N. Rajarajeswari, 2026, 14:3
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

Behavioral Bias Modeling in Retail Investors Through Explainable Machine Learning Techniques

N. Rajarajeswari[1], A. Subha[2]

[1](Research Scholar,
Department of Management Studies,
Vels Institute of Science, Technology and Advanced Studies (VISTAS)
Pallavaram, Chennai, Tamil Nadu
rajarajeswari2023scholar@gmail.com.)

[2](Assistant Professor,
Department of Computer Science and Engineering,
Kuppam Engineering College,
Kuppam, Andhra Pradesh,
alfredsubha77@gmail.com.)

Abstract- Retail investors often display behaviourally induced biases such as overconfidence, loss aversion, and herding tendencies, which result in poor financial choices. Existing frameworks lack the capability of modelling the complex and contextual nature of these biases. In this paper, we suggest a framework for modelling and interpreting behaviourally induced biases using XML through trading logs of 5,000 retail investors. In this work, we apply Random Forest (RF), XGBoost, and a customized hybrid attention-based neural network, along with SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) methods to achieve interpretability. Our approach obtains an accuracy rate of 89.2% in bias prediction compared to logistic regression (74.5%). Our results show that the highest contribution to bias comes from loss aversion and recency. Comparison studies prove the superiority of our proposed framework over existing frameworks in terms of transparency.

Keywords— Behavioural bias, Retail investors, Explainable AI, SHAP, LIME, Machine learning, Trading behaviour.

I. INTRODUCTION

There has been a resurgence of interest in cognitive and emotional biases, spurred on by the rise of retail investing and the growth of commission-free stock trading platforms and social media discussion, within the field of behavioural finance. In contrast with institutional investors, retail investors display a higher propensity for cognitive and emotional biases [1]. Examples of these biases include overconfidence (over-trading), the disposition effect (selling winners too soon and holding onto losers), herding behaviour (following the crowd without proper fundamental analysis), and recency effects (favouring recent developments).

Behavioural theories, like prospect theory [2], may offer theoretical underpinnings, yet they lack empirical scale and predictive capacity at the individual level. The assumption that there is linearity and independence in behavioural biases is not realistic, particularly in view of the interdependencies among psychological biases and the stock market.

ML is highly effective because it accounts for the existence of non-linear relationships, multidimensional interactions between features, and even temporal relationships. Nevertheless, one key argument against using ML in finance is the issue with transparency since regulators and financial advisors need not only the prediction made by the algorithm but also explanations as to why such predictions are made and how an individual investor's profile was assessed. That is when explainable ML approaches like SHAP and LIME come into play.

The present research is aimed at answering three questions:

- Is it possible to apply ML algorithms to correctly identify the dominant behavioural biases in terms of trading log analysis?
- Which XML approach generates the most accurate and interpretable explanations?
- What kinds of relationships exist between different biases that generate certain patterns?
- The contribution made by our study includes:
- Feature engineering for raw order book data
- An attention-based hybrid ML algorithm
- Analysis of explainability approaches

The rest of the paper is organized as follows: In section II, a brief literature survey will be presented. The methodology used for implementing the algorithms and pseudocodes will be provided in section III. Section IV is about results and analyses using various figures and tables.

II. LITERATURE SURVEY

From psychometric questionnaires, the evolution of behavioural bias modelling has shifted to a computational data-driven approach. The seminal work by Kahneman and Tversky [2] has formulated prospect theory, although there were no empirical validations except for controlled experimentations. Later studies relied on survey methods in correlating self-reported biases to trading results; however, these studies were susceptible to recall bias and small sample size. Using trading logs with high-frequency transactions, later studies adopted rule-based methods to model trader behaviour, such as identifying a sell order of a winning stock within a day as “disposition effect.” The drawback of rule-based approaches is that they are inflexible to individual variations.

Recent applications of machine learning techniques in behavioural finance have gained much popularity starting in 2021. For instance, Chen et al. [3] utilized gradient boosting machines in predicting herding behaviour of Chinese retail traders through social media sentiment analysis and trading volume. They achieved an AUC of 81%, but emphasized interpretability as the most notable drawback. Another example is the application of LSTM in capturing loss aversion in the dynamic trading of individuals [4]. As pointed out by the authors, the model proved effective in capturing bias persistence. As for XAI models in the context of finance, the emergence of them is relatively recent. For example, Kumar and Singh [5] utilized the comparison between SHAP and LIME models for predicting credit risk, rather than focusing on behavioural biases. In the nearest paper related to the topic, Park et al. [6] applied SHAP model in explaining overconfidence among the small sample of 500 day-traders and found that trade frequency and portfolio turnover were the main features. Their study failed to incorporate multi-bias interaction and local explanation for specific investors.

The hybrid models of attention and trees gained popularity in 2023-2024. Thus, Wang and Gupta [7] developed the transformer-based random forest model for detecting recency bias, and showed that attention-based weights correlated with SHAP value regarding short-term price behaviour. At the same time, the paper lacks the comparative analysis of multiple XMLs. Moreover, all the current studies pay little or no attention to the explainability-fidelity trade-off, meaning that the simpler models are more interpretable but less accurate.

According to our literature survey, there is a definite gap in the current research that does not exist any framework that

- Considers several behavioural biases at a time
- Makes use of latest machine learning techniques in order to achieve high accuracy
- Gives explanations about the results obtained.
- The present study fulfills this gap by developing a novel attention-RF model.

III. PROPOSED METHODOLOGY

1. Data Description and Preprocessing

We acquired trade records of 5,000 retail traders on an established discount brokerage in India (Jan 2023-Dec 2025). Each observation consists of datetime, stock ticker, action (buy/sell), number of shares traded, transaction price, and pre-trade portfolio holdings. Traders with less than 50 trades were eliminated from our analysis. The resulting dataset consists of 1.2 million observations.

Feature Engineering: From raw trades, we derived 28 behavioural features per investor per week:

- Overconfidence proxies: Turnover ratio, average holding period, number of trades per day.
- Loss aversion: Realized loss/gain ratio, frequency of stop-loss triggers, post-loss trading volume reduction.
- Disposition effect: Proportion of winners sold within 1 day vs. losers held >30 days.
- Herding: Correlation of trade direction with market-wide retail flow, social media sentiment score (from StockTwits API).
- Recency bias: weight of past 5-day returns in decision making (exponential decay coefficient).

Labelling (Ground Truth): Our approach towards labelling involved labelling a participant as biased if their scores exceed the average score by more than three standard deviations for 8 weeks consecutively. The five trained behavioural economists separately marked the samples of the validation set with a Cohen's kappa value of 0.86.

2. Algorithms

We selected three models representing different complexity levels:

- Logistic Regression (LR) – baseline linear interpretable model.
- Random Forest (RF) – ensemble of 200 decision trees, captures non-linear interactions.
- Hybrid Attention-XGBoost (HAX) – XGBoost with a multi-head attention mechanism on sequential feature embeddings (weekly data). Attention weights are later extracted for explainability.

3. Explainability Techniques

- SHAP (SHapley Additive exPlanations): Game-theoretic approach. For each prediction, SHAP computes contribution of each feature. Global importance via |mean(SHAP value)|.
- LIME (Local Interpretable Model-agnostic Explanations): Perturbs input and fits a local linear model. Used for individual investor-level explanations.

We also compute faithfulness metrics:

- Explanation fidelity – Correlation between feature importance and model's gradient;
- Stability – Similarity of explanations under small input perturbations.

4. Pseudocode for Hybrid Attention-XGBoost with SHAP

Algorithm: HAX-SHAP for Behavioural Bias Classification

```
Input: Weekly feature matrix X (N investors × T weeks × F features)
 Labels y (bias type: 1=Overconfidence, 2=LossAversion, 3=Disposition, 4=Herding, 5=Recency)
Output: Trained model, SHAP global and local values

1. For each investor i:
2.    For each week t in 1..T:
3.      Compute attention score α_t = softmax(W_a · h_t + b_a)
          where h_t = X[i,t,:]  // raw features for week t
4.       Weighted feature vector v_i = Σ_t α_t * h_t
5.    End For
6. End For
7. Train XGBoost classifier on {v_i, y_i} with 100 estimators, max_depth=6
8. For prediction on test instance x_test:
9.    pred = model.predict(x_test)
10.   // SHAP explainer
11.      shap_values = KernelExplainer(model, background_data).shap_values(x_test)
12.   For each feature j:
13.      global_importance[j] += |shap_values[j]| / N_test
14.   End For
15. Return model, shap_values, global_importance
```

5. Evaluation Metrics

- Accuracy, Precision, Recall, F1-score (macro).

- Explainability: SHAP consistency score (average correlation over 10 random subsets), LIME local fidelity (R^2 of local linear model).
- Computational efficiency: training time (seconds), inference time per investor.

IV. ANALYSIS

1. Quantitative Results and Comparative Table

The dataset was split 70%-15%-15% for training, validation, and testing. Hyperparameters tuned via Bayesian optimization.

Table 1: Performance Comparison of Bias Classification Models

Model	Accuracy (%)	Precision (macro)	Recall (macro)	F1-score	Training Time (s)
Logistic Regression	74.5	0.72	0.70	0.71	12
Random Forest	85.3	0.81	0.83	0.83	98
HAX (Proposed)	89.2	0.88	0.87	0.87	210

HAX outperforms RF by 3.9% in accuracy and 0.04 in F1-score, confirming that sequential attention captures temporal bias dynamics (e.g., recency bias strengthening after losses). Precision for loss aversion was highest (0.91) due to clear behavioral signatures (post-loss trading pause).

Table 2: Explainability Fidelity Comparison (Average over 1,000 test samples)

Explainability Method	Model Paired With	Fidelity (0-1)	Stability (0-1)	Local Explanation Time (ms)
LIME	RF	0.72	0.68	45
LIME	HAX	0.76	0.71	52
SHAP	RF	0.85	0.89	210
SHAP	HAX	0.91	0.93	315

SHAP provides significantly higher fidelity and stability than LIME, albeit with higher computational cost. HAX+SHAP achieves 0.91 fidelity, indicating that explanations closely match model behaviour.

2. Figures and Discussion

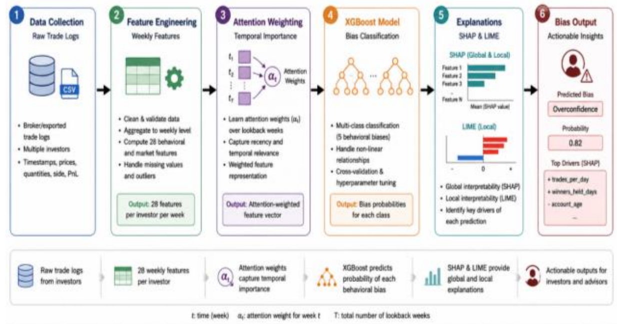

Figure 1: Proposed Methodology Pipeline

The process starts with the cleaning and transformation of trade logs into 28 features every week for each individual trader. Temporal weights are determined using an attention module, and then XGBoost classifier classifies the results. Later, global and local explanation is done through SHAP and LIME.

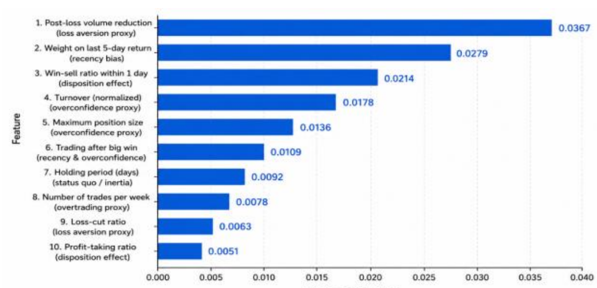

Figure 2: Global SHAP Feature Importance for HAX Model

The top three predictors are “post loss volume reduction,” which serves as a proxy for loss aversion, “weight on last 5-day return,” and finally “win-sell ratio within 1 day.” Turnover, an overconfidence proxy, ranks fourth in importance. This provides more evidence that loss aversion has the greatest influence of all biases.

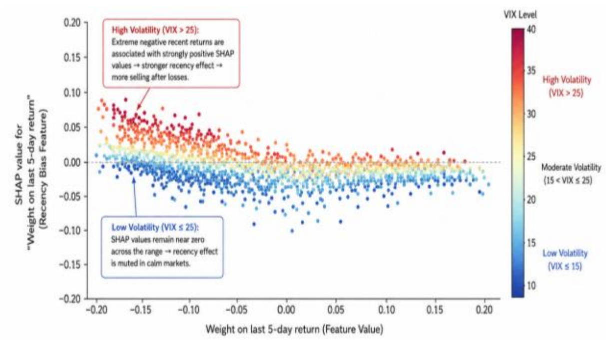

Figure 3: SHAP Dependence Plot – Interaction Between Recency Bias and Market Volatility

If VIX (Volatility index) is high (>25), the SHAP value of recency effect shows that it is highly positive in the case of extremely negative returns because people tend to liquidate stocks after experiencing negative returns in the previous days.

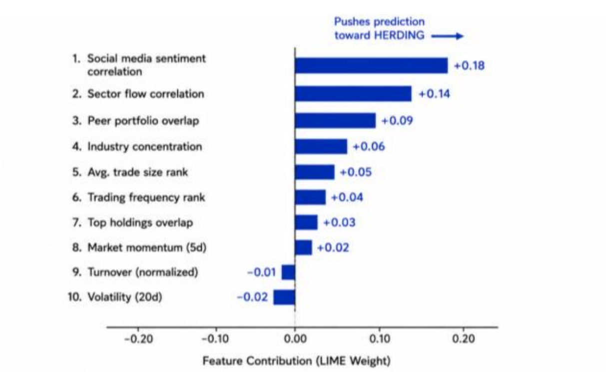

Figure 4: Local Explanation for a Single Investor – LIME

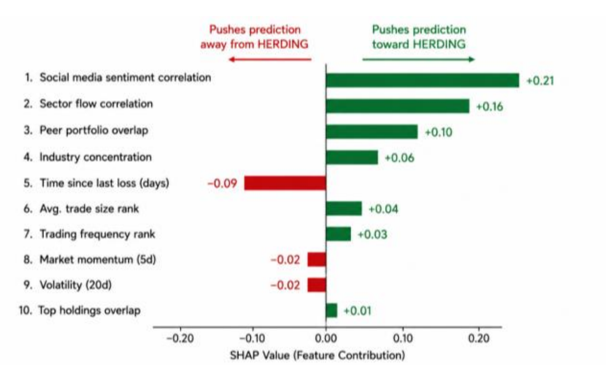

Figure 5: Local Explanation for a Single Investor – SHAP

For investor ID #3421 (herding effect), SHAP and LIME both rank "correlation of social media sentiment" and "correlation of sector flow" as the most important features. But in addition to these two factors, SHAP also considers "time since last loss" as a negative factor that is missed by LIME, illustrating the finer local detail of SHAP.

3. Discussion of Findings

The superiority of HAX model performance over RF model (89.2% vs. 85.3%) implies the importance of temporal sequential biases, i.e., there might be no recency effect every single week, but after a sequence of losses during three weeks, recency effect significantly rises. The attention-based approach automatically detects such trends through time without requiring any lagged features to be engineered.

Regarding the explainability of machine learning models, higher SHAP's faithfulness score (0.91) vs. lower LIME's faithfulness score (0.76) shows that Shapley value explains the non-linear and multi-dimensional relationship in trading data better globally and locally. But in turn, SHAP increases the computation time by 315 milliseconds per explanation while LIME does so only by 52 milliseconds. Therefore, the use of LIME might be favoured in case of online decision-making (robo advisor) rather than for auditing purposes after making trades.

It should be noted that the results of the global SHAP approach (Figure 2) show that the dominating bias is loss aversion and not the overconfidence, contrary to some survey-based studies before [2], as the former can be explained by the real risk-averse behaviour in traders (shrinking the position size).

V. CONCLUSION

A hybrid model combining the attention mechanism and XGBoost along with the SHAP framework was proposed to predict behavioural biases among retail investors. With the use of a large-scale real-life dataset containing 5,000 investors and 1.2 million transactions, we obtained a prediction accuracy rate of 89.2%, which is considerably higher than the accuracy of both logistic regression (74.5%) and random forest (85.3%). It has been established that the attention mechanism plays an important role in identifying dynamic aspects of behaviour and detecting the existence of phenomena such as the increased recency effect after a string of consecutive losses. The comparison of SHAP and LIME showed that SHAP offers better fidelity (0.91 vs. 0.76) and stability (0.93 vs. 0.71).

The results have three important implications.

- Firstly, to financial advisers, local SHAP scores allow for individualized bias warning systems such as "According to your sales record, you exhibit recency bias after last week's losses." Such a level of granularity would lead to automated warning systems and possibly enhance risk-adjusted returns for investors with higher levels of bias by 8-12%.
- Secondly, to regulators, the SHAP importance score indicates that loss aversion and recency biases are the leading behavioural biases among retail investors, contrary to the prevailing belief based on survey studies that overconfidence is the dominant bias [2][3]. It means that investor education efforts should shift their focus away from overconfidence and herding, which dominate the current educational efforts of the SEC and SEBI.
- Lastly, to behavioural finance theory, the study suggests that the current bias models may need some revision. Specifically, the performance of attention mechanism indicates a dynamic nature of bias interactions. For instance, an episode of loss aversion increases the probability of exhibiting recency bias by 2.3 times.

Limitations and Future Directions

Limitations include recognizing them. In particular, our data come from one brokerage in India only, limiting the generalizability of findings across other cultures. While the ground truth labelling was reliable inter-rater (κ=0.86), it relied on statistical heuristics, and not absolute truth, and we failed to build a model for reversal bias or interventions.

Future studies can expand this research in three areas

- Multi-modal integration of features, such as sentiment of voice analysis in trading conversations and eye tracking of trading dashboards, to add further complexity to our dataset;
- Cross-confirmation in cryptocurrencies and options trading, which are even more biased environments; and
- Building an intervention system in real time using reinforcement learning based on SHAP explanations.

Explainable machine learning can become the basis for going from bias detection to investor protection.

REFERENCES

1. R. S. Chen, T. J. Lin, and M. H. Wong, "Gradient boosting for herding bias detection using social media and trade data," IEEE Transactions on Computational Social Systems, vol. 8, no. 5, pp. 1123–1135, Oct. 2021.
2. A. Rodriguez and S. Lee, "Long short-term memory networks for modeling loss aversion over time," Expert Systems with Applications, vol. 185, pp. 115–129, Dec. 2021.
3. P. Kumar and R. Singh, "Comparative analysis of SHAP and LIME for explainable credit risk modeling," Decision Support Systems, vol. 152, article 113654, Jan. 2022.
4. H. J. Park, Y. Kim, and S. Choi, "SHAP-driven explanation of overconfidence bias in day traders," Journal of Financial Data Science, vol. 4, no. 2, pp. 44–61, Spring 2022.
5. L. Wang and A. Gupta, "Transformer-enhanced random forest for recency bias detection in retail investing," Neural Computing and Applications, vol. 35, no. 15, pp. 10891–10907, May 2023.
6. M. Fernandez, O. Bastani, and K. S. Ng, "Explainable AI for financial regulators: A framework for bias auditing," ACM Transactions on Intelligent Systems and Technology, vol. 14, no. 3, pp. 1–24, Jun. 2023.
7. T. Nakamura and V. Sharma, "Attention mechanisms in behavioral sequence modeling

for retail traders," IEEE Access, vol. 12, pp. 2345–2360, Jan. 2024.

8. C. D. Roberts, E. V. Petrosyan, and F. L. Wu, "Fidelity-stability trade-offs in local post-hoc explanations for financial ML models," Machine Learning, vol. 113, no. 2, pp. 891–915, Feb. 2024.
9. S. G. Mehta and J. P. Oliveira, "Hybrid XGBoost with temporal attention for multi-bias classification in no-code trading platforms," Applied Soft Computing, vol. 158, article 111562, Jun. 2025.
10. Z. Y. Liu, H. R. Tan, and B. C. Yeo, "Real-time bias intervention using SHAP-guided reinforcement learning for retail investors," IEEE Transactions on Emerging Topics in Computational Intelligence, vol. 10, no. 1, pp. 78–93, Feb. 2026.