sect. 4, we discuss the theoretical and practical implications derived from the empirical findings, and in Sect. 5, we conclude by presenting the study's limitations and suggesting avenues for future research.

## 2. Methods

In this study, we adopt two methodological approaches to compare the superiority of simple models with that of complex models. First, to identify the model that most effectively satisfies the two conflicting objectives of statistical inference and predictive accuracy, representative analytical models from general statistical methods, namely, ML and DL, were selected, and an empirical analysis was conducted using an empirical trade dataset to evaluate model performance. In addition, as another approach, the Diebold–Mariano test was employed to statistically verify whether the observed differences in predictive power among competing analytical models represent meaningful differences.

### 2.1 Comparison of statistical inference and predictive accuracy

#### 2.1.1 Analytical models

This study is aimed at comparatively analyzing traditional statistical models and recently highlighted ML and DL methods in terms of trade data predictive performance. For this purpose, three major models representing each methodological domain were selected. First, as a general statistical model, pooled OLS, a linear model regarded as a core statistical benchmark model that treats observations across multiple countries and time periods as one integrated dataset, was adopted.<sup>(14)</sup> In addition, to overcome methodological limitations inherent in the pooled OLS model, panel regression—a standard methodology that handles heterogeneity in cross-sectional and time-series data—was used as a robustness check. In this study, pooled OLS, the most basic linear model, is adopted as the core statistical benchmark model for predicting trade data. OLS refers to the most fundamental statistical method for estimating linear relationships, and in pooled OLS, this OLS principle is applied to the panel data of this study. Pooled OLS estimates the model by treating observations across countries and time periods as a single integrated sample. This approach has been criticized for being statistically less efficient than traditional panel regression models because it cannot control the country-specific unobserved heterogeneity (e.g., geographic characteristics) that does not vary over time.<sup>(14)</sup> However, because the objective of this study is to compare predictive performance characteristics, the simplicity of the model and the interpretability of the results offered by pooled OLS become important. Pooled OLS provides the simplest and most transparent linear benchmark for evaluating the predictions of complex models such as RF and Transformer, thereby playing an essential role in clearly verifying the additional predictive superiority of nonlinear models.

Second, as the ML model, RF, a tree-based ensemble learning technique known for strong predictive performance and its ability to capture nonlinear relationships, was selected. RF is an

ensemble method proposed by Breiman<sup>(15)</sup> that combines multiple decision trees to reduce overfitting risk and provide stable predictions. In the ML field, gradient boosting methods (GBMs) such as XGBoost are major competing models that often show slightly higher predictive accuracy by sequentially improving the errors of previous trees through boosting. However, RF is adopted in this study because of two clear academic advantages. First, RF has a bagging-based parallel processing structure, offering superior training speed and computational efficiency compared with GBM, which relies on sequential learning. Second, RF demonstrates strong robustness to overfitting. Because RF reduces variance by randomly selecting data and feature subsets,<sup>(15)</sup> it is advantageous for this study's purpose of comparing multiple models while ensuring model stability and generalization performance. Thus, RF is the most suitable ML benchmark for overcoming the linear limitations of pooled OLS and providing a stable and interpretable comparison with complex DL models.

Third, as the DL model, the Transformer model was adopted instead of the LSTM model, which was originally developed to overcome the limitations of recurrent neural networks (RNNs), to effectively learn the time-series and panel characteristics of trade data. Although LSTM alleviates the long-term dependency problem through its gate structure, it fundamentally relies on sequential processing, which limits computational parallelization. Transformer overcomes these limitations by completely removing recurrence and relying solely on the attention mechanism.<sup>(16)</sup> The core self-attention mechanism captures relationships among all elements of the input sequence simultaneously, regardless of distance, enabling parallel processing and significantly improving training speed and efficiency on large datasets relative to LSTM. Transformer models are particularly strong in learning long-term dependency structures and consistently outperform traditional RNNs (LSTM/RNN) in forecasting trade and macroeconomic indicators characterized by frequent nonlinear shocks.<sup>(17,18)</sup>

Additionally, considering competition among DL models, robustness checks were conducted using the LSTM network, which excels in addressing long-term dependency issues in time-series data. The LSTM network, developed by Hochreiter and Schmidhuber<sup>(19)</sup> to overcome the limitations of conventional RNNs, can learn complex sequence patterns. Through performance comparison among the aforementioned models, in this study, we seek to identify the model that most effectively satisfies the two contrasting objectives of statistical inference and predictive accuracy in trade forecasting.

### 2.1.2 Evaluation metrics

Several major evaluation metrics are employed to compare the predictive performance characteristics of the three models—pooled OLS, RF, and Transformer—in forecasting trade data. Because no single metric can fully capture model performance, three metrics were selected by considering prediction error magnitude, sensitivity to outliers, and interpretability.<sup>(20)</sup> First, *RMSE* is used. *RMSE* squares prediction errors before averaging, thereby imposing a larger penalty on large errors. Thus, it is particularly sensitive to large deviations and is useful when capturing substantial fluctuations in trade forecasts is important. Second, mean absolute error (*MAE*) is used. *MAE* averages the absolute values of errors, responding linearly to error

magnitude. Unlike *RMSE*, it is less sensitive to extreme outliers and provides an intuitive measure of the average size of prediction errors. Third, mean absolute percentage error (*MAPE*) is used. *MAPE* converts errors into percentages relative to actual values, enabling relative comparison across models. It is useful when comparing predictions on different scales and facilitates easy interpretation.

A comprehensive analysis of *RMSE*, *MAE*, and *MAPE* was conducted to evaluate the explanatory power and predictive capability of competing models. Moreover, the OOS  $R^2$  was adopted as an additional performance metric.  $R^2$  traditionally measures how much of the total variation in dependent variable is explained by the model. Ranging from 0 to 1, values closer to 1 indicate higher explanatory power. OOS  $R^2$ , applied to data not used for model training, plays a key role in evaluating overfitting and generalization performance. While in-sample  $R^2$  is always nonnegative, OOS  $R^2$  can take negative values. A negative value indicates that the model performs worse on new data than simply predicting the mean. This suggests that the model failed to capture new data patterns or was overfitted to noise during training. If a model's OOS  $R^2$  is negative, its statistical explanatory power and stability become unreliable. For example, if ML or DL models produce negative OOS  $R^2$  values, it means they do not surpass simple models in terms of statistical stability and explanatory power, serving as a warning against the “arrogance of prediction”. OOS  $R^2$  represents “statistical stability”, while error metrics such as *RMSE* and *MAE* represent “predictive accuracy”. Therefore, comparing these metrics simultaneously is essential for the multidimensional evaluation of model superiority.

## 2.2 Statistical verification of predictive differences: Diebold–Mariano test

The Diebold–Mariano (DM) test is a widely used methodology for evaluating whether the predictive accuracies of two competing forecasting techniques differ statistically. This test is mainly used in time-series analysis and econometrics, and compares the differences in prediction errors between competing models under a given loss function.<sup>(21)</sup> The test examines whether the mean of the loss differential series  $d_t$  between the prediction errors of two models equals zero. The loss differential  $d_t$  is defined as

$$d_t = L(e_{1t}) - L(e_{2t}), \quad (1)$$

where  $e_{1t}$  is the forecasting error of model  $i$  at time  $t$  ( $y_t - \hat{y}_{it}$ ) and  $L(\cdot)$  is a loss function that measures the cost of predictive performance, such as squared error ( $e^2$ ) or absolute error ( $|e|$ ). The null and alternative hypotheses of the DM test are as follows. The null hypothesis ( $H_0$ ) states that “the expected losses of the two forecasts are equal”. That is, there is no statistically significant difference in predictive accuracy ( $H_0: E[d_t] = 0$ ). In contrast, the alternative hypothesis ( $H_A$ ) states that “the expected losses of the two forecasts differ”, meaning that there exists a statistically significant difference in predictive accuracy ( $H_A: E[d_t] \neq 0$ ).

Under a sufficiently large sample size  $T$ , if the null hypothesis  $H_0: E[d_t] = 0$  is true, the DM statistic asymptotically follows the standard normal distribution  $N(0, 1)$ .<sup>(21)</sup> If the absolute value of the computed DM statistic exceeds the critical value ( $Z_{\alpha/2}$ ) or the p-value is less than the

chosen significance level (e.g., 0.05), the  $H_0$  is rejected, and it is concluded that the predictive accuracies of the two forecasts differ statistically. Because the DM test relies on asymptotic properties, it tends to excessively reject the null hypothesis when the sample size is small. To address this issue, Harvey *et al.*<sup>(22)</sup> proposed a modified DM statistic that improves finite-sample properties by using the t-distribution instead of the standard normal distribution.

### 3. Data Collection and Dataset Structure

The dataset used in this study comprehensively includes major elements that explain trade volumes—exports, macroeconomic indicators, exchange rates, commodity prices, and financial instability—and is constructed such that panel model, ML, and DL approaches can be evaluated on an identical data basis. This enables the robust empirical comparison of predictive differences between traditional statistical models and complex models (ML/DL).

Specifically, for the period from 1998 to 2024, a country-year panel dataset integrating annual export data and macro-financial indicators of major trading partners was constructed. The dataset was built by combining export values with the BIS effective exchange rate indices [nominal effective exchange rate (*NEER*), real effective exchange rate (*REER*)], World Bank commodity price indices, global financial instability (*VIX*), international oil prices (*oil\_prices*), and each country's macroeconomic indicators (gross domestic product (*GDP*), population (*POP*), and consumer price index (*CPI*)). The final dataset follows a panel structure with country and year as dimensions and exhibits an unbalanced panel owing to missing observations for certain country-year combinations. Although the dataset used in this study is constructed from macroeconomic and financial indicators, it is important to recognize that these variables are ultimately aggregated outcomes of underlying physical sensing systems in global logistics networks. Trade volumes, commodity price indices, and exchange rate indicators are indirectly shaped by real-time data captures through IoT-enabled monitoring systems, RFID tracking, and automated port infrastructures.

The dependent variable of this study is  $y\_growth$ , defined as the logarithmic export growth rate for the following year. Using growth rates instead of level changes alleviates heteroskedasticity arising from differences in country size and units. Additionally, export levels and their one-year lag were included to reflect trend and autoregressive structure.

Explanatory variables are classified into four categories. First, macroeconomic variables include terms of trade (*TOT*), the natural logarithm of GDP ( $\ln\_gdp$ ), *POP*, and *CPI*. *GDP* represents economic size, *CPI* reflects real purchasing power and inflationary environment, and population proxies economic activity and potential export scale. Second, exchange rate variables include *NEER* and *REER* from BIS, which capture price competitiveness and changing trade conditions—widely used determinants of export fluctuations. Third, seven World Bank commodity price indices (total, energy, non-energy, agriculture, fertilizers, metals and minerals, and precious metals) were used to reflect structural commodity market factors that affect production costs and global demand conditions. Fourth, global financial variables include the *VIX* and *oil\_price*. The *VIX* captures global risk-off sentiment, while oil prices directly affect production costs and logistics expenses, thus affecting export competitiveness.

To reflect lag structure, a one-year lag of export growth ( $y\_growth\_lag1$ ) was included—an important variable for capturing short-term autocorrelation and inertia, used in both statistical models (OLS/panel regression) and ML (RF). For DL models (Transformer and LSTM), instead of conventional regression inputs, time-series characteristics were incorporated by constructing input sequences using the most recent three years of major variables (sequence length = 3). Each sample thus consists of  $(x_{t-2}, x_{t-1}, x_t)$  and the model predicts  $y_{t+1}$ . This structure is advantageous for learning temporal patterns such as trends, cycles, and shock persistence that cannot be captured using static single-period information.<sup>(16)</sup>

### 3. Results

#### 3.1 Basic statistics

Table 1 presents the descriptive statistics of the key variables. The analytical sample consists of 240 country–year observations from 2000 to 2023. The average export volume ( $ln\_exports$ ) is approximately USD 924.5 billion, while the average gross domestic product is around USD 4.46 trillion.  $POP$  exhibits substantial heterogeneity, with an average of 217 million, ranging from small economies with populations of around 19 million to mega-populations exceeding 1.4 billion. The  $VIX$ , which reflects the level of global financial uncertainty, records an average of 19.3, while the  $oil\_price$  averages USD 60.3 per barrel. These values reflect the characteristics of the sample period, which includes major global events such as the 2000s commodity supercycle, the global financial crisis, and the COVID-19 pandemic. The World Bank

Table 1  
Descriptive statistics.

| Variables                    | <i>N</i> (observation) | Mean      | Standard deviation | Min      | Max        |
|------------------------------|------------------------|-----------|--------------------|----------|------------|
| <i>TOT</i>                   | 240                    | 102.014   | 16.553             | 67.478   | 181        |
| <i>POP</i>                   | 240                    | 216509800 | 386853800          | 19028802 | 1412360000 |
| <i>VIX</i>                   | 240                    | 19.278    | 7.164              | 11.054   | 40.001     |
| <i>oil_price</i>             | 240                    | 60.297    | 25.814             | 18.765   | 111.672    |
| <i>ln_exports</i>            | 240                    | 26.38     | 0.796              | 25.114   | 28.942     |
| <i>ln_gdp</i>                | 240                    | 28.392    | 1.794              | 26.662   | 31.251     |
| <i>CPI</i>                   | 240                    | 94.834    | 28.997             | 41.229   | 147.726    |
| <i>is_oecd</i>               | 240                    | 0.625     | 0.485              | 0        | 1          |
| <i>NEER</i>                  | 240                    | 101.774   | 13.287             | 73.459   | 134.705    |
| <i>REER</i>                  | 240                    | 100.517   | 11.208             | 76.738   | 129.034    |
| <i>total_index</i>           | 240                    | 106.623   | 34.99              | 41.147   | 247.791    |
| <i>energy_index</i>          | 240                    | 114.458   | 49.781             | 39.216   | 322.989    |
| <i>nonenergy_index</i>       | 240                    | 102.053   | 22.005             | 45.052   | 172.631    |
| <i>agriculture_index</i>     | 240                    | 98.144    | 22.202             | 48.864   | 181.59     |
| <i>fertilizers_index</i>     | 240                    | 100.269   | 53.943             | 20.503   | 310.636    |
| <i>metals_minerals_index</i> | 240                    | 111.944   | 39.012             | 42.266   | 219.781    |
| <i>precious_metals_index</i> | 240                    | 117.923   | 54.048             | 33.053   | 281.421    |

Notes: *is\_oecd* stands for OECD membership dummy (1 = OECD member; 0 = non-member).

commodity price indices show relatively large standard deviations for the energy and metals indices, indicating a significantly greater volatility than non-energy and agricultural indices.

Figure 1 is a heatmap visualizing the Pearson correlation coefficients between the main variables used in this study. Darker colors signify a larger absolute correlation coefficient [positive (+) correlations are indicated by bright colors and negative (−) correlations are indicated by dark colors], which allows for an intuitive understanding of the correlation structure among the variables. According to the heatmap results, the global commodity index group (such as the total index, non-energy index, agriculture index, fertilizers index, and metals and minerals index) shows strong positive (+) correlations with each other (generally  $|\rho| \geq 0.7$ ), confirming a strong co-movement among raw material markets.

In addition, the effective exchange rates (*neer* and *reer*) show negative correlations with many real-sector indicators. In particular, the moderate negative correlation observed between the *REER* and exports (*reer*–*ln\_exports*) aligns with the theoretical expectation that an appreciation of the real exchange rate weakens export competitiveness. In contrast, GDP (*ln\_gdp*) exhibits a strong positive correlation with exports, reaffirming the empirical regularity that larger economies tend to generate larger volumes of trade. Moreover, the *VIX* demonstrates weak negative correlations with most indicators, whereas *oil\_price* show a very strong positive

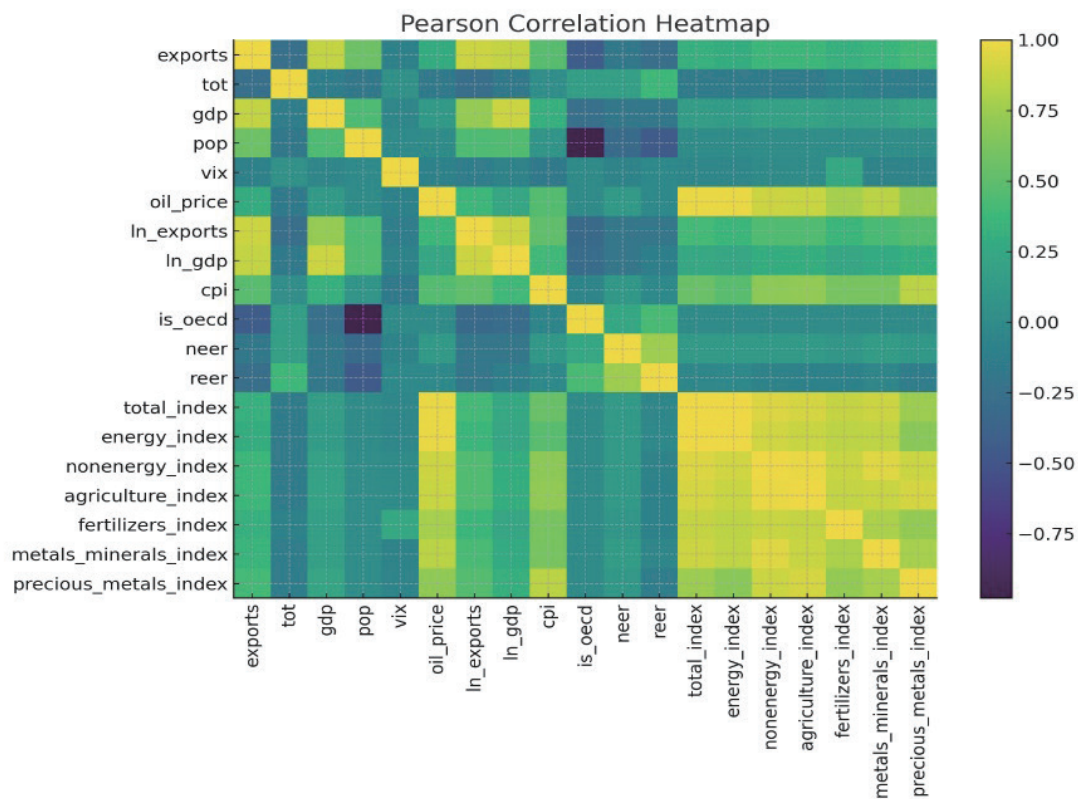


Fig. 1. (Color online) Pearson correlation analysis (heatmap).

correlation with the energy and total commodity indices. This suggests that crude oil prices function as a major common factor that more broadly influence global commodity markets. Overall, the heatmap not only identifies the correlation structure among the key explanatory variables used in this study, but also helps detect potential multicollinearity risks that may arise in linear models (pooled OLS/fixed effects (FE) Panel). Conversely, it provides a methodological rationale for why ML and DL models (RF and Transformer) may display relative advantages under such high-correlation structures.

### 3.2 Comparison of baseline forecasting models

Table 2 presents the out-of-sample (OOS) forecasting performance characteristics of the traditional statistical model (pooled OLS), ML model (RF), and DL model (Transformer encoder). First, the pooled OLS model showed an *RMSE* of 0.456, an *MAE* of 0.391, an *MAPE* of 1840.9, and an OOS  $R^2$  of  $-20.61$ , confirming that its forecasting power virtually collapsed in the OOS period. This was confirmed to be because the linear model failed to capture appropriate patterns owing to the strong nonlinearity and structural changes in trade and macroeconomic time series.

In contrast, the ML-based RF showed a significant improvement in performance compared with pooled OLS, with forecast errors greatly reduced to an *RMSE* of 0.134 and an *MAE* of 0.116. The OOS  $R^2$  also improved from  $-20.61$  to  $-0.85$ , showing a “drastically enhanced relative predictive power compared with the linear model, though still not fully satisfactory”. This means that, in a situation where export growth rates have nonlinear and interactive relationships with variables such as *TOT*, real/nominal exchange rates, and global commodity indices, the tree-based ensemble more efficiently absorbs the complex branching structure and nonlinearity among individual variables.<sup>(15)</sup>

In summary, RF functions as a ‘model that is difficult to explain but practically captures complex patterns that traditional regression misses’, and the performance advantage of ML is strongly supported when the prediction superiority of complex models and nonlinearity are high. The DL-based Transformer encoder recorded the highest predictive accuracy among the three models, with an *RMSE* of 0.109 and an *MAE* of 0.086. The Transformer’s OOS  $R^2$  value of  $-0.44$  still does not exceed 0, but it records a less negative value than RF ( $-0.85$ ), showing improved performance in terms of overall variance explanation and error structure. This suggests that the self-attention-based structure, which directly learns the temporal patterns (sequence) of exports

Table 2  
Performance characteristics of analytic models.

| Model               | <i>RMSE</i> | <i>MAE</i> | <i>MAPE</i> (%) | OOS $R^2$ |
|---------------------|-------------|------------|-----------------|-----------|
| Pooled OLS          | 0.456       | 0.391      | 1840.9          | $-20.61$  |
| RF                  | 0.134       | 0.116      | 568.64          | $-0.85$   |
| Transformer encoder | 0.109       | 0.086      | 346.42          | $-0.22$   |

and macroeconomic variables from three or more past periods, more precisely captures long-term dependency than does RF, which branches based on single time points.<sup>(16)</sup> Therefore, while the DL model shows the highest performance in trade forecasting, a pattern was confirmed where not only the average level of the prediction error but also the magnitude of extreme values (tails) progressively decreased from ML to DL. Although the OOS  $R^2$  was negative for both RF and Transformer, this reflects the fundamental difficulty of macro-forecasting while simultaneously confirming a clear performance hierarchy in relative comparison: Transformer > RF > pooled OLS.

### 3.3 Robustness checks

To ensure the consistency of the empirical findings presented in this study, three types of robustness checks were conducted: (1) replace the traditional statistical model with a fixed-effects panel regression model, (2) replace the DL model with an LSTM model, and (3) statistically verify differences in predictive performance using the DM test.

#### 3.3.1 Comparison of explanatory power in traditional statistical models

While the primary goal of this study is to use pooled OLS as a simple linear benchmark to demonstrate the predictive superiority of the RF and Transformer models, the pooled OLS model has an inherent limitation—it cannot control the country-specific unobserved heterogeneity embedded in panel data.<sup>(14)</sup> Ignoring such heterogeneity may introduce endogeneity problems in coefficient estimation, thereby undermining the statistical validity of the model. Therefore, to evaluate the statistical robustness of the main predictive results, we additionally employ panel data regression models that explicitly control the unobserved heterogeneity.

In particular, an FE model is used to account for time-invariant country-specific characteristics. A Hausman test<sup>(23,24)</sup> was conducted to determine the relative appropriateness of the FE versus random-effect specification. The results reject the null hypothesis (“the RE model is appropriate”, i.e., no systematic correlation between regressors and individual effects) at the 5% significance level, confirming the suitability of the FE specification. If the complex models outperform the simple linear model even after controlling the unobserved heterogeneity through an FE model, this provides strong support that such superior predictive performance is not merely the result of misspecification errors inherent in simple models.

According to the robustness-check results presented in Table 3, the FE model exhibits the lowest predictive performance among all models, including the baseline pooled OLS model ( $RMSE = 1.670$ ,  $MAE = 1.026$ ). Most notably, the OOS  $R^2$ , which indicates explanatory power, is  $-288.10$ , suggesting that although FE may be advantageous for statistical causal inference, it performs extremely poorly in forecasting tasks that require extrapolating complex nonlinear patterns in trade data.

Table 3  
Results of robust checks.

| Model               | <i>RMSE</i> | <i>MAE</i> | <i>MAPE</i> (%) | OOS $R^2$ |
|---------------------|-------------|------------|-----------------|-----------|
| Fixed effect panel  | 1.670       | 1.026      | 3461.3          | −288.10   |
| RF                  | 0.134       | 0.116      | 568.64          | −0.85     |
| Transformer encoder | 0.106       | 0.084      | 323.60          | −0.16     |

### 3.3.2 Comparison of DL model predictive power

From the analysis results of this study, the Transformer model is determined to be the most suitable DL model for accurately predicting complex and long-term trade patterns that pooled OLS and RF models struggle to capture. To ensure the consistency of the findings of this study, the performance of the Transformer model was compared with that of the LSTM model, which overcomes the limitations of RNNs and can effectively assist in learning the time-series and panel characteristics of trade data. Traditional RNNs compress information from previous time steps into a hidden state vector and pass it to the next time step.

However, as the length of the sequence data increases, the vanishing gradient problem occurs, where the gradient needed to update the weights of the initial time steps converges to zero, essentially preventing learning. This leads to the long-term dependency problem, where the network only remembers recent information and forgets very old past information. For example, it becomes difficult to determine the impact of a macroeconomic shock from 10 years ago on current exports in trade data. The LSTM model was devised to solve this long-term dependency problem in time-series and sequence data.

As shown in Table 3, unlike the FE panel analysis model that showed significant limitations in predictive power, the LSTM model achieved the highest predictive accuracy among all competing models, relying on the strength of its DL structure to solve the long-term dependency problem in time-series data.<sup>(19)</sup> LSTM recorded an *RMSE* of 0.106, an *MAE* of 0.084, and an *MAPE* of 323.60, significantly outperforming the predictive performance of the statistically most rigorous FE model. This result clearly shows that the predictive superiority of DL models stems not from statistical estimation errors, but from the ability to capture nonlinear relationships through data-driven learning.

### 3.3.3 Statistical validation of predictive difference: Diebold–Mariano test

Table 4 reports the results of the DM test conducted to examine whether the differences in predictive accuracy among models are statistically significant. In line with the robustness checks, we performed three DM tests: a comparison among the pooled OLS–RF–Transformer models; a comparison among the FE panel–RF–Transformer models, and a comparison among the FE panel–RF–LSTM models.

First, the DM test was used to statistically compare differences in predictive power among the pooled OLS–RF–Transformer models. The results show a very clear performance gap between the traditional panel regression model (pooled OLS) and nonlinear learning-based

Table 4  
Predictive performance characteristics of ML and DL models.

| Model A            | Model B             | Loss | DM_stat  | P_value |
|--------------------|---------------------|------|----------|---------|
| Pooled OLS         | RF                  | MSE  | 10.39701 | 0.00000 |
| Pooled OLS         | Transformer Encoder | MSE  | 10.97136 | 0.00000 |
| RF                 | Transformer Encoder | MSE  | 2.50343  | 0.01355 |
| Model A            | Model B             | Loss | DM_stat  | P_value |
| Fixed Effect Panel | RF                  | MSE  | 4.31846  | 0.00002 |
| Fixed Effect Panel | Transformer Encoder | MSE  | 4.32936  | 0.00001 |
| RF                 | Transformer Encoder | MSE  | 4.92014  | 0.00000 |
| Model A            | Model B             | Loss | DM_stat  | P_value |
| Fixed Effect Panel | RF                  | MSE  | 4.31846  | 0.00003 |
| Fixed Effect Panel | LSTM                | MSE  | 4.32814  | 0.00003 |
| RF                 | LSTM                | MSE  | 4.42389  | 0.00002 |

models (RF and Transformer). In the comparison of pooled OLS versus RF, the DM statistic was 10.397 ( $p < 0.001$ ). This indicates that, in terms of *MSE*, the prediction error of RF is significantly smaller than that of pooled OLS, suggesting that the tree-based ensemble model, which learns nonlinear and interaction structures, overwhelmingly outperforms the linear panel model. In addition, in the comparison of pooled OLS versus Transformer, the DM statistic was 10.971 ( $p < 0.001$ ), a very high value. This provides strong statistical evidence that the DL-based Transformer has markedly superior predictive power relative to pooled OLS and implies that Transformer effectively captures the complex nonlinearity and time-series dependence structure of trade and macroeconomic indicators. Moreover, in the comparison between RF and Transformer, the DM statistic was 2.503 ( $p = 0.01355$ ). This shows that the predictive power of Transformer is statistically significantly superior to that of RF. In other words, even in the comparison between two nonlinear models, Transformer delivers more refined predictive performance than the tree-based model because of its ability to learn long-range dependency and handle time-series structures.

Second, the DM test was used to statistically compare differences in predictive power among the FE panel–RF–Transformer models. The results again reveal a very clear performance gap between the FE panel model and nonlinear learning-based models (RF and Transformer). In the comparison between the FE panel model and RF, the DM statistic was 4.31846 and the p-value was 0.00002. This means that the null hypothesis of equal predictive power between FE and RF is strongly rejected at the 1% significance level, indicating that the difference in predictive accuracy between the two models is statistically significant. Given that the DM statistic is positive, this can be interpreted as RF having a significantly lower prediction error (*MSE*) than the FE model and thus exhibiting superior performance. A similar pattern was observed in the comparison between the FE model and Transformer. The DM statistic was 4.32936 and the p-value was 0.00001, implying that the null hypothesis of equal predictive power between the two models is significantly rejected. This can be interpreted as evidence that Transformer systematically outperforms the FE model and that the DL-based time-series encoder captures country-specific heterogeneity and nonlinear trade patterns much more effectively than the

traditional FE model. Moreover, in the comparison between RF and Transformer, the DM statistic was 4.92014 and the p-value was 0.00000. The difference in predictive power between the two ML·DL models is also statistically significant, and, given the positive value, this implies that Transformer has a greater predictive power than RF. This reflects the fact that the self-attention mechanism of Transformer effectively learns long-term dependence and the propagation patterns of nonstandard shocks in time-series structures.

Third, using the DM test to statistically compare differences in predictive power among the FE panel–RF–LSTM models again revealed very clear performance gaps between the traditional panel regression model and nonlinear learning-based models (RF and LSTM). First, in the comparison of the FE model versus RF, the result was  $DM\_stat = 4.31846$  ( $p = 0.00003$ ), indicating that RF recorded a significantly lower prediction error ( $MSE$ ) than the FE model. This implies that the linear structure of the traditional FE panel model does not sufficiently capture the nonlinearity and volatility of trade and macro time series. In addition, in the comparison of the FE model versus LSTM, the result was  $DM\_stat = 4.32814$  ( $p = 0.00003$ ), which strongly suggests that the DL-based LSTM provides superior predictive power relative to the FE model. Although LSTM is relatively limited compared with Transformer in learning long-term dependency, it can be interpreted that, owing to the inherent capacity of RNN-type models to recognize time-series trends and cyclical patterns, LSTM nonetheless shows higher performance than the FE model. In the RF versus LSTM comparison, the result was  $DM\_stat = 4.42389$  ( $p = 0.00002$ ). This means that LSTM achieves significantly lower prediction error than RF, reaffirming the strength of DL-based architectures in handling time-series dependence in trade and macro indicators.

In summary, the DM tests show that there exists a clear hierarchical difference in predictive power among competing models, leading to the conclusion that predictive performance follows the order Transformer encoder > LSTM > RF > FE > pooled OLS. This implies that pooled OLS or panel regression models are limited in capturing complex nonlinear structures and responses to global shocks, and that ML and DL-based methodologies are more suitable for trade and macro-forecasting.

#### 4. Discussion

In this study, we compared the OOS predictive performance characteristics of three analytical approaches—traditional statistical modeling (OLS), ML (RF), and DL (Transformer encoder)—for forecasting country-level export growth. The empirical analysis yields several theoretical and practical implications.

First, at the theoretical level, the extremely low explanatory power of the traditional pooled OLS model—reflected in an OOS  $R^2$  of  $-20.61$ —suggests that linear models are fundamentally limited in capturing the strong nonlinearities and structural breaks inherent in export growth driven by fluctuations in trade conditions, exchange rates, and commodity prices.<sup>(24,25)</sup> Nevertheless, pooled OLS remains well suited for explaining “direction” and “average effects” through coefficient interpretation and confidence intervals. Thus, OLS functions as a humble baseline model for explanation rather than prediction. By contrast, the RF model demonstrates

substantially improved performance ( $RMSE = 0.1337$ ), highlighting the advantage of ML models that automatically capture nonlinear and interaction effects. The most notable result is the Transformer encoder model with a sequence length of three, which achieves the highest predictive accuracy among all models ( $RMSE = 0.1085$ ;  $MAE = 0.086$ ). The Transformer's attention mechanism enables the model to learn both short- and long-term dependencies within time-series data and capture high-dimensional interactions among macroeconomic and trade-related indicators through its embedding architecture, thereby enhancing predictive performance.

This perspective provides an important methodological implication: ML and DL models such as Transformer and LSTM can be interpreted as tools that learn complex aggregation structures of sensor-generated data. Their superiority in capturing nonlinear relationships may stem from their ability to model the hidden dynamics embedded in high-frequency physical data streams, even when only aggregated macro-level variables are observed. However, the fact that both the RF and Transformer models exhibit negative OOS  $R^2$  values reflects the fundamental difficulty of long-horizon macroeconomic forecasting. A negative OOS  $R^2$  means that the model performs worse than a naïve mean forecast, indicating that despite capturing complex patterns, the models amplified variance excessively, failing to control overfitting. Complex models often display low or even negative OOS  $R^2$  values,<sup>(26)</sup> implying that their statistical predictive power can be limited. This result serves as a reminder that complex models do not guarantee predictive superiority in a statistical sense and underscores the need to guard against the arrogance of prediction. In short, the methodological superiority of a given model is not absolute; it must be evaluated relative to the analytical purpose.

Second, from a practical standpoint, the findings highlight the necessity of an integrated modeling strategy for export forecasting and policymaking—one that balances both interpretability and predictive accuracy. Specifically, a dual-model strategy is recommended. The pooled OLS model should be used to identify average linear effects and provide intuitive policy-relevant explanations of export fluctuations. Complex models such as Transformer—those with superior  $RMSE/MAE$  performance—should be deployed to extract high-frequency predictive signals. When the predictions of OLS and Transformer models diverge, such discrepancies may serve as an early-warning indicator of nonlinear shocks or structural breaks. Operating explanatory and predictive models in parallel enhances the efficiency and reliability of export monitoring systems.

A second practical implication concerns risk management based on prediction errors. Metrics such as  $RMSE$  and  $MAE$  can be used to set accuracy thresholds and manage the acceptable error bounds of predictive systems. The negative OOS  $R^2$  observed in complex models—reflecting both the inherent challenges of macroeconomic forecasting and the risk of overfitting—serves as a practical warning. Even sophisticated ML/DL models should not be relied upon exclusively for long-term policy planning. A cautious risk-control strategy that supplements model outputs with expert qualitative judgment is essential to avoid overconfidence in automated predictions.

Third, from a methodological perspective, it is essential to diagnose the underlying causes of the negative OOS  $R^2$  values observed in complex models. Analogous to identifying noise and instability in physical sensor systems, these predictive failures should not be interpreted solely as

model inadequacy, but rather as the outcome of multiple interacting factors. One possible explanation is that model misspecification arises from a mismatch between model complexity and data scale. The dataset used in this study consists of a relatively limited number of observations, which may not be sufficient for training high-capacity models such as Transformer and LSTM. As a result, these models may overfit to idiosyncratic patterns or noise in the training data, leading to degraded generalization performance in OOS evaluation.

Another important factor is temporal instability in global trade data. Trade and macroeconomic variables are inherently subject to structural breaks, regime shifts, and exogenous shocks, such as financial crises and pandemics. These dynamics violate the stationarity assumption and reduce the reliability of historical patterns for future prediction. In such environments, even highly flexible models struggle to generalize beyond the training period. Additionally, the performance of ML and DL models is highly sensitive to hyperparameter configurations. Model settings such as tree depth, learning rate, attention layers, and sequence length can significantly affect predictive outcomes. Without exhaustive tuning or robust cross-validation, suboptimal hyperparameter choices may lead to unstable predictions and negative OOS  $R^2$  values.

Taken together, these findings suggest that negative OOS  $R^2$  values reflect not only the limitations of complex models but also the intrinsic difficulty of forecasting highly volatile macroeconomic systems. Therefore, model evaluation should go beyond conventional accuracy metrics and incorporate diagnostic perspectives that account for data characteristics, model complexity, and structural uncertainty.

## 5. Conclusion

The negative OOS  $R^2$  values observed in the ML and DL models indicate that, even though these complex models exhibit marginal improvements in  $RMSE$  or  $MAE$ , they ultimately increase statistical risk—reflected in variance—captured by  $R^2$ . This finding reinforces the conclusion that model superiority is not absolute but conditional, depending on a combination of data characteristics, objectives, cost considerations, calibration choices, and threshold settings. Nonetheless, even with low  $R^2$  values, complex models can outperform simpler ones in terms of practical forecasting capabilities—particularly in capturing directional signals and timing in finance, trade, and macroeconomic contexts. Therefore, beyond the comparison of  $RMSE$  differences, the results demonstrate that the observed performance gaps are not random but statistically meaningful.

These main findings provide both theoretical and practical insights for researchers. However, the present study does not offer a fully complete narrative on model choice and methodological design owing to inherent data limitations and fundamental constraints embedded in all analytical frameworks. Consequently, several important avenues for future research remain.

First, additional performance metrics should be examined to evaluate and compare the explanatory and predictive capabilities of competing models. Although the three indicators used in this study— $RMSE$ ,  $MAE$ , and  $MAPE$ —are widely adopted in time-series forecasting and

provide substantial information, alternative metrics may be necessary depending on the research question and model characteristics. *RMSE* and *MAE* quantify forecast errors in distinct ways, highlighting the model's sensitivity to large deviations, while *MAPE* expresses prediction errors in relative percentage terms, facilitating intuitive interpretation. However, other measures such as  $R^2$  or adjusted  $R^2$ , and symmetric *MAPE* (*sMAPE*), may also be appropriate. The former helps evaluate the proportion of variance explained by the model—especially in statistical modeling—while the latter addresses the limitations of *MAPE*, such as divergence when actual values approach zero or oversensitivity to small actual values.

Second, the robustness comparison of LSTM versus Transformer encoder models revealed that, despite using identical input sequences (three-year macroeconomic and trade variables), the Transformer encoder achieved modestly higher *RMSE* and *MAPE* performance. This improvement is likely attributable to the strengths of the self-attention mechanism, which flexibly captures both short- and long-term dependencies and variable interactions. Importantly, however, Transformer was not uniformly superior across all countries. In some cases—particularly those subject to intermittent shocks—the LSTM model displayed more stable performance. This highlights the need for model selection that is sensitive to data characteristics and forecasting objectives, and calls for more nuanced methodological investigation into the comparative applicability of DL architectures.

Finally, model generalization should be assessed more comprehensively by incorporating additional metrics beyond OOS  $R^2$ . While OOS  $R^2$  is among the most stringent indicators of a model's ability to generalize to unseen data, other important measures are widely used in ML and DL research to evaluate predictive risk or generalization error. Such metrics rely on repeated sampling methods—such as cross-validation or bootstrap procedures—to generate multiple OOS datasets. The average *RMSE* or *MSE* across these samples, along with their confidence intervals, provides a more rigorous evaluation of model stability. These approaches mitigate the limitations of relying solely on a single OOS partition and help determine whether a model maintains consistent generalization across diverse data environments. Incorporating these additional metrics complements the limitations of OOS  $R^2$  and enables a more holistic assessment of model superiority.

A complete understanding of the relative merits of simple and complex models requires addressing the future research directions outlined above. It is hoped that this study will serve as a useful catalyst for further methodological debate and development in predictive modeling for trade and macroeconomic analysis.

### Acknowledgments

This research was supported by the ANCHOR program through the Gangwon ANCHOR Center, funded by the Ministry of Education (MOE) and the Gangwon State (G.S.), Republic of Korea. (2026-ANCHOR-10-002).

## References

- 1 J. H. Van Binsbergen, X. Han, and A. Lopez-Lira: *The Rev. Financial Stud.* **36** (2023) 2361. <https://doi.org/10.1093/rfs/hhag017>
- 2 D. Kim, G. Baek, and Y. Kim: *J. Korean Data Inf. Sci. Soc.* **28** (2017) 1217. <https://doi.org/10.7465/jkdi.2017.28.6.1217>
- 3 H. R. Varian: *J. Econ. Perspect.* **28** (2014) 3. <https://doi.org/10.1257/jep.28.2.3>
- 4 S. Mullainathan and J. Spiess: *J. Econ. Perspect.* **31** (2017) 87. <https://doi.org/10.1257/jep.31.2.87>
- 5 H. Kim and M. Lee: *Korean J. Marketing* **36** (2021) 1. <https://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE10532270>
- 6 H. Eom, J. Kim, and S. Choi: *J. Intell. Inform. Syst.* **26** (2020) 105. <http://dx.doi.org/10.13088/jiis.2020.26.2.105>
- 7 S. Kim, Y. Lee, J. Seo, and K. Kim: *Asia-Pacific J. Bus. Venturing Entrepreneurship* **19** (2024) 161. [https://www.kci.go.kr/kciportal/landing/article.kci?arti\\_id=ART003156640](https://www.kci.go.kr/kciportal/landing/article.kci?arti_id=ART003156640)
- 8 K. Kang, C. Park, and H. Na: *Korean J. Bus. Adm.* **36** (2023) 1515. <https://doi.org/10.18032/kaaba.2023.36.9.1515>
- 9 H. Kim, D. Ryu, and H. Cho: *Korean J. Financial Eng.* **18** (2019) 131. [https://www.kci.go.kr/kciportal/landing/article.kci?arti\\_id=ART002510917](https://www.kci.go.kr/kciportal/landing/article.kci?arti_id=ART002510917)
- 10 Y. Park: *J. Global Trade Manage.* **30** (2023) 39. <https://scienceon.kisti.re.kr/srch/selectPORSrchArticle.do?cn=ART002958825>
- 11 C. Yi: *Korea Trade Rev.* **46** (2021) 169. <https://data.doi.or.kr/10.22659/KTRA.2021.46.1.169>
- 12 Y. Lee: *Int. Commerce Inf. Rev.* **27** (2025) 3. <https://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE12136702>
- 13 M. Kwon: *Korea Customs Rev.* **1** (2024) 69. <https://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE11820465>
- 14 W. H. Greene: *Econometric Analysis* (Pearson Education: Harlow, 2018) 8th ed.
- 15 L. Breiman: *Mach. Learning* **45** (2001) 5. <https://doi.org/10.1023/A:1010933404324>
- 16 A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin: *Proc. 2017 31st Conf. Neural Information Processing Systems (NIPS 2017)* 5998–6008.
- 17 B. Lim and S. Zohren: *Philosophical Trans. Royal Soc. A* **379** (2021) 20200209. <https://www.jstor.org/stable/27101134>
- 18 Y. Wu, S. Green, J. Ting, A. Szlam, C. Soumith, and Y. LeCun: *arXiv:2001.08317 [cs.LG]* <https://doi.org/10.48550/arXiv.2001.08317> (accessed September 2025).
- 19 S. Hochreiter and J. Schmidhuber: *Neural Comput.* **9** (1997) 1735. <https://doi.org/10.1162/neco.1997.9.8.1735>
- 20 R. J. Hyndman and G. Athanasopoulos: *Forecasting: Principles Practice* (OTexts, Melbourne, 2021) 3rd ed.
- 21 F. X. Diebold and R. S. Mariano: *J. Business Economic Statistics* **13** (1995) 253. <https://doi.org/10.1198/073500102753410444>
- 22 D. Harvey, S. Leybourne, and P. Newbold: *Int. J. Forecasting* **13** (1997) 455. <https://www.sciencedirect.com/science/article/abs/pii/S0169207096007194?via%3Dihub>
- 23 J. A. Hausman: *Econometrica: J. Econometric Soc.* **46** (1978) 1251. <https://www.jstor.org/stable/1913827?origin=crossref&seq=1>
- 24 J. M. Wooldridge: *Econometric Analysis of Cross Section and Panel Data* (MIT Press, Cambridge, MA, 2010) 2nd ed.
- 25 B. H. Baltagi: *Econometric Analysis of Panel Data* (John Wiley & Sons, Chichester, 2005) 3rd ed.
- 26 O. S. Yang: *Korean Manage. Consulting Rev.* **25** (2025) 461. <https://www.earticle.net/Article/A475686>

## About the Authors



**Oh Suk Yang** received his Ph.D. degree from Seoul National University, Republic of Korea, in 2011. His research interests range from statistics in business administration to software programming in relation to green and red bio innovation, and the application of the big data platform to digital healthcare. ([osyang30@kangwon.ac.kr](mailto:osyang30@kangwon.ac.kr))



**Seong Hun Kim** received his Ph.D. in Business Administration from Kangwon National University, Republic of Korea, in 2023. His research interests include strategic management, marketing, technology transfer, and technology commercialization. ([sh-kim@kangwon.ac.kr](mailto:sh-kim@kangwon.ac.kr))