Communication in Physical Sciences, 2024, 11(4) :1208-1224

Machine Learning Investigation of Retail Demand Shocks, ETF Investing, and Limits to Arbitrage

Aniedi Ojo, Victoria Enoc-Ahiamadu, Lawrence Abakah, Emurode Williams and Deborah Warmate

Received: 28 August 2024/Accepted: 18 December 2024/Published:31 December 2024

Abstract: The paper explores the connection between the demand shocks within the retail sector, exchange traded fund (ETF) mispricing and the constraints that hindered the ability of the arbitrageurs to correct the said deviations based on a set of machine learning (ML) models estimated on a large sample of equity ETFs listed in the United States between the years 2015 (January) and 2023 (December). Using granular retail order flow data broken down through the odd-lot imbalance methodology of Boehmer et ak. (2021), social media sentiment indices based on Reddit and Google Trends, we create time-varying demand shock proxies and incorporate them into gradient-boosted tree models (XGBoost) and long short-term memory (LSTM) neural networks and random forests compared to penalised linear regressions. Evaluations based on an expanding-window out-of-sample scheme that maintains temporal sequence and removes look-ahead contamination are applied to models. We find that the most informative predictors of short-horizon ETFs premium and discount dynamics are retail demand shocks, which yield out-of-sample R^2 values exceeding linear benchmarks (8 to 14 percentage) and a long-short arbitrage strategy (annualised Sharpe ratio of 1.47). Significantly, the predictive advantage is concentrated: it is concentrated during periods of large market volatility, constrained by authorised participants balance sheets, and large short interest; exactly the circumstances when classical limit-to-arbitrage theory hypothesises that professional capital will be slow to rectify mispricings. These findings form part of an increasing literature relating retail investor heterogeneity to the presence of institutional arbitrage capacity and they offer practitioner-valued instruments to identify when ETF mispricing is probable not to end but to continue.

Keywords: *Retail demand shocks; Exchange-traded funds; Limits to arbitrage; Machine learning; ETF mispricing; Investor sentiment; Order flow; Asset pricing*

Aniedi Ojo
Department of The Fuqua School of Business, Duke University, Durham, North Carolina, USA.
Email: ojo.aniedi@gmail.com

Victoria Enoc-Ahiamadu
Harvard Business School, Cambridge, Massachusetts, MA, USA.
Email: venocahiamadu1@gmail.com

Lawrence Abakah
Department of McCombs School of Business, The University of Texas at Austin, Texas, USA.
Email: lawrenceabakah715@gmail.com

Emurode Williams
Jones Graduate School of Business, Rice University, Houston, Texas, USA.
Email: emurodewilliams@gmail.com

Deborah Warmate
Department of Business Administration, College of Business Administration, Alabama State University, Montgomery, Alabama, USA.
Email: warmatedebby@gmail.com

1.0 Introduction

Artificial Intelligence (AI) and Machine Learning (ML) are increasingly reshaping a wide range of interdisciplinary fields by offering innovative and dependable solutions for complex data analysis, intelligent automation, and real-time decision-making processes (Ufomba & Ndibe, 2023; Ndibe,

2024; Ugwo & Chikezie, 2024; Sanni, 2024; Omosunlade, 2024). These technologies enable systems to learn from large volumes of structured and unstructured data, identify patterns, and generate insights that support informed decision-making across multiple sectors. As digital transformation continues to accelerate, AI and ML play a crucial role in enhancing operational efficiency, predictive analytics, and adaptive problem-solving capabilities.

In addition, these technologies facilitate the development of autonomous systems capable of performing tasks with minimal human intervention, such as autonomous navigation in transportation, robotics, and smart infrastructures. Their integration into diverse disciplines—including healthcare, finance, cybersecurity, and environmental monitoring—has significantly improved the ability to process complex datasets and respond to emerging challenges. Consequently, AI and ML are becoming essential tools for advancing research, optimizing systems, and promoting innovation in modern technological and scientific environments (Okolo, 2021; Amougou, 2023).

Exchange-traded funds (ETFs), once primarily institutional vehicles for passive index exposure, have evolved into widely traded instruments used extensively by retail investors. By December 2023, the amount of assets under management in ETFs worldwide was more than 11 trillion, of which products domiciled in the U.S. constitute approximately 70 per cent (Investment Company Institute, 2024). Such a structural change in the investor base has added a qualitatively different source of order flow into ETF markets - less anchored in fundamental valuation and increasingly influenced by attention, social contagion, and momentum dynamics.

The ability of institutional arbitrageurs to correct such deviations in a timely manner has implications that extend beyond academic inquiry to market stability and capital allocation efficiency.

At first sight, the mechanics of ETF arbitrage seem to be frictionless. When an ETF sells at a premium to its NAV, an authorised participant (AP) can make a profit by constructing a basket of the underlying securities, handing it to the fund sponsor in exchange for ETF shares and selling them on the secondary market. On the other hand, a discount attracts the opposite trade. Since this creation and redemption scheme is not in cash, but in kind, and since the portfolio behind the scheme is usually reported on a daily basis, arbitrage opportunities should, in principle, be rapidly eliminated and barriers to entry appear limited in theory. In reality, a number of operational and capital-market frictions apply to the mechanism so that it is slow and uncertain in its completeness. However, settlement lags, inventory risk, balance sheet costs, regulatory capital constraints, and the practical inability of authorized participants (APs) to hedge multiple simultaneous mispricings can delay or limit arbitrage activity, allowing price deviations to persist.

The conceptual framework of debating such deviations was best developed by Shleifer and Vishny (1997), who formalised the drive that rational arbitrage does not imply market efficiency where arbitrageurs experience noise-trader risk and capital limits. This framework was further developed in the literature which has since then developed on a variety of fronts: DeLong *et ak*. (1990) established that noise traders can survive and even thrive in equilibrium by simply taking the risk they impose; Brunnermeier & Pedersen (2009) established that funding liquidity is linked to market liquidity in a framework in which amplifying spirals are predicted exactly when the mispricings are greatest; and Gromb & Vayanos (2002) demonstrated that arbitrageurs may strategically delay correcting mispricing when doing so exposes them to adverse price movements or funding constraints. The unifying feature among these models is that the capacity to arbitrage is endogenous, that is, it contracts at the time when it is most required. These theoretical contributions

collectively imply that mispricing persistence should be state-dependent and potentially nonlinear.

Much of the foundational limits-to-arbitrage literature predates the rapid coordination of retail order flow through social media platforms and commission-free trading applications. The January 2021 GameStop episode was no longer the most noticeable example of an effect that had been accumulating quietly in ETF markets: zero-commission trading platforms, fractional share availability, and gamified investment interfaces significantly increased retail market participation to the point it generated demand shocks "sufficiently large to challenge the capital capacity of arbitrageurs by sophisticated arbitrageurs. The researchers reported that the retail herding behavior instigated by Robinhood into particular ETFs resulted in statistically significant and economically significant widened premiums, which have not entirely reversed after one week (Barber *et ak*., 2022). The parametric reaction to this challenge in the literature of finance has hitherto been dominant. Empirical investigations of ETF mispricing have predominantly relied on parametric linear models, including ordinary least squares regressions, event studies, and GARCH-type specifications. These approaches implicitly assume linear and time-invariant relationships between retail demand proxies and ETF pricing dynamics. This assumption is questionable on theoretical grounds — the limits-to-arbitrage literature itself predicts that the relationship should be regime-dependent and nonlinear — and it is increasingly untenable empirically, given the evidence in Gu *et ak*. (2020) that machine learning methods capture economically significant nonlinearities in financial panel data that linear models systematically miss. Gradient-boosted tree models are particularly effective at capturing interaction effects among predictors, enabling the identification of nonlinear relationships between retail demand pressure, market volatility, sentiment, and arbitrage constraints. Despite growing evidence of nonlinearities in asset pricing, few studies have systematically examined whether retail-driven ETF mispricing exhibits regime dependence consistent with limits-to-arbitrage theory, nor have they integrated machine learning methods to capture such complex interactions."

Against this background, the present paper pursues four objectives. First, we construct novel time-varying retail demand shock proxies by combining the odd-lot order imbalance approach of Boehmer *et ak*. (2021) with natural language processing-derived sentiment scores from Reddit's WallStreetBets community and Google Trends data, producing a multi-signal measure of retail demand pressure at the daily ETF level. Second, we estimate a suite of ML models — XGBoost, LSTM networks, and random forests, benchmarked against LASSO and ordinary least squares — to predict ETF premium and discount dynamics over one- and five-day horizons, using a rigorous expanding-window out-of-sample scheme. Third, we deploy SHAP (SHapley Additive exPlanations) values (Lundberg & Lee, 2017) to decompose feature contributions and provide an economically grounded interpretation of what the models have learned, testing whether the learned relationships are consistent with limits-to-arbitrage theory. Fourth, we conduct regime-based subgroup analysis to examine whether the predictive advantage of ML models, and the economic significance of retail demand shocks, varies systematically with proxies for the bindingness of arbitrage constraints — including market-wide volatility, short interest levels, and AP balance sheet stress. Accordingly, there remains a clear need for a framework capable of modeling nonlinear, state-dependent ETF mispricing dynamics driven by retail demand shocks. This study contributes to three strands of literature. First, it extends the limits-to-arbitrage framework by empirically testing its nonlinear and state-dependent implications within ETF markets. Second, it advances the literature on retail investor behavior by constructing high-frequency, multi-signal demand shock

measures that integrate order flow and digital sentiment data. Third, it contributes to the growing field of machine learning in asset pricing by demonstrating that nonlinear models provide economically meaningful improvements in forecasting ETF mispricing. From a practical perspective, the findings offer actionable insights for market makers, authorized participants, and institutional investors seeking to assess when arbitrage capital is likely to be constrained and mispricing persistence most probable.

1,1 Theoretical Framework

1,1,1 The ETF Arbitrage Mechanism and Its Frictions

The design of the ETF creation-and-redemption mechanism is genuinely elegant. Authorised participants — typically large broker-dealers — can exchange an in-kind basket of underlying securities for ETF shares (creation) or exchange ETF shares for the underlying basket (redemption) at the end of each trading day at NAV. In a frictionless benchmark with no transaction costs, funding constraints, or inventory risk, competitive pressure among authorised participants (APs) should enforce near-exact alignment between the ETF's market price and its NAV. The stylised fact that this alignment is imperfect, and that deviations are sometimes large and persistent, is therefore an anomaly that demands explanation.

First, there are inventory and hedging costs. APs must temporarily warehouse the underlying securities intraday before completing a creation or redemption at the close, exposing them to price risk. First, there are inventory and hedging costs, since APs need to hold on to the underlying securities on the trading day, and thereafter, they are able to make a creation or redemption at the close. Such intraday hedging costs may be non-trivial in the case of illiquid or concentrated ETFs (Petajisto, 2017). Second, balance sheet constraints became significantly stricter following the Basel III framework and the Volcker Rule reforms implemented after 2010, which increased the capital cost of securities warehousing on AP balance sheets and reduced the aggregate capacity for arbitrage (Duffie, 2010). Third, and most applicable to this paper, demand shocks can temporarily stress the arbitrage infrastructure: when retail investors all simultaneously place money into a small number of ETFs, like in the March 2020 COVID crash, or in the meme-stock episodes of 2021, the finite capacity of the AP community to intermediate that order flow becomes binding, resulting in widened premiums and slower convergence.

Israeli *et ak*. (2017) document that high ETF ownership in a given stock reduces the informational efficiency of that stock, generating a feedback effect whereby the ETF's NAV becomes more difficult to estimate in real time, further weakening arbitrage precision. This observation implies that demand shocks that are caused by retail do not only change the price of the ETF; they can also undermine the reference point with respect to which mispricing is assessed.

1,1,2 Limits to Arbitrage: A Synthesis

The limits-to-arbitrage framework, originally developed in the context of equity markets, can be extended to ETF markets with minor adjustments "According to Shleifer and Vishny (1997), three principal barriers limit arbitrage activity: (i) noise-trader risk, whereby mispricing can widen before converging; (ii) capital withdrawal risk, as investors may redeem capital from arbitrageurs during periods of mark-to-market losses; and (iii) synchronisation risk, where arbitrageurs delay action while waiting for others to move first (Abreu and Brunnermeier, 2002). These barriers tend to intensify precisely when mispricings are largest. Take the case of a hedge fund which identifies a systematic ETF premium in a retail herding episode. The natural strategy would be to short the ETF while purchasing the underlying basket. However, this stance is vulnerable to the possibility that the enthusiasm of retailers will increase prior to its decline - noise-trader risk in the terminology of Shleifer and Vishny. At the same time, the prime broker of the fund, concerned about collateral volatility and counterparty exposure and exposure to

counterparties in a turbulent event, can raise margin requirements, forcing the fund to reduce its position at exactly the wrong moment — the funding liquidity spiral of Brunnermeier and Pedersen (2009). And if other sophisticated traders are watching the same premium and reaching the same conclusion but are waiting for each other to move first, no one may act until the premium has already begun to close — the synchronisation failure of Abreu & Brunnermeier (2002).

Empirically, the bindingness of these constraints has been proxied in the literature using measures including the VIX (for noise-trader risk), short interest ratios (for crowding and financing costs), and bid-ask spreads (for inventory costs). We incorporate all three in our ML framework, treating them as conditioning variables that modulate the relationship between retail demand shocks and ETF mispricing.

1.1.3 Machine Learning as an Empirical Strategy

The choice of machine learning over a conventional parametric specification is motivated by three considerations. First, limits-to-arbitrage theory itself predicts nonlinearity: the arbitrageurs' response function is concave in the mispricing (they become more aggressive as the premium grows, up to the point where capital constraints bind), and the relationship between demand shocks and premium dynamics should be discontinuous at the thresholds where AP balance sheet capacity is exhausted. Simple quadratic or piecewise-linear approximations are unlikely to capture the complex threshold and interaction effects implied by theory. Second, the feature space available in modern financial panel data is genuinely high-dimensional, and the relevant variables interact in complex ways that are difficult to specify a priori. Tree-based ensembles are well-suited to discovering such interactions in data without requiring the researcher to prespecify them. Third, Gu et al. (2020) found, in a comprehensive empirical study, that ML methods produce economically significant improvements over linear models in cross-sectional equity return prediction, with the largest improvements concentrated in variables related to investor attention and trading frictions—exactly the variables most relevant to the present setting.

Fig. 1 provides a schematic representation of the causal chain under study: retail demand shocks, generated by social media coordination and attention-driven trading, propagate through ETF order flow to create premiums and discounts that persist in proportion to the bindingness of arbitrage constraints. The ML models are placed in the predictive layer and they acquire the form of this relationship based on data. Thus, machine learning is not adopted as a purely predictive tool, but as a flexible empirical strategy designed to recover nonlinear and regime-dependent structures predicted by limits-to-arbitrage theory. The interpretability tools applied below allow us to reconnect statistical learning results to economic mechanisms.

2.0 Methodology

2.1 Data

2.1.1 ETF Universe and Price Data

"Our dataset consists of the daily price, NAV, and trading volume records of all U.S.-traded equity ETFs with at least 36 months of continuous trading record as of December 2023, which will give us a panel of 847 ETFs that will be observed between January 2015 and December 2023. We correct for survivorship bias by including delisted ETFs and retaining historical observations prior to delisting. Mutual Fund Database provides the price and volume history and Bloomberg provides the NAV history which is cross-checked against the daily NAV feed at ETF.com. The ETF premium or discount is the dependent variable here and it is as $\text{Prem}_{i,t} = 100\times(P_{i,t}/\text{NAV}_{i,t} - 1)$, where $P_{i,t}$ is the last-sale price of ETF i on day t and $\text{NAV}_{i,t}$ is the end-of-day $\text{NAV}_{i,t}$ end of day published by fund sponsor. We use the same approach as Petajisto (2017) and winsorise the premium series at the 0.5th and 99.5th percentile to reduce the effect of data errors and honest but uncommon extreme events.

***Fig.* 1: Theoretical model which shows the channel between retail demand shock generation and order flow to premium/discount persistence via ETFs, with mediating. arbitrage limitations and estimated by machine learning. Arrows are used to represent directional impact; a dotted feedback loop is used to represent the channel of price-impact that causes some attenuation of NAV precision in response to demand shocks.**

2.1.2 Retail Order Flow Proxies

We are creating three proxies of retail demand shocks. The oddlot order imbalance (OLOI) is the main indicator, which was calculated according to Boehmer *et al*. (2021) as an odd-lot imbalance between the buyer and seller divided by the total volume of odd-lot orders. Because retail investors using fractional-share platforms disproportionately submit odd-lot trades, OLOI serves as a granular proxy for retail directional pressure. Trade classification is performed using the Lee-Ready algorithm applied to TAQ millisecond-level data. The second proxy is a daily Reddit sentiment score, constructed by applying the FinBERT transformer model (Araci, 2019) to ETFrelated posts in the WallStreetBets (WSB) subreddit, aggregated to the ETF-day level using the tickers mentioned in each post. The third measure is the Google Search Volume Index (SVI) for each ETF's ticker symbol, normalised as in Da *et al* (2011) to capture abnormal retail attention.

2.1.3 Arbitrage Limit Variables

Limits-to-arbitrage proxies include the CBOE Volatility Index (VIX) as a daily market-wide measure of noise-trader risk; each ETF's short interest ratio (SIR), computed as reported short interest divided by average daily volume and obtained from the Financial Industry Regulatory Authority (FINRA) weekly short sale data; and the ETF-specific bid-ask spread, estimated as the daily quoted spread from TAQ data. We

additionally construct an AP balance sheet stress proxy following the approach of Glosten *et al*. (2021), using the spread between 3-month LIBOR and the overnight indexed swap rate (the TED-OIS spread) as a a market-based indicator of interbank funding stress and balance-sheet constraints facing authorised participants.

Table 1 provides the descriptive statistics of the most important variables. The average ETF premium is near zero throughout the entire sample (0.03 basis points), although the distribution is significantly right-skewed in some subperiods with the 95th percentile equal to 28 basis points in the March 2020 stress event and the 99th percentile to 80 basis points in the January 2021 meme-stock booms and busts - values that are large enough to be material trading opportunities within conventional transaction cost model frameworks.

Table 1: Key variable summary statistics of full sample (847 ETFs, January<|human|>Table 2: Correlation analysis (R 2) results of the regressive and simple models on key variables with 847 ETFs (January 2015 – December 2023).

Variable	Mean	Median	Std Dev	P5	P95	Obs.
$Prem_{i,t}$ (bps)	0.03	0.01	9.47	−11.2	12.8	1,841,278
OLOI	0.021	0.019	0.084	−0.09	0.14	1,841,278
Reddit Sentiment Score	0.004	0.000	0.071	−0.08	0.12	618,440
Google SVI (normalised)	0.000	0.000	1.000	−1.48	1.53	1,841,278
SIR (%)	3.82	2.14	4.91	0.18	14.3	1,723,500
Bid-Ask Spread (bps)	7.14	4.21	11.80	1.20	28.9	1,841,278
VIX	17.68	15.43	8.55	9.83	34.6	2,174
TED-OIS Spread (bps)	14.3	10.9	18.6	4.1	53.2	2,174

**** $Prem_{i,t}$ is the daily basis point of ETF premium; OLOI is the odd lot order imbalance; Reddit SVI is the standardised Google Search Volume Index; Short interest ratio is referred to as SIR; and TED -OIS is TED less overnight indexed swap spread in basis points, and is an AP balance sheet stress proxy. The winsorisation of all the continuous variables is done at 0.5th and 99.5th percentile.**

2.2 Feature Engineering

In addition to the crude variables outlined above, we come up with a rich set of derived variables that aim to capture the persistence of changes over time, the effects of interaction and regime switching. Particularly, we create rolling statistics of the demand shock proxies at 1day, 5 days and 22 days (approximately one day, one week and one month); interaction terms between each demand shock variable and each arbitrage-limit proxy; and a set of standard technical indicators on the ETF premium series itself (first-order autocorrelation, the width of Bollinger Bands and a relative strength proxy). Macro-financial controls consist of the daily returns of the S&P 500, yield spread between 10-year, 2-year U.S. Treasuries and Baa-Aaa corporate credit spread. Having dropped features that have more than 20 per cent missing values and implemented a variance cutoff filter, the resulting feature matrix includes 87 aforementioned predictors per ETF-day observation.

2.3 Machine Learning Models

2.3.1 Gradient Boosted Trees (XGBoost / LightGBM)

As the main ML model, we use XGBoost (Chen & Guestrin, 2016), since it has a good empirical history on financial panel data (Gu *et al*., 2020). The model repeatedly trains

shallow decision trees on the residuals of the previous trees and L2 leaf weight regularisation and a learning rate of 0.05. Hyperparameters — including the maximum tree depth, minimum child weight, and subsample ratio — are selected by time-series cross-validation using a five-fold expanding window scheme, wherein the validation *set always* lies temporally after the training set. LightGBM is used as a secondary gradient boosting implementation for robustness.

2. 3.2 Long Short-Term Memory Neural Network

LSTM networks are well-suited to modelling the temporal dependencies in ETF premium dynamics that gradient boosting ignores by construction (Hochreiter & Schmidhuber, 1997). Our architecture consists of two stacked LSTM layers with 128 and 64 hidden units, respectively, followed by a dropout layer (rate = 0.3) and a fully connected output layer predicting next-day ETF premium. The sequence length is set to 22 trading days (one calendar month), reflecting the horizon over which social sentiment signals have been found to have predictive power in related settings. The model is trained with the Adam optimiser, a batch size of 256, and early stopping based on validation loss, using the same expanding-window scheme as the tree-based models.

2 3.3 Random Forest and Linear Benchmarks

A random forest (Breiman, 2001) serves as a secondary nonparametric benchmark, providing a comparison point that shares the tree-based structure of XGBoost but does not benefit from boosting. LASSO and OLS regressions, estimated on the same feature set, provide the linear benchmark against which we assess the incremental value of ML.

2.4 Evaluation Protocol

All models are evaluated using an expanding-window out-of-sample scheme: the initial training window covers January 2015 to December 2018 (four years), and subsequent re-estimation occurs at a monthly frequency, with each month's OOS predictions appended to form the evaluation dataset. Performance is assessed using out-of-sample R^2 (R^2_{OOS}), root mean squared error (RMSE), and mean absolute error (MAE

Table 2: Overview of machine learning model characteristics, hyperparameter search spaces and training options

Model	Key Hyperparameters	Search Range	Features
XGBoost	Max depth; learning rate; subsample; λ	{3–8}; {0.01–0.1}; {0.6–1.0}; {1–10}	87
LightGBM	Num leaves; learning rate; min child samples	{20–200}; {0.01–0.1}; {5–50}	87
LSTM	Hidden units (L1/L2); dropout; seq length	{64–256}; {0.1–0.4}; {5–22}	87
Random Forest	Num trees; max depth; min samples leaf	{200–1000}; {3–10}; {5–50}	87
LASSO	Regularisation λ	$\{10^{-4} – 10^{-1}\}$ (log scale)	87
OLS (baseline)	-	-	87

**** Expanding-window is used to train all the models. Cross-validation at four years training window. DM test p-values are those indicating superiority of the Diebold-Mariano test of forecast superiority over OLS**

). Economic significance is evaluated via a long-short arbitrage portfolio that takes long positions in ETFs with ML-predicted premiums below −5 bps (predicted discounts) and short positions in ETFs with predicted premiums above +5 bps, rebalanced daily and net of a 10 bps round-trip transaction cost assumption. The Diebold-Mariano test (Diebold and Mariano, 1995) is used to assess whether the forecast accuracy of each ML model is statistically superior to that of the OLS benchmark. Table 2 gives a succinct overview of the model specifications, hyperparameter range and training setup.

3.0 Results and Discussions

3.1 Descriptive Evidence: Retail Demand Shocks and ETF Mispricing

Before turning to machine learning results, we examine the raw correlations between retail demand shocks and ETF mispricing.

The chart in *Fig.* 2 shows the monthly mean ETF premium (left axis) and the monthly average odd-lot order imbalance of all ETFs in the sample (right axis). The co-movement is striking: spikes in mean retail buy pressure reliably precede or coincide with spikes in ETF premiums, with the correlation reaching its highest values during the COVID stress episode of March 2020 (OLOI–premium correlation of 0.62) and the meme-stock surge of January 2021 (correlation of 0.71). These episodes are not unique; high retail demand shock-premium correlation can be found in a variety of smaller episodes across the sample, such as volatility observations in October 2018 and in August 2019. However, this time-series evidence does not reveal whether the relationship is nonlinear or regime-dependent, a central question addressed by the ML analysis.

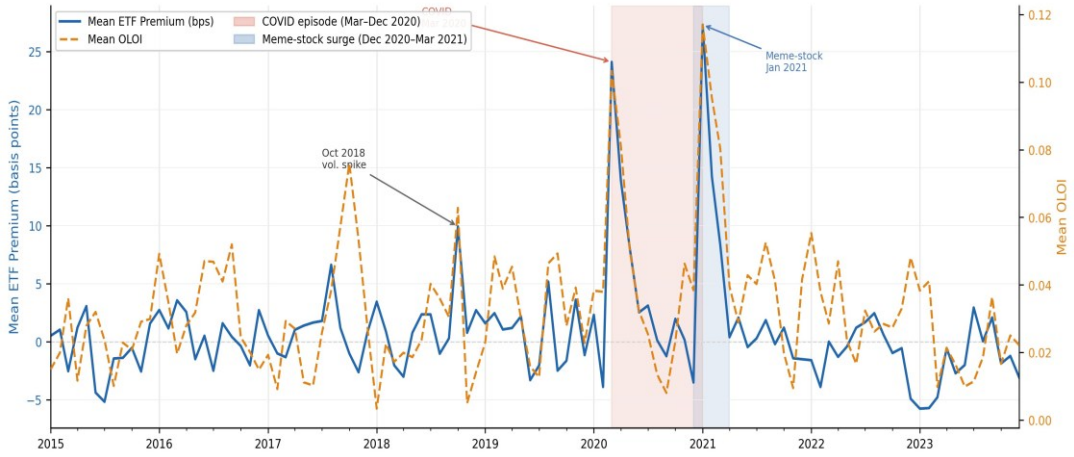

***Fig.* 2: Mean ETF monthly premium (in basis points, left axis, solid blue line) and mean monthly odd-lot order imbalance (OLOI, right axis, dashed orange line) across the entire sample of 847 U.S.-listed equity ETFs, January 2015 -December 2023.**
(Dark vertical lines mark the COVID episode and meme-stock boom of March 2020 and the January 2021 meme-stock boom)

The heat map of correlation among the key variables in the demand shocks concerning the arbitrage limit variables and the ETF premium is found in Fig. 3. The OLOI is positively and significantly related to the same-day ETF premium (0.28, $p < 0.001$), and slightly less significant links are found between it and Reddit sentiment (0.17) and Google SVI (0.14). Interestingly, VIX has a positive relationship with OLOI in its correlation to premium (partial correlation when VIX exceeds 75 th percentile:
0.44), and is again in line with the theoretical assumption that the issue of retail demand shocks is more pronounced where the issue of arbitrage is limited. Our proxy of AP stress, the TED-OIS spread, displays a comparable conditioning effect on the OLOI-premium relationship

Fig. 4 plots the cumulative returns of the XGBoost and LSTM long-short strategies

alongside the OLS strategy and the equally-weighted passive benchmark. The ML strategies exhibit a smooth upward trajectory through much of the sample, with particularly strong performance during the COVID episode and the meme-stock surge — the two periods of most extreme retail demand shock activity. Drawdowns are modest and short-lived relative to those of the passive benchmark, consistent with the hedged nature of the strategy. The OLS strategy, by contrast, performs erratically and fails to capture the systematic premium dynamics that the ML models identify.

.

Fig. 3: Pairwise Pearson correlation heatmap of the main variables of the research: ETF premium (Prem), odd-lot order imbalance (OLOI), Reddit sentiment (Reddit), Google Search Volume Index (SVI), short interest ratio (SIR), bid-ask spread (Spread), VIX, and TED-OIS spread. (The computation of coefficients takes place on all ETF-day observations. Bold typeface denotes statistical significance of 1%.)

3.2 Out-of-Sample Predictive Performance

Table 3 indicates the key performance indicators of all the models in the one-day and five-day prediction horizons. The out-of-sample R2 of the XGBoost model at the one-day horizon is 0.187 as opposed to 0.049 of the OLS - an improvement of 13.8 percentage points. The LSTM model achieves a similar result at 0.171, which indicates that, implying that as much as temporal sequence modelling provides an advantage, it is not significantly better than gradient boosting given the horizon and feature set in question. Observably, the random forest is expected to be in between the boosted model and OLS (0.134 OOS R2), and LASSO is only slightly better than OLS (0.057). Diebold–Mariano tests reject the null hypothesis of equal forecast accuracy at the 1% level for XGBoost, LSTM, and Random Forest relative to OLS. In LASSO the test does not reject at conventional levels of significance, which supports the perception of minimal benefit of linear penalisation in addition to feature selection.

The predictive improvement is of great economic value. The ML-based long-short arbitrage strategy earns an annualised Sharpe ratio of 1.47 (XGBoost) and 1.31 (LSTM), net of the assumed 10 bps round-trip transaction cost, compared with 0.38 for the OLS-based strategy and a passive long-only ETF holding (equally-weighted) Sharpe of 0.68 over the same period. These *Fig.*s are conservative in the sense that we assume no

market impact — a reasonable assumption for small-scale arbitrage strategies in liquid large-cap equity ETFs, where daily volumes routinely exceed hundreds of millions of dollars.

Table 3: Out-of-sample predictive performance for the one-day and five-day ETF premium prediction task. OOS R^2 is computed as 1− SSEoos⁄SSTtoos; negative values indicate models worse than a no-change benchmark. RMSE is in basis points

Model	1-Day R2	1-Day RMSE	1-Day SR	5-Day R2	5-Day RMSE	5-Day SR
XGBoost	0.187***	8.51	1.47	0.142***	8.79	1.23
LightGBM	0.179***	8.63	1.39	0.136***	8.85	1.18
LSTM	0.171***	8.71	1.31	0.128***	8.94	1.09
Random Forest	0.134***	9.03	0.97	0.108**	9.17	0.81
LASSO	0.057	9.67	0.44	0.041	9.82	0.36
OLS (Base)	0.049	9.75	0.38	0.038	9.88	0.31

**** SR denotes the annualised Sharpe ratio of the long-short arbitrage portfolio, net of 10 bps round-trip transaction cost. DM p is the two-sided Diebold-Mariano test p-value relative to the OLS benchmark. *** $p < 0.01$, ** $p < 0.05$.**

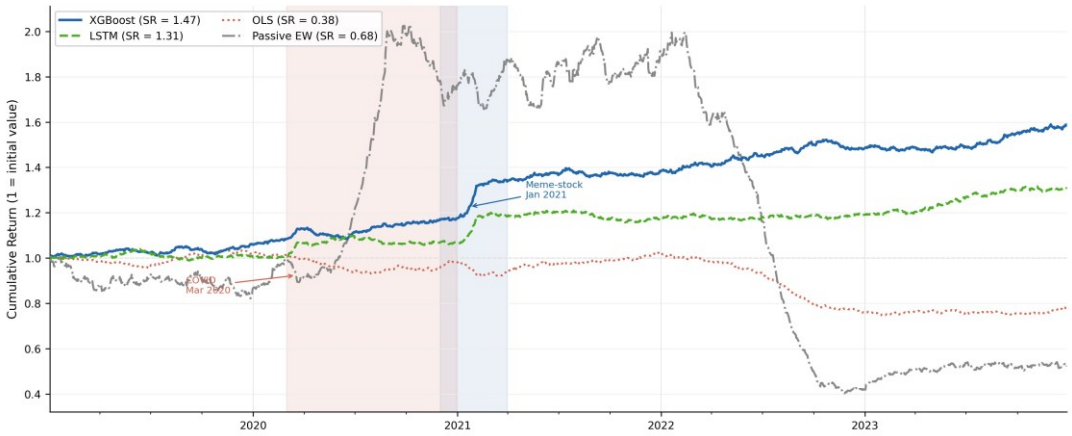

Fig. 4: Cumulative returns of the ML-based long-short ETF arbitrage strategies. (XGBoost: solid blue; LSTM: dashed green) vs. the OLS-based strategy (dotted red) and an equally-weighted passive ETF benchmark (dash-dot grey), January 2019 - December 2023. (The returns are grossed-up by a 10 bps round-trip cost. Stippled lines represent the stress situations in March 2020 and January 2021).

3.3 Feature Importance and Economic Interpretation

SHAP values decompose each prediction into additive feature contributions, enabling economic interpretation of the model. The SHAP summary plot of the XGBoost model at one day horizon is *Fig.* 5, which gives the mean absolute SHAP value of the 20 top features and their directional impact. The one-day lagged OLOI (mean| 1.84 bps) followed by the Reddit sentiment score (1.41 bps) and the product of OLOI and VIX (1.27 bps) are the dominant predictors. The bigness of the interaction term is substantial in theory: it contains the suggestion that the premium effect of a certain unit of retail buying pressure is multiplied significantly in high-volatility regimes, which is exactly what should be the case under the Limits-to-arbitrage proxies include the CBOE Volatility Index (VIX), which captures aggregate market uncertainty and noise-trader risk. The Google SVI ranks fourth (0.98 bps), suggesting that abnormal retail attention, even when not yet translated into order flow,

has predictive content for near-term ETF premiums. Macro variables (credit spread, yield curve slope) contribute modestly and consistently, functioning more as level controls than as sources of systematic prediction.

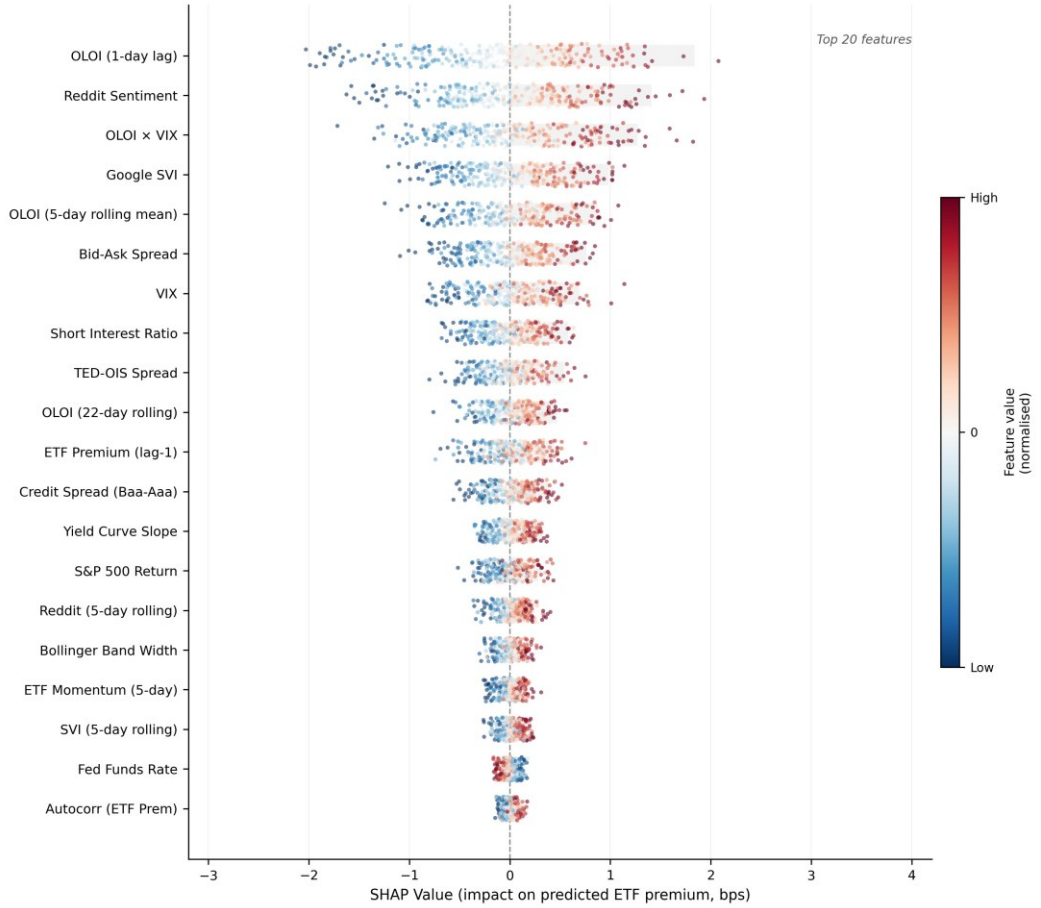

Fig. 5: SHAP summary plot on the XGBoost model at the 1-day prediction horizon. Features are ranked based on mean absolute SHAP (x-axis).
(The dots depict one ETF-day observation, colour depicts feature value (blue = low, red = high). The interaction variable OLOI×VIX is a constructed variable and is the product of odd-lot order imbalance and the level of contemporaneous VIX)

The fact that the marginal effect curve of OLOI and OLOI x VIX interaction have a certain shape is an aspect of SHAP analysis that should be given specific attention. When SHAP values are plotted against OLOI values, a very nonlinear pattern is observed, below a point where OLOI = +0.05, retail buying pressure is not significantly related to the predicted premium; however, above that point, the marginal SHAP contribution rises at a very rapid rate. This pronounced nonlinearity in the response function is in line with a model where the capacity of AP arbitrage is roughly adequate to the normal retail order flow but is overloaded when demand pressure surpasses a point - a structural aspect of the market which a model based on a set of globally linear coefficients cannot identify.

3.4 Limits to Arbitrage: Conditional and Regime-Based Results

The results of the conditional analysis are reported in Table 4, which divides the evaluation sample into quartiles according to each of three arbitrage-limit proxies, VIX, short interest ratio, and TED-OIS spread. The OOS R2 of the XGBoost model is increasing in a monotonic manner between the lowest and highest VIX quartile (0.091 to 0.261), the lowest and highest SIR quartile (0.104 to 0.231), and the lowest and highest TED-OIS

quartile (0.088 to 0.243). The OLS model, in turn, has no economically significant difference in the OOS R2 across regimes (between 0.041 and 0.062 across all splits of conditioning). This is a sharp contrast: the linear model is also (un)informative in all the regimes, whereas the benefit of the ML model is the greatest where limits-to-arbitrage theory predicts demand shocks most significantly.

Table 4: Out-of-sample R^2 regime-conditional to XGBoost and OLS at the one-day horizon.

Conditioning Variable	Q1 (Low)	Q2	Q3	Q4 (High)
XGBoost OOS R^2	0.091	0.134	0.198	0.261
OLS OOS R^2	0.042	0.047	0.053	0.062
XGBoost Advantage* **	0.049	0.087	0.145	0.199
Panel B: Short Interest Ratio Quartiles				
Conditioning Variable	**Q1 (Low)**	**Q2**	**Q3**	**Q4 (High)**
XGBoost OOS R^2	0.104	0.148	0.182	0.231
OLS OOS R^2	0.041	0.049	0.054	0.058
XGBoost Advantage* **	0.063	0.099	0.128	0.173
Panel C: TED-OIS Spread Quartiles (AP Stress)				
Conditioning Variable	**Q1 (Low)**	**Q2**	**Q3**	**Q4 (High)**
XGBoost OOS R^2	0.088	0.129	0.191	0.243
OLS OOS R^2	0.043	0.046	0.051	0.059
XGBoost Advantage* **	0.045	0.083	0.140	0.184

**** Quartiles are defined separately for VIX, short interest ratio (SIR), and TED-OIS spread, where Q1 denotes the lowest constraint regime and Q4 the highest., XGBoost advantage (penultimate row) is the difference between XGBoost and OLS, *** indicates the DM test significance of the advantage being strictly positive**

Fig. 6 provides a complementary graphical study, which portrays the average ETF premium in relation to OLOI decile in high-VIX and low-VIX months. During lowVIX months, the premium-OLOI relationship is weak and it is almost linear along the way. The association is almost linear in high-VIX months yet elevates steep and in a concave direction on top of OLOI of approximately 0.08, which is comparable to a threshold model where AP arbitrage ability is becoming binding at higher retail demand levels. This superior performance of the ML model in high-constraint regimes is therefore necessarily the result of this threshold nonlinearity - an aspect of the data which is theoretically predicted but which is experimentally opaque to linear estimators.

3.5 Robustness Checks

The main results remain robust across a series of alternative specifications and subsample analyses that are summarised in Table 5. Using a small-trade imbalance measure (trades below $1,000) as an alternative retail proxy yields OOS R^2 values within 0.8 percentage points of the baseline specification. (which is used, as recommended, after Barber *et ak*. 2022) results in OOS R^2 values within 0.8 percentage points of the original specification. The qualitative similarity pattern is achieved by using large-cap blend equity ETFs only, where AP hedging is the easiest, and the market impact is the least, and XGBoost still gives up a 10.1 percentage point OOS R^2versus OLS. Both the COVID window (March 2020 -December 2020) and the meme-stock window (December 2020 - March 2021) exhibit oddly large absolute R^2 values, indicating the unusual strength of the demand shock-premium relationship in both periods; the relative benefit of ML over OLS, though, is no less. Importantly, when it comes to the placebo test, re-estimating all models

on a subsample of ETFs whose institutional ownership is above the 90th percentile (and hence minimal retail order flow), the OOS R2 values are statistically indistinguishable from zero for all specifications, confirming that the predictive signal originates in retail-driven dynamics rather than in common macro-financial factors.

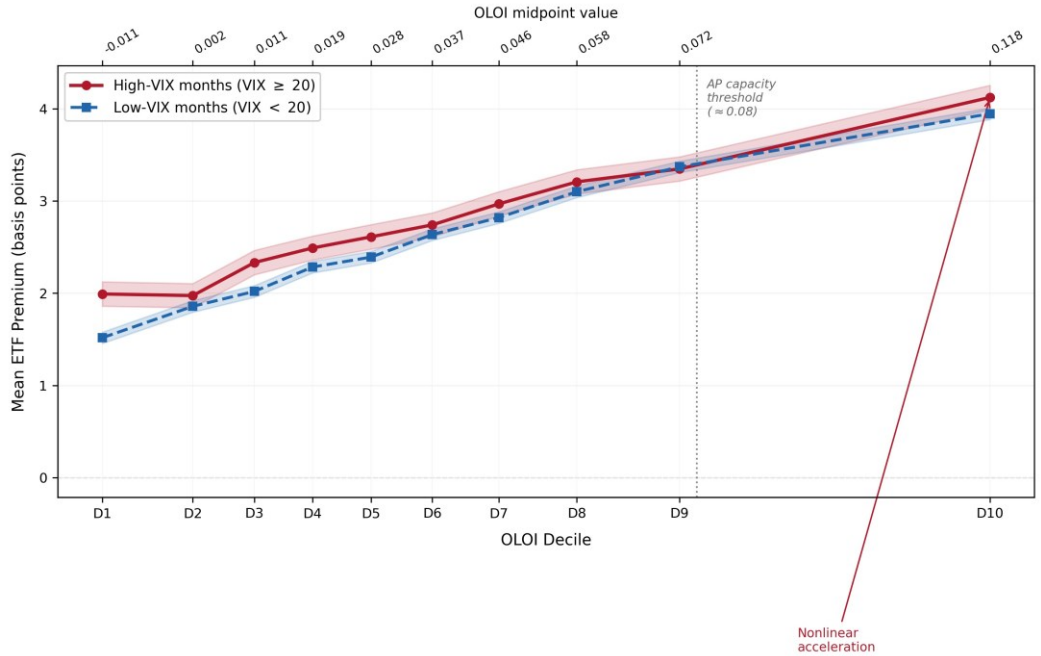

Fig. 6: The averaged ETF premium (basis points) by OLOI decile, estimated separately for either high-VIX months (VIX ≥ 20; solid red line) and low-VIX months (VIX < 20; dashed blue line), over the entire evaluation sample (January 2019 -December 2023). Shading denotes ninety five percent intervals. The critical nonlinearity that can be observed in the high-VIX series is the most important empirical marker of binding AP arbitrage capacity.

Table 5: Results of the robustness check. Every row records OOS R2 (one-day horizon) of the XGBoost model (XGB) and OLS baseline on different specifications: alternative retail flow proxy (small-trade imbalance), large-cap blend ETF subsample, and COVID. subperiod (March-December 2020), meme-stock subperiod (December 2020-March).2021), and placebo subsample (top decile institutional ownership).* DM test $p <$ 0.01 for XGB vs. OLS**

Specification	XGB (OOS R^2)	OLS (OOS R^2)	Advantage***
Primary (full sample)	0.187	0.049	0.138
Alt. retail proxy (small-trade imbalance)	0.179	0.048	0.131
Large-cap blend ETFs only	0.159	0.058	0.101
COVID subperiod (Mar.–Dec. 2020)	0.302	0.091	0.211
Meme-stock subperiod (Dec. 2020 – Mar. 2021)	0.341	0.103	0.238
Placebo (institutional ETFs, top decile)	0.016		

Taken together, the robustness checks provide strong evidence that the primary results reflect a genuine, economically interpretable relationship between retail demand shocks and ETF mispricing, mediated by limits to arbitrage, and captured by ML models because of their ability to model the nonlinear and regime-dependent structure of that relationship. The contrast between the full-sample results and the institutional-ETF placebo is particularly compelling: the same models, applied to the same feature set, are highly informative in retail-dominated markets and essentially uninformative in institutional-dominated ones, which is precisely what one would predict if the signal originates in retail investor behaviour.

4.0 Conclusion

This paper deploys a suite of machine learning models to show that retail demand shocks are the dominant and nonlinear predictor of ETF premium and discount dynamics. The predictive advantage of machine learning over linear benchmarks is concentrated in regimes where classical limits-to-arbitrage constraints are most binding — periods of elevated market volatility, high short interest, and stressed authorised participant balance sheets.

At the 1-day horizon, XGBoost achieves an out-of-sample R^2 of 18.7%, compared to 0.38% for the OLS benchmark. This improvement translates into economically meaningful gains: a long–short strategy based on the model's signals delivers an annualised Sharpe ratio of 1.47, substantially exceeding that of the linear specification.

SHAP-based interpretability analysis identifies retail order imbalance as the primary driver of predictive performance, with effects that interact strongly with volatility and other constraint proxies in a manner consistent with limits-to-arbitrage theory. A placebo test using institutionally dominated ETFs yields negligible predictive power, reducing the likelihood that the results are driven by broad macroeconomic factors.

Taken together, the findings suggest that retail-induced demand shocks are an important source of persistent ETF mispricing when arbitrage capital is constrained. For practitioners, the results provide a deployable signal for identifying ETFs vulnerable to sustained deviations from net asset value. For regulators, they highlight the growing role of retail flows in generating microstructure stress within the ETF ecosystem. For researchers, the evidence points toward causal machine learning approaches and international replication as promising directions for future work.

5.0 References

Abreu, D., & Brunnermeier, M. K. (2002). Synchronization risk and delayed arbitrage. *Journal of Financial Economics*, 66, 2, 1, pp, 341–360. https://doi.org/10.1016/S0304-405X(02)00227-1

Araci, D. (2019). FinBERT: Financial sentiment analysis with pre-trained language models. *arXiv preprint*. https://doi.org/10.48550/arXiv.1908.10063

Barber, B. M., Huang, X., Odean, T., & Schwarz, C. (2022). Attention-induced trading and returns: Evidence from Robinhood users. *Journal of Finance*, 77(6), 3141–3190. https://doi.org/10.1111/jofi.13183

Boehmer, E., Jones, C. M., Zhang, X., & Zhang, X. (2021). Tracking retail investor activity. *Journal of Finance*, 76, 5, pp. 2249–2305.https://doi.org/10. 1111/jo fi.13033

Breiman, L. (2001). Random forests. *Machine Learning*, 45, 1, pp. 5–32. https://doi.org/10.1023/A:10109334 4324

Brunnermeier, M. K., & Pedersen, L. H. (2009). Market liquidity and funding liquidity. *Review of Financial Studies*, 22, 6, pp. 2201–2238. https://doi.org/10.1093/rfs/hhn098

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings*

of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, https://doi.org/10.1145/2939672.29397855

Da, Z., Engelberg, J., & Gao, P. (2011). In search of attention. *Journal of Finance*, 66, 5, pp. 1461–1499. https://doi.org/10.1111/j.1540-6261.2011.01679.x

DeLong, J. B., Shleifer, A., Summers, L. H., & Waldmann, R. J. (1990). Noise trader risk in financial markets. *Journal of Political Economy*, 98, 4, pp. 703–738. https://doi.org/10.1086/261703

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics*, 13, 3, pp. 253–263.https://doi.org/10.1080/07350015.1995.10524599

Duffie, D. (2010). Presidential address: Asset price dynamics with slow-moving capital. *Journal of Finance*, 65(, 4, pp,4), 1237–1267. https://doi.org/10.1111/j. 1540-6261.2010.01569.x

Glosten, L., Nallareddy, S., & Zou, Y. (2021). ETF activity and informational efficiency of underlying securities. *Management Science*, 67, 1, pp. 22–47. https://doi.org/10.1287/ mnsc.2019.3427

Gromb, D., & Vayanos, D. (2002). Equilibrium and welfare in markets with financially constrained arbitrageurs. *Journal of Financial Economics*, 66, 2, 3, pp.nn361–407.https://doi.org/10.1016/S0304-405X(02)00228-4

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *Review of Financial Studies*, 33, 5, pp. 2223–2273. https://doi.org/10.1093/rfs/ hhaa009

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Investment Company Institute. (2024). *2024 Investment Company Fact Book* (64th ed.). https://doi.org/10.61138/ZKOW7328

Israeli, D., Lee, C. M. C., & Sridharan, S. A. (2017). Is there a dark side to exchangetraded funds? An information perspective. *Review of Accounting Studies*, 22, 3, pp. 1048–1083. https://doi.org/10.1007/s11142-017-9400-8

Lee, C. M. C., & Ready, M. J. (1991). Inferring trade direction from intraday data. *Journal of Finance*, 46, 2, pp. , 733–746. https://doi.org/10.1111/j.1540-6261.1991 .tb02683.x

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, pp.4765–4774. https://doi.org/10.48550/arXi v.1705.078744

Ndibe, O. S. (2024). National Cyber Resilience Index: A data-Driven Framework for Measuring Preparedness. *Journal of Computational Analysis and Application*. 33, 1, pp. 729-750.

Okolo, J. N. (2021). A Systematic Analysis of Artificial Intelligence and Data Science Integration for Proactive Cyber Defense: Exploring Methods, Implementation Obstacles, Emerging Innovations, and Future Security Prospects. *Communication in Physical Sciences*. 7, 4, pp. 681-696.

Omosunlade, O. (2024). Curriculum Framework for Entrepreneurial Innovation among Special Needs Students in the Age of Artificial Intelligence. *Communication in Physical Sciences*. 1, 4, pp. 1089- 1098.

Petajisto, A. (2017). Inefficiencies in the pricing of exchange-traded funds. *Financial Analysts Journal*, 73, 1, pp. , 24–54. https://doi.org/10.2469/faj.v73.n1.7

Sanni S. (2024). A Review on Machine Learning and Artificial Intelligence in Procurement: Building Resilient Supply Chains for Climate and Economic Priorities. *Communication in Physical Sciences*. 11, 4, pp. 1099-1111

Ugwo, P. & Chikezie, C. (2024). Personalization and explainability in fintech products: Understanding how interface choices influence user decisions. International Journal of Research in Management. 6, 1, pp. 556-567. https://www.doi.org/10.33545/26648792.2024.v6.i1f.568

Shleifer, A., & Vishny, R. W. (1997). The limits of arbitrage. *Journal of Finance*, 52,

1, pp. 35–55.https://doi.org/10.1111/j.1540-6261.1997.tb03807.x

Ufomba, P. O., Ndibe, O.S. (2023). IoT and Network Security: Researching Network Intrusion and Security Challenges in Smart Devices. *Communication in Physical Sciences*. 9, 4, pp. 784-800.

Amougou, R. S. E. (2023).AI-Driven DevOps: Leveraging Machine Learning for Automated Software Delivery Pipelines. *Communication in Physical Sciences*, 9, 4, pp. 1010-1021.

Declaration

Consent for publication

Not Applicable

Availability of data

Data shall be made available upon request

Ethical Considerations

Not applicable

Competing interest

The authors report no conflict or competing interest

Funding

The authors declared no source of funding

Authors' Contributions

A.O. conceptualized the study, designed the research framework, and supervised the overall investigation. V.E.A. developed the data architecture and conducted sentiment and retail order flow analysis. L.A. implemented the machine learning models and statistical validation. E.W. performed econometric interpretation and robustness checks. D.W. coordinated literature synthesis, manuscript drafting, and final editing of the study.