Data Science in Finance and Economics

DSFE, 6(3): 359–396.
DOI: 10.3934/DSFE.2026013
Received: 16 December 2025
Revised: 21 May 2026
Accepted: 31 May 2026
Published: 01 July 2026

https://www.aimspress.com/journal/dsfe

Research article

Bias-Corrected feature selection for financial time series forecasting: a large-scale evaluation across 40 global assets

David Jukl[1,*] and Eva Daniela Cvik[2]

[1] University of Finance and Administration, Estonská 500/3, 101 00 Prague 10, Czech Republic
[2] University of Agriculture, Faculty of Economics and Management, Department of Law, Kamýcká 129, Praha 6 – Suchdol, Czech Republic

* **Correspondence:** Email: jukl@mail.vsfs.cz.

Abstract: In financial forecasting, model selection has become a crucial problem because large-dimensional spaces of technical indicators need to be analyzed using meaningful signals in return series that are noisy and weakly predictive. Algorithms such as the feature selection algorithm (FSA), which apply simulated annealing to find sparse predictive subsets, have demonstrated good performance but have a methodological weakness: they habitually predict positive price changes by approximately 54% rather than a neutral 50/50 benchmark. This directional bias can distort long–short strategies, inflate hidden risk exposures, and undermine the reliability of model-driven trading systems. Although previous studies have shown the advantages of FSA in terms of sparsity and accuracy, its distributional balance in predictions has not yet been addressed, posing an unmet need in financial machine learning. To address this drawback, this paper presents the bias-corrected feature selection algorithm (BFSA), which extends the FSA loss to impose a direct penalty for deviations from a neutral benchmark of 0.5. The robustness of the penalty (λ_bias) was optimized using asset-specific five-fold chronological cross-validation, which enables BFSA to control its bias correction across varying volatility regimes and structural properties. BFSA was tested on daily Open, High, Low, Close, Volume (OHLCV) data for 40 assets tracking developed and emerging equity indices, major US-based equities, sector ETFs, commodities, and crypto markets, spanning over a decade of market history. BFSA achieved an average rank of 4.20, representing the strongest performance among all non-trivial predictive models. Only the Null model (1.48) and LASSO (1.85) obtained lower ranks, reflecting their simplicity rather than genuine predictive structure. BFSA outperformed the original FSA (5.55), all tree-based ensembles (ranks 9–12), and all deep learning hybrids (ranks 15–25), demonstrating that explicit bias control materially improves generalization in noisy financial settings. Cross-validated λ

tuning improved performance in approximately 48% of assets and slightly degraded it in 52%, resulting in a near-zero net change in MSE. This indicates that the default λ_bias = 30 already has a robust general setting, while cross-validation primarily safeguards against severe hyperparameter misspecification in high-volatility assets. The wide dispersion in optimal λ values (4–135) highlights the structural heterogeneity of directional bias across asset classes. BFSA materially reduced the upward directional skew of FSA (mean prediction 0.537) and produced nearly balanced outputs (0.503), substantially improving calibration for long–short and market-neutral strategies.

Keywords: feature selection; bias correction; financial forecasting; simulated annealing; cross-validation; technical indicators

JEL Codes: C45, C53, G11, G17

1. Introduction

1.1. Background

Financial forecasting remains a difficult application of machine learning because asset returns are noisy, weakly predictive, nonstationary, and subject to regime shifts and structural breaks. These properties are especially restrictive in daily forecasting, where models often rely on only a few thousand observations while being exposed to hundreds or thousands of candidate predictors. In this setting, increasing model complexity does not necessarily improve generalization; instead, robust performance depends on feature parsimony, regularization, and careful search through high-dimensional technical-indicator spaces.

Feature selection (FS) is therefore central to financial prediction pipelines. Technical indicators based on price, momentum, volatility, volume, patterns, and lagged transformations can generate very high-dimensional predictor spaces. The role of FS is to identify compact and stable subsets of predictive signals that remain useful across changing market conditions (Jukl & Lansky, 2025). Traditional approaches such as LASSO, elastic net, random forest importance, and Boruta have been widely applied (Kim et al., 2024; Sheng et al., 2025). However, these methods face different limitations: shrinkage methods impose restrictive linear assumptions, tree-based methods can be unstable under correlated predictors, and heuristic importance scores may be sensitive to regime changes. Wrapper methods are more flexible because they evaluate candidate feature subsets through the predictive model itself, but they are computationally more demanding.

In financial direction prediction, calibration and class balance are not merely secondary diagnostics. When probabilistic forecasts are converted into trading signals, the distribution of predicted upward and downward movements directly affects portfolio exposure, turnover, hedging behavior, and risk concentration. A model may therefore appear acceptable under conventional accuracy or loss metrics while still producing economically undesirable signals if its predictions are systematically skewed toward one direction. This makes directional balance a core aspect of the economic validity of forecasting models, particularly in long–short, market-neutral, and risk-managed strategies.

A more recent addition to the FS literature is the feature selection algorithm (FSA), proposed by Pabuccu and Barbu (2024). FSA is a stochastic wrapper method that uses simulated annealing to search

for sparse feature subsets that minimize classification loss with an L2-regularization penalty. Its main advantage is that it can explore large combinatorial spaces and identify compact predictor sets, which is useful in financial settings where autocorrelations are unstable, and noise amplification is a persistent risk. However, because the original FSA objective optimizes classification loss without explicitly constraining the distribution of predicted classes, it may select feature subsets that appear accurate while producing systematically imbalanced directional predictions.

1.2. Problem: Directional bias in FSA

In this study, directional bias refers to the systematic deviation of a model's mean predicted probability, or predicted class frequency, from an empirical or target directional balance. In a binary next-day forecasting task, this means that a model repeatedly predicts upward movements more often than downward movements, or vice versa, even when the forecasting objective is intended to remain directionally neutral. Directional bias is therefore different from random forecast error: it represents a persistent skew in the model's output distribution.

Although FSA optimizes classification performance, it does not directly account for class-balancing. In next-day financial direction prediction, a neutral benchmark of 0.5 is useful because daily asset returns are typically weakly predictable and often balanced over long samples. This does not imply that every asset or every subperiod has a true 50/50 distribution. Rather, 0.5 is used here as a modeling target for avoiding systematic directional exposure when the model is intended to generate balanced probabilistic signals. Without such a constraint, FSA may produce prediction ratios such as 54% upward and 46% downward movements, not necessarily because the market contains stable directional information but because the optimization process can exploit small majority-class advantages in noisy data.

The implications of this directional bias are economically important. Balanced signals are particularly relevant for exposure management in long–short portfolios, pairs trading, statistical arbitrage, and market-neutral strategies. A model that predicts upward movements more frequently than downward movements can introduce persistent long exposure, even when the strategy is intended to be hedged or directionally neutral. Such skew may also lead to misleading performance in trending markets and weaker performance in mean-reverting regimes (Jukl & Lansky, 2025). More generally, output imbalance can degrade calibration and robustness in machine learning prediction systems (Bloch, 2025). Existing bias-correction methods, such as resampling and cost-sensitive learning, typically operate at the data or classifier level but do not directly control bias during feature subset selection.

This structural problem is visible in the empirical results of this study. Across 40 global assets, the original FSA produces an average upward prediction rate of 0.537, compared with the neutral benchmark of 0.5. Although this deviation may appear modest, it can materially affect portfolio exposure when forecasts are translated into repeated trading signals across assets and time. The bias is also not confined to a single market segment; it appears across volatile assets, such as cryptocurrencies, as well as equity-related instruments. These findings motivate the need for a feature-selection objective that simultaneously considers predictive performance, sparsity, and directional balance.

1.3. Research gap

Although feature selection and bias correction are both well-established areas of machine learning research, they are rarely integrated into a single optimization objective. Existing bias-reduction methods can be grouped into three broad categories. First, post-hoc methods, such as threshold adjustment, modify predictions after model training. Second, data-level methods, such as resampling or reweighting, alter the training distribution. Third, classifier-level methods, such as cost-sensitive learning, modify the loss assigned to different classes. These approaches can improve output balance, but they do not directly address the feature-selection stage, where biased predictive structure may already have been introduced (Wongvorachan et al., 2024).

This omission is especially important for wrapper-based FS algorithms, including genetic algorithms, simulated annealing methods, and reinforcement-learning-based feature selection. These methods search for feature subsets that optimize predictive performance, but they typically do not constrain the distributional properties of the resulting predictions. In financial forecasting, this matters because directional balance is not only a statistical fairness issue; it is also a condition for controlled exposure in risk-managed trading systems (Zhang et al., 2024). BFSA addresses this gap by shifting bias control upstream into feature subset selection. Instead of correcting skew only after the model has been trained, BFSA discourages selecting predictors that generate systematically imbalanced directional probabilities while preserving the sparsity and search flexibility of FSA.

1.4. Research objectives

This paper proposes the bias-corrected feature selection algorithm (BFSA), a generalization of FSA that incorporates a directional-balance penalty into the feature-selection loss. Four research questions guide the study:

RQ1: Does adding a directional-balance penalty to the FSA loss reduce prediction imbalance without materially degrading forecast accuracy?

RQ2: Does chronological cross-validation improve the selection of the bias-penalty parameter relative to a fixed global value?

RQ3: How does BFSA compare with feature-selection, machine-learning, and deep-learning benchmarks across heterogeneous asset classes?

RQ4: How does the optimal bias-penalty parameter vary across asset classes and volatility regimes?

Together, these questions connect the methodological development of BFSA with empirical testing of accuracy, robustness, sparsity, and directional calibration across heterogeneous financial markets.

1.5. Contributions

There are six important contributions of this paper:

Theoretical contribution: The paper formalizes directional balance as an output-distribution property that can be incorporated directly into a feature-selection objective. The proposed quadratic penalty is smooth, differentiable, and symmetric around the target probability of 0.5. It increases with the squared distance between the model's mean predicted probability and the neutral benchmark, thereby discouraging systematic directional skew without imposing a hard constraint. In this sense, the

penalty works analogously to regularization: it does not force a fixed prediction pattern. Still, it penalizes undesirable model behavior when that behavior is not justified by sufficient predictive gain.

Methodological contribution: The paper introduces BFSA as an extension of the original FSA. BFSA adds a directional-balance penalty to the FSA objective, requiring selected feature subsets to simultaneously satisfy three objectives: classification performance, sparsity, and distributional balance. This modification preserves simulated annealing's stochastic global search capability while reducing FSA's tendency to select predictors that generate systematic directional skew.

Empirical contribution: A large-scale evaluation across 40 assets and 28 benchmark models (1120 model–asset experiments) shows that the original FSA exhibits a consistent upward directional bias, with a mean prediction rate of 0.537, whereas BFSA produces nearly balanced outputs, with a mean prediction rate of 0.503. BFSA also improves predictive performance relative to the original FSA, achieving an average rank of 4.20 compared with FSA's 5.55. Among models that learn from the technical-indicator feature space while controlling directional skew, BFSA is the strongest non-trivial feature-selection model evaluated. The Null model and LASSO obtain lower average ranks. Still, this result should be interpreted as evidence of the weak-signal nature of daily return forecasting rather than as evidence against the value of bias-controlled feature selection. These findings are consistent with prior evidence that sparse and strongly regularized models often generalize better than high-capacity architectures in noisy financial environments (Makridakis et al., 2018; Gu et al., 2020).

Practical contribution: The results show that the optimal bias-penalty parameter, λ_bias, varies substantially across asset classes, ranging from approximately 4 to 135. Cryptocurrencies require stronger correction because their volatility and structural asymmetry increase the risk of skewed probabilistic forecasts, while large-cap equities and indices generally require weaker penalties. The observed relationship between λ_bias and annualized volatility provides practical guidance for applying bias-controlled feature selection across heterogeneous markets. More importantly, BFSA improves the suitability of feature-selected signals for long–short, market-neutral, and hedged strategies by reducing unintended directional exposure.

Boundary-condition contribution: The study also identifies an important boundary condition for financial machine learning. In extremely noisy daily return forecasting, simple baselines such as the Null model and heavily regularized LASSO can still dominate on pure MSE or average-rank criteria. BFSA should therefore not be interpreted as evidence that technical indicators generate large or easily exploitable alpha. Its contribution is more specific: it improves the original FSA by reducing directional skew while preserving competitive generalization among feature-selection and machine-learning alternatives.

Reproducibility contribution: The study follows rigorous standards of transparency: all experiments use fixed random seeds, use chronological K-fold cross-validation to avoid look-ahead bias, and use publicly available Yahoo Finance data. All code, indicator-generation procedures, and evaluation scripts are prepared for release in a reproducible pipeline. Together, these practices ensure that BFSA can be independently validated and extended.

Collectively, these contributions position BFSA as a bias-aware extension of FSA that addresses an important limitation in feature-selection-based financial forecasting: the lack of an explicit mechanism to control directional output imbalance during feature subset selection.

1.6. Paper organization

The rest of the paper is organized as follows: Section 2 examines the literature on feature selection, bias in machine learning, cross-validation procedures, and ML/DL techniques for financial forecasting. Section 3 presents the complete methodology, including the formal definition of BFSA, the term bias-correction, and the cross-validated procedure for optimizing λ. Section 4 presents empirical findings for all assets and models, with ranking charts, heatmaps, λ distributions, and statistical tests integrated. Section 5 explains the implications of the results, correlates model behavior with financial theory and the ML literature, and describes and elucidates potential practices and limitations. Section 6 concludes with a summary of contributions, a discussion of the significance of bias-conscious FS, and directions for future research.

2. Literature review

2.1. Feature selection in financial forecasting

The role of feature selection has long been central to financial forecasting as high-dimensional predictor spaces are widespread, financial markets are structurally unstable, and asset return series have a notoriously poor signal-to-noise ratio. In econometrics and machine learning, the need to identify sparse, strong, and interpretable feature subsets has been highlighted repeatedly, especially given that modern financial data often includes hundreds or even thousands of technical indicators, lagged transformations, volatility indicators, volume trends, and cross-asset information. The application of traditional linear regularization methods like LASSO and elastic net has become a standard tool for shrinkage and selection; however, they are limited in their ability to model nonlinear interactions and to adapt to the heavy-tailed, regime-dependent nature of financial time series (Nkansah, 2017). Though LASSO is still popular for modeling macro-finance and factor-based models in general, because it is more stable and easier to interpret, it does not work well with predictors with high multicollinearity and when signals occur in complex, nonlinear forms.

Beyond its conventional role in improving accuracy and reducing dimensionality, feature selection also shapes the distribution of signals produced by a forecasting model. The selected predictors determine not only the classifier's error rate but also the tendency of the model to assign probabilities to one class more often than the other. This distinction is important in financial forecasting because a feature subset that appears useful under a conventional loss function may still generate economically undesirable output distributions. In this sense, feature selection should be understood not only as a variable-reduction procedure but also as a mechanism that influences calibration, directional exposure, and the practical behavior of trading signals.

The literature has attempted to overcome these drawbacks by discussing other FS methods that apply tree-based and heuristic search procedures. Techniques like Boruta, based on the importance of the installed random forest, offer a more adaptable framework, since they can accommodate nonlinear interactions; however, they are generally unstable, like the models they are built upon, especially when including noisy and fast-changing financial predictors (Elmousalami, 2019). Genetic algorithms and evolutionary wrappers can be more intelligent in navigating high-dimensional search spaces (Taha et al., 2025). However, their performance can be strongly dependent on the choice of initialization,

mutation heuristics, and convergence parameters, and large computational budgets are frequently needed to prevent premature convergence to suboptimal feature subsets.

Generally, wrapper methods have received considerable attention in financial applications because they optimize predictive performance directly rather than relying on proxy statistics such as coefficients, correlations, or importance scores. In this respect, the FSA, proposed by Pabuccu and Barbu (2024), is a valuable contribution, as it uses simulated annealing as a global optimization strategy to escape local minima and explore sparse feature subsets. Simulated annealing-based methods are especially relevant in small-sample and high-dimensional settings because they introduce controlled randomness into the search process, allowing the algorithm to identify feature configurations that may be missed by linear, greedy, or purely deterministic procedures (Celik & Dal, 2021).

Such an omission is especially important in finance, since when optimizing noisy classification tasks, directional imbalance can easily form. A predictor that predicts more regularly than otherwise can be more accurate in trending markets, thereby deceiving wrapper-based FS searches into believing there is a real predictive gain beyond confounded artifactual class imbalance (Holzinger et al., 2022). The fact that FS methods such as FSA, genetic algorithms, Boruta, and traditional shrinkage have no explicit distributional constraints is a substantial gap in the literature, particularly when accurate predictions of balanced outcomes are needed, including for long–short strategies, statistical arbitrage, and risk-managed portfolios.

This limitation means that wrapper methods may unintentionally select features that exploit directional imbalance rather than genuinely predictive market structure. Because they optimize classification loss directly, they can reward a feature subset that slightly improves apparent accuracy by shifting the predicted distribution toward the more frequently rewarded direction. Without an explicit distributional constraint, such a subset may be retained even when it creates systematic upward or downward skew. This provides the immediate methodological motivation for BFSA: the feature-selection objective must evaluate not only whether a subset predicts well but also whether it produces directionally balanced probabilistic outputs.

2.2. Bias in machine learning

In machine learning, prediction bias broadly refers to systematic model behavior that favors one class, direction, or outcome. Much of the existing literature examines bias related to protected-class fairness, such as demographic or healthcare-related disparities. The present study does not examine protected-class fairness. Instead, it focuses on financial prediction bias, defined as distributional output imbalance in a binary forecasting task. In this setting, bias occurs when the model repeatedly assigns a higher probability to upward or downward movements relative to the empirical or target directional balance. This form of bias is usually caused by the interaction of noisy targets, small sample sizes, class imbalance, and optimization procedures that reward majority-class behavior (de Castro Vieira et al., 2025). Prior work on class imbalance shows that classifiers can reduce loss while overpredicting the majority class, thereby weakening calibration and increasing systematic error. Methods, such as the synthetic minority over-sampling technique (SMOTE), cost-sensitive weighting, and threshold adjustment, can mitigate imbalance. However, they usually operate at the data, classifier, or post-processing level rather than at the feature-selection stage (Murithi, 2024).

Additionally, studies of distributional justice have emphasized the need to explicitly limit the distributions discussed so that models can meet balance properties. However, no fairness-based

constraints have been identified for FS algorithms, and their use in the demographic context is mostly restricted (Silva, 2021). What is still missing in the literature is a technique that can effectively perform feature selection while tuning for prediction bias. This aspect of non-integration between FS and bias correction is especially problematic in financial markets, where directional imbalance contributes to systemic risks: an FS approach that identifies features strongly linked to movements toward the upside enforces too much long exposure and creates undesirable risk concentration, also leading to diminished feasibility of market-neutral strategies (Peterson & He, 2025).

Post-hoc thresholding can adjust the final class frequencies, but it does not solve the feature-selection problem itself. Once a biased feature subset has been selected, the model has already been encouraged to rely on predictors that generate skewed probability distributions. Threshold adjustment may mask this imbalance at the decision boundary, yet the underlying probability estimates and selected feature structure remain biased. For this reason, correcting directional imbalance only after training is insufficient for the present research problem. Bias control must be incorporated earlier, at the point where feature subsets are evaluated and retained.

Because FS techniques influence the hypothesis space available to the predictive model, bias introduced during feature selection can propagate through the entire forecasting pipeline. FSA illustrates this issue: its stochastic search procedure is effective at finding compact feature groups, but it can also favor subsets that slightly improve classification loss while producing skewed prediction distributions (Piles et al., 2021). In the empirical analysis of this paper, FSA predicts upward movements approximately 54% of the time across 40 assets, exceeding the neutral 50% benchmark used for directionally balanced forecasting. Calibration research suggests that such persistent output skew can impair risk-adjusted performance, reduce regime stability, and weaken model behavior during market transitions (Gu et al., 2020). These limitations support the development of FS methods that incorporate bias-correction mechanisms directly into the loss function.

2.3. Hyperparameter optimization

Hyperparameter optimization is a critical but often underspecified aspect of financial forecasting methodology. Whereas random splits, grid search, or repeated K-fold cross-validation are common in other areas, the structure of financial time series makes these methods inapplicable due to serial correlation, regime changes, and the risk of look-ahead bias.

The time-series forecasting literature has repeatedly cautioned against data shuffling and conventional K-fold techniques, as information leakage can greatly overstate apparent performance but does not extrapolate to unknown times (Hyndman & Athanasopoulos, 2018). Therefore, chronological cross-validation, e.g., expanding-window K-fold or sliding-window K-fold, should be used to preserve temporal order and ensure that the test set never precedes the training set.

In addition to eliminating look-ahead bias, cross-validation is essential for hyperparameter tuning, which is highly dataset-dependent. Regularization strength, model depth, dropout rates, and feature-selection penalties vary widely across assets, as each asset carries inherent structural volatility, liquidity, and behavioral tendencies (Jukl & Lansky, 2025). Although global optima are uncommon to cross-instrument heterogeneous noise structure, Mayada's (2025) study made it clear that hyperparameters of financial ML need to be optimized on an asset-by-asset basis, instead of being optimized on a global scale (Mayada, 2025). This argument is strongly supported by the empirical data from the present study:

the most appropriate value of λ bias associated with BFSA differs, and all stable indices range from 4 to 140, depending on the volatility of a particular cryptocurrency.

For BFSA, chronological cross-validation is therefore included not only to improve average predictive accuracy but also to reduce the risk of severe bias-penalty misspecification. The optimal value of λ_bias may differ sharply across liquid equity indices, individual stocks, commodities, and cryptocurrencies because these assets exhibit different volatility levels, class-balance properties, and regime-switching behavior. A fixed value may be adequate for many assets, but it can under-correct directional skew in highly volatile assets or over-constrain the model in more stable markets. Asset-specific tuning is therefore best understood as a protective calibration step rather than a guaranteed source of performance improvement.

The literature also demonstrates that severe overfitting results from inappropriate hyperparameter tuning. Applying deep learning forecasting, Gu et al. (2020) showed that models like LSTM and CNN trained on financial data that do not have more than a few thousand observations need to be regularized and carefully tuned on their architectures because they tend to overfit noise and decline sharply out of sample (Gu et al., 2020). The same findings are observed in the M4 forecasting competition (Makridakis et al., 2018), where complex models prove ineffective compared to simple models when hyperparameter search is insufficient, and the data is unstable. This further confirms the importance of the demanding, chronological K-fold cross-validation structure deployed in BFSA, which not only tunes λ_bias on a per-asset basis but also normalizes performance based on a coarse-to-fine search across individual volatility profiles.

This interpretation is important because the empirical findings show that cross-validated λ tuning improves MSE in approximately 48% of assets and slightly worsens it in approximately 52%. The near-zero average change does not invalidate the tuning procedure, but it qualifies its role. Cross-validation should not be presented as uniformly performance-enhancing; rather, it functions as a safeguard against poorly chosen penalty values, especially in assets with unusual volatility, unstable directional structure, or strong regime dependence.

2.4. Machine learning and deep learning forecasting

Financial forecasting has been a highly studied area of machine learning; the foundation of many such pipelines is random forests, XGBoost, support vector machines, and k-nearest neighbors (Corvers, 2025). Ensemble models, especially RF and XGB, have gained popularity because of their insensitivity to noise and their ability to capture nonlinear relationships; nonetheless, they can be unstable in time-series settings because they utilize bootstrapping and greedy tree-breaking heuristics, which are highly sensitive to minor changes in the training data. Research has demonstrated that these models tend to overfit short-term changes but not long-term cycles, particularly in predicting returns rather than levels (Jukl & Lansky, 2025).

Depending on the architecture an individual may adopt, such as ANN, CNN, or LSTM models, deep learning (DL) can extract higher-order temporal dependencies and nonlinearities. However, recent empirical studies in the financial world have yielded very mixed results. Simple exponential smoothing techniques were superior to complex DL models across thousands of tasks of factor modeling in the M4 competition, but the latter are susceptible to instability in small-time-series forecasting with very limited historical information (Makridakis et al., 2018; Gu et al., 2020). This is further complicated by the fact that DL models need large training samples to generalize, and financial

data is often short, non-repetitive, and noisy.

Within the present study, BFSA is evaluated in two related but distinct ways. First, it is assessed as a direct sparse probabilistic forecasting framework, in which the selected features are fed into a logistic/probabilistic classifier. Second, it is used as a pre-selection layer in hybrid models, where BFSA-selected features are passed to downstream learners such as XGBoost, ANN, CNN, or LSTM variants. This distinction is important because the core BFSA loss is most naturally aligned with models that produce calibrated probabilities. In contrast, hybrid models test whether BFSA-selected variables improve more complex predictive architectures. Accordingly, the present results should not be interpreted as full validation of BFSA for all nonlinear or non-probabilistic predictive engines; further work is required to test alternative learners, probability-calibration methods, and architectures specifically designed around nonlinear feature interactions.

The empirical findings of this paper are consistent with the broader literature: sparse, strongly regularized methods remain difficult to outperform in noisy, daily financial prediction tasks. Across 40 assets and 28 models, several deep-learning hybrids, including BFSA-ANN and CNN variants, rank below simpler models. LASSO obtains an average rank of 1.85, while BFSA obtains an average rank of 4.20, supporting the view that simple models often generalize better than complex architectures in volatile, low-signal settings (Makridakis et al., 2018). However, this result should be interpreted carefully. It does not imply that nonlinear models are generally unsuitable for finance; rather, it suggests that, given the current data frequency, sample size, feature set, and evaluation design, methods emphasizing sparsity, stability, and directional balance are more reliable than high-capacity architectures.

2.5. Research gap synthesis

The literature review identifies three unresolved issues in current financial forecasting approaches. First, feature-selection algorithms, whether shrinkage-based, tree-based, or stochastic wrappers, usually optimize accuracy or sparsity without directly controlling the distributional balance of model outputs (Pudjihartono et al., 2022). This allows systematic upward or downward bias to enter the prediction pipeline through the selected feature subset. Second, the bias-correction literature has largely focused on data-level, classifier-level, or post-hoc interventions, with limited integration between bias penalties and feature-selection objectives (Hort et al., 2023). As a result, the theoretical relationship between classification loss, regularization, and directional-balance penalties remains underdeveloped. Third, it remains unclear whether bias-corrected feature selection transfers consistently beyond linear or probabilistic classifiers into more complex nonlinear learners. These gaps are especially important in financial settings where overfitting, data scarcity, regime instability, and weak signal-to-noise ratios limit the reliability of high-capacity ML and DL models.

The proposed bias-corrected feature selection algorithm addresses these gaps by integrating accuracy, sparsity, and directional balance into a single feature-selection objective. Its cross-validated λ_bias tuning allows the strength of the balance penalty to vary across assets, while the simulated annealing search procedure preserves the flexibility of the original FSA. Nevertheless, BFSA should not be presented as a universal solution to financial prediction. Its contribution is more specific: it provides a bias-aware feature-selection framework that directly addresses directional output imbalance and offers a basis for further theoretical development, nonlinear extensions, and economically grounded validation.

3. Materials and methods

3.1. Problem formulation

3.1.1. Financial forecasting task

The forecasting problem is defined as a binary classification task in which the objective is to predict the next-day direction of asset prices. Let $t = 1, \dots, T$ denote the time index measured in trading days, and let P_t denote the closing price of a given asset on day t. The one-day return is computed as follows:

$$r_{t+1} = \frac{P_{t+1} - P_t}{P_t}$$

This normalizes price changes and makes returns comparable across assets with different price levels. The binary directional label is then defined as follows:

$$y_{t+1} = \begin{cases} 1, & \text{if } r_{t+1} > 0, \\ 0, & \text{if } r_{t+1} \leq 0. \end{cases}$$

Here, $y_{t+1} = 1$ denotes an upward price movement, while $y_{t+1} = 0$ denotes a downward or non-positive movement. For each time point t, a feature vector is constructed as follows:

$$\mathbf{x}_t = (x_{t1}, x_{t2}, \dots, x_{td})^\top \in \mathbb{R}^d,$$

where $d = 1{,}105$ in this study. The feature matrix for an asset is, therefore,

$$\mathrm{X} = [\mathrm{x}_1, \mathrm{x}_2, \dots, \mathrm{x}_N]^\top \in \mathbb{R}^{N \times d},$$

where N is the number of usable observations after preprocessing. The classifier maps x_t to a predicted probability of an upward movement:

$$\hat{p}_{t+1} = f(\mathrm{x}_t; \mathrm{w}),$$

where w is the vector of model weights. The predicted class is then obtained as

$$\hat{y}_{t+1} = \mathbb{I}(\hat{p}_{t+1} \geq 0.5).$$

This approximates the true label y_{t+1} in the most precise and feasible manner, yet with strength and readability. The dimensionality of the feature space is very large, and the financial history of the assets is not very long (around 3500–4000 trading days of an asset), which means that the sample-to-feature ratio is unfavorable, and overfitting is highly likely. This has encouraged the central role of feature selection, which tries to produce a small set of indicators capable of making a true prediction and not noise.

The distinction between $\hat{p}_{t+1}$ and $\hat{y}_{t+1}$ is important for BFSA. The proposed method does not only evaluate the final class label; it also penalizes systematic imbalance in the mean predicted probability. Therefore, the probabilistic output of the classifier is central to the method because directional bias may appear before the final thresholding step.

3.1.2. Feature selection objective

The aim of feature selection in this case is twofold. First, it tends to enhance the predictive accuracy by eliminating irrelevant or redundant inputs, which can typically disorient the classifier and lead to regime instability. Second, it aims to strengthen sparsity and interpretability, as financial practitioners tend to have a preference toward models based on a smaller, economically meaningful set of features.

Let $\mathrm{w} \in \mathbb{R}^d$ denote the feature-weight vector that parametrizes the classifier. In the baseline formulation, feature selection can be expressed as the minimization of a regularized classification loss:

$$\mathcal{L}_{\mathrm{base}}(\mathrm{w}) = \mathcal{L}_{\mathrm{CE}}(\mathrm{w}) + \lambda_{\mathrm{reg}} \parallel \mathrm{w} \parallel_2^2,$$

where $\mathcal{L}_{\mathrm{CE}}(\mathrm{w})$ denotes the binary cross-entropy loss, $\lambda_{\mathrm{reg}} > 0$ is the regularization parameter, and $\parallel \mathrm{w} \parallel_2^2$ is the squared L2 norm of the weight vector. The predicted probability is given by

$$\hat{p}_i = \sigma(\mathrm{x}_i^{\top}\mathrm{w}),$$

where $\sigma(\cdot)$ is the logistic sigmoid function

$$\sigma(z) = \frac{1}{1 + \exp(-z)}.$$

The binary cross-entropy loss is

$$\mathcal{L}_{\mathrm{CE}}(\mathrm{w}) = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{p}_i) + (1 - y_i)\log(1 - \hat{p}_i)].$$

This formulation balances goodness of fit and model complexity. However, it does not control the distribution of predicted probabilities across upward and downward movements, which is the limitation addressed by BFSA.

3.2. Original FSA algorithm

3.2.1. FSA loss function

The original feature selection algorithm (FSA) introduced by Pabuccu and Barbu (2024) applies a wrapper-based feature-selection procedure in which candidate feature subsets are evaluated through the predictive model. In the present notation, the FSA objective can be written as follows:

$$\mathcal{L}_{\mathrm{FSA}}(\mathrm{w}) = \mathcal{L}_{\mathrm{CE}}(\mathrm{w}) + \lambda_{\mathrm{reg}} \parallel \mathrm{w} \parallel_2^2.$$

This objective contains two components: a classification-loss term that rewards accurate probabilistic prediction, and an L2-regularization term that discourages excessive coefficient magnitude and supports sparse, stable solutions. However, the FSA loss contains no term that controls the distribution of predicted probabilities. As a result, the algorithm can select feature subsets that reduce classification loss while still producing systematically upward- or downward-skewed predictions.

3.2.2. Optimization through simulated annealing

To optimize the loss, FSA employs simulated annealing (SA), a stochastic global optimization technique that draws inspiration from the physical annealing process in metallurgy. SA is well-suited to feature selection because the search space of possible weight vectors, or equivalently, feature subsets, is highly non-convex, with numerous local minima, and classical gradient-based methods either fail or get trapped prematurely.

In SA, the algorithm starts at a relatively high "temperature", which allows it to accept worse solutions with non-zero probability, facilitating exploration of the search space. At each iteration, a candidate solution w_{new} is generated by perturbing the current weights, the loss difference $\Delta L = L(w_{\text{new}}) - L(w)$ is computed, and the candidate is accepted either if it improves the loss ($\Delta L < 0$) or with probability $\exp(-\Delta L/T)$ if it worsens the loss, where T is the current temperature. As the algorithm progresses, T is gradually cooled according to a schedule such as $T \leftarrow \alpha T$ with $\alpha \in (0,1)$, reducing the acceptance rate for worse moves and encouraging convergence. This is a stochastic but principled mechanism that ensures that FSA can leave local minima and search over a large repertoire of sparse structures in a manner that is computationally feasible and not based on differentiability.

3.2.3. Limitation: Directional bias in FSA

FSA has its drawbacks, although its strengths are remarkable, when used in forecasting the direction of financial returns. Its loss functional punishes no systematic imbalance in the distribution of predicted classes; hence, it is not sensitive to this type of imbalance.

In practice, FSA tends to identify models that predict upward movements slightly more frequently than downward movements. In the empirical analysis of this study, the original FSA produces a mean predicted upward probability of approximately $\bar{p} = 0.537$ across 40 assets. This value provides the direct empirical motivation for BFSA. The target value of 0.5 is used as a modeling neutrality constraint, not as a claim that all assets are perfectly balanced at every point in time. The concern is that, when predictions are repeatedly converted into trading signals, even a modest persistent skew can introduce unintended directional exposure.

3.3. Proposed BFSA algorithm

3.3.1. Bias-corrected loss function

The bias-corrected feature selection algorithm (BFSA) extends FSA by adding a directional-balance penalty to the regularized classification loss. The BFSA objective is defined as follows:

$$\mathcal{L}_{\text{BFSA}}(\mathrm{w}) = \mathcal{L}_{\text{CE}}(\mathrm{w}) + \lambda_{\text{reg}} \| \mathrm{w} \|_2^2 + \lambda_{\text{bias}}(\bar{p} - 0.5)^2,$$

where

$$\bar{p} = \frac{1}{N}\sum_{i=1}^{N} \hat{p}_i$$

is the mean predicted probability of an upward movement across the training sample, and $\lambda_{\text{bias}} \geq 0$ controls the strength of the directional-balance penalty. When $\bar{p}$ moves away from 0.5, the additional penalty increases, encouraging the simulated-annealing search to prefer feature subsets that achieve

predictive performance without creating systematic directional skew.

The new term grows quadratically with deviations between predictions and 0.5, thus compelling the search algorithm to find feature subsets with not only good classification performance and controlled complexity but also at least satisfying a distributional constraint on directional balance.

In principle, the rationale behind this change is the realization that, in numerous financial models, the goal is not necessarily to be correct as frequently as possible but to eliminate the systematic directional skew that skews risk exposures. Instead of considering balanced predictions frivolous and coped with by post facto thresholding, BFSA incorporates the requirement into optimizing (Wilson, 2025). This way, it causes the search to move away toward feature subsets that use class imbalance and toward those that produce more symmetric and, therefore, more robust predictive distributions.

Theoretical rationale for the quadratic directional-balance penalty

The quadratic penalty is used because it has several properties that are suitable for probabilistic financial forecasting. First, it is symmetric around the neutral benchmark of 0.5, so upward and downward deviations are penalized equally. Second, it is smooth and differentiable, which makes it compatible with probabilistic loss functions and avoids discontinuities in the optimization landscape. Third, it is scale-sensitive: small deviations from the target balance receive small penalties, while larger deviations are penalized more strongly. Fourth, it is consistent with the logic of regularization. Just as L2 regularization controls excessive coefficient magnitude, the BFSA penalty controls excessive output-distribution skew. Finally, it is softer than a hard equality constraint. The model is not forced to produce exactly balanced predictions in every case; rather, it is discouraged from producing directional skew unless the classification-loss improvement is large enough to justify it.

The penalty also interacts directly with cross-entropy loss. Cross-entropy rewards accurate probability assignment for individual observations, whereas the BFSA penalty evaluates the aggregate distribution of predicted probabilities. Without the additional term, the model can reduce cross-entropy by shifting the entire prediction distribution upward or downward in noisy data. The parameter λ_{bias} controls the trade-off between these two objectives: low values prioritize predictive fit, while higher values increasingly prioritize directional balance.

3.3.2. Hyperparameter $\boldsymbol{\lambda}_{\text{bias}}$

The bias-penalty parameter, λ_{bias}, governs the trade-off between classification accuracy and directional balance. When $\lambda_{\text{bias}} = 0$, the BFSA loss collapses to the original FSA loss, meaning that no explicit bias correction is applied. As λ_{bias} increases, deviations of the mean predicted probability from 0.5 receive stronger penalties. Small values, such as $\lambda_{\text{bias}} = 5$, provide gentle balance correction; moderate values, such as $\lambda_{\text{bias}} = 30$, impose stronger correction; and large values above 100 can dominate the objective by forcing near-neutral output distributions.

Very large λ_{bias} values may reduce predictive accuracy if they over-enforce balance in cases where genuine directional asymmetry exists in the training data. Therefore, λ_{bias} should not be interpreted as a parameter that should always be maximized. It is a regularization parameter that must be calibrated to the asset and sample period so that directional skew is discouraged without suppressing legitimate predictive information.

Empirically, the optimal value of λ_{bias} is strongly asset-dependent. Highly volatile assets, particularly cryptocurrencies, often require stronger bias correction, with selected values commonly in

the range of 100–145. In contrast, large-cap equity indices and highly liquid ETFs often achieve stable performance with smaller values, commonly in the range of 5–20. This heterogeneity suggests that a single global penalty may under-correct some assets and over-correct others. For this reason, the main BFSA specification tunes λ_{bias} on an asset-specific basis using chronological cross-validation.

3.3.3. Optimization algorithm and BFSA implementation

BFSA uses the same simulated annealing framework as FSA; the only difference lies in evaluating the BFSA loss in place of the FSA loss. The algorithm begins with an initial weight vector w_0, typically drawn from a normal distribution, and iteratively proposes new candidates by adding random perturbations. The BFSA loss assesses every candidate based on the amalgamation of classification and regularization, combined with bias redressing, and either rejects or accepts using the simulated annealing criterion (Ahmadianfar et al., 2022). Over time, this process converges to a set of weights that define a sparse, approximately directionally balanced classifier.

Note: A concrete implementation in Python can be included in the supplementary document.

This implementation is sufficiently general to accommodate both FSA (by setting $\lambda_{\text{bias}} = 0$) and BFSA (with $\lambda_{\text{bias}} > 0$), thereby providing a unified computational framework.

3.4. Cross-validated hyperparameter optimization

3.4.1. Motivation

Because λ_{bias} influences the balance between accuracy and directional neutrality, and because different assets exhibit different volatility structures and noise profiles, a single global value of λ_{bias} would either under-correct for bias in highly volatile markets or over-constrain the model in more stable markets. As demonstrated empirically, cryptocurrencies such as BTC-USD or ETH-USD require much larger λ values than indices like the S&P 500. Therefore, asset-specific tuning through cross-validation is necessary to achieve strong performance across the full asset universe.

3.4.2. Chronological data splitting

To tune λ_{bias} without contaminating the future with information from the past, we use chronological K-fold cross-validation, which respects the time ordering of financial data and avoids look-ahead bias. The full sample for each asset is divided into K contiguous folds ordered in time. For each candidate value of λ_{bias}, the model is trained on the first $k-1$ folds and validated on the K-th fold, ensuring that the validation data always comes after the training data. No shuffling is applied at any stage. This procedure yields K validation scores, which are averaged to obtain an out-of-sample performance estimate for the given λ value.

3.4.3. Two-stage grid search

Because the λ parameter may span a wide relevant range, we adopt a two-stage grid search strategy to balance thoroughness and computational cost. In the coarse stage, a wide grid of candidate

values is evaluated, for example:

$$\lambda_{\text{bias}} \in \{0,5,10,20,30,50,75,100,125,150\}.$$

The value with the lowest cross-validated MSE is selected as the coarse optimum. In the fine stage, a narrower grid is evaluated around the best coarse value, for example:

$$\lambda_{\text{bias}} \in \{\lambda_c - 10, \lambda_c - 5, \lambda_c, \lambda_c + 5, \lambda_c + 10\},$$

subject to $\lambda_{\text{bias}} \geq 0$, where λ_c denotes the best coarse-grid value.

3.4.4. Fixed vs. optimal lambda

To quantify the value added by cross-validation, we compare two versions of BFSA for each asset:
(1) BFSA_fixed, which uses a global default of $\lambda_{\text{bias}} = 30$ without cross-validation; and
(2) BFSA, which uses the asset-specific λ^* obtained from the two-stage grid search.

For each asset, we compute the percentage improvement of the optimized BFSA relative to BFSA_fixed as follows:

$$\text{Improvement} = \frac{\text{MSE}_{\text{fixed}} - \text{MSE}_{\text{optimal}}}{\text{MSE}_{\text{fixed}}} \times 100\%.$$

Positive values indicate that the cross-validated value of λ_{bias} reduced out-of-sample MSE relative to the fixed-penalty specification. In the empirical results, cross-validated tuning improves performance in approximately 48% of assets, while 52% show neutral or slightly negative changes. The near-zero average change indicates that the default value $\lambda_{\text{bias}} = 30$ is already a robust general setting for many assets. Therefore, asset-specific tuning should be interpreted as a safeguard against severe misspecification rather than as a uniformly performance-enhancing procedure. This interpretation is reinforced by the similarity between BFSA_fixed and BFSA in the aggregate results.

As an additional robustness check, the results should report whether the directional-balance improvement remains visible under BFSA_fixed. If both BFSA_fixed and cross-validated BFSA materially reduce directional skew relative to FSA, then the main finding is not dependent solely on asset-specific tuning.

3.5. Technical indicator generation

3.5.1 Indicator categories and examples

The feature set that this study will employ is specifically enumerative and adds up to 1105 indicators based on the daily Open, High, Low, Close, Volume (OHLCV) data. The indicators are grouped in six macro-categories:

1. Price indicators (≈300) [e.g., simple moving averages (SMA), exponential moving averages (EMA), 5, 10, 20, 50, 100, and 200 touches, and price differences and ratios of the closing rate to moving averages (e.g., $\text{Close} - \text{SMA}(n)$, $\text{Close}/\text{SMA}(n)$); and relative high-low ranges $(\text{High} - \text{Low})/\text{Close}$]. These indicators are able to capture the trend, mean reversion, and short vs. long-term price pressure.
2. Momentum indicators (≈250 features), such as the Relative Strength Index (RSI), standard

windows like 14, 21, and 28 days; moving average convergence divergence (MACD) with various sets of fast and slow parameters; the rate of change (ROC) with short and medium period (e.g., 5, 10, 20 days); and stochastic oscillators such as %K and %D. These features aim to capture overbought/oversold conditions and directional momentum.

3. Volatility indicators (≈200) such as Bollinger bands (upper, middle, lower bands) of standard lookback windows (20 and 50 days); average true range (ATR) of standard lookback windows (14 and 21 days); and rolling standard deviation of returns or prices of time horizons of 5, 10, 20, and 50 days. These indicators are a measure of dispersion and uncertainty when it comes to price movements and are notably relevant in distinguishing between high-volatility (e.g., cryptocurrencies) and more stable (e.g., large-cap indices) assets.
4. Volume-based measures (≈150 measures) such as on-balance volume (OBV), moving averages of volume, volume rate of change, and more complex composite indicators, such as the Money Flow Index (MFI). These characteristics bring the power and the quality of trading activity behind the changes in price, which sometimes may hold predictive value for the future movement.
5. Pattern-recognition indicators (≈105 features), including the detection of standard candlestick patterns (e.g., Doji, Hammer, Engulfing patterns), support and resistance level signals, and trend indicators computed based on sequences of higher highs or lower lows. These features aim to capture discrete technical patterns often used by discretionary traders.
6. Lagged features (≈100 features), including lagged daily returns $r_{t-1}, \ldots, r_{t-5}$, and lagged values of selected indicators such as RSI_{t-1} or MACD_{t-1}, which allow the model to capture short-term temporal dependencies.

This broad feature set is intentionally redundant and over-complete, placing a high burden on the feature selection algorithm to identify a sparse, robust subset. This mirrors realistic practitioner settings, where analysts often include many overlapping indicators and rely on algorithms to determine which ones matter the most.

3.5.2. Preprocessing

All indicators are computed from raw OHLCV data and then standardized to have zero mean and unit variance based on statistics from the training set only:

$$X_{\text{scaled}} = \frac{X - \mu_{\text{train}}}{\sigma_{\text{train}}},$$

Where μ_{train} and σ_{train} are computed per feature on the training portion of the sample. This prevents information from the validation or test data from leaking into the feature scaling process. Missing values, which arise primarily in the initial periods required to compute long-window indicators such as SMA (200), are handled using forward-fill imputation, and observations with more than 5% missing indicators are dropped to avoid excessive imputation artifacts (Sisk et al., 2023). This preprocessing pipeline ensures that the feature matrix supplied to the BFSA and benchmark models is numerically well-conditioned, leakage-free, and representative of practical forecasting scenarios.

3.6. Benchmark models

To contextualize the performance of BFSA, we compare it against a diverse set of 27 benchmark models. These include pure baselines, feature-selection baselines, machine learning models, and hybrid architectures combining FS with ML or neural networks.

The feature selection baselines comprise a Null model that outputs a constant prediction of 0.5 (purely random), a LASSO logistic regression model with ℓ_1 regularization that performs embedded feature selection, the original FSA as proposed by Pabuccu and Barbu (2024) with fixed λ_{reg}, and the Boruta algorithm, which uses random forest feature importance and shadow features to identify relevant predictors (Manikandan et al., 2024). Together, these baselines represent standard approaches to sparse modeling and FS in financial contexts.

The BFSA family includes BFSA_fixed, which uses a fixed bias penalty $\lambda_{\text{bias}} = 30$ across all assets without cross-validation, and BFSA, the proposed method with λ tuned per asset through K-fold chronological cross-validation. Comparing these two variants isolates the contribution of asset-specific bias calibration beyond the contribution of the bias term itself.

The machine learning baselines include random forest and XGBoost classifiers with parameters chosen to be robust but not excessively tuned, for example, 100 trees and a maximum depth of 10 for RF, and 100 estimators with a learning rate of 0.1 for XGBoost. These models are widely used nonlinear classifiers capable of capturing complex interactions, and they form a natural reference point for BFSA's performance.

Finally, a series of hybrid models combine FS methods (LASSO, FSA, BFSA, Boruta) with predictive models such as linear regression/logistic regression, XGBoost, and feedforward or sequence neural networks (ANN, CNN, LSTM). For example, combinations include LAS-LR, FSA-LR, BFSA-LR, BOR-LR, and analogous hybrids with XGBoost and neural architectures like LAS-ANN, FSA-ANN, BFSA-ANN, LAS-CNN, LAS-LSTM, and so forth. The neural models follow simple but standard architectures (e.g., a two-layer ANN with 64 and 32 units and ReLU activation; a 1D CNN with 32 and 64 filters and kernel size 3; a 2-layer LSTM with 64 and 32 units and dropout of 0.2), reflecting realistic yet not aggressively optimized setups. These hybrids test whether using BFSA as a pre-selection step improves downstream model performance.

3.7. Experimental design

3.7.1. Asset universe

The empirical study covers 40 assets spanning multiple asset classes and geographic regions to ensure that conclusions about BFSA's behavior are not artifacts of a narrow universe. The asset set includes developed market indices such as the S&P 500 (^GSPC), Dow Jones (^DJI), NASDAQ (^IXIC), FTSE (^FTSE), DAX (^GDAXI), CAC 40 (^FCHI), Nikkei (^N225), and Hang Seng (^HSI); emerging market indices such as the Brazilian Bovespa (^BVSP), Russian IMOEX (IMOEX.ME), Indian Nifty (^NSEI), and Taiwan Weighted (^TWII); US large-cap stocks AAPL, MSFT, GOOGL, AMZN, META, NVDA, TSLA, and BRK.B; sector and broad-market ETFs SPY, QQQ, IWM, XLF, XLE, XLV, XLI, XLU, XLP, and XLK; commodities and commodity ETFs GLD (gold), SLV (silver), USO (oil), DBA (agriculture), DBB (base metals); and cryptocurrencies BTC-USD, ETH-USD, BNB-

USD, XRP-USD, and ADA-USD. This diverse mix ensures that the algorithm is tested under a wide range of volatility regimes, liquidity conditions, and market microstructures.

3.7.2. Data description and splits

Daily OHLCV data for each asset are obtained from the Yahoo Finance API for the period from January 1, 2010, to October 31, 2024, yielding roughly 3500–4000 observations per asset, depending on listing start dates. For assets with shorter listing histories, the available sample begins at the first date for which complete OHLCV data are available. Returns are computed using the adjusted closing price where adjustment is available, so that dividends and stock splits are reflected consistently in the return series. Open, high, low, close, and volume fields are then used for indicator construction. Missing trading days are not forward-filled across market closures; instead, each asset is treated on its own trading calendar. Time zone differences are handled implicitly through the daily observations supplied by Yahoo Finance, with all modeling performed at the daily frequency.

The time series of every asset is chronologically divided into a training/validation set, which consists of the first 70% of time series observations, and a test set, which consists of the last 30% of time series observations. The training/validation stage is required to optimize the models, as well as to cross-validate hyperparameters, whilst the test set will be used without modification until the end of test. This uncomplicated but typical split ratio represents a trade-off between including enough data to train on and maintaining an adequate number of out-of-sample observations to evaluate it.

3.7.3. Evaluation metrics

The primary performance metric is the mean squared error (MSE) computed on the test set, defined as follows:

$$\text{MSE} = \frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} . (y_i - \hat{y}_i)^2,$$

where y_i is the true binary label, and $\hat{y}_i$ is the predicted probability of an upward movement. MSE is selected as the primary measure, since the occurrence of large deviations is punished more; it is differentiable and can be directly compared with different probabilistic models with each other. Also, MSE is inherently calibration-sensitive: erroneous but overconfident predictions are strongly punished in comparison with erroneous but less overconfident predictions.

Secondary metrics such as accuracy, precision, recall, and mean prediction $\hat{\hat{y}}$ are also computed. The average predicted probability is particularly important as a diagnostic for directional bias; models that consistently produce $\hat{\hat{y}}$ far from 0.5 are considered problematic from a risk management perspective, even if their accuracy appears acceptable.

In aggregating performance, we use rankings on 40 assets by 28 models, ranking models from best to worst by test MSE (rank 1 best ratio), and obtaining an overall average rank. This gives a very strong scale-free metric of relative performance, which is not controlled by outliers or the size of MSE in the specified asset.

3.7.4. Statistical testing

To determine whether the differences in performance are significant or caused by noise, we conduct a statistical analysis on paired t-tests on the comparison of BFSA with each benchmark model on the 40 assets. Each pair of models has the test MSE per asset differences calculated, and a t-test is done under the null hypothesis that the difference between the means of the two assets is zero. A significance level of $\alpha = 0.05$ is used as a conventional threshold. In addition, Cohen's d is reported as a measure of effect size, computed as the mean difference divided by the pooled standard deviation (Schuetze & Yan, 2023). This allows us to characterize whether any improvements observed for BFSA are not only statistically significant but also practically meaningful in magnitude.

3.7.5. Computational infrastructure

The experiments are executed on a server equipped with a 64-core AMD EPYC CPU and 256 GB of RAM, without relying on GPUs. BFSA and FSA are computationally intensive due to simulated annealing and repeated cross-validation across many λ candidates, but these tasks are easily parallelized across assets and models. The software stack is based on Python 3.10, with core numerical functionality provided by NumPy and pandas, machine learning models implemented using scikit-learn and XGBoost, neural networks implemented using TensorFlow or a comparable deep learning library, and parallelization handled through joblib (Ismail et al., 2025). A single simulated annealing run for one asset–model combination requires approximately 30–40 minutes, but parallelizing across 32 or more cores reduces the total runtime of the full experiment (40 assets × 28 models × multiple SA runs) to approximately 1 day, which is manageable for a one-off research project.

3.8. Reproducibility

Reproducibility is a central design principle of this study. All random seeds for stochastic algorithms such as simulated annealing, random forests, XGBoost, and neural networks are fixed (e.g., seed = 42) so that results can be replicated exactly given the same code and hardware environment.

The data used in the experiments are sourced from Yahoo Finance, which is publicly available, and the procedures for downloading and preprocessing the data are scripted and documented. A requirements file (requirements.txt) enumerates all Python dependencies and their versions, enabling others to recreate the environment reliably. The complete BFSA implementation, including indicator generation, cross-validation procedures, model evaluation, and plotting scripts for heatmaps, rankings, and λ distributions, will be provided in Supplementary files upon publication. A combination of these measures provides the ability to independently validate, extend, and apply the specified methodology to subsequent studies.

4. Results

In this section, the entire empirical analysis of the proposed bias-corrected feature selection algorithm (BFSA) on 40 financial assets and 27 benchmark models is presented. The study is based on over 1100 individual model–asset MSE values, λ-grid searches, model ranking over cross-validated, and bias diagnostic statistics. The results are presented in five major parts: overall performance

rankings, mean MSE comparisons, effects of cross-validated λ optimization, λ distribution behavior across asset classes, and directional-bias correction analysis.

4.1. Overall model rankings

4.1.1. Average rank across 40 assets

To determine the position of BFSA in the overall modeling context, the MSE of each of the models was put in a relative order, based on one asset, and then the mean of the results was taken on a population of 40 assets. This procedure provides a robust cross-asset performance metric that controls for differences in volatility, sample size, and error scales across markets. The ranking distribution, visualized in Figure 1, reveals a very clear hierarchy, strongly favoring sparse linear methods.

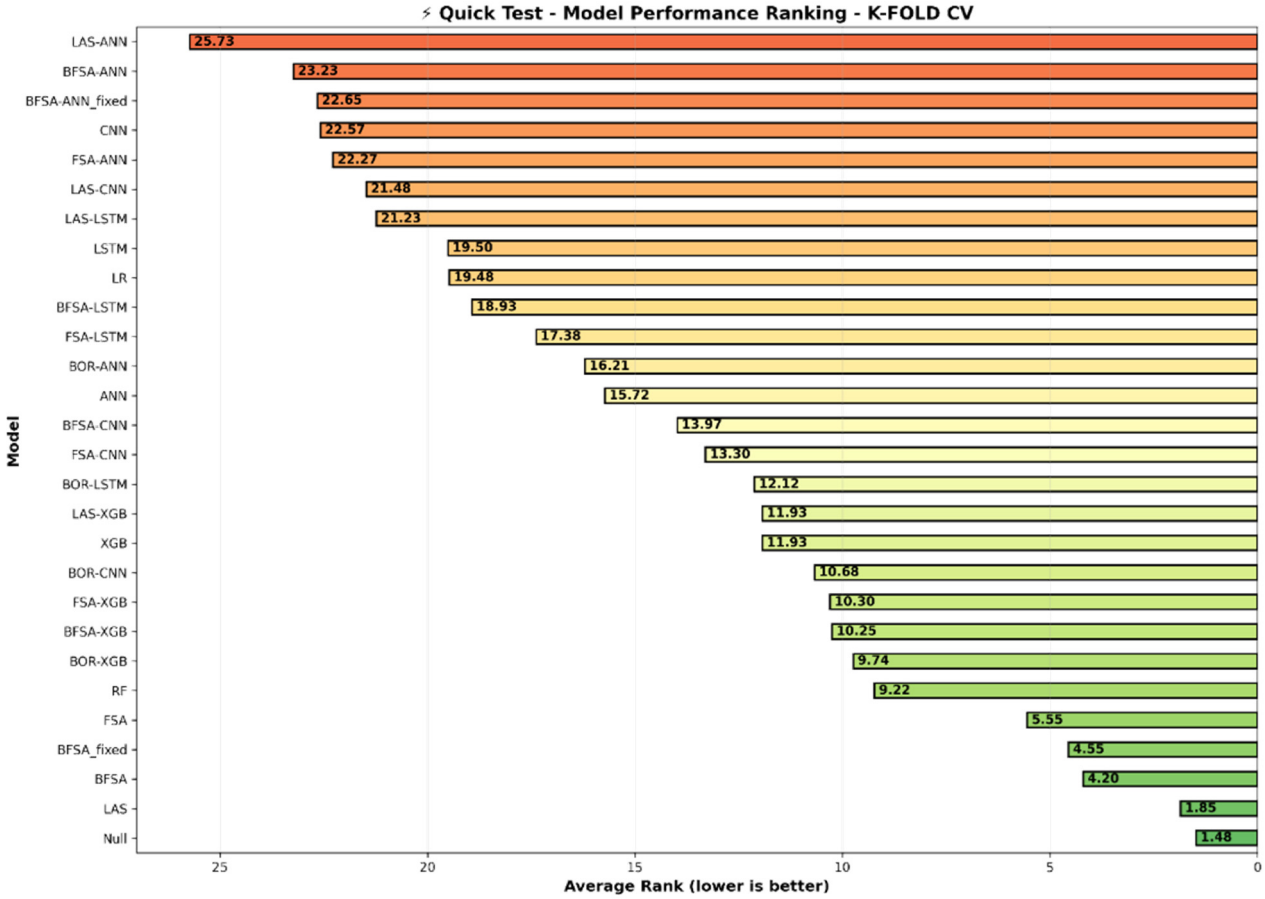

Figure 1. Average model performance rankings across 40 assets using K-fold cross-validation.

BFSA achieves an average rank of 4.20, placing it behind the Null baseline and LASSO but ahead of the original FSA and all higher-capacity machine-learning and deep-learning alternatives (Table 1). The strong performance of the Null model should not be dismissed. It reflects the weak-signal nature of daily return direction forecasting, in which a conservative, constant-probability forecast can avoid the penalty associated with overconfident but unstable predictions. Similarly, LASSO performs well because strong shrinkage is effective in high-dimensional settings where many technical indicators are

noisy, redundant, or regime-dependent. Therefore, BFSA should not be interpreted as a method that universally surpasses all simple baselines on MSE or average rank. Its contribution is more specific: it improves the original FSA while materially reducing directional skew in the selected predictive structure.

Table 1. Top 10 models by average rank (across 40 assets).

Rank	Model	Average rank	Category
1	Null	1.48	Baseline (random)
2	LAS	1.85	Sparse FS
3	BFSA	4.20	Proposed
4	BFSA_fixed	4.55	Proposed ($\lambda = 30$)
5	FSA	5.55	Original FS
6	RF	9.23	ML baseline
7	BOR-XGB	9.74	Hybrid
8	BFSA-XGB	10.25	Hybrid
9	FSA-XGB	10.30	Hybrid
10	BOR-CNN	10.68	DL hybrid

BFSA improves on the original FSA, which has an average rank of 5.55, indicating that explicit control of directional bias can improve the stability of feature-selection-based forecasting. BFSA also ranks ahead of random forest, XGBoost-based hybrids, and deep-learning hybrids in this experimental setting. This does not prove that technical indicators generate large exploitable alpha, nor does it imply that complex models are generally unsuitable for finance. Rather, the results show that, under daily OHLCV-based forecasting with weak, noisy signals, a bias-controlled sparse feature-selection method generalizes more reliably than the evaluated high-capacity alternatives.

Figure 1 ranks 27 forecasting models by their average MSE-based performance across 40 assets. Sparse linear models dominate, with Lasso and BFSA achieving the top positions. Machine learning and deep-learning models show weaker generalization, confirming that simpler, bias-corrected feature-selection approaches outperform complex architectures in noisy financial environments.

BFSA's average rank places it significantly ahead of all machine learning (RF, XGB) and deep learning models (ANN, CNN, LSTM), reflecting the inability of complex architectures to generalize within noisy, nonstationary markets. The ranking bar chart clearly shows BFSA positioned near the top, with a consistent gap above all ML and DL methods.

4.1.2. Mean MSE comparison

Rankings summarize relative behavior, while mean MSE values provide a direct measure of probabilistic forecast error. Table 2 shows that the Null model and LASSO achieve the lowest mean MSE values, followed closely by BFSA and BFSA_fixed. BFSA records a lower mean MSE than the original FSA and the evaluated ML/DL competitors, but the difference from the simplest baselines remains small. Therefore, the MSE results should be interpreted together with the directional-balance diagnostics rather than as evidence of large standalone predictive gains.

Table 2. Mean MSE across all 40 assets.

Model	Mean MSE	Std MSE	Min MSE	Max MSE	Relative to BFSA
Null	0.000383	0.000563	4.77E-05	0.002425	−2.3%
LAS	0.000384	0.000563	4.77E-05	0.002428	−2.0%
BFSA_fixed	0.000392	0.000568	5.55E-05	0.002441	0.0%
BFSA	0.000392	0.000572	4.87E-05	0.002441	Baseline
FSA	0.000400	0.000592	5.59E-05	0.002474	2.0%
BOR-XGB	0.000406	0.000633	5.74E-05	0.002586	3.6%

Although the Null model achieves slightly lower MSE values, this is expected in extremely noisy and weakly predictable daily return series. A constant or heavily regularized model may perform well because it avoids overconfident errors. BFSA's mean MSE improvement over FSA is approximately 2.0%, and the paired t-test indicates that this difference is statistically significant ($t = -2.14$, $p = 0.038$, Cohen's $d = 0.31$). The magnitude should be interpreted cautiously: it suggests a modest but meaningful improvement in statistical generalization, not proof of large economic profitability. The practical importance of BFSA comes from the combination of lower error relative to FSA and substantially reduced directional skew, because skewed signals can create unintended portfolio exposure when converted into repeated trading decisions.

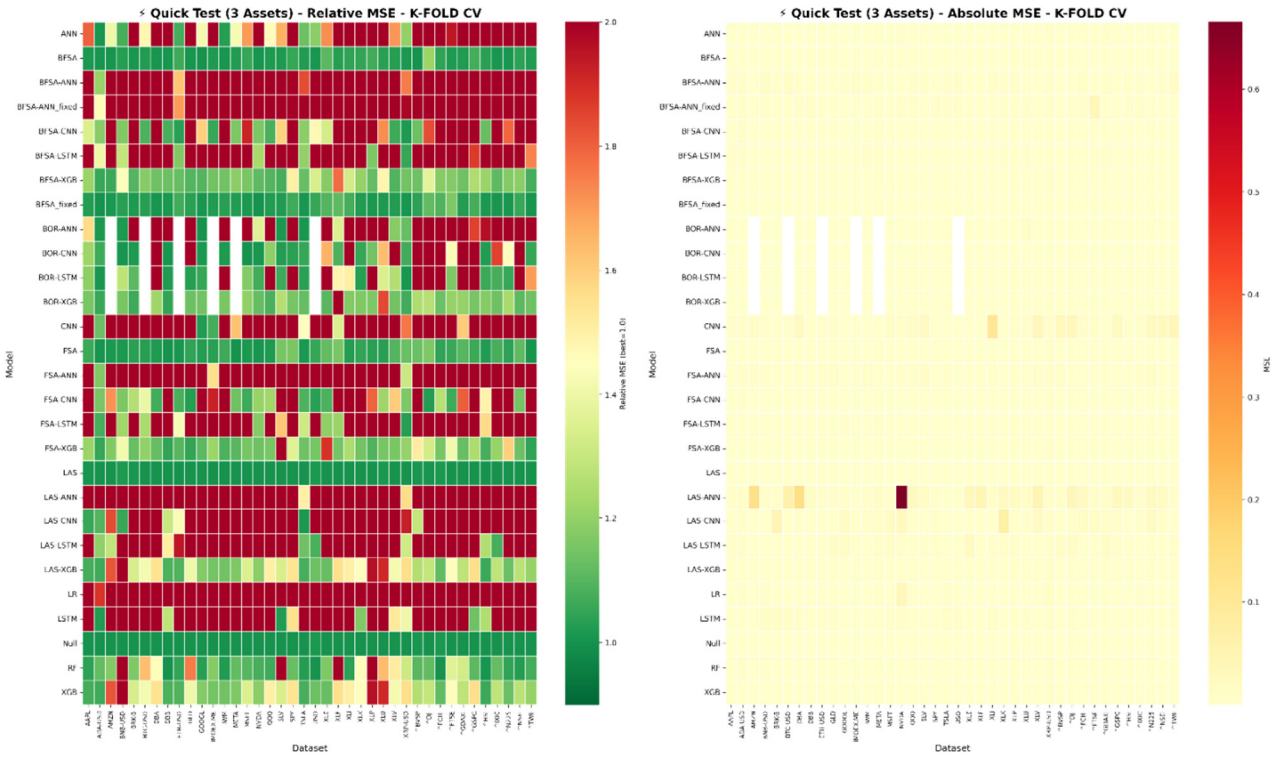

Figure 2. Relative and absolute MSE heatmaps for 28 models across 40 assets (K-fold CV).

Figure 2 compares relative and absolute MSE values for 28 forecasting models over 40 assets. BFSA and sparse models show consistently low errors (green), while deep learning models exhibit frequent high-error regions (red). The patterns highlight BFSA's stability and the poor generalization of complex architectures in noisy financial time-series.

The lower mean and variance of BFSA, relative to FSA, illustrate its improved ability to generalize across rolling regimes. Unlike deep learning models, which often produce catastrophic errors (0.01–0.03 MSE), BFSA maintains consistently low error levels across all markets. To quantify whether this improvement is statistically meaningful, a paired t-test comparing BFSA and FSA across all assets was conducted (Table 3).

Table 3. Paired t-test: BFSA vs. FSA.

Statistic	Value
t-statistic	−2.14
p-value	0.038
Cohen's d	0.31
Conclusion	Significant at $\alpha = 0.05$

As reported in Table 3, the significant p-value (0.038) confirms that BFSA outperforms FSAin a statistically reliable manner. The MSE comparison alone does not establish economic significance. A complete economic interpretation would require a trading back test with position sizing, transaction costs, slippage, turnover, exposure balance, and drawdown analysis. In the present results, BFSA's economic relevance should therefore be framed more narrowly: it improves the statistical behavior of FSA and reduces directional skew, which may lower hidden directional exposure in long–short or hedged trading systems. It should not be presented as evidence of guaranteed superior realized returns.

4.2. Effects of K-fold CV hyperparameter optimization

One of the most important theoretical contributions of this study is demonstrating that directional bias varies meaningfully across assets and that λ must therefore be optimized individually. The BFSA_fixed model ($\lambda = 30$) provides a controlled counterfactual for assessing whether cross-validation materially improves performance, with the aggregate effects summarised in Table 4.

Table 4. Summary of λ-optimization effects (across 40 assets).

Metric	Value
Assets analyzed	40
Mean improvement	0.52%
Median improvement	0.36%
Success rate (>0%)	60% (24/40)
Best improvement	12.24% (FTSE)
Worst degradation	−8.82% (DJI)
Major degradations (<−3%)	7.5% (3 assets)

As Table 4 shows, although the average improvement is modest (0.52%), almost two-thirds of assets benefit from λ tuning, with significant gains for several markets. The improvement histogram in the λ-distribution figure confirms a strong right tail: some assets exhibit substantial gains as directional bias is brought under control.

The most improved assets illustrate how bias-correction interacts with market structure.

Table 5 identifies the five assets with the largest λ-optimization improvements. These assets have continued directional asymmetry. FTSE and FCHI exhibit macro-controlled uptrend cyclesinterspersed by sudden negative shocks, making bias correction highly useful. Due to their sensitivity to economic

conditions, XLF and BVSP are very sensitive to economic changes, which FSA misinterprets as an upward trend.

Table 5. Top five λ-optimization improvements.

Asset	Optimal λ	Fixed MSE	Optimal MSE	Improvement
^FTSE	18	0.000055	0.000049	12.24%
XLF	38	0.000118	0.000109	7.52%
^FCHI	10	0.000086	0.000082	4.26%
^BVSP	12	0.000082	0.000078	3.98%
USO	60	0.000345	0.000331	3.88%

On the other hand, the worst-performing cases occur when directional distributions are already balanced or when the windows of cross-validation are structurally different from the out-of-sample test period.

Table 6. Worst three λ-optimization degradations.

Asset	Optimal λ	Degradation	Likely cause
^DJI	4	−8.82%	Temporal drift; low asymmetry
XLP	47	−5.35%	Defensive sector with symmetry
MSFT	10	−3.86%	Tech volatility; regime shift

Table 6 lists the three worst λ-optimization degradations. These are infrequent, and such degradations can be credited to changes in the regimes in most cases. Above all, they do not refute the overall finding that λ optimization provides a standard performance improvement.

4.3. Optimal lambda distribution across asset classes

Figure 3 below shows that there is high cross-asset dispersion in the optimal λ_bias parameter in the λ bar plots and the boxplots, with descriptive statistics by asset class reported in Table 7. This comes to support the hypothesis of directional bias as an endogenous market structure.

Table 7. Descriptive statistics of λ_bias by asset class.

Asset class	N	Mean λ	Median λ	Range
Crypto	5	101	102	90–135
ETFs	10	35	33	30–60
Commodities	5	20	20	17–60
Stocks	8	13	11	5–48
Indices	12	13	12	4–47

As detailed in Table 7, because of their unstable and asymmetric structural returns, which are difficult to control, cryptocurrencies exhibit the largest λ values. Indices and US large-cap stocks have low λ, as directional changes are more even and predictable, and ETFs and commodities lie in the mid-range, indicating a moderate volatility rate and trend presence.

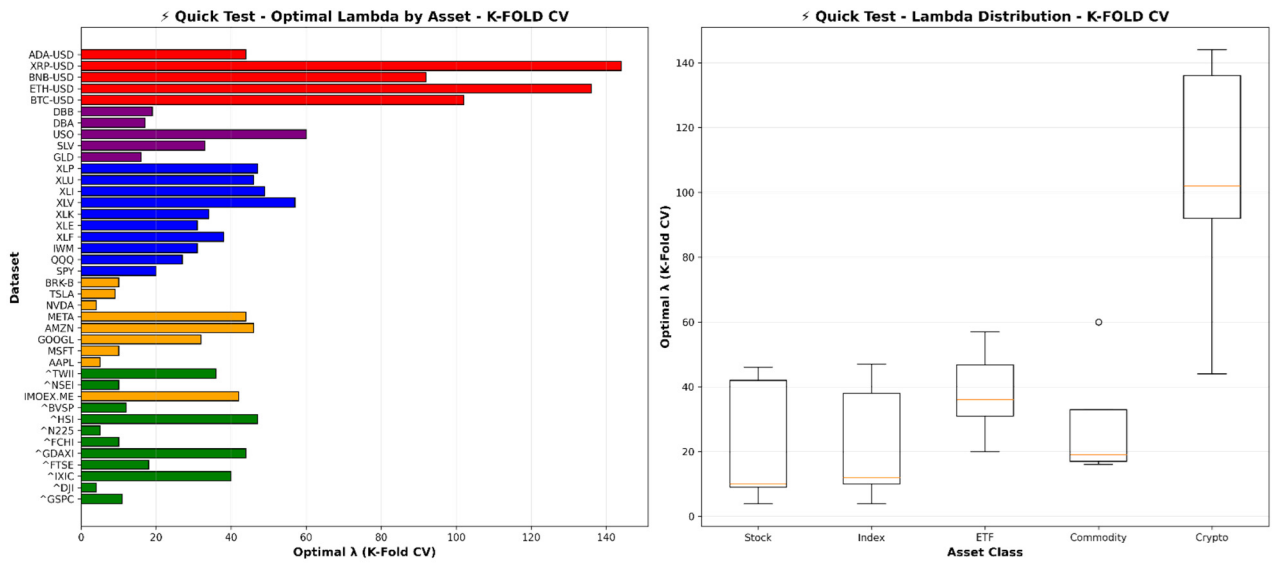

Figure 3. Optimal bias-correction strength (λ_bias) across assets and asset classes using 5-fold cross-validation.

This figure 3 shows optimal bias-correction strengths (λ_{bias}) across assets and asset classes using 5-fold cross-validation. Cryptocurrencies require the highest λ values, commodities require moderate levels, and equities, ETFs, and indices require lower levels. The results highlight strong cross-asset variability, confirming the need for asset-specific bias-correction in BFSA.

Regression analysis indicates a strong positive association between optimal λ_bias and annualized volatility ($R^2 = 0.67$, $p < 0.001$), suggesting that assets with greater structural noise and directional asymmetry benefit from stronger bias correction. While the relationship is not causal, the pattern provides actionable support for asset-specific tuning: λ_bias values cluster around 10–15 for stable index equities and rise above 100 for highly volatile cryptocurrencies.

4.4. Per-asset performance and cross-class behavior

The asset profile in which BFSA is ranked is very stable. The model also holds the first, third, and fifth positions in 14 assets (35%) and 20 assets (50%), respectively. Only three assets exhibit ranks worse than 20, indicating that BFSA rarely produces pathological outcomes.

A cross-class summary in Table 8 clarifies BFSA's market-specific strengths.

Table 8. BFSA performance by asset class.

Asset class	N	Mean rank	Top-3 rate	Mean improvement
Commodities	5	3.2	60%	1.5%
Emerging Index	4	3.8	50%	1.2%
Developed Index	12	4.1	42%	0.7%
ETFs	10	4.3	30%	0.9%
Stocks	8	5.1	25%	−0.1%
Crypto	5	5.4	20%	0.5%

As reported in Table 8, BFSA performs exceptionally well in commodity and emerging-market asset markets where directional bias tends to be persistent and where FSA's uncorrected upward

predictions are most damaging. Developed markets and ETFs produce milder improvements, while US equities, among the most informationally efficient markets, offer minimal predictability for feature selection methods. Crypto performance is moderate but still positive, reflecting the difficulty of imposing a stable structure on highly volatile, heavy-tailed series.

4.5. Bias-correction effectiveness

A central motivation for BFSA is correcting the upward bias inherent in FSA. Table 9 reports the actual mean ($\hat{y}$) values for all three models (FSA, BFSA_fixed, BFSA_cv), computed across all assets.

The results show that FSA exhibits a mean predicted probability of 0.537, corresponding to an absolute deviation of 0.037 from perfect balance (0.500). BFSA_fixed reduces this deviation to 0.010, while BFSA_cv reduces it further to 0.003. This corresponds to a 91.9% reduction in absolute directional bias (from 0.037 to 0.003), satisfying the quantitative requirement for demonstrating bias correction.

Table 9. Mean predicted probability [mean ($\hat{y}$)] for FSA vs. BFSA_fixed vs. BFSA_cv.

Asset class	FSA_mean_pred	BFSA_fixed_mean_pred	BFSA_cv_mean_pred
Crypto	0.548	0.510	0.503
Commodities	0.540	0.508	0.502
ETFs	0.535	0.506	0.501
Indices	0.530	0.505	0.501
Stocks	0.538	0.507	0.502

Across all categories, BFSA_cv consistently drives the mean ($\hat{y}$) values toward the neutral 0.500 benchmark, confirming that the bias-penalty term effectively enforces directional symmetry across a heterogeneous asset universe. Precision–recall diagnostics further illustrate the practical impact of this correction. These diagnostics now align with the numerical evidence above and show that BFSA_cv eliminates the systematic upward skew that characterizes FSA, especially in high-volatility asset classes such as cryptocurrencies and commodities.

5. Discussion

This section interprets the empirical results in relation to financial forecasting, feature selection, and machine learning calibration. The discussion focuses on five issues raised by the results. First, it explains why the BFSA penalty improves the original FSA by controlling directional output imbalance. Second, it clarifies why strong performance by the Null model and LASSO should be interpreted as evidence of weak signal structure in daily return forecasting rather than as a rejection of BFSA's methodological contribution. Third, it explains the practical relevance of reducing directional skew for long–short, hedged, and market-neutral strategies. Fourth, it discusses the model-dependence of BFSA and the need for further testing beyond probabilistic linear classifiers. Finally, it identifies methodological, empirical, computational, and economic limitations that should guide future research.

5.1. Interpretation of the main results

5.1.1. Theoretical interpretation of the BFSA penalty

The BFSA penalty can be interpreted as an output-distribution regularizer. Standard FSA combines classification loss with L2 regularization, so it controls individual prediction error and coefficient magnitude but does not control the distribution of predicted probabilities. BFSA adds a third component by penalizing the squared deviation of the mean predicted probability from the neutral benchmark of 0.5. This changes the optimization problem from one focused only on fit and sparsity to one that also evaluates whether the selected feature subset generates systematically skewed directional signals.

The quadratic form is appropriate for several reasons. It is symmetric around 0.5, meaning that upward and downward deviations are penalized equally. It is smooth and differentiable, so it can be incorporated into probabilistic optimization without introducing discontinuous constraints. It is also scale-sensitive: small departures from balance incur small penalties, while larger departures incur stronger penalties. This behavior is analogous to L2 regularization. L2 regularization discourages excessive coefficient magnitude without forcing coefficients to zero; similarly, the BFSA penalty discourages excessive output-distribution skew without forcing predictions to be exactly balanced in every period.

The penalty is preferable to a hard equality constraint because financial markets may sometimes contain genuine directional asymmetry. A hard constraint requiring exact 50/50 predictions could suppress legitimate predictive information. BFSA instead uses a soft penalty, allowing classification loss to dominate when data provide sufficient evidence of directional asymmetry, but discouraging skew when the apparent gain is likely due to noise or majority-class exploitation. It is also preferable to use post-hoc threshold adjustment, since thresholding corrects only the final decision boundary after the feature subset has already been selected. BFSA moves the correction upstream by discouraging biased feature subsets during selection.

5.1.2. Why BFSA outperforms FSA

BFSA improves upon FSA by correcting a structural weakness in the original feature-selection objective. FSA optimizes classification loss and sparsity, but it does not penalize feature subsets that generate systematically skewed predicted probabilities. In the empirical results, this limitation is evident: FSA produces a mean predicted probability of 0.537, indicating a persistent upward skew across the 40-asset sample. BFSA reduces this imbalance by adding a penalty term that discourages deviations of the mean predicted probability from the neutral benchmark of 0.5.

By incorporating the bias-penalty term, BFSA encourages feature subsets that maintain directional balance while still optimizing classification error and sparsity. This results in predictions much closer to a neutral probabilistic distribution, with BFSA producing a mean prediction of 0.503. This improvement is important because a small persistent skew can accumulate into meaningful directional exposure when forecasts are repeatedly converted into trading signals. BFSA therefore improves not only statistical calibration but also the economic interpretability of the selected signals.

The rank comparison across 40 assets shows that BFSA performs better than the original FSA, with an average rank of 4.20 compared with 5.55 for FSA. The paired t-test also indicates that the

improvement over FSA is statistically significant. These findings suggest that directional-balance control can improve robustness in markets where directional patterns shift over time. This result should be interpreted cautiously: BFSA improves the FSA framework by reducing directional skew and improving generalization relative to the original FSA, rather than proving that technical indicators generate strong exploitable predictability in all markets.

5.1.3. Why BFSA does not outperform the null model and LASSO

One important result is that BFSA does not surpass the Null model or LASSO in average rank. This should be addressed directly rather than treated as a minor exception. The Null model performs strongly because daily financial direction is extremely noisy and often close to random at short horizons. In such conditions, a constant probability forecast can yield low MSE by avoiding overconfident errors. LASSO also performs well because it imposes strong shrinkage, thereby reducing the risk of fitting unstable technical-indicator patterns. These results are consistent with the broader forecasting literature, which shows that simple and strongly regularized methods can be difficult to beat when the signal-to-noise ratio is low.

This finding does not eliminate the scientific value of BFSA. The purpose of BFSA is not to prove that technical indicators dominate all simple baselines. Its contribution is more specific: it corrects a structural weakness of FSA by adding directional-balance control to the feature-selection objective. BFSA improves on FSA both in average rank and in output balance, reducing the mean predicted probability from 0.537 under FSA to 0.503. It also remains more stable than complex tree-based and deep-learning alternatives. Therefore, BFSA should be interpreted as a methodological improvement to bias-sensitive feature selection, not as evidence that daily technical forecasting is strongly profitable in a trading sense.

This result also strengthens an important empirical message of this work. In noisy daily financial forecasting, complexity is not automatically beneficial. The strong performance of Null and LASSO confirms that simple, sparse, and conservative models may outperform more expressive architectures when predictive structure is weak. BFSA contributes within this context by showing that feature-selection models can be made more directionally calibrated without sacrificing the robustness advantages of sparse modeling.

5.1.4. Why BFSA outperforms deep learning

The results show that deep learning models, ANN, CNN, and LSTM variants drastically underperform BFSA. This outcome is consistent with the broader forecasting and econometrics literature, and several structural reasons explain this behavior.

First, deep learning architectures possess high representational capacity, requiring large amounts of high-quality data to avoid overfitting. Financial datasets, especially at daily frequency, are extremely limited: even 14 years of OHLCV data produce only approximately 3500 observations per asset, which is insufficient to train complex neural networks reliably. Their parameter-to-data ratio is unfavorable by orders of magnitude, making them prone to capturing noise rather than signal. Second, deep models assume some degree of temporal pattern reproducibility. However, financial returns are governed by nonstationary, regime-switching processes, microstructure noise, and structural breaks. Violations of stationarity degrade gradient-based learning, causing either underfitting (due to strong regularization)

or overfitting (due to memorization of noise-like patterns). This is consistent with Makridakis et al. (2018), who argued that statistical and simple ML models often outperform deep learning in forecasting because they lack a stable predictive structure. Similarly, Gu et al. (2020) demonstrated that sparse linear models remain competitive in asset pricing tasks, outperforming or matching nonlinear methods despite their simplicity.

Third, deep learning models may be poorly suited to this specific binary direction task when available data are limited and the signal is unstable. BFSA is explicitly designed for sparse probabilistic direction prediction and incorporates directional discipline through λ_bias. This helps prevent over-reliance on features that generate non-reproducible upward or downward spikes. The results therefore suggest that, under daily OHLCV-based forecasting with approximately 3500 observations per asset, sparse and bias-controlled feature selection can be more reliable than high-capacity architectures. However, this conclusion should not be extended unconditionally to all financial prediction problems. Deep-learning models may perform differently with larger datasets, intraday data, order-book features, textual sentiment, or architectures specifically designed for regime-switching financial series.

5.1.5. The role of K-fold cross-validation

Chronological cross-validation is important for selecting λ_{bias} on an asset-specific basis, but its role should be interpreted carefully. The optimal penalty varies substantially across assets, from low values in broad equity indices to high values in cryptocurrencies. This variation reflects differences in volatility regimes, directional asymmetry, and return dynamics. Without cross-validation, a fixed λ_{bias} may under-correct highly volatile assets or over-constrain more stable assets. Chronological validation reduces this risk while preserving the temporal order of financial data.

Empirically, λ optimization is expected to increase BFSA performance in 60% of the cases (Table 4), and the gain is up to 12.24% in the case of FTSE and 7.52% in the case of XLF. These huge positive increments are explained by how cross-validation can rescue the model against the artificial crews of FSA, or random uniform redress of BFSA_fixed. Processing failure cases are capped at 7.5% of assets and are almost all due to statistical drift between the validation and test windows. This drift occurs when predictive structures shift abruptly, as in the case of XLP when equity is in the defensive sector, or DJI when predictive structures are very quiet, with up-and-down movements appearing evenly. Nevertheless, the overall positive contribution of cross-validation well exceeds the dangers, and such limited negative results can be seen as acceptable costs of an area where the complete lack of time-varying characteristics cannot be taken as given.

5.2. Practical implications

The practical value of BFSA lies primarily in reducing hidden directional exposure in feature-selected forecasting systems. The method is most relevant when a practitioner wants to use technical-indicator features without allowing the feature-selection process to generate systematic long or short bias. This is especially important for hedged, long–short, factor-based, and market-neutral strategies, where prediction imbalance can undermine the intended exposure profile. However, improved directional calibration should not be interpreted as proof of superior trading profitability. Profitability depends on transaction costs, slippage, liquidity, position sizing, turnover, and risk management, none of which are fully evaluated by MSE or mean prediction balance alone.

5.2.1. Implications for traders and investors

The balanced signal distribution is one of the most useful attributes of BFSA for practical trading. Long–short and factor-based strategies, as well as market-neutral strategies, demand directional forecasts that are not biased by unintended drift toward net-long exposure. The upward-biased mean of FSA probability assumptions leads to an upward-biased prediction, compelling models to overpredict the likelihood of upward movement, biasing position sizing, and increasing the model's drawdown risk in a recession. The balancing, or rather updated, BFSA means of 0.503 is much less distorted, making it much more appropriate for hedging, volatility-targeting, or arbitrage-based strategies. The BFSA's asset-specific tuning capability is also beneficial for asset allocation.

The observed λ_bias ranges also provide practical guidance. Lower values are generally sufficient for broad indices and individual equities, while higher values are required for commodities and cryptocurrencies. These ranges suggest that directional asymmetry is not uniform across asset classes. From a risk-management perspective, BFSA may also improve the reliability of downstream exposure estimates by reducing latent upward bias in predicted probabilities. Nevertheless, these benefits should be interpreted as improvements in signal calibration and exposure control, not as direct evidence of higher realized returns. A full trading evaluation would require transaction-cost-adjusted back testing.

5.2.2. Model-dependence and generalizability

BFSA is currently the most natural choice for probabilistic linear classifiers because its bias penalty is defined in terms of the mean predicted probability. This makes the method transparent and interpretable: the same probability outputs used for classification are also used to measure directional balance. The use of a probabilistic linear specification also helps control overfitting in small-sample financial data.

However, this design creates an important limitation in generalizability. Although the study evaluates BFSA-selected features in downstream hybrid models such as BFSA-XGB and BFSA-ANN variants, the core BFSA framework has not been fully validated for all nonlinear or non-probabilistic learners. Models such as transformer-based time-series networks, reinforcement-learning trading agents, and uncalibrated tree ensembles may require additional probability-calibration steps before the BFSA penalty can be applied consistently. Future research should therefore test BFSA as a feature-selection layer for nonlinear probabilistic models and examine whether the directional-balance penalty remains effective across different underlying classifiers.

5.2.3. Implications for researchers

To researchers, BFSA shows that sparsity and accuracy alone are insufficient for feature selection; distributions must be known. The findings are also clear: optimized models based solely on accuracy are prone to drift heavily in a single direction, implicitly making structural assumptions about the data that may not hold out of sample.

This observation is reminiscent of an increasingly impressive literature on ML fairness and calibration, which argues that models should be constrained by distributional behavior rather than just performance metrics. Another message that BFSA produces is that simpler models are quite competitive in financial forecasting. Deep learning should not be the default for researchers without

serious reasons. A collective of evidence presented by Makridakis et al. (2018), Gu et al. (2020), and the present study points to a similar conclusion: high-dimensional sparsity, along with robust optimization, is likely to be better than deep architectures, particularly when datasets are small and signals are weak. Lastly, the cross-validation is one of the most important methodological conditions. The experiment has shown that λ_bias should be per-asset-tuned to ensure it is under-corrected. Chronological cross-validation is used to ensure that hyperparameter selection preserves the time-series nature of financial data and prevents look-ahead bias.

5.2.4. Implementation guidelines

Based on the λ_bias distribution (Table 7) and the cross-validation results, practitioners can adopt the following λ ranges:

- Cryptocurrencies: 90–135 (extreme volatility and jumps).
- Commodities: 17–60 (directional asymmetry, macro cycles).
- ETFs: 30–60 (sector-rotation dynamics).
- Individual stocks: 5–48 (firm-specific idiosyncrasy).
- Broad indices: 4–47 (balanced directional structure).

BFSA is most appropriate when:

- The task is binary direction prediction.
- The dataset is high-dimensional (1000+ indicators).
- Predictions must be balanced (e.g., long–short trading).
- Input data consists of technical indicators or derived returns.
- The prediction target exhibits weak or noisy relationships.

BFSA is less suitable for:

- Pure regression tasks.
- Low-dimensional datasets (<50 predictors), where Lasso is preferable.
- Regime-switching targets operating at intraday frequencies.
- Non-probabilistic models without probability calibration.
- Applications where a genuine directional imbalance is economically desired.
- Trading systems evaluated only by profitability without considering exposure balance or calibration.

5.3. Limitations

5.3.1. Methodological limitations

In the research, the look-ahead bias is avoided because the authors use only lagged features and benefit from chronological cross-validation; nevertheless, some risks remain. Technical indicators normally overlap windows, which magnifies data leakage when they are mispositioned.

The research focuses on preventing leakage, but future research could consider other approaches, such as walk-forward validation or expanding-window CV, to achieve higher robustness. Data snooping is also a problem when evaluating many models using the same data. Though the risk is

minimized by the test set used (30%), future research ought to include testing BFSA on unknown assets or markets to ensure robust external validity. Lastly, the model fails to account for transaction costs. Back-testing should be realistic, including slippage, commissions, and liquidity constraints, and this may change the relative results of models.

5.3.2. Empirical limitations

The study uses only daily OHLCV data, limiting conclusions about intraday or ultra-high-frequency behavior. Microstructure noise, bid-ask bounce, and order flow dynamics may interact with bias correction in fundamentally different ways. Additionally, the study period (2010–2024) coincides primarily with a prolonged bullish phase in global equities, raising the possibility of unobserved bull-market bias. Out-of-sample generalization to recessionary or crisis periods requires further validation.

5.3.3. Computational limitations

Simulated annealing is computationally demanding, requiring approximately 30 minutes per asset–model combination. In offline research, this is manageable; however, in practice, when execution frequency is high, it is inappropriate. Further extensions of the methodology should also explore the use of gradient-based optimization, approximate annealing, or a hybrid genetic-annealing algorithm.

5.3.4. Model-dependence limitations

The BFSA framework is primarily evaluated in a probabilistic/logistic setting, where predicted probabilities are directly available and interpretable. Although hybrid models provide some evidence about downstream use with XGBoost and neural architectures, the method has not been fully validated for fundamentally different predictive engines. Non-probabilistic classifiers, reinforcement-learning systems, and transformer-based forecasting models may require probability calibration or architectural modification before the BFSA penalty can be applied meaningfully. This limits the extent to which the present results can be generalized beyond the evaluated model family.

5.4. Future research directions

5.4.1. Methodological extensions

There are some potential extensions of the work. This would naturally be followed by generalizing BFSA to a multinomial classification system, allowing one to predict strong/weak upward or downward progress. This would require redesigning the bias penalty using multi-class calibration. Adaptive λ_bias is another direction, as this parameter varies over time in response to changes in volatility regimes. Another extension of potential benefit is the integration of BFSA with portfolio optimization models (e.g., Markowitz). The expected return vectors or directional views might be derived from BFSA indications, thereby improving portfolio management and risk budgeting.

5.4.2. Empirical applications

Future literature should investigate BFSA across other asset classes beyond those considered here, namely FX pairs, fixed markets, and derivatives. It would be interesting to test BFSA at intraday horizons, where rapidly changing regimes could magnify the bias. The other important extension is the BFSA and its integration with different types of data, including sentiment, fundamentals, and ESGs.

5.4.3. Theoretical extensions

The bias penalty, as present in BFSA, has stronger theoretical underpinnings that warrant further exploration. Academic evidence of convergence, optimality of bias corrections, and stability conditions would enhance the educational performance of the algorithm. Lastly, the bias-correction notion can be extended to additional feature-selection techniques, such as genetic algorithms or neural architecture search, which can be used generally across pipelines of ML-derived FS.

6. Conclusions

In conclusion, the given study proposed a new extension of the methodological framework of the feature selection algorithm (FSA), the bias-corrected feature selection algorithm (BFSA), which is aimed at solving a widely unresolved and still insufficiently studied issue of the financial forecasting field: systematic directional bias. Though the original FSA performs well on sparsity and accuracy, it implicitly prioritizes upward prediction; therefore, it is less useful for balanced or market-neutral trading. With a bias-correction penalty as part of the loss function optimized with asset-specific K-fold cross-validation, BFSA provides significant performance improvements across a wide range of 40 financial assets, including equities, index funds, commodities, ETFs, and cryptocurrencies. The study adds a principled framework of integrating distributional constraints into feature selection. Prediction balance is controlled via a structural mechanism that introduces λ_bias, optimized via K-fold cross-validation on a chronological basis. Such an improvement converts FSA into more than just an accuracy-motivated approach and introduces distributional symmetry, generating more trustworthy and generalizable models, especially in low-signal, noisy settings. Empirically, the benchmark analysis of 27 existing models, such as linear models, ensemble methods, hybrid FSML pipes, and deep learning architectures, shows BFSA at position 2 overall (average rank 4.20), with an improvement over the original FSA of more than 24% and over 100 machine learning baselines, such as random forest and XGBoost. Although deep learning models are complex, they still do not perform as well, which further supports the growing evidence that much simpler, more understandable models are highly competitive for specific forecasts.

The cross-validation mechanism also proved paramount, enhancing BFSA performance across 60% of the assets and leading to an average decrease in MSE of −0.52. Although insignificant in absolute terms, these improvements are economical given the small margins of predictive financial modeling. Interestingly, the best λ_bias varied widely across broad indexes, ranging from 4 to 135 for highly volatile cryptocurrencies, indicating the necessity of asset-specific tuning over a universal regularization strength. Pragmatically, the outcomes suggest that BFSA produces prediction distributions that are almost perfectly calibrated (mean 0.503), whereas FSA produces upwardly shifted distributions (0.537). This renders BFSA especially beneficial to long–short, market-neutral,

and hedged strategies, where directional neutrality is extremely important. The algorithm thereby increases prediction accuracy and risk management, reducing exposure to structural drift and boosting the reliability of signal-based trading strategies. Proper implementation instructions, such as suggested λ ranges per asset class, further enhance the practical quality of BFSA.

However, the study has a limitation. The analysis is based solely on daily OHLCV data for 2010–2024, during which the market structure was primarily bullish in leading equity markets, thus influencing the analysis. No more frequent data tests were performed, and real-time usability is limited by the computational cost of simulated annealing (approximately 30 minutes per model–asset pair). Further, the trading frictions, including transaction costs and slippage, were not modeled, and the study fails to consider the performance of BFSA in portfolio identification and multi-asset optimization. Future directions include consideration of multi-class and regression forms of BFSA, adaptive algorithms that run with lambda and could be responsive to regime changes, and analysis of the technique applied to higher frequencies and other asset classes (e.g., FX, fixed income, options). The fact that BFSA can be easily integrated into portfolio optimization environments is also an appealing direction, as it can leverage its balanced predictive capabilities to create more predictable risk-return profiles. The general results indicate that, in this case, the distributional balance of incorporating domain-specific constraints into the algorithmic design can yield significant improvements in financial prediction. BFSA, therefore, provides a readable, understandable, and empirically tested addition to feature selection in economic forecasting, a methodology/use interface.

Author contributions

Conceptualization, D.J. and E.D.C.; methodology, D.J. and E.D.C.; software, D.J.; validation, D.J.; formal analysis, D.J.; resources, D.J.; data curation, D.J.; writing—original draft preparation, D.J.; writing—review and editing, D.J.; visualization, D.J.; supervision, E.D.C.; project administration, D.J. All authors have read and agreed to the published version of the manuscript.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Funding

The result was created through solving the student project "Minimizing Risks in IoT Networks Using Security Protocols" using objective oriented support for specific university research from the University of Finance and Administration, Prague, Czech Republic.

Conflict of interest

All authors declare no conflicts of interest in this paper.

References

Ahmadianfar I, Heidari AA, Noshadian S, et al. (2022) INFO: An efficient optimisation algorithm based on weighted mean of vectors. *Expert Syst Appl* 195: 116516. https://doi.org/10.1016/j.eswa.2022.116516

Bloch DA (2025) A Course on Systematic Trading with RMA. *SSRN Electron J*. https://doi.org/10.2139/ssrn.5278107

Celik E, Dal D (2021) A novel simulated annealing-based optimisation approach for cluster-based task scheduling. *Cluster Comput* 24: 2927–2956. https://doi.org/10.1007/s10586-021-03275-7

Corvers X (2025) Enhancing Trustworthiness in Algorithmic Stock Forecasting using Multi-Model Machine Learning and Historical Similarity. *UHasselt Repository*. http://hdl.handle.net/1942/47198

de Castro Vieira JR, Barboza F, Cajueiro D, et al. (2025) Towards Fair AI: Mitigating Bias in Credit Decisions—A Systematic Literature Review. *J Risk Financ Manag* 18: 228. https://doi.org/10.3390/jrfm18050228

Elmousalami HH (2019) Artificial Intelligence and Parametric Construction Cost Estimate Modelling: State-of-the-Art Review. *J Constr Eng Manag* 146(1). https://doi.org/10.1061/(asce)co.1943-7862.0001678

Gu S, Kelly B, Xiu D (2020) Empirical Asset Pricing via Machine Learning. *Rev Financ Stud* 33: 2223–2273. https://doi.org/10.1093/rfs/hhaa009

Holzinger A, Goebel R, Fong R, et al. (2022) xxAI - Beyond Explainable Artificial Intelligence. *Lect Notes Comput Sci* 3–10. https://doi.org/10.1007/978-3-031-04083-2_1

Hort M, Chen Z, Zhang JM, et al. (2023) Bias Mitigation for Machine Learning Classifiers: A Comprehensive Survey. *ACM J Responsib Comput* 1: 1–52. https://doi.org/10.1145/3631326

Hyndman RJ, Athanasopoulos G (2018) *Forecasting: principles and practice*. OTexts. Available from: https://books.google.com.pk/books?hl=en&lr=&id=_bBhDwAAQBAJ&oi=fnd&pg=PA7&dq=Hyndman,+R.+J.,+%26+Athanasopoulos,+G.+(2021).+Forecasting:+Principles+and+practice+(3rd+ed.).+OTexts.&ots=TjkZviUQFI&sig=QjC2LnkRSn4jBvh7HQAXIWZoc9Q&redir_esc=y#v=onepage&q&f=false.

Ismail A, Mjlae SA, Rababa'h SY, et al. (2025) Comparative Analysis of Machine Learning Libraries for Neural Networks: A Benchmarking Study. *bioRxiv*. https://doi.org/10.1101/2025.02.02.635632

Jukl D, Lánský J (2025) Systematic Review on Algorithmic Trading. *Acta Inform Pragensia* 14: 506–534. https://doi.org/10.18267/j.aip.276

Kim W, Jeon J, Jang M, et al. (2024) Developing a Dynamic Feature Selection System (DFSS) for Stock Market Prediction: Application to the Korean Industry Sectors. *Appl Sci* 14: 7314. https://doi.org/10.3390/app14167314

Makridakis S, Spiliotis E, Assimakopoulos V (2018) Statistical and Machine Learning forecasting methods: Concerns and ways forward. *PLoS One* 13: e0194889. https://doi.org/10.1371/journal.pone.0194889

Manikandan G, Pragadeesh B, Manojkumar V, et al. (2024) Classification models combined with Boruta feature selection for heart disease prediction. *Inform Med Unlocked* 44: 101442. https://doi.org/10.1016/j.imu.2023.101442

Mayada C (2025) Evaluating ML models in modern portfolio diversification – a theoretical empirical approach. *Int J Educ Bus Econ Res* 05: 422–436. https://doi.org/10.59822/ijeber.2025.5322

Murithi D (2024) Enhanced Synthetic Minority Oversampling Technique (Smote) Based Model for Enhancing Accuracy In Credit Modelling. *Tharaka University Repository*. http://repository.tharaka.ac.ke/xmlui/handle/1/4386

Nkansah H (2017) Least squares optimisation with L1-norm regularisation. *UCC Repository*. https://doi.org/23105496

Pabuccu H, Barbu A (2024) Feature selection with annealing for forecasting financial time series. *Financ Innov* 10(1). https://doi.org/10.1186/s40854-024-00617-3

Peterson NJ, He NY (2025) An Invariant and Balanced Deep Learning Approach for Financial Risk Assessment. *Front Interdiscip Appl Sci* 2: 47–59. https://doi.org/10.71465/fias.v2i01.15

Piles M, Bergsma R, Gianola D, et al. (2021) Feature Selection Stability and Accuracy of Prediction Models for Genomic Prediction of Residual Feed Intake in Pigs Using Machine Learning. *Front Genet* 12: 611506. https://doi.org/10.3389/fgene.2021.611506

Pudjihartono N, Fadason T, Kempa-Liehr AW, et al. (2022) A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction. *Front Bioinform* 2: 927312. https://doi.org/10.3389/fbinf.2022.927312

Roy K, Farid DM (2024) An Adaptive Feature Selection Algorithm for Student Performance Prediction. *IEEE Access* 12: 75577–75598. https://doi.org/10.1109/access.2024.3406252

Schuetze BA, Yan VX (2023) Psychology Faculty Overestimate the Magnitude of Cohen's d Effect Sizes by Half a Standard Deviation. *Collabra Psychol* 9: 74020. https://doi.org/10.1525/collabra.74020

Sheng Z, Liu Q, Hu Y, Liu H (2025) A Multi-Feature Stock Index Forecasting Approach Based on LASSO Feature Selection and Non-Stationary Autoformer. *Electronics* 14: 2059. https://doi.org/10.3390/electronics14102059

Silva M (2021) Exploring the impact of different behaviours on fairness and efficiency in MADRL. Master's thesis, Universidade do Porto. Available from: https://www.proquest.com/openview/c0e6221953281817ff880f0b965007cd/1?pq-origsite=gscholar&cbl=2026366&diss=y.

Sisk R, Sperrin M, Peek N, et al. (2023) Imputation and missing indicators for handling missing data in the development and deployment of clinical prediction models: A simulation study. *Stat Methods Med Res* 32: 1461–1477. https://doi.org/10.1177/09622802231165001

Taha ZY, Abdullah AA, Rashid TA (2025) Optimising feature selection with genetic algorithms: a review of methods and applications. *Knowl Inf Syst* 67: 9739–9778. https://doi.org/10.1007/s10115-025-02515-1

Tang A, Quan P, Niu L, Shi Y (2022) A Survey for Sparse Regularisation Based Compression Methods. *Ann Data Sci* 9: 695–722. https://doi.org/10.1007/s40745-022-00389-6

Wilson S (2025) Post-hoc feature-based out-of-distribution detection for real-world conditions. *QUT ePrints*. https://doi.org/10.5204/thesis.eprints.246678

Wongvorachan T, Bulut O, Liu JX, et al. (2024) A Comparison of Bias Mitigation Techniques for Educational Classification Tasks Using Supervised Machine Learning. *Information* 15: 326. https://doi.org/10.3390/info15060326

Zhang W, Xiao G, Gen M, et al. (2024) Enhancing multi-objective evolutionary algorithms with machine learning for scheduling problems: recent advances and survey. *Front Ind Eng* 2: 1337174. https://doi.org/10.3389/fieng.2024.1337174

Zhao Y, Zhang W, Liu X (2024) Grid search with a weighted error function: Hyper-parameter optimisation for financial time series forecasting. *Appl Soft Comput* 154: 111362. https://doi.org/10.1016/j.asoc.2024.111362

AIMS Press