Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Mitigating Sequence-of-Returns Risk with Regime-Switching Asset Allocation: A Comparison of Econometric and Deep Learning Models in Portfolio Decumulation

Tom Maurer

mail@maurer-tom.de

IU Internationale Hochschule

Christian Müller-Kett

IU Internationale Hochschule

Thomas Zöller

IU Internationale Hochschule

Research Article

Keywords: Regime-switching, Dynamic asset allocation, Sequence-of-returns risk, Econometrics, Deep learning, Stationary bootstrap

Posted Date: November 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11243482/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Mitigating Sequence-of-Returns Risk with Regime-Switching Asset Allocation: A Comparison of Econometric and Deep Learning Models in Portfolio Decumulation

Tom Maurer[1*], Christian Müller-Kett[1] and Thomas Zöller[1]

[1]Data Science, IU Internationale Hochschule, Juri-Gagarin-Ring 152, Erfurt, 99084, Thuringia, Germany.

*Corresponding author(s). E-mail(s): mail@maurer-tom.de;
Contributing authors: christian.mueller-kett@iu.org;
thomas.zoeller@iu.org;

Abstract

Retirement portfolios under regular withdrawals face sequence-of-returns risk: early losses can exhaust capital even when average returns are adequate. We examine whether regime-switching dynamic asset allocation mitigates this risk for a 60/40 stock-bond portfolio, and whether deep learning identifies regimes better than classical econometrics. Five models (a Markov-switching model, a multivariate and a univariate Gaussian hidden Markov model, an LSTM, and an encoder-only Transformer) are compared under identical data, tuning, and walk-forward validation. Their signals, net of transaction costs, feed 10,000 stationary-bootstrap retirement paths in three withdrawal scenarios. At 4.0% the MSM, the univariate HMM, and the LSTM cut the depletion rate from 3.30% to 0.57%, 0.60%, and 0.60% with disjoint Wilson intervals; at 6.0% only the LSTM reduces it (17.01% against 33.04%), the univariate filters raising it to about 49% as their return sacrifice becomes the dominant ruin cause, and the multivariate HMM raising it in every scenario. Only the econometric models turn defensive in all five crisis windows; neither network identifies regimes more accurately, and the LSTM holds the highest median terminal capital by staying invested rather than by timing regimes. Classification quality and search diagnostics do not predict decumulation value. Regime-switching allocation in decumulation is a state-contingent tail hedge: model choice determines the protected risk profile, and above a critical withdrawal rate the premium exceeds the payout.

Keywords: Regime-switching, Dynamic asset allocation, Sequence-of-returns risk, Econometrics, Deep learning, Stationary bootstrap

JEL Classification: G11 , G17 , C45 , C58

1 Introduction

Pay-as-you-go pension systems in aging societies face sustained demographic pressure, and a growing share of retirement income must be financed from individually accumulated capital (Organisation for Economic Co-operation and Development, 2024). Market risk thereby shifts into the decumulation phase, in which households withdraw from a portfolio instead of contributing to it. Under regular withdrawals, portfolio outcomes become path dependent: two return sequences with an identical mean can produce sharply different terminal wealth and ruin probabilities, because losses realized early in retirement interact with continued withdrawals and permanently shrink the capital base from which any recovery must compound (Pfau, 2013; Soni, 2024). This sequence-of-returns risk (SORR) is invisible to mean-variance summary statistics and only partially addressed by withdrawal-side rules such as the 4% rule of Bengen (1994). On the allocation side, the 60/40 stock-bond portfolio is the conventional reference allocation for retirement decumulation. Its risk containment, however, rests on a negative stock-bond correlation that is itself regime dependent: during the 2022 monetary tightening both asset classes lost value simultaneously, and the diversification underlying the 60/40 design failed precisely when it was needed (Šikić, 2025).

A static allocation cannot respond to such state changes by construction; it holds the same risk exposure in a calm bull market and in the early stage of a drawdown that may be irreversible for a decumulation portfolio. Regime-switching models offer a data-driven alternative. Following Hamilton (1989), returns can be described as realizations of a small number of latent market regimes, and regime inference matters for portfolio choice (Ang & Bekaert, 2004). Regime-switching dynamic asset allocation (DAA) operationalizes this idea: a model estimates the probability of a bear regime, and the portfolio moves into a risk-free asset whenever this probability exceeds a threshold. Economically, such a strategy acts as insurance against bear phases arriving early in retirement, paying a premium in transaction costs and missed recoveries in exchange for a thinner ruin tail.

Whether this insurance works, and which model class provides it best, sits at the intersection of three literatures that rarely meet. The SORR literature concentrates on withdrawal rules and partial annuitization while keeping the allocation essentially static (Bengen, 1994; Clare et al., 2021; Soni, 2024). The regime-based allocation literature documents drawdown reduction from regime and trend signals (Li & Mulvey, 2021; Shu et al., 2024), and the deep learning literature nonlinear forecasting gains (Fischer & Krauss, 2018; Taneva-Angelova & Granchev, 2025), but both evaluate on single historical backtests, without withdrawal dynamics or ruin measures. To our knowledge, no study has compared econometric and deep learning regime models under identical conditions using decumulation-native outcome metrics.

This paper addresses that gap along two research questions. *RQ1: Does regime-switching DAA reduce the ruin probability of a 60/40 portfolio in the decumulation phase? RQ2: Do deep learning models identify regime switches more accurately than classical econometric models, and does higher classification quality translate into better decumulation outcomes?* Ruin is measured by the depletion rate, the share of simulated paths on which capital is exhausted, with maximum drawdown (MDD) as a secondary metric.

We compare five regime models under fully identical conditions: a Markov-switching model (MSM), a multivariate Gaussian hidden Markov model (HMM), a univariate HMM included as an ablation, a long short-term memory network (LSTM) (Hochreiter & Schmidhuber, 1997), and an encoder-only Transformer (Vaswani et al., 2017), the latter two trained on Pagan–Sossounov bull/bear labels (Pagan & Sossounov, 2003). All models emit daily bear-regime probabilities for a 60/40 S&P 500/long-term Treasury portfolio over 1990 to 2026 and are validated in a walk-forward design (10-year training and 12-month test windows; 26 folds spanning 25.7 out-of-sample years), with hyperparameters tuned on development folds only. The resulting allocation signals, net of transaction costs, feed a retirement simulation of 10,000 stationary-bootstrap return paths (Politis & Romano, 1994) in each of three withdrawal scenarios, evaluated with Wilson confidence intervals on depletion rates and paired Wilcoxon tests on path-wise outcomes.

Three findings emerge. First, regime-switching DAA reduces ruin risk conditionally on the withdrawal rate rather than on the model class: at 4.0% the univariate econometric filters cut the benchmark's depletion rate by a factor of five to six, at 6.0% they raise it by about half, as their return sacrifice becomes the dominant ruin cause. Second, the two networks separate: the LSTM lowers the depletion rate in all three scenarios and holds the highest median terminal capital, but by staying invested rather than by timing regimes, whereas the Transformer reacts late or not at all in the recent crisis windows and exceeds the benchmark's depletion rate at the two higher rates. Third, classification quality and search diagnostics do not predict decumulation value: the model with the field's best-equal precision-recall ranking and its best search diagnostics ranks second worst on depletion. Regime-switching DAA is therefore a state-contingent tail hedge bounded by the withdrawal rate.

The paper makes four contributions. First, it provides the first systematic comparison of econometric and deep learning regime models in a decumulation setting under identical conditions. Second, it delivers distribution-level evidence, based on 10,000 bootstrap paths per scenario with formal inference, for the withdrawal-rate dependence of the tail-versus-median asymmetry of regime-based allocation. Third, the univariate HMM ablation disentangles architecture effects from feature-set effects and overturns the intuition that richer multivariate inputs improve regime detection; it also shows the two univariate specifications to be one piece of evidence rather than two, agreeing on 99.75% of out-of-sample days. Fourth, the complete pipeline, from data ingestion to evaluation, is released as an open-source artifact for replication and extension.

The remainder of the paper is organized as follows. Section 2 reviews related work on regime identification, deep learning for financial time series, and SORR. Section 3 describes the data, the five models, the walk-forward and tuning protocol, and the

retirement simulation. Section 4 reports classification quality, crisis-window performance, depletion results, and robustness checks. Section 5 discusses mechanisms, implications, and limitations, and Section 6 concludes.

2 Related Work

Econometric regime models form the established reference paradigm for regime-based asset allocation. Building on the Markov-switching framework of Hamilton (1989) and the hidden Markov model with its standard estimation machinery (Rabiner, 1989), a substantial body of work shows that latent bull and bear states can be estimated reliably enough to yield economically exploitable allocation signals (Arimond et al., 2026; Fons et al., 2021; Wang et al., 2020). Collin-Dufresne et al. (2020) find that actively timing volatility and liquidity regimes dominates static strategies, and Erlwein-Sayer et al. (2020) show that filter-based HMM portfolios can beat buy-and-hold benchmarks when time-varying correlations are modeled explicitly. Closest to our allocation design, Shu et al. (2024) report a reduction of the maximum drawdown from 55% to 27% relative to a passive benchmark using regime signals from a statistical jump model. The limits of the paradigm are equally well documented: parametric distributional assumptions, delayed reactions to abrupt nonlinear breaks (Agal et al., 2025; Jung et al., 2022), and unstable transition-probability estimates in turbulent phases (Dua & Tuteja, 2021). Throughout this literature, performance is measured with accumulation-oriented metrics on single historical backtests.

The deep learning literature follows a trajectory from recurrent to attention-based architectures. LSTM networks established the case for deep learning in financial prediction (Fischer & Krauss, 2018), and recent variants combine them with differentiable risk-budgeting layers, reporting drawdown reductions of around 40% against classical methods (Agal et al., 2025). Comparative evidence is mixed, however: Taguchi et al. (2022) find that econometric regime models capture market structure better while recurrent networks predict returns more accurately. Transformer architectures are the current focus. Taneva-Angelova and Granchev (2025) report the lowest volatility-forecasting errors and the strongest generalization across structural breaks for Transformer models, and Song et al. (2025) find that an adaptive Transformer framework outperforms ARIMA and GARCH benchmarks by a wide margin. Two caveats recur: deep models risk memorizing historical idiosyncrasies rather than generalizable regime structure (Kumar et al., 2026; Rahaman, 2026), and results are sensitive to hyperparameter choices (Agal et al., 2025). A smaller hybrid strand combines both paradigms, for example recurrent networks that learn no-trade zones around HMM allocations (Li & Mulvey, 2021), HMM-based allocation within a generative forecasting framework (Oliveira & Costa, 2025), and jump models combined with gradient boosting for crash protection (Shu et al., 2025). These are isolated model combinations, though: systematic comparisons of the econometric and deep learning paradigms under identical data, labels, and backtest protocols remain rare, and none of the cited studies evaluates regime signals in a decumulation setting.

The SORR literature, by contrast, engages directly with the decumulation problem introduced in Section 1 but treats the allocation as essentially given. Its instruments

are withdrawal rules and insurance products: revisions of the 4% rule showing that static withdrawal rates raise ruin risk under changed market conditions (Leggio et al., 2024), dynamic withdrawal schemes that match observed retiree behavior (Kieren & Weber, 2022), and partial annuitization as a buffer against early losses (Soni, 2024). Clare et al. (2021) are an exception on the allocation side, showing that trend-following filters reduce drawdowns and stabilize sustainable withdrawal rates, extended to the full life cycle in Clare et al. (2025). Common to these approaches is that they respond to SORR through time (glide paths), rules (withdrawal rates), or products (annuities), without using the current market state as an input. Jung et al. (2022) provide the only regime-based exception, an HMM framework for surplus optimization in occupational pension plans, which targets institutional funding ratios rather than household ruin risk. Model-based regime identification combined with withdrawal dynamics and formal inference on ruin probabilities is, to our knowledge, absent from this strand.

A final issue cuts across the supervised branch of the comparison: bull and bear regimes have no unique ground truth. Rule-based dating algorithms such as Pagan and Sossounov (2003), which adapt the Bry-Boschan business-cycle dating (Bry & Boschan, 1971) to financial prices, impose minimum phase lengths and amplitudes; threshold-based criteria such as Lunde and Timmermann (2004) trigger regime changes on cumulated returns; business-cycle chronologies like the NBER's date the real economy rather than the market, and even their real-time reproduction is nontrivial (Chauvet & Piger, 2008; Harding & Pagan, 2002). The choice among these constructs determines which crises exist in a supervised training signal at all; minimum-duration filters, in particular the four-month minimum phase used below, can exclude short, V-shaped crashes by construction. This construct dependence is directly relevant for the deep learning results reported in Section 4.

Relative to these strands, the present study differs in three respects. First, it compares five regime models spanning both paradigms under fully identical conditions: the same data, the same label definition for the supervised models, the same walk-forward protocol, tuning budget discipline, and cost assumptions. Second, it evaluates the resulting strategies with decumulation-native metrics, at the level of the outcome distribution and with formal inference, rather than with accumulation metrics on the single realized history. Third, it links classification quality, signal dynamics, and portfolio outcomes within one design, including an ablation that separates architecture effects from feature-set effects. The study thereby connects the regime-allocation, deep learning, and SORR strands whose intersection the literature has left unoccupied.

3 Data and Methodology

3.1 Data

The empirical basis consists of daily adjusted closing prices from Yahoo Finance, retrieved through the `yfinance` interface, version 1.5.1 (Aroussi, 2026), and covering January 2, 1990 to July 31, 2026, or 9,211 trading days after cleaning. The universe is restricted to the two instruments of the benchmark introduced in Section 1, the S&P 500 index (^GSPC) for equity and the Vanguard Long-Term Treasury Fund (VUSTX) for long-duration bonds, with the 13-week Treasury bill rate (^IRX) as the risk-free

rate and the return on cash, and the CBOE Volatility Index (^VIX) and the 10-year Treasury yield (^TNX) entering the feature set only. The feature frame supplied to the models begins on October 16, 1990 and holds 9,012 rows, because the 200-day moving average consumes the earliest observations. The sample spans multiple complete market cycles, including the five crisis windows of Section 3.5; the two-asset design isolates the effect of the regime signal on the allocation decision.

Daily log returns are computed on adjusted closes, $r_t = \ln(A_t/A_{t-1})$. The benchmark return is $r_t^{60/40} = 0.6\, r_t^{\mathrm{GSPC}} + 0.4\, r_t^{\mathrm{VUSTX}}$ with constant weights, and the cash position accrues at $r_t^{\mathrm{cash}} = \ln\left(1 + \mathrm{IRX}_t/(100 \cdot 252)\right)$. From these series, six model features are derived: the 60/40 log returns, a 10-day rolling mean return as momentum, the normalized distance to the 200-day simple moving average, the 20-day rolling standard deviation of returns, the VIX, and the term spread, defined as the 10-year minus the 3-month yield. Pairwise Pearson correlations stay below $|r| = 0.6$ for all feature pairs except realized versus implied volatility ($r = 0.80$), and the term spread is nearly orthogonal to all other features (maximum $|r| = 0.14$). The implications of the collinear volatility block for the multivariate HMM are taken up in Section 4.

Table 1 reports descriptive statistics and augmented Dickey-Fuller (ADF) tests. The return series show the familiar deviations from normality, with excess kurtosis between 4.5 and 10.9 and negative skewness for the equity and portfolio returns, which motivates the drawdown- and downside-oriented evaluation metrics used later. The ADF test rejects a unit root for all return series and the VIX; the two yield levels are non-stationary but enter the feature set only through their difference. Gaps caused by diverging publication calendars of the level series are forward-filled, and two rows without a fill anchor are dropped at the sample start.

Table 1 Descriptive statistics and augmented Dickey-Fuller (ADF) tests of the daily series, January 2, 1990 to July 31, 2026

Series	Mean	SD	Min	Max	Skew.	Ex. kurt.	ADF	p
S&P 500 returns	0.00033	0.01135	−0.1277	0.1096	−0.36	10.86	−17.56	< 0.001
VUSTX returns	0.00021	0.00725	−0.0605	0.0654	−0.03	4.52	−18.72	< 0.001
60/40 returns	0.00028	0.00688	−0.0662	0.0584	−0.28	7.65	−21.01	< 0.001
VIX (level)	19.45	7.74	9.14	82.69	2.21	8.75	−7.32	< 0.001
10Y yield (level)	4.25	1.93	0.50	9.09	0.33	−0.63	−2.34	0.161
3M yield (level)	2.72	2.20	−0.11	7.99	0.19	−1.25	−2.35	0.157

Note: Returns are daily log returns on adjusted closing prices; the VIX is in index points and yields are in percentage points. $N = 9{,}211$ trading days. ADF: test statistic against a 5% critical value of −2.86; the null hypothesis is a unit root. Data: Yahoo Finance.

An automated quality gate governs ingestion. Observed coverage against a weekday calendar is at least 96.2% per series with a maximum gap of one trading day, consistent with exchange holidays rather than data loss; raw missing-value shares stay below 0.5% per series; and the largest daily moves coincide with documented crisis days, supporting the plausibility of the adjustment chain. Yahoo Finance allows

license-free reproduction of the entire pipeline; its residual measurement risk relative to commercial vendors is a limitation (Section 5).

3.2 Regime identification models

All five models solve the same task: at each trading day t, they map the information available up to t to a probability $p_t = P(S_t = \text{bear} \mid \mathcal{F}_t)$ for a binary regime variable $S_t \in \{\text{bull}, \text{bear}\}$. The two-regime space is deliberate. It matches the binary risk-on/risk-off character of the allocation decision in decumulation and avoids the fragmentation of finer regime partitions. Each probability stream is converted into an allocation signal by a model-specific threshold θ: the portfolio switches to the defensive position when $p_t \geq \theta$. The threshold is treated as a hyperparameter and tuned jointly with each model (Section 3.3) rather than fixed at 0.5, which places models with differently calibrated probability scales on an equal footing.

The Markov-switching model follows Hamilton (1989) as a two-state specification with switching mean and switching variance, estimated on the univariate 60/40 return series:

$$r_t^{60/40} = \mu_{S_t} + \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}\left(0, \sigma_{S_t}^2\right), \qquad P(S_t = j \mid S_{t-1} = i) = p_{ij}. \tag{1}$$

The switching variance is essential given the volatility clustering documented in Section 3.1. Within each training window, the bear regime is identified as the state with the larger estimated variance, using training parameters only. Out-of-sample probabilities are filtered probabilities $P(S_t \mid r_1, \ldots, r_t)$ from the forward recursion; smoothed probabilities are never used, so the signal at day t contains no look-ahead.

The hidden Markov model shares the latent first-order Markov chain but emits a multivariate observation vector with Gaussian emissions (Rabiner, 1989):

$$\mathbf{x}_t \mid S_t = k \;\sim\; \mathcal{N}\left(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right), \qquad \mathbf{x}_t = \left(r_t^{60/40},\, \text{VIX}_t,\, \text{Spread}_t\right)', \tag{2}$$

where $\boldsymbol{\Sigma}_k$ is state specific and restricted to be diagonal, selected by the hyperparameter search over full, diagonal, and tied variants. The feature subset combines price dynamics, implied volatility, and the macroeconomic term spread. Estimation uses the Baum-Welch algorithm on observations scaled by median and interquartile range fitted on the respective training window, with ten random restarts per training window, keeping the fit with the highest training likelihood to tame the initialization sensitivity of expectation maximization; the same restart count applies inside the hyperparameter search. The bear state is identified as the state whose training days exhibit the higher realized return volatility under the decoded state sequence, again using training information only, which keeps the rule identical across covariance parametrizations and specifications; filtered inference follows the MSM protocol, with the training window serving as burn-in context for the forward pass on the test window. The univariate HMM ablation is identical in every respect except that the observation vector reduces to the 60/40 return, which is exactly the input space of the MSM; it fixes a state-specific variance and searches only the threshold (Appendix A).

The resulting design separates architecture from information: MSM against univariate HMM, and univariate against multivariate HMM.

The two deep learning models are trained as supervised classifiers on rule-based bull/bear labels generated by the Pagan and Sossounov (2003) dating algorithm from the S&P 500 price path (8-month turning-point window, 4-month minimum phase, 16-month minimum cycle, with the standard 20% amplitude override). Pagan–Sossounov is an ex-post algorithm whose centered window revises recent labels once later prices arrive. Labels are therefore recomputed at each fold cutoff from the history available through it and aligned to the training window, so no test-period price enters a training target; this changes 628 fold-training rows relative to one label series computed on the full sample. Six candidate schemes (MSM- and HMM-implied states, Pagan–Sossounov, peak-to-trough, Lunde–Timmermann, and the NBER chronology) were compared ex post on the full sample, as a descriptive diagnostic rather than as the training construction. The three rule-based schemes agree closely with each other and with the NBER chronology (Pagan–Sossounov $\kappa = 0.56$ on 91% of days, peak-to-trough 0.57 and 91%, Lunde–Timmermann 0.50 and 90%), so the choice within this near-congruent block does not drive the outcome; Pagan–Sossounov was taken from it as the established market-dating algorithm with explicit duration and amplitude criteria. HMM-implied labels were rejected for fragmentation (64%, $\kappa = 0.21$); Appendix B reports the full matrix. Both networks receive the same six features of Section 3.1, scaled by median and interquartile range fitted on the respective training window only; the sequence length is a tuned hyperparameter and differs between them (100 trading days for the LSTM, 220 for the Transformer). Both are trained with a class-weighted binary cross-entropy whose positive-class weight $\sqrt{n_{\text{bull}}/n_{\text{bear}}}$ compensates the scarcity of bear days while damping the overweighting a linear correction would induce (Arimond et al., 2026), and both emit a bear probability that is thresholded exactly like the econometric ones.

The LSTM (Hochreiter & Schmidhuber, 1997) serves as the recurrent baseline, in the role it has established in financial prediction (Fischer & Krauss, 2018): two stacked LSTM layers of 32 and 256 units with dropout of 0.25, followed by a dense sigmoid head. The Transformer implements an encoder-only architecture (Vaswani et al., 2017): a linear projection of the six-dimensional input into an embedding space of dimension 128, sinusoidal positional encoding, four encoder layers with eight attention heads and a feed-forward width of 32, dropout of 0.05, and a linear head on the representation of the final time step. All architecture sizes, window lengths, learning rates, and thresholds are outcomes of the hyperparameter optimization described in Section 3.3; search spaces and final configurations are reported in Appendix A.

3.3 Allocation rule, backtesting, and hyperparameter optimization

The allocation rule is binary. On a defensive signal ($p_t \geq \theta$), the portfolio holds 100% cash accruing r_t^{cash}; otherwise it holds the full 60/40 mix. Signals are executed with a delay of one trading day, so the position on day $t+1$ reflects only information available at t. Each switch incurs an assumed transaction cost of 10 basis points of portfolio value; the sensitivity of all results to this parameter is examined over a range from

0 to 200 basis points in Section 4.4. Out-of-sample evaluation follows a rolling walk-forward protocol (Figure 1): a 10-year training window, the subsequent 12 months as the test window, and a 12-month step, yielding 26 non-overlapping test windows from October 2000 to July 2026 (6,484 trading days, 25.7 years). All five models, including their labels, scalers, and bear-state identification, are re-estimated in every fold on the respective training window (Section 3.2). The deep learning models additionally warm-start from the previous fold's weights for ten epochs and average the predicted probabilities of five independently seeded trainings per fold, which dampens retraining variance; each fold then trains under a model-specific epoch budget (30 for the LSTM, 50 for the Transformer) with early stopping on validation loss at a patience of five.

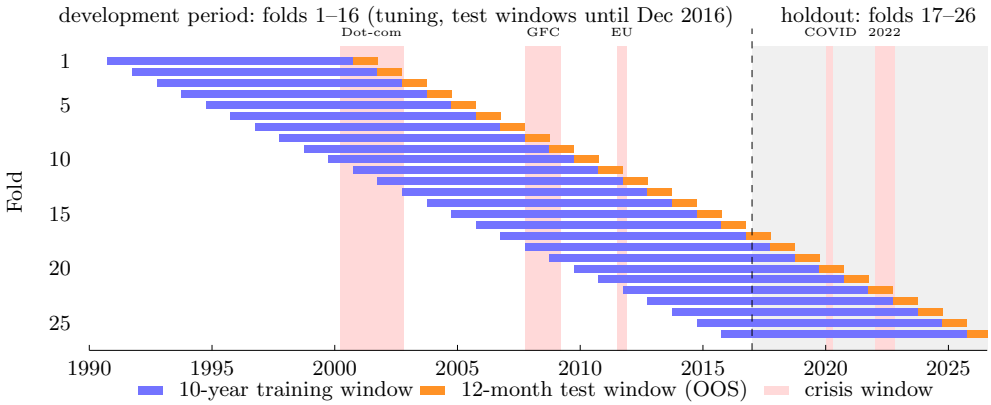


Fig. 1 Walk-forward validation scheme: 26 rolling folds of 10-year training windows and non-overlapping 12-month out-of-sample test windows, October 2000 to July 2026. The hyperparameter search observes only the 16 development folds whose test windows end by December 2016 (dashed line); folds 17 to 26 form the selection-free holdout. Shaded bands mark the five crisis windows of Section 3.5: the dot-com, GFC, and EU debt crisis windows fall into the development period, while the COVID crash and the 2022 rate shock lie in the holdout.

Hyperparameters are tuned with Optuna (Akiba et al., 2019) in an inner cross-validation that reuses the exact fold generator of the final evaluation. The fold structure makes one departure from common practice necessary: 20 of the 26 test windows contain no bear-market day at all and only two a complete Pagan–Sossounov bear phase, so a fold-wise median of any metric is nearly blind to crisis behavior. Each trial therefore concatenates the daily net out-of-sample returns of its folds into one pooled series and maximizes its Martin ratio, the compound annual growth rate divided by the Ulcer index (Martin & McCann, 1989). The concatenation is chronologically continuous: the lagged signal, the portfolio position, and the transaction cost are carried across fold boundaries, so the pooled evaluation is identical to a single backtest on the concatenated out-of-sample signal, verified as bit-identical for all five models. The objective builds its equity path as the exponential of the cumulative pooled log returns, so neither the growth rate nor the Ulcer index inherits a volatility drag from the return convention. Because the Ulcer index integrates both depth and duration of drawdowns,

the rare crisis segments drive the selection. The low-dimensional econometric spaces (the signal threshold, plus the covariance parametrization for the multivariate HMM) are searched exhaustively on a grid of 36, 108, and 36 configurations, so the selected configurations are global optima within their grids. The seven- and eight-dimensional deep learning spaces are searched with a multivariate tree-structured Parzen estimator under a fixed seed, with 300 trials for the LSTM and 400 for the Transformer, of which the first 50 and 60 are random exploration. Trial pruning is disabled: a median pruner would eliminate crisis-protective configurations on the chronologically early, mostly bullish folds. The LSTM study completed 300 trials and the Transformer 400, none pruned. The search observes only folds whose test window ends by December 2016, a development period that includes the dot-com and global financial crises, while the final walk-forward run applies the frozen configurations to all 26 folds. The COVID crash and the 2022 rate shock therefore remain selection-free and test generalization to structurally different crises.

The search is accompanied by four validity diagnostics: a convergence review combining fANOVA hyperparameter importances with an edge-of-range check, with bounds widened and re-run wherever an optimum sat on a boundary; an objective-sensitivity analysis across the candidate metrics (Sharpe, Sortino, Calmar, Ulcer, maximum drawdown); deflated Sharpe ratios (Bailey & López de Prado, 2014) together with the probability of backtest overfitting, estimated by combinatorially symmetric cross-validation (Bailey et al., 2017); and a multi-seed re-evaluation of the top deep learning configurations. Their outcomes are reported alongside the robustness results in Section 4.4; search spaces and final configurations are listed in Appendix A.

3.4 Retirement simulation

The walk-forward backtest yields one realized out-of-sample path per strategy. For SORR, this single path is insufficient: ruin is a tail event whose probability cannot be inferred from one realization. For each strategy, the 25.7-year series of daily net out-of-sample returns (transaction costs included) and the accompanying regime signals are resampled jointly with the stationary bootstrap of Politis and Romano (1994), which concatenates blocks of geometrically distributed length with a mean of 20 trading days. Blocks of on average roughly one calendar month preserve the short-range autocorrelation and the volatility clustering documented in Section 3.1 while still producing diverse multi-decade histories, and the joint resampling of returns and signals keeps the two aligned, which the withdrawal mechanics below require. Per strategy and scenario, 10,000 paths of 7,560 trading days (30 years at 252 trading days) are drawn. All scenario-strategy combinations share one common random stream (common random numbers), which underpins the paired path-wise comparisons of Section 3.5.

Each path evolves a capital account. Capital compounds with the resampled daily return, and every 21 trading days a fixed nominal withdrawal is deducted. An assumed liquidity fee of 0.1% applies to withdrawals taken while the portfolio is invested, reflecting the liquidation of fund shares. Capital is absorbing at zero: once exhausted, a path remains ruined. The three scenarios in Table 2 span the relevant risk space. Standard anchors on the 4.0% withdrawal rate of the safe-withdrawal-rate literature; Aggressive

keeps the same capital at a stressed 6.0% rate; Low Capital combines a reduced capital base with an intermediate 5.0% rate. The 30-year horizon is the standard of that literature (Bengen, 1994; Cooley et al., 1998), which keeps the 4.0% anchor directly comparable and makes ruin observable at plausible withdrawal rates rather than only at an overstretched one. The withdrawals are deliberately non-adaptive: fixed nominal amounts isolate the effect of the allocation from the withdrawal-rule effects studied in the literature of Section 2, and the sensitivity to the withdrawal rate itself is examined in Section 4.4.

Table 2 Withdrawal scenarios of the retirement simulation

Scenario	Initial capital (EUR)	Monthly withdrawal (EUR)	Withdrawal rate (% p.a.)
Standard	500,000	1,666.67	4.0
Aggressive	500,000	2,500	6.0
Low Capital	300,000	1,250	5.0

Note: Fixed nominal withdrawals every 21 trading days over a 30-year horizon (7,560 trading days); the withdrawal rate is stated relative to initial capital. A liquidity fee of 0.1% of the withdrawal applies when the portfolio is invested on the withdrawal date.

The primary outcome of the simulation is the depletion rate, the share of the 10,000 paths on which capital is exhausted before the end of the 30-year horizon. It operationalizes the ruin probability of RQ1 at the level of the outcome distribution rather than of the single historical realization, and it is the object on which the inference of Section 3.5 operates. The second outcome, the path-wise maximum drawdown, is measured on the capital path after withdrawals. A depleted path therefore registers -100% by construction, and the higher the withdrawal rate, the more the mechanical outflow dominates the market component. This decumulation drawdown is not a pure market-risk statistic, which is why it can rank strategies differently from the maximum drawdown of the realized backtest path in Section 4.1. As a robustness check, the entire simulation is repeated with the fixed-length moving block bootstrap of Künsch (1989), which draws blocks of constant length of 20 trading days. Across the 18 scenario-strategy cells, the depletion rate changes by at most 3.88 percentage points between the two resampling schemes, the sign of every regime model's depletion difference against buy-and-hold is preserved in all 15 model-scenario comparisons, and the ranking by depletion rate is identical in two of the three scenarios; the stationary bootstrap remains the primary specification.

3.5 Evaluation framework

The two research questions of Section 1 are formalized as one-sided hypotheses at a significance level of $\alpha = 0.05$. Both are evaluated in all three withdrawal scenarios and reported in full, with the Aggressive scenario retained as the pre-registered headline cell, because the scenario grid spans the range over which the finding changes sign. H1 operationalizes RQ1: regime-switching DAA reduces SORR relative to the buy-and-hold benchmark. It is evaluated on two levels. At the tail, depletion rates are compared

through 95% Wilson confidence intervals (Wilson, 1927), which remain calibrated for proportions near zero where Wald intervals degenerate; disjoint intervals serve as a conservative sufficient criterion for a significant difference. At the distribution body, one-sided paired Wilcoxon signed-rank tests (Wilcoxon, 1945) compare the path-wise maximum drawdowns of each model against buy-and-hold on the 10,000 common-random-number paths of Section 3.4; these are the decumulation drawdowns defined there, not market drawdowns, and the paired design removes the between-history variance component from the comparison. H2 operationalizes RQ2: the attention-based architecture translates its representational capacity into better decumulation outcomes. It is tested by one-sided paired Wilcoxon tests on path-wise terminal capital, comparing the Transformer separately against each competing active strategy (MSM, multivariate and univariate HMM, and LSTM). Nonparametric tests are chosen because terminal wealth and drawdown distributions are heavily skewed and, on ruined paths, carry a point mass at zero.

Beyond the hypothesis tests, the evaluation reports three groups of descriptive metrics. At the backtest level, annualized performance and risk figures (CAGR, volatility, Sharpe ratio (Sharpe, 1966), Sortino ratio (Sortino & van der Meer, 1991), maximum drawdown, Calmar ratio, Ulcer index, and time to recovery for drawdowns exceeding 5%) summarize the single realized out-of-sample path; the Sharpe and Sortino ratios are computed against a zero risk-free rate, since the cash yield already enters each strategy's return series through the allocation rule. Signal-dynamics statistics (number of regime switches, share of whipsaws, defined as reversals within five trading days, and cumulative transaction costs) quantify the economic cost of signal instability. At the classification level, precision, recall, and F1 scores against the NBER recession chronology (National Bureau of Economic Research, 2023) provide a model-independent external reference for regime identification. The chronology enters through the FRED series `USREC` of the Federal Reserve Bank of St. Louis, which is a monthly 0/1 indicator; it is forward-filled to daily frequency and reindexed onto the trading-day calendar, so every recession day counted in Section 4.1 and Appendix C is a forward-filled monthly value, and the positive class is aligned to month boundaries rather than to daily turning points. Together with the fact that the chronology dates the real economy rather than the market (Section 2), this is why it serves as a common yardstick across models rather than as an absolute accuracy measure. At the crisis level, cumulative return and maximum drawdown are evaluated within five historical windows chosen to span distinct crisis morphologies, complemented by the episode-based recovery durations of Section 4.2: the slowly unfolding dot-com bear market (2000 to 2002), the global financial crisis (2007 to 2009), the externally triggered EU debt crisis episode with its fast recovery (2011), the V-shaped COVID crash (2020), and the 2022 rate shock, in which equities and bonds fell jointly. Because crisis metrics are computed on the out-of-sample result path, the dot-com window is truncated at the start of that path, October 16, 2000. The first three windows fall into the development period of the hyperparameter search, whereas COVID and the 2022 episode lie in the selection-free holdout (Section 3.3); crisis behavior in the last two windows therefore measures generalization to unseen crisis types.

4 Results

This section reports the empirical results in the order of the evaluation framework of Section 3.5: classification quality and signal stability first, then performance in the five crisis windows, the distribution-level decumulation outcomes, and finally sensitivity and robustness checks. All figures refer to the 25.7 out-of-sample years of the walk-forward protocol, net of transaction costs.

4.1 Classification accuracy and signal stability

Table 3 reports precision, recall, and F1 of the daily bear signals against the NBER recession chronology, pooled over the 26 test windows. The MSM and the univariate HMM detect 92.2% of recession days at a precision of 0.228. The multivariate HMM pushes recall to 0.985 at precision 0.200, flagging a bear state on 2,886 of the 6,485 classified out-of-sample days against a recession base rate of 9.0%, which is close to a constant-alarm classifier; the Transformer sits in the same corner (recall 0.980, precision 0.243). The LSTM is the only model with a sparse profile, missing 38% of recession days (recall 0.618) at the highest precision of the field (0.329) and the highest F1 (0.430). Two patterns stand out. First, the univariate HMM reproduces the MSM almost cell by cell (identical F1 of 0.366, confusion counts differing by four days) and the two agree on 99.75% of all out-of-sample days: the regime information in the univariate 60/40 return series is architecture-invariant, so the two specifications are one piece of evidence rather than two, and Tables 4 to 7 report them separately only for completeness. Augmenting the observation vector with implied volatility and the term spread, by contrast, degrades classification rather than sharpening it. Second, the low recall of the LSTM is consistent with its training signal: labels that survive the Pagan–Sossounov minimum-phase filter cover only prolonged bear phases, so the network learns those episodes and stays silent otherwise (Section 3.2). The Transformer fails in the opposite direction: its tuned threshold of $\theta = 0.10$ (Table A2) produces near-constant bear signaling, so it is not label-constrained but threshold-degenerate. Threshold-independent ROC and precision-recall curves separate the models differently (Appendix C): the ranking above is partly a thresholding effect.

Table 3 Classification of daily bear signals against the NBER recession chronology, pooled over the 26 out-of-sample test windows

Model	Precision	Recall	F1	TN	FP	FN	TP
MSM	0.228	0.922	0.366	4,071	1,828	46	540
HMM (multivariate)	0.200	0.985	0.332	3,590	2,309	9	577
HMM (univariate)	0.228	0.922	0.366	4,075	1,824	46	540
LSTM	0.329	0.618	0.430	5,162	737	224	362
Transformer	0.243	0.980	0.389	4,110	1,789	12	574

Note: Positive class: NBER recession days. TN, FP, FN, and TP count trading days. The NBER chronology serves as a common external yardstick across models, not as an absolute accuracy measure (Section 3.5).

Table 4 quantifies signal stability, and the sensitivity of the univariate econometric models carries a substantial price: the MSM switches regimes 324 times, 48.9% of its phases are whipsaws, the median phase lasts five trading days, and the per-switch charges cumulate to 32.4 percentage points; the univariate HMM behaves near-identically. Filtered state probabilities of a daily two-state chain contain no intrinsic smoothing mechanism, consistent with the signal instability documented by Dua and Tuteja (2021). The multivariate HMM switches 129 times with a mean phase of 49.9 days, a 40.0% whipsaw share and 12.9 percentage points of cumulated charges, so its near-perfect recall is bought with churn rather than with stability. The LSTM produces the calmest trajectory of the field: 20 switches, no whipsaws, a median phase of 80 days and 2.0 percentage points of charges. The Transformer sits at the opposite extreme, with 52 switches, the highest whipsaw share of all models (50.9%) and a three-day median phase, so the smoothing often attributed to sequence models is a property of the LSTM configuration (100-day window, $\theta = 0.30$) rather than of deep architectures as such. Stability therefore tracks the tuned threshold and the sequence length rather than the model paradigm.

Table 4 Signal stability of the regime models over the out-of-sample period

				Phase (days)		
Model	Switches	Whipsaws	Share (%)	Mean	Median	Costs (pp)
MSM	324	159	48.9	20.0	5	32.4
HMM (multivariate)	129	52	40.0	49.9	6	12.9
HMM (univariate)	318	153	48.0	20.3	5	31.8
LSTM	20	0	0.0	308.8	80	2.0
Transformer	52	27	50.9	122.4	3	5.2

Note: Whipsaws are regime phases shorter than five trading days (Section 3.5); Share is the whipsaw share of all regime phases, that is, of the number of switches plus one. Mean and Median refer to regime-phase lengths in trading days. Costs are the per-switch charges of 10 basis points summed over all switches in the 25.7 out-of-sample years, stated in percentage points; they are a sum of cost rates over time, not a fraction of portfolio value.

Table 5 consolidates both dimensions into the annualized balance of the single realized out-of-sample path. Despite its cost drag, the MSM delivers the best risk-adjusted profile of the field: a Sharpe ratio of 0.78 against 0.55 for buy-and-hold, a maximum drawdown of −10.91% against −35.08%, and an Ulcer index of 3.77 against 9.05. The LSTM is the only active strategy whose CAGR exceeds the benchmark (6.58% against 5.67%), at intermediate risk; the Transformer falls below it at 4.35%, and their identical maximum drawdown of −27.71% reflects that both remain fully invested through the same episode, the 2022 rate shock. The multivariate HMM is the weakest active strategy on return (CAGR 2.95%, Calmar 0.17) and the Transformer the weakest in risk-adjusted terms (Calmar 0.16, Ulcer 7.16), because both pay switching costs without a return offset. Figure 2 adds the time dimension: the LSTM tracks the benchmark through the 2002 and 2022 declines and the Transformer only through 2022, while the benchmark's deepest drawdown occurs in the global financial

crisis. Period-aggregated metrics average crisis and calm phases into a single number; the next subsection therefore disaggregates performance into the five crisis windows.

Table 5 Annualized out-of-sample performance and risk metrics, net of transaction costs, October 2000 to July 2026

Strategy	CAGR	Vol.	Sharpe	Sortino	MDD	Calmar	Ulcer
Buy-and-hold	5.67%	11.16%	0.55	0.72	−35.08%	0.16	9.05
MSM	4.59%	6.03%	0.78	0.87	−10.91%	0.42	3.77
HMM (multivariate)	2.95%	5.50%	0.56	0.57	−17.02%	0.17	5.76
HMM (univariate)	4.56%	6.04%	0.77	0.87	−11.01%	0.41	3.78
LSTM	6.58%	9.53%	0.72	0.87	−27.71%	0.24	6.36
Transformer	4.35%	7.39%	0.61	0.63	−27.71%	0.16	7.16

Note: CAGR: compound annual growth rate; Vol.: annualized volatility; MDD: maximum drawdown; Ulcer: Ulcer index in percent. All metrics are computed on the realized walk-forward path of 6,484 trading days (25.7 years).

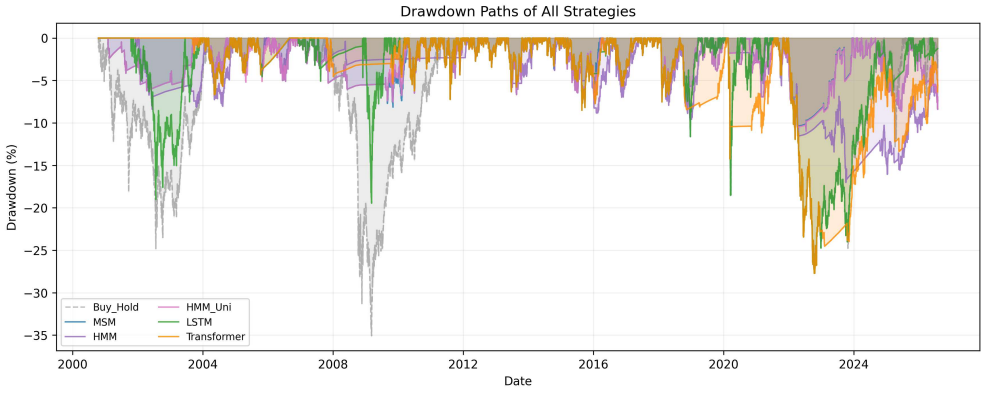


Fig. 2 Out-of-sample drawdown trajectories of all strategies, October 2000 to July 2026. The univariate econometric strategies (MSM, univariate HMM) remain above roughly −12%, while the deep learning strategies incur their deepest drawdowns in the 2022 rate shock.

4.2 Crisis performance

Table 6 reports cumulative return and maximum drawdown within the five crisis windows defined in Section 3.5. The MSM and the univariate HMM produce identical crisis figures because their first defensive signals coincide in all five windows, consistent with the 99.75% signal agreement of Section 4.1; they are therefore discussed jointly.

The three slowly unfolding crises differentiate the models most clearly. In the global financial crisis, every model turns defensive within the first three weeks of October

Table 6 Crisis-window performance: cumulative net return and maximum drawdown within each window

Strategy	Dot-com 2000–2002	GFC 2007–2009	EU debt 2011	COVID 2020	Rate shock 2022
Panel A: Crisis return					
Buy-and-hold	−15.77%	−26.99%	+4.10%	−8.55%	−24.20%
MSM	−2.45%	−2.99%	+2.32%	+0.73%	−6.33%
HMM (multivariate)	−1.60%	−1.39%	+0.01%	+0.73%	−7.66%
HMM (univariate)	−2.45%	−2.99%	+2.32%	+0.73%	−6.33%
LSTM	−4.82%	−2.35%	+4.10%	−8.55%	−24.20%
Transformer	+6.04%	−1.74%	+4.10%	−8.16%	−24.20%
Panel B: Maximum drawdown					
Buy-and-hold	−24.81%	−35.08%	−7.24%	−18.53%	−26.98%
MSM	−6.22%	−4.04%	−4.74%	−1.81%	−8.34%
HMM (multivariate)	−8.40%	−2.80%	0.00%	−1.81%	−8.74%
HMM (univariate)	−6.22%	−4.04%	−4.74%	−1.81%	−8.34%
LSTM	−19.02%	−19.46%	−7.24%	−18.53%	−26.98%
Transformer	−0.08%	−4.56%	−7.24%	−14.25%	−26.98%

Note: Window definitions: dot-com, March 2000 to October 2002; global financial crisis (GFC), October 2007 to March 2009; EU debt crisis, July to November 2011; COVID crash, February to March 2020; rate shock, January to October 2022. Crisis metrics are computed on the out-of-sample path, so the dot-com column effectively starts at its first out-of-sample day, October 16, 2000, rather than in March 2000. The first three windows lie in the development period of the hyperparameter search, COVID and the rate shock in the selection-free holdout (Section 3.3). Identical values for MSM and univariate HMM result from coinciding allocation signals in all five windows.

2007, between 507 and 525 days before the March 2009 trough, and every model beats the benchmark: the MSM and the univariate HMM cap the drawdown at −4.04% against −35.08%, while the multivariate HMM posts the best crisis return of the field (−1.39% against −26.99%), ahead of the Transformer (−1.74%) and the LSTM (−2.35%). Buy-and-hold needs 780 trading days beyond the trough to recover its prior peak. In the dot-com window, all five models first signal on the opening out-of-sample day, October 16, 2000, so the window measures containment rather than detection speed (Section 3.5). The Transformer stays defensive throughout and ends positive (+6.04%, drawdown −0.08%), the best result of the field, which both its low threshold and the window boundary favor; the univariate pair contains the drawdown at −6.22%; the multivariate HMM contains the loss (−1.60%) at a deeper drawdown (−8.40%); and the LSTM re-enters during the decline and gives back most of the protection (−4.82%, drawdown −19.02%). In the 2022 rate shock, the univariate econometric models switch to cash on January 5, 2022, 282 days before the trough, and limit the drawdown to −8.34%; the multivariate HMM follows on January 18 and limits it to −8.74%, while both deep learning models remain fully invested through the entire episode (−24.20% return, −26.98% drawdown, identical to the benchmark).

The two fast episodes expose both sides of the insurance trade. In the COVID crash, all three econometric models turn defensive on February 24, 2020, 23 days before the trough, and convert the benchmark's −18.53% drawdown into a positive crisis return (+0.73%, drawdown −1.81%). The Transformer reacts only on March 13,

five days before the trough, and takes most of the loss; the LSTM produces no signal at all. The EU debt crisis episode shows the premium side, because the recovery falls inside the window: the multivariate HMM spends it entirely in cash (zero drawdown, but +0.01% against the benchmark's +4.10%), the univariate models give up part of the rebound (+2.32%), and the deep learning models, again without a signal, collect the full benchmark outcome.

The timing pattern behind these outcomes is systematic rather than episodic. Only the three econometric models turn defensive in every window, including the two outside the hyperparameter development period, because they respond to shifts in the return distribution itself and require no labeled precedent: leads before the respective trough are 645, 525, 4, 23 and 282 days for the univariate pair and 645, 507, 38, 23 and 269 days for the multivariate HMM, so the two specifications differ in cost and precision rather than in timing. The supervised models protect in the two prolonged, equity-driven bear markets of their development period, exactly the crisis type their training labels contain, and the Transformer additionally, late and partially, in the COVID window. Their silence elsewhere follows the label construction: the COVID crash never enters the label set because the Pagan–Sossounov minimum-phase filter censors a 26-trading-day decline, and the 2022 joint equity-bond decline appears in the labels only after the fact, so no trained pattern was available when it unfolded. The LSTM's failures in the holdout windows are therefore at least partly a label artifact rather than an architecture failure. Whether the observed protection translates into lower ruin probabilities under withdrawals is a distributional question, which the next subsection addresses.

4.3 Sequence-of-returns risk reduction

In the outcome distributions of Section 3.4, ruin over 30 years is material at the Standard scenario's 4.0% rate: buy-and-hold depletes on 3.30% of paths [2.97, 3.67]. The MSM (0.57%), the univariate HMM (0.60%), and the LSTM (0.60%) cut this by a factor of five to six with intervals disjoint from the benchmark; the Transformer (2.97% [2.65, 3.32]) overlaps it; the multivariate HMM raises it to 8.88%. Median terminal capital ranks them differently: LSTM EUR 1,533,695, buy-and-hold 1,040,838, MSM 643,115, univariate HMM 634,032, Transformer 566,457, multivariate HMM 230,545, so the defensive filters pay their crisis insurance in foregone compounding.

Figure 3 places the three scenarios side by side. In the Standard and the Low Capital panel the MSM, the univariate HMM, and the LSTM sit left of the buy-and-hold reference and the multivariate HMM far to its right; in the Aggressive panel only the LSTM remains left of it. The ordering relative to buy-and-hold reverses at the stressed one.

Table 7 quantifies the tail, and the three scenarios do not agree. At the 6.0% Aggressive rate, buy-and-hold depletes on 33.04% of paths and only the LSTM improves on it (17.01%); the MSM, the univariate HMM, the Transformer, and the multivariate HMM all exceed it with disjoint intervals. At the 5.0% Low Capital rate the Standard ordering returns, with the LSTM, the MSM, and the univariate HMM below the benchmark's 14.07% and the Transformer and the multivariate HMM above it. Table 5 identifies the mechanism behind the disagreement: at 6.0% the ruin ordering

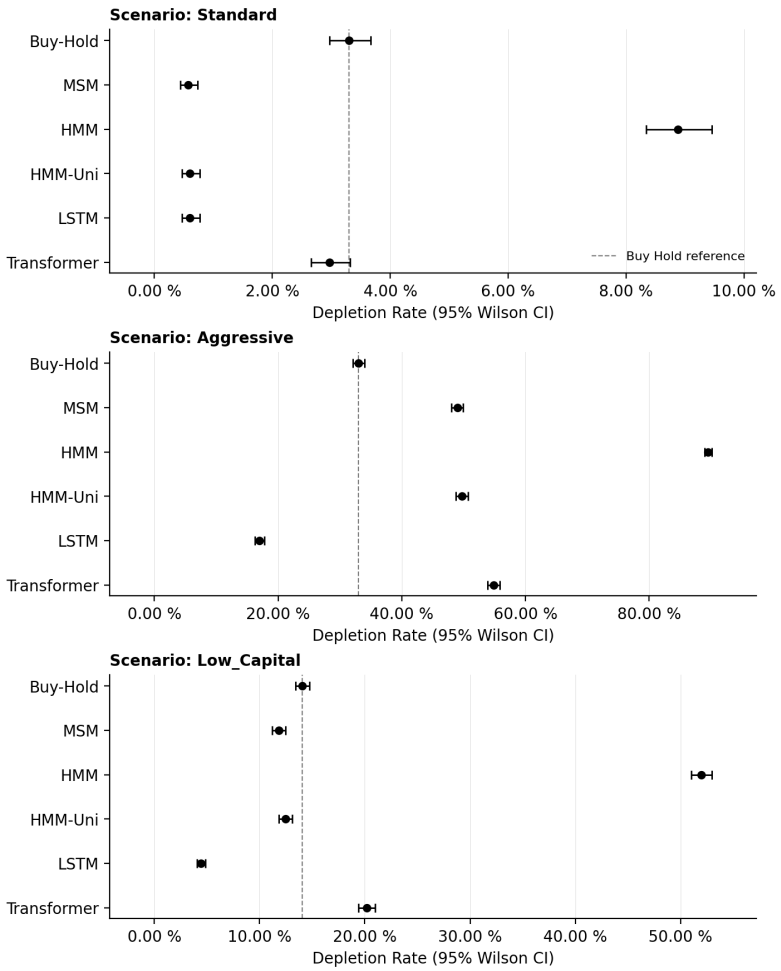


Fig. 3 Depletion rates with 95% Wilson confidence intervals in the three withdrawal scenarios, 10,000 stationary-bootstrap paths per strategy and scenario. The dashed line marks the buy-and-hold reference of the respective panel.

reproduces the CAGR ordering without exception (LSTM 6.58%, buy-and-hold 5.67%, MSM 4.59%, univariate HMM 4.56%, Transformer 4.35%, multivariate HMM 2.95%), whereas at 4.0% it does not. Below a critical withdrawal rate, bracketed between 5.0% and 6.0% on this sample, the variance reduction of a defensive overlay dominates its return sacrifice; above it the sacrifice itself becomes the dominant cause of ruin. On RQ1, H1 at the tail is therefore confirmed for the MSM, the univariate HMM, and the LSTM at 4.0% and 5.0%, at 6.0% for the LSTM alone, and rejected with reversed sign for the multivariate HMM everywhere and for the MSM, the univariate HMM, and the Transformer at 6.0%. The ablation survives in narrower form: the multivariate HMM is worse than the univariate one in every scenario, so adding implied volatility

and the term spread to the observation space destroys rather than dilutes the protection. The wider claim that protection derives from univariate return dynamics does not survive, because the LSTM protects best at 6.0% on six multivariate features.

Table 7 Depletion rates with 95% Wilson confidence intervals, 10,000 stationary-bootstrap paths per scenario and strategy

Strategy	Standard	Aggressive	Low Capital
Buy-and-hold	3.30% [2.97, 3.67]	33.04% [32.12, 33.97]	14.07% [13.40, 14.77]
MSM	0.57% [0.44, 0.74]	48.97% [47.99, 49.95]	11.82% [11.20, 12.47]
HMM (multivariate)	8.88% [8.34, 9.45]	89.54% [88.92, 90.12]	51.99% [51.01, 52.97]
HMM (univariate)	0.60% [0.47, 0.77]	49.73% [48.75, 50.71]	12.46% [11.83, 13.12]
LSTM	0.60% [0.47, 0.77]	17.01% [16.29, 17.76]	4.43% [4.04, 4.85]
Transformer	2.97% [2.65, 3.32]	54.89% [53.91, 55.86]	20.18% [19.40, 20.98]

Note: Depletion rate: share of paths on which capital is exhausted before the end of the 30-year horizon; brackets contain 95% Wilson confidence intervals in percent. Scenario definitions in Table 2.

H2 addresses the opposite end of the distribution, and it is rejected: in all three scenarios the one-sided paired Wilcoxon tests on path-wise terminal capital confirm the Transformer only against the multivariate HMM, not against the MSM, the univariate HMM, or the LSTM. In the headline Aggressive cell the differences in medians are −9,310 EUR against the MSM ($p = 0.951$), −1,657 against the univariate HMM ($p = 0.444$), and −615,996 against the LSTM ($p > 0.999$); against the multivariate HMM both medians are zero and the confirmation ($p < 0.001$) rests on the path-wise differences. The first two are failures to reject, not significant reversals; only against the LSTM is the ordering decisive. The Standard and Low Capital cells repeat the pattern, and the highest median terminal capital in all three scenarios belongs to the LSTM. With Section 4.1, the answer to RQ2 is negative on both margins. The Transformer is not a weak classifier, matching the field's best precision-recall AUC (0.40, Appendix C) and its best search diagnostics (Section 4.4); the failure is in translating ranking quality into allocation value.

The second level of H1, path-wise drawdowns, follows the same rate dependence. Buy-and-hold's median decumulation drawdowns are −37.53%, −62.00%, and −43.41% at 4.0%, 6.0%, and 5.0%. Against them the LSTM improves in all three scenarios (+9.29, +23.07, +11.75 pp), the Transformer only at 4.0% (+4.23, −38.00, −13.17 pp), the multivariate HMM in none (−20.91, −38.00, −56.59 pp), and the MSM and the univariate HMM at 4.0% (+12.63 and +12.23 pp) but not at 6.0% (−36.27 and −37.70 pp). At 5.0% the paired test confirms both univariate filters ($p = 8.8 \times 10^{-7}$ and $p = 0.0228$) although their medians are 2.51 and 3.84 pp deeper than the benchmark's: it compares the path-wise difference distribution under common random numbers, not the marginal medians. Two properties qualify these verdicts. The measure is the decumulation drawdown of Section 3.4, so at 6.0% it approaches a survival test and its ranking there is a perfect rank correlation with the depletion rates above: H1 and the depletion analysis are one finding there, not two. And it is not the realized-path maximum drawdown of Section 4.1, near −11% for the MSM,

which contains no withdrawals. Two conclusions follow. The protection profile is rate-conditional rather than asymmetric across model classes, since at sustainable rates the econometric filters and the LSTM improve both the tail and the typical drawdown, at the stressed rate only the LSTM does and the filters convert their return sacrifice into ruin. And the decisive quantity is the joint effect of variance reduction and return sacrifice. The robustness of these findings to transaction costs, withdrawal rates, and the resampling scheme is examined next.

4.4 Sensitivity and robustness

Figure 4 varies the 10 basis point cost calibration of the headline results. At zero cost the MSM and the univariate HMM end above the benchmark multiple of 4.14, and the crossing lies between 0 and 5 basis points, so already the 10 basis point calibration converts their signal advantage into the wealth gap documented in Table 5 (multiples of 3.18 and 3.16). Their entire return disadvantage is thus a churning artifact, whereas their tail protection at sustainable withdrawal rates, measured net of the same costs, survives it. The LSTM, switching 20 times in 25.7 years, is nearly cost-invariant and stays above the benchmark up to roughly 100 basis points, falling below it by 150; the Transformer, with 52 switches and a 50.9% whipsaw share, is below the benchmark at every cost level including zero, as is the multivariate HMM, whose gap reflects signal quality, not execution costs. A causal signal-smoothing layer, a five-day minimum holding period with a confidence buffer around the threshold, is implemented but disabled in the headline configuration; the reported results are unsmoothed, and the churning of the univariate econometric models is a genuine economic property rather than a post-processing choice.

No strategy exhausts its capital when the realized path is decumulated at 3.5%, 4.0%, and 5.0% with the Standard scenario's initial capital and liquidity fee. The LSTM ends above the benchmark at every rate and the Transformer only at 5.0%, while the ordering between the benchmark and the univariate econometric models flips with the rate: at 3.5% buy-and-hold retains more terminal capital (EUR 855,415 against 740,590 for the MSM), at 5.0% the relation reverses (334,470 against 376,406). On a path without a ruin event, where the return sacrifice cannot cause ruin, drawdown protection gains value with the withdrawal rate; the grid stops below the 6.0% rate at which the distribution reverses, and with one path observed it complements rather than replaces Section 4.3.

The validity diagnostics of the search (Section 3.3) do not follow the paradigm split. The deflated Sharpe ratio is significant at the 5% level for the MSM (0.984), the univariate HMM (0.980), the LSTM (0.979), and the Transformer (0.996), and not for the multivariate HMM (0.862). The probability of backtest overfitting is 0.488 for the MSM, 0.484 for the univariate HMM, 0.159 for the LSTM, and 0.091 for the Transformer, but 0.972 for the multivariate HMM, so the only choice the data flag as overfit is the multivariate HMM's. The multi-seed re-evaluation of the top five deep learning configurations shows the optimism of a best trial without changing the ranking: the LSTM's best Martin ratio of 2.31 becomes 1.67 ± 0.11 across five seeds, the Transformer's 2.53 becomes 2.26 ± 0.31. Retraining stability of the deployed configurations, measured on the five-member seed ensemble of the headline run, is high for the

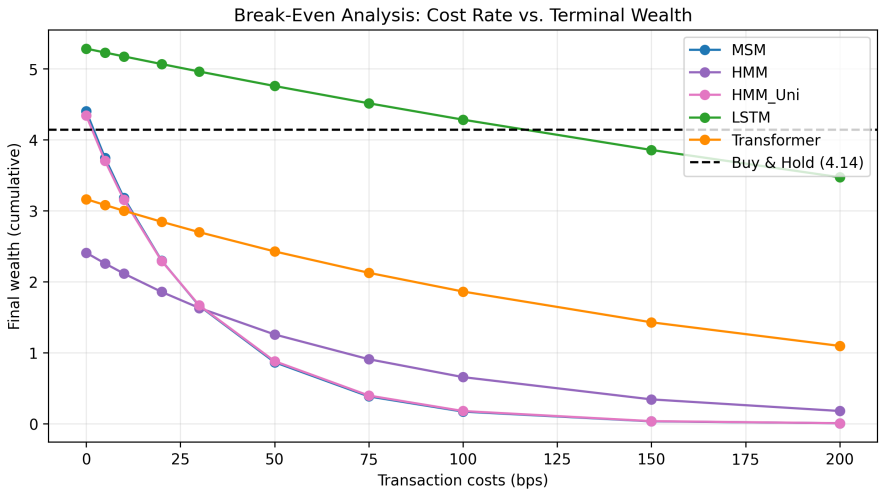


Fig. 4 Break-even analysis: final out-of-sample wealth as a multiple of initial capital under varying transaction-cost levels. The dashed line marks the buy-and-hold multiple of 4.14. Break-even is reported as the lowest grid cost at which a strategy no longer exceeds the benchmark: 5 basis points for the MSM and the univariate HMM, 150 for the LSTM, and 0 for the Transformer and the multivariate HMM, which stay below the benchmark at every cost level.

econometric models (the MSM deterministic, the univariate HMM at a CAGR coefficient of variation of 0.003, the multivariate HMM 0.014) and lower for the networks (LSTM 0.037 with a Martin coefficient of variation of 0.170, Transformer 0.045 and 0.127), so a single deep learning run must not be reported as a point estimate. Under objective sensitivity, only the LSTM shifts its selected configuration with the candidate objective, and the univariate HMM only under the Ulcer index. Taken together, the selection diagnostics do not predict the decumulation ranking: the best-diagnosed model, the Transformer, is second worst on depletion and the worst-diagnosed, the multivariate HMM, is worst, so they identify one failure and miss the other. That is an argument for evaluating regime overlays on decumulation-native outcomes rather than on search-quality statistics.

The resampling scheme, finally, does not drive the distributional findings (Section 3.4): the fixed-length block bootstrap shifts no depletion rate by more than 3.88 percentage points, preserves the sign of every difference against buy-and-hold in all 15 comparisons, and leaves the ranking identical in two of the three scenarios. The hierarchy is therefore robust to the resampling scheme and, for the LSTM, to the cost calibration, but not to the withdrawal rate, which is the substantive result of Section 4.3 rather than a robustness failure. The next section discusses what these regularities imply for regime models in decumulation portfolios.

5 Discussion

Regime-switching DAA in decumulation is a state-contingent tail hedge that the insurance analogy prices once premium and payout are separated. The univariate

econometric filters pay their premium in 32.4 and 31.8 percentage points of switching costs, in recoveries foregone while defensive, and in about one percentage point of CAGR in bull phases, which leaves their median terminal capital in the Standard scenario (4.0%) 38 and 39% below the benchmark's. The payout is rate-conditional: there the MSM and the univariate HMM turn 330 ruined paths per 10,000 into 57 and 60, in Low Capital (5.0%) 1,407 into 1,182 and 1,246, and in Aggressive (6.0%) 3,304 into 4,897 and 4,973. The premium is paid in return and the payout is bounded by the variance reduction that buys it, so above the critical rate of Section 4.3 the policy causes the loss it insures against.

The ablation separates what a plain MSM-versus-HMM comparison confounds: within the HMM architecture the observation space, not the estimation machinery, makes the difference. Two architectures as different as maximum-likelihood Markov-switching and Baum-Welch expectation maximization produce nearly indistinguishable signals, classifications, and decumulation outcomes when fed the same univariate return series, agreeing on 99.75% of out-of-sample days. The same architecture fails once implied volatility and the term spread enter the observation vector, and it fails by over-signaling rather than by delay, consistent with the added variables decoupling the latent state from the portfolio's own return: the multivariate HMM flags a bear state on 44.5% of days against a 9.0% base rate, churns away 12.9 percentage points in costs, and depletes 89.54% of Aggressive paths (Section 4.3). This runs against the expectation that richer conditioning information improves regime detection, and it suggests that reported gains from multivariate specifications, as in Erlwein-Sayer et al. (2020), do not transfer automatically to daily risk-off switching.

The deep learning results are conditional on the Pagan–Sossounov labels, but only one network fails in the way that dependence predicts. The LSTM matches it exactly: sparse, high-precision signaling, 20 switches, and no defensive signal in the EU debt, COVID, and 2022 windows, the three episodes the minimum-phase filter either censors or leaves without a labeled precedent (Section 4.2). That is the memorization failure mode raised for deep financial models (Kumar et al., 2026; Rahaman, 2026) and the asymmetry Taguchi et al. (2022) report in favor of econometric specifications. The Transformer fails in the opposite direction: a tuned threshold of $\theta = 0.10$ leaves it signaling on 36% of out-of-sample days at a 50.9% whipsaw share, churn without timing, so its silence in 2022 sits beside near-constant low-conviction signaling elsewhere. Its constraint is threshold degeneracy under the pooled Martin objective, not label scarcity. The conclusion for RQ2 is accordingly narrow: no evidence emerges that recurrent or attention-based capacity improves regime identification for decumulation allocation, and the LSTM's better outcomes follow from its CAGR and low turnover rather than from regime timing, given its silence in three of five windows; the constraint may still be supervision rather than architecture. Whether threshold-based labels such as Lunde and Timmermann (2004), which admit short crises by construction, unlock the pattern-recognition advantages that Section 2 documents for other tasks is an open empirical question.

A methodological finding cuts across the model comparison: drawdown-based and ruin-based evaluations rank the same strategies differently, even within one experiment. The MSM posts the best realized-path maximum drawdown and the best

depletion rate at 4.0%, but a worse one than the benchmark at 6.0%; the Transformer holds the best search diagnostics, the field's best-equal precision-recall ranking, and the second-worst depletion rate. Maximum drawdown, whether measured on the single history or as a distribution median, does not encode when losses occur relative to withdrawals, how long recovery takes under continued outflows, or whether capital survives at all, which is exactly the information SORR is about, and the decumulation drawdown of Section 3.4 fails in the opposite direction, by absorbing the withdrawal schedule into the risk measure. Single-backtest drawdown reductions of the kind commonly reported in the regime-allocation literature (Clare et al., 2021; Shu et al., 2024) are therefore insufficient evidence for decumulation claims, and so are classification and search-quality statistics (Section 4.4). Depletion rates with formal confidence bounds on resampled paths are a workable minimal standard.

The temporal structure of the results fits the adaptive markets perspective (Lo, 2004), under which a strategy's profitability waxes and wanes with the environment rather than being a fixed property. Between 2012 and 2019, a period without a single crisis window, every regime model pays its premiums without collecting payouts; payouts concentrate in a handful of episodes whose morphology decides which model collects. Slowly unfolding equity bears reward every competent filter, supervised or not; fast and cross-asset shocks were detected only by the unsupervised filters, whose signal is a function of the return distribution rather than of a labeled precedent. The 2022 episode illustrates why this distinction matters: it broke both the stock-bond diversification underlying the static 60/40 design (Šikić, 2025) and the learned crisis patterns of the supervised models, while the return-distribution filters adapted, and still lost the Aggressive scenario on compounding.

The practical recommendation is conditional on the withdrawal plan rather than on a model ranking. At rates the portfolio can sustain, 4.0 to 5.0% on this sample, a univariate econometric filter gives households or providers whose binding concern is ruin economically large and statistically defensible protection, at the price of median wealth and provided execution costs stay below roughly 5 basis points per switch. At an overstretched rate no overlay repairs the plan: the defensive filters make depletion more likely, and the only strategy that helps is the one that stays invested, so the instrument to reach for is the withdrawal rate rather than the allocation (Section 2). Retirees who prioritize expected bequest over ruin protection are better served by the passive benchmark or by a low-turnover overlay of the LSTM type; the Transformer configuration, below the benchmark even at zero cost, is not viable at any retail cost level. The simplicity of the univariate filters also carries regulatory weight: a two-state model on portfolio returns is auditable in a way a seed-ensembled Transformer is not, in retail decumulation products where MiFID II, the EU AI Act, and the GDPR favor inspectable decision logic (Azzutti, 2026) and where acceptance of deep learning fails on absent explainability standards (Khan et al., 2025).

Several limitations bound the conclusions. The data are free Yahoo Finance series, whose measurement risk relative to commercial vendors the ingestion quality gate (Section 3.1) reduces but does not eliminate. All deep learning claims are scoped to the Pagan–Sossounov labeling schema, and the fold design is thin where it matters: of the 26 out-of-sample windows, 2 contain a complete bear phase, 6 overlap one, and 20 no

bear day, so the crisis-window comparisons rest on few and not independent episodes. The design restricts attention to two regimes, one two-asset US portfolio, a rolling rather than expanding walk-forward window, and fixed nominal withdrawals; adaptive withdrawal rules, taxes, and multi-asset universes are excluded by construction, which isolates the allocation effect but limits external validity. Because the withdrawals are never indexed, their real value falls by about 45% over the 30-year horizon at an assumed 2% annual inflation rate, so the simulation understates the pressure a real-terms plan would impose. Because the decumulation drawdown absorbs the withdrawal schedule, the second level of H1 is at the stressed rate closer to a survival test than to an independent criterion. The multivariate HMM configuration carries documented selection uncertainty (overfitting probability 0.972, deflated Sharpe ratio 0.862), whereas the networks carry retraining rather than selection variance (Section 4.4), so a single training run of either is not a point estimate. Finally, the backtest abstracts from trading frictions beyond proportional costs: short-selling constraints, trading halts, and minimum holding periods are not modeled.

6 Conclusion

This paper asked whether regime-switching dynamic asset allocation mitigates sequence-of-returns risk in portfolio decumulation, and whether deep learning identifies the underlying regimes better than classical econometrics. RQ1 is answered affirmatively, conditional on the withdrawal rate rather than on the model class. At 4.0% the MSM, the univariate HMM, and the LSTM cut the depletion rate from 3.30% to 0.57%, 0.60%, and 0.60% with disjoint confidence intervals; at 5.0% the same three stay below the benchmark; at 6.0% only the LSTM does, while the two univariate econometric filters raise it from 33.04% to about 49% as their return sacrifice becomes the dominant ruin cause; the multivariate HMM raises it in all three scenarios. The same conditionality holds on the second level, path-wise decumulation drawdown: four of five models improve on the benchmark at 4.0%, one at 6.0%. RQ2 is answered negatively on both margins. The supervised models do not rank regimes more accurately than the univariate filters (Appendix C), the LSTM produces no defensive signal in three of five crisis windows and the Transformer in two, and the Transformer loses the terminal-capital comparison to the MSM, the univariate HMM, and the LSTM in every scenario, so H2 is rejected. The best decumulation outcomes belong to the LSTM, which obtains them by staying invested at the highest CAGR of the field rather than by timing regimes, a different claim from the one RQ2 asks about.

Two lessons follow. First, the univariate return filters buy ruin insurance with unstable, transaction-intensive signals, and their viability stands or falls with execution costs near 5 basis points per switch. The LSTM overlay offers a stable, cost-robust signal, the highest median terminal capital in every scenario and the only depletion improvement at the stressed rate, but no crisis signal at all in three of five windows. The Transformer combines the best search diagnostics of the field with the highest whipsaw share and the second-worst depletion rate. Enriching the information set makes the econometric model worse. Choosing a model is choosing which part of the outcome distribution to protect, and for the retiree whose binding constraint is not

outliving their capital at a sustainable withdrawal rate, the empirically supported choice is the simplest model in the field: a two-state filter on the portfolio's own returns. Second, neither classification metrics, search diagnostics, nor backtest drawdowns are sufficient statistics for decumulation value. The highest F1 belongs to the model with the weakest probability ranking (LSTM, 0.430 at ROC AUC 0.76) and the best search diagnostics to the model with the second-worst depletion rate (Transformer, PBO 0.091); the best realized-path drawdown coexists with a collapsed median drawdown under resampling at the stressed rate; and only the depletion rate, measured on the outcome distribution with formal confidence bounds across withdrawal rates, separates the strategies in the dimension SORR is about.

The research agenda follows from the limitations. First, sequence-of-returns risk deserves its own outcome metric: a continuous measure weighting losses by their timing relative to the withdrawal schedule would grade strategies between the coarse endpoints of drawdown and ruin and separate the market from the withdrawal component, which the decumulation drawdown used here conflates. Second, the label-artifact explanation of the deep learning failures is testable: retraining the same architectures on threshold-based labels such as Lunde and Timmermann (2004), which admit short crises by construction, or on multi-class partitions would separate supervision from architecture effects. Pretrained time-series foundation models (Ansari et al., 2024) offer a third route that bypasses labeling altogether. Beyond the model layer, the framework extends naturally to multi-asset universes, inflation-indexed and adaptive withdrawal rules, and an expanding-window walk-forward variant, and a replication on commercial data would remove the residual measurement risk of the free data used here. A further direction is the operational resilience of deployment: latency, throughput, and the cost of retraining and inference, considerable for the seed-ensembled Transformer, decide whether a regime overlay is viable in production (Koganti & Anne, 2026). The complete pipeline is available as an open-source artifact (Maurer, 2026).

Declarations

Funding. The authors did not receive support from any organization for the submitted work.

Competing interests. The authors have no competing interests to declare that are relevant to the content of this article.

Ethics approval and consent to participate. Not applicable. The study uses publicly available financial market data and does not involve human participants or animals.

Consent for publication. Not applicable.

Data availability. All input data are publicly available: the price and yield series from Yahoo Finance, and the NBER recession chronology used as the external classification reference from the FRED database of the Federal Reserve Bank of St. Louis (series `USREC`, https://fred.stlouisfed.org/series/USREC). The published pipeline retrieves and prepares both automatically. No proprietary data are used.

Materials availability. Not applicable.

Code availability. The complete pipeline, including configuration, model implementations, and evaluation code, is available at https://github.com/regime-switching-daa/regime-switching-daa-paper.

Author contributions. T.M. developed the research design, implemented the pipeline, performed the analyses, and wrote the original draft. C.M.-K. supervised the work and contributed to the research design and the revision of the manuscript. T.Z. contributed to the revision of the manuscript. All authors read and approved the final manuscript.

Appendix A Hyperparameter search spaces and selected configurations

Table A1 lists the per-model search spaces of the hyperparameter optimization described in Section 3.3. The econometric spaces are searched exhaustively, so the selected configurations are global optima within their grids; the deep learning spaces are searched with a multivariate tree-structured Parzen estimator under a fixed seed. The number of training epochs is not a search dimension: trials run under a 150-epoch ceiling with early stopping on validation loss, which removes a redundant dimension; the final run uses the shorter budgets of Section 3.3.

Table A2 reports the configurations deployed in the final walk-forward run together with the pooled development-period Martin ratio of each selected configuration.

Appendix B Regime label concordance

Figure B1 reports the pairwise Cohen's κ between the six candidate labeling schemes of Section 3.2, computed ex post on daily bull/bear classifications of the full sample; the supervised models' training labels are recomputed causally at each fold cutoff and are not the series compared here. The three rule-based schemes (Pagan–Sossounov, peak-to-trough, Lunde–Timmermann) form a nearly congruent block ($\kappa \geq 0.93$); the HMM-implied labels show the lowest concordance with every other scheme, which motivated their exclusion as a training target.

Table A1 Hyperparameter search spaces

Model	Parameter	Search space
MSM	threshold θ	0.100 to 0.975, step 0.025 (grid, 36 configurations)
HMM (multivariate)	covariance type	{full, diagonal, tied}
	threshold θ	0.100 to 0.975, step 0.025 (grid, 108 configurations)
HMM (univariate)	threshold θ	0.100 to 0.975, step 0.025 (grid, 36 configurations)
LSTM	window size	20 to 250, step 10
	units (layer 1)	{16, 32, 64, 128}
	units (layer 2)	{32, 64, 128, 256}
	batch size	{32, 64, 128}
	learning rate	10^{-5} to 10^{-2}, log scale
	dropout	0.00 to 0.60, step 0.05
	threshold θ	0.10 to 0.90, step 0.05
Transformer	window size	20 to 250, step 10
	(d_{model}, heads)	{16-2, 32-2, 32-4, 64-4, 64-8, 128-4, 128-8}
	encoder layers	1 to 4
	feed-forward width	{32, 64, 128, 256}
	batch size	{32, 64, 128}
	learning rate	10^{-5} to 10^{-2}, log scale
	dropout	0.00 to 0.50, step 0.05
	threshold θ	0.10 to 0.90, step 0.05

Note: Deep learning spaces are sampled with a multivariate tree-structured Parzen estimator (seed 42) after 50 (LSTM) and 60 (Transformer) random startup trials; the covariance parametrization is not searched for the univariate HMM: for one-dimensional Gaussian emissions the full, diagonal, and spherical parametrizations are equivalent state-specific variances, while the tied variant imposes one variance shared across states; the ablation fixes the state-specific variant and searches only the threshold. The (d_{model}, heads) pair is encoded as a single categorical of valid combinations so that no trials are spent on divisibility violations.

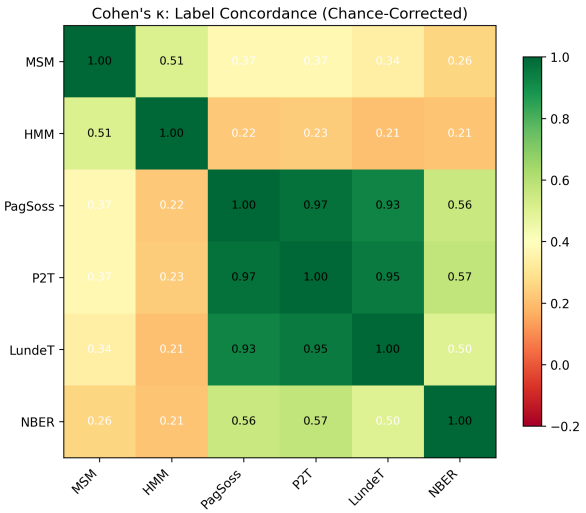


Fig. B1 Chance-corrected label concordance (Cohen's κ) between the six candidate labeling schemes: MSM- and HMM-implied states, Pagan–Sossounov (PagSoss), peak-to-trough (P2T), Lunde–Timmermann (LundeT), and the NBER recession chronology.

Table A2 Selected configurations of the final walk-forward run

Model	Selected configuration	Martin (dev.)
MSM	$\theta = 0.175$	1.64
HMM (multivariate)	diagonal covariance, $\theta = 0.975$	0.89
HMM (univariate)	$\theta = 0.175$	1.63
LSTM	window 100, units 32/256, batch 64, learning rate 2.44×10^{-5}, dropout 0.25, $\theta = 0.30$	2.31
Transformer	window 220, $d_{\text{model}} = 128$, 8 heads, 4 layers, feed-forward width 32, batch 128, learning rate 1.11×10^{-5}, dropout 0.05, $\theta = 0.10$	2.53

Note: Martin (dev.): Martin ratio of the chronologically continuous pooled out-of-sample return path (Section 3.3) over the development folds. The reported configurations are the Martin-optimal configurations selected by the final hyperparameter search.

Appendix C Threshold-independent classification diagnostics

Figures C1 and C2 complement Table 3 with curves computed before thresholding. Ranking quality is close across four of the five models: MSM 0.89 ROC AUC and 0.40 PR AUC, univariate HMM 0.88 and 0.39, Transformer 0.88 and 0.40, multivariate HMM 0.87 and 0.36, LSTM 0.76 and 0.30. The F1 ordering of Section 4.1 is therefore largely a thresholding effect: the LSTM holds the highest F1 on the weakest probability ranking, and the Transformer's tuned $\theta = 0.10$ spends its best-equal precision-recall ranking on near-constant signaling.

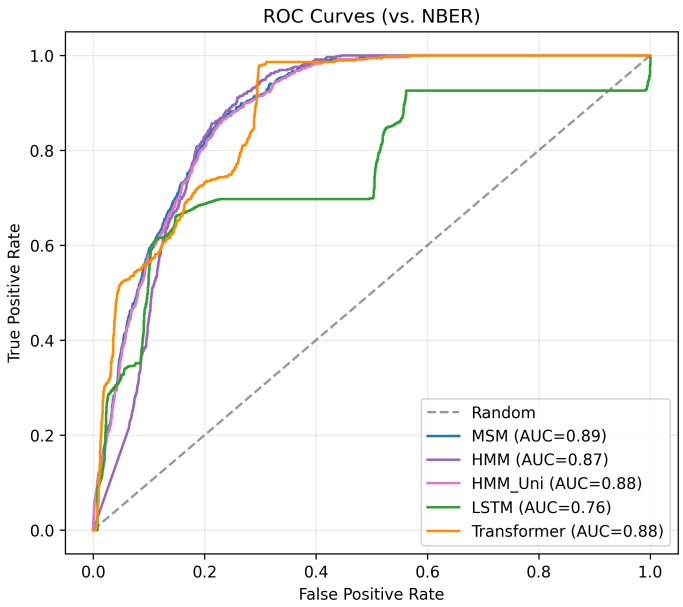


Fig. C1 ROC curves of the daily bear probabilities against the NBER recession chronology, pooled over the 26 out-of-sample test windows.

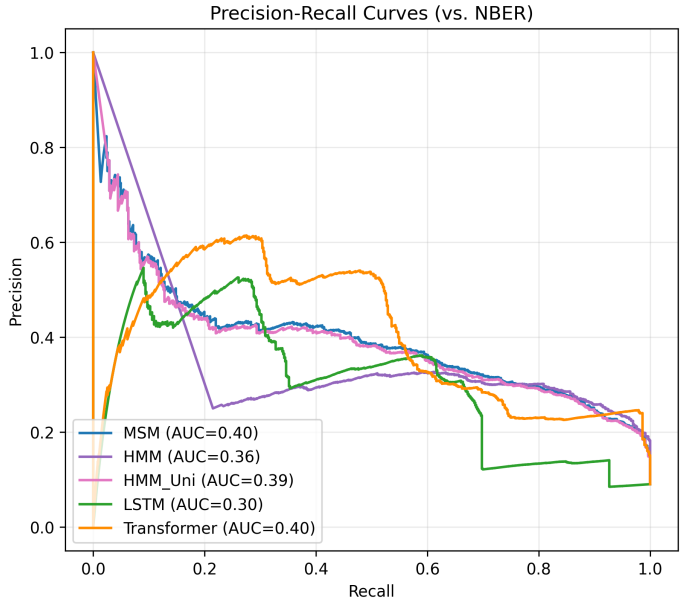


Fig. C2 Precision-recall curves of the daily bear probabilities against the NBER recession chronology, pooled over the 26 out-of-sample test windows.

References

Agal, S., Raulji, K., Odedra, N.D. (2025). A machine learning approach to risk based asset allocation in portfolio optimization. *Scientific Reports*, *15*(1), 42263, https://doi.org/10.1038/s41598-025-26337-x

Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 2623–2631). (https://doi.org/10.1145/3292500.3330701)

Ang, A., & Bekaert, G. (2004). How regimes affect asset allocation. *Financial Analysts Journal*, *60*(2), 86–99, https://doi.org/10.2469/faj.v60.n2.2612

Ansari, A.F., Stella, L., Türkmen, C., Zhang, X., Mercado, P., Shen, H., ... others (2024). *Chronos: Learning the language of time series.* Transactions on Machine Learning Research. (https://openreview.net/forum?id=gerNCVqqtR)

Arimond, A., Borth, D.S., Garcia-Vega, S., Harjoto, M., Hoepner, A.G., Klawunn, M., Weisheit, S. (2026). Neural networks and value at risk in asset management. *Review of Quantitative Finance and Accounting*, *67*(1), 277–316, https://doi.org/10.1007/s11156-025-01460-y

Aroussi, R. (2026). *yfinance.* (Software, https://pypi.org/project/yfinance/)

Azzutti, A. (2026). AI governance after MiFID II: beyond (mere) technological neutrality? *ERA Forum*, *27*(1), 7–31, https://doi.org/10.1007/s12027-026-00871-1

Bailey, D.H., Borwein, J., López de Prado, M., Zhu, Q.J. (2017). The probability of backtest overfitting. *The Journal of Computational Finance*, *20*(4), 39–69, https://doi.org/10.21314/JCF.2016.322

Bailey, D.H., & López de Prado, M. (2014). The deflated Sharpe ratio: Correcting for selection bias, backtest overfitting, and non-normality. *The Journal of Portfolio Management*, *40*(5), 94–107, https://doi.org/10.3905/jpm.2014.40.5.094

Bengen, W.P. (1994). Determining withdrawal rates using historical data. *Journal of Financial Planning*, *7*(4), 171–180,

Bry, G., & Boschan, C. (1971). *Cyclical analysis of time series: Selected procedures and computer programs.* New York: National Bureau of Economic Research. (https://www.nber.org/books-and-chapters/cyclical-analysis-time-series-selected-procedures-and-computer-programs)

Chauvet, M., & Piger, J. (2008). A comparison of the real-time performance of business cycle dating methods. *Journal of Business & Economic Statistics*, *26*(1), 42–49, https://doi.org/10.1198/073500107000000296

Clare, A., Seaton, J., Smith, P.N., Thomas, S.H. (2021). Can sustainable withdrawal rates be enhanced by trend following? *International Journal of Finance & Economics*, *26*(1), 27–41, https://doi.org/10.1002/ijfe.1774

Clare, A., Seaton, J., Smith, P.N., Thomas, S.H. (2025). Saving to decumulate: A lifetime journey based on the perfect contribution rate. *The Journal of Retirement*, *13*(1), 28–45, https://doi.org/10.3905/jor.2025.1.180

Collin-Dufresne, P., Daniel, K., Sağlam, M. (2020). Liquidity regimes and optimal dynamic asset allocation. *Journal of Financial Economics*, *136*(2), 379–406, https://doi.org/10.1016/j.jfineco.2019.09.011

Cooley, P.L., Hubbard, C.M., Walz, D.T. (1998). Retirement savings: Choosing a withdrawal rate that is sustainable. *AAII Journal*, *20*(2), 16–21,

Dua, P., & Tuteja, D. (2021). Regime shifts in the behaviour of international currency and equity markets: A Markov-switching analysis. *Journal of Quantitative Economics*, *19*(S1), 309–336, https://doi.org/10.1007/s40953-021-00273-9

Erlwein-Sayer, C., Grimm, S., Ruckdeschel, P., Sass, J., Sayer, T. (2020). Filter-based portfolio strategies in an HMM setting with varying correlation parametrizations. *Applied Stochastic Models in Business and Industry*, *36*(3), 307–334, https://doi.org/10.1002/asmb.2491

Fischer, T., & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research*, *270*(2), 654–669, https://doi.org/10.1016/j.ejor.2017.11.054

Fons, E., Dawson, P., Yau, J., Zeng, X., Keane, J. (2021). A novel dynamic asset allocation system using feature saliency hidden Markov models for smart beta

investing. *Expert Systems with Applications*, *163*, 113720, https://doi.org/10.1016/j.eswa.2020.113720

Hamilton, J.D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, *57*(2), 357–384, https://doi.org/10.2307/1912559

Harding, D., & Pagan, A. (2002). Dissecting the cycle: A methodological investigation. *Journal of Monetary Economics*, *49*(2), 365–381, https://doi.org/10.1016/S0304-3932(01)00108-8

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

Jung, H., Lee, Z., Cheong, D.Y. (2022). Surplus optimization in defined benefit pensions using the regime-switching model: Occupational pension plans in South Korea. *Asia-Pacific Journal of Financial Studies*, *51*(5), 709–734, https://doi.org/10.1111/ajfs.12396

Khan, F.S., Mazhar, S.S., Mazhar, K., AlSaleh, D.A., Mazhar, A. (2025). Model-agnostic explainable artificial intelligence methods in finance: a systematic review, recent developments, limitations, challenges and future directions. *Artificial Intelligence Review*, *58*(8), 232, https://doi.org/10.1007/s10462-025-11215-9

Kieren, P., & Weber, M. (2022). When saving is not enough: Wealth decumulation in retirement. *Journal of Pension Economics and Finance*, *21*(3), 446–473, https://doi.org/10.1017/S1474747221000196

Koganti, H., & Anne, S. (2026). Operational excellence through AI and ML in financial services: A comprehensive review of applications, challenges, and future directions. *International Journal of Computational Intelligence Systems*, *19*(1), 181, https://doi.org/10.1007/s44196-025-01110-0

Kumar, P.N., Umeorah, N., Alochukwu, A. (2026). Regime-dependent graph neural networks for enhanced volatility prediction in financial markets. *Mathematics*, *14*(2), 289, https://doi.org/10.3390/math14020289

Künsch, H.R. (1989). The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, *17*(3), 1217–1241, https://doi.org/10.1214/aos/1176347265

Leggio, K.B., Lien, D., Dai, Y. (2024). The death of the 4% rule: New market conditions and retiree needs require rethinking retirement spending. *The Journal of Retirement*, *11*(4), 40–60, https://doi.org/10.3905/jor.2024.1.154

Li, X., & Mulvey, J.M. (2021). Portfolio optimization under regime switching and transaction costs: Combining neural networks and dynamic programs. *INFORMS Journal on Optimization*, *3*(4), 398–417, https://doi.org/10.1287/ijoo.2021.0053

Lo, A.W. (2004). The adaptive markets hypothesis: Market efficiency from an evolutionary perspective. *The Journal of Portfolio Management*, *30*(5), 15–29, https://doi.org/10.3905/jpm.2004.442611

Lunde, A., & Timmermann, A. (2004). Duration dependence in stock prices: An analysis of bull and bear markets. *Journal of Business & Economic Statistics*, *22*(3), 253–273, https://doi.org/10.1198/073500104000000136

Martin, P.G., & McCann, B.B. (1989). *The investor's guide to Fidelity funds.* New York: John Wiley & Sons.

Maurer, T. (2026). *Regime-switching DAA: Complete research pipeline.* (Software, https://github.com/regime-switching-daa/regime-switching-daa-paper)

National Bureau of Economic Research (2023). *US business cycle expansions and contractions.* Business Cycle Dating Committee, National Bureau of Economic Research. (https://www.nber.org/research/data/us-business-cycle-expansions-and-contractions)

Oliveira, M., & Costa, G. (2025). Quantitative portfolio optimization framework with market regimes classification, probabilistic time series forecasting, and hidden Markov models. *Digital Finance*, *7*(3), 553–603, https://doi.org/10.1007/s42521-025-00153-4

Organisation for Economic Co-operation and Development (2024). *OECD pensions outlook 2024: Improving asset-backed pensions for better retirement outcomes and more resilient pension systems.* Paris: OECD Publishing.

Pagan, A.R., & Sossounov, K.A. (2003). A simple framework for analysing bull and bear markets. *Journal of Applied Econometrics*, *18*(1), 23–46, https://doi.org/10.1002/jae.664

Pfau, W.D. (2013). *The lifetime sequence of returns: A retirement planning conundrum.* SSRN. (https://doi.org/10.2139/ssrn.2544637)

Politis, D.N., & Romano, J.P. (1994). The stationary bootstrap. *Journal of the American Statistical Association*, *89*(428), 1303–1313, https://doi.org/10.1080/01621459.1994.10476870

Rabiner, L.R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257–286, https://doi.org/10.1109/5.18626

Rahaman, S.U. (2026). Dynamic neuroplastic networks for financial decision making: A self-adaptive approach for mitigating catastrophic forgetting in continual learning. *Computational Economics*, *68*(1), 457–474, https://doi.org/10.1007/s10614-025-11057-1

Sharpe, W.F. (1966). Mutual fund performance. *The Journal of Business*, *39*(1), 119–138, https://doi.org/10.1086/294846

Shu, Y., Yu, C., Mulvey, J.M. (2024). Downside risk reduction using regime-switching signals: A statistical jump model approach. *Journal of Asset Management*, *25*(5), 493–507, https://doi.org/10.1057/s41260-024-00376-x

Shu, Y., Yu, C., Mulvey, J.M. (2025). Dynamic asset allocation with asset-specific regime forecasts. *Annals of Operations Research*, *346*(1), 285–318, https://doi.org/10.1007/s10479-024-06266-0

Šikić, L. (2025). The breakdown of the classic portfolio hedge: A Markov-switching analysis of the US stock-bond correlation. *Notitia*, *11*(1), 97–111, https://doi.org/10.32676/n.11.1.7

Song, Z., Tsang, H.S.-H., Hsung, R.T.-C., Zhu, Y., Lo, W.-L. (2025). From market volatility to predictive insight: An adaptive transformer-RL framework for sentiment-driven financial time-series forecasting. *Forecasting*, *7*(4), 55, https://doi.org/10.3390/forecast7040055

Soni, A. (2024). The role of annuities in managing sequence of returns risk approaching and in retirement. *The Journal of Retirement*, *11*(4), 26–39, https://doi.org/10.3905/jor.2023.1.147

Sortino, F.A., & van der Meer, R. (1991). Downside risk. *The Journal of Portfolio Management*, *17*(4), 27–31, https://doi.org/10.3905/jpm.1991.409343

Taguchi, R., Watanabe, H., Sakaji, H., Izumi, K., Hiramatsu, K. (2022). Constructing equity investment strategies using analyst reports and regime switching models. *Frontiers in Artificial Intelligence*, *5*, 865950, https://doi.org/10.3389/frai.2022.865950

Taneva-Angelova, G., & Granchev, D. (2025). Deep learning and transformer architectures for volatility forecasting: Evidence from US equity indices. *Journal of Risk and Financial Management*, *18*(12), 685, https://doi.org/10.3390/jrfm18120685

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., ... Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems* (Vol. 30, pp. 5998–6008).

Wang, M., Lin, Y.-H., Mikhelson, I. (2020). Regime-switching factor investing with hidden Markov models. *Journal of Risk and Financial Management*, *13*(12), 311, https://doi.org/10.3390/jrfm13120311

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, *1*(6), 80–83, https://doi.org/10.2307/3001968

Wilson, E.B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, *22*(158), 209–212, https://doi.org/10.1080/01621459.1927.10502953