mathematics

MDPI

Article

Information Sets and Learning Models for First Exits from Dynamic Price Zones: An Empirical Competing-Risk Study

Giwon Kim [1], Yoojeong Song [2,*] and Sang Ik Han [2,*]

[1] Department of Computer Software Engineering, Soonchunhyang University, 22 Soonchunhyang-ro, Asan-si 31538, Chungcheongnam-do, Republic of Korea; giwon9108@sch.ac.kr
[2] Department of Software Convergence, Soonchunhyang University, 22 Soonchunhyang-ro, Asan-si 31538, Chungcheongnam-do, Republic of Korea
* Correspondence: yoojeong@sch.ac.kr (Y.S.); sihan@sch.ac.kr (S.I.H.)

Abstract

We evaluate the information and model complexity needed to forecast first closing-price exits from pre-issued dynamic zones. Established competing-risk principles encode no exit within 60 min and first upper/lower exits in four intervals as nine outcomes. Annual historical analysis uses 502,443 DZ-LSTM-zone and 474,591 causal moving-quantile (MQ60) test origins for BTC, ETH, and XRP from 2020 through June 2026. Multinomial logit, XGBoost, survival forest, and an auxiliary MLP receive 22 non-zone predictors, 2 added distances, or all 32 inputs. Exact-calendar reconstruction repeats all 504 learner/information-set fits on unchanged historical targets and samples. Full inputs have lower Brier loss than non-zone inputs in 4/4 DZ-LSTM and 4/4 MQ60 learners, with signed relative reductions of 10.02–11.85% and 13.58–14.53%, respectively. Two distances recover most historical input gains. A locked follow-up applies stored 2026 MQ60 models and training scalers without tuning to 19,056 unused origins from 30 June through 30 September 2026; 4/4 learners have lower full-input Brier, and 4/4 reductions remain Holm-significant under all tested date-block choices. The contribution is empirical evidence about boundary information and learner choice. These retrospective checks do not establish a new output layer or universal neural superiority. Existing volatility benchmarks and negative external-market results delimit interpretation. Episode alerts remain optional illustrations under assumed policies and costs.

Keywords: dynamic price zones; competing risks; information sets; probability forecasting; cryptocurrency; early-warning evaluation

MSC: 35J25

Check for updates

Academic Editors: Anatoliy Swishchuk and Zengjing Chen

Received: 23 September 2026
Revised: 15 October 2026
Accepted: 26 October 2026
Published: 30 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

A forecast of departure from a price boundary requires information about the boundary itself. Similar returns and volatility can accompany different distances to an upper or lower boundary and therefore different short-horizon exit probabilities. The empirical question is how much boundary-relative information improves forecasts beyond ordinary market and clock information, and which learners use it most effectively.

We study the first closing-price exit from a two-sided, time-varying price zone issued before the forecast origin. At completion of each eligible 15 min bar, the outcome is no exit within 60 min or a first upper or lower exit in one of four subsequent intervals. A model-issued zone provides a forecast reference. Departure from this reference is not a

margin call, liquidation, position-limit violation, or an independently established measure of market instability. An alert may prompt closer monitoring or reassessment of a forecast; an institution-specific intervention requires boundaries and losses appropriate to that decision.

The statistical representation has established precedents. Discrete event-history and competing-risk models describe grouped event times with multiple causes [1,2]. Financial triple-barrier labels record which upper, lower, or time barrier is reached first [3]. DeepHit estimates a joint cause-time distribution with a neural softmax and constructs cumulative incidence from its masses [4]. The nine-state construction used here specifies this financial application and its observation rule. Its coherence is shared by any valid probability mass function on these states; it is not a new output layer or an empirical discovery.

The empirical contribution is to quantify the information and model complexity needed for this boundary-relative forecast. Distance from a boundary is an intuitive predictor; the unresolved empirical issue is how much two distances recover of the full zone representation, whether the remaining variables add useful accuracy, and which learner benefits under a common target and sample. A non-zone set contains 22 time, market, and completed intrahour-path variables. Adding two boundary distances provides a parsimonious reference; adding all ten zone-state variables supplies the full 32-input representation. Comparing these sets within each learner separates the value of the inputs from differences between learners. Comparing learners within a fixed set separates model choice from information availability. Together, these comparisons can identify a compact forecasting representation even when a neural model is not the most accurate learner. Conditional-volatility models provide additional benchmarks whose input mechanisms differ from those of the tabular learners.

The original evaluation reconstructs the upstream zone generator out of sample, supplies the downstream train, validation, and test partitions from that ledger, and freezes challenger selection before each test fold. The revision extends this evaluation with a four-learner information-set comparison, direct group removals, contact proper scores, subset uncertainty, episode-level alerts, and numerical GARCH sensitivity. These are retrospective revision analyses on a period already examined during development. They strengthen characterization of the observed results but are not independent prospective confirmation.

A separate robustness analysis replaces the original zones with an unfitted causal MQ60 rule, refits all 252 learner conditions on its new labels and risk set, and recalculates two GARCH comparators for every validation and test origin. Both generators use the full three-asset, 21-fold scope through the original June 2026 cutoff. This historical comparison broadens the assessed boundary constructions. A further exact-calendar reconstruction repeats all 504 information-set fits while preserving their targets and samples, and a separately locked follow-up tests fixed MQ60 models on 30 June–30 September 2026. The follow-up adds dates in the same three cryptocurrencies; it does not add assets or institutional boundaries.

The evidence shows an additional predictive value for zone information under the evaluated conditions, together with substantial gains from simple boundary distances and lower Brier loss from competing learners. Upstream generator quality, prediction conditional on a generator, and decision-policy performance are assessed separately. The new reconstruction and fixed-period follow-up assess whether these conclusions depend on retained-row time windows or the original observation period. Negative external-market results still delimit the conclusions. We do not assert that DZ-LSTM or MLP reference is necessary or universally preferable.

The primary output is a probability distribution over exit time, direction, and no exit. Supplying boundary information and comparing these forecasts is the main empirical task. Converting the probabilities into alerts is an optional downstream use, evaluated

separately under declared policy and cost assumptions. An illustrative alert evaluation does not establish that an institution should take an action or that the assumed costs represent its losses.

Section 2 positions the application and states its research questions. Section 3 specifies data, zones, information timing, and the risk set. Section 4 describes learners and evaluation. Section 5 reports the information-set comparison and supporting assessments. Sections 6 and 7 discuss the scope of the evidence and conclude.

2. Related Literature and Research Questions

2.1. Market Information and Probabilistic Forecasts

ARCH and GARCH describe changing conditional variance and volatility persistence [5,6]; GJR-GARCH and Student-t innovations accommodate leverage and heavy tails [7,8]. Cryptocurrency comparisons document the relevance of distributional and model choices for volatility forecasting [9–11]. Conditional variance measures movement magnitude, while quantile regression describes selected distributional points [12]. Neither summary alone determines the first direction and interval of an exit from a supplied boundary.

Neural sequence models and tree boosting allow nonlinear patterns in financial inputs [13,14]. Evidence from empirical asset pricing illustrates the usefulness of interactions [15] but does not establish which learner best forecasts this boundary-relative outcome. Quantile neural prediction also motivates an evaluated zone construction [16]; using that construction as a reference does not imply that it has superior downstream predictive value.

2.2. First Passage, Moving Barriers, and Financial Labels

First-passage analysis supplies relevant analytical precedents. Kunitomo and Ikeda study option pricing with curved boundaries, and Geman and Yor develop a probabilistic treatment of double-barrier valuation [17,18]. Time-dependent boundaries are therefore not outside the scope of existing analytical work. Such valuation results depend on specified processes and pricing objectives. Here, the objective is empirical prediction of observed close exits from a pre-issued path.

Triple-barrier labeling records the first encounter with an upper, lower, or time barrier [3]. Our target retains this first-boundary/no-event logic and records four exit intervals. The distinctions are the issued time-varying boundary, monitoring at completed closes, and the explicit finite-horizon probability distribution. These are application specifications rather than a new first-passage principle. Continuous crossings and trade execution remain outside the close-observation target.

2.3. Discrete Competing Risks and Joint Distributions

Allison discusses event times observed in discrete intervals; Tutz and Schmid treat discrete time-to-event models, including competing risks [1,2]. Cause-specific and subdistribution methods supply alternative descriptions of competing events [19,20]. DeepHit directly estimates a joint cause-time distribution, and Dynamic-DeepHit extends neural survival prediction to longitudinal information [4,21]. Competing-risk forests estimate coherent cumulative incidence through tree ensembles [22].

The present finite-horizon target has four fully observed future closes for every retained origin. No exit by 60 min is an observed outcome of this target; it does not imply that no exit occurs later. Missing future paths are excluded rather than coded as no exit. A multinomial representation, its likelihood, and cumulative sums are established constructions. Their role is to provide a common basis for empirical information and learner comparisons.

2.4. Forecast Accuracy and Early-Warning Evaluation

Brier loss and the logarithmic score assess probability forecasts [23,24]. Calibration, discrimination, and threshold performance answer different questions; bin-based calibration errors cannot replace proper scores [25,26]. Class prevalence also affects precision–recall summaries [27]. Forecast comparisons must accommodate dependent observations [28–30].

Early-warning usefulness depends on preferences over missed events and false alarms. Sarlin relates evaluation to policymaker losses, event frequencies, and observation weights [31]. This motivates an optional use-case evaluation rather than establishing operational value from probability accuracy alone. We apply common validation budgets, cooldowns, and causal daily limits to actual observed outcomes and fitted forecasts. The policy rules and relative costs are assumed; the event counts are computed from the data. Our episode-delay loss is a separate specification and neither Sarlin's measure nor an estimate of institutional losses. The 1:5:0.5 weights are illustrative assumptions for this task rather than a cost calibration drawn from this literature.

2.5. Research Questions

Existing theory supplies ways to define barrier outcomes and competing-event probabilities, but it does not determine the relative value of an information set or learner for the issued zones and observed market period. We therefore ask:

1. Does adding zone-state information improve first-exit probability forecasts beyond the specified 22-variable non-zone set?
2. How much of the observed improvement is obtained by adding only two boundary distances, and how do learners differ within the same information set?

Secondary questions concern sensitivity to group removal, observation rules, evaluation subsets, simulation settings, alert policies, and whether the information gains persist with exact-calendar features and in an unused follow-up period. Separate generator and external assessments delimit dependence on the chosen boundary. This study addresses these questions empirically; it does not claim the absence of prior integrated designs or a causal effect of zone variables on market prices.

3. Data and Research Design

3.1. Market Data and Evaluation Sample

The historical analyses and Table 1 use the preserved original data snapshot. The separate 30 June–30 September 2026 acquisition and evaluation are described in Section 4.10; these additional observations do not replace or enlarge the historical cohort.

We use Binance 15 min BTC/USDT, ETH/USDT, and XRP/USDT OHLCV bars. BTC and ETH begin at 04:00 UTC on 17 August 2017; XRP begins at 08:00 UTC on 4 May 2018. The raw series end at 06:45 UTC on 29 June 2026. The frozen evaluation applies a 29 June exclusive cutoff, so its 2026 fold is a partial year. Data before 2020 support initial generator fitting and the 2019 out-of-sample zone ledger. Table 1 describes the raw series; its counts are not risk-set counts.

Table 1. Frozen historical cryptocurrency series and return summaries.

Asset	Raw Bars	Return SD (bp)	Excess Kurtosis
BTC	310,295	39.77	71.04
ETH	310,295	49.04	51.95
XRP	285,495	56.03	133.06

Statistics use 15 min log returns over the raw-data periods. The primary common test risk set contains 502,443 origins.

3.2. Issued Reference Zones and Terminal Alignment

The zone generator acts as a preprocessing module that supplies pre-issued reference boundaries. The downstream learning model is a separate choice. Figure 1 distinguishes boundary-derived feature extraction from return-process simulation, while all forecasts are scored against the same first-exit outcomes within a generator.

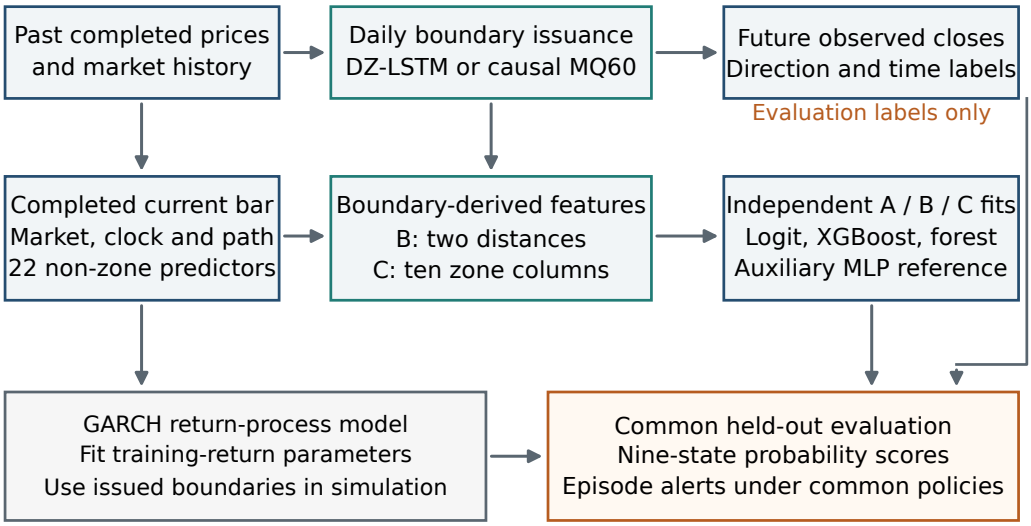


Figure 1. Study modules and comparison roles. The preprocessing module issues a boundary path and supplies boundary-relative predictors. Logit, XGBoost, forest and an auxiliary MLP are fitted independently with nested information sets. Ordinary GARCH models learn a return process and use the issued boundaries to map paths to exit probabilities. The added GARCH-X extensions also estimate A/B/C covariate coefficients in that process (Section 4.8). Future observed closes enter evaluation labels only.

The primary ledger uses an evaluated two-stage LSTM construction, denoted DZ-LSTM. It is retained to preserve the original conditional target and permit a matched information comparison; this choice is not an empirical selection of the best generator. At 00:00 UTC, completed information fixes the reference close and a daily lower, median, and upper path. A daily quantile LSTM uses the preceding 60 daily observations to forecast the 0.1, 0.5, and 0.9 quantiles of the current day's daily candle closing price. Two 64-unit LSTM layers and a 32-unit output layer are used; cumulative softplus enforces quantile ordering.

An intraday-path LSTM uses 13 same-clock market and path variables from the preceding 30 days. Its 48-unit representation combines with an eight-dimensional time-slot embedding and daily terminal quantiles to generate 23 nominal hourly knots from 01:00 to 23:00 UTC. The alignment weight at hour h is $h/23$, placing the terminal quantiles at the 23:00 knot. Pinball loss is minimized, and scaling and checkpoint selection use only the applicable training/validation history.

The daily target is the daily candle close at completion of the UTC day. The 23:00 knot is indexed by an hourly candle's opening stamp and is not the physical daily closing time. The implementation places the daily target forecast at that nominal knot; it is an alignment convention and can affect late-day event definitions. Linear interpolation supplies boundaries between the reference and hourly knots. The fully cross-fitted dataset excludes nominal rows after 23:00; it does not carry the final knot forward through 23:45. Consequently, a complete four-bar primary test path has its latest nominal origin at 22:00. These exclusions preserve the audited implementation, and conclusions apply to this retained intraday window.

Once issued, boundaries are fixed for the day. Future realized prices never redraw them. The 0.1 and 0.9 terminal quantiles do not establish pointwise 80% coverage at every interpolated knot or simultaneous 80% coverage of the whole path. They are neither institutional exposure limits nor verified liquidation levels.

Figure 2 illustrates the shape of the issued reference and its interpolation using the audited 2024 BTC/USDT out-of-sample ledger. This case explains the construction; it is not evidence about aggregate coverage or model accuracy.

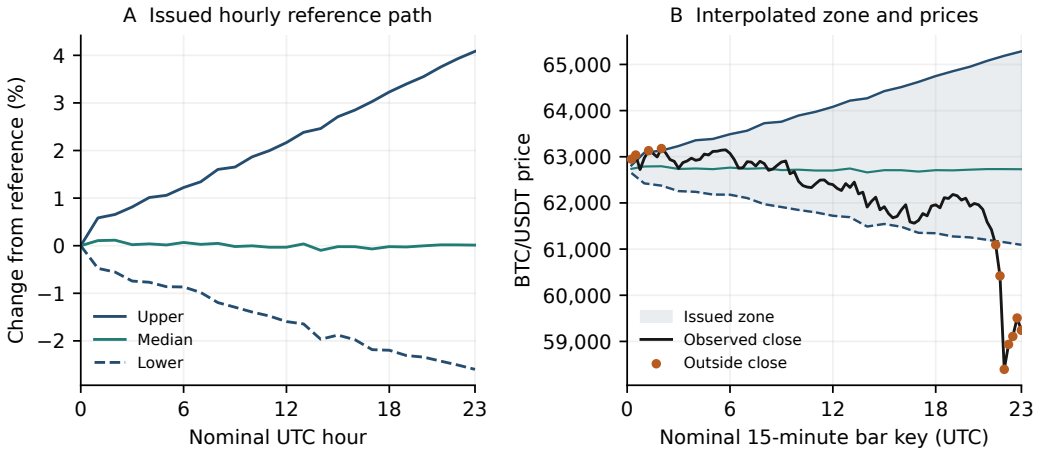


Figure 2. Pre-issued zone shape for BTC/USDT on 27 August 2024. (**A**) The 23 hourly lower, median and upper reference values relative to the fixed issuance reference; the daily terminal quantiles are aligned to the nominal 23:00 knot. (**B**) Their linear interpolation and the observed 15 min closes. Orange points are strict outside closes. The horizontal keys are nominal bar opening timestamps; their closes are observed 15 min later. The path is fixed at daily issuance. A nominal 80% terminal interval does not guarantee pointwise or simultaneous intraday coverage.

3.3. Observation Rule and Risk Set

Let a_t be the opening timestamp of the current 15 min bar and $d_t = a_t + 15$ min its completion time. The forecast can be issued at d_t, when its OHLCV values are available. Stored nominal keys use a_t. A boundary value attached to that key is a predetermined comparator for that bar; it is not a newly issued forecast at the opening timestamp.

An origin is eligible when the completed current close C_t lies within its issued lower and upper boundaries, $L_t \le C_t \le U_t$, and four exactly consecutive future bars within the same zone and chronological partition are available. For $j = 1, \ldots, 4$, an upper exit is the first strict inequality $C_{t+j} > U_{t+j}$ and a lower exit the first $C_{t+j} < L_{t+j}$. Only the earliest exit determines the label. Future close observation times are $d_t + 15j$ minutes. The outcomes are

$$\Omega = \{0, U_1, U_2, U_3, U_4, L_1, L_2, L_3, L_4\}, \tag{1}$$

where 0 denotes no close exit within 60 min. Already-outside origins and incomplete paths are excluded. The primary cohort has 502,443 common test origins across 21 asset-year folds. Its cumulative close-event rates are 2.4727%, 4.3925%, 6.0365%, and 7.4914% at 15, 30, 45, and 60 min. These rates differ from the earlier 523,575-origin supporting ledger.

Figure 3 illustrates the first-close-exit labeling rule.

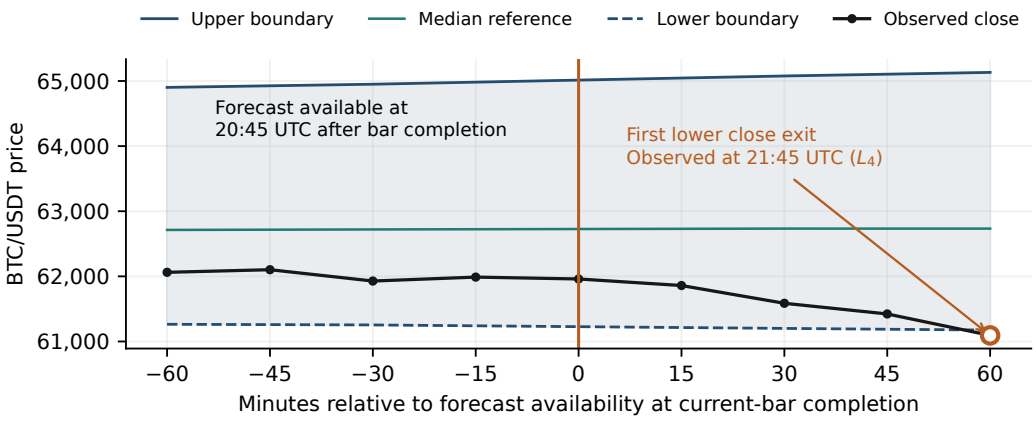


Figure 3. First-close-exit labeling in the same audited BTC/USDT day. The current nominal 20:30 bar completes at 20:45 UTC, when the forecast becomes available. The next four close observations occur at 15, 30, 45 and 60 min. Their first strict lower departure is observed at 21:45 UTC, giving state L_4. Predetermined boundaries at the corresponding nominal keys remain fixed; later realized prices do not enter current predictors. Intrabar contact is a different outcome.

3.4. Predictors and Information Sets

The audited input has four groups: time/calendar (9), market (8), completed intrahour path (5), and zone state (10). Time includes UTC clock and weekday cycles and time remaining in the current hour. Market inputs summarize lagged returns, return variability, and volume. Path inputs describe the completed portion of the current hour, not the cumulative day. Zone inputs express current price relative to the issued boundaries, their width, position, and current-bar breaches. Appendix A gives exact definitions.

The nested sets are A, all 22 non-zone variables; B, A plus upper and lower boundary distances (24); and C, all 32 variables. The original label "Market-only" denotes A in a neural model; it does not represent the eight market variables alone. Removing the zone group is identical to A. Other neural removals delete the time, market, or intrahour group and refit on the same risk set.

For the preprocessing comparison, the two distance features are $\Delta_U = (U_t - C_t)/C_t$ and $\Delta_L = (C_t - L_t)/C_t$. Set A withholds boundary predictors but uses the same boundary-defined outcomes as B and C. This matched target is necessary to estimate the value of supplying boundary state to a learner; it does not claim that the derived columns are independent new market risk factors. The upstream generator is therefore evaluated separately on observed-price zone quality, and the downstream feature effect is evaluated within each supplied generator.

The zone group is not ten independent information sources. Lower-ratio and lower-distance variables coincide, as do their upper counterparts. Width and normalized position are derived from the same boundaries and current price. The current-close breach flag is constant at zero in the inside-zone risk set; the wick flag records the completed current bar. Group improvements therefore assess a supplied representation rather than independent causal contributions of ten variables.

Feature reconstruction revealed a temporal naming limitation. Lag and rolling operations are performed after excluding the final three nominal bars of each day and count retained rows. Variables named 30 min, 60 min, one-hour, and four-hour are consequently not always fixed elapsed-time windows across day or data gaps. Earlier values, not future data, enter these operations. We preserve the original matrix for comparability and describe the row-based operations in Appendix A; no result is presented as a corrected calendar-window experiment. Scaling uses chronological training rows only.

3.5. Chronological Splits and Cross-Fitting

Cross-fitting separates the estimation of an upstream model from the observations on which its outputs are subsequently used. In the usual K-fold construction, an upstream estimator is fitted outside a held-out fold and evaluated on that fold; the held-out outputs are assembled across folds [32]. Using predictions for observations excluded from the first-stage fit also underlies second-stage learning in stacked generalization [33]. Here, the upstream output is a price-zone path that defines both predictors and the first-exit outcome rather than a prediction of the same response combined by a stacking ensemble. Because future observations cannot be used to construct historical forecasts, we use successive annual prediction blocks with expanding past histories, following the temporal ordering of rolling-origin out-of-sample evaluation [34].

Step 1: fit the upstream generator before each prediction block. For each asset and year $b = 2019, \ldots, 2026$, let a_b be 1 January at 00:00 UTC and $v_b = a_b - 60$ calendar days. The daily quantile LSTM and path LSTM form one generator bundle, $\widehat{G}_b$. Daily training examples have daily target timestamps before v_b; daily validation targets lie in $[v_b, a_b)$. For the path component, training requires the last active path-target timestamp to precede v_b, while validation requires an origin at or after v_b and the entire target path to end before a_b. Thus, partition eligibility depends on target completion, not just the input timestamp; the downstream 60 min purge is not substituted for these upstream target-end checks. Input and target scaling use the corresponding upstream training examples, and early stopping and checkpoint selection use only pre-block validation. The selected generator parameters are then fixed throughout the prediction year, which ends early at the June 2026 cutoff in the final block.

Step 2: construct the out-of-sample zone ledger. At each UTC-midnight issuance within block b, $\widehat{G}_b$ receives the completed history available at that issuance and outputs the day's lower, median, and upper path. Advancing the available input history within a block does not refit its generator. Only predictions for that held-out annual block enter the canonical ledger: outputs for the generator's own training or validation observations are excluded. The annual ledgers are concatenated, and each downstream row receives the entry associated with its issuance identifier; unavailable zones are excluded rather than replaced by fitted-history values. The same issued path supplies current zone predictors and the four future boundary values used for labeling. The ledger records issuance and target times, upstream fold, fitting cutoffs, and source identifiers. Audited entries satisfy

$$\begin{aligned} &\max(\text{upstream training target time}) < v_b, \\ &\max(\text{upstream validation target time}) < a_b \leq \text{zone issuance time}. \end{aligned}$$

The recorded prediction role is out of sample, with no fitted-history entry or source origin included in the generating fit.

Step 3: fit and evaluate the downstream exit model. Seven annual test blocks, 2020–2026, are evaluated for each asset. For test year Y, the preceding 60 calendar days provide validation and earlier eligible ledger rows provide training. A 60 min purge separates neighboring downstream partitions; an origin is retained only when all four future bars are continuous, use the same issued zone, and belong to the same partition. Downstream scaling uses training rows, while checkpoint selection, original challenger selection, and alert thresholds use validation. Training, validation, and test rows all use their own historical out-of-sample zones. For example, the 2020 downstream training and validation sets use eligible 2019 ledger entries generated by $\widehat{G}_{2019}$, whereas the 2020 test set uses entries from $\widehat{G}_{2020}$. The latter is fitted and selected before 2020 and is not used to regenerate fitted zones for the 2019 downstream history. Completed earlier test years can enter the expanding training history of later folds.

In this study, "fully cross-fitted" therefore means that every supplied downstream zone is out of sample relative to the complete upstream generator bundle. It does not mean that the daily anchors used internally to train the path component are themselves generated by an additional nested cross-fit. Nor is this an implementation of double/debiased machine learning: that framework combines cross-fitting with Neyman-orthogonal scores under stated assumptions [32], whereas our downstream objective is predictive likelihood. We do not import its causal or bias-removal guarantees. The present procedure excludes in-sample upstream outputs and future fitting/selection information from the downstream ledger; serial dependence and systematic generator error can remain. A generator that consistently issues narrow zones, for example, can produce a predictable pattern of exits that the downstream learner learns even out of sample. The estimand remains conditional on the issued boundaries. The separately refitted MQ60 experiment in Sections 4.7 and 5.9 assesses this dependence under one alternative boundary rule rather than proving its removal.

The original input manifest, keys, labels, annual cutoffs, and common model origins were verified. The original 32 features were reconstructed for 737,472 dataset rows, with float32 differences at most 2.33×10^{-10}. Chronological splitting does not make the previously studied market period prospective data, and the uncertainty assessed below remains conditional on the fitted models and observed procedure. Figure 4 summarizes the chronological training, validation, and test construction.

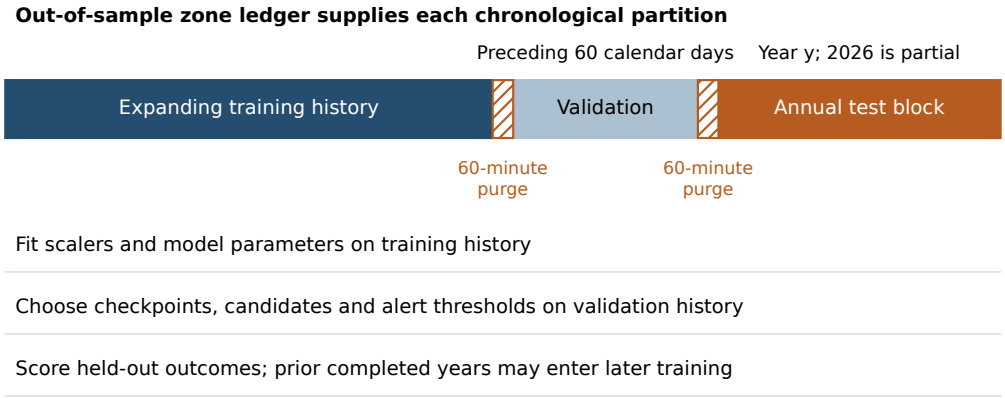


Figure 4. Chronological training, validation and test partitions. Each asset uses seven annual test blocks from 2020 to the June 2026 cutoff. Validation occupies the preceding 60 days, with 60 min purges at neighboring boundaries. Every downstream partition uses its own out-of-sample zone-ledger entries. Checkpoints, original challenger selection and alert thresholds are determined without current-fold test labels. Revision analyses are retrospective on previously studied periods.

4. Methods

4.1. Joint PMF and the Auxiliary Neural Reference

For input x_t, any learner producing nonnegative masses summing to one on Equation (1) defines a common forecast. In the neural specification, logits $z_c(x_t)$ yield

$$p_c(x_t) = \frac{\exp z_c(x_t)}{\sum_{c' \in \Omega} \exp z_{c'}(x_t)}, \qquad c \in \Omega. \tag{2}$$

The unweighted multinomial negative log-likelihood is

$$\mathcal{L}(\theta) = -\frac{1}{N} \sum_{i=1}^{N} \log p_{y_i}(x_i; \theta). \tag{3}$$

The multilayer perceptron (MLP) reference, called DZ-CRPMF in the original submission, has 96- and 64-unit ReLU hidden layers with dropout 0.10 and 0.08, followed by a 48-unit

ReLU head and nine logits; the 32-input version has 12,937 parameters. AdamW uses learning rate 0.001, weight decay 0.0001, batches of 512, and at most 80 epochs. Validation NLL early stopping has patience 10 and minimum improvement 10^{-5}; the best checkpoint is restored. Reduced-input conditions retain this procedure and are trained anew rather than masking inputs after fitting.

For $h = 1, \ldots, 4$, cumulative incidence and finite-horizon survival are

$$F_k(h \mid x_t) = \sum_{j=1}^{h} p_{k,j}(x_t), \quad k \in \{U, L\}, \tag{4}$$

$$F_{\text{any}}(h \mid x_t) = F_U(h \mid x_t) + F_L(h \mid x_t), \quad S(h \mid x_t) = 1 - F_{\text{any}}(h \mid x_t). \tag{5}$$

Nonnegative masses imply nondecreasing cumulative incidence and directional additivity. These are algebraic properties, also shared by coherent logit, XGBoost, forest, and simulated-path forecasts. Equations (2)–(5) implement established discrete competing-risk principles [2,4]; they are not assessed as a unique neural performance contribution.

4.2. Learners, Selection, and Comparison Roles

The revision factorial comparison crosses multinomial logit, nine-state XGBoost, a competing-risk forest, and an auxiliary MLP reference with A/B/C on the original 21 folds: 252 fits. Three additional auxiliary neural group removals contribute 63 fits for 315 independently fitted conditions. All information conditions, including A and C references, use the current runtime and original matrices. This avoids attributing software-version differences to added information. Frozen original scores are reported separately.

Multinomial logit uses L2 regularization, with $C \in \{0.1, 1, 10\}$ chosen from pre-2020 validation NLL for each asset/set and then frozen. XGBoost uses the original 12-candidate grid and pre-2020 NLL/complexity rule, followed by annual validation NLL early stopping. The forest uses native randomForestSRC logrank-CR splitting, 300 trees, the original 50,000-row training sample cap, and 4 candidate settings. Its cause-time CIFs are converted to interval masses with an explicit no-exit complement. Tabular learner inputs, target, test origins, and split rules are matched; training objectives and model complexity are not identical. Configuration and seed details are given in Appendix B and the Supplementary Records.

The frozen original comparison evaluates Full (C neural), Non-zone NN (A neural), DeepHit, XGBoost, competing-risk forest, Gaussian GARCH, and Student-t GJR-GARCH on the common primary cohort. The strongest-challenger pool contains *DeepHit, XGBoost, survival forest, and Student-t GJR-GARCH*. XGBoost and the forest were included, not excluded. In each fold, minimum validation nine-state Brier selects the candidate; ties within 10^{-12} use validation NLL and then lexicographic model name. The verified selections are 11 XGBoost, 7 GJR, and 3 forest folds. This selected composite differs from an individual model's pooled score. Gaussian GARCH was a secondary comparator outside that frozen pool, which is not expanded or reselected using test results.

Both GARCH models are fitted to training 15 min log returns with zero conditional mean. Gaussian GARCH(1,1) and Student-t GJR-GARCH(1,1) retain admissible converged fits with finite likelihood, nonnegative coefficients, and persistence below 0.999. Parameters remain fixed while conditional variance updates causally through the origin. Future closing-price paths are compared with four issued boundary values, and first-event frequencies form the nine-state PMF. For M paths, state c receives $(n_c + 0.5)/(M + 4.5)$. The original setting uses 2048 scrambled Sobol paths.

These are estimated statistical learning models. With zero conditional mean, the GJR specification used by the simulation is

$$r_t = \sigma_t z_t, \qquad \sigma_t^2 = \omega + \alpha r_{t-1}^2 + \gamma r_{t-1}^2 \mathbf{1}\{r_{t-1} < 0\} + \beta \sigma_{t-1}^2. \tag{6}$$

Gaussian GARCH sets $\gamma = 0$ and uses standard-normal innovations; GJR uses standardized Student-t innovations. Training-return maximum likelihood estimates the process coefficients and, for Student-t, the degrees of freedom. These fitted coefficients do not include the supplied zone-state columns. The known future boundary path enters the subsequent first-exit simulation. Consequently, this experiment compares final forecast accuracy, while the A/B/C fits compare additional feature information. These ordinary GARCH comparisons remain separate from the new covariate-trained extensions in Section 4.8.

An independent multi-horizon sigmoid is only an auxiliary comparator. Unlike the PMF neural risk-set fit, its original training used all chronological rows, positive-weighted binary cross-entropy, and direction, continuation, and consistency objectives. Any-event positive weights were clipped to 1–8 and directional weights to 1–12. Thus, even a common evaluation sample does not isolate output-layer coherence. The earlier supporting mean any-exit ECE of 0.0948 indicates a probability-calibration problem; being above event prevalence is not a mathematical impossibility. The reported 72.8% supporting Brier gap is not used as evidence that coherence alone improves forecasts.

4.3. Proper Scores and Paired Uncertainty

The principal score is the nine-state Brier loss,

$$\mathrm{BS} = \frac{1}{N} \sum_{i} \sum_{c \in \Omega} (p_{ic} - \mathbf{1}\{y_i = c\})^2. \tag{7}$$

Unclipped log loss is $-N^{-1} \sum_i \log p_{i,y_i}$. Assigning zero probability to an observed state makes it infinite. A historical 10^{-300} floor is retained only as an explicitly floored sensitivity, not as the unmodified proper logarithmic score. Binary any/upper/lower horizon scores follow from the CIFs.

Paired differences use common test origins. Stationary bootstrap resamples UTC calendar days synchronized across assets, using expected blocks of 3, 7, and 14 days and 20,000 replicates per block [30]. Seven days is the reporting reference. Ninety-fifth percentile intervals use uncentered differences; two-sided tests use null-centered differences. Holm adjustments are applied within declared comparison families [35]. The four C–A tests form one family; the eight B–A/C–B contrasts another; six learner pairs form a family within each set; four neural group removals form another. Supplementary Files include all blocks and the analogous log-loss families. Of 40 planned log-loss contrasts, 32 are finite, and 8 involving observed zero mass are undefined. Such rows are not dropped or clipped for significance tests; the planned family size is retained with undefined members contributing no rejection.

The frozen original Full–Non-zone and Full–selected-challenger tests remain a separate two-test family with their original decision criteria (1% relative Brier improvement, negative upper interval bound, Holm $p < 0.05$). Revision families are retrospective diagnostics, not new confirmatory hypotheses. Inference conditions on the fitted models, observed generator and cohort; it omits fitting uncertainty and is not evidence of generator-independent or prospective effects.

4.4. Contact, Subset, and Simulation Sensitivity

Contact sensitivity holds close-trained probabilities fixed and labels the first future high touching/exceeding the upper boundary or low touching/falling below the lower boundary. Any-contact scores retain all 502,443 origins. If the first contact bar touches both sides, OHLC does not identify their order; 63 such origins are excluded from directional nine-state scores, leaving 502,380. This is not contact-target retraining.

Leave-one-asset and leave-one-year analyses omit each of three assets or seven years from evaluation while holding the original fitted models fixed. Both original paired effects

are estimated with the same synchronized bootstrap. Brier and log-loss omission families are separated by metric and omission type (6 asset and 14 year tests). These are subset diagnostics, not leave-out retraining or independent replications.

GARCH sensitivity fixes parameters, causally filtered variances, boundaries, labels, and test origins. Each of 42 distribution/fold combinations first reproduces its original 2048-path PMF to maximum difference 2×10^{-7}. Counts $512, 1024, 2048, 4096, 8192$ and scrambling seeds 42–46 then yield 1050 fold cells. Nested prefixes within a seed support numerical comparisons. The 0.5 pseudocount per state remains fixed, so its relative weight $4.5/(M + 4.5)$ changes with M. Seed ranges describe numerical variability, not confidence intervals or parameter uncertainty. The original 11/7/3 challenger choices are kept fixed when replacing only the selected GJR fold probabilities.

4.5. Illustrative Episode-Alert Use Case

These optional alert scenarios retain their originally frozen forecasts. The exact-calendar refits and unused-period probability check in Sections 4.9 and 4.10 are separate; no alert rule is selected on their test outcomes. Refitting corrected predictors can change probabilities and proper scores, whereas changing only a post-processing policy leaves fixed PMFs unchanged.

The learned forecast and the alert rule are separate objects. The neural model is fitted by the unweighted likelihood in Equation (3); the alert-cost coefficients do not enter that fit. A threshold converts the 60 min any-exit probability into a candidate warning, while cooldown and daily capacity determine which candidates are issued. Keeping the forecast PMFs fixed while changing these rules changes alerts but not the Brier or log-loss scores on the same evaluation origins. Larger alert volume alone is not an optimum: additional detections can accompany additional false or repeated alerts. The appropriate trade-off depends on the declared use and constraints. These policies illustrate one possible application of the probability forecasts rather than a necessary component of the forecasting model. An episode starts at an outside close immediately following an inside close on contiguous bars within the same issued zone. A return inside permits a later episode; a new issuance or gap breaks continuity. An eligible episode has at least one retained forecast origin strictly before onset and within 60 min. There are 12,580 eligible onsets. Each alert is matched to the nearest strictly future onset within 60 min; alerts at or after onset are not advance warnings. An episode receives one detection credit, using its earliest warning. Later matching alerts are recorded as repeats.

For each fold/model, the validation distribution of 60 min any-exit probabilities determines a threshold at budget fractions 0.5%, 1%, 2.5%, 5%, or 10%; strict exceedance is fixed on test. Cooldowns are 0, 60, or 120 min, and daily caps are absent, one, or three alerts per asset. Both controls are applied chronologically without future test ranking. These are common validation budgets and maximum capacities, not hindsight rules that force identical realized test alert totals. All seven coherent models and the auxiliary sigmoid are compared, giving 7560 fold policy and 360 pooled rows.

The illustrative 1:5:0.5 specification assigns relative weights to false alerts, missed episodes, and reduced warning lead. These weights were assumed for the use-case evaluation and were not estimated from institutional cost data. The delay term penalizes a shorter advance warning; alerts at or after onset do not count as detections. For a fixed alert policy, changing these weights changes its loss score without changing its warnings. Using weights to select among policies can change the selected threshold, cooldown, or capacity

but does not identify an economically optimal weighting from forecast outcomes alone. The assumed loss is

$$L = c_{\mathrm{FA}} N_{\mathrm{FA}} + c_{\mathrm{miss}} N_{\mathrm{miss}} + c_{\mathrm{delay}} \sum_{e \in \mathcal{D}} (1 - \ell_e/60), \tag{8}$$

where ℓ_e is the first-alert lead in minutes and $\mathcal{D}$ the detected episodes. Let D_m be the detected episode count, F_m the false-alert count, $Q_m = \sum_{e \in \mathcal{D}_m} (1 - \ell_e/60)$ the observed delay units, and E the common eligible episode count for a fixed model-policy comparison. For $c_{\mathrm{FA}} > 0$, define $r = c_{\mathrm{miss}}/c_{\mathrm{FA}}$ and $s = c_{\mathrm{delay}}/c_{\mathrm{FA}}$. Relative to never alert, the normalized cost reduction is

$$S_m(r, s) = rD_m - F_m - sQ_m.$$

Consequently, a model with $D_m > 0$ improves on never alert exactly when $r > F_m/D_m + (Q_m/D_m)s$. For two models evaluated on the same episode set and policy rule, model m has a lower observed cost than n exactly when

$$r(D_m - D_n) > (F_m - F_n) + s(Q_m - Q_n).$$

These affine inequalities describe the entire nonnegative cost-ratio domain rather than requiring a selected representative triplet. They express the trade-off implied by the observed alerts and leads. Costs must be supplied by a decision-maker; prediction performance alone cannot identify them. This follows the preference-dependent evaluation motivation of Sarlin [31], while our episode and delay formula is a separate specification.

The 1:5:0.5 triplet serves only as an assumed scenario. Tables reported in the results give empirical break-even boundaries, and an appendix figure shows descriptive lowest-cost regions under the same fixed policy rule, including never alert. The display covers the upper endpoints of the previously assessed cost ranges; the supplied inequalities apply beyond it. Every one of the 360 original and 270 MQ60 model-policy combinations has a supplied break-even formula and pairwise comparisons. Zero detections cannot yield a strictly positive reduction for nonnegative costs.

Supplementary discrete sensitivities use the following assumed weights: original-zone missed-episode weights are 2, 5, 10, and 20, while MQ60 weights are 1, 2, 5, and 10; both use delay weights 0, 0.5, 1, and 2 and unit false-alert cost. These grids contribute 5760 and 4320 cost rows, respectively. We verify all 10,080 rows against the affine cost identity. They are scenarios, not estimates of institutional loss. The analytic regions are retrospective descriptions of fixed test counts; they do not select an optimal model, threshold, cooldown, capacity, or cost ratio for deployment. No utility confidence interval is supplied. Detected repeats have no separate monetary penalty, and intervention or true-alert processing costs are absent from this formula. Such costs would alter the boundaries. Economic profitability or realized savings is not inferred. Appendix E reports a separate exploratory check in which assumed penalty weights propose policies and a fixed cost-independent detection/burden criterion selects among them. It preserves the fitted probabilities and does not supply an institutional cost estimate.

4.6. Generator and External Assessments

Changing a zone changes the target, prevalence, and inside-zone risk set. Existing DZ-LSTM (DZL), same-clock moving-quantile (STQ), and volatility-band (VB) comparisons therefore report each generator's own training-climatology Brier skill, not a same-target ranking of raw losses. STQ/VB settings are selected before 2020; same-slot rolling returns are shifted by one daily occurrence. Importantly, these are earlier supporting results from a different ledger and are not the 502,443-origin fully cross-fitted factorial experiment. That provenance prevents treating them as proof of generator-independent information gains.

External analyses adapt zones to New York trading sessions and daylight-saving changes for EUR/USD, USD/JPY, XAU/USD, and USA500 (an index-CFD quote series). Synthetic bars do not fill closures; evaluation requires a continuous future 60 min interval. The end-to-end assessment re-estimates a session-aware DZ-LSTM construction and the downstream model within each market and compares zone WIS with session-matched moving quantiles (MQ). This is distinct from the earlier application of a fixed cryptocurrency predictor to MQ zones. Positive within-market refitted skill is not proof that the fixed cryptocurrency predictor transfers. The original upstream 2% WIS non-inferiority criterion and negative results are preserved.

4.7. Causal MQ60 Replication and Matched Volatility Comparators

The added reference issues its entire day path at UTC midnight from the preceding 60 calendar days of same-clock cumulative log returns. The reference price is the immediately preceding completed 23:45 nominal bar. Slot-specific 0.1, 0.5, and 0.9 quantiles require at least 30 finite historical observations; missing days remain in the calendar window. No current-day or future return enters the zone. There is no estimated generator model or test-selected window. MQ30 and MQ120 are upstream-quality sensitivities only.

The 22 original non-zone columns are preserved, while 10 zone columns and all first-exit labels are reconstructed. The annual splits, 60-day validation period, 60 min purge, original tuning grids, and training-only scaling are retained. A total of 4 learners by 3 information sets by 21 folds produce 252 independent refits, with 474,591 common test origins. This is the complete existing fold scope, not a capped sample or an extension beyond the June 2026 cutoff. Future-deletion checks and independent label reconstruction verify the causal implementation. Changing the generator changes the inside-zone risk set and event labels; raw downstream Brier scores across generators are not same-target rankings.

Gaussian GARCH and Student-t GJR-GARCH are re-evaluated against the four known MQ60 future boundaries on exactly these validation and test indices. Their train-only return-process parameters can be reused because the close history and chronological training split are unchanged by replacing the zone. We verify identical timestamps and closes, reproduce the fixed-parameter training likelihood, reconstruct causal variances through each origin, and reproduce 256 original test PMFs per fold within 2×10^{-7}. Each new PMF uses 2048 Sobol paths, the frozen fold seed, and pseudocount 0.5 per state. A 64-origin batch-size check is bitwise identical. Old DZ-LSTM exit probabilities and old alert thresholds are never reused. The matched analysis comprises 42 GARCH fold/distribution calculations. GARCH is a structural comparator with a different information mechanism, not an A/B/C feature-removal condition.

The added GARCH comparisons form eight paired contrasts per score (two GARCH specifications against four full-input learners). UTC-day-synchronized stationary bootstrap uses 20,000 replicates and expected blocks of 3, 7, and 14 days, with Holm correction within each eight-test score family. Uncertainty is conditional on the fitted learners and frozen return-process estimates. All six comparators use the same validation budgets, cooldowns, daily caps, and episode rule as the original analysis. The added 270 pooled policies and 4,320 cost rows use missed-episode costs 1/2/5/10 and delay costs 0/0.5/1/2; these are explicitly illustrative. The original cost grid is kept separately.

All 63 saved MQ60 neural models are additionally evaluated on every validation and test row using unchanged float32 forward logits and float64 log-softmax. Probability reconstruction against stored PMFs must agree within 5×10^{-7}. Stable-logit scores are disclosed separately; stored probabilities, fits, Brier tables, and prior inference are not silently replaced.

4.8. Covariate-Trained GARCH-X Extensions

We additionally fit Gaussian GARCH-X and Student-t GJR-GARCH-X with each A/B/C covariate set on both the original and MQ60 references: 2 distributions by 3 sets by 21 folds by 2 generators, or 252 new conditions. Signed covariates enter a positive log-intercept link, distinct from an additive specification restricted to nonnegative covariates. For percentage log return r_t and conditional variance h_t, the fitted recursion is

$$q_{j,t} = 8\tanh\left(\frac{x_{j,t} - \mu_j}{8s_j}\right),$$
$$b_t = \exp\left(c + 8\tanh\left(\frac{\delta^\top q_t}{8}\right)\right),$$
$$h_{t+1} = b_t + \alpha r_t^2 + \gamma r_t^2 \mathbf{1}\{r_t < 0\} + \beta h_t.$$

Gaussian innovations set $\gamma = 0$; standardized Student-t innovations estimate degrees of freedom $\nu \in [2.05, 100]$. A softmax parameterization keeps $\alpha + \gamma/2 + \beta < 0.999$. Means μ_j, scales s_j, and constant-column decisions use chronological training rows only. Nonconstant duplicate columns are retained, with a fixed penalty $0.001\|\delta\|^2/2$ on the mean training-return negative log likelihood. The process coefficients and covariate coefficients are jointly refitted rather than reusing ordinary-GARCH exit probabilities. Optimizer success, projected gradient at most 10^{-4}, and an objective no worse than the best training-only initialization are required. All 252 fits pass; finite-difference derivatives, a null-GARCH likelihood comparison, saved-model probability reconstruction, and future-data perturbation checks are supplied.

Fitting uses finite training returns and the preceding completed retained-row covariates; it differs from the nine-state learners' inside-zone outcome likelihood. The existing row-based history and returns across retained-row gaps are preserved, not described as a corrected calendar-time experiment. Through a forecast origin, variance is updated with observed history. During the four-step simulation, its origin covariate vector is held fixed: observed future covariates are never used. The issued boundary path, 2048 scrambled Sobol paths, original fold seed, and 0.5 category pseudocount are common mapping choices. Validation/test keys, outcomes and risk sets match the corresponding generator's existing experiment.

Every GARCH-X condition, including A, still uses the issued boundaries to translate price paths into first-exit probabilities. Its C–A contrast therefore measures the effect of additional *learned* zone covariates beyond structural boundary use; it is not an experiment that removes all boundary information from the forecasting procedure. These are retrospective extensions specified after earlier study results, not prospective confirmation. Three-, seven-, and 14-day synchronized date bootstrap uses 20,000 replicates. Per generator and score, C–A has a two-distribution Holm family and B–A/C–B a four-comparison family, separate from the original learner families. Ordinary-GARCH challenger selection and episode-policy analyses remain unchanged; no new GARCH-X policy or institutional-utility claims are made.

4.9. Exact-Calendar Reconstruction and Matched Refits

The legacy ledger excludes nominal times after 23:00 before constructing market histories. Consequently, retained-row lags may span more than their named elapsed times, and its previous-hour volume pool includes partially retained hours. We reconstruct predictors on the complete raw UTC 15 min grid before applying any zone or risk-set restriction. A missing bar remains missing. Returns use exact 15, 30, and 60 min endpoints; volatility uses 4 or 16 consecutive 15 min log returns with sample standard deviation. The volume z-score uses sixteen complete consecutive bar volumes. Intrahour open, extrema, and cumulative

volume use only the prefix observed by current-bar completion. The volume denominator is the median of complete four-bar totals in the preceding 168 calendar hours, excluding the current hour, with at least 24 complete hours. Later-bar perturbation, exact-lag, standard-deviation, and missing-bar fixtures verify the implementation.

The nine nominal-clock variables and ten supplied zone variables are unchanged. All 245,824 historical candidate rows per asset have finite corrected predictors; no test origin is lost. Original first-exit labels, boundary ledgers, 502,443 DZL and 474,591 MQ60 test origins, annual folds, and chronological purges are preserved. Four learners by three information sets by three assets by seven years by two generators give five hundred and four fresh fits. Existing pre-2020 hyperparameter choices, objectives, architecture, seeds, and validation stopping rules are reused. Scaling is recomputed from chronological training rows only. This is a retrospective implementation check on previously studied data; it is not prospective preregistration. Existing fitted models and their displayed legacy results remain intact.

Primary C–A comparisons use synchronized UTC-date stationary bootstrap with expected blocks of 3, 7, and 14 days and 20,000 replicates. Holm correction covers four learners separately by generator and score. B–A/C–B form a separate eight-comparison family. Corrected-minus-legacy comparisons form a separate twelve-comparison family, paired on identical origins. Bootstrap inference conditions on fits and choices. Published log loss is unmodified; an observed zero probability yields infinity. Nonfinite log-score contrasts receive no finite inferential claim, and planned multiplicity is retained for the finite contrasts.

4.10. Locked Unused-Period MQ60 Follow-Up

Before acquiring the extension OHLCV or scoring it, the protocol fixes 30 June 2026 through 30 September 2026, the three original assets, the MQ60 rule, all A/B/C model specifications and training rules, and the evaluation procedure. We subsequently fit weights solely on pre-2026 training/validation data and seal them before extension scoring. Public Binance 15 min data are retrieved into separate files. All 9024 bars per asset from the acquisition interval beginning 29 June are present; 28 overlapping legacy bars per asset agree on OHLCV. No legacy CSV is replaced. The evaluated period begins on 30 June, beyond the original data snapshot and its 29 June exclusive evaluation cutoff.

MQ60 keeps its fixed preceding-60-calendar-day same-clock quantiles and preceding completed 23:45 reference close. Incoming earlier days can update its declared rolling history; future target days are excluded. Exact-calendar feature construction is unchanged. Nominal origins retain the original 00:15–23:00 support, with complete labels requiring a latest 22:00 origin. The current close must be inside, and all four future closes must be consecutive within one daily issue. There are 19,056 common eligible origins: 6392 BTC, 6494 ETH, and 6170 XRP. Generator future-deletion checks verify daily issuance causality.

The 36 corrected MQ60 models for the 2026 fold use training data ending before the pre-2026 validation interval and checkpoints chosen only on that validation interval. Model and scaler hashes are sealed before follow-up scoring. Replay predictions on 512 fixed historical test origins per model must agree with the saved probabilities within 10^{-6}. We then apply those stored scalers and models without retraining, tuning, or selecting a learner, information set, threshold, or favorable date slice on the extension. All twelve learner/information-set combinations are reported. Paired inference uses the same 3/7/14-day procedure and planned Holm families. This is a retrospective follow-up on an unused period, not publicly preregistered live forecasting; it does not evaluate DZ-LSTM generation on the extension or repair its external-market failures.

OpenAI Codex assisted with drafting and executing the analysis scripts and preparing tables and figures from recorded market data. Computational consistency checks compared input and output hashes, chronological partitions, target labels, probability masses, and independently calculated Brier scores. These checks assess the recorded computations; they do not provide external validation of the research conclusions.

5. Results

Sections 5.1–5.9 retain the earlier analyses and their declared feature definitions for traceability. The new matched reconstruction and unused-period results are presented separately in Sections 5.10 and 5.11.

5.1. Information-Set and Learner Comparison

Table 2 reports the four nine-state learners and two GARCH-X extensions on the same 502,443 test origins after independent refitting of A/B/C in the current runtime. C reduces Brier loss relative to A by 9.95% for logit, 10.07% for MLP reference, 11.35% for XGBoost, and 11.85% for the forest. All four seven-day paired intervals exclude zero and Holm-adjusted p values are approximately 0.0002 (Table 3); their direction and rejection persist with 3- and 14-day blocks. This supports an additional predictive value for the supplied zone group within these fitted conditions. It does not identify the independent value of each group member.

Table 2. Nine-state Brier scores by input set, including covariate-trained GARCH-X extensions.

Learner	A: 22	B: 24	C: 32	A–C Reduction
Multinomial logit	0.141970	0.128263	0.127848	9.95%
XGBoost	0.140044	0.124600	0.124148	11.35%
Survival forest	0.140415	0.124543	0.123770	11.85%
MLP reference	0.140313	0.128018	0.126188	10.07%
Gaussian GARCH-X	0.126147	0.125755	0.125563	0.46%
Student-t GJR-GARCH-X	0.122336	0.122331	0.122346	−0.01%

Lower loss is better. A contains all non-zone inputs; B adds two distances; C includes the full zone group. All conditions use 21 folds and 502,443 common origins. For the two GARCH-X rows, A/B/C denote fitted covariates in the return-variance equation. All three conditions also use the issued boundaries in path-to-exit mapping; objectives and training populations differ from the nine-state learners. Negative percentage reduction denotes deterioration. See Section 4.8. The MLP row is an auxiliary reference. These are revision refits, not replacements for frozen original scores; the separate frozen GARCH accuracy comparison is reported later.

The added Gaussian GARCH-X has C–A Brier difference −0.000584 (0.46% relative reduction), with seven-day interval [−0.000787, −0.000345] and Holm $p = 0.000100$. For Student-t GJR-GARCH-X, the difference is 0.000011, interval [−0.000124, 0.000167], $p = 0.884756$: no Brier improvement is established. Its log loss increases from 0.296239 to 0.297429; the seven-day difference is 0.001190, interval [0.000606, 0.001854], Holm $p = 0.000500$, with rejection also at 3 and 14 days. Gaussian C log loss is likewise not lower than A, but its difference is not significant. These findings make the learned zone-covariate effect model- and score-dependent.

The MLP is an auxiliary architecture reference rather than the claimed preprocessing contribution. With C, forest and XGBoost have lower Brier scores than MLP reference (0.123770 and 0.124148 versus 0.126188). MLP reference minus XGBoost is 0.002040 [0.001659, 0.002442], and minus forest is 0.002418 [0.001809, 0.003078]; each seven-day Holm p is approximately 0.0003. These results concern Brier loss. The full-input forest assigns exactly zero to the

observed state at one test origin, making its unmodified log loss infinite (Table 4). A low Brier score therefore does not establish superiority under every proper score.

Table 3. Paired C minus A Brier differences; seven-day stationary bootstrap.

Learner	Difference	95% Interval	Holm p
Multinomial logit	−0.014121	[−0.015980, −0.012131]	0.000200
XGBoost	−0.015896	[−0.017267, −0.014575]	0.000200
Survival forest	−0.016645	[−0.018034, −0.015293]	0.000200
MLP reference	−0.014126	[−0.015405, −0.012919]	0.000200
Gaussian GARCH-X	−0.000584	[−0.000787, −0.000345]	0.000100
Student-t GJR-GARCH-X	0.000011	[−0.000124, 0.000167]	0.884756

Negative differences favor C. The four nine-state learner comparisons form one family; the two GARCH-X comparisons form a separate family. Each uses 20,000 replicates. Inference is conditional on fitted models.

Table 4. Unclipped nine-state log loss for the information-set comparison.

Learner	A: 22	B: 24	C: 32
Multinomial logit	0.410003	0.320824	0.321622
XGBoost	0.398420	0.309303	0.308755
Survival forest	0.399030	0.303385	∞
MLP reference	0.400202	0.322893	0.317551
Gaussian GARCH-X	0.306173	0.305802	0.306306
Student-t GJR-GARCH-X	0.296239	0.296557	0.297429

The GARCH-X rows retain unclipped scores, including the deterioration under Student-t innovations. The forest C score is infinite because one observed outcome has zero predicted mass. No test row is removed to obtain a finite score. Floored sensitivities and finite/undefined inference flags are supplied separately.

Figure 5 compares the four nine-state learners across the three information sets.

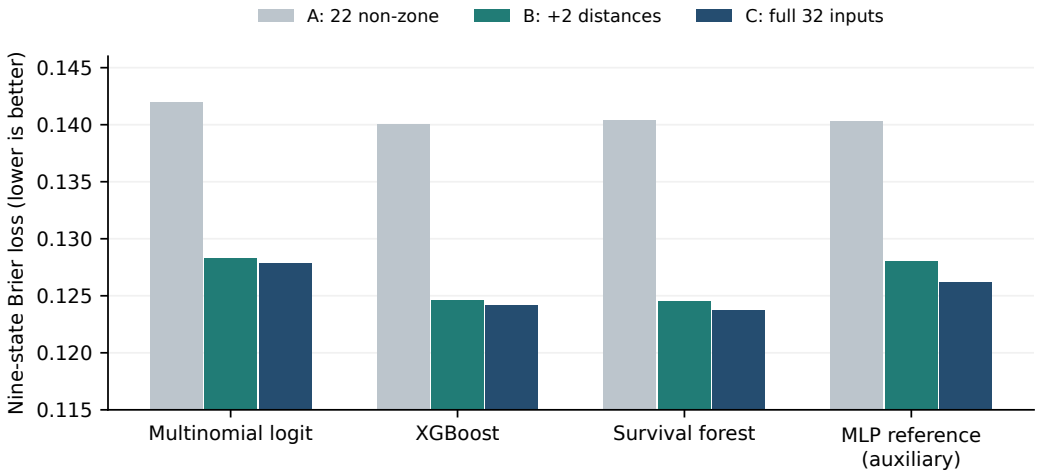


Figure 5. Brier scores for the four nine-state learners by information set on the common primary cohort. Information gain within a learner and final loss across learners are separate comparisons.

5.2. Boundary Distance and Direct Group Removals

The B–A improvement is approximately 97.07%, 87.04%, 97.16%, and 95.35% of the C–A improvement for logit, MLP reference, XGBoost, and forest, respectively. This is a descriptive ratio of score reductions across refitted conditions, not a causal decomposition

or fraction of explained variance. Thus, two simple boundary distances obtain most of the observed Brier reduction. Additional C–B Brier improvement is small in absolute terms and varies by learner; the additional log-loss benefit is not universal (logit B is 0.320824 and C is 0.321622).

Direct neural group removals (Table 5) show the largest deterioration when the zone group is removed. Removing time or intrahour inputs instead slightly lowers loss. The small benefit of retaining market inputs is sensitive to the bootstrap block: its Holm p is 0.0458 at 7 days and 0.0561 at 14. These patterns do not support the claim that all four groups, or all 32 columns, are necessary or jointly optimal.

Table 5. Auxiliary neural predictor-group removals with independent retraining.

Condition	Inputs	Brier	Log Loss
Full	32	0.126188	0.317551
Without zone	22	0.140313	0.400202
Without time	23	0.125754	0.312261
Without market	24	0.126397	0.318159
Without intrahour	27	0.125868	0.314193

All conditions share the original test origins and chronological procedure. Group removal does not mask inputs of an already fitted model.

Figure 6 shows the corresponding neural group-removal scores.

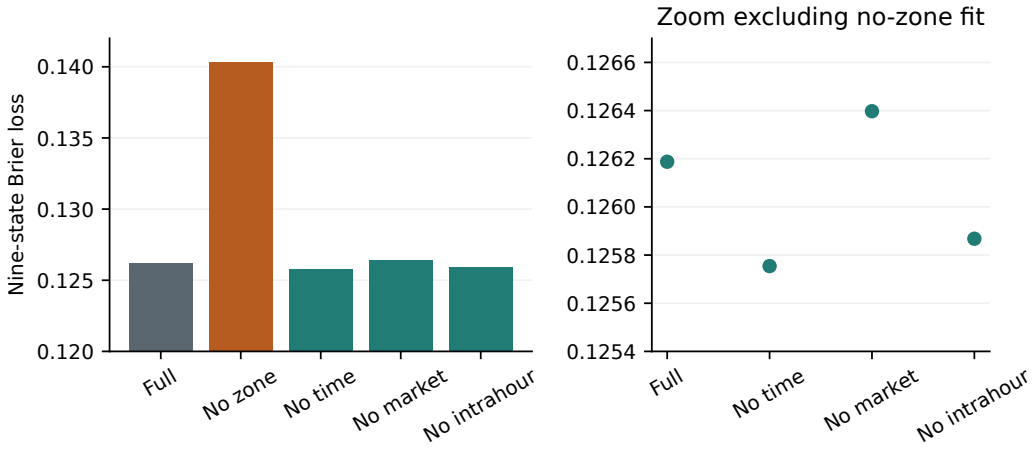


Figure 6. Auxiliary neural group-removal Brier scores. The right panel distinguishes the small differences among non-zone removals; lower is better.

5.3. Preserved Original Benchmark Comparison

The original frozen Full and Non-zone NN Brier scores reproduce as 0.126183 and 0.140341, a 10.09% relative reduction. The validation-selected challenger scores 0.123524. Full is 2.15% higher when the challenger score is the denominator; equivalently, the challenger is 2.11% lower when Full is the denominator. These values are kept separate from revision refits (Table 6). XGBoost and forest selection is part of the original four-model pool, with verified 11/7/3 counts. Low absolute ECE or coherent output does not override this proper-score disadvantage.

Table 6. Reproduced frozen original nine-state scores.

Model	Brier	Unclipped Log Loss
MLP reference	0.126183	0.317010
Non-zone NN	0.140341	0.400416
DeepHit	0.128081	0.319573
XGBoost	0.124165	0.308817
Survival forest	0.123770	∞
Student-t GJR	0.124610	0.302724
Gaussian GARCH	0.127932	0.311376
Validation-selected composite	0.123524	0.299610

N = 502,443. Full = original C neural; Non-zone NN = original A neural. The forest has one observed zero; the selected composite has none in its chosen folds. Gaussian GARCH is outside the original selection pool.

5.4. Intrabar-Contact Proper Scores

At 60 min, prevalence rises from 7.4914% for close exit to 12.1658% for contact; at 15 min, it rises from 2.4727% to 5.3258%. Table 7 uses the primary cohort and fixed close-trained probabilities. All seven models have larger nine-state Brier losses under contact, and finite log losses also increase. Preserved discrimination therefore cannot establish preserved probability accuracy. The forest has three observed zero masses under contact and infinite unmodified log loss. Directional scores exclude the 63 ambiguous first-contact origins; the all-origin binary scores in Appendix D include them.

Table 7. Fixed-probability contact sensitivity on 502,380 unambiguous origins.

Model	Contact Brier	Contact Log Loss
MLP reference	0.192220	0.473394
Non-zone NN	0.222067	0.607124
DeepHit	0.191880	0.462496
XGBoost	0.188194	0.455156
Survival forest	0.188758	∞
Student-t GJR	0.177770	0.420676
Gaussian GARCH	0.175469	0.423044

These probabilities were fitted for close exit, not retrained for contact. Ambiguous within-bar direction is unidentified. The earlier 523,575-origin analysis and its 57 ambiguous contacts are not combined with this cohort.

5.5. Leave-One-Asset and Leave-One-Year Inference

The original Full–Non-zone Brier difference is negative and Full–selected-challenger difference positive in every asset and year omission. Their Holm-adjusted significance persists with 3-, 7-, and 14-day blocks. Appendix D reports each effect, confidence interval, sample size, and seven-day adjusted test. This quantifies sign stability but does not supply independent replications: the same fixed fits and overlapping data generate the subsets, and 2026 is only a partial year. Log-loss tests follow the disclosed zero-mass rule and are reported in the supplementary inference file.

5.6. GARCH Path-Count and Scrambling Sensitivity

All 42 original-PMF reproduction gates pass before the 1050 sensitivity cells. Across every tested path-count/seed combination, Student-t GJR retains a lower Brier score than frozen Full and Gaussian GARCH a higher one. Table 8 reports five-seed means and ranges. From 2048 to 8192 paths, mean Brier changes by only −0.0000428 for GJR and −0.0000512

for Gaussian GARCH. With the original fold selections fixed, all 25 challenger mixtures retain a lower Brier score than Full; Full's disadvantage ranges from 2.081% to 2.164%.

This establishes Brier-ranking stability over the tested settings, not an optimal 2048-path choice or complete convergence of every probability or score. Gaussian mean log loss increases from 0.311376 at 2048 to 0.312973 at 8192 paths, so its logarithmic score is nonmonotone. The fixed pseudocount's total relative weight decreases from 0.2192% to 0.0549%; numerical and smoothing effects have not been separately identified. Five-seed ranges are descriptive numerical spreads, not confidence intervals.

Table 8. Sobol sensitivity: pooled score means and Brier ranges across five seeds.

Distribution	Paths	Mean Brier	Seed Range	Mean Log Loss
Student-t GJR	512	0.124849	[0.124747, 0.124952]	0.305993
Student-t GJR	1024	0.124658	[0.124595, 0.124716]	0.303622
Student-t GJR	2048	0.124608	[0.124572, 0.124638]	0.302804
Student-t GJR	4096	0.124576	[0.124564, 0.124581]	0.302385
Student-t GJR	8192	0.124565	[0.124560, 0.124575]	0.302233
Gaussian GARCH	512	0.128271	[0.128113, 0.128346]	0.312785
Gaussian GARCH	1024	0.128007	[0.127942, 0.128064]	0.311302
Gaussian GARCH	2048	0.127922	[0.127896, 0.127957]	0.311376
Gaussian GARCH	4096	0.127889	[0.127879, 0.127896]	0.312073
Gaussian GARCH	8192	0.127871	[0.127864, 0.127880]	0.312973

All entries use the same 502,443 test origins and fixed original fitted parameters. Seeds 42–46; pseudocount 0.5 per state. Ranges are not inferential intervals.

Figure 7 displays the path-count and scrambling sensitivity.

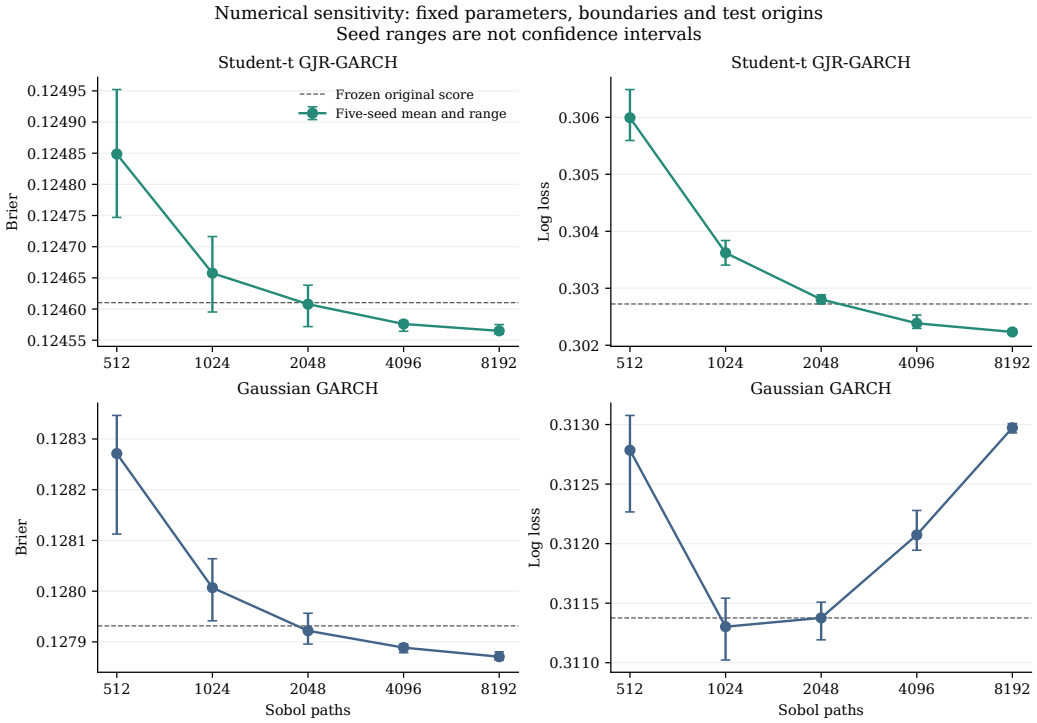


Figure 7. Sensitivity to simulation paths and scrambling. Green curves denote Student-t GJR-GARCH and blue curves denote Gaussian GARCH. Solid curves show five-seed means, error bars show seed minima and maxima, and dashed horizontal lines show the frozen original scores. Brier ordering is stable across the tested grid, while Gaussian log loss is nonmonotone. Spread across seeds measures numerical variability only.

5.7. Illustrative Common-Policy Episode Alerts

Table 9 is an optional use-case illustration with a common validation 1% budget, 60 min cooldown, and no additional daily cap; this slice is not an estimated institutional rule or a validated optimum. Every eligible episode is counted once. MLP reference detects 2142 of 12,580 episodes with 1909 false alerts. XGBoost, forest, and GJR detect more episodes under the same policy rule. The last column reports each model's break-even cost boundary rather than a saving evaluated at one assumed cost ratio. Actual test alert counts differ because validation budget equality does not imply identical test probability distributions. The full grid, including identical maximum daily capacities, is supplied rather than choosing test thresholds to equalize counts after the fact.

Table 9. Optional episode-alert use case under a common assumed policy.

Model	Alerts	False	Detected	Mean Lead (min)	Break-Even r
MLP reference	4051	1909	2142	25.6	$0.891 + 0.573s$
Non-zone NN	3809	2760	1049	30.6	$2.631 + 0.490s$
DeepHit	3885	1918	1967	26.5	$0.975 + 0.558s$
XGBoost	4337	1922	2415	24.9	$0.796 + 0.585s$
Survival forest	4729	2062	2667	24.6	$0.773 + 0.591s$
Student-t GJR	5174	2380	2794	26.3	$0.852 + 0.562s$
Gaussian GARCH	5092	2492	2600	26.7	$0.958 + 0.555s$
Independent sigmoid	3755	1714	2041	24.9	$0.840 + 0.585s$

Budget 1%, cooldown 60 min, no cap; eligible onsets=12,580, origins=502,443. Improvement over never alert requires r above the last-column expression at supplied s. Rounded point estimates; exact formulas in the Supplementary Materials.

The MLP boundary is $r > 0.8912 + 0.5731s$, directly calculated from 1909 false alerts, 2142 detections, and 1227.5 delay units. Appendix D Figure A1 compares assumed-cost surfaces, including never alert; these are conditional loss comparisons rather than realized monetary savings. Preferred models and positive-gain conditions depend on costs; the 1:5:0.5 scenario is included in the Supplementary Sensitivity Records. Cooldown and daily caps reduce repeated alerts and constrain capacity while sacrificing detections in some settings. The analysis does not demonstrate operational dominance of MLP reference. No uncertainty interval is assigned to the episode utility point estimates, and the grid is a sensitivity assessment, not an institution-calibrated decision study. Appendix D supplies additional policy and cost summaries; the complete 360 policies and 5760 cost rows accompany the source.

5.8. Generator Dependence and External Results

The earlier generator comparison (Table 10) reports DZL coverage 0.8228 versus STQ 0.7795 at similar mean log widths (0.0519 versus 0.0514), and lower descriptive quantile losses for DZL. However, downstream mean Brier skill is 0.1369 for DZL, 0.1546 for STQ, and 0.1582 for VB. The larger skills of the simpler generators argue against claiming a uniquely beneficial DZ-LSTM/downstream combination. Each skill has its own target and climatology, and the DZL source ledger differs from the fully cross-fitted primary ledger. These supporting results cannot establish the generator-independent effect of the revised information sets or justify DZ-LSTM necessity.

In the existing session-aware external end-to-end assessment, DZL WIS exceeds MQ WIS by 12.6–26.6%. All four markets fail the original 2% non-inferiority criterion (Table 11). Positive downstream Brier skill for XAU/USD and USA500 follows within-market re-estimation and is conditional on that market's supplied zones. EUR/USD and USD/JPY skill intervals include zero. These results do not support the complete pipeline's general-

ization or universal transfer of a fixed cryptocurrency predictor. We retain the negative upstream finding and favor simpler moving-quantile reference zones where that assessed zone-quality comparison supports them.

Table 10. Earlier supporting generator comparison (21 fold means).

Generator	Coverage	Log Width	Lower Pinball	Median Pinball	Brier Skill
DZL	0.8228	0.0519	0.00444	0.00813	0.1369
STQ	0.7795	0.0514	0.00470	0.00860	0.1546
VB	0.7615	0.0479	0.00472	0.00858	0.1582

This analysis uses its own earlier ledger and generator-specific risk sets, not the 502,443-origin matched factorial cohort. Skill is relative to each fold-training climatology. No raw same-target superiority or synergy is inferred.

Table 11. Existing external end-to-end zone assessment and refitted downstream skill.

Market	WIS Increase	Brier Skill	Reported 95% Interval	2% Zone Margin
EUR/USD	26.6%	−0.0527	[−0.4413, 0.1771]	Failed
USD/JPY	25.8%	−0.0638	[−0.2158, 0.0613]	Failed
XAU/USD	13.2%	0.1339	[0.0838, 0.1698]	Failed
USA500	12.6%	0.1060	[0.0446, 0.1642]	Failed

WIS increase is relative to session-matched MQ on paired targets. Downstream skill uses market-specific refitting and climatology. Positive skill does not override upstream failure. This is a separate external pipeline, not the primary cryptocurrency cohort.

5.9. Causal Moving-Quantile Replication on the Complete Fold Scope

5.9.1. Information Value Under a Simpler Boundary

Table 12 repeats the information experiment on newly constructed MQ60 labels and 474,591 common test origins. The four original nine-state learners comprise 252 independently fitted conditions. For these four learners, full zone information reduces Brier loss by 13.47–14.52% relative to the 22-input non-zone condition; each reduction is significant after Holm adjustment under all three date-block choices. Two boundary distances account for approximately 94–98% of the reduction. This extends the observed pattern to a simple causal boundary rule, but it does not imply that all zone variables are necessary or that information gains hold for every generator.

Table 12. MQ60 information-set replication on the same new first-exit target.

Learner	A Brier	B Brier	C Brier	C vs. A Reduction
Multinomial logit	0.181593	0.156848	0.155453	14.39%
XGBoost	0.179596	0.154908	0.154384	14.04%
Survival forest	0.180121	0.154561	0.153969	14.52%
MLP reference	0.180433	0.157633	0.156132	13.47%
Gaussian GARCH-X	0.148339	0.147403	0.146889	0.98%
Student-t GJR-GARCH-X	0.144211	0.144072	0.143911	0.21%

A = 22 non-zone inputs; B = A plus two distances; C = 32 inputs. All 21 folds and 474,591 origins. For the two GARCH-X rows, A/B/C denote fitted covariates in the return-variance equation. All three conditions also use the issued boundaries in path-to-exit mapping; objectives and training populations differ from the nine-state learners. Negative percentage reduction denotes deterioration. See Section 4.8. The MLP row is an auxiliary reference. These are within-generator gains, not improvements over original DZ-LSTM Full; the matched GARCH accuracy comparison is reported below.

For MLP reference, C–A is −0.024300, with seven-day 95% interval [−0.025984, −0.022682] and Holm $p = 0.000200$. The complete A/B/C contrasts and all three block lengths are supplied.

For the added GARCH-X fits, Gaussian Brier falls from 0.148339 (A) to 0.146889 (C), a 0.98% reduction; Student-t GJR falls from 0.144211 to 0.143911, a 0.21% reduction. Seven-day C–A intervals are $[-0.001806, -0.001138]$ and $[-0.000399, -0.000211]$, respectively; both Holm $p = 0.000100$, with rejection under all three block choices. Log losses also fall, from 0.344647 to 0.341909 for Gaussian and 0.333290 to 0.332775 for Student-t; their seven-day Holm p values are 0.000100 and 0.000600. Thus, the small additional learned-covariate effects on MQ60 coexist with the non-improving Student-t result under the original boundary reference. Complete A/B/C log losses and paired contrasts are supplied.

5.9.2. Paired Zone Quality on Observed Prices

Table 13 uses 641,976 identical observed price times rather than a generator-specific inside-zone risk set. At the same nominal 80% interval, WIS is $(0.5|r - m| + 0.1\,\mathrm{IS}_{0.2})/1.5$. MQ60 WIS is 5.81% higher than the frozen DZ-LSTM value. The seven-day paired difference is 0.000650, with 95% interval $[0.000501, 0.000796]$. This is evidence of lower DZ-LSTM zone prediction error in the assessed cryptocurrency sample, not its necessity for information gains or external superiority. The earlier ledger comparisons and negative external findings remain separate.

Table 13. Paired upstream zone quality at identical price times.

Reference	WIS	Coverage	Mean Log Width
DZL_frozen	0.011201	81.69%	0.0510
MQ30	0.011942	75.22%	0.0491
MQ60	0.011851	77.29%	0.0505
MQ120	0.011784	78.59%	0.0513

641,976 price observations; common nominal 80% interval. MQ30/MQ120 were not used to select the reported downstream MQ60 experiment.

5.9.3. Matched GARCH Probability Comparisons

Both ordinary volatility-process models have lower nine-state Brier loss than the four original full-input nine-state learners on the MQ60 target (Table 14). Each of the eight Brier differences has a negative 95% interval and remains Holm-significant under all three block lengths. The Student-t GJR value is 0.147176, approximately 5.74% lower than MLP reference's 0.156132. This is a matched model comparison within MQ60; it is not an A/B/C information-effect estimate or a comparison of generator quality.

Table 14. Full-input learners and volatility comparators on matched MQ60 outcomes.

Model	Brier	Log Loss	Any-Exit Brier	Brier Skill
Multinomial logit	0.155453	0.368893	0.067859	0.1553
MLP reference	0.156132	0.367204	0.068307	0.1516
XGBoost	0.154384	0.366804	0.067018	0.1611
Survival forest	0.153969	0.363087	0.066834	0.1634
Student-t GJR	0.147176	0.341411	0.064683	0.2003
Gaussian GARCH	0.150896	0.351159	0.067401	0.1801
Gaussian GARCH-X	0.146889	0.341909	0.064183	0.2018
Student-t GJR-GARCH-X	0.143911	0.332775	0.062065	0.2180

All eight rows share 474,591 test origins. Ordinary GARCH uses frozen train-only return-process estimates and newly simulated MQ60 PMFs; GARCH-X C rows use new joint covariate/process fits. Table 15 concerns only the two ordinary GARCH specifications; GARCH-X A/B/C inference is reported separately. Published scores have no probability floor.

Table 15. Paired MQ60 nine-state Brier differences with seven-day date blocks.

GARCH	Compared Learner	Difference	95% Interval	Holm p
Student-t GJR	Multinomial logit	−0.008278	[−0.009591, −0.006943]	0.000400
Student-t GJR	MLP reference	−0.008957	[−0.010306, −0.007588]	0.000400
Student-t GJR	XGBoost	−0.007208	[−0.008471, −0.005930]	0.000400
Student-t GJR	Survival forest	−0.006794	[−0.008024, −0.005547]	0.000400
Gaussian GARCH	Multinomial logit	−0.004557	[−0.006240, −0.002739]	0.000400
Gaussian GARCH	MLP reference	−0.005237	[−0.006933, −0.003371]	0.000400
Gaussian GARCH	XGBoost	−0.003488	[−0.005090, −0.001717]	0.000400
Gaussian GARCH	Survival forest	−0.003073	[−0.004681, −0.001300]	0.000900

Difference = GARCH minus learner; negative favors GARCH. Eight tests per score family; 20,000 synchronized date-bootstrap replicates. All score families and block lengths are in the Supplementary Materials.

5.9.4. Illustrative Common-Policy Episodes

Table 16 illustrates the same assumed validation 1% budget and 60 min cooldown for all six MQ60 comparators as a downstream use case. The eligible episode set is defined from the new boundaries and common origins. Each episode receives one detection credit; repeated true alerts and realized test alert totals are separately recorded. The supplied policy grid varies cooldown and causal daily capacity. The final column reports break-even cost ratios derived from the observed counts and leads. MLP requires $r > 0.7031 + 0.5831s$ under this rule. No institutional cost calibration or operational dominance is inferred.

Table 16. MQ60 episode alerts under the common illustrative 1%/60 min policy.

Model	Alerts	False	Detected	Mean Lead (min)	Break-Even r
Multinomial logit	4148	1631	2517	24.7	$0.648 + 0.588s$
MLP reference	3682	1520	2162	25.0	$0.703 + 0.583s$
XGBoost	4007	1554	2453	24.6	$0.634 + 0.589s$
Survival forest	3717	1352	2365	24.0	$0.572 + 0.600s$
Student-t GJR	3513	681	2832	22.2	$0.240 + 0.629s$
Gaussian GARCH	3622	797	2825	22.7	$0.282 + 0.621s$

No additional daily cap. Eligible episodes = 16,852; origins = 474,591. Improvement over never alert requires r above the final-column expression for the supplied delay weight s. Rounded coefficients are descriptive; no utility confidence interval or institutional cost estimate is supplied.

5.9.5. Full-Cohort Neural Numerical Audit

Table 17 distinguishes probabilities stored after float32 softmax/CIF mapping from stable log-softmax evaluation of every saved MQ60 neural model. The distance-only B condition contains one stored zero at the observed class, giving infinite as-stored log loss; its full-cohort stable-logit loss is 0.370996. The corresponding finite-logit diagnostic explains this representation failure. Full-input C has no observed stored zero and stable loss 0.367204. Brier changes from the stable representation are below 3×10^{-11} in the pooled means, and the previous Brier tables and inference are preserved. This audit covers all 63 neural fits and both validation and test partitions rather than diagnosing only one extreme observation.

Table 17. MQ60 neural log-loss representations across all test origins.

Information Set	Stored Log Loss	Stable-Logit Loss	Stored Observed Zeros
A	0.495211	0.495211	0
B	∞	0.370996	1
C	0.367204	0.367204	0

Each row uses all 474,591 test origins from 21 fitted models. Stable evaluation changes no weights, labels, or stored PMF files. Validation results and reconstruction gaps accompany the source.

5.10. Matched Exact-Calendar Information Comparisons

Tables 18 and 19 report the completed 504-refit reconstruction on the unchanged historical samples. Under DZL, all four learners have lower C than A Brier loss; signed relative reductions range from 10.02% to 11.85%. All four C–A reductions remain Holm-significant with a negative interval under all three block choices. Under MQ60, all four learners have lower C than A Brier loss; signed relative reductions range from 13.58% to 14.53%. All four C–A reductions remain Holm-significant with a negative interval under all three block choices.

Table 18. Exact-calendar refits on the unchanged historical target and sample.

Zone	Learner	A Brier	B Brier	C Brier	C–A Reduction (%)	C Log Loss
DZL	Multinomial logit	0.141760	0.127928	0.127550	10.02	0.319836
DZL	MLP reference	0.140437	0.128049	0.126114	10.20	0.316287
DZL	XGBoost	0.140041	0.124584	0.124099	11.38	0.308674
DZL	Survival forest	0.140388	0.124446	0.123749	11.85	∞
MQ60	Multinomial logit	0.181582	0.156558	0.155201	14.53	0.367749
MQ60	MLP reference	0.180493	0.157425	0.155982	13.58	0.366507
MQ60	XGBoost	0.179498	0.154865	0.154351	14.01	0.366738
MQ60	Survival forest	0.180092	0.154507	0.153943	14.52	∞

A: 22 non-zone variables; B: A plus two distances; C: 32 inputs. Positive relative reduction means lower C Brier. Raw scores compare learners within a generator, not generator quality across different labels. Tables 2–17 remain separately identified legacy results.

Table 19. Paired C–A Brier contrasts for corrected histories; 14-day date blocks.

Zone	Learner	Difference	95% Interval	Holm p
DZL	Multinomial logit	−0.014211	[−0.016497, −0.011828]	0.00020
DZL	MLP reference	−0.014323	[−0.015976, −0.012816]	0.00020
DZL	XGBoost	−0.015941	[−0.017604, −0.014364]	0.00020
DZL	Survival forest	−0.016638	[−0.018311, −0.015017]	0.00020
MQ60	Multinomial logit	−0.026381	[−0.028633, −0.024268]	0.00020
MQ60	MLP reference	−0.024511	[−0.026508, −0.022605]	0.00020
MQ60	XGBoost	−0.025148	[−0.027240, −0.023179]	0.00020
MQ60	Survival forest	−0.026150	[−0.028266, −0.024146]	0.00020

All block choices, B–A/C–B contrasts, per-fold scores, unmodified log scores, and paired corrected-minus-legacy results are supplied. Relative to the corresponding legacy full-input forecasts, corrected C Brier point estimates are lower in all eight generator/learner combinations. This comparison describes the implementation change; the C–A contrasts assess whether the information gain persists after that change. The compact B representation and full C representation are compared directly, without asserting necessity of all ten zone variables.

5.11. Fixed-Model Results on the Unused MQ60 Period

All four learners have lower C than A Brier loss; signed relative reductions range from 14.02% to 14.77%. All four C–A reductions remain Holm-significant with a negative interval under all three block choices. Tables 20 and 21 use the same 19,056 extension origins for every learner and information set. No model is selected for deployment from these scores.

The follow-up concerns the temporal stability of information effects for the declared MQ60 task in the same three assets. It supplies an additional period without establishing cross-asset, institutional-boundary, or complete-pipeline generalization. All results, including nonsignificant and unfavorable contrasts when present, are retained. The interval is short relative to the historical sample, and bootstrap inference remains conditional on fitted models.

Table 20. Stored 2026 MQ60 models evaluated on 30 June–30 September 2026.

Learner	A Brier	B Brier	C Brier	C–A Reduction (%)	C Log Loss
Multinomial logit	0.182758	0.158602	0.156619	14.30	0.368389
MLP reference	0.180125	0.156291	0.154872	14.02	0.358229
XGBoost	0.180975	0.154972	0.154614	14.57	0.359499
Survival forest	0.181934	0.155371	0.155068	14.77	0.362249

Table 21. Paired C–A Brier contrasts on the unused period; 14-day date blocks.

Learner	Difference	95% Interval	Holm p
Multinomial logit	−0.026138	[−0.034995, −0.018319]	0.00020
MLP reference	−0.025253	[−0.033401, −0.018281]	0.00020
XGBoost	−0.026361	[−0.035780, −0.018157]	0.00020
Survival forest	−0.026867	[−0.036422, −0.018544]	0.00020

6. Discussion

6.1. What the Information Comparison Supports

The principal evidence concerns the additional predictive value of boundary-relative inputs for a specified conditional target. Across four learners, adding the zone group lowers Brier loss by approximately 10–12%. The two-distance condition obtains most of that reduction. The relevant contribution is therefore identifying useful information and comparing learners under the same target and sample. The softmax, likelihood, and CIF identities supply the established framework for this comparison.

This result has an intuitive basis: the event is defined relative to a boundary, and a forecast without that boundary omits part of the event's state. It is not surprising that distance is useful. Its measured magnitude and persistence across learners are empirical findings. This study does not establish that the zone encodes a new independent market risk factor, nor that all transformed coordinates are needed. Duplicate columns and constant close-breach flags reinforce this limitation. Small C–B gains and group-removal improvements favor consideration of simpler representations.

Model choice remains consequential. XGBoost and forest have lower Brier scores than MLP reference, while zero-mass log loss prevents calling the forest best on all probability criteria. The original selected challenger and new information matrix agree that neural superiority is unsupported. Coherence is desirable for compatible probabilities but is shared by coherent competitors and cannot explain their accuracy ranking. The auxiliary sigmoid comparison is confounded by training masks, class weighting, and other objectives.

The GARCH-X extension qualifies any learner-independent interpretation: Gaussian Brier improves under both references, whereas Student-t GJR improves under MQ60 but has no established Brier gain and worse log loss under DZ-LSTM. All its variants already use boundaries to calculate exit outcomes. Additional learned zone covariates are therefore useful in some assessed specifications rather than necessary or uniformly beneficial.

6.2. Role of the Reference Generator

DZ-LSTM supplies a specific audited boundary ledger for the main conditional study. It is not the paper's claimed methodological advance or a preferred generator inferred from downstream scores. The earlier coverage/pinball observations describe one possible upstream role; their legacy provenance and the external WIS failures restrict their strength. The earlier STQ/VB results do not establish a unique generator–learner advantage; their different labels, risk sets, and fit provenance prevent controlled attribution or refutation of such synergy. Moving quantiles are viable simpler alternatives.

The additional causal MQ60 experiment now provides a separately refitted four-learner information comparison on its own labels and risk set. The 13.47–14.52% improvements reproduce the direction of the DZ-LSTM information effect under one simpler boundary specification. Paired observed-price WIS is lower for DZ-LSTM in these cryptocurrencies, while MQ60 supports a simpler unfitted reference. These findings do not establish generator necessity, invariance across all boundaries, or external fixed-model transfer. Cross-fitting and causal issuance control information timing; they do not remove systematic generator characteristics. GARCH performance on MQ60 further shows that useful boundary information does not imply superiority of a neural learner.

6.3. Scope of the Optional Alert Use Case

This study's main question concerns the predictive value of boundary information and the accuracy of the time–direction probability distribution. An alert is one possible use of that distribution. The price observations and predictions used in the illustration are actual recorded data; the assumptions concern the action rule and relative costs. Episode consolidation prevents repeated warnings of one departure from receiving multiple detection credits. Shared validation rules allow comparison of warning burden and lead but do not force identical realized test alert totals. The break-even expressions describe preferences under which an assumed loss is lower. They do not establish institutional usefulness or justify deploying a model selected from the test-cost surface.

More alerts can increase detections while also increasing false and repeated warnings; fewer alerts can reduce burden while missing more events. No universally best weighting follows from alert volume alone. Appendix E finds gains and losses that depend on the comparator and supplied generator, not a consistently improved replacement ratio. The monitored zones remain model-issued forecast references. Actual collateral, liquidation, or hedging actions would require appropriate institutional boundaries, measured losses, exposure and execution costs, and separate decision validation. The 1:5:0.5 ratio is retained only as an assumed scenario and has no representative or economically optimal status. Outcomes exclude trade-level first passage and within-bar ordering.

6.4. Limitations and Further Evidence

First, the period was already examined during development. The frozen original protocol and revision diagnostics preserve chronological information controls, but neither is publicly preregistered prospective evidence on unused data. Bootstrap uncertainty is conditional on fixed fits and does not include parameter, tuning, or generator estimation uncertainty. Overlapping origins and common market dates are addressed through synchronized blocks, but finite block choices do not remove every form of dependence.

Second, the event is a strict first close exit in a retained window ending at the last nominal knot. Its daily target/hourly knot alignment is a modeling convention. Intrabar contact has higher prevalence and worse fixed-probability proper scores; 63 first two-sided contacts have unidentified direction. A contact-trained model or trading-level event study would address a different task.

Third, the legacy feature matrix uses retained-row lags and partly retained hourly totals; its names and interpretation remain qualified. The new exact-calendar experiment now reconstructs histories from the full raw grid and repeats all 504 four-learner A/B/C fits on identical targets and origins (Sections 4.9 and 5.10). It directly evaluates this implementation sensitivity. Legacy neural group removals, ordinary GARCH and GARCH-X fits, contact diagnostics, and optional alert results are not relabeled as corrected-clock refits. Further model specifications and observation rules still need their own reconstruction and training checks.

Fourth, the MQ60 information comparison evaluates one causal alternative boundary. The corrected-history replication and locked 30 June–30 September follow-up extend that assessment within the same three cryptocurrencies using fixed 2026 models on the new period. This is retrospective unused-period evidence, not publicly preregistered prospective confirmation. Generator replacement still changes labels and risk sets, preventing raw cross-generator score rankings. The extension does not test DZ-LSTM generation or new assets, and the full external pipeline still fails its upstream criterion.

Finally, numerical Sobol sensitivity preserves tested Brier rankings but shows nonmonotone Gaussian log loss and changing smoothing weight. Forest zero mass distinguishes Brier from log-score performance. The episode cost boundaries are descriptive functions of observed counts and leads, omit uncertainty in those counts, and exclude institution-specific economic exposures and repeat-alert monetary costs. These limitations prevent conclusions about profitability, universal accuracy, or demonstrated operational dominance.

7. Conclusions

This study compares information sets and learning models for first closing-price exits from pre-issued dynamic zones. Established discrete competing-risk methods provide a common nine-state outcome for no exit and upper/lower exits in four intervals. Under the evaluated ledger, cohort, and chronological procedure, adding zone-state information lowers Brier loss across four learners, with much of the observed gain obtained from two boundary distances. Additional representation and model complexity are not uniformly beneficial. After exact-calendar reconstruction, full inputs improve Brier loss in 4/4 DZL and 4/4 MQ60 learners on the same historical outcomes. The fixed MQ60 follow-up adds 4/4 positive C–A contrasts on the unused period, of which 4/4 remain significant under all tested date-block choices. These checks strengthen a conditional information result within the studied assets and boundary rules.

MLP reference is a coherent neural implementation of this task, but XGBoost and forest have lower Brier scores, and the original validation-selected challenger retains its advantage. Episode-policy comparisons illustrate an optional application and do not establish neural operational superiority or realized economic value. Tested GARCH simulation settings preserve Brier ordering, while log-score and zero-mass limitations remain. Existing external results fail to support the complete pipeline's generalization.

The original retained-row MQ60 analysis (Section 5.9) gives 13.47–14.52% Brier reductions, with most gains captured by two distances. The exact-calendar refits (Section 5.10) give 13.58–14.53% reductions on the same historical targets and origins. On the original matched MQ60 outcomes, Student-t GJR-GARCH has lower Brier loss than the four full-input learners. These analyses concern boundary information and comparative evidence rather than the primacy of MLP reference.

The empirical contribution of the boundary preprocessing and supplied representation is a conditional assessment of what information helps and which learners predict well under a common financial boundary task. Threshold-based alerts remain an optional illustration of using those probabilities under assumed policies and costs. Claims of new output-layer novelty, indispensable DZ-LSTM generation, realized institutional value, or universal transfer are unsupported. The new GARCH-X comparisons additionally show that learned zone-covariate gains are model-, generator-, and score-dependent. The added exact-calendar reconstruction and fixed MQ60 follow-up directly assess two remaining implementation and temporal questions (Sections 5.10 and 5.11). Prospective evaluation, further boundary constructions, genuinely new assets, and institutional decision validation are needed to broaden the conclusions.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/math14213963/s1.

Author Contributions: Conceptualization, G.K., Y.S. and S.I.H.; methodology, G.K.; software, G.K.; validation, G.K.; formal analysis, G.K.; investigation, G.K.; resources, Y.S. and S.I.H.; data curation, G.K.; writing—original draft preparation, G.K.; writing—review and editing, G.K., Y.S. and S.I.H.; visualization, G.K., Y.S. and S.I.H.; supervision, Y.S. and S.I.H.; project administration, Y.S. and S.I.H.; funding acquisition, Y.S. and S.I.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Research Foundation of Korea (NRF) through grants funded by the Korean government (MSIT), grant number RS-2026-25478322, and by the Soonchunhyang University Research Fund.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. The raw cryptocurrency data are openly available through Binance Public Data at https://data.binance.vision/ (accessed on 25 October 2026). A reproducibility package is available from the corresponding author upon reasonable request.

Acknowledgments: During the preparation and revision of this manuscript, the authors used the generative AI tool OpenAI Codex (OpenAI; accessed October 2026) to assist with analysis-script drafting and execution, preparation of figures and manuscript files, and computational consistency checks. The tool also assisted with English translation and language editing. The authors reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Appendix A. Predictor Definitions

Table A1 lists the predictor definitions. This appendix audits the original retained-row implementation. The new exact-calendar definitions are specified in Section 4.9; they retain these variable names and groups while rebuilding the affected market and intrahour histories before zone exclusions.

Let C, L, M, U be the current close and the issued lower, median, and upper values; $W = U - L$. All market and wick values become available only at bar completion. Clock variables use the nominal bar key. "Retained rows" means the already aligned dataset after exclusion of nominal 23:15, 23:30, and 23:45 rows. Ratios below reproduce the original matrix rather than an improved feature specification.

Table A1. Exact predictor groups and audited operations.

No.	Variable(s)	Group	Definition
1	minute_of_day	Time	UTC hour × 60 plus minute of the nominal key
2–3	minute_sin, minute_cos	Time	Sine/cosine of 2π times minute-of-day/1440
4–5	hour_sin, hour_cos	Time	Sine/cosine of 2π times UTC hour/24
6–7	day_of_week_sin/cos	Time	Sine/cosine of 2π times weekday/7
8–9	remaining_minutes_in_hour; remaining_hour_ratio	Time	60 minus (nominal minute + 15); that value divided by 60
10–11	return_15m; log_return_15m	Market	Simple/log return over one retained row

Table A1. *Cont.*

No.	Variable(s)	Group	Definition
12–13	return_30m; return_60m	Market	Simple returns over two/four retained rows; not always 30/60 elapsed minutes
14–15	rolling_volatility_1h/4h	Market	Sample SD of one-row log returns over four/16 retained rows
16	volume_change_15m	Market	Simple volume change over one retained row
17	volume_z_4h	Market	Volume minus its 16-row mean, divided by its 16-row SD
18	intrahour_return	Path	Current close divided by first open of the current nominal hour, minus one
19–21	intrahour_high/low/range_so_far_ratio	Path	Completed hourly high/low relative to hour open, and completed range divided by hour open
22	intrahour_volume_cumsum_ratio	Path	Hourly volume so far divided by the shifted median of previous 168 retained-hour totals (minimum 24); some totals contain fewer than four bars
23	zone_low_ratio	Zone	$(C - L)/C$; identical to lower distance
24	zone_mid_ratio	Zone	$(C - M)/C$
25	zone_high_ratio	Zone	$(U - C)/C$; identical to upper distance
26	zone_width_ratio	Zone	W/C
27	zone_position	Zone	$(C - L)/W$
28	zone_mid_distance_ratio	Zone	$(C - M)/W$
29–30	distance_to_upper/lower_ratio	Zone	$(U - C)/C$ and $(C - L)/C$
31–32	close_breach; wick_breach	Zone	Current-bar strict close/wick breach flags; no cumulative-day interpretation; close flag zero in the risk set

Appendix B. Training and Execution Details

The revision runtime uses Python 3.12.14, NumPy 2.3.5, pandas 3.0.1, SciPy 1.18.0, PyTorch 2.14.1 (CPU), XGBoost 3.4.1, and scikit-learn 1.9.1. Native forest fits use R 4.6.1 and randomForestSRC 3.6.2. Original frozen results retain their original runtime; they are not silently substituted by refits. Every input set is selected and standardized without test information. Neural fold seeds follow 28,000 + year + 100 × asset index; the original separate five-initialization analysis is not the Sobol scrambling study. Forest seeds and pre-2020 complexity preferences are preserved in the fit records.

All 315 fit outputs have common origin/label checks, train-only scaler and artifact hashes. Held-out probabilities were reconstructed from every saved neural/logit/XGBoost model. All forest output keys, ordering, masses, and hashes were checked; nine saved forest models (three assets by three information sets in 2020) were loaded for held-out prediction reconstruction. This is the actual scope of model-load verification, not a claim that every forest model was reloaded. Sobol reproduction and output aggregation checks precede reporting.

The XGBoost grid uses depths 3/5, learning rates 0.03/0.07, minimum child weights 200/800, row fractions 0.8/1.0, column fractions 0.7/1.0, and L2 penalties 1/10 in the original 12 fractional combinations supplied in `learner_candidate_grids.json`; it is not their full Cartesian product. The objective is nine-class soft probability, histogram trees, a maximum 2000 boosting rounds, and validation early-stopping patience 100. Candidates within 0.001 of minimum pre-2020 validation NLL are ordered by lower depth, larger minimum child weight, larger L2 penalty, lower learning rate, then fewer trees. Forest candidates are $(\text{mtry}, \text{nodesize}) \in \{(6, 250), (6, 750), (11, 250), (11, 750)\}$, with 300 trees, at most 50,000 sampled training rows, and 10 candidate split points. Its 0.001 NLL tolerance favors smaller mtry, followed by larger nodesize. Candidate-tuning NLL retains the original 10^{-12} probability floor; published test log scores are unmodified. These controls differ from the separate validation-Brier rule selecting the original four-model challenger.

The Supplementary Files preserve information-set scores, paired differences for every block, finite/undefined log-loss flags, group ablation scores, original fold counts, fixed-probability contact scores, omission inference, all alert policies/cost scenarios, and simulation summaries. Candidate-specific pre-2020 selections and complete models remain in the reproducibility archive available on request.

Appendix C. Provenance and Interpretation of Evidence

The fully cross-fitted original run is `crypto_crossfit_v1_fully_crossfitted_20260810`. Its input manifest begins `c85aa673894d9be6`; full input hashes and table-source hashes accompany the source archive. A total of 9 frozen artifacts and 147 source inputs were checked before reconstruction of the 502,443 common origins. The three original datasets each contain 245,824 aligned rows. Raw-bar counts, aligned rows, and eligible risk origins are separate quantities.

Original comparisons and revision refits share the primary ledger/cohort but have separate fit provenance. Contact, omission, and policy analyses use fixed original probabilities. Sobol varies simulation while preserving original fits. Earlier generator alternatives use the `stage2_8_zone_generator_robustness` ledger from before the fully cross-fitted rebuild; its fitted-history limitation is not resolved by the primary ledger audit. External session-aware fits are separate again. The additional MQ60 analysis instead constructs a causal unfitted reference, independently refits 252 learner conditions, and recomputes 42 GARCH PMFs on the new risk set. Its same-price WIS table uses 641,976 observations. None of these analyses is presented as an untouched prospective replication.

The source audit corrected nominal/completion timestamps, final-knot handling, current-bar breach flags, intrahour rather than daily path definitions, and duplicate distance variables. An early-day exclusion diagnostic retains the two original Brier effect directions but uses already fitted probabilities and cannot validate corrected calendar-window features. The chronological cross-fit treats the daily/path generator as one bundle. Its annual OOS outputs supply every downstream partition; there is no additional nested cross-fitting of daily anchors within upstream training. Ledger checks require upstream training targets to end before upstream validation and validation targets to end before the prediction block. The recorded fitted-history and source-origin-in-fit flags are false. Source hashes and statistical computation checks demonstrate reproducibility of recorded operations, not scientific validity of every modeling assumption.

Appendix D. Additional Scores and Policy Diagnostics

Table A2 reports the original neural close-exit and contact scores by horizon.

Table A2. Original neural any-exit/contact scores at each horizon.

Minutes	Close Rate	Contact Rate	Close Brier	Contact Brier	Contact Log Loss
15	2.473%	5.326%	0.020872	0.041373	0.153704
30	4.393%	8.073%	0.035030	0.059099	0.215329
45	6.037%	10.275%	0.046813	0.073170	0.263176
60	7.491%	12.166%	0.056959	0.084805	0.301171

All 502,443 origins are retained, including direction-ambiguous contacts. There are no observed zero probabilities in these neural binary scores. The Supplementary CSV contains all seven models.

Table A3 reports the paired leave-one-asset and leave-one-year diagnostics.

Table A3. Paired original Brier omission diagnostics; seven-day blocks.

Omitted	Comparison	Origins	Difference	95% Interval	Holm p
asset BTC	Full–Non-zone	335,370	−0.014419	[−0.015798, −0.013122]	0.000300
asset BTC	Full–Selected	335,370	0.002744	[0.002041, 0.003500]	0.000300
asset ETH	Full–Non-zone	333,925	−0.014705	[−0.016026, −0.013436]	0.000300
asset ETH	Full–Selected	333,925	0.002903	[0.002111, 0.003813]	0.000300
asset XRP	Full–Non-zone	335,591	−0.013354	[−0.014728, −0.012042]	0.000300
asset XRP	Full–Selected	335,591	0.002332	[0.001565, 0.003133]	0.000300
year 2020	Full–Non-zone	430,347	−0.014882	[−0.016316, −0.013542]	0.000700
year 2020	Full–Selected	430,347	0.001180	[0.000722, 0.001640]	0.000700
year 2021	Full–Non-zone	436,901	−0.012644	[−0.013769, −0.011551]	0.000700
year 2021	Full–Selected	436,901	0.002680	[0.001894, 0.003560]	0.000700
year 2022	Full–Non-zone	422,516	−0.014474	[−0.015918, −0.013118]	0.000700
year 2022	Full–Selected	422,516	0.003110	[0.002353, 0.003967]	0.000700
year 2023	Full–Non-zone	415,598	−0.015751	[−0.017150, −0.014434]	0.000700
year 2023	Full–Selected	415,598	0.003104	[0.002274, 0.004006]	0.000700
year 2024	Full–Non-zone	423,043	−0.013892	[−0.015345, −0.012501]	0.000700
year 2024	Full–Selected	423,043	0.002801	[0.001998, 0.003686]	0.000700
year 2025	Full–Non-zone	422,800	−0.013606	[−0.015055, −0.012240]	0.000700
year 2025	Full–Selected	422,800	0.002915	[0.002093, 0.003816]	0.000700
year 2026	Full–Non-zone	463,453	−0.013946	[−0.015270, −0.012689]	0.000700
year 2026	Full–Selected	463,453	0.002838	[0.002094, 0.003654]	0.000700

Models are held fixed. Asset family = 6; year family = 14. All three block lengths and log-loss records are in the Supplementary Materials. Positive Full–Selected difference favors the challenger.

Table A4 summarizes the illustrative cooldown comparisons.

Table A4. Illustrative cooldown effects at the same validation 1% budget.

Model	Cooldown (min)	Alerts	Detected	Repeat Hits	Mean Lead (min)
MLP reference	0	6102	2360	834	27.5
XGBoost	0	6036	2627	730	26.1
Student-t GJR	0	8291	3029	1318	28.3
MLP reference	60	4051	2142	0	25.6
XGBoost	60	4337	2415	0	24.9
Student-t GJR	60	5174	2794	0	26.3
MLP reference	120	3316	1740	0	25.6
XGBoost	120	3676	2057	0	24.7
Student-t GJR	120	4350	2344	0	26.3

No daily cap. Detection credit is once per episode; repeat hits are shown separately. The full file includes every model, budget, and daily cap condition.

Table A5 summarizes the descriptive policy and cost sensitivity.

Table A5. Descriptive sensitivity across all 720 policy/cost settings per model.

Model	Settings	Minimum Saving/1000	Maximum Saving/1000	Lower Cost than Full
MLP reference	720	−50.52	311.50	0
Non-zone NN	720	−78.65	140.59	0
DeepHit	720	−47.30	299.72	21
XGBoost	720	−60.24	334.68	593
Survival forest	720	−65.54	335.08	595
Student-t GJR	720	−83.63	329.34	466
Gaussian GARCH	720	−83.83	326.33	384
Independent sigmoid	720	−44.96	313.30	600

Savings are relative to never alert in illustrative units. Extremes span different policy and cost assumptions and are not tuned recommendations. Counts use identical policy/cost rows; no statistical dominance is implied.

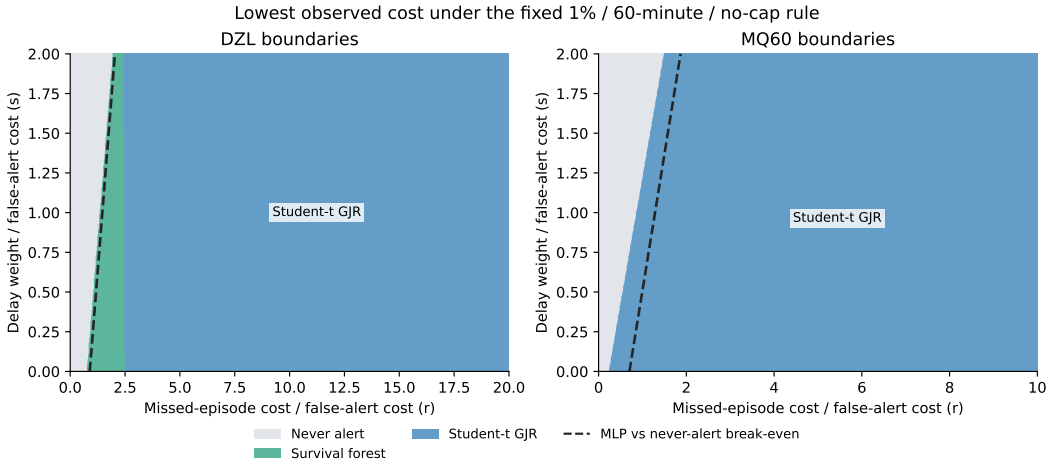


Figure A1. Illustrative assumed-cost regions under the existing fixed 1% validation-budget, 60 min cooldown, no-cap rule. Colors identify the model or never-alert policy with the lowest observed cost. Dashed lines show the MLP versus never-alert break-even boundary, which differs from being the lowest-cost model. All original comparators, including the auxiliary sigmoid, and all six MQ60 comparators enter their respective comparisons. Display limits follow the upper endpoints of the existing sensitivity ranges; supplied affine inequalities cover all nonnegative ratios. No ratio, policy, or model is optimized for deployment from this test surface.

Appendix E. Exploratory Selection of Optional Alert Policies

This retrospective check asks whether assumed loss weights can be used as policy-selection parameters. It does not estimate institutional losses or refit a forecast model. False-alert weight is fixed at one, missed-episode weights are $1, 2, 5, 10, 20$, and delay weights are $0, 0.25, 0.5, 1, 2$: 25 combinations. Each uses the same 45 threshold-budget/cooldown/cap candidates and never alert. All eight original alert comparators and all six MQ60 comparators share the procedure.

The existing 60-day validation interval is divided chronologically at 30 days before each test-year start. Earlier-part origins must have their entire 60 min label path ending strictly before the division. Probability ranks in that part define the candidate thresholds. Each penalty ratio chooses the earlier-part minimum-loss policy subject to the declared actual alert ceiling. That unchanged policy is evaluated in later validation; if it exceeds the later ceiling, it is replaced by never alert. Later-part episode detections under the same ceiling select the ratio. Ties prefer fewer false and repeated alerts, longer aggregate first-warning lead, fewer alerts, and the declared policy order. Ratios yielding the same selected policy

prefer weights closest to 1:5:0.5. The primary ceiling is 1%; 0.5%, 2.5%, 5%, and 10% are supplied sensitivities. Choices are frozen before test probability arrays are evaluated.

The fixed-ratio control applies the same new split and gates with 1:5:0.5 held constant. Direct policy selection instead selects from all feasible policies using the same later-part detection/burden criterion without penalty weights. The original illustrative rule uses the full validation interval for its budget threshold and a 60 min cooldown without a daily cap. Hence, differences from that rule also involve validation use and policy construction; they cannot be attributed solely to changing a ratio.

Table A6. Neural-reference outcomes in the exploratory optional-alert comparison at the 1% validation ceiling.

Zone	Policy Procedure	Detected	False Alerts	All Alerts
DZL	Original illustrative rule	2142	1909	4051
DZL	Selected penalty ratio	1659	1803	3621
DZL	Matched fixed 1:5:0.5	1362	1531	3047
DZL	Matched direct policy selection	1792	1814	3654
MQ60	Original illustrative rule	2162	1520	3682
MQ60	Selected penalty ratio	1759	1218	3059
MQ60	Matched fixed 1:5:0.5	1494	1,100	2768
MQ60	Matched direct policy selection	1681	1168	2871

DZL uses 502,443 origins and 12,580 eligible episodes; MQ60 uses 474,591 origins and 16,852 episodes. Counts compare policies within each target. Shared validation ceilings do not impose identical realized test alert counts.

The selected-ratio procedure detects more episodes than the matched fixed-ratio control but fewer than the original illustrative rule under both generators (Table A6). False and total alerts also differ. Comparison with matched direct policy selection is mixed. The 14-day paired date-bootstrap recall differences versus the original rule are −3.839 percentage points for DZL (95% interval [−5.105, −2.659]) and −2.391 for MQ60 ([−2.851, −1.927]); primary Holm-adjusted $p < 0.001$. All 3-, 7-, and 14-day results and matched-control contrasts are supplied. No uniformly superior replacement weighting is identified.

This is a follow-up on previously studied data, not a pristine prospective test. Underlying checkpoints and hyperparameters already used the original full validation interval; the two parts separate policy construction from ratio selection only. Bootstrap inference conditions on fixed fits and choices. Multiple ratios can produce the same policy, so selection frequency does not identify a true economic coefficient. No test-best ratio is adopted. All-model policy metrics, every ratio, capacity-gate outcomes, frozen choices, input hashes, and inference accompany the source. Forecast PMFs and their primary proper scores are unchanged.

References

1. Allison, P.D. Discrete-Time Methods for the Analysis of Event Histories. *Sociol. Methodol.* **1982**, *13*, 61–98. [CrossRef]
2. Tutz, G.; Schmid, M. *Modeling Discrete Time-to-Event Data*; Springer: Cham, Switzerland, 2016. [CrossRef]
3. López de Prado, M. *Advances in Financial Machine Learning*; Wiley: Hoboken, NJ, USA, 2018; ISBN 978-1-119-48208-6.
4. Lee, C.; Zame, W.R.; Yoon, J.; van der Schaar, M. DeepHit: A deep learning approach to survival analysis with competing risks. In *Proceedings of the AAAI Conference on Artificial Intelligence*; AAAI Press: Palo Alto, CA, USA, 2018; Volume 32, pp. 2314–2321. [CrossRef]
5. Engle, R.F. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica* **1982**, *50*, 987–1007. [CrossRef]
6. Bollerslev, T. Generalized autoregressive conditional heteroskedasticity. *J. Econom.* **1986**, *31*, 307–327. [CrossRef]
7. Glosten, L.R.; Jagannathan, R.; Runkle, D.E. On the relation between the expected value and the volatility of the nominal excess return on stocks. *J. Finance* **1993**, *48*, 1779–1801. [CrossRef]

8. Bollerslev, T. A conditionally heteroskedastic time series model for speculative prices and rates of return. *Rev. Econ. Stat.* **1987**, *69*, 542–547. [CrossRef]
9. Katsiampa, P. Volatility estimation for Bitcoin: A comparison of GARCH models. *Econ. Lett.* **2017**, *158*, 3–6. [CrossRef]
10. Catania, L.; Grassi, S. Forecasting cryptocurrency volatility. *Int. J. Forecast.* **2022**, *38*, 878–894. [CrossRef]
11. Wang, Y.; Andreeva, G.; Martin-Barragan, B. Machine learning approaches to forecasting cryptocurrency volatility: Considering internal and external determinants. *Int. Rev. Financ. Anal.* **2023**, *90*, 102914. [CrossRef]
12. Koenker, R.; Bassett, G. Regression quantiles. *Econometrica* **1978**, *46*, 33–50. [CrossRef]
13. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
14. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; Association for Computing Machinery: New York, NY, USA , 2016; pp. 785–794. [CrossRef]
15. Gu, S.; Kelly, B.; Xiu, D. Empirical asset pricing via machine learning. *Rev. Financ. Stud.* **2020**, *33*, 2223–2273. [CrossRef]
16. Cheung, J.; Rangarajan, S.; Maddocks, A.; Chandra, R. Quantile deep learning models for multi-step ahead time series prediction. *Appl. Soft Comput.* **2026**, *186*, 114043. [CrossRef]
17. Kunitomo, N.; Ikeda, M. Pricing Options with Curved Boundaries. *Math. Finance* **1992**, *2*, 275–298. [CrossRef]
18. Geman, H.; Yor, M. Pricing and Hedging Double-Barrier Options: A Probabilistic Approach. *Math. Finance* **1996**, *6*, 365–378. [CrossRef]
19. Cox, D.R. Regression models and life-tables. *J. R. Stat. Soc. Ser. B (Methodol.)* **1972**, *34*, 187–202. [CrossRef]
20. Fine, J.P.; Gray, R.J. A proportional hazards model for the subdistribution of a competing risk. *J. Am. Stat. Assoc.* **1999**, *94*, 496–509. [CrossRef]
21. Lee, C.; Yoon, J.; van der Schaar, M. Dynamic-DeepHit: A deep learning approach for dynamic survival analysis with competing risks based on longitudinal data. *IEEE Trans. Biomed. Eng.* **2020**, *67*, 122–133. [CrossRef] [PubMed]
22. Ishwaran, H.; Gerds, T.A.; Kogalur, U.B.; Moore, R.D.; Gange, S.J.; Lau, B.M. Random survival forests for competing risks. *Biostatistics* **2014**, *15*, 757–773. [CrossRef] [PubMed]
23. Brier, G.W. Verification of forecasts expressed in terms of probability. *Mon. Weather. Rev.* **1950**, *78*, 1–3. [CrossRef]
24. Gneiting, T.; Raftery, A.E. Strictly proper scoring rules, prediction, and estimation. *J. Am. Stat. Assoc.* **2007**, *102*, 359–378. [CrossRef]
25. Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K.Q. On calibration of modern neural networks. In Proceedings of the 34th International Conference on Machine Learning, PMLR, Sydney, Australia, 6–11 August 2017; Volume 70, pp. 1321–1330. Available online: https://proceedings.mlr.press/v70/guo17a.html (accessed on 22 September 2026).
26. Van Calster, B.; Nieboer, D.; Vergouwe, Y.; De Cock, B.; Pencina, M.J.; Steyerberg, E.W. A calibration hierarchy for risk models was defined: From utopia to empirical data. *J. Clin. Epidemiol.* **2016**, *74*, 167–176. [CrossRef] [PubMed]
27. Saito, T.; Rehmsmeier, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE* **2015**, *10*, e0118432. [CrossRef] [PubMed]
28. Diebold, F.X.; Mariano, R.S. Comparing predictive accuracy. *J. Bus. Econ. Stat.* **1995**, *13*, 253–263. [CrossRef]
29. Künsch, H.R. The jackknife and the bootstrap for general stationary observations. *Ann. Stat.* **1989**, *17*, 1217–1241. [CrossRef]
30. Politis, D.N.; Romano, J.P. The stationary bootstrap. *J. Am. Stat. Assoc.* **1994**, *89*, 1303–1313. [CrossRef]
31. Sarlin, P. On policymakers' loss functions and the evaluation of early warning systems. *Econ. Lett.* **2013**, *119*, 1–7. [CrossRef]
32. Chernozhukov, V.; Chetverikov, D.; Demirer, M.; Duflo, E.; Hansen, C.; Newey, W.; Robins, J. Double/debiased machine learning for treatment and structural parameters. *Econom. J.* **2018**, *21*, C1–C68. [CrossRef]
33. Wolpert, D.H. Stacked generalization. *Neural Netw.* **1992**, *5*, 241–259. [CrossRef]
34. Tashman, L.J. Out-of-sample tests of forecasting accuracy: An analysis and review. *Int. J. Forecast.* **2000**, *16*, 437–450. [CrossRef]
35. Holm, S. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **1979**, *6*, 65–70. Available online: https://www.jstor.org/stable/4615733 (accessed on 22 September 2026).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.