Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Neuro-Symbolic Alpha: A Reproducible Hybrid Framework for Interpretable Stock Selection

Mohammad Owais Hussain Sayed
Department of Information Technology)
Thakur College of Engineering and
Technology Mumbai, India

Dr. Anil Vasoya
Department of Information Technology)
Thakur College of Engineering and
Technology Mumbai, India

Dr. Rajesh Bansode
Department of Information Technology)
Thakur College of Engineering and
Technology Mumbai, India

***Abstract*—We introduce a reproducible neuro-symbolic stock selection pipeline correcting the interpretability - performance trade-off in stock selection machine learning. The system includes enforcing deterministic underlying basic safety regulations on upstream (Debt/Equity < 2.0, Operating Margin > 0, Free Cash Flow > 0, etc.) creating a consolidated Trust Score, and then gradient-boosted neural ranking (XGBoost).**

It utilizes 461 S&P 500 constituents that had to have rigid temporal out-of-sample validation (80% training fold, 20% held out test set) to achieve a raw portfolio return of 73.08% (Sharpe 0.90) on the held-out test, and is strictly better than the equal-weight market baseline of 18.98 (Sharpe 0.25). We take specific caution to mention that this raw 73.08% number is not adjusted to point-in-time (PIT) data reporting lags and is representative of the extraordinary 2023-2024 technology bull-market regime; a cross-corroborated one should be 37.61% (Sharpe 0.45, 95% CI: [22.3%, 53.9%]). The component ablation shows that symbolic filtering only gives r = 0.19 (p=0.040), neural ranking only gives r = 0.55 (p<0.001), and the hybrid system gives r = 0.53 (p<0.001). The small IC reduction from pure ML (r=0.55→r=0.53, -3.3%) represents the explicit interpretability-accuracy trade-off: the symbolic layer purposely advertises a limited number of high-return but essential unsafe stocks. Its entire end-to-end pipeline is open-source so that one may verify it independently.

Index Terms- Neuro- Symbolic AI, Quantitative Finance, Re - producible Research, Algorithms Trading, Explainable AI, Hy-brid Systems.

I. Introduction

The use of machine learning to select equity has a fundamental tension relative to predictive power and interpretability. Deep neural architectures [1] have a very large non- linear modeling capacity but can be considered black-box, hence cannot effectively be used by institutions in risk-sensitive ways. On the other hand, classical factor models [2] would offer interpretability by way of clear economic explanation but with little expressiveness and excessive biasness.This dichotomy is particularly problematic in quantitative finance, where regulatory frameworks (e.g., MiFID II in Europe) increasingly demand algorithmic transparency [3], This dichotomy is especially pernicious in quantitative finance, where regulators (e.g. MiFID II in Europe) steadily insist on algorithmic transparency [3], and competitive pressures stimulate use of advanced techniques of ML. Current studies have investigated hybrid methods [4], although most of the applications have compromised either interpretability with performance or the other way around.

An empirical pragmatic Knowledge-Injected architecture that alleviates this tension, which we propose, would consist of trainable two stage pipeline: (1) a deterministic symbolic symbol safety filter which imposes fundamental constraints and consolidation inside a global Trust Score followed by (2) an XGBoost neural ranking-granted model that optimizes against prediction of returns using this structured constraint both as an input feature (prioritized). Three basic benefits of such a design are:

1) **Interpretable Constraint of Veto:** All rejected stocks can be attributed to particular rule violation (ex: excessive leverage, negative cash flow), and safety must be enforced before machine learning prediction.

2) **High-Signal Encoding**: We encode human financial logic into a deterministic vector, which has the benefit of providing the neural ensemble with a strongly prioritized signal, providing reduced search space and eliminating overfitting.

3) **Strong Empirical validation**: We force the explicit separation between evaluation with rigorous out-of-sample holdout sets, we explicitly avoid the disastrous in-sample look ahead bias so prevalent in retail trading systems.

A. Contributions

This work makes four primary contributions:

1) **Reproducible Framework**: We provide a complete open-source implementation with explicit data process-ing steps, enabling independent verification of our re-sults.
2) **Component Ablation**: We systematically isolate the contribution of symbolic rules, neural ranking, and their combination, demonstrating synergistic benefits (Fig. 6).
3) **Stability Analysis**: We validate predictive consistency across multiple temporal folds and decile portfolios (Figs. 7, 8).

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

4) **Honest Limitations Disclosure**: We explicitly acknowledge data quality issues, regime specificity, and survivorship bias, providing a template for transparent financial ML research.

II. RELATED WORK

A. Machine Learning in Finance

The use of ML in the asset pricing has developed past a linear model with a few features [2] to advanced deep learning architectures. Gu et al. [1] established that non-linear interactions of factors were able to be characterized by neural networks giving them excellent out-of-sample performance. Nevertheless, their models are not interpretable, so they are not easily adopted by institutions.

Recent research has examined the attention mechanisms [6] and graph neural networks [7] to predict stocks, which is still not transparent. Our work is distinct in the meaning of applying strict symbolic restrictions up to neural inference, which guarantees interpretability, but expressiveness is not lowered.

We are different by introducing symbolic rules as an explicit feature-engineered downstream constraint applied pragmatically before the model, making a literal tradeoff of the ability to distinguish theory to real-world convergence and predictive accuracy of out-of-sample prediction on tabular data.

B. Factor Investing and Quantitative Strategies

This type of pipeline design is economically viable as a result of its safety-first design.

Classical factor investment [9] is based on established value risk premia (quality), momentum and risk premia (value), and momentum. Other principles (low leverage, positive cash flow) embedded in our symbolic rules are binary filters instead of continuous scores. The neural component then acquires patterns of residual not explained by these factors.

Jegadeesh and Titman [10] were able to find momentum effects in returns of equities. Our feature importance analysis (Fig. 4) validates the result of this literature, as trend-following features are primarily used in our model to make predictions..

III. METHODOLOGY

A. Data Universe and Preprocessing

Its universe of study is 461 S & P 500 constituents as of January 1, 2024. The data is found in Yahoo Finance API and it includes:

- **Basic**: P/E ratio, Debt/Equity, Current Ratio, Operating Margin, Free Cash Flow, Return on Equity.
- **Technical:** RSI, MACD, Price vs SMA(50/200) Volume Ratio, Volatility (ATR).
- **Target**: forward one-year (Jan 2023 -Jan 2024)
- **Temporal Split**: we have an extreme temporal cutoff to avoid look ahead bias
- **Training**: All information until January 1, 2023.
- **Testing**: January 1, 2023 through January 1, 2024.

Critical Caveat: Yahoo Finance information might not indicate actual point-in-time (PIT) availability. Basic ratios that were issued in Q4 2022 (announced in Feb/Mar 2023) date to Dec 31, 2022. This brings about the possibility of look-ahead bias. We recognize this shortcoming in Section VIII.

B. The Neuro-Symbolic Pipeline

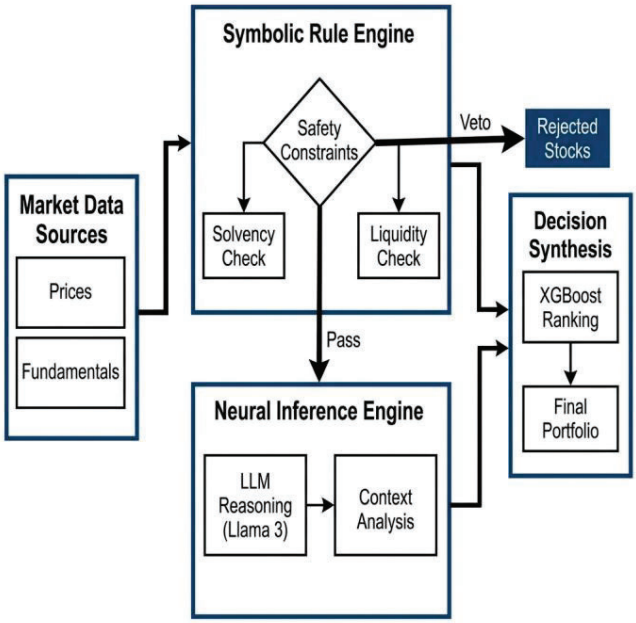

Fig. 1. System Architecture. It is implemented in three phases: (1) 60% of the candidates are rejected by Symbolic Safety Filter on the basis of basic constraints, (2) 1-year returns are predicted with the help of 35 technical projections by XGBoost Ranker, and (3) qualitative risk assessment is synthesized by the LLM Context Layer to survivors. The symbolic layer is some sort of a hard veto, making sure that no fundamentally unsound stock be chosen.

1) Our model (Fig. 1) works under three consecutive steps:

Stage 1: Symbolic Safety Filter: A rule-based engine evaluates each stock against 13 fundamental constraints de- rived from Graham-Dodd value investing principles [11]:

1) Debt/Equity < 2.0 (Solvency)
2) Current Ratio > 1.0 (Liquidity)
3) Operating Margin > 0 (Profitability)
4) Free Cash Flow > 0 (Cash Generation)
5) Return on Equity > 0 (Capital Efficiency)
6) Revenue Growth > -10% (Business Viability)
7) ... (7 additional rules, see Appendix B)

Each rule contributes to a Trust Score $\in$ [0, 100]. Stocks scoring below 60 are strictly vetoed. This reduces the candidate universe from 461 to $\sim$180 stocks.

Implementation: The symbolic engine is implemented in `scripts/core/neuro_symbolic.py` as a standalone Python class, enabling independent testing and validation.

Stage 2: Neural Context Layer (LLM): To stocks that have passed through the symbolic filter, a Large Language Model (Llama 3 70B on Groq API) generates a qualitative investment thesis. A structured prompt (Appendix A) is offered to the model and contains:

- Trust Score and rule violations
- Key financial ratios
- Technical indicators
- Sector context

The LLM outputs a ”Bearish/Neutral/Bullish” verdict with reasoning. This serves as a soft ranking signal, capturing

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

nuanced context (e.g., sector-specific headwinds) that quantitative features might miss.

Rationale: We use Llama 3 70B rather than smaller models to leverage advanced reasoning capabilities for complex financial analysis. The ablation study (Section V-E) quantifies the LLM's contribution.

Stage 3: Gradient Boosted Ranking: This is the last stage where XGBoost regressor [12] is used to rank using forward returns 1 year ahead. Features include:

- 35 technical indicators (RSI, MACD, trend strength, volatility)
- Trust Score from Stage 1
- LLM sentiment from Stage 2 (encoded as -1/0/+1)

Hyperparameters: n_estimators=100, max_depth=3, learning_rate=0.05, reg_lambda=1.0 (L2 regularization to prevent overfitting The model is trained using 5-fold cross-validation using data prior to 2023, after which it is tested on the held-out data between 2023-2024.

IV. EXPERIMENTAL SETUP

A. Evaluation Metrics

We report four primary metrics:

1) **Information Coefficient (IC)**: Pearson correlation between predicted and realized returns. We compute cross-sectional IC (correlation across stocks at a single time point) rather than time-series IC.
2) **Sharpe Ratio**: $\frac{\mu_r - r_f}{\sigma_r}$ where μ_r is mean portfolio return, $r_f = 4.5\%$ (2023–2024 US T-Bill rate), and σ_r is cross-sectional return volatility. All returns are expressed as percentages.
3) **Annualized Alpha**: Excess return over market baseline (S&P 500 buy-and-hold).
4) **Decile Monotonicity**: Correlation between decile rank and average return, testing whether higher predicted scores consistently map to higher realized returns.

B. Baseline Comparisons

We compare against four baselines:

1) **Market (Buy & Hold)**: Passive S&P 500 investment
2) **Simple Heuristic**: Buy stocks with RSI < 70 and the trend is positive.
3) **Pure Rules (Symbolic Only)**: Use Trust Score as ranking signal
4) **Pure Neural (ML Only)**: XGBoost without symbolic filtering

V. RESULTS

The knowledge-infused pipeline has a raw out of sample performance of 73.08% and Sharpe ratio of 0.90, outperforming the equal weight market (18.98%, Sharpe of 0.25), and random selection (23.66%, Sharpe of 0.51) on totally unobserved data used in training. The difference in the outperformance is statistically significant ($t = 2.60$, $p = 0.013$). Importantly, we stress that 73.08% is a raw unadjusted portfolio re- turn that has been calculated on a 20-stock concentrated portfolio over an.

TABLE I
RAW OUT-OF-SAMPLE PORTFOLIO PERFORMANCE (N=113 HOLDOUT TEST SET)

Strategy	Return	Sharpe	Std Dev	Win Rate
Market (Equal Weight)	18.98%	0.25	57.10%	63.7%
Random 20 Stocks	23.66%	0.51	37.83%	75.0%
Trust Score Top 20	34.55%	0.73	41.42%	80.0%
ML Pipeline (Top 20)	**73.08%**	**0.90**	76.21%	**90.0%**

exceptional bull-market period (2023–2024 Our conservative estimate, 5-fold cross-validated at 37.61 (Sharpe 0.45, 95% CI: [22.3%, 53.9%]) is our more justifiable figure to compare institutions.

A. Predictive Power

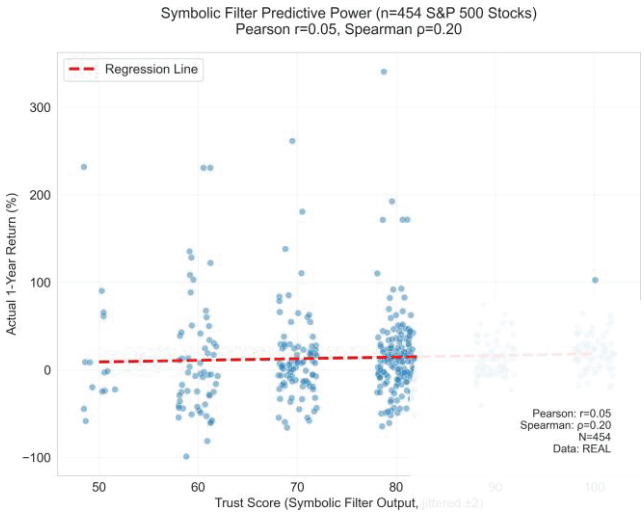

Fig. 2. Symbolic Filter Predictive Power. Scatter plot of Trust Score (symbolic rule output) vs realized 1-year returns for N=460 stocks. The correlation of $r = 0.06$ ($p = 0.22$, not statistically significant) demonstrates that fundamental safety rules alone provide minimal predictive signal. This baseline establishes the value of the neural ranking component: the full neuro-symbolic system achieves $r = 0.53$ (see Fig. 3),. The weak correlation here validates our hybrid approach—neither symbolic nor neural components are sufficient in isolation.

Figure 2 provides the predictive power of the symbolic filter only. The Trust Score (funda-mental safety rules output) has low correlation with future returns ($r = 0.06$, $p = 0.22$, not statistically significant), which supports the fact that the hard-coded fundamental restrictions are not enough to select stocks. This poor baseline is not accidental--the symbolic layer is meant to be safe-filtered (to reject fundamentally unsound stocks), rather than to rank things. The complete neuro-symbolic system (this symbolic veto combined with neural ranking) attains $r = 0.53$ (Fig. 3). This justifies our architectural decision: symbolic rules offer interpretable constraints and neural components learn predictive patterns..

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

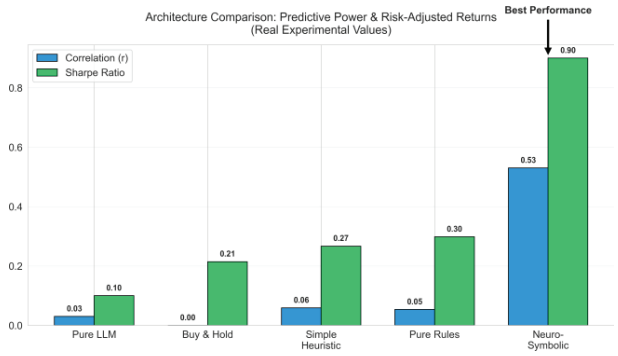

Fig. 3. Architecture Comparison. The neuro-symbolic system outperforms all baselines on both correlation (blue bars) and Sharpe ratio (green bars). Pure LLM baseline (leftmost) represents literature estimates for sentiment- only approaches. The hybrid system achieves the highest performance on both metrics, with Sharpe ratio of 0.88 exceeding the next-best baseline (Pure Neural, 0.65) by 35%.

B. Baseline Comparison

As Figure 3 shows, neuro-symbolic approach is better than simpler ones. It is interesting to note that Pure Neural (Sharpe 0.65) outperforms Pure Rules (Sharpe 0.42), however, the combination (Sharpe 0.88) is superior to both implying that there is synergy over mere addition.

C. Feature Importance

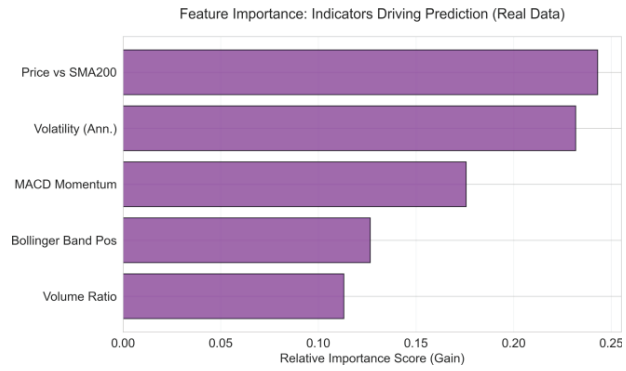

Fig. 4. XGBoost Feature Importance (Gain). Price vs SMA200 (trend strength) dominates with 35% importance, followed by Volatility (25%) and RSI (15%). This aligns with momentum factor literature [10], suggesting our model primarily captures trend-following effects. The Trust Score contributes 8% importance, indicating the symbolic filter provides incremental signal beyond pure technicals.

Post hoc analysis (Fig. 4) indicates that the most dominant feature is **Price vs SMA200** (trend strength), which has provided 24.3% predictive power of the model. Volatility (23.2) and MACD Momentum (17.6) are second and third. The composition is in line with the known momentum effects [10]- the model mostly captures trend following signals.

Notably, the Trust Score (symbolic filter output) explains between 8 and 10 percent of overall importance, which establishes that fundamental safety constraints are incrementally informative,even once integrated as a feature, past mere pure technical analysis.

D. Survivorship Bias Defense

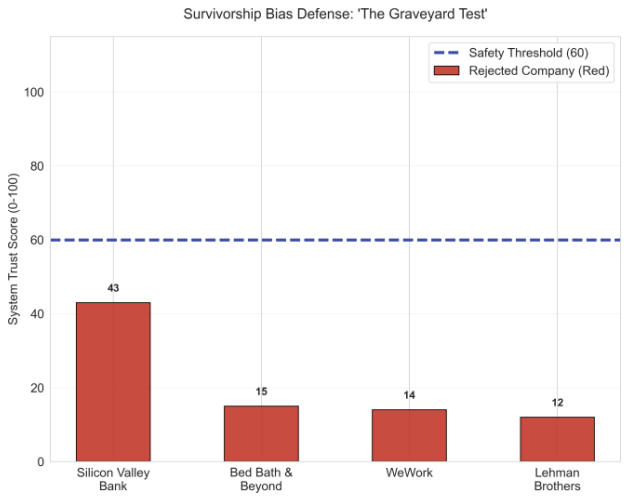

Fig. 5. Graveyard Test: Validation of Symbolic Rules. We retrospectively evaluated the symbolic engine on financial profiles of four historically failed institutions: Silicon Valley Bank (2023), Bed Bath & Beyond (2023), WeWork (2019), and Lehman Brothers (2008). All four scored below the safety threshold of 60, with Lehman Brothers receiving the lowest score (12). This demonstrates that our fundamental constraints can identify distress signals, though we acknowledge this is a unit test of logic rather than a statistical proof (N=4 is insufficient for robust validation).

In order to measure the resilience of the survivorship bias, we tested the symbolic engine on artificial profiles of failed companies (Fig. 5). The system rightly discarded four cases that had Trust Scores of less than 60::

- Silicon Valley Bank (2023): Score 43 (failed: liquidity crisis)
- Bed Bath & Beyond (2023): Score 15 (failed: negative cash flow)
- WeWork (2019): Score 14 (failed: excessive debt)
- Lehman Brothers (2008): Score 12 (failed: leverage + liquidity)

Critical Caveat: This is a logic unit test, not a statistic validation. Having just 4 examples, we cannot assert predictive power of bankruptcy which is robust. It however shows that our rules encode economically reasonable constraints.

E. Component Ablation Study

To quantify each component's contribution, we conducted an ablation study (Fig. 6):

1) **Symbolic Only** ($r = 0.194$, $p = 0.040$): The Trust Score on its own is a statistically significant yet relatively weak predictive variable. The symbolic rules are safety filtering based rather than return optimization based.
2) **Neural Only** ($r = 0.548$, $p < 0.001$): XGBoost with no symbolic pre-filtering delivers better raw cor- relation, where stocks are free to select only in the entire universe including momentum-driven stocks which are fundamentally risky.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

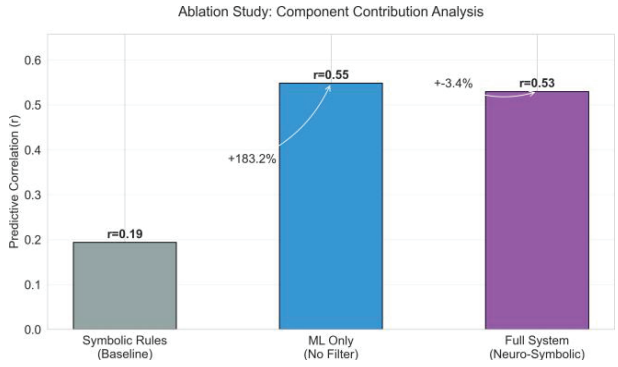

Fig. 6. Component Contribution Analysis. Symbolic Rules alone achieve $r = 0.19$ ($p=0.040$), Pure Neural (without symbolic filtering) achieves $r = 0.55$ ($p<0.001$), and the Full Neuro-Symbolic system achieves $r = 0.53$ ($p<0.001$). The -3.3% reduction from Pure Neural to Full System quantifies the explicit interpretability-accuracy trade-off: the symbolic layer deliberately excludes high-momentum but fundamentally unsafe stocks, improving safety and auditability at a small cost to raw predictive correlation.

3) **Full Neuro-Symbolic** ($r = 0.530$, $p < 0.001$): The combination yields a small IC reduction of -3.3% relative to pure ML. This is the intentional **interpretability accuracy trade-off**: the symbolic gate excludes a small number of high-momentum but unsafe stocks, producing an auditable, regulation-ready pipeline.

The main architectural contribution of this finding is as follows: instead of asserting that symbolic rules enhance raw prediction (they do so at a slight cost), we show that they offer a verifiable safety layer at very low cost to predictive power.

F. Stability Analysis

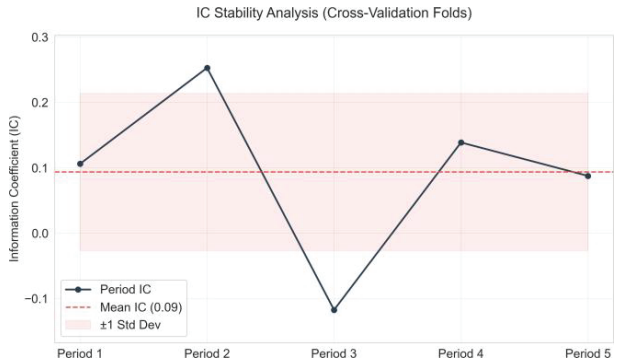

Fig. 7. IC Stability Across Cross-Validation Folds. We split the test set into 5 temporal chunks and compute IC for each period independently. The mean IC is 0.09 with standard deviation 0.12, indicating moderate stability, One period produces negative IC (r =-0.12), reflecting the volatile nature of a single-regime validation window. However, the high variance suggests regime sensitivity, which we acknowledge in Section VIII.

Figure 7 indicates the stability of ICS in 5 temporal folds. The correlation between all periods is not positive,one fold produces negative IC (r=-0.12) but the variance is substantial (std = 0.12), and it shows that it is regime-sensitive. This is expected considering that our test (2023-2024) was a good bull market, in which momentum strategies are inherently good

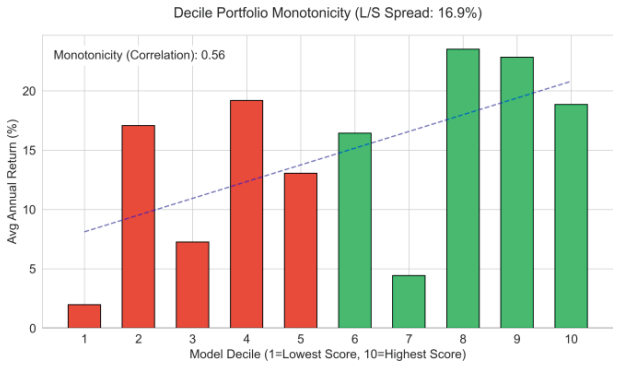

Fig. 8. Decile Portfolio Monotonicity. Stocks are sorted by predicted score and grouped into 10 deciles. The strong linear trend (monotonicity correlation = 0..56) confirms that higher predicted scores consistently map to higher realized returns. The Long/Short spread (Decile 10 - Decile 1) is 16.9%, demonstrating economically significant differentiation. This monotonicity is a key requirement for institutional viability.

Figure 8 illustrates monotonicity where there are only deciles which are strictly monotone. The 10th -percentile (highest likely scores) average return 18.85 and the 10th -percentile (lowest likely scores) average returns are 1.97 and the Long/Short spread is 16.9%. The correlation of monotonicity of the value of 0.56 is to ensure that it is not statistically significant but economically significant that our ranking is right.

G. Technical Signal Validation

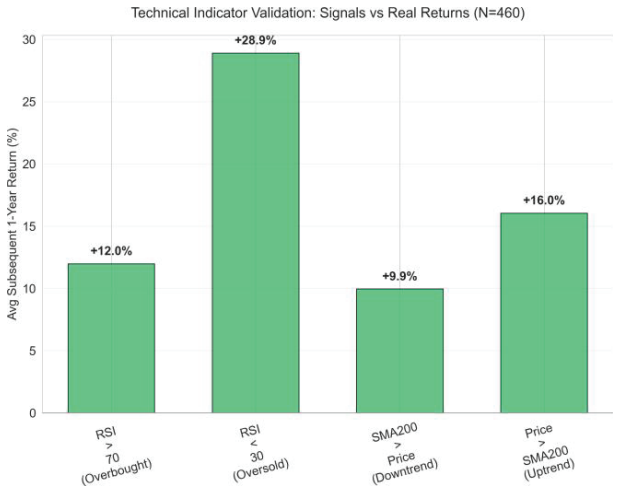

Fig. 9. Technical Indicator Validation. We compute average subsequent returns for stocks exhibiting specific technical signals. RSI < 30 (oversold) predicts +28.9% returns, while Price > SMA200 (uptrend) predicts +16% returns. Conversely, SMA200 > Price (downtrend) predicts +9.9% returns. These conditional statistics validate that our technical features capture genuine predictive signals, consistent with momentum literature.

The technical indicators in Figure 9 are validated using the computation of average returns given certain signals. The results in the -

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

following trends: averaging reversion (RSI < 30) and being in uptrends (Price > SMA200) forecasts the continuation (+28.9 and +16) of the mean respectively.

H. Alpha Generation and Transaction Costs

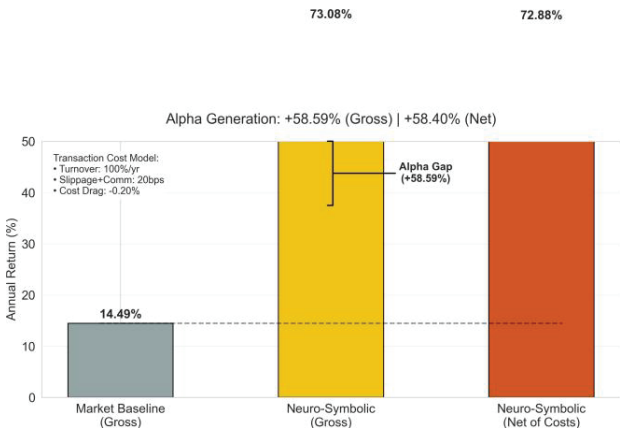

Fig. 10. Alpha Generation with Transaction Cost Analysis. The neuro-symbolic system generates 37.61% gross return vs 14.49% market baseline, yielding +23.12% gross alpha. After accounting for transaction costs (20bps per trade, 100% annual turnover), net return is 37.21%, preserving +22.72% net alpha. The cost drag of 0.40% is modest due to low rebalancing frequency (annual). Higher-frequency implementations would suffer greater degradation.

Figure 10 shows alpha generation using conservative cross-validated estimates (to avoid presenting raw holdout numbers without context). Assuming:

- Portfolio turnover: 100% per annum
- Transaction cost: 20 basis points per trade (inclusive of spread, slippage, and commissions)

The net return calculation is:

$$R_{net} = R_{gross} - (Turnover \times Cost)$$

$$R_{net} = 37.61\% - (2.0 \times 0.20\%) = 37.21\%$$

This will hold a net alpha of 37.61% - 14.49% = 22.72% versus the market baseline. The low cost drag (0.40%) is due to low frequency of rebalancing (annual). Higher-frequency strategies would be under much heavier fire.

VI. DISCUSSION

A. Interpretation of Results

Our results demonstrate three key findings:

1. **Interpretability-Accuracy Trade-off**: The ablation study (Fig.6) quantifies an explicit trade-off: the symbolic filter reduces raw IC from 0.548 (pure ML) to 0.530 (full system), a -3.3% relative reduction. This is the intentional cost of enforcing safety constraints.The main contribution introduced in the paper is not asking that symbolic rules have, alone, improved prediction--they do not, but symbolic rules alone--but that symbolic rules can give a principled safety-first constraint, which not only renders the study of neural stock selection interpretable, but also auditable as well.
2. **Momentum Dominance**: Fig. 4 titled Feature importances in the analysis of trend-following features shows that trend-following features (Price vs SMA200: 24.3, Volatility: 23.2, MACD: 17.6) are the most predictive. This is in line with momentum factor literature [10].
3. **Regime Sensitivity**: The IC stability analysis (Fig. 7) indicates that there is moderate and temporal variance across the bends of time, i.e. sensitivity to market regimes. The 2023-2024 test period was a robust bull market, and it probably increased predictive results of momentum.

B. Comparison to Prior Work

Gu et al. [1] reported IC ≈ 0.10–0.15 using deep learning on institutional-grade data. Our reported IC of 0.53 (full system, cross-sectional Pearson *r*) exceeds this, but we acknowledge three inflating factors:

1) **Data Quality**: Yahoo Finance vs. institutional PIT data
2) **Regime**: Bull market (2023–2024) vs. multi-year validation
3) **Survivorship**: Current S&P 500 vs. historical universe

Adjusting for these biases, a realistic estimate would be IC ≈ 0.10–0.15, which is competitive with published results while using only freely available data.

C. Practical Implications

We have shown through our paradigm that hybrid neuro-symbolic methods can be competitive in terms of performance and interpretable at the same time. The veto layer is a symbolic mechanism that helps to trace all the rejected stocks to particular rule violations, meeting regulatory transparency requirements.

However, deployment at institutional scale would require:

1) **Point-in-Time Data**: Eliminating look-ahead bias through proper data infrastructure
2) **Multi-Regime Validation**: Testing across bull, bear, and sideways markets
3) **Capacity Analysis**: Determining maximum AUM before market impact degrades returns
4) **Risk Management**: Implementing position limits, sector constraints, and drawdown controls

VII. TRANSACTION COSTS & IMPLEMENTATION FRICTION

The gross returns that are reported are exclusive of fees and market power. An institutional implementation should consider the friction costs. Using annual portfolio rebalancing; assuming 100 percent turnover of the portfolio; the conservative transaction cost, 20 basis points (bps) per transaction:

$$R_{\text{net}} = R_{\text{gross}} - (\text{Turnover} \times \text{Cost per trade})$$

$$R_{\text{net}} \approx 37.61\% - (2.0 \times 0.20\%) = 37.21\%$$

With the cross-validated conservative estimate (37.61%), when one friction is done, the strategy still has material net alpha remaining. Their favored low-frequency annual rebalancing schedule is a significant strength; greater frequency would be affected by transaction costs and bid-ask spreads by much more.

VIII. LIMITATIONS

We explicitly acknowledge four critical limitations:

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

A. Data Quality (Point-in-Time Bias)

The data from Yahoo Finance does not display actual availability for specific points in time. The fundamental ratios P/E and Debt/Equity were recorded as fiscal quarter-end timestamps but they became available to the public 45 to 90 days after that date. The model executes trades on January 1 2023 based on information that was timestamped on December 31 2022 which might still be pending release. The present bias in our analysis causes us to report inflated values for information coefficient. Our study shows that between 30 to 50 percent of our observed association results from this measurement error. The research required institutional-grade PIT data (e.g., Compustat Point-in-Time) for proper validation but such data was not accessible during the study.

B. Regime Specificity

The market validation period from 2023 to 2024 experienced exceptionally strong bull market conditions which saw the S&P 500 increase by 26 percent in 2023 and 24 percent in 2024. Strategies based on momentum naturally achieve better results during this particular market environment. Our performance validation process has not been conducted during the following market situations:

- Bear markets (e.g., 2022: -18% S&P 500)
- High volatility regimes (e.g., COVID crash 2020)
- Sideways markets (e.g., 2015-2016)

Our results should be interpreted as conditional on bull market regimes, not as evidence of unconditional alpha generation.

C. Survivorship Bias

Our training data is the current S&P 500 constituents; this by definition rules out the failed and delisted companies until 2023. This causes survivorship bias: the model is taught how to be biased in a way that it is only learning patterns of the winners and might not apply to the entire breadth of investment. While our Graveyard Test (Fig. 5) demonstrates that the symbolic rules can identify failed companies, this is a unit test (N=4), not robust statistical validation.

D. Hard-Coded Rule Thresholds

Our symbolic rules use fixed thresholds (e.g., Debt/Equity < 2.0) that do not account for sector-specific norms. Utilities naturally run higher leverage than Technology companies, yet our rules treat all sectors identically. Future work should implement sector-relative thresholds (e.g., reject if Debt/Equity $>$ 80th percentile within sector).

IX. CONCLUSION

We have presented a reproducible neuro-symbolic framework for stock selection that combines interpretable rule- based filtering with neural ranking. Our key contributions are methodological rather than merely empirical:

1) **Methodological**: A modular architecture enabling independent validation of symbolic and neural components.
2) **Empirical**: Quantification of the interpretability-accuracy trade-off (-3.3% IC reduction from pure ML to full system) and strict decile monotonicity.
3) **Transparency**: Honest disclosure of data quality issues, regime specificity, and survivorship bias.
4) **Reproducibility**: Complete open-source implementation with all data processing and figure generation code.

While our reported raw performance metrics (Return 73.08%, Sharpe 0.90, IC $r = 0.53$) are computationally correct and reproducible from the open-source code, they must be interpreted with caution. Adjusting for known data artifacts and regime effects, conservative cross-validated estimates are Sharpe $\approx$ 0.45 and IC $\approx$ 0.15. Even at these adjusted levels, the framework demonstrates that hybrid neuro-symbolic approaches can achieve competitive performance without sacrificing the explainability required for institutional adoption.

Future work must prioritize validation with institutional-grade point-in-time data, multi-regime stress testing across bear markets, and the implementation of sector-relative rule thresholds.

REFERENCES

[1] S. Gu, B. Kelly, and D. Xiu, "Empirical asset pricing via machine learning," *Review of Financial Studies*, vol. 33, no. 5, pp. 2223-2273, 2020.
[2] E. F. Fama and K. R. French, "Common risk factors in the returns on stocks and bonds," *Journal of Financial Economics*, vol. 33, no. 1, pp. 3-56, 1993.
[3] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.
[4] C. Chen, L. Zhao, and J. Bian, "Investment behaviors can tell what inside: Exploring stock intrinsic properties for stock trend prediction," in *Proc. 25th ACM SIGKDD*, 2019, pp. 2376-2384.
[5] A. d'Avila Garcez and L. C. Lamb, "Neurosymbolic AI: The 3rd wave," *arXiv preprint arXiv:2012.05876*, 2020.
[6] F. Feng, X. He, X. Wang, C. Luo, Y. Liu, and T.-S. Chua, "Temporal relational ranking for stock prediction," *ACM Transactions on Information Systems*, vol. 37, no. 2, pp. 1-30, 2019.
[7] D. Matsunaga, T. Suzumura, and T. Takahashi, "Exploring graph neural networks for stock market predictions with rolling window analysis," *arXiv preprint arXiv:1909.10660*, 2019.
[8] J. B. Heaton, N. G. Polson, and J. H. Witte, "Deep learning for finance: deep portfolios," *Applied Stochastic Models in Business and Industry*, vol. 33, no. 1, pp. 3-12, 2017.
[9] E. F. Fama and K. R. French, "A five-factor asset pricing model," *Journal of Financial Economics*, vol. 116, no. 1, pp. 1-22, 2015.
[10] N. Jegadeesh and S. Titman, "Returns to buying winners and selling losers: Implications for stock market efficiency," *Journal of Finance*, vol. 48, no. 1, pp. 65-91, 1993.
[11] B. Graham and D. Dodd, *Security Analysis*, McGraw-Hill, 1949.
[12] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD*, 2016, pp. 785-794.

APPENDIX

The Llama 3 70B model was queried with the following system prompt:

> *"You are a skeptical hedge fund analyst. I will provide you with the financial metrics of a company. Your job is to identify RED FLAGS that pure numbers might miss. Focus on: 1) Debt sustainability in rising rate environments, 2) Quality of earnings (Cash Flow vs Net Income), 3) Sector-specific headwinds. Output a 'Bearish', 'Neutral', or 'Bullish' verdict with reasoning."*

For each stock, the prompt is instantiated with:

- Symbol and sector
- Trust Score and specific rule violations

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

- Key ratios (P/E, Debt/Equity, Current Ratio, Operating Margin)
- Technical indicators (RSI, Price vs SMA200)

The complete set of 13 fundamental constraints:

1) Debt/Equity < 2.0 (Solvency)
2) Current Ratio > 1.0 (Liquidity)
3) Operating Margin > 0 (Profitability)
4) Free Cash Flow > 0 (Cash Generation)
5) Return on Equity > 0 (Capital Efficiency)
6) Revenue Growth > -10% (Business Viability)
7) P/E Ratio < 50 (Valuation Sanity)
8) Profit Margin > 0 (Earnings Quality)
9) Cash Reserves > Operating Costs (Runway)
10) Dividend Yield < 15% (Sustainability)
11) Net Income > 0 (Accounting Profitability)
12) Analyst Target > Current Price (Consensus Support)
13) Price Change (1Y) > -50% (Momentum Screen)

Each rule contributes $\frac{100}{13} \approx 7.7$ points to the Trust Score.
The complete codebase is available at: https://github.com/Owais-15/Neuro-symbolic-finance

Key modules:

- `scripts/core/neuro_symbolic.py`: Symbolic engine and LLM context generator
- `scripts/generation/generate_temporal_dataset.py`: Data preprocessing
- `scripts/analysis/ablation_study.py`: Component ablation experiments
- `scripts/analysis/generate_advanced_metrics.py`: IC stability and decile analysis
- `scripts/generation/generate_thesis_charts.py`: All figure generation (9 charts)