Original Research Article

Numerical Features from Candlestick Chart Structures for ETF Return Prediction: A Comparison with OHLC Prices and Technical indicators

Brennan Park

Seoul International School, 15 Seongnam-daero 1518 beon-gil, Sujeong-gu, Seongnam-si, Gyeonggi-do, 13113, South Korea

ABSTRACT

This study examined whether transforming relationships among OHLC prices into relative and normalized numerical features can support short-term ETF return prediction. Candlestick charts display relationships among the open, high, low, and close prices through components such as the body, shadows, closing position, and opening gap. Based on these price relationships, this study defined eight numerical candlestick features and compared their prediction performance with raw OHLC information and eight technical indicators. OHLC refers to open, high, low, and close prices, and technical indicators are commonly used features in stock prediction research. The experiment used eight ETFs: SPY, QQQ, DIA, IWM, XLF, XLK, XLE, and XLV and predicted the cumulative log return over the next five trading days from 2014 to 2025. Ridge regression and random forest regressor were used as prediction models. Principal component analysis was also applied to examine whether reducing redundancy within the features affected prediction performance. The results showed that the numerical candlestick-based representations generally recorded lower observed RMSE and MAE values than the raw OHLC and technical indicators representations. Numerical candlestick with PCA also recorded the highest observed directional accuracy among the five feature representations in both prediction models. These findings suggest that explicitly encoding relative and normalized relationships within existing OHLC information may provide a more suitable input representation for short-term ETF return prediction.

Keywords: Candlestick Chart; Exchange Traded Fund (ETF); numerical candlestick Feature; principal component analysis (PCA); random forest; Return Prediction; ridge regression; technical indicator

INTRODUCTION

Financial market prices move in response to various factors, including economic conditions, corporate performance, investor sentiment, and overall market uncertainty. Short-term returns contain a large amount of noise that is difficult to predict, making it challenging to accurately forecast future stock or ETF returns. Nevertheless, short-term return prediction has important implications for investment decisions, risk management, and portfolio construction. Therefore, identifying useful information related to future price movements from financial data is an important issue in stock prediction research.

Previous stock prediction studies have often used OHLC information, consisting of open, high, low, and close prices, either directly or by transforming it into technical indicators according to mathematical rules (1–

Corresponding author: Brennan Park, E-mail: brennanpark7@gmail.com.

Accepted August 10, 2026
https://doi.org/10.70251/HYJR2348.44903920

4). OHLC provides basic price information over a trading day, but because ETFs have different price levels, using absolute prices directly may limit the comparability of movements across different assets. Technical indicators, on the other hand, have the advantage of summarizing characteristics of price movements, such as moving averages, momentum, and volatility. For example, a moving average represents the average price level over a certain period, while the Relative Strength Index shows the relative strength of recent price increases and decreases. In this way, technical indicators are frequently used as auxiliary features in stock prediction because they can express price trends and volatility more explicitly than raw prices.

OHLC information can also be represented through candlestick charts (5, 6). A candlestick organizes the open, high, low, and close prices of a trading day into components such as the body, upper and lower shadows, closing position, and opening gap. These components describe relative relationships among the same OHLC values and provide a compact graphical representation of the intraday price structure. Based on this representation, this study converts the price relationships displayed by candlestick components into numerical features. Specifically, this study defines eight numerical candlestick features using the open, high, low, close, and previous close prices. These features include the relationship between the open and close prices, the daily price range, the gap from the previous close, the relative position of the close price, and the relative sizes of the body and shadows. In other words, relationships that remain implicit in raw OHLC values are expressed as relative and normalized features that can be used directly by prediction models.

The main research questions of this study are as follows. First, how does the predictive performance of numerical candlestick features compare with that of raw OHLC information? Second, how does their performance compare with that of technical indicators? Third, how does prediction performance change when principal component analysis is applied to reduce redundancy within the features? To examine these questions, this study used OHLC data from 2010 to 2025 for eight ETFs representing the U.S. stock market. The prediction target was defined as the cumulative log return over the next five trading days for each ETF. This target was selected to represent short-term return movements over a multi-day horizon rather than a single next-day change. Ridge regression and random forest regressor were used as prediction models. Ridge regression is a linear model with regularization, while random forest regressor is an ensemble model that can learn nonlinear relationships and interactions among features. By using these two models, this study aimed to examine whether the usefulness of numerical candlestick features depends on a specific model or appears relatively consistently across different types of models.

The main contribution of this study is that it provides a systematic feature-engineering approach that explicitly represents relationships implicit in OHLC prices as mathematically defined numerical features. This study compares numerical candlestick features with raw OHLC and technical indicators under the same experimental conditions, thereby examining whether prediction performance can differ depending on how the same price information is represented. Through this analysis, the study examines whether explicitly encoding relative and normalized OHLC relationships may provide a suitable input representation for the prediction models evaluated in this study.

METHODS AND MATERIALS

Experimental Setup

ETF Universe

This study used eight ETFs listed in Table 1. SPY, QQQ, DIA, and IWM were selected because they track major U.S. market indices: the S&P 500, Nasdaq-100, Dow Jones Industrial Average, and Russell 2000. These ETFs represent different parts of the U.S. stock market, including large-cap stocks, growth-oriented stocks, blue-chip stocks, and small-cap stocks. XLF, XLK, XLE, and XLV were also included to cover major sectors such as financials, technology, energy, and health care. By using both broad market ETFs and sector ETFs, this study examines whether numerical candlestick features provide a useful alternative representation to raw OHLC prices and technical indicators across different types of ETFs.

Daily OHLC data for the eight ETFs were collected from Yahoo Finance using the Python yfinance package on March 15, 2026. Although the actual experimental period extended from January 4, 2010, to December 31, 2025, data from January 2, 2009, to January 30, 2026, were collected to obtain the historical observations required for the initial calculation of technical indicators and the future closing prices required to construct the five-trading-day target variable at the end of 2025. The data were downloaded with auto_adjust=False. The Open, High, Low, and Close fields were used in the analysis, whereas Adjusted Close was not used. No

Table 1. Exchange-traded fund (ETF) universe used in the study. *The table reports the ticker, full fund name, and represented market segment for the eight U.S.-listed ETFs included in the analysis. SPY, QQQ, DIA, and IWM represent broad U.S. equity indices covering large-cap, growth-oriented, blue-chip, and small-cap stocks, respectively. XLF, XLK, XLE, and XLV represent the financial, technology, energy, and health care sectors, respectively.*

Ticker	Full Name	Market Segment
SPY	SPDR S&P 500 ETF Trust	Broad U.S. large-cap equities
QQQ	Invesco QQQ Trust	Nasdaq-100 growth-oriented equities
DIA	SPDR Dow Jones Industrial Average ETF Trust	Dow Jones blue-chip equities
IWM	iShares Russell 2000 ETF	U.S. small-cap equities
XLF	Financial Select Sector SPDR Fund	U.S. financial sector
XLK	Technology Select Sector SPDR Fund	U.S. technology sector
XLE	Energy Select Sector SPDR Fund	U.S. energy sector
XLV	Health Care Select Sector SPDR Fund	U.S. health care sector

additional price adjustment was applied by the researcher. After the collected data were sorted chronologically and examined for duplicate dates and missing OHLC values, no duplicate or incomplete observations were found. Weekends and market holidays were already treated as non-trading days in the downloaded data, and no additional observations were generated for these dates. Therefore, prices were neither interpolated nor replaced with previous prices.

Prediction Target

The prediction target of this study was the cumulative log return over the next five trading days for each ETF. Although it is possible to directly predict the future closing price, closing prices are not always comparable across ETFs because different ETFs have different price levels. For example, the same prediction error may not have the same economic meaning for SPY and IWM if their price levels are different. Therefore, this study used return, rather than price level, as the prediction target.

Let P_t be the closing price at time t. The one-day log return is defined as follows:

$$r_t = log\left(\frac{P_t}{P_{t-1}}\right)$$

Here, P_t is the closing price at time t, and P_{t-1} is the closing price on the previous trading day. Log return was used because returns over multiple periods can be added easily. For example, the cumulative log return from time to time t + 5 can be written as the sum of five daily log returns:

$$\sum_{k=1}^{5} r_{t+k} = log\left(\frac{P_{t+1}}{P_t}\right) + log\left(\frac{P_{t+2}}{P_{t+1}}\right) + \cdots + log\left(\frac{P_{t+5}}{P_{t+4}}\right)$$

Using the property of logarithms, this expression can be simplified as follows:

$$\sum_{k=1}^{5} r_{t+k} = log\left(\frac{P_{t+1}}{P_t} \cdot \frac{P_{t+2}}{P_{t+1}} \cdots \frac{P_{t+5}}{P_{t+4}}\right) = log\left(\frac{P_{t+5}}{P_t}\right)$$

Therefore, the prediction target y_t is defined as:

$$y_t = log\left(\frac{P_{t+5}}{P_t}\right)$$

This value represents how much the closing price increases or decreases over the next five trading days. If $y_t > 0$, the ETF price increased over the next five trading days. If $y_t < 0$, the ETF price decreased.

The input of the model is a feature vector constructed using one of three main representations: raw OHLC prices, technical indicator-based features, or numerical candlestick features. The model uses this input vector to predict y_t:

$$\hat{y}_t = f(x_t)$$

Here, x_t is the input feature vector at time t, and $\hat{y}_t$ is the predicted cumulative log return over the next five trading days.

Rolling Train-Test Split

This study used a rolling train-test split instead of a random train-test split to preserve the time order of the data. In financial time-series data, using future information during training can cause data leakage. Therefore, the training period was always placed before the test period. Specifically, this study used a rolling structure with a four-year training period and a one-year test period. For example, when 2014 was used as the test year, the data from 2010 to 2013 were used for training, and the data from 2014 were used for testing. Then, the test year was moved forward by one year, and the same process was repeated.

In this study, the test years ranged from 2014 to 2025. Therefore, each ETF had 12 rolling folds. Since this study used eight ETFs and twelve test years, the experiment included 96 rolling folds in total. This rolling split structure is shown in Table 2. All models and feature sets were evaluated under the same rolling split structure so that the results could be compared under the same data conditions.

Training was conducted independently for each ETF, and observations from different ETFs were not pooled into a single training sample. Accordingly, a separate model was trained for each combination of ETF, test year, algorithm, and feature representation. In addition, to prevent the five-trading-day target variable from using price information from the test period, observations near the end of each training period were excluded when their target dates fell on or after the first trading day of the corresponding test period. The exact first and last trading dates and the corresponding numbers of training and test observations for all 12 rolling folds are reported in Table 3.

Table 2. Rolling training and testing periods used for each ETF. *Each rolling fold consists of a four-year training period followed chronologically by a one-year test period. The first fold uses observations from 2010–2013 for training and observations from 2014 for testing. The training and test periods are then advanced by one calendar year until the final fold, which uses 2021–2024 for training and 2025 for testing. The same 12-fold structure was applied independently to each of the eight ETFs, resulting in 96 ETF-year test folds. No random train-test splitting was used.*

Fold Number	Training Period	Test Period
1	2010–2013	2014
2	2011–2014	2015
3	2012–2015	2016
4	2013–2016	2017
…	…	…
12	2021–2024	2025

Table 3. Training and testing samples for each rolling fold. *The table reports the first and last trading dates and the numbers of observations after excluding training observations whose five-day target dates fell on or after the first trading day of the corresponding test period.*

Fold	Training sample			Test sample		
	First trading date	Last trading date	Observations (N)	First trading date	Last trading date	Observations (N)
1	2010-01-04	2013-12-23	1,001	2014-01-02	2014-12-31	252
2	2011-01-03	2014-12-23	1,001	2015-01-02	2015-12-31	252
3	2012-01-03	2015-12-23	1,001	2016-01-04	2016-12-30	252
4	2013-01-02	2016-12-22	1,003	2017-01-03	2017-12-29	251
5	2014-01-02	2017-12-21	1,002	2018-01-02	2018-12-31	251
6	2015-01-02	2018-12-21	1,001	2019-01-02	2019-12-31	252
7	2016-01-04	2019-12-23	1,001	2020-01-02	2020-12-31	253
8	2017-01-03	2020-12-23	1,002	2021-01-04	2021-12-31	252
9	2018-01-02	2021-12-23	1,003	2022-01-03	2022-12-30	251
10	2019-01-02	2022-12-22	1,003	2023-01-03	2023-12-29	250
11	2020-01-02	2023-12-21	1,001	2024-01-02	2024-12-31	252
12	2021-01-04	2024-12-23	1,000	2025-01-02	2025-12-31	250

Evaluation Metrics

This study used RMSE, MAE, and directional accuracy to evaluate prediction performance. Since the prediction target is a continuous return value, RMSE and MAE were used as the main regression performance metrics. Directional accuracy was also used because the direction of return is important in financial prediction.

First, RMSE measures the square root of the average squared difference between the actual and predicted values. For test observations, let be the actual value and be the predicted value. RMSE is defined as follows:

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

RMSE gives a larger penalty to large prediction errors. Therefore, if the predicted value is far from the actual return, RMSE increases more strongly. In this study, a lower RMSE indicates better prediction performance.

MAE is the average absolute difference between the actual and predicted values.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left| y_i - \hat{y}_i \right|$$

MAE is less sensitive to extreme errors than RMSE and shows how far the predictions are from the actual values on average. Like RMSE, a lower MAE indicates better performance.

Finally, directional accuracy measures the proportion of cases in which the predicted return and the actual return have the same direction.

$$DA = \frac{1}{n}\sum_{i=1}^{n} I\left(sign\left(y_i\right) = sign\left(\hat{y}_i\right)\right)$$

Here, $I(\cdot)$ is an indicator function that equals 1 if the condition is true and 0 otherwise. Directional accuracy evaluates whether the model correctly predicts the upward or downward direction of returns, rather than the exact size of the return. For example, if the actual five-day cumulative log return is positive and the predicted value is also positive, the prediction is counted as directionally correct.

RMSE, MAE, and directional accuracy were first calculated separately for each combination of ETF and test year. Specifically, all daily predictions within each ETF-year rolling fold were used to calculate one value for each evaluation metric. The overall results were then calculated as the unweighted arithmetic mean of the 96 ETF-year-level metric values, consisting of eight ETFs and 12 test years from 2014 to 2025. Thus, the overall metrics represent averages of the fold-level metrics rather than metrics recalculated after pooling all daily predictions. This aggregation procedure ensured that each ETF-year fold received equal weight regardless of the number of trading days included in the fold, preventing folds with more observations from receiving greater weight solely because of their sample size. The ETF-level results were calculated by averaging the 12 annual metric values for each ETF, while the year-level results were calculated by averaging the metric values of the eight ETFs for each test year.

Also, because consecutive five-day cumulative returns share up to four future daily returns, a paired moving-block bootstrap was additionally conducted to account for dependence among daily prediction outcomes. Within each test year, the daily observations were resampled with replacement in moving blocks of five consecutive trading days until a bootstrap sample with the same length as the original test sample was obtained. The same resampling indices were applied to the candidate and comparator feature representations. Observations for all eight ETFs corresponding to the same dates were also resampled jointly to retain contemporaneous dependence across ETFs.

This procedure was repeated 10,000 times to calculate 95% percentile bootstrap confidence intervals for differences in RMSE, MAE, and directional accuracy. For RMSE and MAE, the performance difference was calculated by subtracting the candidate value from the comparator value. For directional accuracy, the difference was calculated by subtracting the comparator value from the candidate value. Therefore, a positive difference indicates a result favoring the numerical candlestick-based representation for all three metrics.

Technical indicators

A technical indicator is a variable that transforms stock price information, such as OHLC data, according to a mathematical rule to summarize price trend, momentum, volatility, and other market conditions. In stock prediction studies, technical indicators are often used as additional features instead of using only raw OHLC prices. This is because technical indicators can express the direction of price movement, the speed of change, the distance from an average level, and the structure of volatility more explicitly than raw price levels.

In this study, eight representative technical indicators were used to construct technical indicator-based features. These features served as a comparison representation against raw OHLC prices and numerical candlestick features. The number of technical indicators was set to eight, the same as the number of numerical candlestick features. This was done to reduce comparison bias caused by different numbers of features. In addition, because this study uses only raw OHLC information and does not use volume, volume-based indicators were excluded.

However, the numerical candlestick features used in this study are mostly defined as log ratios or ratios relative to the daily price range. Therefore, they are less sensitive to the absolute price level of each ETF. For a fairer comparison, the technical indicators were also used in relative or normalized forms, when necessary, while keeping their original calculation structures. This was done to reduce the possibility that differences in ETF price levels would act as unnecessary noise in the prediction process.

The eight technical indicator-based features used in this study are defined below. At time t, let the open, high, low, and close prices be O_t, H_t, L_t, and C_t, respectively. Also, ϵ is a small positive constant added to avoid division by zero. The first feature is Simple Moving Average Distance, which is based on the 20-day simple moving average. The n-day simple moving average, denoted as $SMAn,t$, is defined as the average closing price over the most recent n trading days:

$$SMA_{n,t} = \frac{1}{n}\sum_{i=0}^{n-1} C_{t-i}$$

Since the raw moving average itself is affected by the price level of an ETF, this study transformed it into a relative distance between the current close price and the 20-day simple moving average:

$$TI_t^{(1)} = \frac{C_t}{SMA_{20,t}} - 1$$

A positive value means that the current close price is above its 20-day average, while a negative value means that the current close price is below its 20-day average.

The second feature is Exponential Moving Average Distance, which is based on the 20-day exponential moving average. The -day exponential moving average, denoted as $EMAn,t$, gives more weight to recent prices and is defined as follows:

$$EMA_{n,t} = \alpha_n C_t + (1 - \alpha_n) EMA_{n,t-1}, \alpha_n = \frac{2}{n+1}$$

To reduce the effect of the absolute price level, this study used the relative distance between the current close price and the 20-day EMA:

$$TI_t^{(2)} = \frac{C_t}{EMA_{20,t}} - 1$$

This feature is like Simple Moving Average Distance, but it reacts more strongly to recent price changes because EMA gives greater weight to more recent observations. The other technical indicator-based features were also constructed by applying relative or normalized forms to the standard indicators when needed, as summarized in Table 4.

From Candlestick to Numerical Features

A candlestick is a visual price structure formed by the combination of the open, high, low, and close prices over one trading day. From a numerical data perspective, a candlestick is made from four price values. These four price values can also be expressed through relative relationships, including the candlestick body, upper and lower shadows, close location, and opening gap. Therefore, a candlestick can be viewed as a compact graphical representation of the intraday relationships among the same OHLC prices.

Many stock prediction studies use raw OHLC prices or technical indicators as input features. However, raw OHLC prices represent the four prices as absolute levels and do not explicitly encode the relative relationships displayed by candlestick components. To address this gap, this study transformed the structural information of candlesticks into numerical features and compared this representation with raw OHLC prices and technical indicator-based features. Specifically, eight numerical features were defined using the open, high, low, close, and previous close prices. In this study, these features are called numerical candlestick (NC) features.

The NC features were not chosen by randomly listing all possible candlestick-related variables. Instead, they were designed to correspond to the main geometric components of the candlestick representation. First, the relationship between the open and close prices was used to measure intraday direction. Second, the relationship between the high and low prices was used to measure the daily price range. Third, the location of the close price within the daily range was measured. Fourth, the relative sizes of the body, upper shadow, and lower shadow were used to quantify their proportions within the daily price range. In addition, the opening gap from the previous close was included to capture price discontinuity at the start of the trading day.

Table 4. Technical indicator features used as input variables. *The table presents the names and mathematical definitions of the eight technical indicators evaluated in the study: Simple Moving Average Distance, Exponential Moving Average Distance, Normalized Moving Average Convergence Divergence Momentum, Relative Strength Momentum, 10-Day Rate of Change, Commodity Channel Deviation, Bollinger Band Width, and Normalized Average True Range. All indicators were calculated separately for each ETF using only price information available on or before the corresponding prediction date.*

No.	TI Feature Name	Formula
1	Simple Moving Average Distance	$TI_t^{(1)} = \frac{C_t}{SMA_{20,t}} - 1$ $SMA_{n,t} = \frac{1}{n}\sum_{i=0}^{n-1} C_{t-i}$
2	Exponential Moving Average Distance	$TI_t^{(2)} = \frac{C_t}{EMA_{20,t}} - 1$ $EMA_{n,t} = \alpha_n C_t + (1-\alpha_n)EMA_{n,t-1}, \alpha_n = \frac{2}{n+1}$
3	Normalized MACD Momentum	$TI_t^{(3)} = \frac{MACD_t}{C_t} = \frac{EMA_{12,t} - EMA_{26,t}}{C_t}$
4	Relative Strength Momentum	$TI_t^{(4)} = 100 - \frac{100}{1 + RS_t}$ $RS_t = \frac{\bar{G}_{14,t}}{\bar{D}_{14,t} + \epsilon}$ $\bar{G}_{14,t} = \frac{1}{14}\sum_{i=0}^{13} G_{t-i}$ $\bar{D}_{14,t} = \frac{1}{14}\sum_{i=0}^{13} D_{t-i}$ $G_t = max(\Delta C_t, 0), D_t = max(-\Delta C_t, 0)$ $\Delta C_t = C_t - C_{t-1}$
5	10-Day Rate of Change	$TI_t^{(5)} = \frac{C_t}{C_{t-10}} - 1$
6	Commodity Channel Deviation	$TI_t^{(6)} = \frac{TP_t - SMA(TP)_{20,t}}{0.015 \times MAD(TP)_{20,t} + \epsilon}$ $MAD(TP)_{20,t} = \frac{1}{20}\sum_{i=0}^{19} \lvert TP_{t-i} - SMA(TP)_{20,t} \rvert$ $SMA(TP)_{20,t} = \frac{1}{20}\sum_{i=0}^{19} TP_{t-i}$ $TP_t = \frac{H_t + L_t + C_t}{3}$

Continued Table 4. Technical indicator features used as input variables. *The table presents the names and mathematical definitions of the eight technical indicators evaluated in the study: Simple Moving Average Distance, Exponential Moving Average Distance, Normalized Moving Average Convergence Divergence Momentum, Relative Strength Momentum, 10-Day Rate of Change, Commodity Channel Deviation, Bollinger Band Width, and Normalized Average True Range. All indicators were calculated separately for each ETF using only price information available on or before the corresponding prediction date.*

No.	TI Feature Name	Formula
7	Bollinger Band Width	$TI_t^{(7)} = \frac{UpperBB_t - LowerBB_t}{MiddleBB_t + \epsilon}$ $UpperBB_t = MiddleBB_t + 2\sigma_{20,t}$ $LowerBB_t = MiddleBB_t - 2\sigma_{20,t}$ $MiddleBB_t = SMA_{20,t}$ $\sigma_{20,t} = \sqrt{\frac{1}{20}\sum_{i=0}^{19}\left(C_{t-i} - SMA_{20,t}\right)^2}$
8	Normalized Average True Range	$TI_t^{(8)} = \frac{ATR_{14,t}}{C_t}$ $ATR_{14,t} = \frac{1}{14}\sum_{i=0}^{13} TR_{t-i}$ $TR_t = max(H_t - L_t, \lvert H_t - C_{t-1}\rvert, \lvert L_t - C_{t-1}\rvert)$

At time t, let the open, high, low, and close prices be O_t, H_t, L_t, and C_t, respectively. Let the previous close price be C_{t-1}. The total daily price range is defined as:

$$R_t = H_t - L_t$$

Here, R_t represents the total price movement within the trading day. The body, upper shadow, and lower shadow of a candlestick are all defined within this daily range. When $H_t = L_t$, the denominator may become zero, so a small positive constant ϵ was added when necessary. The NC features proposed in this study are presented in Table 5.

Dimensionality Reduction and Prediction Models

This study used ridge regression and random forest regressor to predict the five-day cumulative log return of each ETF. These two models were used to examine whether the relative usefulness of each feature representation was consistent across a linear model and a nonlinear model. In this process, principal component analysis (PCA) was also used to examine whether reducing redundancy within each representation affected prediction performance.

PCA is a dimensionality reduction method that transforms several features that may be correlated with each other into new axes that are not correlated with each other (7). Usually, many features can be compressed into a smaller number of components. That is, PCA does not remove some of the original features, but creates new features called principal components by using linear combinations of the original features.

The basic idea of PCA is to find the direction that explains the largest variance in the data. The first principal component is the direction that explains the largest variance among the original features, and the

Table 5. Numerical candlestick features used as input variables. *The table presents the names, interpretations, and mathematical definitions of the eight numerical candlestick features constructed from daily open, high, low, close, and previous-close prices. The features represent the intraday open-to-close return, high-to-low range, opening gap, closing-price location within the daily range, relative candlestick-body size, signed body size, upper-shadow size, and lower-shadow size.*

No.	NC Feature Name	Description	Formula
1	Intraday Directional Return	Log return from open to close within the same trading day	$NC_t^{(1)} = log\left(\frac{C_t}{O_t}\right)$
2	High-Low Range Return	Log ratio between the daily high and low prices	$NC_t^{(2)} = log\left(\frac{H_t}{L_t}\right)$
3	Opening Gap Return	Log return from the previous close price to the current opening price	$NC_t^{(3)} = log\left(\frac{O_t}{C_{t-1}}\right)$
4	Close Location Ratio	Relative position of the closing price within the daily high-low range	$NC_t^{(4)} = \frac{C_t - L_t}{H_t - L_t + \epsilon}$
5	Body Strength Ratio	Relative size of the candlestick body within the daily high-low range	$NC_t^{(5)} = \frac{\lvert C_t - O_t \rvert}{H_t - L_t + \epsilon}$
6	Signed Body Ratio	Relative body size with price direction	$NC_t^{(6)} = \frac{C_t - O_t}{H_t - L_t + \epsilon}$
7	Upper Shadow Ratio	Relative size of the upper shadow within the daily range	$NC_t^{(7)} = \frac{H_t - max(O_t, C_t)}{H_t - L_t + \epsilon}$
8	Lower Shadow Ratio	Relative size of the lower shadow within the daily range	$NC_t^{(8)} = \frac{min(O_t, C_t) - L_t}{H_t - L_t + \epsilon}$

second principal component is the direction that explains the largest remaining variance while being orthogonal to the first principal component. In this way, PCA compresses several related features into a smaller number of components while minimizing information loss.

In this study, PCA was used because both the technical indicator-based features and the numerical candlestick features are calculated from OHLC price data. Therefore, there is a possibility that the features may contain overlapping information. Such redundancy may make prediction performance unstable in some models or may act as unnecessary noise.

Therefore, this study used PCA to preserve as much information as possible from the original features while reducing redundancy. In this process, the number of components was selected as the minimum number that preserves at least 95% of the cumulative explained variance. PCA was fitted only on the training period of each rolling fold, and the fitted transformation was applied to the test period. This was done to prevent data leakage, where information from the test period is included in the training process.

The prediction models used in this study were ridge regression and random forest regressor. Ridge regression is a model that adds L2 regularization to linear regression (8). If the feature vector of sample i is x_i, and the actual prediction target, the five-day cumulative log return, is y_i, the linear prediction model can be expressed as follows:

$$\hat{y}_t = x_i^T \beta + \beta_0$$

Here, $\hat{y}$ is the predicted value of the model, β is the vector of regression coefficients, and β_0 is the intercept. Ordinary linear regression minimizes the sum of squared errors between the actual and predicted values, but ridge regression adds a penalty for the size of the coefficients.

$$\min_{\beta,\, \beta_0}\left[\sum_{i=1}^{n}(y_i-\beta_0-x_i^T\beta)^2+\lambda\sum_{j=1}^{p}\beta_j^2\right]$$

Here, λ is a hyperparameter that controls the regularization strength. As λ becomes larger, the size of the regression coefficients is more strongly constrained. In this study, λ = 1.0 was used.

Ridge regression limits the size of the coefficients through L2 regularization. Therefore, when multicollinearity exists among features, it can provide more stable predictions than ordinary linear regression. In this study, ridge regression was used because it allows the effect of different feature representations on prediction performance to be interpreted relatively simply.

However, it is difficult to assume that the relationship between stock returns and input features is always linear. Therefore, this study also used random forest regressor as a nonlinear model to evaluate the feature representations under a different model structure. Random forest is an ensemble model that trains multiple decision trees and averages the predictions of each tree to make the final prediction (9). When there are B regression trees, the prediction of random forest can be expressed as follows:

$$\hat{y}_i=\frac{1}{B}\sum_{b=1}^{B}T_b(x_i)$$

Here, $T_b(x_i)$ is the prediction produced by the b-th regression tree for input x_i. In this study, B = 300 was used. The maximum depth of each tree was set to 5, and the minimum number of samples in a leaf node was set to 10 to reduce overfitting. Since financial time-series data contain a large amount of noise, a model that is too complex can easily overfit the training data. Therefore, this study aimed to obtain stable predictions by combining several relatively shallow trees.

Finally, this study compared five input feature sets: OHLC, technical indicator-based features, technical indicator-based features with PCA, numerical candlestick features, and numerical candlestick features with PCA. These feature sets were used to compare raw price information, traditional technical indicator representation, and candlestick-based numerical representation. The PCA versions were included to examine whether reducing redundancy within each feature group improves prediction performance. All feature sets were evaluated using both ridge regression and random forest regressor under the same rolling train-test split structure.

RESULTS AND DISCUSSION

Overall Prediction Performance

This study compared five feature representations: raw OHLC, technical indicators, technical indicators with PCA, numerical candlestick, and numerical candlestick with PCA. Each feature representation was applied to both ridge regression and random forest regressor. The overall prediction performance was calculated by averaging the results across eight ETFs and 12 rolling test years from 2014 to 2025. The overall results are presented in Table 6, and Figure 1, Figure 2, and Figure 3 visualize the prediction performance by feature representation in terms of RMSE, MAE, and directional accuracy, respectively.

As shown in Table 6, numerical candlestick achieved the lowest RMSE in ridge regression. Its RMSE was 0.026390, which was lower than that of OHLC at 0.027326 and technical indicators at 0.026914. For MAE and directional accuracy, numerical candlestick with PCA showed the best results, with values of 0.019818 and 0.549088, respectively. Taken together, the numerical candlestick-based representations recorded the lowest observed RMSE and MAE and the highest observed directional accuracy among the feature representations evaluated with ridge regression.

A similar pattern was observed for the random forest regressor. Numerical candlestick with PCA recorded the best observed performance across all three evaluation metrics among the five feature representations, with an RMSE of 0.026529, an MAE of 0.019913, and a directional accuracy of 0.547787. In contrast, using only OHLC resulted in the highest prediction errors, with an RMSE of 0.028573 and an MAE of 0.022031. The technical indicators representation produced lower observed RMSE and MAE values than OHLC, although both errors remained higher than those obtained with numerical candlestick and numerical candlestick with PCA.

Figure 1 shows the overall comparison based on RMSE. As shown in Figure 1, numerical candlestick achieved the lowest RMSE in ridge regression, while numerical candlestick with PCA achieved the lowest RMSE in random forest regressor. Overall, the candlestick-derived numerical representations produced lower average RMSE values than the raw OHLC and technical indicators representations.

Figure 2 shows the overall comparison based on MAE. The MAE results were broadly consistent with the RMSE results. In both models, OHLC produced the highest MAE, while technical indicators produced a lower MAE than OHLC. However, the lowest MAE was observed in the numerical candlestick feature group. In both ridge regression and random forest regressor, numerical candlestick with PCA achieved the lowest MAE.

Figure 3 presents the overall performance based on

Table 6. Overall out-of-sample prediction performance by model and feature representation. *The table compares raw open, high, low, and close prices (OHLC), eight technical indicators (TI), technical indicators transformed using principal component analysis (TI + PCA), numerical candlestick features (NC), and numerical candlestick features transformed using PCA (NC + PCA). Each feature representation was evaluated using ridge regression and random forest regressor. Lower RMSE and MAE values and higher DA values indicate better prediction performance.*

Model	Feature Set	RMSE	MAE	DA
Ridge regression	OHLC	0.027326	0.020863	0.486014
	TI	0.026914	0.020287	0.520257
	TI + PCA	0.026573	0.019922	0.533297
	NC	**0.026390**	0.019820	0.548111
	NC + PCA	0.026402	**0.019818**	**0.549088**
Random forest regressor	OHLC	0.028573	0.022031	0.513857
	TI	0.027275	0.020453	0.519882
	TI + PCA	0.027128	0.020350	0.521096
	NC	0.026579	0.019947	0.542274
	NC + PCA	**0.026529**	**0.019913**	**0.547787**

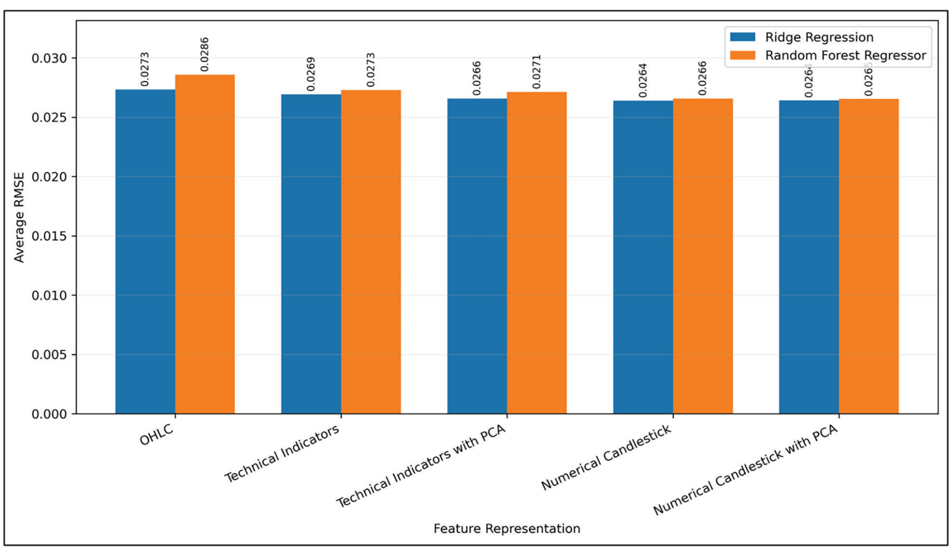

Figure 1. Overall root mean squared error (RMSE) for the five feature representations evaluated using ridge regression and random forest regressor. *The representations include raw open, high, low, and close prices (OHLC), eight technical indicators, technical indicators transformed using principal component analysis (PCA), numerical candlestick features, and numerical candlestick features transformed using PCA. Lower RMSE values indicate better prediction performance.*

directional accuracy. Directional accuracy evaluates whether the model correctly predicts the upward or downward direction of returns, rather than the exact size of the return. As shown in Figure 3, numerical candlestick with PCA recorded the highest directional accuracy among the five feature representations in both models, with 0.549088 in ridge regression and 0.547787 in random forest regressor.

As an additional naive benchmark analysis, the class balance and the directional accuracy of a historical mean forecast were examined. Across the ETF-year test folds, positive, negative, and zero five-day returns accounted for 57.4661%, 42.2938%, and 0.2402% of the observations, respectively. The historical mean

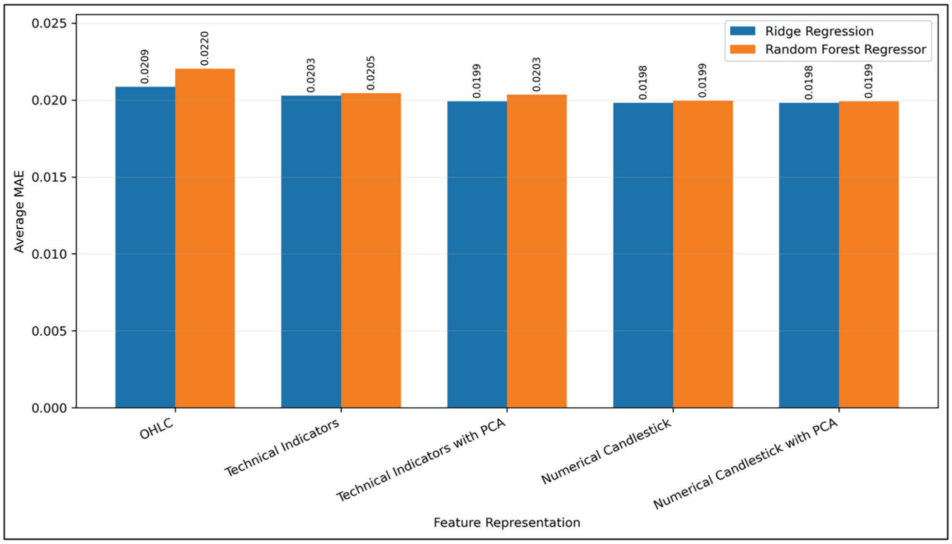

***Figure 2.* Overall mean absolute error (MAE) for the five feature representations evaluated using ridge regression and random forest regressor.** *The representations include raw open, high, low, and close prices (OHLC), eight technical indicators, technical indicators transformed using principal component analysis (PCA), numerical candlestick features, and numerical candlestick features transformed using PCA. Lower MAE values indicate better prediction performance.*

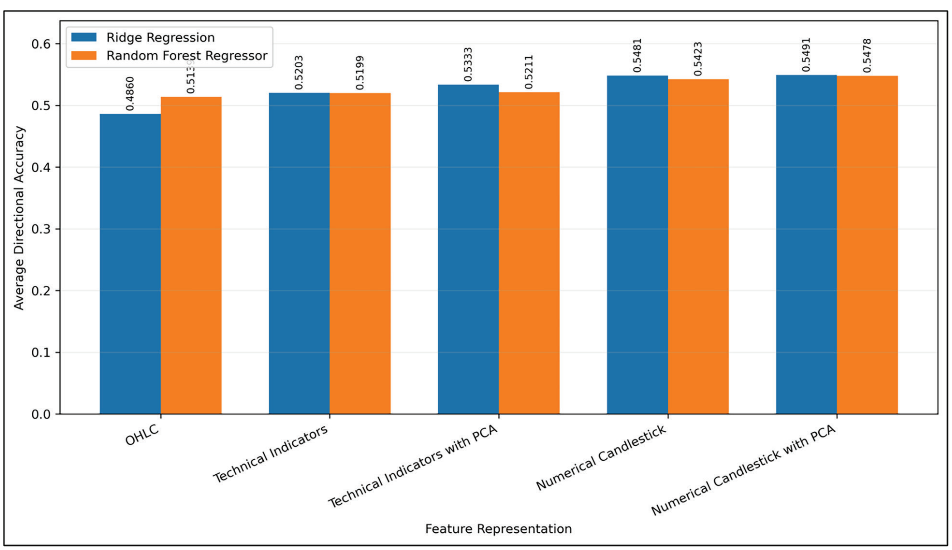

***Figure 3.* Overall directional accuracy for the five feature representations evaluated using ridge regression and random forest regressor.** *The representations include raw open, high, low, and close prices (OHLC), eight technical indicators, technical indicators transformed using principal component analysis (PCA), numerical candlestick features, and numerical candlestick features transformed using PCA. Higher values indicate better directional prediction performance.*

benchmark, which used the mean five-day return from each ETF's corresponding training period as the prediction for the test period, recorded a directional accuracy of 0.568241. This was higher than the directional accuracy of numerical candlestick with PCA in both ridge regression and random forest regressor. Nevertheless, numerical candlestick with PCA recorded higher observed directional accuracy than raw OHLC, which achieved 0.486014 with ridge regression and 0.513857 with random forest regressor. Therefore, although the numerical candlestick representation did not outperform the historical mean benchmark, it provided a higher observed directional accuracy than the raw OHLC representation within both evaluated prediction models.

To account for dependence arising from overlapping five-day cumulative returns, an additional analysis was conducted using a paired moving-block bootstrap with a block length of five trading days. Table 7 reports the observed performance differences between the numerical candlestick-based representations and the OHLC and technical indicator-based representations, together with their 95% bootstrap confidence intervals.

For ridge regression, numerical candlestick recorded lower RMSE and MAE and higher directional accuracy than OHLC and technical indicators. The 95% confidence intervals for these differences did not include zero. When numerical candlestick with PCA was compared with technical indicators with PCA, the confidence interval for the directional accuracy difference did not include zero, whereas those for the RMSE and MAE differences included zero. Thus, after accounting for dependence among overlapping observations, the prediction-error differences between the two PCA-based representations were not clearly distinguishable for ridge regression.

For the random forest regressor, the confidence intervals for all RMSE, MAE, and directional accuracy differences were entirely above zero when numerical candlestick and numerical candlestick with PCA were compared with OHLC and the corresponding technical indicator-based representations. Overall, the block-based analysis suggests that the relatively low observed prediction errors and relatively high observed

Table 7. Paired moving-block bootstrap confidence intervals for differences in prediction performance between feature representations. *NC, TI, and PCA denote numerical candlestick, technical indicators, and principal component analysis, respectively. RMSE, MAE, DA, and CI denote root mean squared error, mean absolute error, directional accuracy, and confidence interval, respectively. For ΔRMSE and ΔMAE, the candidate value was subtracted from the comparator value; for ΔDA, the comparator value was subtracted from the candidate value. Positive values therefore favor the NC-based candidate representation. The 95% CIs were calculated from 10,000 paired moving-block bootstrap replications using five-trading-day blocks.*

Model	Candidate vs. comparator	ΔRMSE [95% CI]	ΔMAE [95% CI]	ΔDA [95% CI]
Ridge regression	NC vs. OHLC	0.000936 [0.000728, 0.001155]	0.001043 [0.000843, 0.001233]	0.062097 [0.040247, 0.082311]
	NC vs. TI	0.000524 [0.000278, 0.000762]	0.000467 [0.000269, 0.000666]	0.027854 [0.012294, 0.043202]
	NC with PCA vs. OHLC	0.000924 [0.000718, 0.001143]	0.001045 [0.000844, 0.001234]	0.063074 [0.041428, 0.082229]
	NC with PCA vs. TI with PCA	0.000171 [−0.000085, 0.000409]	0.000104 [−0.000051, 0.000264]	0.015791 [0.003560, 0.028146]
Random forest regressor	NC vs. OHLC	0.001994 [0.001699, 0.002359]	0.002084 [0.001790, 0.002409]	0.028417 [0.010987, 0.046346]
	NC vs. TI	0.000695 [0.000389, 0.001004]	0.000506 [0.000302, 0.000724]	0.022392 [0.010682, 0.033746]
	NC with PCA vs. OHLC	0.002044 [0.001752, 0.002406]	0.002118 [0.001826, 0.002442]	0.033930 [0.016191, 0.052416]
	NC with PCA vs. TI with PCA	0.000599 [0.000322, 0.000890]	0.000436 [0.000261, 0.000634]	0.026691 [0.014565, 0.038444]

directional accuracy of the numerical candlestick-based representations were retained in most comparisons after accounting for dependence arising from overlapping five-day returns.

Overall, these results show that prediction performance can vary depending on how the same OHLC information is represented. OHLC uses the absolute levels of open, high, low, and close prices directly. In contrast, numerical candlestick transforms daily price movements into log ratios, relative locations within the daily price range, and relative sizes of the candlestick body and shadows. Therefore, numerical candlestick can be viewed as a representation that more directly captures the structural shape of daily price movements rather than the price level itself. These results suggest that the deterministic transformation of OHLC prices into relative and normalized features may provide a more suitable input representation.

Year-by-Year Performance Comparison

Overall average performance summarizes the observed out-of-sample performance of each feature representation. However, it is also necessary to examine whether the observed performance of a representation was concentrated in a few specific years or appeared across multiple test years. For this purpose, this study calculated the average RMSE, MAE, and directional accuracy across the eight ETFs for each test year. The year-by-year results are presented in Figures 4, 5, and 6, where panel A corresponds to ridge regression and panel B corresponds to random forest regressor.

Figure 4 shows the year-by-year RMSE results. In both models, RMSE increased sharply in 2020 compared with other years. This increase coincided with the heightened market volatility associated with the COVID-19 shock in 2020. During this period, the numerical candlestick representations recorded RMSE values comparable to those of the other feature representations. For ridge regression, numerical candlestick or numerical candlestick with PCA showed the lowest or relatively low RMSE in many test years. In particular, the numerical candlestick representations recorded relatively low observed RMSE values in multiple years, including 2015–2017, 2019–2021, and 2023–2025. A similar pattern was observed for the random forest regressor, where numerical candlestick or numerical candlestick with PCA recorded low RMSE in most years. Thus, the observed RMSE pattern was not limited to a single test year but appeared across multiple periods.

Figure 5 shows the year-by-year MAE results. The

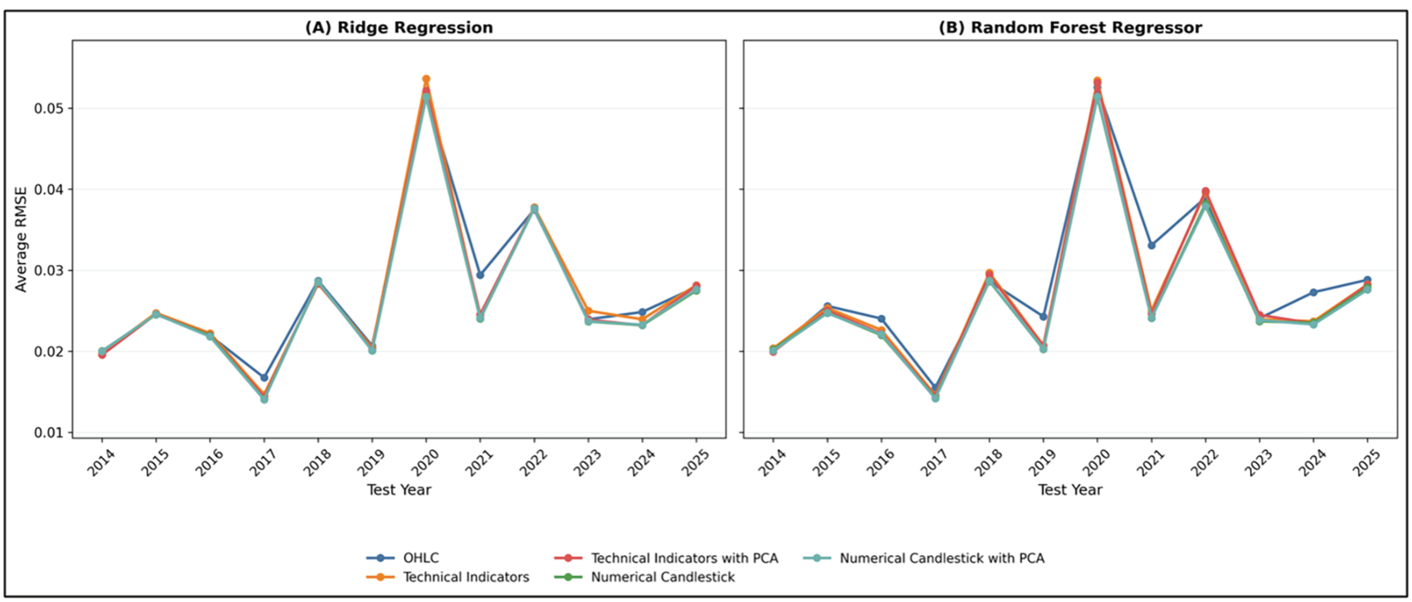

Figure 4. Year-by-year root mean squared error (RMSE) for the five feature representations over the 12 rolling test years from 2014 to 2025. *Panel A presents the results from ridge regression, and Panel B presents the results from random forest regressor. The five lines represent raw open, high, low, and close prices (OHLC), eight technical indicators, technical indicators transformed using principal component analysis (PCA), numerical candlestick features, and numerical candlestick features transformed using PCA. Lower RMSE values indicate better prediction performance.*

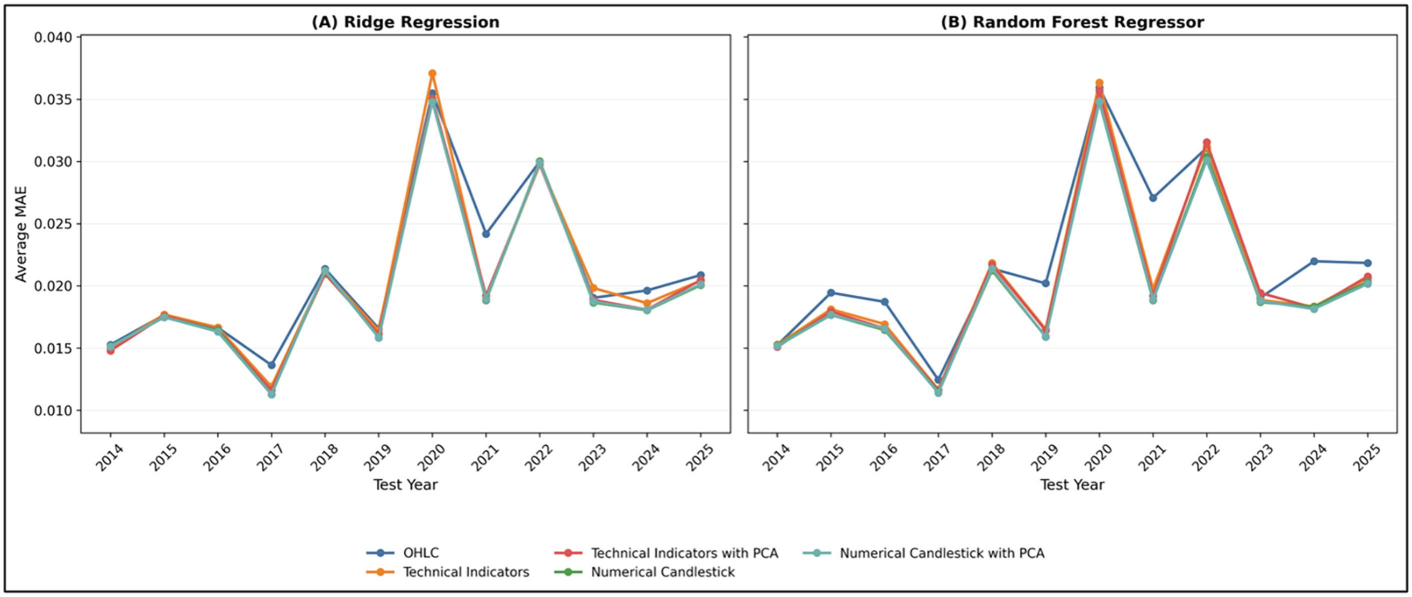

Figure 5. Year-by-year mean absolute error (MAE) for the five feature representations over the 12 rolling test years from 2014 to 2025. *Panel A presents the results from ridge regression, and Panel B presents the results from random forest regressor. The five lines represent raw open, high, low, and close prices (OHLC), eight technical indicators, technical indicators transformed using principal component analysis (PCA), numerical candlestick features, and numerical candlestick features transformed using PCA. Lower MAE values indicate better prediction performance.*

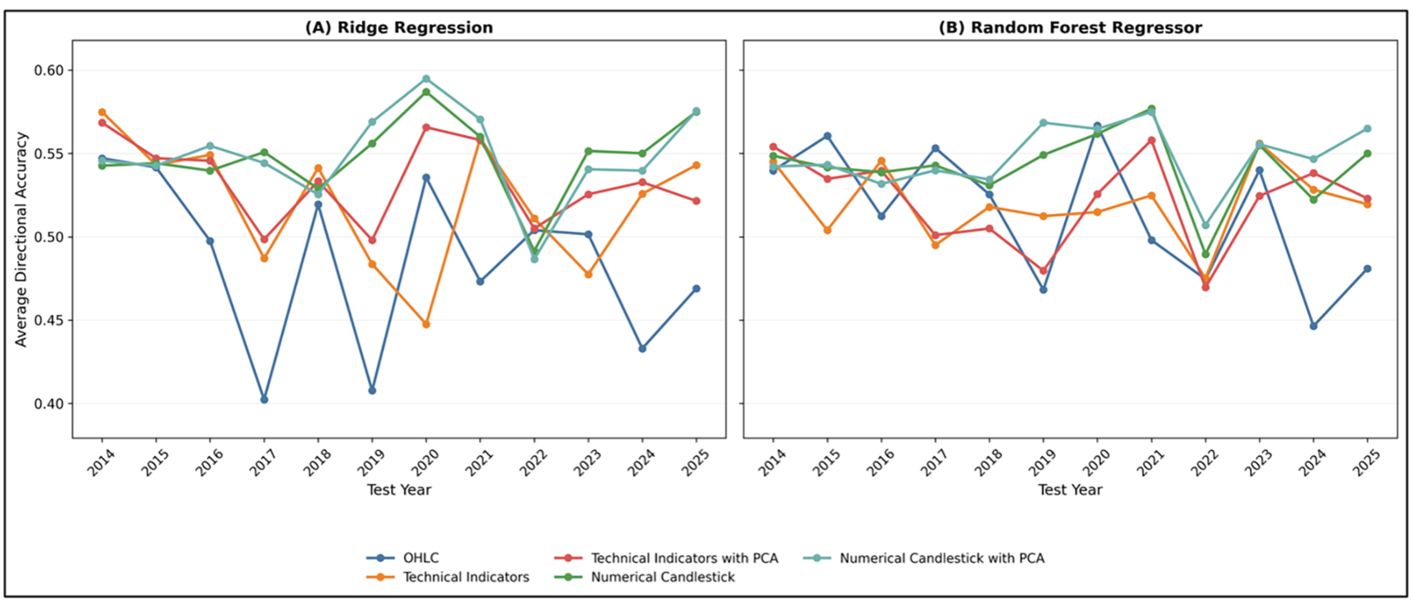

Figure 6. Year-by-year directional accuracy for the five feature representations over the 12 rolling test years from 2014 to 2025. *Panel A presents the results from ridge regression, and Panel B presents the results from random forest regressor. The five lines represent raw open, high, low, and close prices (OHLC), eight technical indicators, technical indicators transformed using principal component analysis (PCA), numerical candlestick features, and numerical candlestick features transformed using PCA. Higher values indicate better directional prediction performance.*

year-by-year MAE results were broadly consistent with the RMSE results. Prediction errors generally increased in 2020 and 2022, and in some years, the differences among feature representations were relatively small. Nevertheless, numerical candlestick and numerical candlestick with PCA often showed lower MAE than OHLC. Thus, the relatively low observed MAE values of the numerical candlestick representations were found across multiple test years rather than in only one specific year.

Figure 6 shows the year-by-year directional accuracy results. Directional accuracy varied more strongly across years than RMSE or MAE. OHLC with ridge regression showed noticeably low directional prediction performance in several years. In contrast, numerical candlestick and numerical candlestick with PCA recorded relatively high observed directional accuracy in multiple years. For random forest regressor, the numerical candlestick representations also recorded relatively high directional accuracy in multiple test years, although their relative performance varied across years.

The year-by-year analysis showed that the numerical candlestick representations recorded relatively low prediction errors and relatively high directional accuracy across multiple test years, rather than only in the aggregate results. This recurring observed pattern suggests that the relative and normalized transformation of OHLC prices warrants further investigation as a feature-engineering approach for ETF return prediction.

Discussion of Main Findings

The main finding of this study is that the numerical candlestick-based representations recorded relatively low prediction errors and relatively high directional accuracy compared with raw OHLC and technical indicators in the tested sample. In ridge regression, numerical candlestick achieved the lowest RMSE, while numerical candlestick with PCA achieved the lowest MAE and the highest directional accuracy. For random forest regressor, numerical candlestick with PCA recorded the lowest observed RMSE and MAE and the highest observed directional accuracy.

The additional naive benchmark analysis places these feature-level comparisons in a broader return-forecasting context. The historical mean benchmark recorded a directional accuracy of 0.568241, which was higher than the values of 0.549088 for ridge regression and 0.547787 for random forest regressor using numerical candlestick with PCA. In addition, because positive five-day returns accounted for 57.4661% of the test observations, an always-positive prediction would have recorded a Directional accuracy of 0.574661. Therefore, the relatively high directional accuracy of the numerical candlestick representations compared with raw OHLC and technical indicators should not be interpreted as evidence that they outperform simple directional benchmarks. Historical mean forecasts are widely used as a stringent no-predictability benchmark in the stock-return forecasting literature, and previous studies have shown that many predictive models have difficulty outperforming this benchmark in out-of-sample evaluation (10, 11). Although the prediction horizons and assets examined in those studies differ from those used here, the benchmark result further limits the interpretation of the present findings to the relative performance of alternative feature representations within the evaluated prediction models.

Because both technical indicators and numerical candlestick consisted of eight features, the observed performance differences cannot be attributed simply to a difference in the number of input features. Instead, the observed differences may reflect how the same OHLC information is represented mathematically for the prediction models.

Technical indicators are useful for summarizing statistical characteristics of past prices, such as moving averages, momentum, and volatility. Technical indicators generally recorded lower observed prediction errors than raw OHLC in this study. However, technical indicators mainly summarize price movements over a certain historical window. In contrast, numerical candlestick directly represents the intraday price structure formed within a single trading day. For example, it includes the relationship between the open and close prices, the range between the high and low prices, the position of the close price within the daily range, the relative sizes of the candlestick body and shadows, and the opening gap relative to the previous close. These numerical candlestick features are calculated directly from the same OHLC prices and describe their relative intraday relationships without introducing additional market information.

The effect of PCA differed depending on the feature type. For technical indicators, applying PCA reduced RMSE and MAE and increased directional accuracy. This suggests that multiple technical indicators may contain overlapping information related to price trends, momentum, and volatility. In contrast, the performance difference before and after PCA was relatively small for numerical candlestick. This may be because numerical candlestick is already constructed using log ratios or ratios relative to the daily price range, which reduces scale-related problems and redundancy. In addition, each numerical candlestick feature represents a different structural component of the candlestick.

From the model perspective, the random forest regressor did not always outperform ridge regression. Random forest regressor has the advantage of learning nonlinear relationships and interactions among features. However, financial return prediction is a noisy problem with weak signals. Therefore, a more complex nonlinear model may overfit unstable patterns in the training period, which may not lead to lower prediction errors in the test period. In contrast, ridge regression uses a simple linear structure with regularization, which may provide more stable performance in a noisy prediction problem such as short-term ETF return forecasting.

Overall, the results suggest that numerical candlestick features provide a more structured representation of existing OHLC information. Although they do not introduce information beyond the underlying OHLC prices, they explicitly encode relational patterns that remain implicit in the raw price levels, including intraday direction, price range, close location, body and shadow proportions, and opening gaps. By making these relative and normalized relationships directly available to the model, numerical candlestick may provide a more suitable input representation for the models evaluated in this study.

CONCLUSION

This study examined whether numerical candlestick features, which express relational structures implicit in OHLC prices as relative and normalized numerical features, provide a suitable input representation for predicting short-term ETF returns. To examine this question, five feature representations were compared: raw OHLC, technical indicators, technical indicators with PCA, numerical candlestick, and numerical candlestick with PCA. Ridge regression and random forest regressor were used to predict the cumulative log return over the next five trading days. The results showed that the numerical candlestick-based representations generally recorded lower observed RMSE and MAE values than the raw OHLC and technical indicators representations. Numerical candlestick with PCA also recorded the highest observed directional accuracy among the five feature representations in both prediction models.

One possible explanation is that, although numerical candlestick features do not add new market information, they reorganize the existing OHLC information into a more structured form for prediction. OHLC uses the absolute values of open, high, low, and close prices. However, because ETFs have different price levels and different ranges of price movement, using absolute prices alone may make it difficult for the model to compare price structures across different assets. In contrast, numerical candlestick features express the relationship between the open and close prices, the daily price range, the relative position of the close price, the ratios of the body and shadows, and the opening gap from the previous close in the form of log ratios or range-based ratios. This transformation reduces the effect of price-level differences among ETFs and allows the daily price structure to be compared from a more relative and normalized perspective. By explicitly encoding these relative and normalized relationships, numerical candlestick features may allow the models to use the existing OHLC information more directly than when using raw price levels.

From a feature-engineering perspective, numerical candlestick features make relational patterns that remain implicit in raw OHLC price levels explicitly available to prediction models. These patterns include intraday direction, price range, close location, body and shadow proportions, and opening gaps. Although these features do not add information beyond the underlying OHLC prices, their relative and normalized form may provide a more suitable input representation for the models.

However, this study also has several limitations. First, the experiment used only eight ETFs. Therefore, further research is needed to examine whether similar results can be found in individual stocks, ETFs from other countries, commodity ETFs, bond ETFs, or other asset classes. Second, the prediction models used in this study were limited to ridge regression and random forest regressor. These models are useful for comparing the basic effect of different feature representations, but they may not fully capture long-term dependencies or complex nonlinear patterns in time-series data. Future studies could extend this approach to neural network-based time-series models, such as LSTM, GRU, or Transformer, to test whether numerical candlestick features are also useful in more complex models.

Third, this study represented the candlestick structure using eight features, but additional OHLC-derived relationships may be considered beyond the eight features examined in this study. For example, future studies could consider the sequence of candlesticks over multiple trading days, changes in the body and shadow sizes, the persistence of gaps, recent changes in the closing position, or numerical versions of specific candlestick patterns. Future research could also increase the number of numerical candlestick features and examine which combination is most helpful for predicting returns based on OHLC information. Feature selection methods, PCA, or LASSO could be used to identify the most important candlestick-based features among a larger set of candidates.

Finally, this study focused on predictive performance and did not evaluate whether the observed prediction differences would lead to economically meaningful trading outcomes. Accordingly, the economic significance of the observed prediction differences was not established in this study. Before drawing conclusions about practical investment performance, future studies should evaluate explicit portfolio construction and rebalancing rules, portfolio turnover, transaction costs, bid–ask spreads, slippage, and risk-adjusted returns through realistic out-of-sample back-testing.

CONFLICT OF INTEREST

The author declares that there are no conflicts of interest related to this work.

REFERENCES

1. Lo AW, Mamaysky H, Wang J. Foundations of technical analysis: computational algorithms, statistical inference, and empirical implementation. *J Finance.* 2000; 55 (4): 1705–1765. doi:10.1111/0022-1082.00265.
2. Mostafavi SM, Hooman AR. Key technical indicators for stock market prediction. *Mach Learn Appl*. 2025; 20: 100631. doi:10.1016/j.mlwa.2025.100631.
3. Shynkevich Y, McGinnity TM, Coleman SA, Belatreche A, Li Y. Forecasting price movements using technical indicators: investigating the impact of varying input window length. *Neurocomputing*. 2017; 264: 71–88. doi:10.1016/j.neucom.2016.11.095.
4. Moodi F, Jahangard-Rafsanjani A. Integrating feature selection and regression methods with technical indicators for predicting Apple Inc. stock prices (Preprint). arXiv. 2025 Oct 15. arXiv:2310.09903v5 (q-fin.ST). doi:10.48550/arXiv.2310.09903. Available from: https://arxiv.org/abs/2310.09903v5 (accessed on 2026-04-26).
5. Lu TH, Shiu YM, Liu TC. Profitable candlestick trading strategies—The evidence from a new perspective. *Rev Financ Econ*. 2012; 21 (2): 63–68. doi:10.1016/j.rfe.2012.02.001.
6. Chen Z, Huang W. Can candlestick charts help predict stock price movement? A machine learning approach. SSRN (Preprint). 2024 Oct 26. SSRN 5000007. doi:10.2139/ssrn.5000007. Available from: https://ssrn.com/abstract=5000007 (accessed on 2026-04-13).
7. Jolliffe IT. “principal component analysis.” 2nd ed. Springer New York, 2002. ISBN: 978-0-387-95442-4.
8. Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*. 1970; 12 (1): 55–67. doi:10.2307/1267351.
9. Breiman L. Random forests. *Mach Learn*. 2001; 45 (1): 5–32. doi:10.1023/A:1010933404324.
10. Welch I, Goyal A. A comprehensive look at the empirical performance of equity premium prediction. *Rev Financ Stud.* 2008 Jul; 21 (4): 1455–1508. doi:10.1093/rfs/hhm014.
11. Campbell JY, Thompson SB. Predicting excess stock returns out of sample: can anything beat the historical average? *Rev Financ Stud*. 2008 Jul; 21 (4): 1509–1531. doi:10.1093/rfs/hhm055.