Article

A Risk-Optimized Framework for Data-Driven IPO Underperformance Prediction in Complex Financial Systems

Mazin Alahmadi 🆔

Department of Industrial Engineering, Faculty of Engineering, King Abdulaziz University, Jeddah 21589, Saudi Arabia; mfalahmadi@kau.edu.sa

Abstract: Accurate predictions of Initial Public Offerings (IPOs) aftermarket performance are essential for making informed investment decisions in the financial sector. This paper attempts to predict IPO short-term underperformance during a month post-listing. The current research landscape lacks modern models that address the needs of small and imbalanced datasets relevant to emerging markets, as well as the risk preferences of investors. To fill this gap, we present a practical framework utilizing tree-based ensemble learning, including Bagging Classifier (BC), Random Forest (RF), AdaBoost (Ada), Gradient Boosting (GB), XGBoost (XG), Stacking Classifier (SC), and Extra Trees (ET), with Decision Tree (DT) as a base estimator. The framework leverages data-driven methodologies to optimize decision-making in complex financial systems, integrating ANOVA F-value for feature selection, Randomized Search for hyperparameter optimization, and SMOTE for class balance. The framework's effectiveness is assessed using a hand-collected dataset that includes features from both pre-IPO prospectus and firm-specific financial data. We thoroughly evaluate the results using single-split evaluation and 10-fold cross-validation analysis. For the single-split validation, ET achieves the highest accuracy of 86%, while for the 10-fold validation, BC achieves the highest accuracy of 70%. Additionally, we compare the results of the proposed framework with deep-learning models such as MLP, TabNet, and ANN to assess their effectiveness in handling IPO underperformance predictions. These results demonstrate the framework's capability to enable robust data-driven decision-making processes in complex and dynamic financial environments, even with limited and imbalanced datasets. The framework also proposes a dynamic methodology named Investor Preference Prediction Framework (IPPF) to match tree-based ensemble models to investors' risk preferences when predicting IPO underperformance. It concludes that different models may be suitable for various risk profiles. For the dataset at hand, ET and Ada are more appropriate for risk-averse investors, while BC is suitable for risk-tolerant investors. The results underscore the framework's importance in improving IPO underperformance predictions, which can better inform investment strategies and decision-making processes.

Keywords: data-driven decision making; complex financial systems; IPO prediction; ensemble learning; investor risk preferences; machine learning

1. Introduction

Predicting the aftermarket performance of Initial Public Offerings (IPOs) is a critical area of research in finance, as it can provide valuable insights for investors, companies, and policymakers [1]. The ability to forecast IPO outcomes is essential for data-driven decision-making in complex financial systems, enabling stakeholders to navigate uncertainties and

optimize strategies. The performance of IPOs post-listing serves as a significant indicator of market dynamics and investor sentiment, impacting decisions around investment, issuance strategies, and regulatory policies [2]. Accurate prediction models help mitigate risks and inform better decision-making, making this an area of high importance and continuous development.

Traditionally, IPO performance prediction has relied on various statistical methods, such as regression analysis, which offer foundational insights into the factors influencing IPO outcomes [3]. However, these methods often fall short of capturing the complex, non-linear relationships inherent in financial markets. Despite advancements in machine-learning models, existing approaches still face limitations in handling small, imbalanced IPO datasets, particularly in emerging markets, where data scarcity and class imbalance present significant challenges. Many studies have either overlooked these issues or relied on limited features, reducing their generalizability and predictive robustness. As financial systems grow more intricate, there is an increasing need for advanced data-driven methodologies, such as machine learning, to uncover hidden patterns and improve predictive accuracy [4].

Despite the progress made, predicting IPO performance remains challenging due to issues such as limited data availability and class imbalance. This issue is particularly pronounced in emerging markets, where IPO datasets are typically small, making it difficult to train robust models. Additionally, in these markets, the number of successful IPOs often significantly outnumbers those that underperform, leading to skewed predictions. Addressing these challenges is critical for developing reliable, data-driven models capable of managing the complexities and imbalances of financial systems, particularly in emerging markets.

This study addresses key gaps in IPO underperformance prediction through a comprehensive data-driven framework specifically designed for challenging financial environments. Unlike previous approaches, the framework integrates feature selection, data balancing techniques, and risk-based model evaluation to improve generalizability and decision-making. Leveraging publicly available financial and prospectus data reduces reliance on proprietary datasets, enhancing the practical applicability of the model, especially in emerging markets where data scarcity is prevalent. The methodology employs various ensemble methods, including BC, RF, Ada, GB, XG, ET, and SC, with Decision Trees (DT) as the base estimator. It incorporates ANOVA F-value for feature selection, Randomized Search for hyperparameter optimization, and the Minority Over-Sampling Technique (SMOTE) for class balance to optimize predictive accuracy. Additionally, the study introduces a dynamic methodology that tailors evaluation metrics based on investor risk preferences, ensuring adaptability to different risk profiles. This comprehensive approach aims to provide a robust, versatile tool for IPO performance prediction, offering valuable insights into the field of financial forecasting. Thus, the major contributions of this paper are summarized as follows:

(1) It proposes a unique data-driven framework that is tailored to handle IPO underperformance predictions in complex financial systems, focusing on small and imbalanced datasets relevant to emerging markets by conducting the following:

    a. Depending solely on the publicly accessible pre-listing prospectus and firm-specific financial data.

    b. Utilizing SMOTE to handle class imbalances and ANOVA to manage feature selection.

    c. Incorporating various tuned ensemble classifiers to handle small datasets in the context of IPO underperformance predictions.

(2) It proposes a risk-optimized methodology for classifier selection based on investor's risk preferences.

The remainder of the paper is organized as follows: Section 2 provides a review of the relevant literature, Section 3 describes the dataset used, Section 4 outlines the proposed framework, Section 5 presents the results obtained from the framework, and Section 6 offers conclusions.

2. Literature Review

Predicting the aftermarket performance of IPOs is a critical area of research in finance, as it can provide valuable insights for investors, companies, and policymakers. A wide range of data-driven methodologies has been explored to understand and forecast the behavior of IPO stocks post-listing, spanning from traditional regression analysis to advanced machine-learning techniques tailored for complex financial systems. This literature review examines the key approaches and findings in this domain, emphasizing the evolution of predictive modeling and the role of data-driven decision-making in optimizing IPO performance predictions.

2.1. Regression Approaches

The performance of IPOs has been the subject of extensive research, with various studies employing regression analysis to uncover the determinants and implications of IPO performance across different markets, as shown in Table 1.

Table 1. Comparison of Regression-based Approaches.

Ref	Year	Technique Name	Limitation	Advantage
[5]	2021	Multiple Regression	Limited to the Australian Stock Exchange; lacks broader generalizability to other markets or conditions.	Incorporates industry and listing year as dummy variables, capturing market-specific effects.
[6]	2020	Multiple Regression	Did not find a significant impact on key variables (e.g., firm size, issue size, leverage), limiting actionable insights.	Examines the long-term financial and operating performance of IPOs over a ten-year period.
[7]	2020	Regression and Statistical Tests	Small sample size (12 IPOs); limited to Nairobi Securities Exchange, reducing applicability to larger datasets.	Identifies a high average underpricing rate and post-IPO performance trends.
[8]	2021	Regression Analysis	Focused only on the non-linear relationship of public float, ignoring other potential determinants of underpricing.	Provides empirical evidence supporting the fixed allocation hypothesis for IPO underpricing.
[9]	2022	OLS and Stepwise Regression	Relies on natural logarithm transformations; lacks exploration of alternative models for improved accuracy.	Enhances model explanatory power by applying logarithmic transformations.
[10]	2021	Univariate and Multiple OLS Regression	Focused on Malaysian IPOs; conclusions may not generalize to other markets or book-building mechanisms.	Demonstrates a strong relationship between IPO price multiples and comparable firms, highlighting market valuation dynamics.

Using multiple regression with industry and listing year as dummy variables, Ferdous et al., [5] analyzed 211 IPOs on the Australian Stock Exchange from 2011 to 2015. They found underpricing in the total market return and overpricing in the secondary market, influenced by the year of listing and industry settings. Rafique et al., [6] also used multiple regression to investigate 51 IPOs on the Pakistan Stock Exchange over ten years and concluded that prior IPO demand, firm size, issue size, and leverage do not significantly impact financial and operating performance. Mutai [7] employed regression and traditional statistical tests to examine 12 IPOs on the Nairobi Securities Exchange (NSE) from 1996 to 2013, discovering an average underpricing of 55.36% and a significant post-IPO decline in Cumulative Abnormal Returns (CAR) and Return on Equity (ROE), suggesting the need for investors to consider more financial determinants beyond ROA and ROE. Michel et al., [8] applied regression analysis to explore the relationship between IPO underpricing and public float, the portion of a company's shares that are available for trading by the public, with data from 1996 to 2008, finding a non-linear relationship that supports the hypothesis that firms allocate a fixed amount for underpricing. Mittal and Verma [9] used ordinary least squares (OLS) and stepwise regression methods to analyze 335 book-built IPOs in the Indian capital market from 2006 to 2015, finding that natural logarithm transformations significantly improved model explanatory power. Lastly, Ong et al., [10] utilized univariate and multiple OLS regression analyses to study 467 Malaysian IPOs listed from 2000 to 2017, discovering a positive relationship between IPO price-multiples and those of comparable firms, with lower-valued firms underpricing their IPOs to attract investors and book-built IPOs generating higher initial returns, highlighting the book-building mechanism's role in mitigating misvaluation.

These studies underscore both the strengths and limitations of using regression analysis to predict IPO aftermarket performance. The advantages include identifying key determinants of IPO performance, assessing the impact of various factors such as industry, listing year, and investor demand, and improving model predictability through variable transformation. However, the traditional regression method's limitations include the assumption of a linear relationship between predictors and the outcome, which may not always hold true, leading to potential model misspecification. It also struggles with capturing complex, non-linear interactions compared to advanced techniques like machine learning. Regression models also can be sensitive to outliers, which can disproportionately affect the model estimates and reduce predictive accuracy. Nevertheless, regression analysis remains a pivotal tool in the empirical examination of IPO dynamics, offering valuable insights for investors, underwriters, and regulators.

2.2. Machine-Learning Approaches

The prediction of IPO aftermarket performance has recently been extensively explored through machine-learning (ML) models, as shown in Table 2. This literature review examines several key studies that employ different machine-learning techniques to enhance the prediction accuracy of IPO outcomes.

Table 2. Comparison of Machine-Learning-based Approaches.

Ref	Year	Technique Name	Limitation	Advantage
[1]	2020	Random Forest	Focused only on Borsa Istanbul; lacks comparison with other ensemble methods.	Demonstrates RF's robustness in handling outliers and key predictors like IPO proceeds and trading volume.
[11]	2017	Random Forest	Benchmarked against traditional ML algorithms but excluded newer methods like XGBoost or LightGBM.	RF outperforms traditional ML models in predictive accuracy and error variance,
[12]	2022	Random Forest	Limited to prospectus content; does not include broader financial or market factors.	Achieves 71% accuracy in IPO outcome prediction based on prospectus content.
[13]	2024	Random Forest and Gradient Boosted Trees	Found no significant differences between models; lacks exploration of why performance is similar.	Provides empirical validation of RF and GBT performance similarities in IPO prediction.
[3]	2022	Random Forest and XGBoost	XGBoost achieves higher accuracy, but the study does not explain the disparity in detail.	XGBoost achieves 87.89% accuracy, outperforming RF's 80.25% in IPO forecasting.
[14]	2023	Random Forest and Logistic Regression	Focused only on post-IPO financial performance; lacks broader IPO metrics or market-level insights.	RF outperforms logistic regression in post-IPO performance prediction.
[15]	2022	Random Forest and Logistic Regression	Logistic regression outperforms RF in Hong Kong; lacks exploration of RF's strengths in this market.	RF performs well for long-term predictions.
[16]	2023	Artificial Neural Networks (ANN)	ANN achieves moderate accuracy (68.11%); limited comparison with other advanced deep-learning methods.	ANN outperforms RF and other models in predicting IPO underperformance.
[17]	2024	LightGBM	Focused only on tree-based models; lacks comparison with non-tree-based ensemble methods.	LightGBM achieves an F1 score of 82.3%, excelling in regression and classification tasks.
[18]	2023	Deep Neural Networks (DNN) and Stochastic Frontier Analysis	Complex non-linear models but limited to pricing efficiency, not broader IPO outcomes.	DNN-based model estimates significant premarket underpricing, highlighting pricing inefficiencies.

The random forest algorithm emerges as a prominent method in several studies due to its robustness and high predictive accuracy. Baba and Sevil [1] emphasize the superiority of the random forest algorithm over traditional linear regression models in predicting IPO initial returns on Borsa Istanbul. They attribute this to the algorithm's ability to handle outliers effectively, with key predictors including IPO proceeds and trading volume. Similarly, Quintana et al. [11] benchmark random forests against eight traditional machine-learning algorithms, finding that random forests outperform others in terms of mean and median predictive accuracy and exhibit the second smallest error variance. Emidi and Galán [12] also find that random forests achieve the highest performance (accuracy

of 71%) in predicting IPO outcomes based on prospectus content. In contrast, Dhini and Sondakh [13] and Munshi et al., [3] compare random forests with other ensemble learning methods, such as gradient-boosted trees and XGBoost. While Dhini and Sondakh find no significant performance differences between random forest and gradient-boosted trees, Munshi et al. report that XGBoost achieves the highest average accuracy of 87.89%, compared to 80.25% for random forest.

Methods such as logistic regression and random forest algorithms were explored by Supsermpol et al. [14], who concluded that random forest outperforms logistic regression in predicting post-IPO financial performance. This finding aligns with the results of Emidi and Galán [12] and Ni [15], who also observed superior performance of random forests over logistic regression in various contexts. However, Ni [15] notes that logistic regression produces superior outcomes for predicting IPO performance in the Hong Kong stock market, with random forest also performing well, particularly for longer prediction horizons (Days 10, 20, and 30).

The exploration of advanced ML models reveals varied results. Sonsare et al. [16] find that artificial neural networks (ANN) outperform other models, including random forest, with an accuracy of 68.11% in predicting IPO underperformance. In a similar vein, Neghab et al. [17] demonstrate that tree-based models, particularly LightGBM, outperform other models in both regression and classification tasks, achieving an average F1 score of 82.3%. Neghab et al. [18] introduce a non-linear approach using deep neural networks (DNN) and stochastic frontier analysis to estimate IPO pricing efficiency. Their DNN-based method identifies significant premarket underpricing, with IPO offer prices being, on average, 12.43% lower than the estimated maximum offer prices.

Machine-learning models offer several advantages in predicting IPO outcomes. They exhibit robustness to outliers, with models like random forests handling these anomalies more effectively than traditional linear regression models. Additionally, machine-learning models, including random forests, XGBoost, and artificial neural networks (ANNs), consistently demonstrate higher predictive accuracy compared to traditional statistical methods, with reported accuracy ranging from 68.11% for ANNs to 87.89% for XGBoost for IPO performance predictions. These advanced models are also capable of capturing complex, non-linear relationships, providing deeper insights into the determinants of IPO performance. However, machine-learning models come with certain disadvantages. Their complexity and interpretability can be challenging, especially with advanced models like deep neural networks (DNNs), which are often less transparent than traditional models, complicating their practical application. Moreover, some machine-learning models, such as logistic regression and DNNs, require large datasets and extensive assumptions, which may not always be feasible. There is also a potential risk of overfitting, particularly with complex models like ANNs and ensemble methods, necessitating careful model validation and tuning to avoid this issue.

2.3. Determinants of IPO Underpricing and Performance

The literature on IPO underpricing identifies various determinants influencing initial and long-term performance, as shown in Table 3. Lubis et al. [19] focus on the Indonesian market, finding that inflation and interest rates significantly boost initial returns, while firm-specific factors like ROA, size, and age do not. Oliveira et al., [4] emphasize informational asymmetry as a primary theory, highlighting underwriter and issuer reputations, corporate governance, and offering size as key determinants. Arora and Singh [20] explore Indian SME IPOs, noting that factors such as issue size and oversubscription negatively affect long-run performance, whereas auditor reputation, underwriter reputation, and market conditions positively influence it. Kumar and Sahoo [2] analyze the impact of

anchor investor regulations in India, finding that anchor-backed IPOs underperform less severely in the long run, with offer size, grade, and promoter holding being significant variables. Hussein et al. [21] investigate ChiNext IPOs, revealing that risk factors like ongoing litigation, policy changes, and capital expenditures significantly affect initial returns, indicating the importance of disclosed risk factors. Collectively, these studies underscore the multifaceted macroeconomic conditions, firm characteristics, market perceptions, and regulatory environments.

Table 3. Overview of Determinants of IPO Underpricing and Performance.

Ref	Year	Technique Name	Limitation	Advantage
[19]	2023	Statistical Analysis (Indonesian Market)	Focused only on macroeconomic factors (inflation, interest rates); neglects broader firm-specific and market variables.	Identifies inflation and interest rates as significant determinants of initial returns in the Indonesian market.
[4]	2023	Informational Asymmetry Analysis	Limited to qualitative insights; lacks quantitative validation of determinants.	Highlights underwriter and issuer reputations, corporate governance, and offering size as key factors.
[20]	2020	Regression Analysis (Indian SME IPOs)	Focused on SME IPOs; conclusions may not generalize to larger or global IPOs.	Demonstrates how issue size and oversubscription negatively impact long-run performance, while auditor reputation and market conditions have a positive influence.
[2]	2021	Anchor Investor Analysis	Focused on Indian anchor-backed IPOs; lacks applicability to markets without similar regulations.	Finds that anchor-backed IPOs experience less severe long-term underperformance, with offer size, grade, and promoter holding as significant variables.
[21]	2020	Risk Factor Analysis (ChiNext IPOs)	Limited to ChiNext IPOs; risk factors may vary significantly in other markets or regulatory contexts.	Identifies litigation, policy changes, and capital expenditures as major risk factors affecting IPO initial returns.

2.4. Sentiment and Textual Analysis Approaches

Recent research has extensively explored the utility of sentiment and textual analysis in predicting IPO aftermarket performance, as shown in Table 4. Ly and Nguyen [22] demonstrate that sentiment analysis of IPO prospectuses can predict stock price movements with up to 9.6% greater accuracy than random chance, highlighting the significance of sentiment in short-term IPO performance. Chi and Li [23] find that the readability of IPO prospectuses, assessed through a gradient boost decision tree model, significantly predicts IPO underpricing, suggesting that clearer prospectuses lead to less underpricing due to reduced information asymmetry. Katsafado et al. [24] extend this analysis by incorporating both textual and financial data from S-1 filings, using various machine-learning algorithms to predict IPO underpricing. Their models show a 6.1% improvement in accuracy over financial-only models, with sophisticated approaches outperforming traditional methods.

Zou et al. [25] examine the impact of media coverage on IPOs in China, finding a negative relationship between media coverage and IPO underpricing, as well as heightened investor sensitivity to negative news. Fedorova et al. [26] use advanced techniques like Latent Dirichlet Allocation (LDA) and BERT to analyze news sentiment and key topics, finding that media sentiment and specific themes significantly influence IPO underpricing.

Table 4. Comparison of Textual Analysis-based Approaches.

Ref	Year	Technique Name	Limitation	Advantage
[21]	2020	Sentiment Analysis of IPO Prospectuses	Limited to short-term predictions; lacks integration with broader financial metrics.	Demonstrates that sentiment analysis can improve IPO stock movement prediction accuracy by 9.6%.
[23]	2021	Readability Analysis (Gradient Boost Trees)	Focused only on prospectus readability; does not incorporate sentiment or external media factors.	Shows that clearer IPO prospectuses reduce underpricing by mitigating information asymmetry.
[24]	2023	Textual and Financial Data Analysis	High computational cost; potential overfitting with complex machine-learning models.	Combining textual and financial data improves IPO underpricing prediction accuracy by 6.1% over financial-only models.
[25]	2020	Media Coverage Analysis	Focused on Chinese IPOs; findings may not be generalized to other markets or cultural contexts.	Reveals a negative relationship between media coverage and IPO underpricing, highlighting investor sensitivity to news.
[26]	2022	Sentiment and Topic Modeling (LDA, BERT)	Requires high-quality textual data; complex algorithms may be resource-intensive and difficult to implement.	Finds that media sentiment and specific topics significantly influence IPO underpricing using advanced NLP techniques.

Utilizing sentiment and textual analysis offers several key advantages. Enhanced predictive accuracy is achieved by incorporating textual data alongside financial metrics, providing a more comprehensive understanding of IPO dynamics. Additionally, improved readability and nuanced media coverage analysis help reduce information asymmetry, leading to more efficient market outcomes. Furthermore, techniques such as sentiment analysis and topic modeling provide advanced insights into investor sentiment and market trends, capturing elements that traditional financial metrics may overlook. These advantages underscore the growing importance of sentiment and textual analysis in financial research, particularly in the context of IPO performance prediction. Despite their advantages, utilizing sentiment and textual analysis for predicting IPO aftermarket performance faces challenges like inconsistent data quality, subjective text interpretation, complex and resource-intensive algorithms, overfitting risks, unpredictable investor behavior, varying media influence, practical integration difficulties, and regulatory and ethical concerns.

2.5. Data-Driven Approaches

Recent studies have explored various data-driven predictive models to forecast IPO aftermarket performance, as shown in Table 5. Kang et al. [27] examine the relationship between online search volumes and post-IPO stock returns, finding that lower pre-IPO

search volumes correlate with higher post-IPO returns, suggesting that less pre-IPO attention may indicate undervaluation. Sorkhi and Paradi [28] introduce a methodology combining Bayesian inference and Data Envelopment Analysis (DEA) to estimate the probability density function (PDF) of IPO stock prices in the short-term, addressing the challenge of predicting price uncertainty for firms with limited market history. Their approach iteratively updates prior beliefs using DEA to find comparable IPOs and Bayesian inference to refine the IPO's prior PDF, validated through backtesting. Turpanov [29] investigates the impact of optimistic analyst forecasts on long-run abnormal returns for South Korean IPOs, revealing an upward bias in earnings forecasts and a positive correlation between risk-adjusted returns and earnings forecast revisions, which supports the long-run underperformance hypothesis when controlling for risk.

Table 5. Comparison of Data-Driven Approaches.

Ref	Year	Technique Name	Limitation	Advantage
[27]	2021	Online Search Volume Analysis	Focused on pre-IPO attention; limited applicability in markets with low internet penetration.	Identifies a negative correlation between pre-IPO search volumes and post-IPO returns, suggesting undervaluation signals.
[28]	2020	Bayesian Inference and Data Envelopment Analysis (DEA)	Computationally intensive; relies on the availability of comparable IPOs for accurate predictions.	Provides a probabilistic framework to estimate short-term IPO price movements by iteratively refining prior beliefs.
[29]	2022	Analyst Forecast Impact Analysis	Limited to markets with a strong presence of analyst coverage; may not generalize to all regions.	Demonstrates that optimistic analyst forecasts lead to long-run IPO underperformance, supporting the long-run underperformance hypothesis.

Overall, while data-driven predictive models offer significant benefits in terms of accuracy and innovative data usage, they also present challenges related to data quality, model complexity, and inherent biases. These must be carefully managed to ensure reliable and effective IPO performance predictions.

After reviewing current methods, our approach will utilize ensemble learning classifiers, known for their superior accuracy and ability to handle non-linear interactions. Selection will focus on classifiers inherently capable of managing small datasets. Furthermore, we will apply several optimization techniques such as SMOTE, ANOVA F-value, and hyperparameter tuning. Lastly, determinants will be based on pre-listing prospectus characteristics and financial ratios, as these features are accessible to investors and decision-makers in advance.

3. Data

This research study utilizes a comprehensive dataset derived from the original prospectus documents of 68 companies listed on the Saudi stock market between 2004 and 2023. The dataset was meticulously collected and organized to serve the objective of this study, which is to utilize machine-learning techniques to predict the short-term underperformance of IPO.

The dataset encompasses two primary components:

1. **Prospectus Characteristics**: This part of the dataset includes features specific to the initial public offering of the companies. These features include the total number of offer shares, the total number of issued shares, the offer price, the total value of offer shares, the nominal value per share, the number of substantial shareholders, the total direct ownership of substantial shareholders pre- and post-offering (expressed in percentage), and the number of days allocated for individual subscribers. These features provide insights into the structure of the IPO and the company's ownership around the time of the offering.

2. **Financial Ratios**: This section comprises financial ratios calculated for a full fiscal year immediately preceding the IPO year. These ratios include gross profit margin, net profit margin, return on equity, return on assets, current assets-to-current liabilities ratio, liability-to-equity ratio, and earnings per share. These ratios show the company's financial health and performance before the public offering.

To ensure comparability, financial metrics were collected for each company's full audited fiscal year prior to their IPO. This approach allows for a standardized comparison across different companies, providing a complete picture of a company's financial performance in the period leading up to the IPO.

One important note regarding the data collection process is the exclusion of insurance and banking companies from the dataset. This decision was made due to the unique financial structure of such companies, which exhibit distinct financial ratios that are not directly comparable to those of other industries. For instance, insurance companies do not have comparable ratios such as profit margin or current assets-to-current liabilities ratio, among others. We ensure the analysis is based on a more homogeneous and comparable dataset by excluding insurance and banking companies, with a final dataset of 55 records. The details of the features extracted for the dataset are listed in Table 6.

Table 6. Details of features used for the dataset.

Feature	Description
Total Number of Offer Shares (TNOOS)	The total shares offered for sale during the IPO, indicating the IPO size and public equity distribution.
Total Number of Issued Shares (TNOIS)	The total shares issued, including those offered during the IPO and retained by original owners.
Offer Price (OP)	The price at which each share is offered during the IPO, determined by the company based on various factors.
Total Value of Offer Shares (TVOS)	The overall value of shares offered during the IPO, calculated as the offer price multiplied by the number of offer shares.
Nominal Value per Share (NVPS)	The face value of a share as stated in the company's corporate charter.
Number of Substantial Shareholders (NOSS)	The count of shareholders holding a significant portion of the company's shares.
Total Direct Ownership of Substantial Shareholders Pre- and Post-Offering (TDOS)	Indicates the percentage of shares held by substantial shareholders before and after the IPO.
Number of Days for Individual Subscribers (NDIS)	The duration of the subscription of shares by individual investors during the IPO.
Gross Profit Margin (GPM)	The proportion of revenue that remains after deducting the cost of goods sold, serving as an indicator of a company's profitability.
Net Profit Margin (NPM)	The percentage of revenue exceeding all company costs, including indirect expenses.
Return on Equity (ROE)	Measures company profitability by revealing profit generated with shareholder investments.
Return on Assets (ROA)	Indicates company profitability relative to total assets.
Current Assets to Current Liabilities (CACR)	Also referred to as the current ratio, this metric evaluates a company's capability to cover its short-term obligations using its short-term assets.

4. Methodology

The framework aims to predict IPO underperformance within the first month post-listing based on investor risk preference. Figure 1 illustrates that following typical data preprocessing steps such as cleansing, the framework first implements SMOTE, a widely used technique in data-driven decision-making for complex systems [30]. SMOTE creates synthetic instances for the minority class by interpolating between existing examples rather than merely duplicating them. This process balances the class distribution, thereby mitigating model bias towards the majority class and enhancing overall performance metrics, which are essential for robust decision-making in imbalanced and data-scarce environments. Additionally, by introducing variability through synthetic samples, SMOTE reduces overfitting, leading to more generalizable and robust models capable of accurately predicting IPO underperformance in data-scarce environments.
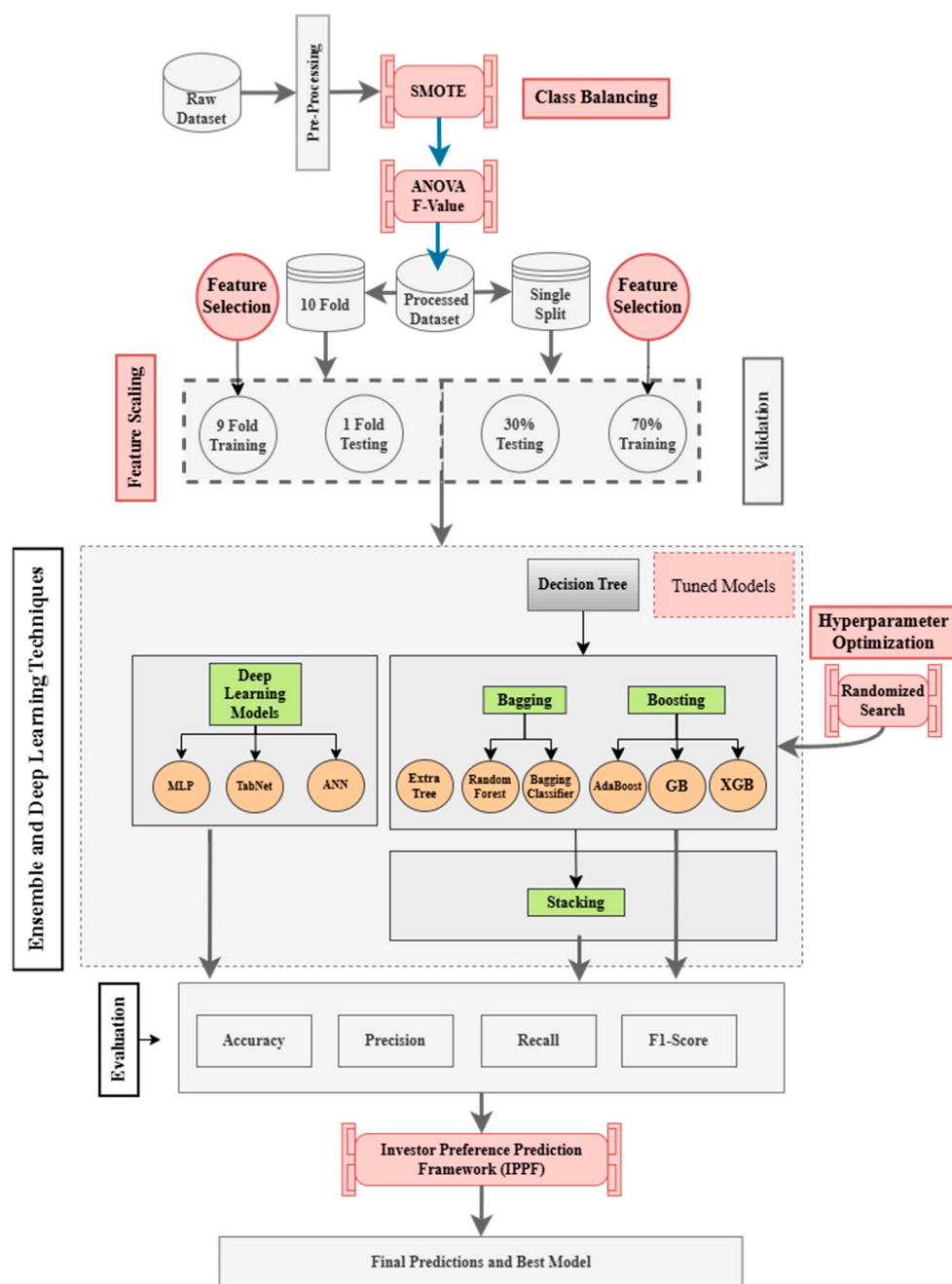

Figure 1. The Proposed Framework.

Utilizing SMOTE can also enhance the effectiveness of ANOVA F-value for feature selection, which is the next major step for our framework, by helping to attain the normality assumption required for ANOVA. By generating synthetic samples to balance the class distribution, SMOTE reduces the skewness and imbalance in the dataset, which in turn promotes a more normal distribution of data. To further validate feature importance, ANOVA F-value analysis is employed to assess variance among groups and determine statistical significance. If no significant variance is detected, it confirms the absence of a meaningful relationship between factors, preventing the inclusion of redundant or non-informative features and enhancing the credibility of our findings. A threshold of 0.2 is set to exclude less predictive features, improving model interpretability and preventing overfitting. This threshold was selected based on a grid search from 0.1 to 0.9 with a step of 0.1 to ultimately obtain 0.2 as the chosen level for best accuracy. Consequently, the combination of SMOTE and ANOVA F-value facilitates the identification of the most significant features, thereby improving the predictive performance and robustness of models, especially in scenarios involving small, imbalanced datasets such as those found in emerging financial markets.

Moreover, outlier analysis using the interquartile range (IQR) method confirmed that no significant outliers were present in the dataset. Additionally, Principal Component Analysis (PCA) was applied to manage feature correlations, reducing redundancy and enhancing the model's stability.

Next, the framework splits the dataset into 80% for training and 20% for testing (single-split validation), providing a straightforward and efficient means to evaluate model performance on unseen data. The same training and test datasets are used across all models to ensure consistency in evaluation. This method allocates a substantial part of the data for training purposes yet retains sufficient data to effectively evaluate the model's generalizability. Feature scaling is performed only on the training set after splitting and then applied to the test set to prevent data leakage. To increase the reliability of the assessment, the framework utilizes k-fold cross-validation, designating nine folds for training and one fold for testing. The k-fold cross-validation is a standard practice in data-driven modeling, reducing overfitting risks and ensuring comprehensive performance evaluation across diverse data subsets. By averaging the results from all folds, k-fold validation offers a comprehensive and reliable measure of the model's performance, which is particularly crucial for small datasets where variability can significantly impact outcomes.

The framework then starts training data on a set of thoroughly selected ensemble models known to handle small datasets more efficiently. Each model is trained on the same training dataset and subsequently evaluated on the test dataset, ensuring a uniform assessment protocol. The set includes Bagging Classifier (BC), Random Forest (RF), AdaBoost (Ada), Gradient Boosting (GB), XGBoost (XG), Stacking Classifier (SC), and Extra Trees (ET), with Decision Tree (DT) as a base estimator. Ensemble models are particularly effective for small datasets as they combine the strengths of multiple learning algorithms to improve predictive performance and robustness. By aggregating the outputs of various base learners, ensemble methods reduce the risk of overfitting and enhance generalization, which is crucial when dealing with limited data. In addition to ensemble learning models, we also evaluate deep-learning-based approaches such as Multi-Layer Perceptron (MLP), TabNet, and Artificial Neural Networks (ANN). Deep-learning models offer the advantage of automatically learning complex representations from raw data, potentially capturing intricate patterns that traditional ensemble models might overlook. However, deep-learning models typically require larger datasets for optimal performance and are more prone to overfitting when trained on limited data. To mitigate this issue, we employ regularization techniques, dropout layers, and hyperparameter tuning to enhance their generalization ca-

pability. Hyperparameter optimization plays a crucial role in enhancing model performance by fine-tuning parameters to achieve optimal results. Randomized Search is particularly effective for this task, as it efficiently navigates a broad range of hyperparameter values, identifying the best configurations without the prohibitive computational expense of grid search [31]. This approach exemplifies a strategic balance between efficiency and accuracy, making it well-suited for computationally constrained environments. After selecting the optimal hyperparameters, the models were trained using these configurations to ensure peak performance.

Metrics such as accuracy, precision, recall, F1 scores, and area under the receiver operating characteristic curve (AUC) are used to evaluate and compare the efficacy of different models. Accuracy determines the overall correctness of the model by calculating the proportion of correct predictions out of the total number of predictions (i.e., accuracy in predicting underperformance). In situations where datasets are imbalanced, relying solely on accuracy might not provide a clear picture. Precision and recall provide deeper insights. Precision measures the fraction of true positive predictions within all positive predictions, showing how well the model avoids false positives, while recall quantifies the fraction of true positives detected among all actual positives, emphasizing the model's capacity to identify true positives.

The F1 score, as the harmonic mean of precision and recall, mitigates the trade-offs between these two metrics, making it a crucial measure when both false negatives and false positives carry significant consequences. The AUC offers a comprehensive evaluation of the model's performance at various classification thresholds, reflecting its ability to effectively differentiate between classes. These metrics collectively ensure a robust and data-driven evaluation framework for decision-making under uncertainty.

At the heart of our proposed framework, we strategically propose a dynamic methodology—named Investor Preference Prediction Framework (IPPF)—to improve the decision-making process for IPO investments. Recognizing the underlying link between investing decisions and risk preferences, the framework acknowledges a wide range of investors, from risk-averse to risk-tolerant. It ranks the evaluated ensemble models based on investor risk preference, thereby tailoring investment strategies to individual risk profiles.

This dynamic methodology is particularly important in the context of IPO short-term underperformance prediction because the outcomes of such predictions can vary significantly based on different risk metrics. For example, a risk-tolerant investor seeks higher returns and wants to avoid false alarms about underperforming stocks. Hence, they value precision, which ensures most predicted underperformers are truly underperforming. On the other hand, a risk-averse investor prioritizes safety and wants to avoid missing any underperforming stocks. Hence, they value recall, which ensures most actual underperformers are correctly identified.

By employing IPPF, the framework dynamically adjusts the model evaluation criteria based on these risk preferences, ensuring that the selected model aligns with the investor's risk tolerance. This adaptability enhances the relevance and applicability of the model's predictions, as it allows investors to make more informed decisions that align with their risk appetite. Furthermore, this approach improves the overall robustness and flexibility of the prediction framework, making it more responsive to the diverse needs of different investors. In the volatile and unpredictable environment of IPO investments, such tailored decision-making support is crucial for optimizing returns and managing risks effectively.

The subsequent section will discuss each component of the proposed framework in further detail.

4.1. Class Imbalance, Feature Selection, and Hyperparameter Tuning

Various strategies have been implemented to enhance the proposed framework. One of the key approaches is the use of SMOTE, which effectively tackles the problem of class imbalance within our datasets. This technique generates synthetic samples for the under-represented minority class, helping to balance the dataset [30]. SMOTE creates synthetic examples along the line segments joining existing minority class instances. This strategy reduces bias toward the majority class and increases prediction accuracy in scenarios where instances belong to a minority class. The class distribution after applying SMOTE is shown in Table 7.

Table 7. Class Distribution Before and After Applying SMOTE.

Class Label	Original Distribution	Balanced Distribution (After SMOTE)
0	35	35
1	20	35

The proposed framework employs the ANOVA F-value, a statistical measure that assesses the significance of the differences in means among multiple groups [32]. In the machine-learning domain, the ANOVA F-value reveals important variables for predictive modeling by examining the variances of distinct classes. A higher ANOVA F-value indicates a more substantial impact of a factor on the target. Setting a threshold, such as 0.2 in this framework, eliminates less predictive characteristics, improving model interpretability and lowering the danger of overfitting.

Hyperparameter tuning is a crucial stage in enhancing machine-learning model performance. A hyperparameter tuning approach, Randomized Search, offers effective parameter space search techniques [31]. In contrast to other hyperparameter tuning techniques, Randomized Search selects a random subset of configurations, reducing processing costs and time [31]. This method offers a comprehensive search in the hyperparameter space for optimal settings that strike a reasonable balance between model variance and bias.

4.2. Random Search for Hyperparameter Optimization

In order to improve the performance of tree-based ensemble models and deep-learning models, we use Randomized Search Cross-Validation (RandomizedSearchCV) to perform hyperparameter tuning. Since random search explores a wide range of hyperparameter values to find optimal configurations, it is computationally efficient compared to exhaustive grid search. The algorithm optimizes hyperparameters based on multiple metrics, including accuracy, recall, precision, and F1 score, ensuring a balanced evaluation of model performance, particularly for handling imbalanced datasets. Each model is tuned based on the key parameters, such as the number of estimators, maximum depth, feature selection strategies, learning rate, etc., as shown in Table 8.

Table 8. Best Hyperparameters for Tree-Based Models.

Model	Best Parameters
Decision Tree (DT)	min_samples_split = 3, min_samples_leaf = 2, max_depth = None, criterion = 'entropy'
Random Forest (RF)	n_estimators = 20, min_samples_split = 7, max_features = 'log2', max_depth = 10
Bagging Classifier (BC)	n_estimators = 10, max_samples = 1.0, max_features = 0.5, bootstrap_features = True, bootstrap = False
AdaBoost (Ada)	n_estimators = 200, learning_rate = 1.0, algorithm = 'SAMME.R'
Gradient Boosting (GB)	subsample = 0.5, n_estimators = 100, min_samples_split = 9, min_samples_leaf = 5, max_features = 'sqrt', max_depth = 3, learning_rate = 0.2
XGBoost (XG)	subsample = 0.8, scale_pos_weight = 1.0, reg_lambda = 0, reg_alpha = 0.1, n_estimators = 100, min_child_weight = 2, max_depth = 5, learning_rate = 0.1, gamma = 0.3, colsample_bytree = 0.6
Extra Trees (ET)	n_estimators = 900, min_samples_split = 10, min_samples_leaf = 1, max_features = 'log2', max_depth = None
MLP	hidden_layer_sizes = (128, 64, 32), activation = 'relu', solver = 'adam', alpha = 0.0001, learning_rate = 'adaptive', batch_size = 32, max_iter = 500
TabNet	n_d = 8, n_a = 8, n_steps = 3, gamma = 1.3, lambda_sparse = 0.001, momentum = 0.02, optimizer_params = {'lr': 0.02}, max_epochs = 100
ANN	layers = [64, 32, 16], activation = 'relu', optimizer = 'adam', dropout_rate = 0.2, batch_size = 32, epochs = 200

4.3. Base Estimator

In machine learning, base estimators are essential as they serve as the foundational models for advanced ensemble methods. These initial models process the data first, and their outputs are typically integrated or improved using different ensemble strategies to boost the accuracy of predictions. The choice of base estimator significantly influences the overall effectiveness of the ensemble model. A well-chosen base estimator can capture essential patterns in the data, providing a robust foundation upon which ensemble methods can build. Among the various types of base estimators, decision trees have emerged as a popular and powerful choice due to their unique characteristics and adaptability.

The appeal of decision trees lies in their simplicity and interpretability [33]. They provide a clear and intuitive representation of how decisions are made, which is valuable for understanding the underlying patterns in the data. Additionally, decision trees can handle both numerical and categorical data, making them versatile tools for diverse datasets. Their ability to capture non-linear relationships and interactions between features further enhances their utility in complex domains such as finance. As base estimators, decision trees are not only effective on their own but also serve as the cornerstone for more advanced ensemble methods like Random Forests and boosting algorithms. These

ensemble techniques leverage the strengths of individual decision trees, combining them to produce models with improved accuracy and robustness. To summarize, DT is selected as the base estimator for several reasons:

1. **Simplicity and interpretability:** Decision trees are straightforward to understand and interpret, making them ideal for initial model building and understanding feature importance.
2. **Handling non-linearity:** DTs can capture non-linear relationships between features, which are common in financial datasets. This allows them to model complex interactions without requiring extensive data preprocessing or transformation.
3. **Versatility:** Decision trees can handle both numerical and categorical data, making them versatile tools in diverse datasets. This versatility is particularly valuable in financial data, where features can vary widely in type and scale.
4. **Foundation for ensembles:** As a base estimator, decision trees form the foundation for more complex ensemble methods like Random Forests and Boosting algorithms. Their ability to be combined into ensembles allows for improved predictive performance and robustness.
5. **Efficiency:** DTs are relatively fast to train and evaluate, which is crucial when conducting multiple iterations of model training and hyperparameter tuning, such as in Randomized Search and cross-validation processes.

4.4. Selection of Ensemble and Deep-Learning Classifiers

The choice of specific ensemble classifiers in this research is driven by their proven effectiveness in handling small and imbalanced datasets, as well as their ability to model complex relationships in financial data. The ensemble methods selected—BC, RF, Ada, GB, XG, SC, and ET—each bring unique strengths to the framework, enhancing predictive reliability and robustness in various market conditions.

- **Bagging Classifier:** Bagging helps to lower variance and prevent overfitting by creating multiple models, each trained on distinct subsets of the dataset. This technique is especially beneficial in environments with small datasets, where the risk of overfitting is typically higher.
- **Random Forest:** RF, which builds on the concept of bagging, constructs several decision trees and combines their outputs. It is renowned for its strong performance and capability to manage data with many dimensions, making it a suitable option for predicting IPO performance, particularly when dealing with complex interactions among numerous features.
- **AdaBoost:** AdaBoost, a type of boosting approach, trains models sequentially, each time concentrating on the instances that were previously misclassified. Its capability to learn from mistakes and adjust accordingly makes it effective for enhancing results on imbalanced datasets, a key factor in accurately predicting IPO underperformance.
- **Gradient Boosting and XGBoost:** Both GB and XG are powerful boosting methods that build models sequentially, each new model correcting the errors of its predecessor. XGBoost, in particular, is optimized for speed and performance, making it highly effective for large-scale data analysis. These methods are chosen for their ability to capture subtle patterns and interactions within the data.
- **Stacking Classifier:** Stacking leverages the strengths of multiple base models by combining their predictions using a metamodel. This approach is selected for its ability to synthesize diverse model insights, thereby improving prediction accuracy and robustness.
- **Extra Trees:** ET is similar to RF but differs in the way it splits nodes. It uses random splits rather than the best splits, which can lead to lower variance and improved

generalization on small datasets. ET is chosen for its efficiency and effectiveness in handling diverse feature sets.

- **Multi-Layer Perceptron (MLP):** MLP is a deep-learning-based artificial neural network that consists of multiple hidden layers with non-linear activation functions. It is particularly effective in capturing intricate relationships within data and learning hierarchical representations. MLP is well-suited for IPO performance prediction as it can uncover deep feature interactions that traditional machine-learning models may overlook. However, careful tuning of hyperparameters, such as the number of layers, neurons, and learning rate, is required to achieve optimal performance.

- **TabNet:** TabNet is a deep-learning model specifically designed for tabular data, leveraging attention mechanisms to perform feature selection dynamically during training. Unlike traditional tree-based methods, TabNet allows for interpretability by identifying which features contribute the most to predictions. Its ability to focus on relevant aspects of the dataset makes it a promising candidate for IPO forecasting, especially when feature importance is a crucial aspect of decision-making.

- **Artificial Neural Network (ANN):** ANN is a versatile deep-learning architecture composed of interconnected layers of neurons that learn patterns through backpropagation. It is particularly powerful for modeling complex and non-linear relationships within financial data. When applied to IPO performance prediction, ANN can effectively capture interactions between financial indicators and company-specific attributes, providing a robust alternative to conventional machine-learning approaches. Its effectiveness depends on factors such as network depth, activation functions, and regularization techniques.

The selected models were chosen for their ability to handle structured data, robustness against overfitting, and strong generalization capabilities, as shown in Table 9. Unlike heuristic methods like Binary Ant Colony Optimization (BACO), which focuses on feature selection and optimization [34], ensemble models provide superior classification accuracy and stability by leveraging multiple learners. Additionally, boosting techniques such as AdaBoost, Gradient Boosting, and XGBoost offer advantages in improving weak learners, while stacking enhances predictive performance by combining multiple models.

4.5. Validation and Evaluation Metrics

In this study, we utilize both single train/test split and k-fold cross-validation methods. The effectiveness of our classifiers is evaluated using metrics such as accuracy, precision, recall, and F1 score, which are commonly employed in classification tasks to provide a detailed assessment of model performance. These metrics are calculated based on the confusion matrix, as shown in Table 10. The confusion matrix is built on the following values:

Table 9. Comparison of Selected Classification Models and Alternative Heuristic Methods.

Model	Selection Criteria	Advantages	Limitations	Comparison to BACO and Heuristic Methods
Decision Tree (DT)	Simple, interpretable baseline	Easy to interpret, fast training	Prone to overfitting	Heuristic methods often optimize DT-based splits but may not generalize well
Bagging Classifier (BC)	Reduces variance, improves stability	Reduces overfitting, handles noise well	Not ideal for small datasets	BACO focuses on feature selection, whereas bagging improves stability
Random Forest (RF)	Strong performance on structured data	Robust, reduces overfitting, handles missing values	Computationally expensive	BACO selects optimal features, but RF is more robust for classification
AdaBoost (Ada)	Strong boosting performance	Good for weak learners, enhances accuracy	Sensitive to noisy data	BACO may struggle with boosting weak models effectively
Gradient Boosting (GB)	Handles non-linearity, improves accuracy	High predictive accuracy, handles complex patterns	Computationally intensive	GB builds models iteratively, unlike BACO's feature-based optimization
XGBoost (XG)	Best for structured data, fast training	Optimized for speed, regularized to avoid overfitting	Requires careful tuning	More efficient than BACO for structured datasets
Stacking Classifier (SC)	Combines multiple models for enhanced performance	Leverages strengths of different models	Complex to train and tune	BACO does not provide ensemble-based learning benefits
Extra Trees (ET)	Improves variance reduction	Faster than RF, robust to noise	Less interpretability	BACO does not offer variance reduction but optimizes feature selection
Multi-Layer Perceptron (MLP)	Deep-learning approach for tabular data	Learns complex patterns, handles high-dimensional data	Requires large datasets, tuning is challenging	BACO optimizes feature selection, while MLP extracts hierarchical features
TabNet	Attention-based learning for tabular data	Feature interpretability, automatic selection	Computationally expensive	BACO makes explicit feature selection, while TabNet learns feature importance dynamically
Artificial Neural Network (ANN)	General-purpose deep-learning model	Captures non-linear relationships, flexible architecture	Requires significant tuning, overfitting risk	BACO focuses on feature selection, whereas ANN builds deep representations

1. **True Positives (*tp*)**—correctly predicted underperforming stocks. In our context, a true positive occurs when the model predicts that a stock will underperform (positive

prediction), and the actual outcome is that the stock underperformed (positive label). This means that the model correctly identified and predicted stocks that experienced underperformance.

2. **True Negatives (*tn*)**—correctly predicted non-underperforming stocks. In this paper, a true negative occurs when the model predicts that a stock will not underperform (negative prediction), and the actual outcome is that the stock did not underperform (negative label). This means that the model correctly identified and predicted stocks that did not underperform (or performed well).

3. **False Positives (*fp*)**—incorrectly predicted underperforming stocks when they actually did not underperform, are also known as *Type* I *Errors*. A false positive occurs when the model predicts that a stock will underperform (positive prediction) when the actual outcome is that the stock did not underperform (negative label). This means that the model made an incorrect positive prediction, indicating that the stock would experience underperformance when this was not actually the case. For risk-tolerant investors who are making decisions that should maximize profits, this represents a possible missed opportunity.

4. **False Negatives (*fn*)**—incorrectly predicted non-underperforming stocks when they actually underperformed, are also known as *Type* II *Error*. A false negative occurs when the model predicts a stock will not underperform (negative prediction) when the actual outcome is that the stock underperforms (positive label). In our context, this means that the model made an incorrect negative prediction, failing to identify that the stock would experience underperformance when it was actually the case. For risk-averse investors who are making decisions that should limit loss, this represents a higher risk of actual losses.

Table 10. Confusion Matrix Structure.

Actual	0	*tn*	*fp* (Type I Error)
	1	*fn* (Type II Error)	*tp*
		0 (Non-Underperforming) Negative Prediction	1 (Underperforming) Positive Prediction
		Predicted	

In order to efficiently assess and compare the performance of IPO underperformance prediction models, this study employs a number of key performance metrics that capture different dimensions of model robustness and suitability to different investor risk preferences. Since these metrics concern trade-offs between false positives and false negatives, they allow us to understand how the model can generate accurate predictions while balancing.

Accuracy (A_c): Accuracy measures the overall correctness of the predictions or the proportion of total correct predictions. It is calculated as follows:

$$A_c = \frac{tp + tn}{tp + tn + fp + fn} \tag{1}$$

While a high accuracy score means that a model correctly predicts many instances, it fails to specify the types of prediction errors (false positives vs. false negatives) or account for how classes are distributed (such as underperforming vs. non-underperforming

stocks). In situations where there is a prediction of IPO underperformance, there is often a class imbalance, with one class (like underperforming stocks) being much rarer than the other (non-underperforming stocks). Under these conditions, a model could reach a high accuracy simply by predominantly predicting the more frequent class. For instance, in a scenario where 5% of stocks underperform, and 95% perform well, a model predicting "non-underperforming" for all cases would achieve 95% accuracy, yet it would be utterly ineffective in recognizing any underperforming stocks, resulting in no true positives.

Accuracy can be misleading because it ignores class proportions, hides critical errors, and lacks specificity. It does not account for the imbalance between classes, leading to a false sense of performance when the minority class is rarely predicted. For risk-averse investors who prioritize avoiding losses, high accuracy does not necessarily mean low false negatives. A model can achieve high accuracy by correctly identifying a large number of non-underperforming stocks (true negatives) while still missing many underperforming ones (false negatives). Moreover, accuracy alone does not provide insight into how well the model identifies underperforming stocks (recall) or the reliability of its positive predictions (precision). Metrics like recall and precision should be considered to obtain a true picture of the model's performance in predicting IPO underperformance.

Recall (R_e): Recall, also known as sensitivity, measures the model's ability to correctly identify actual positives (stocks that went below IPO price). It is calculated as follows:

$$R_e = \frac{tp}{tp + fn} \tag{2}$$

Risk-averse investors, who aim to limit losses, would prefer recall. Recall focuses on the proportion of true positives among all actual positives, minimizing false negatives. This ensures that most underperforming stocks are correctly identified, reducing the risk of missing out on stocks that would lead to losses. This aligns with their preference for avoiding the risk of actual losses by ensuring that underperformance is detected whenever it occurs.

Precision (P_r): Precision measures the proportion of positive identifications that were actually correct, or the proportion of correctly predicted positive instances out of all instances predicted as positive by the model. It is calculated as follows:

$$P_r = \frac{tp}{tp + fp} \tag{3}$$

Risk-tolerant investors who aim to maximize profits and can accept more risk would prefer precision. Precision measures the proportion of true positives among all positive predictions, minimizing false positives. This ensures that when the model predicts underperformance, it is highly likely to be correct, thus avoiding missed opportunities due to incorrect predictions of underperformance. This aligns with their preference for maximizing returns by accurately identifying stocks that are unlikely to underperform.

F1 Score (F_1): The F_1 score is the harmonic mean of precision and recall, providing a balance between them. It is calculated and simplified as follows:

$$F_1 = 2 \times \left(\frac{P_r \times R_e}{P_r + R_e} \right) = \frac{2 \times tp}{2 \times tp + fp + fn} \tag{4}$$

The F1 score differs from other statistical measures, such as the F-statistic, which is used in hypothesis testing. The F1 score is instrumental in balancing precision and recall. For risk-averse investors, a high F1 score indicates a model that effectively identifies stocks likely to decrease in value (thus should be avoided) while minimizing the incorrect labeling of stocks as risky (thus not missing out on potential gains).

A general or broader form for the F_1 score is the so-called F_β score [35], and it is formulated as follows:

$$F_\beta = \frac{(\beta^2 + 1) \times P_r \times R_e}{\beta^2 \times P_r + R_e} = \frac{(\beta^2 + 1) \times tp}{(\beta^2 + 1) \times tp + fp + \beta^2 fn} \tag{5}$$

where β is a positive real constant that allows unequal weighting between precision and recall. When $\beta = 1$, precision and recall are evenly balanced, leading to the regular F1 score. $\beta < 1$ favors precision, while $\beta > 1$ favors recall.

The decision-making process about investments in IPOs is fundamentally tied to an investor's risk preference. Investors typically span a spectrum from risk-averse, favoring the minimization of the probability of loss, to risk-tolerant, often willing to accept higher probabilities of loss for the potential of greater returns. Thus, in our study of predicting IPO underperformance based on the preference of the investor, the infinite space for β would make it more difficult to assess or quantify an investor's risk. We, therefore, define a new parameter r, called the risk preference factor, which takes any real number between 0 (favoring precision, i.e., risk-averse investors) and 1 (favoring recall, i.e., risk-tolerant investors). The real constant β can then be redefined as a function of r, ensuring consistency with the following interpretation:

$$\beta(r) = 10^{(1-2r)} \qquad 0 \leq r \leq 1 \tag{6}$$

The investor's risk preference now determines β. For a risk-averse investor, $r = 0$ would favor recall, whereas $r = 1$ would emphasize precision. Finally, $r = 0.5$ represents completely balanced precision and recall, leading back to the F_1 score. We also can use any real number in between as per investor risk bias. As a result, the model selection guarantees that the score appropriately reflects the trade-off between *fn* and *fp*, as evaluated by various investor profiles. We call this method IPPF, a systematic strategy for balancing P_r (the significance of avoiding *fp* predictions) and R_e (the value of avoiding *fn* predictions) to evaluate class-based machine-learning models for modeling investor risk in investment decisions.

We can also assess the discrepancy of a certain model in response to investor risk preference by drawing a straight line between P_r and R_e utilizing the risk preference factor r as follows:

$$L(r) = rP_r + (1-r)R_c \tag{7}$$

The slope of this straight line is mathematically the difference between P_r and R_e derived as follows:

$$\frac{\partial L}{\partial r} = P_r - R_c \tag{8}$$

The absolute value Equation (8) would range between 0, indicating a robust model or a model that is insensitive to investor risk preference, and 1, indicating a fragile model that is highly sensitive to investor risk preference. This would measure the discrepancy between precision and recall.

If we do not know the investor risk preference, we would prefer a model that minimizes the absolute difference between P_r and R_e, leading us to the following objective:

$$\Delta_{PR} = |P_r - R_c| \tag{9}$$

Thus, we would want a model that has a maximum F score, but we can penalize models that have high Δ as follows to assess robustness (ρ):

$$\rho = \frac{F_\beta}{\Delta_{PR}} \tag{10}$$

This measure, however, becomes problematic when $\Delta = 0$, leading to an undefined outcome. Thus, we would simply use the geometric mean [36] to measure the discrepancy between precision and recall as follows:

$$\Delta'_{PR} = \sqrt{P_r \times R_c} \tag{11}$$

This measure naturally penalizes large discrepancies between precision and recall. The model selection criteria, named robustness ratio, become the following:

$$\rho = \frac{F_\beta}{\Delta'_{PR}} \tag{12}$$

We can now assess the best models across investors' preference levels to find the best model when exact risk preference is unknown with the following measure:

$$\min_x \left(\prod_{j=1}^{m} \sum_{i=1}^{n} \rho_{ij} \cdot x_{ij} \right) \forall i \in \{1, 2, \ldots, n\}, \forall j \in \{1, 2, \ldots, m\} \tag{13}$$

where ρ_{ij} is the robustness ratio for model i at risk level j, n is the number of models, and m is the number of risk levels. x_{ij} is the binary decision variable that is equal to 1 if model i is selected for risk level j; otherwise, it is 0.

In summary, to find the best tree-based ensemble model for IPO predicting, the IPPF technique evaluates models through the perspective of the investor's risk preferences. We identify the most appropriate model by analyzing each model's recall and precision over the risk preference continuum, balancing these metrics based on the investor's risk tolerance.

Receiver Operating Characteristic (ROC): The ROC curve serves as a method to assess the performance of binary classification systems by displaying the compromise between the true positive rate and the false positive rate at different threshold levels. The area under the curve (AUC) summarizes these data into a single number. A perfect prediction is denoted by an AUC of 1, while an AUC of 0.5 indicates a performance no better than random guessing. Models with AUC values approaching 1 are considered more accurate, and a higher AUC value signifies better overall model performance.

4.6. Investor Preference Prediction Framework (IPPF)

IPPF is built around risk sensitivity, which is critical in financial decision-making. By quantitatively measuring risk preference from 0 (risk-averse) to 1 (risk-tolerant), our methodology provides a systematic strategy for balancing precision (the significance of avoiding false positives) and recall (the value of avoiding false negatives) in our prediction models.

The framework combines these ensemble methods with decision trees as base estimators to take advantage of the strengths of each of these methods to better predict a small, imbalanced dataset. The aim of this strategic selection of these models is to address the particular challenges of IPO underperformance prediction in emerging markets, ensuring that these models fit the individual preferences of the investors and help them make appropriate investment decisions.

4.7. Algorithm of Proposed Framework

Let $X \in R^{n \times p}$ denotes the feature matrix for n samples and p features and let $y \in R^n$ be the target vector. The set Θ comprises a finite collection of learning algorithms considered for model training. The scalar α signifies the predetermined threshold for feature selection based on performance metrics. We define M as the collection of models obtained from training algorithms in Θ on the dataset (X, y). The model yielding the highest performance according to predefined criteria is represented as M_{best}. Then, SMOTE is applied for data balancing to y. The functions S, F, R, and P correspond to standard scaling, feature selection (using ANOVA F-value), hyperparameter optimization (using Randomized Search), and performance evaluation, respectively. Single-split and k-fold cross-validation strategies are encapsulated by SS and CV, correspondingly. Sensitivity for investors' risk level (r) is conducted utilizing IPPF by calculating the adjusted f_β.

The algorithm of the proposed framework is shown in Algorithm 1:

Algorithm 1 Proposed Framework with IPPF

Input: Feature matrix X, target vector y, set of algorithms Θ, threshold α, investor's risk preference r
Output: Best performing model Mbest, performance metrics
procedure EvaluateModels(X, y, Θ, α, r)
    Xscaled $\leftarrow$ S(X)
    Apply SMOTE to balance classes in (Xscaled, y)
    Split (Xscaled, y) into training (Xtrain, ytrain) and testing (Xtest, ytest) using single split (SS)
    Xselected $\leftarrow$ F(Xtrain) ANOVA F-value feature selection with threshold α
    Initialize an empty list for Results
    for each θ **in** Θ **do**
        Perform k-fold cross-validation (CV) on Xselected, ytrain
        M $\leftarrow$ hyper-parameter tuning by Random Search R using θ on **Xselected, ytrain** with SS/CV
        Metrics $\leftarrow$ evaluate M on Xtest, ytest after final model training on the entire Xselected, ytrain
        Append (M, Metrics) to Results
    end for
    Mbest $\leftarrow$ P(M) evaluates models with best Metrics from Results
    Calculate fβ for each model using investor's risk preference r
    Update Mbest based on fβ and r using IPPF
    return Mbest and its Metrics
end procedure

---

5. Results and Discussion

The Results and Discussion section provides an overall evaluation of the model's predictive ability, using both single-split training and 10-fold cross-validation approaches. The analysis covers the in-depth analysis of the confusion matrix and corresponding results obtained during training as well as testing phases, providing essential insights into the model's efficacy in correctly identifying true positives, true negatives, false positives, and false negatives. The examination of single-split training results explains the model's behavior on the training set and also evaluates the model's generalization on new, unseen data. Furthermore, this section also explores the findings of 10-fold cross-validation, which

gives some insight into how the model can remain consistent and resilient across different regions of the dataset.

5.1. Models Training

In the training phase, the ensemble models performed well across various metrics. The DT model showed an exceptional true negative rate of 53.57% and achieved a solid balance in counts for false negatives (3.57%) and true positive outcomes (42.86%). RF follows the same trend, with high true negative and positive rates of 51.79% and 42.86%, respectively. The true negative rates for BC, AdaBoost Classifier, GB Classifier, and XGBoost are around 53.57%. Most notably, the ET Classifier demonstrated a significantly different trend, achieving 46.43% true positive and 48.21% true negative rates simultaneously. The SC presented a balanced performance, achieving a true negative rate of 50.00% while effectively managing false positives and false negatives.

Deep-learning models exhibited varied performance, with MLP, TabNet, and ANN showing different strengths and weaknesses. MLP achieved a moderate true positive rate of 37.5% but struggled with a higher false positive rate of 10.7%. TabNet, leveraging attention-based learning, showed a lower true negative rate (35.7%) and the highest false negative rate (21.4%), indicating difficulty in correctly identifying positive instances. ANN performed similarly to ET in terms of true negatives (48.21%) but had the highest false negative rate (26.79%), indicating a struggle in correctly classifying positive cases. The confusion matrix details of trained models are provided in Table 11.

Table 11. Confusion Matrix Details During Training.

Model	TN	TN%	FP	FP%	FN	FN%	TP	TP%
Decision Tree	30	53.57%	0	0.00%	2	3.57%	24	42.86%
Random Forest	29	51.79%	1	1.79%	2	3.57%	24	42.86%
Bagging Classifier	30	53.57%	0	0.00%	0	0.00%	26	46.43%
AdaBoost Classifier	30	53.57%	0	0.00%	0	0.00%	26	46.43%
Gradient Boosting Classifier	30	53.57%	0	0.00%	0	0.00%	26	46.43%
XGBoost Classifier	29	51.79%	1	1.79%	0	0.00%	26	46.43%
Stacking Classifier	28	50.00%	2	3.57%	2	3.57%	24	42.86%
Extra Trees Classifier	27	48.21%	3	5.36%	0	0.00%	26	46.43%
MLP	24	42.8%	6	10.7%	5	9%	21	37.5%
TabNet	20	35.7%	10	17.9%	12	21.4%	14	25%
ANN	27	48.21%	3	5.36%	15	26.79%	11	19.64%

The training results show that ensemble models perform well across all evaluation metrics. The BC, AdaBoost Classifier, and GB Classifier all achieve 100% accuracy, precision, recall, and F1 score. It shows that these models made flawless predictions on the training

data, accurately capturing both positive and negative cases. The DT, RF, XGBoost Classifier, and ET Classifier have somewhat lower accuracy, but their performance is exceptional, particularly regarding recall for detecting true positive cases. The SC's 93% accuracy, balanced recall, precision, and F1 scores show that it can effectively combine diverse base models.

In contrast, deep-learning models exhibited varied performances. MLP achieved 80% accuracy with balanced recall (80%) and F1 scores (79%), demonstrating its ability to capture underlying patterns but with some misclassifications. TabNet performed the worst among all models, achieving only 61% accuracy with a recall of 54%, indicating its difficulty in distinguishing true positive cases. ANN also struggled, with an accuracy of 67% and the lowest recall of 42%, suggesting challenges in capturing positive cases effectively despite a relatively high precision of 79%. The detailed evaluation metrics of trained models are provided in Table 12

Table 12. Evaluation Metrics Results Details During Training.

Model	Accuracy	Recall	Precision	F1
Decision Tree	96%	92%	100%	96%
Bagging Classifier	100%	100%	100%	100%
Random Forest	95%	92%	96%	94%
AdaBoost Classifier	100%	100%	100%	100%
Gradient Boost Classifier	100%	100%	100%	100%
XGBoost Classifier	98%	100%	96%	98%
Stacking Classifier	93%	92%	92%	92%
Extra Trees Classifier	95%	100%	90%	95%
MLP	80%	80%	77%	79%
TabNet	61%	54%	58%	56%
ANN	67%	42%	79%	55%

5.2. Models Testing

The testing confusion matrix provides insights into the models' generalization abilities by demonstrating how well they function on untested data. Notably, the BC, AdaBoost Classifier, and GB Classifier exhibit consistent and impressive results across TN%, FP%, FN%, and TP%. These models strike an outstanding balance between avoiding false positives and false negatives, which is crucial for accurate predictions. The highest TP% (57%) and TN% (35.71%) are found in the ET Classifier, indicating its proficiency in correctly predicting instances. The DT and XGBoost Classifier, although having lower TN% and TP%, display a very balanced performance. The SC, representing the union of models, demonstrates robustness with 28.57% TN% and 50.00% TP%, highlighting its ability to synthesize predictions effectively on the testing data.

Deep-learning models exhibit mixed generalization performance. MLP achieves a TN% of 35.7% with a balanced TP% of 35.7%, demonstrating its ability to generalize moderately well but with a higher FN% (28.6%), indicating misclassifications in positive cases. TabNet struggles with a low TP% of 21.4% and the highest FN% (42.9%), suggesting

difficulty in correctly predicting positive instances. ANN performs similarly, with a TP% of 21.43% and an FN% of 42.86%, highlighting its limitations in capturing true positives. The confusion matrix details of tested models are provided in Table 13.

Table 13. Confusion matrix Details During Testing.

Model	TN	TN%	FP	FP%	FN	FN%	TP	TP%
Decision Tree	3	21.43%	2	14.29%	2	14.29%	7	50.00%
Random Forest	4	28.57%	1	7.14%	3	21.43%	6	42.86%
Bagging Classifier	5	35.71%	0	0.00%	2	14.29%	7	50.00%
AdaBoost Classifier	3	21.43%	2	14.29%	2	14.29%	7	50.00%
Gradient Boosting Classifier	4	28.57%	1	7.14%	3	21.43%	6	42.86%
XGBoost Classifier	3	21.43%	2	14.29%	3	21.43%	6	42.86%
Stacking Classifier	4	28.57%	1	7.14%	2	14.29%	7	50.00%
Extra Trees Classifier	4	28.57%	1	7.14%	1	7.14%	8	57.14%
MLP	5	35.7%	0	0.0%	4	28.6%	5	35.7%
TabNet	4	28.6%	1	7.1%	6	42.9%	3	21.4%
ANN	5	35.71%	0	0.00%	6	42.86%	3	21.43%

The models' evaluation is critical for determining their practical usefulness in forecasting whether a newly traded stock would trade below its IPO price within one month. The BC and ET Classifiers emerge as standout performers with high accuracy (86%) and balanced recall, precision, and F1 scores. The SC, representing a combination of diverse models, achieves a commendable 79% accuracy and exhibits balanced recall, precision, and F1 scores. While the DT and AdaBoost Classifier achieve moderate accuracy (71%), their recall, precision, and F1 scores suggest room for improvement, especially in avoiding false negatives and maintaining precision.

Deep-learning models show mixed performance in this evaluation. MLP achieves 71% accuracy with a high precision of 100%, but its recall (56%) suggests that it struggles with capturing true positive cases, leading to a less balanced performance. TabNet performs the weakest, with only 50% accuracy and a recall of 33%, highlighting its difficulty in correctly predicting stocks trading below their IPO price. ANN performs slightly better than TabNet, with 57% accuracy and 33% recall, but its 100% precision suggests that while it correctly identifies some positive cases, it misclassifies many others. The detailed evaluation metrics of tested models are provided in Table 14.

The above discussion shows that BC is the best performer with high accuracy and precision. It shows its ability to correctly identify positive cases while limiting false positives to a minimum. Moreover, the ET Classifier performs admirably, with excellent accuracy, recall, precision, and F1 scores. These models differ from the others in that they can predict whether a newly listed stock will trade below its Initial Public Offering (IPO)

price within one month of its listing. On the other hand, the XGBoost Classifier displays competitive numbers but falls somewhat below top performance. In contrast, the DT model underperformed across evaluation metrics, suggesting limits in its capacity to represent the complexity of IPO price fluctuations.

Table 14. Evaluation Metrics Results Details During Testing.

Model	Accuracy	Recall	Precision	F1
Decision Tree	71%	78%	78%	78%
Bagging Classifier	86%	78%	100%	88%
Random Forest	71%	67%	86%	75%
AdaBoost Classifier	71%	78%	78%	78%
Gradient Boost Classifier	71%	67%	86%	75%
XGBoost Classifier	64%	67%	75%	71%
Stacking Classifier	79%	78%	88%	82%
Extra Trees Classifier	86%	89%	89%	89%
MLP	71%	56%	100%	71%
TabNet	50%	33%	75%	46%
ANN	57%	33%	100%	50%

BC and SC perform better because BC trains individual models on slightly varied data samples, creating a diverse ensemble. This strategy proved valuable in mitigating the risk of overfitting the limited dataset associated with IPO predictions. The ensemble's predictions, aggregated through averaging or majority voting, contribute to more stable and reliable outcomes, which is crucial in scenarios with small-scale data. The ET Classifier, employed in the ensemble, also excels in addressing challenges related to variance and overfitting, which are common concerns in the context of limited data. Its randomized decision-making process and comprehensive analysis of the entire learning sample result in de-correlated DT, effectively reducing variance and improving the model's ability to capture underlying patterns in the data.

In contrast, deep-learning models exhibit varying levels of effectiveness in handling IPO predictions. MLP achieves moderate accuracy with high precision but struggles with recall, indicating its difficulty in identifying all relevant positive cases. This suggests that while MLP is effective at making confident predictions, it may not generalize well in capturing all variations in IPO performance. TabNet, despite its architectural advantages for tabular data, underperforms significantly, demonstrating the lowest accuracy and recall. This suggests that its feature representation may not align well with the stock market data structure. ANN, while slightly better than TabNet, also struggles to balance precision and recall, indicating challenges in effectively learning from the limited dataset.

5.3. 10-Fold Cross-Validation

In predicting stock performance following IPOs, applying a 10-fold cross-validation technique adds a crucial level of consistency to the assessment procedure. In this method, the dataset is divided into 10, nine of which are used to train the models and one to test them. Each subset is utilized as the validation data precisely once during the ten repetitions of this process. This approach allows for an overall evaluation of the models' generalization abilities over distinct data portions, reducing biases that may occur based on a sole split. For stakeholders in the financial sector, 10-fold cross-validation is a dependable indicator of the consistency and dependability of models when it comes to IPO stock

prediction. The following indicators have been used to evaluate the models using 10-fold cross-validation: Mean, Median Interquartile Range (IQR), First Quartile Q1, Third Quartile (Q3), Whisker Low (WhisLo), Whisker High (WhisHi), and Fliers (Outliers above and below $Q1 \pm 1.5 \times IQR$)

The models perform consistently across measures, with mean accuracy ranging from 57% to 70%. The BC and ET models demonstrate outstanding accuracy, with averages of 70% and 69%, respectively. The recall values range from 63% to 78%, reflecting the balance sensitivity in forecasting stocks that fall below their IPO prices. Precision ratings for positive predictions are typically good, ranging from 61% to 76%, with the bagging model achieving an impressive precision mean of 76%, emphasizing its ability to minimize false positives. Lastly, the F1 scores range between 60% and 69%, indicating a well-rounded performance, with BC achieving the highest value.

Beyond accuracy, recall, precision, and F1 score, the analysis of additional metrics provides a more comprehensive understanding of each model's performance in predicting stock movements in the IPO market. The DT has a consistent and equal distribution across metrics with a moderate IQR for accuracy, recall, precision, and F1 score. The RF model has a balanced distribution across all measures, indicating dependability in multiple aspects of prediction. AdaBoost Classifier, despite a lower IQR in accuracy, demonstrates robust performance in recall, precision, and F1 score. The GB Classifier, XGBoost Classifier, and SC all have varied degrees of IQR, demonstrating trade-offs between different measures. While the ET Classifier has an accuracy IQR of 0%, other metrics show significant variability, indicating distinct strengths and shortcomings.

Deep-learning models exhibit mixed results in 10-fold cross-validation. MLP performs competitively, achieving a mean accuracy of 67%, along with strong precision (77%) and well-balanced F1 scores (68%), making it a viable alternative for IPO predictions. However, TabNet struggles significantly, yielding the lowest accuracy (44%) and recall (27%), suggesting that its feature representation may not effectively capture the patterns necessary for stock movement prediction. ANN also underperforms with a 53% accuracy and a recall of 42%, indicating difficulties in learning meaningful patterns from the dataset. The lower performance of deep-learning models, particularly ANN and TabNet, suggests that traditional ensemble-based models may still be more suited for IPO prediction tasks where data availability is limited and interpretability is essential. The detailed results of the 10-fold cross-validation are shown in Tables 15 and 16.

Table 15. Results for 10-Fold Cross-Validation.

	Accuracy	Recall	Precision	F1
Decision Tree	64%	63%	70%	62%
Bagging Classifier	70%	69%	76%	69%
Random Forest	63%	70%	64%	66%
AdaBoost Classifier	63%	78%	61%	67%
Gradient Boost Classifier	57%	67%	58%	60%
XGBoost Classifier	60%	68%	63%	63%
Stacking Classifier	64%	63%	66%	62%
Extra Trees Classifier	69%	74%	69%	68%
MLP	67%	64%	77%	68%
TabNet	44%	27%	35%	25%
ANN	53%	42%	38%	33%

Table 16. Variation Analysis for 10-Fold Cross-Validation Results.

Model	Metric	Mean	Median	IQR [1]	Q1 [2]	Q3 [3]	WhisLo [4]	WhisHi [5]	Fliers [6]
Decision Tree	Accuracy	64%	57%	14%	57%	71%	43%	86%	0
	Recall	63%	71%	38%	38%	75%	25%	100%	0
	Precision	70%	71%	15%	60%	75%	50%	75%	3
	F1	62%	62%	21%	52%	73%	33%	86%	0
Bagging Classifier	Accuracy	70%	71%	14%	57%	71%	43%	71%	2
	Recall	69%	67%	44%	50%	94%	33%	100%	0
	Precision	76%	71%	46%	54%	100%	50%	100%	0
	F1	69%	62%	22%	57%	79%	50%	100%	0
Random Forest	Accuracy	63%	57%	11%	57%	68%	43%	71%	3
	Recall	70%	71%	8%	67%	75%	67%	75%	4
	Precision	64%	60%	23%	50%	73%	33%	100%	0
	F1	66%	62%	16%	57%	73%	50%	75%	3
AdaBoost Classifier	Accuracy	63%	64%	14%	57%	71%	57%	86%	1
	Recall	78%	75%	31%	69%	100%	33%	100%	0
	Precision	61%	63%	7%	60%	67%	50%	75%	1
	F1	67%	67%	8%	67%	75%	57%	86%	1
Gradient Boost Classifier	Accuracy	57%	50%	29%	43%	71%	29%	86%	0
	Recall	67%	67%	44%	50%	94%	25%	100%	0
	Precision	58%	55%	32%	41%	73%	33%	100%	0
	F1	60%	59%	25%	50%	75%	29%	89%	0
XGBoost Classifier	Accuracy	60%	57%	11%	57%	68%	43%	71%	3
	Recall	68%	71%	8%	67%	75%	67%	75%	4
	Precision	63%	59%	21%	50%	71%	33%	100%	0
	F1	63%	67%	17%	58%	74%	40%	80%	1
Stacking Classifier	Accuracy	64%	64%	14%	57%	71%	57%	86%	1
	Recall	63%	67%	38%	38%	75%	25%	100%	0
	Precision	66%	67%	22%	53%	75%	33%	100%	0
	F1	62%	62%	21%	52%	73%	29%	89%	0
Extra Trees Classifier	Accuracy	69%	71%	0%	71%	71%	71%	71%	4
	Recall	74%	75%	33%	67%	100%	25%	100%	0
	Precision	69%	71%	8%	67%	75%	67%	80%	3
	F1	68%	71%	12%	67%	79%	50%	89%	1
MLP	Accuracy	50%	50%	14%	43%	57%	29%	71%	0
	Recall	61%	50%	73%	27%	100%	25%	100%	0
	Precision	57%	50%	20%	43%	63%	33%	67%	2
	F1	51%	59%	30%	35%	65%	29%	67%	0
TabNet	Accuracy	41%	43%	29%	29%	57%	14%	57%	0
	Recall	25%	29%	33%	0%	33%	0%	75%	0
	Precision	31%	50%	50%	0%	50%	0%	60%	0
	F1	27%	37%	40%	0%	40%	0%	67%	0
ANN	Accuracy	46%	50%	14%	43%	57%	29%	57%	1
	Recall	27%	0%	33%	0%	33%	0%	33%	2

[1]: IQR: Interquartile Range, [2]: Q1: First Quartile, [3]: Q3: Third Quartile, [4]: WhisLo: Whisker Low, [5]: WhisHi: Whisker High, and [6]: Fliers: Outliers above and below Q1 $\pm$ 1.5 $\times$ IQR.

In 10-fold cross-validation results, the BC is a good choice among the classifiers. It achieves the highest accuracy of 70%, indicating effectiveness in identifying stocks likely to

decline while maintaining a balanced trade-off between recall and precision. On the other hand, the ET Classifier demonstrates the second-highest accuracy of 69%. However, the existence of fliers and a non-zero IQR implies that its efficacy may vary between folds. BC appears as a noteworthy alternative with the most outstanding mean F1 score, indicating a solid balance between accuracy and recall despite the slightly lower recall score.

On the other hand, the RF model performs poorly compared to ET and BC, particularly regarding the recall rate with many fliers. Based on these results, cautious investors looking for a model capable of reliably detecting stocks that may fall below their IPO price while generating the fewest false alarms should pick the AdaBoost Classifier. However, the BC and ET Classifier should be considered viable alternatives, particularly for investors prepared to accept somewhat lower recall in exchange for higher accuracy and fewer false positives. For risk-averse investors seeking a model that consistently identifies stocks at risk of falling below their IPO price with minimal false negatives, the AdaBoost Classifier is the top recommendation based on these findings. However, practical consideration should also be given to the BC and ET Classifier as viable secondary options, especially for investors willing to trade off some recall for improved precision and fewer false positives.

Deep-learning models present a mixed picture in IPO stock prediction. The MLP model achieves a competitive accuracy of 67% and the highest precision of 77%, making it a strong contender for investors prioritizing precise predictions with minimal false positives. However, its slightly lower recall (64%) suggests some limitations in capturing all potential declining stocks. In contrast, TabNet significantly underperforms, with an accuracy of only 44% and a recall of 27%, indicating its struggle to extract meaningful patterns from IPO data. Similarly, ANN falls behind with a 53% accuracy and 42% recall, suggesting that deep-learning models may require more extensive feature engineering or larger datasets to perform optimally in this context. While MLP remains a viable choice for precision-focused investors, traditional ensemble models like BC and ET still offer better overall reliability, especially in handling small-scale IPO prediction tasks.

5.4. Receiver Operating Characteristic Curve (ROC)

In the context of this study, ROC analysis is an essential technique for assessing the binary classification performance of various machine-learning and deep-learning models. ROC curves are visual representations of each model's ability to balance true positive rates and false positive rates across different classification thresholds, providing insights into their discriminative power in predicting whether a newly listed stock will fall below its Initial Public Offering (IPO) price within a month of trading.

The BC has the highest AUC of 0.90, showing a more remarkable ability to balance true positive and false positive rates across different classification thresholds. RF and ET Classifiers follow closely, with AUC values of 0.89, indicating strong performance in the study's context. The SC shows constant discriminative capability with an average AUC of 0.83, whereas the GB Classifier obtains an AUC of 0.78, indicating significantly weaker discriminating power. The AdaBoost and XGBoost Classifiers had lower AUC scores of 0.73 apiece, indicating inferior performance. Finally, the DT Classifier falls behind with a score of 0.69, indicating the limits in its capacity to distinguish between positive and negative cases in the given predicting job. Among deep-learning models, MLP achieved an AUC of 0.87, showing competitive performance close to RF and ET, while ANN obtained an AUC of 0.80, demonstrating moderate discriminative power. However, TabNet performed poorly with an AUC of 0.69, aligning with DT in its weaker ability to differentiate between cases. The ROC curve of all the classifiers is shown in Figure 2.
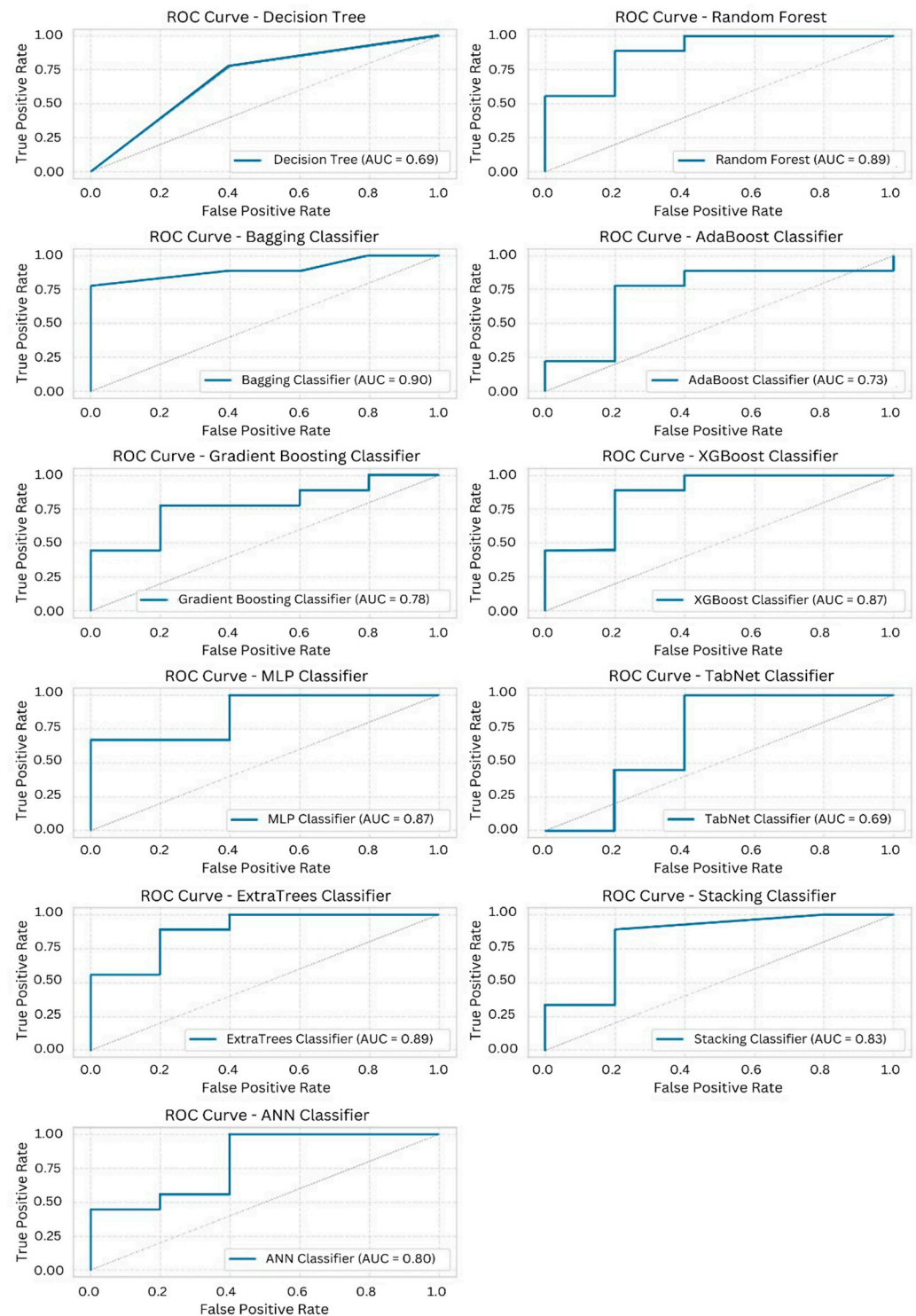
Figure 2. ROC curves of all the classifiers during testing.

5.5. Comparison

In this study, we propose a new framework to predict if a recently listed stock will trade below its IPO price in a month of trading using tree-based ensemble learning techniques. An essential component of our research is the limited utilization of tree-based ensemble methods in existing research on stock market prediction. We validate our suggested framework by comparing the results of our best model with those of Ampomah and Nyame [37], who also used tree-based ensemble approaches. Since their research used a different dataset, the comparative analysis of results is not prohibited because they used

the same tree-based ensemble methods. Further, this comparative study also shows that our framework is not only efficient but also unique in providing better results, notably for small datasets, which is an area underexplored in most studies.

As illustrated in Figure 3, the Extra Trees (ET) Classifier in our study outperformed the ET Classifiers from Ampomah and Nyame [37] across all evaluation metrics. Our ET Classifier achieved an accuracy of 86%, compared to 83.75% in their study. Notably, the recall improved from 81.25% to 88%, precision increased from 86.25% to 89%, and the F1 score rose from 83.69% to 88.48%. These results underscore the robustness and reliability of our framework, especially in predicting IPO underperformance in the short term.
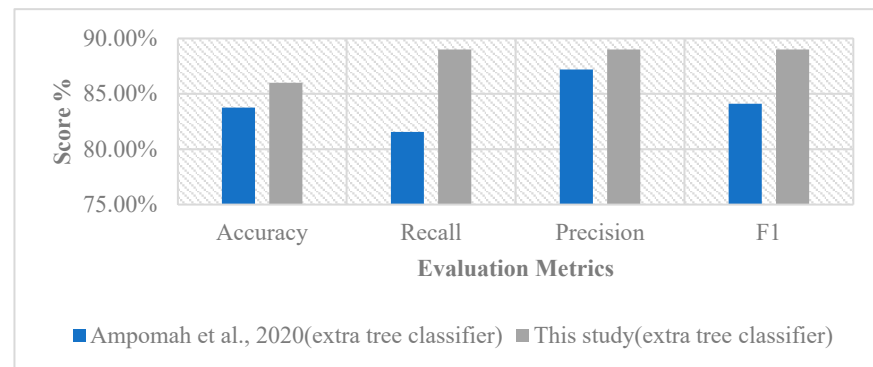
Figure 3. Comparison with existing studies using the test dataset [37].

Several factors may explain this performance gap. First, our framework incorporates advanced feature selection through ANOVA F-value and hyperparameter optimization using Randomized Search, which likely enhanced model accuracy and generalization. Additionally, the use of SMOTE to address class imbalance has contributed to improved recall, reflecting better sensitivity in identifying underperforming IPOs.

5.6. Risk Sensitivity and f_β Score Calculation Using IPPF

This study introduces IPPF to explore the effect of different investors' risk preferences on the outcome of the framework. It will be used in both single-split and 10-fold cross-validation to evaluate our suggested approach's risk sensitivity and robustness. By adding IPPF, we aim to enhance the decision-making process in complex financial systems, offering a more thorough knowledge of the model's performance across multiple data splits and assuring its usefulness in capturing the dataset's underlying patterns.

5.6.1. IPPF with Single-Split Validation

The results in Table 17 and Figure 4 indicate a clear shift in the chosen model when the risk level rises. The following is a discussion and interpretation based on the findings of single-split validation.

The Extra Trees Classifier outperforms other models in the risk range of 0 to 0.5, as determined by the f_β score. The f_β score remains constant at 0.89, indicating that the model's precision–recall balance does not change considerably or is not sensitive to β parameter alterations within the risk range. The BC is the preferable model, starting at a risk level of 0.6, with its f_β score continuously increases with each rise in risk. This trend continues until the maximum risk level is 1.0. The shift suggests that the Bagging Classifier is better suited to handling scenarios in which precision is increasingly essential— for example, missing out on predicting a price increase above the IPO level, which is costly (i.e., the opportunity cost is high) and a common concern for risk-tolerant investors. Transitioning from the ET Classifier to the BC at a risk threshold of 0.6 is an essential milestone in the use of the IPPF method. This means that as the risk level grows, the

BC model will likely outperform the ET Classifier in terms of precision, which becomes increasingly important as we focus more on minimizing false positives.

Table 17. Different risk levels with calculated measure scores for single-split validation.

Risk Level	*Max F*	Selected Model	Δ'_{PR}	Robustness Ratio σ
0	89.0%	ET	89.0%	1.00
0.1	89.0%	ET	89.0%	1.00
0.2	89.0%	ET	89.0%	1.00
0.3	89.0%	ET	89.0%	1.00
0.4	89.0%	ET	89.0%	1.00
0.5	89.0%	ET	89.0%	1.00
0.6	93.0%	BC	88.0%	1.06
0.7	96.0%	BC	88.0%	1.09
0.8	98.0%	BC	88.0%	1.11
0.9	99.0%	BC	88.0%	1.13
1	100.0%	BC	88.0%	1.14
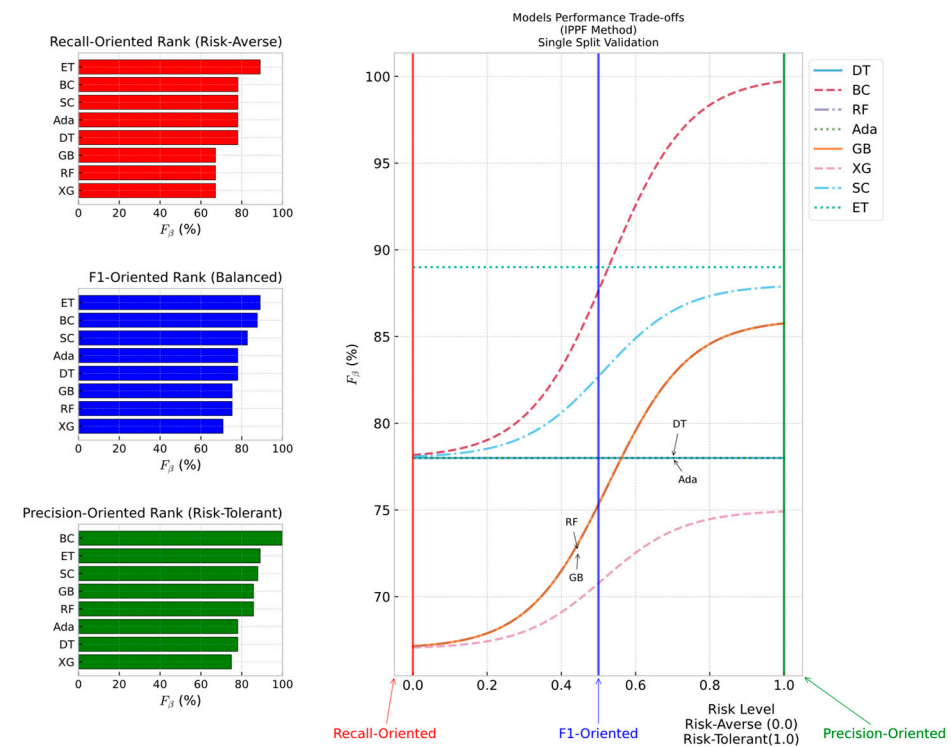
Figure 4. Representation of model selection adjusted for investor's risk level for single-split validation.

5.6.2. IPPF with 10-Fold Cross-Validation

For risk levels 0 to 0.3, the AdaBoost Classifier is consistently selected. At lower risk levels, when misclassifying a negative event (a price drop below the IPO level) is more penalized (lower β), this model performs best according to the f_β score, as shown in Figure 5. AdaBoost may be suitable for risk-averse investors, as the f_β score decreases with increased risk. At risk level 0.4, the ET Classifier is preferred, indicating a fair trade-off at this stage.
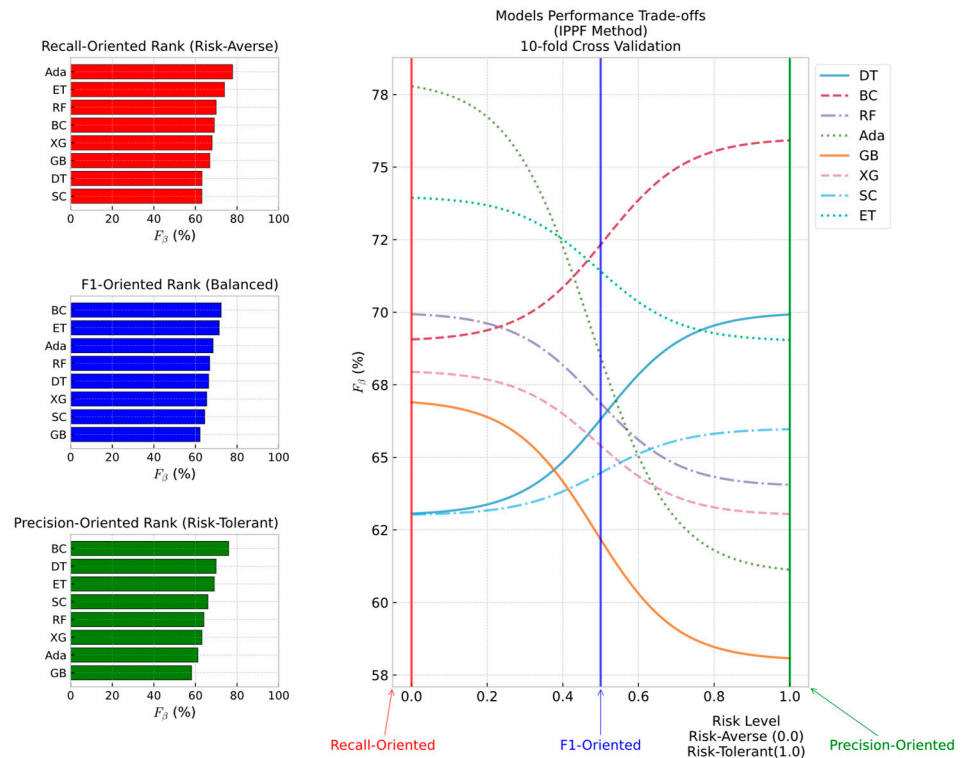
Figure 5. Representation of model selection adjusted for investor's risk level for 10-fold validation.

This preference, however, is fleeting, as the BC takes the lead from risk level 0.5 onward, showing that it handles the increasing focus on precision better than the ET and AdaBoost Classifiers. From a risk threshold of 0.5 to 1.0, the BC's f_β score improves and stays the preferred model, suggesting its relative strength in cases where missing a positive result (e.g., predicting a price gain beyond the IPO level) is more punished. This steady choice demonstrates the model's resilience in high-risk situations where precision becomes critical. The detailed risk levels with calculated measure scores for 10-fold validation are shown in Table 18.

Table 18. Different risk levels with calculated measure scores for 10-fold cross-validation.

Risk Level	*Max F*	Selected Model	Δ'_{PR}	Robustness Ratio σ
0	77.8%	Ada	69.0%	1.13
0.1	77.5%	Ada	69.0%	1.12
0.2	76.7%	Ada	69.0%	1.11
0.3	75.1%	Ada	69.0%	1.09
0.4	72.5%	ET	71.0%	1.02
0.5	72.3%	BC	72.0%	1.00
0.6	73.9%	BC	72.0%	1.03
0.7	75.0%	BC	72.0%	1.04
0.8	75.5%	BC	72.0%	1.05
0.9	75.8%	BC	72.0%	1.05
1	75.9%	BC	72.0%	1.05

Figure 6 shows the robustness ratios for both single-split and 10-fold validations for analysis.
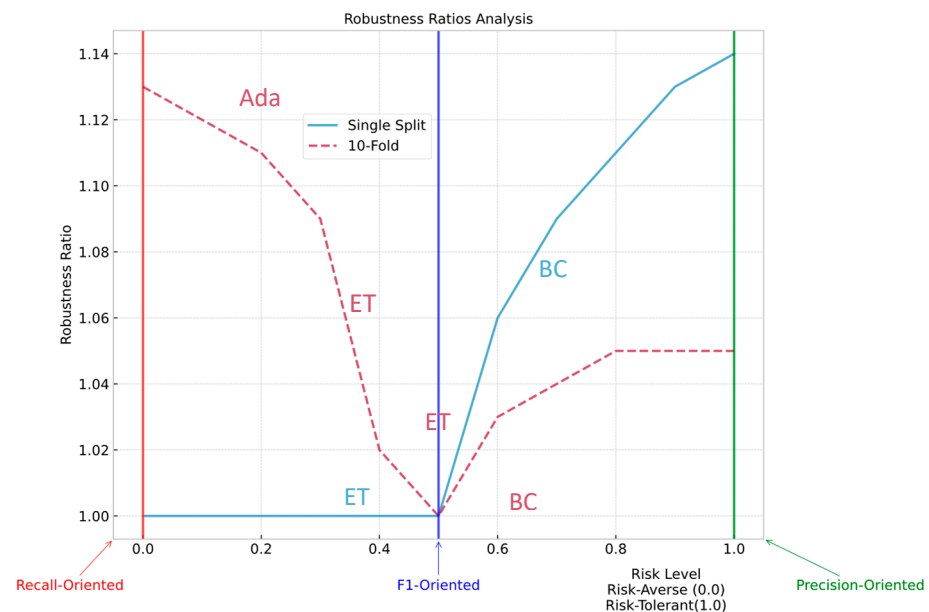
Figure 6. Robustness Ratio Curves for Both Single-Split and 10-Fold Validations.

The diagram in Figure 6 demonstrates how different models perform under different investor risk profiles that span from risk-averse (focused on recall only at r = 0) to risk-tolerant (focused on precision only at r = 1). The robustness ratio shows how much a forecasting system reacts to changes between accurate detection and total finding. Lower ρ ratings mean our models work better under all conditions, whereas higher ρ readings indicate they react strongly to changes. ET proves to be the most dependable for all risk levels below 0.5, while BC takes over as the best model for risk-tolerant investors. Another interpretation is that we should select AdaBoost only if we know for sure the investor is risk-averse since it is less robust with a higher ρ value. Also, we can interpret that if the risk preference of an investor is unknown, then it would be best to select either ET or BC for balanced and more robust results.

6. Research Limitations

A limitation of our study is the underlying assumption that the application of SMOTE will help attain the normality assumption required for ANOVA F-value feature selection. While SMOTE effectively balances class distributions and promotes normality, it may not always fully achieve the normality required for ANOVA, particularly in datasets with complex or highly skewed distributions. Although SMOTE mitigates class imbalance, it might not address the underlying skewness of the data, which could influence the feature selection process. If this assumption does not hold, the validity of the feature selection process could be compromised. To mitigate this risk, future work should explore alternative data preprocessing methods or evaluate the robustness of the feature selection process under different assumptions. Additionally, non-parametric methods, such as mutual information or decision tree-based approaches, could be employed to provide more accurate feature selection in cases where normality is difficult to achieve, even after applying SMOTE.

The dataset used in this study is limited to 55 records after excluding insurance and banking companies. While this decision was made to ensure consistency and relevance in the context of IPOs, the relatively small sample size may affect the generalizability of the results. A larger dataset would enhance the reliability and robustness of the findings, offering more statistical power and more representative insights. Furthermore, the study

focuses on data from the Saudi stock market, which may limit the external validity of the results to other markets with different economic conditions and regulatory environments.

7. Conclusions and Future Directions

This paper explores the challenging area of forecasting IPO results, particularly in overcoming the obstacles associated with insufficient data and class imbalance. The study significantly improves predicted accuracy by applying a modified framework that integrates SMOTE for class balancing with ensemble learning methods. The ensemble contains a variety of classifiers, including DT, RF, BC, AdaBoost, GB, XGBoost, ET, and SC. The results show that the ET Classifier performs better than other models in terms of accuracy and well-balanced recall, precision, and F1 scores in the single-split evaluation. Also, the BC achieves a high accuracy of 70% and well-balanced recall, precision, and F1 scores in 10-fold validation. In contrast, the MLP model showed the best performance among the deep-learning models, achieving an accuracy of 67% and a strong recall rate of 77%, indicating its effectiveness in this context.

The proposed framework outperforms existing tree-based ensemble learning techniques in single-split evaluation. This validation illustrates the impact of data-driven decision-making in complex systems and the improvements achieved in this domain by using ensemble approaches in our proposed framework for predictive modeling in stock market prediction.

Furthermore, IPPF reveals insights into the dynamic nature of decision-making based on varying investor risk preferences. In single-split validation, the ET Classifier is resilient for investors with low to moderate risk tolerance. However, the BC is preferable as risk tolerance increases due to its greater recall rate. This demonstrates a model's flexibility to varying investor choices. Lastly, the AdaBoost Classifier performs well in 10-fold cross-validation for risk-averse investors but loses efficacy as risk tolerance increases. The ET Classifier strikes a balance at a moderate risk level, but the BC performs best for moderate to high-risk tolerance. The dynamic change between these classifiers highlights the need to understand model behavior across a wide range of risk preferences and the relevance of cross-validation in creating stable and generalizable models.

While the current study yields noteworthy results and a promising conceptual framework, future research might explore more dimensions. Firstly, an analysis of the influence of feature engineering and the inclusion of domain-specific financial metrics might improve the model's prediction abilities. More advanced ensemble techniques, hybrid models integrating deep-learning algorithms, or using different base estimators may allow for more significant results. Additionally, future work will involve using a larger dataset to further validate the robustness and generalizability of the proposed framework. Lastly, exploring interpretability and explainability in the context of IPO prediction models would contribute to fostering trust and understanding among stakeholders, promoting data-driven decision-making in complex systems.

References

1.   Baba, B.; Sevil, G. Predicting IPO initial returns using random forest. *Borsa Istanb. Rev.* **2020**, *20*, 13–23. [CrossRef]

2. Kumar, A.; Sahoo, S. Do anchor investors affect long run performance? Evidence from Indian IPO markets. *Pac. Account. Rev.* **2021**, *33*, 322–346. [CrossRef]

3. Munshi, M.; Patel, M.; Alqahtani, F.; Tolba, A.; Gupta, R.; Jadav, N.K.; Tanwar, S.; Neagu, B.-C.; Dragomir, A. Artificial Intelligence and Exploratory-Data-Analysis-Based Initial Public Offering Gain Prediction for Public Investors. *Sustainability* **2022**, *14*, 13406. [CrossRef]

4. de Oliveira, C.H.F.; Rodrigues, C.L.; Juca, M.N. Determinants of IPO's Underpricing: A Systematic Review. *Contemp. Econ.* **2023**, *17*, 252–274. [CrossRef]

5. Ferdous, L.T.; Withanalage, N.P.; Zaman, A.A.Q. Review of short-run performance of initial public offerings in Australia. *Corp. Ownersh. Control* **2021**, *18*, 188–200. [CrossRef]

6. Rafique, A.; Quddoos, M.U.; Khadim, I.; Tariq, M. Financial and Operating Performance of Initial Public Offerings in Pakistan. *iRASD J. Econ.* **2020**, *2*, 35–42. [CrossRef]

7. Mutai, J. Effects of firm characteristics on the post-IPO performance by listed companies on the Nairobi Securities Exchange (NSE), Kenya. *Afr. J. Educ. Sci. Technol.* **2019**, *5*, 160–170.

8. Michel, A.; Oded, J.; Shaked, I. What determines institutional investors' holdings in IPO firms? *Int. Rev. Financ.* **2021**, *21*, 1302–1333. [CrossRef]

9. Mittal, S.; Verma, S. The Model Predictability Power to explain Underpricing in Bookbuild IPOs: A Study of Indian Capital Market. *Financ. India* **2022**, *36*, 663–680.

10. Ong, C.Z.; Mohd-Rashid, R.; Taufil-Mohd, K.N. Do institutional investors drive the IPO valuation? *Borsa Istanb. Rev.* **2020**, *20*, 307–321. [CrossRef]

11. Quintana, D.; Sáez, Y.; Isasi, P. Random forest prediction of IPO underpricing. *Appl. Sci.* **2017**, *7*, 636. [CrossRef]

12. Emidi, C.; Galán, S. Prospectus Content as Predictor of IPO Outcome: A Topic Model Approach. May 2022. Available online: https://lup.lub.lu.se/student-papers/record/9083567/file/9083574.pdf (accessed on 10 February 2025).

13. Dhini, A.; Sondakh, L. Predicting stock return of initial public offering in Indonesia stock exchange. *AIP Conf. Proc.* **2024**, *2710*, 050011.

14. Supsermpol, P.; Huynh, V.N.; Thajchayapong, S.; Chiadamrong, N. Predicting financial performance for listed companies in Thailand during the transition period: A class-based approach using logistic regression and random forest algorithm. *J. Open Innov. Technol. Mark. Complex.* **2023**, *9*, 100130. [CrossRef]

15. Ni, S. Predicting IPO Performance from Prospectus Sentiment. *BCP Bus. Manag.* **2023**, *38*, 3063–3075. [CrossRef]

16. Sonsare, P.M.; Pande, A.; Kumar, S.; Kurve, A.; Shanbhag, C. A Comparative Analysis of Machine Learning Algorithms for Android Malware Detection. *Procedia Comput. Sci.* **2023**, *220*, 763–768. [CrossRef]

17. Neghab, D.P.; Cevik, M.; Basar, A. Identifying the Factors Influencing IPO Underpricing using Explainable Machine Learning Techniques. In Proceedings of the 36th Canadian Conference on Artificial Intelligence, Montreal, QC, Canada, 5–9 June 2023. [CrossRef]

18. Neghab, D.P.; Bradrania, R.; Elliott, R. Deliberate premarket underpricing: New evidence on IPO pricing using machine learning. *Int. Rev. Econ. Financ.* **2023**, *88*, 902–927. [CrossRef]

19. Lubis, R.; Sadalia, I.; Irawati, N. The Influences of Prospectus Information and Macroeconomics on Initial Returns to Companies that Undergo Initial Public Offering (IPO) on the Indonesia Stock Exchange (IDX). In Proceedings of the 7th Global Conference on Business, Management, and Entrepreneurship (GCBME 2022), Bandung, West Java, Indonesia, 8 August 2022; Atlantis Press International BV: Dordrecht, The Netherlands, 2024; Volume 255. [CrossRef]

20. Arora, N.; Singh, B. Corporate governance and underpricing of small and medium enterprises IPOs in India. *Corp. Gov.* **2020**, *20*, 503–525. [CrossRef]

21. Hussein, M.; Zhou, Z.G.; Deng, Q. Does risk disclosure in prospectus matter in ChiNext IPOs' initial underpricing? *Rev. Quant. Financ. Account.* **2020**, *54*, 957–979. [CrossRef]

22. Ly, T.H.; Nguyen, K. Do words matter: Predicting IPO performance from prospectus sentiment. In Proceedings of the 2020 IEEE 14th International Conference on Semantic Computing (ICSC), San Diego, CA, USA, 3–5 February 2020; pp. 307–310. [CrossRef]

23. Chi, K.; Li, K. IPO Underpricing and Prospectus Readability: A Machine Learning Approach. *J. Appl. Bus. Econ.* **2021**, *23*, 151–159. [CrossRef]

24. Katsafados, A.G.; Androutsopoulos, I.; Chalkidis, I.; Fergadiotis, G.N.; Leledakis, E.; Pyrgiotakis, E.G. Textual information and IPO underpricing: A machine learning approach. *J. Financ. Data Sci.* **2023**, *5*, 100–135. [CrossRef]

25. Zou, G.; Li, H.; Meng, J.G.; Wu, C. Asymmetric Effect of Media Tone on IPO Underpricing and Volatility. *Emerg. Mark. Financ. Trade* **2020**, *56*, 2474–2490. [CrossRef]

26. Fedorova, E.; Druchok, S.; Drogovoz, P. Impact of news sentiment and topics on IPO underpricing: US evidence. *Int. J. Account. Inf. Manag.* **2022**, *30*, 73–94. [CrossRef]

27. Kang, H.G.; Bae, K.; Shin, J.A.; Jeon, S. Will data on internet queries predict the performance in the marketplace: An empirical study on online searches and IPO stock returns. *Electron. Commer. Res.* **2021**, *21*, 101–124. [CrossRef]

28. Sorkhi, S.; Paradi, J.C. Measuring short-term risk of initial public offering of equity securities: A hybrid Bayesian and Data-Envelopment-Analysis-based approach. *Ann. Oper. Res.* **2020**, *288*, 733–753. [CrossRef]

29. Turpanov, A. The Long-Run Performance of Initial Public Offerings South Korea Case. Master's Thesis, M. Narikbayev KAZGUU University, Nur-Sultan, Kazakhstan, 2022; pp. 356–363.

30. Polydouri, A.; Vathi, E.; Siolas, G.; Stafylopatis, A. An efficient classification approach in imbalanced datasets for intrinsic plagiarism detection. *Evol. Syst.* **2020**, *11*, 503–515. [CrossRef]

31. Bergstra, J.; Bengio, Y. Random search for hyper-parameter optimization. *J. Mach. Learn. Res.* **2012**, *13*, 281–305.

32. Song, X.; Liu, X.; Liu, F.; Wang, C. Comparison of machine learning and logistic regression models in predicting acute kidney injury: A systematic review and meta-analysis. *Int. J. Med. Inform.* **2021**, *151*, 104484. [CrossRef] [PubMed]

33. Sharma, K.; Bhalla, R. Stock Market Prediction Techniques: A Literature Review. *Int. J. Res. Appl. Sci. Eng.* **2023**, *11*, 1793–1800. [CrossRef]

34. Wan, Y.; Wang, M.; Ye, Z.; Lai, X. A feature selection method based on modified binary coded ant colony optimization algorithm. *Appl. Soft Comput. J.* **2016**, *49*, 248–258. [CrossRef]

35. Wegier, W.; Koziarski, M.; Wozniak, M.; Wegier, W. Multicriteria Classifier Ensemble Learning for Imbalanced Data. *IEEE Access* **2022**, *10*, 16807–16818. [CrossRef]

36. Lu, Y.T.; Chao, H.J.; Chiang, Y.C.; Chen, H.Y. Explainable Machine Learning Techniques to Predict Amiodarone-Induced Thyroid Dysfunction Risk: Multicenter, Retrospective Study With External Validation. *J. Med. Internet Res.* **2023**, *25*, e43734. [CrossRef] [PubMed]

37. Ampomah, E.K.; Qin, Z.; Nyame, G. Evaluation of tree-based ensemble machine learning models in predicting stock price direction of movement. *Information* **2020**, *11*, 332. [CrossRef]