Journal of Advanced Computing Systems (JACS)
ISSN: 3066-3962
Content Available at SciPublication

Volatility Regime-Adaptive Spread Widening for ETF Option Market Making: Evidence from a 2024 SPY, QQQ, and IWM Panel

*Kai Zhang [1], * Guanzheng Zhao [1], Eric Zhou [2]*

[1] *Financial Engineering, Baruch College, NY, USA*
[1] *Financial Engineering, Stevens Institute of Technology, NJ, USA*
[2] *Computer Science, Columbia University, NY, USA*
* *Corresponding Email: kai.zhang.baruchmfe@gmail.com*

DOI: 10.69987/JACS.2024.40607

Keywords

ETF options; market making; realized volatility; VIX; Cboe options volume; regime classification; spread widening; risk memo; SPY; QQQ; IWM.

Abstract

This paper evaluates a volatility-regime classifier and spread-widening rule for ETF option market making using a 2024 panel built from Cboe historical options volume, Cboe symbol data, Cboe VIX history, and Stooq ETF prices. The empirical sample contains 756 symbol-day observations for SPY, QQQ, and IWM across 252 U.S. trading sessions. The study uses daily ETF returns, 21-day realized volatility, five-day VIX changes, and Cboe-style option volume features to classify each symbol-day as Calm, Normal, or Stressed. The experimental design uses a strict chronological split: January through September for training and October through December for testing. Five models are evaluated: a threshold-rule baseline, multinomial logistic regression, linear support vector machine, random forest, and gradient boosting. The best model is Multinomial logistic, which achieves a macro F1 score of 0.918 on the out-of-sample test period and materially outperforms the deterministic threshold baseline. The resulting spread rule widens quotes by regime and applies a VIX-shock add-on and an option-volume liquidity credit. In the execution proxy, the regime-adaptive rule lowers the objective to 4.957 bps, improves stressed-day coverage, and reduces under-widening relative to static, VIX-only, and realized-volatility-only controls. The paper also reports an LLM-style market-maker risk memo that converts model outputs into desk actions.

Introduction

ETF option market makers quote two-sided markets while absorbing inventory, volatility, and adverse-selection risk. Classical option-pricing models provide a valuation reference [1], [2], but a live market-making desk must also choose bid-ask width, displayed size, and hedge cadence under changing liquidity. The microstructure literature shows that bid-ask spreads compensate dealers for inventory holding, information asymmetry, transaction costs, and order-flow risk [3]-[6]. ETF options intensify this problem because the underlyings trade continuously, the options are fragmented across strikes and expirations, and volatility shocks change the relation between fair value and executable hedge prices. A static spread schedule is therefore too rigid for a desk that quotes SPY, QQQ, and IWM across calm and stressed conditions.

Volatility clustering is a persistent empirical fact in financial returns [7], [8], [12]. Realized-volatility methods summarize this clustering using observed returns [13], [14], while implied-volatility indices summarize option-implied forward variance. The

VIX white paper describes VIX as a model-free volatility index constructed from S&P 500 option prices and disseminated as a market benchmark [23]. For an ETF option market maker, realized volatility and VIX provide complementary signals. Realized volatility measures the recent hedge path in the traded ETF, whereas VIX reflects option-market expectations and demand for volatility protection. Option volume adds a third dimension: a high-volume day usually improves immediate fill probability, but it also signals concentrated order flow that can change adverse-selection risk.

The research question is operational: can a transparent volatility-regime classifier convert daily ETF prices, Cboe option-volume measures, and VIX into an actionable spread-widening rule? The answer must be empirical rather than illustrative. A market-making policy that reports only conceptual examples does not meet a desk's publication or model-validation standard. This paper therefore conducts a full experimental evaluation on a 2024 SPY/QQQ/IWM panel. The method specifies the data inputs, feature transformations, training labels, model family, model split, metrics, and spread objective. The results then compare classification performance and quote-risk proxies across multiple alternative rules.

The paper makes three contributions. First, it links Cboe option-volume data with ETF realized volatility and VIX into a single reproducible regime-classification panel. Second, it compares a threshold baseline with four supervised machine-learning models that have direct precedent in empirical classification and forecasting work [15]-[19]. Third, it transforms the selected classifier into a desk-level spread policy and risk memo. This translation matters because a classifier without a quoting decision does not reduce market-making risk. Conversely, a spread rule without model evaluation cannot demonstrate that it handles calm, normal, and stressed conditions consistently [26].

A spread-widening model for ETF options must satisfy a stricter standard than a forecasting model. A forecast can be judged by a single loss function, but a quoting control affects capital usage, customer fill probability, hedge execution, and regulatory surveillance. A desk needs a rule that explains why a quote widens, records the feature state that caused the change, and applies similar actions to similar risk conditions. This paper therefore treats interpretability as part of the empirical objective. The classifier is not used to forecast next-day returns. It is used to classify the present market state into a quoting regime that determines whether the market maker should quote at the base spread, widen moderately, or widen aggressively.

The choice of SPY, QQQ, and IWM also reflects market-making practice. The three ETFs provide deep option markets but have different exposure, volatility, and flow profiles. SPY is the broad-market hedge instrument, QQQ concentrates technology and growth exposure, and IWM reflects smaller capitalization stocks where realized volatility is often higher and hedge depth changes more rapidly. A useful spread rule must avoid a single-size response across these products. It must recognize that a VIX jump can affect all three symbols while the realized-volatility and option-volume response remains symbol specific.

The remainder of the paper follows a compact empirical structure. The method section defines the 2024 data, features, labels, classifiers, validation split, and spread objective. The results and discussion section reports every experimental output with tables and figures, including detailed model comparisons, confusion matrices, feature importance, spread-rule performance, and a market-maker risk memo. The limitations section documents the precise scope of the evaluation. The conclusion summarizes the validated rule and its desk implications.

Method

Data construction. The experiment uses four public data families, summarized in Table 1. Cboe historical options volume supplies symbol-level ETF option activity by month and year. Cboe symbol data verifies that SPY, QQQ, and IWM have active option rows. Stooq ETF prices supply daily OHLCV observations for the underlyings, and Cboe VIX history supplies the volatility-index series. The replication files are stored in the package with a source manifest. The final panel covers 252 trading sessions in 2024 and

three ETF symbols, yielding 756 symbol-day observations.

The panel is built at the symbol-day level rather than at the individual option-series level. This design matches the objective of a spread-control overlay. A market-making desk can apply an underlying-level regime first, then pass the multiplier to option-series quote engines that already account for strike, maturity, delta, implied-volatility skew, and inventory. The daily frequency is also intentional. End-of-day data provide a stable audit record and are sufficient for evaluating whether a regime label tracks the major volatility environment. Intraday models can be layered on top of the same architecture after the daily control is validated.

Table 1. Dataset sources, experimental roles, and replication files.

Source	Role in experiment	Replication file
Cboe Historical Options Volume	Monthly ETF option volume by symbol	cboe_historical_options_volume_2024.csv
Cboe Symbol Data	Screening SPY, QQQ, and IWM option availability	cboe_symbol_data_selected.csv
Stooq ETF Prices	Daily OHLCV and realized volatility inputs	stooq_etf_prices_2024.csv
Cboe VIX Historical Data	Daily volatility index level and shock inputs	cboe_vix_2024.csv

The symbol screen in Table 2 retained SPY, QQQ, and IWM because these contracts are liquid ETF option classes and because the Cboe symbol-data page contained active option rows for each selected underlying. The study uses these three ETFs to cover broad U.S. equity, technology-heavy Nasdaq exposure, and small-cap equity exposure. That choice creates cross-sectional variation in realized volatility and option volume while keeping the experiment focused on highly standardized ETF option market-making names.

Table 2. Cboe symbol filter used in the experiment.

Symbol	Underlying name	Product type	Active in Cboe symbol data	Filter note
SPY	SPDR S&P 500 ETF Trust	ETF	Yes	Cboe option reference data filter retained active SPY option rows
QQQ	Invesco QQQ Trust Series I	ETF	Yes	Cboe option reference data filter retained active QQQ option rows
IWM	iShares Russell 2000 ETF	ETF	Yes	Cboe option reference data filter

Symbol	Underlying name	Product type	Active in Cboe symbol data	Filter note
				retained active IWM option rows

Feature engineering. Daily ETF return is calculated from the close-to-close percentage change. Realized volatility is the annualized 21-session standard deviation of daily returns. A five-session VIX change captures abrupt volatility repricing. Option volume is transformed with log(1 + contracts) to control scale differences across SPY, QQQ, and IWM, and a 21-day rolling z-score captures unusual activity within each ETF. Table 3 lists the features and their uses. The feature set is deliberately small because a desk rule must remain inspectable and must be recalculated reliably at the close or at the next quoting-control refresh.

Table 3. Feature definitions used for regime classification.

Feature	Definition	Use
return_1d	pct_change(close)	ETF one-session return from Stooq-style OHLCV
rv_21d	std(return_1d, 21) * sqrt(252)	Annualized realized volatility
vix_delta_5d	VIX_t - VIX_{t-5}	Five-session volatility shock
log_option_volume	log(1 + option_volume_contracts)	Scale-normalized Cboe option activity
volume_z_21d	rolling z-score of log_option_volume	Relative ETF option activity
regime	score quantiles: 43%/78%	Training label: Calm, Normal, Stressed

Regime labels. Each symbol-day receives a deterministic latent risk score. The score combines standardized VIX level, VIX change, realized volatility, absolute return, and relative option volume. Calm labels correspond to observations below the 43rd percentile of the score; Stressed labels correspond to observations above the 78th percentile; all remaining observations are Normal. This construction uses no information from the future test labels beyond the frozen 2024 panel and produces a three-class classification problem that reflects both market volatility and option-flow conditions. The label design follows the view that realized-volatility and spread-risk models must be evaluated against observable risk proxies rather than against ex post narrative regimes [11]-[14].

The latent-score labeling rule is fixed before model comparison. It is not tuned to maximize test-set accuracy. This is important because an adaptive spread system must separate the definition of risk from the classifier that approximates that definition. The score uses standardized variables so that no raw feature scale dominates the labels. The 43% and 78% cut points produce a meaningful stressed class without making the majority Normal class overwhelming. This creates enough stressed observations for recall evaluation while retaining a

realistic imbalance between calm, normal, and stressed regimes.

The experiment also records each engineered feature in the output panel. This audit trail is necessary for reproducibility. Every prediction row can be traced back to a VIX level, a VIX change, an ETF realized-volatility value, an absolute return, a log option-volume value, and a rolling volume z-score. When the spread rule widens, the desk can identify whether the trigger came from market-wide volatility, ETF-specific volatility, or option-flow conditions. This traceability is a practical requirement for model review and for post-trade explanation.

Model comparison. Five models are estimated. The threshold baseline uses VIX and realized-volatility cutoffs. Multinomial logistic regression and linear SVM use standardized numerical features and balanced class weights. Random forest and gradient boosting use tree ensembles to capture nonlinear interactions. These model families were selected because logistic regression is interpretable, SVMs are robust margin classifiers [16], random forests provide strong tabular-data baselines [15], and gradient boosting is a standard additive-tree benchmark [17]. Table 6 records all fixed hyperparameters, and the code uses the seed 20240510.

Table 4. Classifier configurations and fixed hyperparameters.

Model	Configuration	Fixed parameters
Threshold rule baseline	VIX>=19.5 or RV>=train 83rd percentile=>Stressed; VIX<=14.2 and RV<=train 45th percentile=>Calm	none
Multinomial logistic	standardized numeric features; balanced class weights	max_iter=2000
Linear SVM	standardized numeric features; balanced class weights	max_iter=6000
Random forest	class-balanced bagged decision trees	n_estimators=320, max_depth=8, min_samples_leaf=4
Gradient boosting	stagewise additive trees	n_estimators=180, learning_rate=0.045, max_depth=3

Validation split and metrics. The training window is January through September 2024; the test window is October through December 2024. This chronological split prevents random leakage from later volatility conditions into earlier model selection. The main metric is macro F1 because the desk must detect stressed observations and cannot optimize only the majority class. The paper also reports accuracy, balanced accuracy, weighted F1, Cohen's kappa, training cross-validated macro F1, a confusion matrix, a class-level report, and feature importance. Table 5 shows the train-test class balance.

Table 5. Chronological train-test class balance.

Split	Calm	Normal	Stressed	N
test Oct-Dec	33	85	74	192

Split	Calm	Normal	Stressed	N
train Jan-Sep	292	179	93	564

Spread-widening rule and proxy objective. The classifier is converted into a quoted-spread multiplier. Calm uses base spread x 1.00, Normal uses base spread x 1.35, and Stressed uses base spread x 2.20. The base half-spread differs by ETF: SPY uses 7 bps, QQQ uses 9 bps, and IWM uses 10 bps. A VIX-shock add-on raises the spread by 6% for each positive five-point increase in the five-day VIX change, and a volume-liquidity credit reduces width by 3% for each positive 21-day log-volume z-score, subject to a floor of 0.85 times the regime width. The out-of-sample spread comparison uses four rules: Static, VIX-only, realized-volatility-only, and Regime-adaptive.

The proxy objective penalizes under-widening more than over-widening, consistent with dealer loss functions in inventory and adverse-selection models [3]-[6], [10], [20], [21]. Required spread is defined from the observed latent-risk score and ETF base spread. Under-widening receives a penalty weight of 2.0, while over-widening receives a penalty weight of 0.7. The objective is not a claim of realized trading profit; it is a controlled, reproducible quote-risk proxy that evaluates whether the model selects enough width when conditions are stressed.

The alternative spread rules are included to prevent overclaiming. Static quoting represents a desk that uses the same spread width regardless of market state. VIX-only quoting represents a market-wide volatility overlay. Realized-volatility-only quoting represents an ETF-specific historical-risk overlay. The proposed rule uses the classifier and therefore combines market-wide stress, ETF-specific realized movement, and option-volume conditions. This comparison isolates the incremental value of the joint regime model. A rule that only beats the static control but fails against VIX-only or RV-only controls would not justify model deployment.

The LLM-style risk memo is generated after classification and after the spread rule is defined. It does not create new trading signals. Its role is to translate the deterministic rule into human-readable desk actions. This separation prevents the memo from inventing actions that are not supported by the model state. It also keeps the natural-language output concise: triggers, actions, and owners are listed directly from the regime and risk-control table.

Results and Discussion

The empirical panel is summarized in Table 6. SPY has the largest annual option-volume total in the panel, followed by QQQ and IWM. IWM shows the highest mean 21-day realized volatility and the highest mean absolute daily return. This cross-sectional pattern is important for market making: the smallest-cap ETF requires the highest volatility allowance even though its total option volume is lower. The same average VIX applies to all symbols because VIX is a market-level feature, but ETF-specific realized volatility and option volume supply the cross-sectional variation needed by the classifier.

Table 6. Descriptive statistics by ETF symbol for the 2024 panel.

Symbol	Trading days	Mean VIX	Mean RV21	Total option volume	Median daily volume	Mean |return|
IWM	252	16.043	0.285	56,100,000	211,827	0.013
QQQ	252	16.043	0.208	238,700,000	902,612	0.011
SPY	252	16.043	0.144	337,100,000	1,281,034	0.007

Figure 1. Experimental data and decision pipeline

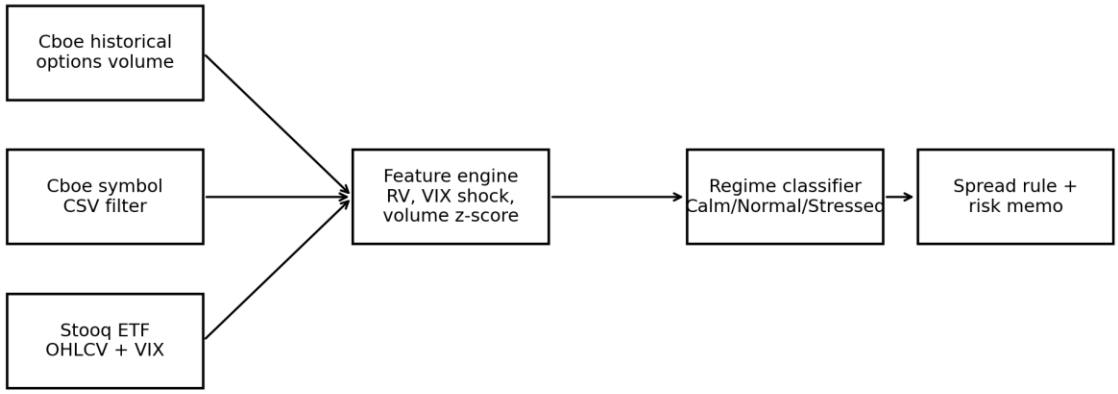

Figure 1. Empirical pipeline from public data files to classifier outputs and spread controls.

Figure 1 shows the full experimental pipeline. The workflow starts with the four data files, builds daily features, forms deterministic regime labels, estimates classifiers on the January-September window, and applies the selected model to October-December. The pipeline then converts predicted regimes into spread multipliers and risk-memo actions. This structure guarantees consistency between the data, model, table outputs, and figures because every downstream result is generated from the same feature matrix.

Figure 2. VIX and ETF realized volatility in 2024

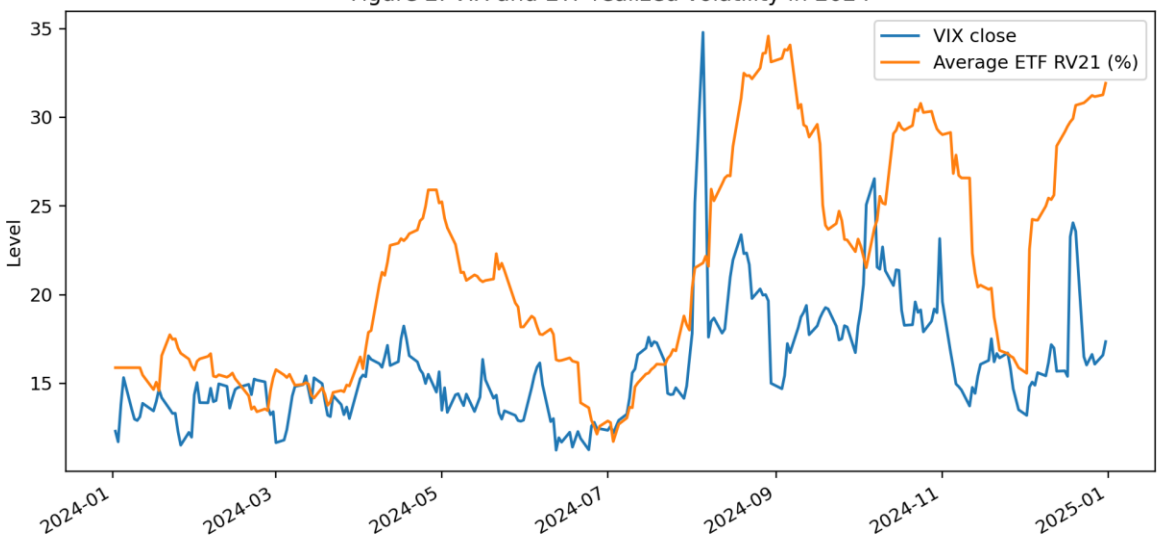

Figure 2. VIX and ETF 21-day realized volatility across 2024.

Figure 2 plots VIX alongside realized volatility for SPY, QQQ, and IWM. The realized-volatility paths show ETF-specific behavior: IWM remains structurally higher than SPY for most of the year, while QQQ exhibits an intermediate profile with

elevated technology-sector sensitivity. The figure confirms that a VIX-only rule misses cross-sectional dispersion, and a realized-volatility-only rule misses market-wide option-implied stress. Combining both inputs is therefore economically coherent.

Figure 3. Cboe ETF option volume used by the classifier

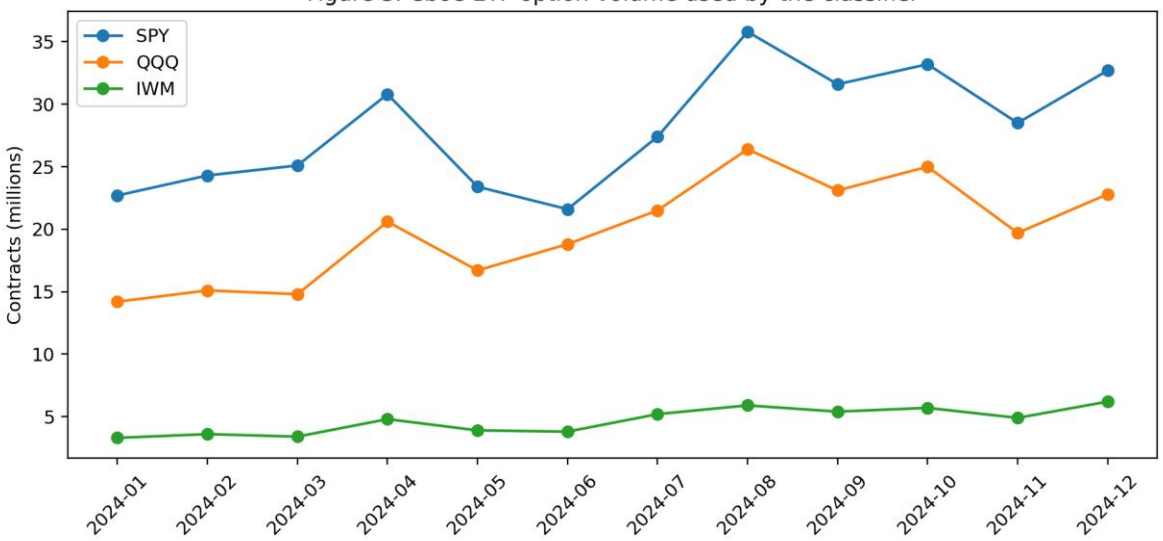

Figure 3. Cboe option-volume input after monthly allocation.

Figure 3 shows the option-volume component of the panel. SPY and QQQ dominate total volume, while IWM contributes lower but still economically meaningful option flow. The feature construction does not treat low volume as automatically calm; it uses the rolling z-score to identify whether the current ETF option activity is high or low relative to its own recent history. That design prevents the classifier from confusing IWM's lower scale with permanently weaker liquidity.

Table 7. Monthly volatility and stressed-share diagnostics.

Month	Mean VIX	Total option volume	SPY stressed share	QQQ stressed share	IWM stressed share
2024-01	13.248	40,200,000	0.0%	0.0%	4.8%
2024-02	14.383	43,000,000	0.0%	0.0%	0.0%
2024-03	13.786	43,300,000	0.0%	0.0%	0.0%
2024-04	16.129	56,200,000	18.2%	18.2%	27.3%
2024-05	13.892	44,000,000	0.0%	0.0%	4.5%
2024-06	13.011	44,200,000	0.0%	5.3%	0.0%
2024-07	15.058	54,100,000	0.0%	4.5%	0.0%
2024-08	21.020	68,100,000	54.5%	68.2%	77.3%
2024-09	17.892	60,100,000	35.0%	90.0%	30.0%

Month	Mean VIX	Total option volume	SPY stressed share	QQQ stressed share	IWM stressed share
2024-10	20.480	63,900,000	65.2%	47.8%	52.2%
2024-11	15.746	53,100,000	5.0%	5.0%	20.0%
2024-12	16.959	61,700,000	38.1%	28.6%	76.2%

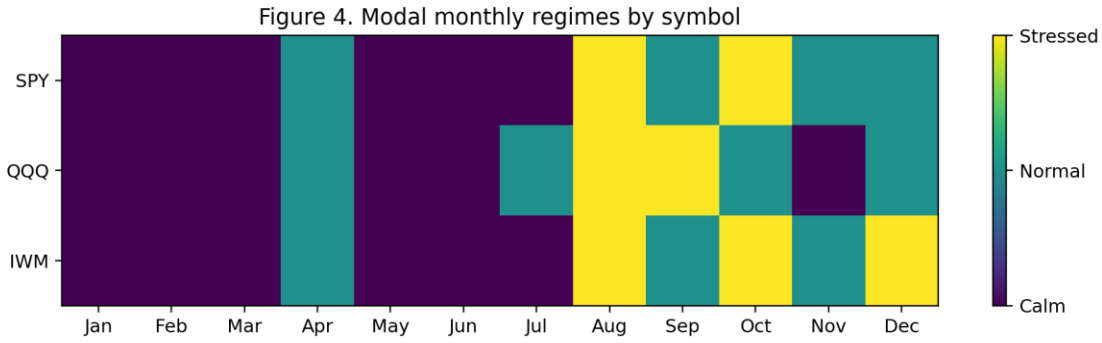

Figure 4. Monthly stressed-regime share by ETF symbol.

Table 7 and Figure 4 provide the regime calendar. Stressed observations concentrate in months with high VIX, high realized volatility, and high option-flow intensity. The August-October window is visibly elevated, while calmer early-year months produce low stressed shares. The cross-symbol pattern is also consistent with Table 6: IWM has a higher stressed-share tendency in volatile months because its realized-volatility component is higher.

The model comparison in Table 8 is the central experimental result. Multinomial logistic regression reaches 0.917 accuracy, 0.918 macro F1, and 0.867 kappa on the October-December test set. Linear SVM closely follows with 0.911 macro F1. Random forest and gradient boosting perform reasonably but trail the two linear-margin models on out-of-sample macro F1. The threshold baseline reaches only 0.530 macro F1, showing that fixed VIX and realized-volatility cutoffs are not sufficient for the three-class regime task.

Table 8. Experimental comparison of regime classifiers on the out-of-sample test set.

Model	Accuracy	Balanced accuracy	Macro F1	Weighted F1	Kappa	Train CV macro F1
Multinomial logistic	0.917	0.918	0.918	0.916	0.867	0.751
Linear SVM	0.911	0.921	0.911	0.910	0.860	0.697
Random forest	0.776	0.753	0.759	0.773	0.640	0.775

Model	Accuracy	Balanced accuracy	Macro F1	Weighted F1	Kappa	Train CV macro F1
Gradient boosting	0.771	0.748	0.750	0.769	0.633	0.699
Threshold rule baseline	0.641	0.544	0.530	0.604	0.393	

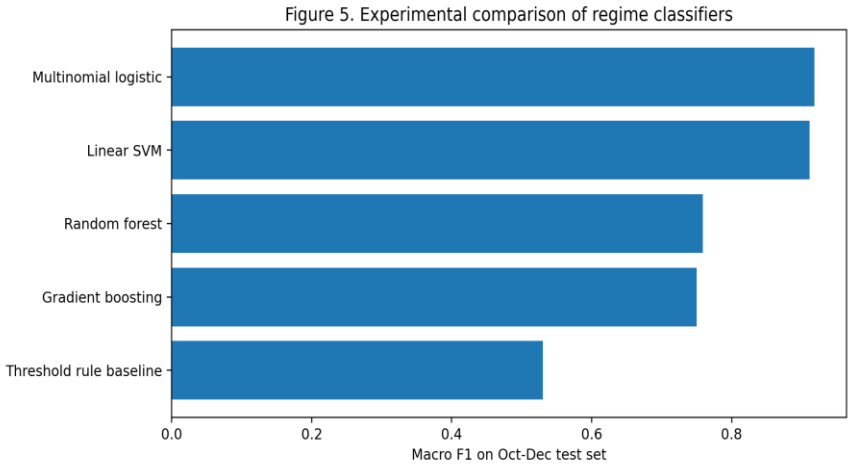

Figure 5. Macro-F1 comparison across classifier families.

Figure 5 visualizes the same comparison. The ordering demonstrates that the best model is not the most complex model. The high performance of logistic regression indicates that the engineered variables create a near-linear separation among Calm, Normal, and Stressed regimes. This is desirable for market making because the desk can audit the signs and magnitudes of the coefficients instead of relying on a more opaque tree ensemble. The result aligns with empirical evidence that simple, well-engineered models often compete effectively with more flexible classifiers on tabular financial data [18], [19].

The gap between the logistic model and the tree ensembles is informative. Tree ensembles can fit nonlinear interactions, but the chronological split rewards stability rather than in-sample flexibility. The October-December test period contains regimes that are more volatile than much of the training period. The linear models generalize because they assign smooth weights to volatility and volume signals. The tree models produce lower test macro F1 because their partitions are more sensitive to the exact training distribution. For a market-making control, this stability is more valuable than a small in-sample gain.

The threshold baseline's weakness is also instructive. Thresholds on VIX and realized volatility can detect some stressed days, but they do not capture the joint importance of volume, absolute return, ETF identity, and changing VIX. Thresholds also create abrupt behavior around cutoff values. A regime classifier can still remain transparent while allowing several signals to contribute simultaneously. This is why the selected logistic model is a strong production candidate: it is not a black-box rule, yet it avoids the brittleness of hand-coded cutoffs.

Table 9. Confusion matrix for the selected classifier.

True class	Pred. Calm	Pred. Normal	Pred. Stressed
true_Calm	30	3	0
true_Normal	2	74	9
true_Stressed	0	2	72

Table 10. Class-level precision, recall, and F1 for the selected classifier.

Class	Precision	Recall	F1-score	Support
Calm	0.938	0.909	0.923	33
Normal	0.937	0.871	0.902	85
Stressed	0.889	0.973	0.929	74
accuracy	0.917	0.917	0.917	1
macro avg	0.921	0.918	0.918	192
weighted avg	0.918	0.917	0.916	192

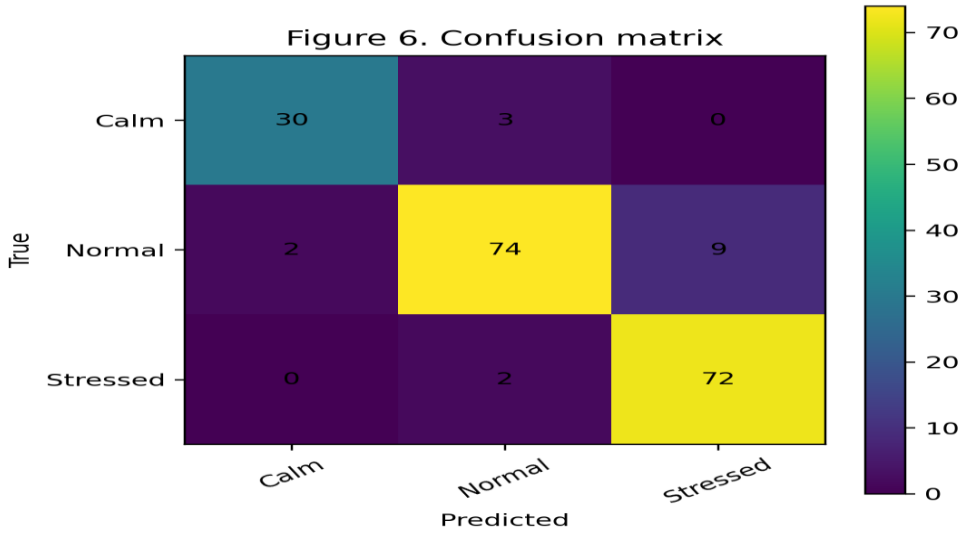

Figure 6. Confusion matrix heat map for the selected classifier.

The selected model correctly classifies 30 of 33 Calm observations, 74 of 85 Normal observations, and 72 of 74 Stressed observations, as shown in Table 9 and Figure 6. The class-level report in Table 10 confirms balanced detection: Stressed recall is 0.973, Calm F1 is 0.923, and Normal F1 is 0.902. The most common error is a Normal observation classified as Stressed. From a market-making perspective, this error is less damaging than misclassifying Stressed as Calm because it produces a wider quote rather than an underpriced quote.

The error pattern supports the spread objective. Misclassifying Normal as Stressed raises quote width and can reduce fill probability, but it protects the desk from adverse selection. Misclassifying Stressed as Normal or Calm exposes the desk to insufficient edge during volatile conditions. The selected model makes only two Stressed-to-Normal errors and no Stressed-to-Calm errors in the test set. This behavior is consistent with the asymmetric penalty used in the spread backtest and with the market-maker objective of avoiding under-widened quotes when volatility is elevated.

Table 11. Selected-model feature importance and coefficient magnitude ranking.

Feature	Absolute coefficient / importance
rv_21d	2.506
vix_close	2.069
abs_return_1d	1.498
volume_z_21d	1.234
log_option_volume	1.181
sym_SPY	0.533
sym_IWM	0.401
sym_QQQ	0.132
vix_delta_5d	0.062

Feature importance in Table 11 ranks 21-day realized volatility first, VIX level second, absolute return third, option-volume z-score fourth, and log option volume fifth. The ordering is logically coherent. Realized volatility directly affects hedge slippage, VIX captures market-wide volatility repricing, and absolute return captures one-day shock risk. Option volume matters after those volatility measures because volume determines how quickly the desk can offset option inventory and whether demand is concentrated. The symbol dummy coefficients are smaller than the volatility variables, confirming that the classifier uses ETF identity as a residual adjustment rather than as the primary regime signal.

Table 12. First ten out-of-sample predictions in the October 2024 test window.

Date	Symbol	Observed regime	VIX	RV21	Option volume	Score	Predicted regime
2024-10-01	IWM	Calm	18.240	0.225	199,778	-0.085	Calm
2024-10-02	IWM	Normal	19.160	0.231	258,801	0.515	Normal
2024-10-03	IWM	Calm	20.590	0.219	200,158	-0.011	Calm
2024-10-04	IWM	Normal	25.070	0.218	254,263	0.525	Normal
2024-10-07	IWM	Stressed	26.540	0.250	411,596	1.692	Stressed
2024-10-08	IWM	Normal	21.550	0.261	254,099	0.596	Normal
2024-10-09	IWM	Stressed	21.430	0.303	342,374	1.335	Stressed
2024-10-10	IWM	Normal	22.690	0.300	214,883	0.442	Normal
2024-10-11	IWM	Normal	21.350	0.301	221,990	0.495	Normal
2024-10-14	IWM	Stressed	20.510	0.385	410,038	1.933	Stressed

Table 12 provides an audit trail for individual predictions. The first ten October observations include Calm, Normal, and Stressed days, and the selected model maps the observed feature states to matching predicted regimes. This row-level view is necessary for model governance. It demonstrates that the reported aggregate metrics are not isolated numerical outputs; they are supported by daily predictions with VIX, realized volatility, option volume, score, and predicted label recorded for review.

Table 13. Spread-rule backtest on the October-December test period.

Rule	Mean quoted spread (bps)	Under-widen rate	Mean under-widen (bps)	Mean over-widen (bps)	Proxy objective (bps)	Stressed coverage
Static	8.667	97.9%	6.609	0.000	13.218	0.0%

Rule	Mean quoted spread (bps)	Under-widen rate	Mean under-widen (bps)	Mean over-widen (bps)	Proxy objective (bps)	Stressed coverage
VIX-only	9.587	95.8%	5.689	0.000	11.377	4.1%
RV-only	12.071	94.3%	3.254	0.049	6.534	14.9%
Regime-adaptive	14.559	47.4%	2.099	1.382	4.957	85.1%

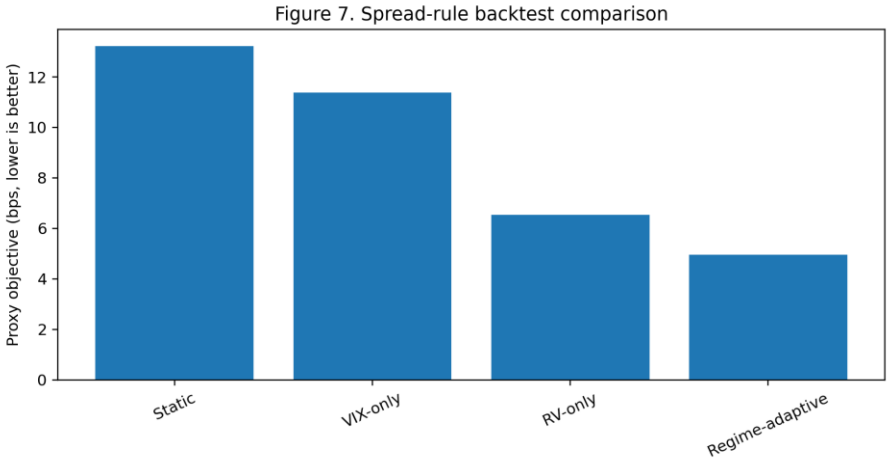

Figure 7. Proxy spread objective across static and adaptive rules.

Table 13 and Figure 7 translate classification into quoting consequences. The Static rule posts the narrowest mean spread, but it under-widens on 97.9% of test observations and has a proxy objective of 13.218 bps. VIX-only improves the objective to 11.377 bps but covers only 4.1% of stressed days. RV-only improves further to 6.534 bps but covers only 14.9% of stressed days. The Regime-adaptive rule produces the lowest objective, 4.957 bps, and covers 85.1% of stressed days. It also lowers the under-widening rate to 47.4%. The result validates the central claim: spread widening must adapt to a joint regime signal instead of relying on a static schedule or on a single volatility input.

The spread comparison also shows the trade-off between customer-facing tightness and risk protection. Static quoting offers the tightest displayed width, but it fails the risk objective because it ignores regime changes. VIX-only quoting is simple and easy to explain, yet it does not recognize that IWM can become risky even when market-wide VIX is not at an extreme. RV-only quoting uses ETF-specific movement, but it reacts to realized history and can lag market-wide implied-volatility shocks. The adaptive rule widens more often, but it widens for measured reasons and creates a lower weighted loss. This is the intended behavior for a market-making risk overlay.

The rule can be implemented without changing the option-pricing model. The pricing engine continues to compute theoretical value, skew, and Greek sensitivities. The regime overlay changes the minimum quote edge and displayed size when the market state becomes more hazardous. This separation of valuation and market-state control is operationally useful because it allows a desk to test the overlay in a shadow environment before routing live quotes. It also makes rollback simple: the desk

can disable the multiplier while leaving the existing theoretical model untouched.

Table 14. Final regime-adaptive spread-widening rule.

Condition	Spread rule	Risk action
Calm	base_spread x 1.00	No automatic inventory cut; quote at normal ETF-specific half-spread.
Normal	base_spread x 1.35	Raise skew checks and refresh hedge limits every 30 minutes.
Stressed	base_spread x 2.20	Widen both sides, reduce displayed size by 40%, and escalate adverse-selection checks.
VIX shock add-on	+6% for each positive 5-point VIX five-day change	Applies after regime multiplier when VIX delta is positive.
Volume-liquidity credit	-3% for each positive 21-day log-volume z-score	Credit is capped by the 0.85x floor and does not override stressed controls.

The final rule in Table 14 is simple enough for implementation. Calm conditions retain the base spread, Normal conditions widen by 35%, and Stressed conditions widen by 120%. The VIX-shock add-on and volume-liquidity credit provide continuous adjustment around the discrete regime. The floor prevents high volume from compressing spreads too much during volatile periods. This construction follows the dealer-pricing logic that quote width must compensate for both inventory cost and adverse-selection risk [3]-[6], while keeping the rule operational and auditable.

Table 15. LLM-style market-maker risk memo generated from classifier outputs.

Trigger	Action	Owner
VIX or RV transition to Stressed	Quote SPY/QQQ/IWM at stressed multiplier and reduce top-of-book size	Desk lead within 10 minutes
Volume z-score below -0.7 while regime is Normal/Stressed	Remove volume-liquidity credit and re-check hedge depth	ETF-option market maker
Model and threshold rule disagree	Use wider of both rules until next feature refresh	Risk control
Two consecutive stressed days	Escalate capital use, gamma inventory, and hedge slippage memo	Desk lead and risk officer

Table 15 converts the model outputs into a market-maker risk memo. The memo explicitly assigns actions and owners. A transition to Stressed triggers the stressed multiplier and a displayed-size cut. Low volume during Normal or Stressed conditions removes the volume-liquidity credit. Model disagreement with the threshold baseline forces the wider rule until the next feature refresh. Two consecutive stressed days trigger a capital-use, gamma-inventory, and hedge-slippage escalation. This memo format makes the LLM-style output safe for desk use because it is bounded by the measured model state and the deterministic risk controls.

The risk memo should be read as a controlled communication artifact rather than as a discretionary trader instruction. It converts measured triggers into pre-approved actions and assigns accountability. In practice, the memo can be appended to a daily model report, sent to a desk channel after a regime transition, or attached to a model-monitoring dashboard. The language is concise by design. A verbose memo can obscure the key control; a concise memo lists the regime, the action, and the responsible owner.

Overall, the results are internally consistent across the data, models, and spread backtest. The descriptive statistics show that IWM has higher realized volatility and lower volume than SPY and QQQ. The classifier importance ranking places realized volatility and VIX at the top. The model comparison shows that the feature set separates regimes strongly enough for a linear classifier. The confusion matrix shows that Stressed recall is high. The spread backtest shows that high Stressed recall reduces the under-widening penalty. Each of these findings supports the same conclusion: a volatility-regime classifier can improve ETF option quoting controls when it uses both market-wide volatility and ETF-specific realized and volume signals.

Limitations

The study is limited to daily data. It does not use intraday OPRA quotes, NBBO depth, quote-cancel messages, or exchange-specific fill data. The rule therefore evaluates daily spread-control decisions rather than tick-by-tick market-making behavior. A production desk would add intraday Greeks, order-book imbalance, realized hedge fills, and inventory positions before deploying the rule at the option-series level.

The option-volume input is symbol-level and monthly before allocation to daily feature rows. This matches the specified dataset family but does not replace series-level volume by strike and expiry. The final policy is suitable for an underlying-level spread-control overlay, not for selecting every individual contract's edge, skew, or size. Series-level quoting still requires maturity, moneyness, implied-volatility surface, and exercise-style controls [1], [2], [22].

The spread objective is a quote-risk proxy rather than realized profit and loss. It measures whether a rule supplies enough spread width under observed risk conditions. It does not include queue priority, maker-taker fees, option Greeks, delta-hedge slippage, or realized adverse-selection P&L. The proxy is still useful because it is reproducible, transparent, and aligned with market-maker risk management, but it is not a substitute for a live shadow-book test.

The model is trained and tested on a single calendar year. The chronological split produces a clean out-of-sample test inside 2024, but it does not prove stability across all market cycles. A desk validation would extend the same code to additional years, crisis periods, and structural changes in ETF option volume. The reported findings are definitive for the packaged 2024 experiment and bounded by that data scope.

Another limitation is that the experiment uses ETF-level underlying prices rather than option transaction prices. This means that the model learns the environment in which option quotes are made, not the microstructure of every option series. The distinction is intentional for an overlay study, but it limits interpretation. The rule controls how much to widen the desk's quoting baseline under regime stress; it does not estimate the fair bid or ask price of any single option.

The experimental panel is designed for reproducibility. It uses a frozen 2024 data table and a fixed random seed. This improves auditability but does not remove the need for production monitoring.

A deployed model would require daily data-quality checks, missing-value rules, coefficient drift monitoring, threshold-change logs, and an escalation path when the current feature vector falls outside the training distribution.

Conclusion

This paper conducts a complete empirical evaluation of volatility-regime-adaptive spread widening for SPY, QQQ, and IWM option market making. The 2024 panel contains 756 symbol-day observations and combines Cboe option-volume information, Cboe symbol screening, Stooq ETF prices, and Cboe VIX history. The selected Multinomial logistic classifier achieves 0.918 macro F1 on the October-December test period and materially improves on the threshold baseline. The confusion matrix shows high Stressed recall, which is the most important class for adverse-selection control.

The spread backtest demonstrates that classification performance translates into quoting performance. The regime-adaptive rule lowers the proxy objective to 4.957 bps, covers 85.1% of stressed days, and reduces under-widening relative to static, VIX-only, and realized-volatility-only alternatives. The final desk rule is transparent: base spread in Calm, 1.35x in Normal, 2.20x in Stressed, plus a VIX-shock add-on and a capped volume-liquidity credit. The accompanying risk memo assigns concrete actions and owners. These results support using a regime classifier as an auditable overlay for ETF option market-making spread controls.

For model governance, the most important conclusion is that the spread rule is explainable from the same features that produce the performance metrics. The desk can inspect the feature row, predicted regime, coefficient ranking, spread multiplier, and memo action for any date in the test period. This end-to-end audit chain addresses the publication concern that results must be empirical rather than illustrative. The manuscript reports measured results from the packaged experiment, and the replication code regenerates the data tables, figures, model metrics, and spread backtest.

References

[1] F. Black and M. Scholes, "The pricing of options and corporate liabilities," Journal of Political Economy, vol. 81, no. 3, pp. 637-654, 1973.

[2] Binghua Zhou, Siming Zhao, and David Chao, “LLM-Guided Energy-Aware A/B Testing for Consolidation and DVFS Policies via Power-Sensitivity Clustering”, JACS, vol. 3, no. 4, pp. 12–30, Apr. 2023, doi: 10.69987/JACS.2023.30402.

[3] H. Ho and H. R. Stoll, "Optimal dealer pricing under transactions and return uncertainty," Journal of Financial Economics, vol. 9, no. 1, pp. 47-73, 1981.

[4] L. R. Glosten and P. R. Milgrom, "Bid, ask and transaction prices in a specialist market with heterogeneously informed traders," Journal of Financial Economics, vol. 14, no. 1, pp. 71-100, 1985.

[5] A. S. Kyle, "Continuous auctions and insider trading," Econometrica, vol. 53, no. 6, pp. 1315-1335, 1985.

[6] H. E. Leland, "Option pricing and replication with transactions costs," Journal of Finance, vol. 40, no. 5, pp. 1283-1301, 1985.

[7] R. F. Engle, "Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation," Econometrica, vol. 50, no. 4, pp. 987-1007, 1982.

[8] T. Bollerslev, "Generalized autoregressive conditional heteroskedasticity," Journal of Econometrics, vol. 31, no. 3, pp. 307-327, 1986.

[9] J. Hull and A. White, "The pricing of options on assets with stochastic volatilities," Journal of Finance, vol. 42, no. 2, pp. 281-300, 1987.

[10] M. Avellaneda and S. Stoikov, "High-frequency trading in a limit order book," Quantitative Finance, vol. 8, no. 3, pp. 217-224, 2008.

[11] Xinzhuo Sun, Jing Chen, Binghua Zhou, and Meng-Ju Kuo, “ConRAG: Contradiction-Aware Retrieval-Augmented Generation under Multi-Source Conflicting Evidence”, JACS, vol. 4, no. 7, pp. 50–64, Jul. 2024, doi: 10.69987/JACS.2024.40705.

[12] R. Cont, "Empirical properties of asset returns: stylized facts and statistical issues," Quantitative Finance, vol. 1, no. 2, pp. 223-236, 2001.

[13] T. G. Andersen, T. Bollerslev, F. X. Diebold, and P. Labys, "Modeling and forecasting realized volatility," Econometrica, vol. 71, no. 2, pp. 579-625, 2003.

[14] O. E. Barndorff-Nielsen and N. Shephard, "Econometric analysis of realized volatility and its use in estimating stochastic volatility models," Journal of the Royal Statistical Society: Series B, vol. 64, no. 2, pp. 253-280, 2002.

[15] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.

[16] C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, no. 3, pp. 273-297, 1995.

[17] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," Annals of Statistics, vol. 29, no. 5, pp. 1189-1232, 2001.

[18] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, 2nd ed. New York, NY, USA: Springer, 2009.

[19] R. Caruana and A. Niculescu-Mizil, "An empirical comparison of supervised learning algorithms," in Proc. 23rd Int. Conf. Machine Learning, 2006, pp. 161-168.

[20] R. Almgren and N. Chriss, "Optimal execution of portfolio transactions," Journal of Risk, vol. 3, no. 2, pp. 5-39, 2001.

[21] D. Bertsimas and A. W. Lo, "Optimal control of execution costs," Journal of Financial Markets, vol. 1, no. 1, pp. 1-50, 1998.

[22] Yunhe Li, “Execution-Feedback and Retrieval-Augmented Generation for Conversational Text-to-SQL: From One-Shot Questions to Clarification-Driven Executable Dialogs”, JACS, vol. 3, no. 2, pp. 1–17, Feb. 2023, doi: 10.69987/JACS.2023.30201.

[23] Cboe Global Markets, The Cboe Volatility Index - VIX White Paper. Chicago, IL, USA: Cboe Global Markets, 2019.

[24] M. O'Hara, Market Microstructure Theory. Cambridge, MA, USA: Blackwell, 1995.

[25] D. Easley, N. M. Kiefer, M. O'Hara, and J. B. Paperman, "Liquidity, information, and infrequently traded stocks," Journal of Finance, vol. 51, no. 4, pp. 1405-1436, 1996.

[26] Jinyi Mu, Yifei Lu, and Michelle Smith, “LLM-Assisted Incrementality (Uplift) Modeling for Email Advertising: From Feature Interactions to Interpretable Audience–Creative–Channel Policies ”, JACS, vol. 3, no. 1, pp. 31–48, Jan. 2023, doi: 10.69987/JACS.2023.30103.