Realised Volatility Forecasting using Neural Networks and Statistically Disciplined Model Memory

A research essay presented in part fulfilment of the requirements
for the degree of Bachelor of Commerce (Honours)
in the Department of Accounting and Finance at The University of
Auckland.

Rickey Lee

Supervised by Dr. Justin Case

2025

Acknowledgements

I would like to express my sincere gratitude to all those who supported the completion of this dissertation throughout the most challenging, yet the most rewarding year of my academic journey.

First and foremost, I am deeply grateful to my supervisor, Dr. Justin Case, for invaluable guidance, constructive feedback, and constant encouragement throughout this research. His expertise and patience has been instrumental to the development of this work.

I would also like to thank the academic staff at the Faculty of Business and Economics, whose teaching and mentorship have shaped my research and intellect. Specifically, I would like to thank Professor Henk Berkman for early guidance in crafting the research proposal, as well as Professor Dimitris Margaritis, Professor Steven Cahan, Associate Professor Ryan Greenaway-Mcgrevy, Dr. Xing Han, Dr. John Lee, and Dr. Gertjan Verdickt for their teaching and support throughout the year.

Finally, I extend my heartfelt appreciation to the unwavering support and encouragement from my family, friends, and partner, whose belief in my has made this achievement possible and worthwhile.

Contents

List of Tables

List of Figures

Abstract

This paper studies whether a statistically disciplined choice of lags (its “memory”) improves one-day-ahead forecasts of realised volatility in both linear and nonlinear neural network (NN) machine learning (ML) model candidates. Specifically, I consider the disciplined choice of lags in the Heterogeneous Autoregressive (HAR) model, the Multilayer Perceptron (MLP) neural network, and the Long Short-Term Memory (LSTM) neural network. Using daily realised volatility (RV) for S&P 500 constructed from high-frequency trades, I compare the standard set of lags given by the heterogeneous investor theory, compared to a statistically disciplined set of lags chosen by minimising a Bayesian Information Criterion (BIC). I evaluate out-of-sample (OOS) forecasts by mean squared error (MSE) and quasi-likelihood (QLIKE) losses and compare models with Diebold-Mariano (DM) tests, showing improvements with statistical significance at the 1% level. The LSTM_{BIC} attains the lowest QLIKE as well as the lowest MSE. Gains are concentrated in the Global Financial Crisis and COVID-19 episodes, when autoregressive dependence lengthens. Realised utility (RU) calculations show systematic economic value, corresponding to certainty-equivalent improvements of 0.27 to 1.70 basis points, closing up to 5.2% of the residual gap to the theoretical upper bound for a mean-variance investor. Findings remain robust to different sectors, and alternative information criteria.

Chapter 1

Introduction

Forecasting the variance of equity returns is central to risk management, derivatives pricing, and portfolio allocation, and therefore has been of interest to both financial practitioners and academics. A key empirical challenge lies in determining how far back the model should look, or equivalently, how to define its effective memory. In stable markets, short memory designs often suffice, yet during crises, volatility persistence lengthens and fixed-lag models tend to under-react. Neural networks (NN) address nonlinearity in volatility dynamics but often rely on ad hoc choices of memory, exposing them to overfitting and poor generalisation. This study proposes a statistically disciplined framework that links model memory directly to a data-driven lag selection process. At each re-estimation date, the autoregressive (AR) order that minimises the Bayesian Information Criterion (BIC) is selected on a rolling validation sample and then mapped into inputs for both linear and neural network models before generating out of sample (OOS) forecasts. This procedure allows the model's memory to adapt to changing dependence structures while preserving transparency.

Using daily realised volatility (RV) for the S&P 500 constructed from high-frequency trades between 1996 and 2024, the study evaluates one-day-ahead forecasts across three model candidates. Specifically, I consider the Heterogeneous Autoregressive (HAR) model, a Feedforward Neural Network (FNN) in the form of a Multilayer Perceptron (MLP), and a Recurrent Neural Network (RNN) in the form of a Long Short-Term Memory (LSTM) network. The BIC guided specification uniformly improves predictive accuracy. Relative to fixed lags, mean squared error (MSE) decreases by 3.1% for the linear model, 15.8% for the MLP, and 3.4% for the LSTM, while QLIKE loss falls by 2.7%, 4.2%, and 0.5%, respectively. Diebold-Mariano (DM) tests confirm the statistical significance of these improvements, with the strongest gains for the MLP.

The results show that the BIC rule expands effective memory during high-dependence regimes such as the Global Financial Crisis and the COVID-19 shock, while maintaining parsimony in tranquil periods. These findings highlight the importance of a statistically disciplined memory structure in reducing forecast errors precisely when they are most costly, in the "black box" that is machine learning.

I also translate statistical gains into economic value using a Realised Utility (RU) framework introduced by Bollerslev et al. (2018). It represents the certainty equivalent performance of a mean-variance investor who positions an asset against a risk-free asset based on model implied volatility. Intuitively, RU increases when forecasts avoid undersizing the asset in calm periods, and especially oversizing in turbulent markets. BIC variants of each model candidate systematically improve RU: 1.70 basis points for the HAR, 1.46 basis points for the MLP, and 0.27 basis points for the LSTM, closing up to 5.2% of the residual gap to the theoretical upper bound.

I replicate the exercise across liquid S&P sector ETFs (Technology, Consumer Discretionary, Health Care) and also implement the Akaike Information Criterion (AIC) and Hannan-Quinn Information Criterion (HQC) in the model memory selection step. The same pattern holds: statistically disciplined lags reduce loss within each model candidate, with the strongest gains coming from the two strongest complexity penalising information criterion, the HQC and BIC.

The next step is to extend this memory-selection discipline into a true model-selection criterion by replacing the crude parameter count in BIC with an effective degrees-of-freedom (edf) measure that captures the learner's sensitivity to data. Building on recent developments in statistical learning, this extension would allow network architectures to be selected through a single unified score, providing a principled penalty that scales with model complexity and serial dependence. This extension would generalise BIC to modern nonlinear learners such as FNNs and RNNs and allow dynamic architecture selection within a time-series validation framework, especially in higher dimensional settings with more predictors than predictions, and more pronounced nonlinear relationships.

The remainder of this dissertation is set out as follows. Chapter 2 reviews related work with an evaluation of the research gap, and states our research questions and hypotheses. Chapter 3 details the methodology, including the BIC based lag selection and its mapping into inputs for HAR, MLP, and LSTM. Chapter 4 presents data and OOS forecasting results and statistical significance. Chapter 5 assesses economic value using the RU framework. Chapter 6 reports robustness checks, including liquid Sector ETFs outside of SPY, and alternative information criterion such as the AIC and HQC. Chapter 7 concludes the research, including limitations of this study, future work, and overall evaluation of the research.

Chapter 2

Related Literature and Contributions

2.1 Literature Review

Volatility modelling, and the literature in volatility forecasting builds primarily on the foundations from the early work in ARCH laid by Engle (1982) and its extensions (e.g. Bollerslev, 1986) which uses daily returns to infer conditional variance. However, any single squared return provides a "weak signal" of the true volatility, so GARCH models inherently adjust slowly to shocks. In contrast, Andersen and Bollerslev (1998) introduced realised volatility measures from high frequency data, showing that aggregating intraday returns yields far more accurate information about current volatility levels. This insight spurred a wave of models that incorporate intraday data. Andersen et al. (2003) demonstrated that simple long memory time series models on log realised volatility perform admirably, and likewise, the HAR model by Corsi (2008), built on this idea by regressing realised variance on its daily, weekly and monthly lags. Importantly, Corsi mapped three distinct horizons based on investor heterogeneity, in 1, 5 and 22 days. These values were intuitive proxies for one trading day, trading week, and trading month, with later guides describing these lags as "standard lag indexes" following the heterogeneous markets hypothesis.

Early studies applied simple feed-forward neural networks (FNNs) to single-asset realised variance series and reported modest gains over HAR and GARCH, hinting that flexible function approximators might distil additional predictive signals from the same inputs. Bucci (2020), for example, shows that recurrent architectures such as long short-term memory (LSTM) and nonlinear autoregressive exogenous model (NARX) significantly out-forecasted HAR and GARCH for S&P 500 volatility, especially during the 2008 Global Financial Crisis, suggesting that a network's ability to model long memory and regime shifts is economically meaningful. Despite these advances, recent empirical asset pricing work underscores that the same flexibility that makes neural networks attractive can also produce fragile, overfitted forecasts. Gu et al. (2020) stress that "enhanced flexibility … comes with a higher propensity of overfitting the data," and therefore modern ML toolkits must pair rich functional forms with explicit regularisation to ensure OOS stability. A parallel concern arises in the asset pricing "factor zoo", where Feng et al. (2020) show that without disciplined model selection rules, hundreds of seemingly significant factors melt away once one accounts for model selection mistakes, leading to spurious in-sample success but poor generalisation.

While not explicitly finance research, computer science (CS) and ML journals sharing similar concerns developed literature regarding stable model selection, such as Matsuda et al. (2022) proposing the Noise Contrastive Information Criterion (NCIC) and Score Matching Information Criterion (SMIC), which demonstrated that information criteria can scale to modern tasks and offered a solution to reducing potential overfitting risk. Similarly, recent statistics literature by McInerney and Burke (2024) proposed a BIC (Schwarz, 1978) style model selection with a modified degrees of freedom measure in FNNs resulting in competitive OOS accuracy.

2.2 Evaluation and Hypothesis Development

The foregoing literature establishes that ML models can improve volatility forecasts relative to linear models, especially in volatile markets. However, it also highlights that the flexibility and “black box” nature may magnify overfitting risk in the noisy and serially dependent financial data. Recent developments in computer science, statistics, and mathematics literature has information criterion, and similar regularisations in machine learning becoming more prevalent - however, the application to the financial time-series in a neural network context, to the best of my knowledge, is underdeveloped.

The contribution of this proposal is therefore a model’s memory selection discipline, in selecting a model’s memory with superior generalisation in the OOS, without relying on ad hoc grids or heuristics to the training data. I apply this to neural networks, using theory that has empirically been shown to generalise better in traditional financial econometrics literature, as well as recent developments in the mathematical and computational sciences. I develop a mapping from BIC selected AR orders to heterogeneous lag structures that preserves the interpretability of the HAR framework while adapting to data driven estimates of persistence. Accordingly, this paper proposes three testable hypotheses based on the foregoing literature:

1. H_1**:** Nonlinear neural networks will outperform linear models in average OOS forecasting accuracy.

2. H_2**:** Models with a BIC minimising feature set will exhibit superior OOS forecasting accuracy than otherwise identical models.

3. H_3**:** The forecasting advantage of information criterion tuned models over otherwise identical models will be largest during high volatility market regimes, when nonlinear dynamics and structural breaks are most pronounced.

The research question of this study is therefore: *Can a statistically disciplined set of realised volatility predictors, as selected by a Bayesian Information Criterion minimising process, improve out of sample forecasting performance for linear and non-linear modelling candidates?*

Chapter 3

Methodology

This chapter introduces and motivates three complementary model candidates that will be used in this study for realised volatility forecasting: the linear Heterogeneous Autoregressive (HAR) model, the feedforward (FNN) Multi-layer Perceptron (MLP) neural network, and the recurrent (RNN) Long Short-Term Memory (LSTM) neural network. HAR serves as an easily interpretable, linear benchmark that captures multi-horizon dependence and the apparent long memory of volatility (Corsi, 2008; Andersen et al., 2003). The MLP relaxes linearity and can approximate complex, nonlinear mappings from lagged volatility and auxiliary predictors to the conditional mean (Hornik et al., 1989), while recent empirical work reports that MLP type architectures often improve upon linear models such as HAR and GARCH as well as tree-based ML alternatives (Zhang et al., 2023; Christensen et al., 2022). To address temporal nonlinearities and long range dependence directly, I also employ LSTM NNs, whose gating and memory mechanisms are well suited to persistent volatility dynamics. Recent studies find that LSTMs based models deliver superior forecasts across equities and indices (e.g., Liu, 2025; Souto and Moradi, 2023; Lin et al., 2022). Together, these models allow this study to benchmark a transparent linear structure against flexible nonlinear and sequence learning approaches within a common feature set. The sections that follow specify estimation and training details for each.

3.1 Model Candidates

3.1.1 Heterogeneous Autoregressive (HAR)

The HAR model, suggested by Corsi (2008), mimics the heterogeneous market hypothesis, which suggests that realised volatility (RV) exhibits long memory and heterogeneous dependence across daily, weekly, and monthly horizons. In its daily form, it treats RV of asset i on day t as the weighted arithmetic for each of the three investor horizons, given by:

$$RV_{i,t}^{(1)} = RV_{i,t-1}, \quad RV_{i,t}^{(5)} = \frac{1}{5}\sum_{j=1}^{5} RV_{i,t-j}, \quad RV_{i,t}^{(22)} = \frac{1}{22}\sum_{j=1}^{22} RV_{i,t-j} \tag{3.1}$$

$$\hat{RV}_{i,t+1} = \beta_0 + \beta_d RV_{i,t}^{(1)} + \beta_w RV_{i,t}^{(5)} + \beta_m RV_{i,t}^{(22)} + \epsilon_{i,t+1} \tag{3.2}$$

with the β's - the weights of each horizon - being estimated using ordinary least squares (OLS). It is easily interpretable, robust with limited samples, and provides a linear benchmark against which the neural network models' nonlinear gains can be judged.

3.1.2 Multilayer Perceptron (MLP)

Volatility dynamics often show nonlinear state dependence that linear models, such as HAR, struggle to capture. An MLP is a class of FNN that is a "universal approximator" of complex mappings, being composed of an input layer to receive the raw features, and an output layer that makes forecasts about the input - with an arbitrary number of "hidden layers" in between the two that transform the nonlinear dynamics (Hornik et al., 1989). Formally, a standard three hidden-layer MLP is given by:

$$\hat{y}_{t+1|t} = W_4\phi(W_3\phi(W_2\phi(W_1x_t + b_1) + b_2) + b_3) + b_4 \tag{3.3}$$

where W and b are the weights and biases, x is the predictors available at time t, and ϕ is the Rectified Linear Unit (ReLU) activation.

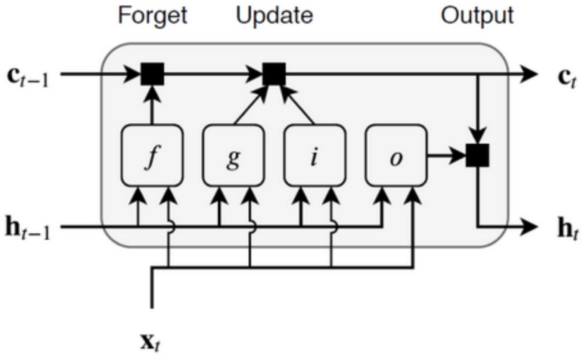


Figure 3.1: Architecture of an LSTM cell illustrating the flow of information through forget, input, and output gates. The horizontal cell state C_1 retains long term memory, while the hidden state h_1 captures short term dependencies, allowing volatility persistence to be modelled effectively.

3.1.3 Long Short-Term Memory (LSTM)

An LSTM NN is an RNN that, unlike the aforementioned MLP FNN, can reset, update, and keep long-term information; which is especially helpful, since volatility often shows persistence and long range dependence, allowing it to learn temporal patterns directly from windows of past observations (Hochreiter and Schmidhuber, 1997). For reference, an illustration of an LSTM cell is pictured in Figure 3.1.

Structurally, memory cell C runs horizontally through the unit, regulated by a forget gate, input gate, and output gate, and the hidden state h underneath, as demonstrated by Figure 3.1. Formally, the LSTM is given by:

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \tag{3.4}$$

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \tag{3.5}$$

$$\bar{C}_t = tan\ h(W_c x_t + U_c h_{t-1} + b_c) \tag{3.6}$$

$$C_t = f_t \circ C_{t-1} + i_t \circ \bar{C}_t \tag{3.7}$$

with x_t and h_{t-1} representing the input vector and previous hidden state respectively. f_t is the forget gate's activation vector, i_t is the input gate's activation vector, $\bar{C}_t$ is the cell input activation vector, and C_t is the hidden state vector. The weights W and U influence which elements of the input vector and previous hidden state influences forgetting, sigmoid function σ pushes values between 0 (completely forget) and 1 (completely keep), and b is the bias. Finally, the output gate is given by:

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \tag{3.8}$$

$$h_t = o_t \circ tan\ h(C_t) \tag{3.9}$$

where then a single dense layer with a ReLU activation transforms h_t into a one-day-ahead volatility forecast.

3.2 Data Splitting and Hyperparameter Selection

For each dataset, the observations (further discussed in Chapter 4.1) are divided into three, non-overlapping periods while maintaining chronological order. These three periods will be known as the training window, the validation window, and the testing window. The training window is used to fit the data to the model, the validation window is used to tune the model's hyperparameters, and the testing window is a truly OOS period that is used to create realised volatility forecasts used for evaluation of model performance.

The initial training window will consist of 5 years from January 1st, 1996 to January 1st, 2000, the initial validation window will consist of the next 1 year from January 2nd, 2000 to January 2nd, 2001, and the first testing window will be a truly OOS period ranging January 3rd, 2001 to January 3rd, 2002. Following Zhang et al. (2023), and Gu et al. (2020), due to limited computational resources, models are updated annually: following the first testing window, the following training window will be expanded by 1 year to encompass the data from the previous training window as well as new data, while the validation window will be rolled forward by 1 year to tune hyperparameters in the most recent period. The OOS testing window will similarly be rolled forward by 1 year.

For the linear HAR model, the training and validation windows are combined into one single training window, as there no requirements for hyperparameter tuning. Following Gu et al. (2020), I apply an ensemble approach with different random seeds to train the MLP and LSTM NNs due to their stochastic nature, improving robustness and model stability by averaging the outputs of different seeds.

3.3 Bayesian Information Criterion (BIC) for Adapting Model Memory

In order to statistically discipline a model's memory to the data for effective adaptation of changing market conditions, I select lag lengths using BIC (Schwarz, 1978). For a model with a maximised likelihood $L(\hat{\theta})$, k free parameters and n observations, the BIC is given by:

$$BIC = -2 \cdot log L(\hat{\theta}) + k \cdot log(n), \tag{3.10}$$

which penalises complexity at rate $log(n)$, making it a stronger penalty than comparable information criterion (Lütkepohl, 2005). Minimising BIC is consistent for order selection and imposes a penalty that increases with sample size, thereby favouring parsimonious specifications in finite samples. With each re-estimation window, I fit $AR(p)$ models for the realised volatility series y_t, over a grid of admissible orders p, given by:

$$AR(p) = \beta_0 + \beta_1 RV_{t-1} + \cdots + \beta_p RV_{t-p} + \epsilon_t, t = 1, ..., p_{max}, \tag{3.11}$$

where the order $p*$ that minimises $AR(p)$'s BIC value is retained, i,e.,

$$\{p_1^*, \ldots, p_k^*\} = \underset{\{p_1,\ldots,p_k\}}{\arg\min} \{\mathrm{BIC_{p*}}\} \tag{3.12}$$

$$BIC = -2 \log L(\hat{\theta}) + \log n\,(k+1), \tag{3.13}$$

where $\hat{\sigma}^2$ is the residual variance from estimation, and $k+1$ accounts for the $k = p$, $AR(p)$ coefficients plus the intercept.

For this study, I consider $p_{max} = 44$, or two trading months. To avoid look-ahead bias while preserving recency in selection, the information set used for this selection is based on the most recent 1 year prior to the OOS testing period, equivalent to the validation set data; for the HAR model, because specification has no validation set, the BIC is computed from the most recent one year of training observations prior to the testing set, reflecting the combined training and validation window.

The mapping compresses the p^* lagged regressors of an AR(p^*) into a single regressor $RV_t^{(p*)}$, thereby retaining the information content of the selected memory length while avoiding multicollinearity and overfitting from feeding many adjacent lags into a linear model. Intuitively, if the BIC chooses $AR(p) = 4$ (i.e., dependence extending over four trading days), the four-day average

$$RV_t^{(4)} = \frac{1}{4}\sum_{j=1}^{4} RV_{t-j} \tag{3.14}$$

serves as a lag component, summarising the information that an AR(4) would extract from $\{RV_{t-1}$

$, \ldots, RV_{t-4}\}$. Averaging imposes uniform weights over the last p^* observations, which is a conservative shrinkage relative to freely estimated AR weights; it is particularly attractive when adjacent lags are highly collinear. Using this, the mapping of features is given by:

$$\{1, p_1, p_2\} = \begin{cases} \{1, \lfloor p^* \cdot \frac{5}{22} \rfloor, p^*\}, & p^* > 10, \\ \{1, p^*, \lfloor p^* \cdot \frac{22}{5} \rfloor\}, & p^* \leq 10, \end{cases} \tag{3.15}$$

with the tuple subsequently sorted to ensure $1 < p_1 < p_2$. In practice, rounding and deduplication are enforced so that $1 \neq p_1 \neq p_2$, and p_2 remains within the admissible range.

This is done through limiting the minimum $AR(p)$, p_{min} to 2 to avoid perfectly collinear lag candidates such as $\{1, 1, 4\}$. Under the mapping for $p^* \leq 10$, it treats $p*$ as the 'medium' horizon and scales the long horizon to $p_2 = (22/5)\, p^*$, i.e., about a month's worth given the weekly anchor. Conversely, when $p^* > 10$, the mapping sets $p_2 = p^*$ and scales the weekly horizon to $p_1 \approx (5/22)p^*$, so that the monthly component reflects the longer memory directly selected by BIC while preserving the daily-to-weekly ratio of trading days. The pivot at $p* = 10$ acts as a practical boundary between a "weekly scale" dependence, as roughly two trading weeks, ensuring the monthly component reflects genuinely longer memory and the weekly component preserves the daily:weekly proportion.

To avoid test leakage, (p_1, p_2) are re-selected at each re-estimation date using only the designated selection sample, then the model in Equation (3.16) is re-fitted on the full training window and rolled forward to produce OOS forecasts. The selected order is then mapped into a triplet of three memory lengths. Let $\{1, p_1, p_2\}$ denote the first-, second-, and third order lags used as predictors. The mapping preserves the canonical trading-day ratios while allowing the lags to adapt to the estimated BIC minimising lag length. The next three subchapters will discuss how this triplet will be implemented into the different model candidates to form next period volatility forecasts.

3.3.1 Theoretical Justification for Heterogeneous Lag Selection

The mapping from p^* to the triplet $\{1, p_1, p_2\}$ formalises a tradeoff between data driven adaptivity and structural interpretability. In principle, one could feed all p^* lags $RV_{t-1}...RV_{t-p*}$ directly into the NN, allowing the model to learn optimal weights, However, this approach suffers from three drawbacks in the time series forecasting context. Firstly, adjacent volatility lags exhibit persistence with rates often exceeding 0.9 (Andersen et al., 2003) causing multicollinearity, with studies such as Zeng et al. (2022) attributing poor performance of their MLP during high collinearity between predictors. Secondly, the curse of dimensionality, as described by Bellman (1961), states that as the input size grows (with p^* in this study), it increases overfitting risk in finite samples despite regularisation. Section 3.3's mapping addresses these concerns by compressing the $AR(p^*)$ dimensional lag vectors into three averaged components, $RV^{(1)}$, $RV^{(p_1)}$, and $RV^{(p_2)}$, as described in Equation 3.14.

The specific mapping preserves the 1:5:22 ratio to maintain compatibility with the heterogeneous autoregressive framework, which itself derives from the heterogeneous agent hypothesis in finance (Müller et al., 1997). Under this hypothesis, short-, medium-, and long-term investors interact to generate volatility clustering and long memory. The ratio 1:5:22 reflects approximate calendar structure (1 day, 1 week, 1 month of trading days) and has been empirically validated across multiple asset classes (Corsi, 2008). By scaling p^* to preserve this ratio, I ensure that when BIC selects longer lags in high volatility regimes, the model adapts within the framework.

The pivot at p^* distinguishes two empirical regimes. When $p^* \leq 10$, volatility persistence appears short-lived ($<$ 2 calendar weeks), consistent with microstructure noise or transient liquidity shocks (Hansen and Lunde, 2006). I therefore treat p^* as a "weekly" signal and extrapolate to monthly signals. When $p^* > 10$, persistence extends beyond two weeks, consistent with macroeconomic shocks or risk premium dynamics (Engle and Lee, 1999) representing true long term persistence - I therefore treat p^* as a "monthly" signal and back out weekly lags.

3.3.2 HAR Implementation of BIC

The HAR framework models daily realised volatility y_t as a function of past values at heterogeneous horizons. In this implementation, the predictors are based on the BIC minimising autoregressed order

$$\hat{RV}_{i,t+1} = \beta_0 + \beta_1 RV_{i,t}^{(1)} + \beta_2 RV_{i,t}^{(p_1)} + \beta_3 RV_{i,t}^{(p_2)} + \epsilon_{i,t+1} \tag{3.16}$$

where $\{1, p_1, p_2\}$ are updated at every re-estimation date by the BIC procedure explained in Section 3.3. Estimation proceeds by ordinary least squares on the series. The benchmark is estimated according to Equation 3.2.

3.3.3 MLP Implementation of BIC

The MLP uses the same heterogeneity principle to construct its inputs. For each expanding window iteration, the AR based BIC on the most recent 1 year before testing yields $p*$ and, via the mapping in Section 3.3, a triplet $\{1, p_1, p_2\}$. The feature vector for date t is then

$$x_t = (RV_t^{(1)}, RV_t^{(p_1)}, RV_t^{(p_2)}), \tag{3.17}$$

scaled by the training statistics to have zero mean and unit variance, and forecasts $\hat{RV}_{t+1}$. Because the horizons are determined exclusively from the in-sample data within each window, the procedure avoids test leakage and ties the complexity of the input space to OOS performance, which is the relevant criterion for forecasting. This design preserves a clear economic interpretation of short-, medium-, and longer-run components of volatility, while delegating the combination of these components to a dynamic learner based on market conditions.

3.3.4 LSTM Implementation of BIC

For the recurrent specification, the selection is used to determine the sequence length, i.e., the lookback period hyperparameter, which is typically a fixed value in the literature integrating LSTMs. After computing $p*$ on the validation fold and mapping to $\{1, p_1, p_2\}$, the lookback window is set to the largest horizon, $L = p_2$. The input to the LSTM at date t is therefore the sequence

$$x_t = (RV_t^{(1)}, RV_t^{(2)} ... RV_t^{(L)}), \tag{3.18}$$

standardised using training statistics and outputs one day ahead $\hat{RV}_{t+1}$. Determining L from the BIC-minimising order links the recurrent memory of the network to a statistically disciplined estimate of the series' effective dependence length. This avoids arbitrarily excessively large sequences, which has been shown to reduce performance (Leites et al., 2024), while allowing the lookback length to expand in periods when longer memory is empirically supported. The resulting procedure marries the economic rationale of heterogeneous horizons with a sequence model capable of learning distributed temporal representation, with the size of that representation guided by an information criterion estimated without reference to the test set.

Algorithm 1 summarises the rolling BIC-guided horizon selection and forecasting procedure applied to all model candidates. The algorithm operates on an expanding window basis with annual re-estimation, selecting AR orders on the most recent 1 year prior to the testing period, mapping them to heterogeneous horizons, and generating OOS forecasts.

Algorithm 1 Rolling BIC-Guided Horizon Selection and Forecasting

```
Require: Realised volatility series $\{RV_t\}_{t=1}^{T}$;
Require: Initial windows: Train (expanding, start with 5y), Valid (1y), Test (1y); annual updates.
Require: AR search bounds $p_{\min} = 2$, $p_{\max} = 44$; model $M \in \{\mathrm{HAR}, \mathrm{MLP}, \mathrm{LSTM}\}$.
Require: Ensemble size $E$ for neural networks ($E = 10$); standardisation computed on Train.

for each re-estimation date $\tau$ in annual steps do
    Define $\mathrm{Train}_\tau$, $\mathrm{Valid}_\tau$, $\mathrm{Test}_\tau$ (chronological, non-overlapping).
    if $M = \mathrm{HAR}$ then
        SelWindow ← last 1 year of $\mathrm{Train}_\tau$        ▷ HAR has no separate validation set
    else
        SelWindow ← $\mathrm{Valid}_\tau$
A. BIC order selection on the selection window
    for $p \leftarrow p_{\min}, \dots, p_{\max}$ do
        Fit AR($p$): $RV_t = \beta_0 + \sum_{j=1}^{p} \beta_j RV_{t-j} + \varepsilon_t$ on SelWindow by OLS.
        $L(\hat{\theta}) \leftarrow$ maximised likelihood;   $n \leftarrow$ length(SelWindow).
        $\mathrm{BIC}(p) \leftarrow -2 \cdot \log L(\hat{\theta}) + (p+1)\log n$        ▷ $k = p$ AR params + intercept
    $p^\star \leftarrow \arg\min_{p \in \{p_{\min}, \dots, p_{\max}\}} \mathrm{BIC}(p)$        ▷ Eq. (3.12)
B. Map $p^\star$ to heterogeneous horizons
    Define $RV_t^{(m)} = \frac{1}{m}\sum_{j=1}^{m} RV_{t-j}$ for $m \in \{1, p_1, p_2\}$        ▷ Eq. (3.14)
    $\{p_1, p_2\} \leftarrow \textsc{MapToTriplet}(p^\star; p_{\min}, p_{\max})$        ▷ Eq. (3.15), Alg. 2
C. Build model-specific inputs and fit on $\mathrm{Train}_\tau$ (then predict on $\mathrm{Test}_\tau$)
    if $M = \mathrm{HAR}$ then        ▷ Eq. (3.16)
        Fit OLS on $\mathrm{Train}_\tau$:  $\hat{RV}_{i,t+1} = \beta_0 + \beta_1 RV_{i,t}^{(1)} + \beta_2 RV_{i,t}^{(p_1)} + \beta_3 RV_{i,t}^{(p_2)} + \varepsilon_{i,t+1}$
        Generate $\hat{RV}_{i,t+1}$ for all $t \in \mathrm{Test}_\tau$.
    else if $M = \mathrm{MLP}$ then        ▷ Eq. (3.17)
        Features: $x_t = (RV_t^{(1)}, RV_t^{(p_1)}, RV_t^{(p_2)})$; standardise using $\mathrm{Train}_\tau$ statistics.
        for $e = 1, \dots, E$ do
            Train $\mathrm{MLP}_e$ on $\mathrm{Train}_\tau$ with early stopping on $\mathrm{Valid}_\tau$.
        $\hat{RV}_{t+1} = \frac{1}{E}\sum_{e=1}^{E} \mathrm{MLP}_e(x_t)$ for all $t \in \mathrm{Test}_\tau$.
    else if $M = \mathrm{LSTM}$ then        ▷ Eq. (3.18)
        Set lookback $L \leftarrow p_2$; sequences $x_t = (RV_t^{(1)}, RV_t^{(2)}, \dots, RV_t^{(L)})$; standardise on $\mathrm{Train}_\tau$.
        for $e = 1, \dots, E$ do
            Train $\mathrm{LSTM}_e$ on $\mathrm{Train}_\tau$ with early stopping on $\mathrm{Valid}_\tau$.
        $\hat{RV}_{t+1} = \frac{1}{E}\sum_{e=1}^{E} \mathrm{LSTM}_e(x_t)$ for all $t \in \mathrm{Test}_\tau$.
```

```
**Algorithm 2** MapToTriplet($p^\star; p_{\min}, p_{\max}$)
**Require:** $p^\star \in \{p_{\min}, \dots, p_{\max}\}$; set $p_{\min} = 2$ to avoid duplicates with the daily lag.
if $p^\star \le 10$ then                                  ▷ Weekly-scale dependence
    $p_1 \leftarrow p^\star; \quad p_2 \leftarrow \left\lfloor \frac{22}{5} p^\star \right\rfloor$
else                                                      ▷ Monthly-scale dependence
    $p_2 \leftarrow p^\star; \quad p_1 \leftarrow \left\lfloor \frac{5}{22} p^\star \right\rfloor$
return $\{1, p_1, p_2\}$
```

3.4 Performance Evaluation

To evaluate performance of the models, the following functions are used on the OOS testing window predictions, with the target being one day ahead volatility forecasts. Lower values indicate better performance for both functions.

$$\text{Mean Squared Error (MSE)} = \frac{1}{T}\sum_{t=1}^{T}(\hat{R}V_t - RV_t)^2 \tag{3.19}$$

$$\text{Quasilikelihood (QLIKE)} = \frac{1}{T}\sum_{t=1}^{T}[\frac{\sigma_t}{\hat{\sigma}_t} - ln(\frac{\sigma_t}{\hat{\sigma}_t}) - 1] \tag{3.20}$$

where T represents the total number of OOS trading days used for evaluation. MSE provides a transparent quadratic penalty, while QLIKE is a strictly proper scoring rule for volatility and variance forecasts which penalises underprediction more heavily than overprediction, and remains well behaved when RV_t is a noisy measure (A. Patton, 2010).

To empirically compare significance in the forecasting accuracy between models, the Diebold-Mariano (DM) test will be used (Diebold and Mariano, 1995). Letting $L_t^{(a)}$ and $L_t^{(b)}$ be the forecast error for models a and b respectively, the differential becomes $d_t^{(a-b)} = L_t^{(a)} - L_t^{(b)}$, and its sample mean, $\bar{d}_t^{(a-b)} = \frac{1}{T}\sum_{t=1}^{L} d_t^{(a-b)}$. Loss differentials will be computed using QLIKE losses, following A. J. Patton and Sheppard (2009) demonstrating that measuring by QLIKE has the highest explanatory power in the DM test. The DM test assumes that d_t is covariance stationary. Given a null hypothesis of $E(d_t^{(a-b)}) = 0$, the DM test statistic is given by:

$$\text{DM} = \frac{\hat{d}}{\sqrt{V\hat{a}r(d)}} \tag{3.21}$$

A significantly negative (positive) value indicates that model a (model b) delivers a statistically significant smaller average loss. DM statistics use a Heteroskedasticity and Autocorrelation robust (HAC) standard errors for the long run variance, using the standard Newey-West truncation lag in the volatility forecasting literature of $q = h - 1$. However since this study is based on one day ahead forecasts, $q = h - 1 = 0$. As robustness, I also report DM tests using an automatic Newey-West bandwidth $q = \lfloor 4(T/100)^{2/9} \rfloor$, where T is the amount of observations per d_t. Performance metrics will be evaluated across different models and specifications, including HAR, HAR$_{BIC}$, MLP, MLP$_{BIC}$, LSTM, and LSTM$_{BIC}$.

Chapter 4

Empirical Study

4.1 Empirical Data and Realised Volatility

I use daily, annualised volatility series disseminated through Dacheng Xiu's Risk Lab. The database provides up to date estimators for individual equities and ETFs constructed from high frequency transactions. In particular, I work with the QMLE "Trades" series for the SPDR S&P 500 ETF (SPY), which I log transform prior to modelling. The sample runs from January 1, 1996 to April 29, 2024, yielding 7,301 daily observations after cleaning and filtering. A graph of the sample's true realised volatility is shown in Figure 4.1, and the summary statistics are listed in Table 4.1. Methodological details and data construction are documented on the Risk Lab site and in Xiu (2010), and validity of this data is discussed in Appendix A.

Statistic	N	**Mean**	**Median**	**Std. dev.**	**Min**	**P1**	**P99**	**Max**	**Skew.**	**Ex. Kurt.**	**JB** p
Value	7,301	0.1300	0.1091	0.0839	0.0004	0.0379	0.4586	1.2407	3.10	17.72	$< 10^{-6}$

Table 4.1: Summary statistics for QMLE realised volatility of SPY, January 1996 to April 2024. All values are annualised. Percentiles are empirical. Jarque-Bera (p) reported for normality.

SPY is among the most liquid cash instruments for U.S. market exposure, trading at extremely high frequency with tight spreads and deep depth, which are conditions that reduce the impact of microstructure frictions and support the use of high frequency based volatility

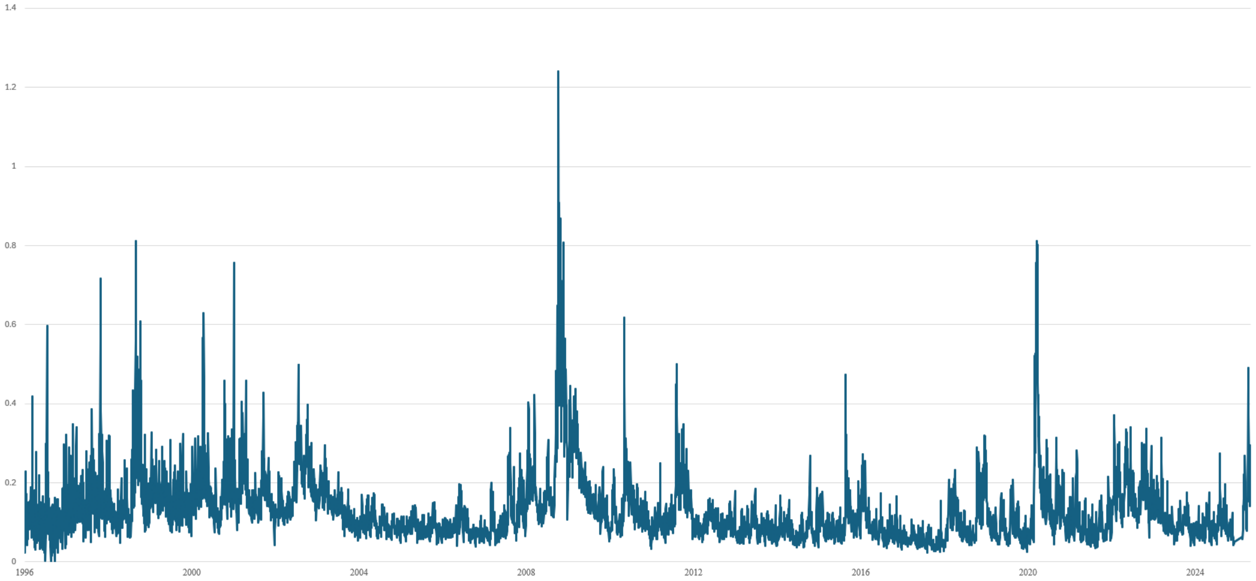


Figure 4.1: Daily realised volatility of the SPY ETF (QMLE Series), January 1996 to April 2024. Volatility spikes correspond to major market crises, including the Global Financial Crisis (2008-2009), and the COVID-19 Shock (2020).

estimators. Empirically, SPY is a standard object of study in volatility forecasting (often alongside large cap constituents), and prior work documents that its exceptional liquidity implies relatively low levels of microstructure noise. For example, Bu et al. (2022), forecast SPY volatility with extended HAR, and emphasise that “SPY is a very liquid asset with a low level of microstructure noise”, using it as the primary market proxy. Marshall et al. (2013) study intraday arbitrage in the two most liquid S&P 500 ETFs (SPY and IVV), further motivating ETF based proxies for market wide volatility. Using SPY therefore aligns this study with the literature while maximising data quality for high frequency estimators.

Standard daily realised volatility is a nonparametric estimator defined as the sum squared of intraday returns, consistent for quadratic variation as sampling becomes dense in the absence of microstructure noise. In practice, however, bid ask bounce, discreteness, and asynchrony induce noise that biases high frequency RV upward. Xiu’s QMLE offers a complementary, likelihood based route that models the noise explicitly rather than avoiding it by coarser sampling. The Risk Lab implementation treats observed non-zero transaction returns as the sum of latent efficient returns plus a discrete-time moving average noise process associated with trade arrival; it then fits an $MA(q)$ to the intraday return sequence, selects q via an Akaike Information Criterion (AIC), and estimates daily integrated variance by quasi-maximum-likelihood. Only days with at least 12 observations are retained. Xiu (2010) shows that the resulting quasi-MLE remains consistent and efficient for integrated volatility under stochastic volatility and a wide class of noise processes, while Da and Xiu (2021) further provide uniform inference when the noise is serially correlated and its magnitude varies with the sample. For comparison, the Risk Lab also reports 5- and 15- minute RV, highlighting the conceptual difference that RV is nonparametric and depends on a sampling choice. QMLE is maximum-likelihood based, and explicitly accounts for the noise dynamics.

4.2 Empirical Results

Table 4.2 reports OOS accuracy for one-day-ahead forecasts of the SPY ETF's realised volatility.

Metric	HAR	HAR$_{\text{BIC}}$	MLP	MLP$_{\text{BIC}}$	LSTM	LSTM$_{\text{BIC}}$
MSE	0.001634	0.001583	0.001823	0.001535	0.001474	0.001424
QLIKE	0.030623	0.029787	0.030820	0.029520	0.029062	0.028924

Table 4.2: Out of sample forecast accuracy for one day ahead SPY realised volatility, evaluated using MSE and QLIKE. BIC guided specifications outperform their fixed horizon counterparts within each model candidate, with the largest gain observed for the MLP. Blue text indicates best performance within each model candidate, and red text indicates best overall performance.

The BIC step lowers loss within every model candidate. Relative to the fixed horizon benchmarks, HAR$_{BIC}$ reduces MSE from 0.001634 to 0.001583 (3.1%) and QLIKE from 0.030623 to 0.029787 (2.7%). The gains are largest for the MLP, where MSE falls from 0.001823 to 0.001535 (15.8%) and QLIKE from 0.030820 to 0.029520 (4.2%). Even the recurrent model benefits, where the LSTM$_{BIC}$ trims MSE from 0.001474 to 0.001424 (3.4%) and QLIKE from 0.029062 to 0.028924 (0.5%). Furthermore, the rankings seem consistent with the loss function. On both MSE and QLIKE metrics, the LSTM$_{BIC}$ model attains the lowest average losses across the entire OOS test set, and overall, the BIC guided horizons consistently lower the average loss functions for each model candidate.

In short, BIC guided horizon selection is uniformly helpful, with especially material improvements for the FNN, and it is evident, at first glance, that best-in-class model is the recurrent LSTM network.

Panel A: $q = 0$ **(standard)**

	HAR	HAR$_{\text{BIC}}$	MLP	MLP$_{\text{BIC}}$	LSTM	LSTM$_{\text{BIC}}$
HAR	—	15.40***	-10.53***	20.66***	59.32***	32.01***
HAR$_{\text{BIC}}$		—	-19.95***	8.85***	11.60***	20.45***
MLP			—	27.40***	71.62***	37.91***
MLP$_{\text{BIC}}$				—	7.94***	21.24***
LSTM					—	4.30***
LSTM$_{\text{BIC}}$						—

Panel B: $q = 10$ **(Newey-West automatic)**

	HAR	HAR$_{\text{BIC}}$	MLP	MLP$_{\text{BIC}}$	LSTM	LSTM$_{\text{BIC}}$
HAR	—	5.09***	-3.58***	6.84***	19.47***	19.59***
HAR$_{\text{BIC}}$		—	-6.62***	2.86***	3.82***	6.55***
MLP			—	9.17***	22.78***	12.59***
MLP$_{\text{BIC}}$				—	2.64***	6.75***
LSTM					—	1.45*
LSTM$_{\text{BIC}}$						—

Table 4.3: Diebold-Mariano one-sided pairwise tests of predictive accuracy for SPY, measured by QLIKE loss differential. Positive (negative) statistics indicate the column (row) model outperforms the row (column) model. Asterisks (*, **, ***) denote significance at the 10%, 5%, and 1% levels. BIC guided designs yield statistically significant improvements across all model candidates. Panel B shows DM test statistics using HAC standard errors computed through automatic Newey-West truncation lag $q = \lfloor 4(T/100)^{2/9} \rfloor$.

Table 4.3 pairs models in a DM "horse race", where positive (negative) statistics indicate the column (row) model is preferred. The results confirm the patterns in Table 4.2. Within each class, BIC guided designs dominate their fixed horizon counterparts with statistical significance; HAR vs HAR$_{BIC}$ (DM = 15.3998***), MLP vs MLP$_{BIC}$ (DM = 27.4014***), and LSTM vs LSTM$_{BIC}$ (DM = 4.3048***) indicating statistically meaningful OOS gains. Across classes, LSTM$_{BIC}$ leads on QLIKE, significantly beating LSTM, MLP$_{BIC}$ and HAR$_{BIC}$. Statistical significance is also tested using a Newey-West truncation lag of $q = \lfloor 4(T/100)^{2/9} \rfloor = 10$, where the BIC variants of each model candidate continue to outperform their baselines with statistical significance. Overall, the BIC step both improves models within each class with statistical significance at the 1% level.

In sum, these results are consistent with the paper's core motivation. In tranquil periods, the canonical HAR lags or short MLP input window already approximate the signal well, so the BIC step adds little. However when the dependence lengthens in crises, disciplined horizon expansion curbs the largest errors. Empirically, this manifests as class consistent improvements from the BIC step (Table 4.2), and statistically significant wins in the DM horse race (Table 4.3). These patterns align with H_1 (nonlinear models can outperform linear models), and crucially, with the prediction that information criterion guided designs enhance OOS performance by regularising the horizon choice (H_3).

4.3 Predictive Periods

To locate when BIC selected horizons add value, I track the cumulative sum of squared error (CSSE) differences between each model against its non-BIC counterpart, as given by ($M \in \{\text{HAR}, \text{MLP},$
$\text{LSTM}\}$), and plot its running sum. Let

$$e_{M,t} = y_t - \widehat{y}_{M,t}, \qquad e_{M,\text{BIC},t} = y_t - \widehat{y}_{M,\text{BIC},t}, \tag{4.1}$$

denote one-step-ahead forecast errors (in levels). The pointwise improvement is

$$\Delta_t^{(M)} \;=\; e_{M,t}^2 \;-\; e_{M,\text{BIC},t}^2, \tag{4.2}$$

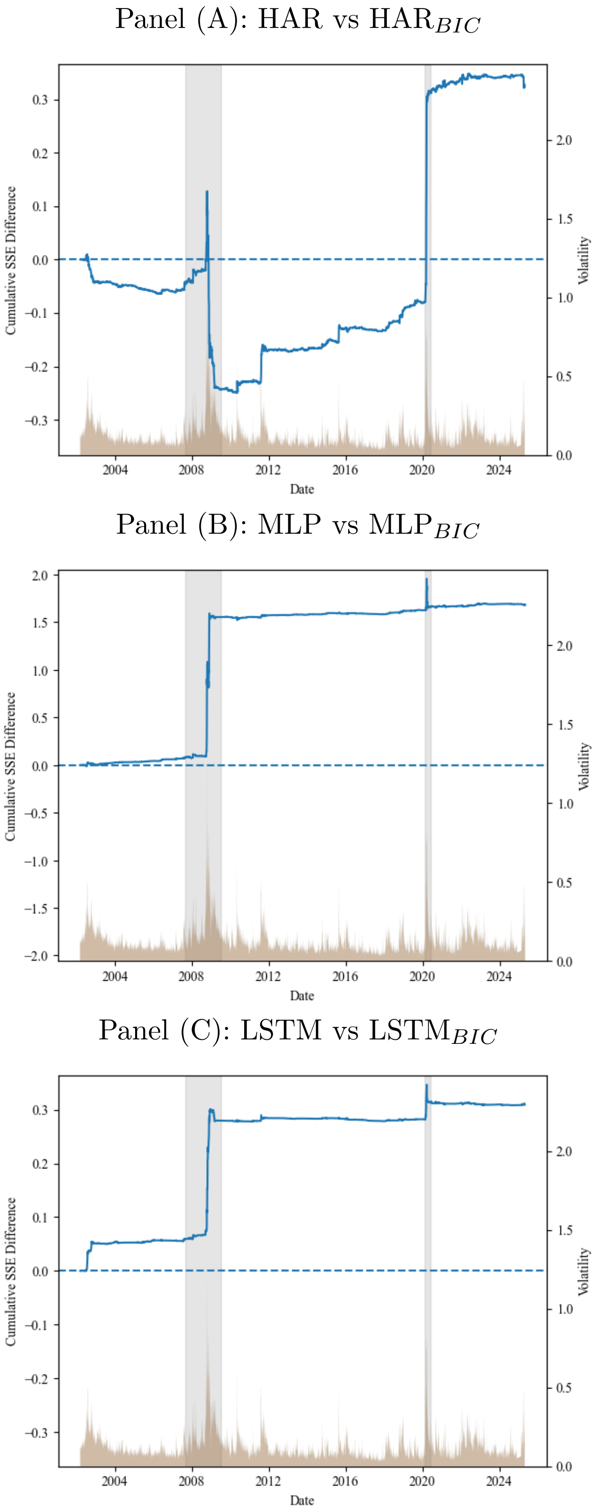
so that $\Delta_t^{(M)} > 0$ indicates a smaller squared error for the BIC version at t. The cumulative curve reported in Panels (A) to (C) of Figure 4.2 is

$$C_t^{(M)} = \sum_{s \leq t} \Delta_s^{(M)} = \sum_{s \leq t} e_{M,s}^2 - \sum_{s \leq t} e_{M,\mathrm{BIC},s}^2, \tag{4.3}$$

so an upward segment indicates periods in which the BIC specification reduces loss. The figures also overlay realised volatility on the right axis to relate improvements to market conditions.

Panel (A) represents the CSSE for HAR vs. HAR$_{BIC}$. The cumulative improvement is essentially flat to mildly negative throughout the tranquil pre 2007 period, implying that the canonical HAR$(1, 5, 22)$ is hard to improve upon when volatility is low and dependence is short. A pronounced step occurs during the Global Financial Crisis (GFC), with a sharp negative decline. Since BIC is estimated once every training set, this suggests that the HAR$_{BIC}$ model struggles to properly readjust for the "recovery period" after a high volatility spike such as the GFC given its purely linear structure. This is followed by a gradual, persistent rise and a second discrete jump in the COVID-19 episode. By and large, step increases coincide with spikes in realised volatility and are consistent with the mechanism in Section 3.3: when the autocorrelation of volatility lengthens in stress, the BIC chooses larger AR orders p^* allowing it to capture longer lags without compromising generalisation of the model, thereby reducing large forecast errors precisely when persistence is strongest. Between crises, when realised volatility subsides, the curve is largely horizontal, indicating that the two specifications are nearly observationally equivalent.

Panel (B) represents the CSSE for MLP vs. MLP$_{\mathrm{BIC}}$. For the FNN the gains are even more concentrated in high volatility regimes. The cumulative series remains near zero until the GFC, at which point it jumps discretely and stays elevated thereafter, with a modest additional step around COVID. This pattern indicates that disciplined, BIC guided input horizons matter most when the target exhibits long memory and level shifts. In calm markets



Panel (A): HAR vs HAR$_{BIC}$

Panel (B): MLP vs MLP$_{BIC}$

Panel (C): LSTM vs LSTM$_{BIC}$

Figure 4.2: The cumulative squared error differences between BIC-guided and non-BIC models over time. Left axis represents the cumulative squared error difference, and the right axis shows realised volatility. The x axis represents the date. Upward (downward) slopes indicate periods where BIC (non-BIC) models achieve lower forecast errors, particularly notable during high-volatility regimes such as the GFC (2008-2009) and COVID-19 (2020), as indiciated by the shaded regions.

the MLP approximates the mapping from short horizon inputs adequately even if horizons are misspecified; in turbulent markets the BIC driven expansion of (p_1, p_2) supplies inputs commensurate with the longer run persistence, reducing the large errors that dominate the sum of squares in Equation (4.3). The dominance of MLP_{BIC} under QLIKE in the DM horserace is therefore attributable to material gains concentrated in the same stress windows.

Panel (C) represents the CSSE for LSTM vs. LSTM_{BIC}. The recurrent architecture exhibits the smallest, yet economically meaningful improvements. Outside the GFC and COVID windows the cumulative curve is nearly flat, consistent with the LSTM's gating providing an internally adaptive memory so that the exact lookback length is second order in placid conditions. During the crises, however, the curve moves upward, indicating that setting the sequence length $L = p_2$ via BIC supplies a more generalisable lookback period. This behaviour reconciles the modest but significant DM statistics for LSTM_{BIC} with the visual evidence that gains are episodic rather than diffuse.

Taken together, Figure 4.2 indicates that the BIC mechanism functions as a crisis adaptation device, where it stretches the model's effective memory when the dependence structure of realised volatility lengthens, to the extent of not overfitting, and retracts it when conditions normalise. Because the objective in (4.3) is quadratic, improvements accrue primarily by curbing large forecast misses, which are concentrated in high volatility episodes, as anticipated by the initial hypothesis of this study.

Chapter 5

Economic Value and Realised Utility

This chapter asks whether the statistical gains from Chapter 4 actually matter for an investor who sizes positions using model implied volatility. In order to measure this gain in economic magnitude, Bollerslev et al. (2018) proposed a realised utility framework, which measures the utility based benefits of an investor with mean-variance preferences, investing in an asset with time varying volatility with a constant Sharpe ratio. This allows for an easily interpretable economic application, as opposed to OOS performance via statistical metrics.

Typically, the expected utility of an investor at time t may be approximated as:

$$E_t(u(W_{t+1})) = E_t(W_{t+1}) - \frac{1}{2}\gamma^A Var_t(W_{t+1}), \tag{5.1}$$

where W_t denotes the wealth, and γ^A denotes the absolute risk aversion of the investor. Assume that the investor allocates a fraction x_t of current wealth to a risky asset with return r_{t+1} and the rest to a risk-free money market earning r_t^f. Then, wealth at $t+1$ becomes $W_{t+1} = W_t(1 + r_t^f + x_t r_{t+1}^e)$, where $r_{t+1}^e = r_{t+1} - r_t^f$. This results in an expected utility of

$$\begin{aligned} U(x_t) &= W_t \left(x_t E_t(r_{t+1}^e) - \frac{\gamma}{2} x_t^2 Var_t(r_{t+1}^e) \right) \\ &= W_t \left(x_t E_t(r_{t+1}^e) - \frac{\gamma}{2} x_t^2 E_t(RV_{t+1}) \right) \end{aligned} \tag{5.2}$$

where $\gamma = \gamma^A W_t$ is the investor's relative risk aversion.

To focus on risk modelling, Bollerslev et al. (2018) assume the conditional Sharpe ratio as constant, given by $SR = E_t(r^e_{t+1})/\sqrt{E_t(RV_t + 1)}$. Under this assumption, the expected utility, depending on position x_t, together with expected realised volatility $E_t(RV_{t+1})$ is given by:

$$U(x_t) = W_t(x_t SR\sqrt{E_t(RV_{t+1})}) - \frac{\gamma}{2} x_t^2 E_t(RV_{t+1})). \tag{5.3}$$

The optimal portfolio that maximises this utility is obtained by investing into the risky asset at the following rate:

$$x_t^* = \frac{SR/\gamma}{\sqrt{E_t(RV_{t+1})}}. \tag{5.4}$$

In order to give an explicit quantification of the different models' utility gains, let $E_t^\theta(\cdot)$, and let $E_t(\cdot)$ denote the expectations from the true (unknown) model. Assuming the investor uses model θ, the position $x_t^\theta = (SR/\gamma)/\sqrt{E_t^\theta(RV_{t+1})}$ is chosen. Given this, the expected utility per unit of wealth, otherwise known as realised utility (RU), is given by:

$$RU_t = \frac{SR^2}{\gamma} \times \frac{\sqrt{E_t(RV_{t+1})}}{\sqrt{E_t^\theta(RV_{t+1})}} - \frac{SR^2}{2\gamma} \times \frac{E_t(RV_{t+1})}{E_t^\theta(RV_{t+1})} \tag{5.5}$$

If a model is ideal - i.e., predicts the RV for period $t + 1$ perfectly - then its RU is $SR^2/2\gamma$. In other words, the investor is willing to give up $SR^2/2\gamma$ of wealth in order to use the "perfect model" instead of making an investment in the risk-free asset. Following Bollerslev et al. (2018), a Sharpe ratio of 0.4, and coefficient of risk aversion of 2 is applied, making the theoretical maximum RU 4%. RU is evaluated using the same OOS test forecasts as previous sections.

	HAR	HAR$_{\text{BIC}}$	MLP	MLP$_{\text{BIC}}$	LSTM	LSTM$_{\text{BIC}}$
Realised Utility (%)	3.714916	3.731914	3.718395	3.732959	3.729043	3.731782

Table 5.1: Realised utility gains (certainty equivalent, %, SR = 0.4, $\gamma = 2$). Although absolute differences are modest, BIC guided variants consistently deliver higher realised utility, indicating economically meaningful benefits for mean variance investors. Blue text indicates best performance within each candidate model, and red text indicates best overall performance.

Table 5.1 reports certainty equivalent realised utility (RU) for the six forecasting configurations under mean variance preference, as given in Equation 5.5. The estimates lie in a narrow band, tightly clustered between 3.7149% and 3.7330%, but the ordering is unambiguous. For each model candidate, the BIC guided specification delivers higher RU than its non-BIC counterparts - HAR improves by 1.70 bps (3.7149% to 3.7319%), MLP by 1.46 bps (3.7184% to 3.7330%), and the LSTM by 0.27 bps (3.7290% to 3.7318%). In levels, the RU rankings is MLP$_{BIC}$ $>$ HAR$_{BIC}$ $>$ LSTM$_{BIC}$ $>$ LSTM $>$ MLP $>$ HAR. Measured against the theoretical upper bound $SR^2/2\gamma = 4\%$, the BIC step closes 5.96% of the residual gap for HAR, 5.17% for the MLP, and 1.01% for LSTM. These values are economically modest, but systematic and directionally consistent across model candidates. These patterns establish that statistically disciplined horizon selection translates into tangible gains for a mean variance investor who sizes positions using model implied volatility.

The fact that the RU ordering does not mirror the statistical loss rankings from Section 4 is entirely coherent once one recognises how portfolio sizing propagates forecast errors into utility. With the constant Sharpe ratio rule, the investor's position obeys $x_t^* = (SR/\gamma)/\sqrt{E_0(RV_{t+1})}$, implying that the certainty equivalent in Equation (5.5) is a nonlinear functional of both the level and the dispersion of the volatility forecast. Underprediction of volatility mechanically induces oversizing and is penalised twice in Equation (5.5). This inflates the variance term and reduces the first order gain, whereas comparable over prediction only scales exposure down. Consequently, RU is more sensitive to calibration and stability of scale than to marginal improvements in average statistical loss; small, infrequent downward bias episodes can overturn an apparent advantage in MSE or QLIKE. The BIC step raises RU precisely because it curtails these economically costly mis-sizing events by adapting memory when dependence lengthens, while restraining effective model complexity in placid regimes.

Chapter 6

Robustness Checks

For robustness, this chapter probes external validity along two dimensions. First, I replicate the one day ahead forecasting exercise for liquid S&P sector ETFs to test whether BIC guided horizon selection continues to increase forecasting performance when dependence structures differ across industries. Second, I implement two other commonly used information criterion: the Akaike Information Criterion (AIC), and the Hannan-Quinn Information Criterion (HQC), whose methodologies will be detailed in Section 6.2. Throughout, I keep same expanding window evaluation, loss functions, and RU evaluation so that any changes in accuracy is attributable to the horizon discipline rather than altered training or evaluations. Hyperparameter selection for the NNs remain consistent as explained in Appendix B. Throughout the section, I apply a modified DM test metric appropriate multi-asset pairwise comparisons following Gu et al. (2020), given by:

$$d_t^{(a-b)} = \frac{1}{N}\sum_{i=1}^{N}(L_{i,t}^{(a)} - L_{i,t}^{(b)}), \tag{6.1}$$

$$\text{DM} = \frac{\hat{d}}{\sqrt{\hat{Var}(d)}}, \tag{6.2}$$

where $L_{i,t}^{(a)}$ refers to the forecast error for asset i at time t, for N total assets. This test compares the cross-sectional average of prediction errors from each forecast, rather than the errors for each forecast individually. Again, the DM test uses HAC consistent standard errors using the Newey-West truncation lag following the standard $q = h - 1$, as well as $q = \lfloor 4(T/100)^{2/9} \rfloor$.

6.1 Alternative Sector ETFs

For the alternative industry ETFs, I focus on XLK (Technology), XLY (Consumer Discretionary), and XLV (Health Care), as together, they span growth sensitive cyclicals and a classic defensive sector while remaining among the most liquid SPDR sectors, making for an ideal stress test for whether horizon selection adapts to heterogeneous persistence. Technology's valuations are documented to feature elevated uncertainty and episodic volatility around innovation cycles, which amplifies state dependence in various dynamics (Pástor and Veronesi, 2009). In contrast, standard cyclical/defensive classifications place Consumer Discretionary squarely in the cyclical bucket and Health Care among defensives, letting performance be evaluated across cash flow profiles that react very differently to the business cycle (MSCI, 2018). Using sector ETFs also preserves the microstructure advantages discussed earlier for SPY, given by tight spreads and high trade counts, so that high frequency QMLE inputs are comparable to the baseline. Table 6.1 reports OOS accuracy for one day ahead volatility in XLK, XLY, and XLV.

XLY shows the largest and most systematic gains from BIC, with QLIKE falling by 19.1% for HAR (0.092854 to 0.075123), 25.1% for MLP (0.098709 to 0.073928), and 2.3% for LSTM (0.064767 to 0.063298). MSE drops 12.9%, 11.3%, and 1.4% respectively. These are economically meaningful reductions concentrated in the cyclical sector where persistence tends to lengthen in stress. Again, the LSTM and LSTM$_{BIC}$ shows class-leading performance, due to its ability to capture long term persistence and short term memory within its gates. Second, XLV exhibits QLIKE improvements for all model candidates, showing improvements

Sector	Metric	HAR	HAR$_{\text{BIC}}$	MLP	MLP$_{\text{BIC}}$	LSTM	LSTM$_{\text{BIC}}$
XLK							
	MSE	0.004185	0.003814	0.004020	0.003956	0.003762	0.003813
	QLIKE	0.043150	0.038985	0.041405	0.038973	0.038471	0.038650
XLY							
	MSE	0.007156	0.006236	0.006505	0.005770	0.005230	0.005155
	QLIKE	0.092854	0.075123	0.098709	0.073928	0.064767	0.063298
XLV							
	MSE	0.004382	0.004081	0.003469	0.003483	0.002357	0.002700
	QLIKE	0.067957	0.059933	0.053190	0.046639	0.040531	0.039196

Table 6.1: Out of sample forecast accuracy for one day ahead XLK, XLY, and XLV realised volatility, evaluated using MSE and QLIKE. BIC guided specifications outperform their non-BIC counterparts for each model candidate. Blue text indicates best performance within each model candidate, and red text indicates best overall performance.

Realised Utility (%)	HAR	HAR$_{\text{BIC}}$	MLP	MLP$_{\text{BIC}}$	LSTM	LSTM$_{\text{BIC}}$
XLK	3.473024	3.584412	3.505861	3.568772	3.527980	3.523877
XLY	2.600691	3.000359	2.442899	3.017611	3.142559	3.198858
XLV	2.897912	3.096859	3.230271	3.370701	3.450969	3.480068

Table 6.2: Realised utility gains (certainty equivalent, %, SR = 0.4, $\gamma = 2$). Although absolute differences are modest, BIC guided variants consistently deliver higher realised utility, indicating economically meaningful benefits for mean variance investors (with the exception of LSTM XLK). Blue text indicates best performance within each model candidate, and red text indicates best overall performance.

of 11.8% for HAR, 12.3% for MLP, and 3.3% for LSTM, with again, first-class performance from the LSTM$_{BIC}$ within the sector. On MSE however, the non-BIC LSTM attains the sector low at 0.002357, while it's BIC variant shows a 3.43 bps increase. Third, XLK shows clear BIC gains for HAR (-8.9% MSE, -9.7% QLIKE) and MLP (-5.9% QLIKE), while the LSTM baseline already leads within the sector and the LSTM$_{BIC}$ shows no clear gains. This mirrors the SPY results where the recurrent network saw the smallest incremental benefit from horizon tuning, due to its gates already providing adaptive memory. Fixed horizon learners, in HAR and MLP, gain most from disciplined expansion. Overall, BIC reduces loss in most sector model pairs, with the largest improvements in the cyclical XLY, moderate but broad gains in defensive XLV, and limited, but notable gains on the XLK with the exception of the recurrent model. On average however, all BIC variants exhibit lower MSE and QLIKE than their non-BIC counterparts.

Table 6.2 again translates the empirical results into RU, and shows the same directional pattern as previously: BIC variants almost always deliver higher RU within class, with the lone exception of LSTM$_{BIC}$ in XLK (-0.41 bps). Levels differ across industries, but the pattern remains consistent. XLK sits highest (3.52 to 3.58%), XLV close behind (3.23 to 3.48%), and XLY materially lower (2.44 to 3.2%), consistent with their cyclical/defensive mix. The largest economic gains from BIC accrue in XLY, where HAR$_{BIC}$ (+39.97 bps), MLP$_{BIC}$ (+57.47 bps), and LSTM$_{BIC}$ (+5.63 bps) close roughly 29 to 37% of the residual gap to the 4% theoretical upper bound. Following Section 5, HAR$_{BIC}$ shows the highest RU for XLK, yet for XLY and XLV, LSTM$_{BIC}$ leads in RU.

Table 6.3 reports the DM test statistics for the alternative sector ETF forecasts. The results closely mirror the findings shown in Section 4.2, with the BIC variants outperforming its non-BIC counterparts, given by: HAR vs HAR$_{BIC}$ (DM = 22.3426***), MLP vs MLP$_{BIC}$ (24.0037***), and LSTM vs LSTM$_{BIC}$ (2.7913***) reflecting outperformance with statistical significance at the 1% level. Notably, one difference is the base MLP's outperformance of the canonical HAR model, which is more closely in line with the findings in the financial machine

Panel A: $q = 0$ (standard)

	HAR	HAR_{BIC}	MLP	MLP_{BIC}	LSTM	LSTM_{BIC}
HAR	—	22.34^{***}	16.90^{***}	25.31^{***}	35.88^{***}	29.08^{***}
HAR_{BIC}		—	-17.07^{***}	22.46^{***}	27.49^{***}	30.11^{***}
MLP			—	24.00^{***}	34.60^{***}	28.06^{***}
MLP_{BIC}				—	17.03^{***}	28.36^{***}
LSTM					—	2.79^{***}
LSTM_{BIC}						—

Panel B: $q = 9$ (Newey-West automatic)

	HAR	HAR_{BIC}	MLP	MLP_{BIC}	LSTM	LSTM_{BIC}
HAR	—	7.24^{***}	5.20^{***}	8.10^{***}	11.41^{***}	9.27^{***}
HAR_{BIC}		—	-5.57^{***}	7.26^{***}	8.87^{***}	9.90^{***}
MLP			—	7.73^{***}	10.96^{***}	8.96^{***}
MLP_{BIC}				—	5.50^{***}	8.99^{***}
LSTM					—	0.91
LSTM_{BIC}						—

Table 6.3: Modified Diebold-Mariano one-sided pairwise tests of predictive accuracy for industry sector ETFs, measured by QLIKE loss differential. Positive (negative) statistics indicate the column (row) model outperforms the row model. Asterisks (*, **, ***) denote significance at the 10%, 5%, and 1% levels. BIC guided designs yield statistically significant improvements across all model candidates. Panel B shows DM test statistics using HAC standard errors computed through automatic Newey-West truncation lag $q = \lfloor 4(T/100)^{2/9} \rfloor$.

learning literature. Using the automatic Newey West bandwidth for q, all models retain statistical significance at the 1% level, with the exception of the LSTM$_{BIC}$, presumably due to LSTM's gated memory nature allowing for more similar forecasts even with an endogenously defined lookback period. However, the point estimates directionally support the LSTM$_{BIC}$'s forecasting advantage, albeit with a one-sided p-value of only 0.16. Overall, sector ETFs adds robustness to the claim in Section 5 that BIC disciplined models translates into tangible economic gains, throughout multiple sectors of differing characteristics.

6.2 Alternative Information Criterion

In order to test for robustness under alternative information criterion, I implement two other commonly used information criterion in econometrics, in addition to the BIC: the Akaike Information Criterion (AIC) (Akaike, 1974), and the Hannan-Quinn Information Criterion (HQC) (Hannan and Quinn, 1979). Ordered by the weakest to strongest in terms of complexity penalty, they are given by:

$$AIC = -2 \cdot \log L(\hat{\theta}) + 2k \tag{6.3}$$

$$HQC = -2 \cdot \log L(\hat{\theta}) + 2k \cdot \log(\log(n)) \tag{6.4}$$

$$BIC = -2 \cdot \log L(\hat{\theta}) + k \cdot \log(n) \tag{6.5}$$

where identically to the BIC, $L(\hat{\theta})$ is the maximised likelihood, k is the parameter count, and n is the number of observations, penalising complexity at the rates 2, $2log(log(n))$, and $log(n)$ respectively, making BIC the most penalising criterion - and theoretically, offer the most parsimonious model. Again, within each re-estimation window, I fit $AR(p)$ models for the realised volatility series y_t, over a grid of admissible orders p, given by:

$$AR(p) = \beta_0 + \beta_1 RV_{t-1} + \cdots + \beta_p RV_{t-p} + \epsilon_t, t = 1, ..., p_{max}, \tag{6.6}$$

where for each IC candidate, AIC, HQC, and BIC, an IC minimising value is retained:

$$\{p_1^*, \ldots, p_k^*\} = \underset{\{p_1,\ldots,p_k\}}{\arg\min} \{\mathrm{IC_{p*}}\} \tag{6.7}$$

$$\mathrm{IC} = \begin{cases} \mathrm{AIC} & = -2\log L(\hat{\theta}) + 2(k+1), \\ \mathrm{HQC} & = -2\log L(\hat{\theta}) + 2\,\mathrm{loglog}\, n\,(k+1), \\ \mathrm{BIC} & = -2\log L(\hat{\theta}) + \log n\,(k+1), \end{cases} \tag{6.8}$$

where $\hat{\sigma}$ would be the residual variance from estimation, and $k+1$ accounts for the $k = p$, $AR(p)$ coefficients plus the intercept. The mapping into $\{1, p_1, p_2\}$ is identical to Section 3. Given this, Table 6.4 details the forecasting accuracy of each model candidate with it's four different lag selection candidates, as measured by MSE and QLIKE.

Metric		**MSE**	**QLIKE**	**RU (%)**
HAR	Standard	0.001634	0.030623	3.714916
	AIC	0.001588	0.029935	3.729582
	HQC	0.001576	0.029721	3.732011
	BIC	0.001583	0.029787	3.731914
MLP	Standard	0.001823	0.030820	3.718395
	AIC	0.001525	0.029754	3.729472
	HQC	0.001536	0.029538	3.732711
	BIC	0.001535	0.029520	3.732950
LSTM	Standard	0.001474	0.029062	3.729043
	AIC	0.001420	0.028942	3.731294
	HQC	0.001420	0.028922	3.731969
	BIC	0.001424	0.028924	3.731782

Table 6.4: Out of sample forecast accuracy for one day ahead SPY realised volatility, evaluated using MSE, QLIKE, and Realised Utility (RU). First column represents the model candidate, and second represents the alternative information criterion used for forecast evaluation. IC-guided variants outperform the standard variants . Blue text indicates best performance within each model candidate, and red text indicates best overall performance.

Firstly, any statistically disciplined horizon choice improves accuracy: for every model candidate and for both loss functions, all IC guided specifications strictly dominate the standard variants. Second, the IC that minimises loss depends on the model candidate. For the linear HAR, HAR$_{HQC}$ attains the lowest MSE and QLIKE (0.001576 and 0.029935), followed closely by HAR$_{BIC}$ (0.001583, 0.029787). For the MLP, the stronger BIC penalty of MLP$_{BIC}$ attains the best QLIKE (0.02952), while the MLP$_{AIC}$ is narrowly best on MSE (0.001525), with MLP$_{HQC}$ sitting in between on both metrics. For the LSTM, LSTM$_{HQC}$ again edges out on both MSE and QLIKE (0.001420, 0.028922), and LSTM$_{BIC}$ delivering a close second. The improvements are small in absolute magnitude but systematic, and they replicate the central message from earlier sections, where forecasting gains flow from adapting the model's effective memory to the data rather than from any particular architecture.

The model candidate DM tests in Table 6.5 confirm the statistical significance of these rankings under QLIKE. In the HAR family (Panel A), every IC guided variant significantly outperforms the standard benchmark. Interestingly, HQC is significantly preferred to both AIC ($DM = 28.62^{***}$) and BIC ($DM = -11.77^{***}$), establishing HQC as the most accurate IC for the linear specification. In the MLP family (Panel B), the ordering is strict: BIC dominates HQC ($DM = 16.23^{***}$), AIC ($DM = -20.55^{***}$), and the non-IC MLP ($DM = 27.40^{***}$), underscoring that the more flexible FNN benefits from the stronger complexity penaltiy embodied in BIC. In the LSTM family (Panel C), all three IC guided designs beat the standard LSTM at the 1% level, but HQC is preferred to BIC ($DM = -3.75^{***}$) and AIC ($DM = 17.13^{***}$), pointing to an intermediate penalty as the sweet spot for the recurrent architecture, thanks in part to its internal gating and long memory allowing for a more lenient penalisation.

Using $q = 10$, results remain robust for all three model candidates in essentially every combination. However, LSTM$_{AIC}$'s DM stat indicates that while it produces directionally superior forecasts, it is statistically indistinguishable from the standard LSTM at typical levels (p=0.104). The remaining tests again illustrates that while the HQC and BIC penalties

Panel A: HAR

$q = 0$

	HAR_{AIC}	HAR_{HQC}	HAR_{BIC}
HAR	12.74***	16.75***	15.40***
HAR_{AIC}	—	28.62***	19.84***
HAR_{HQC}		—	-11.77***

$q = 10$

	HAR_{AIC}	HAR_{HQC}	HAR_{BIC}
HAR	4.04***	5.31***	4.88***
HAR_{AIC}	—	27.90***	8.68***
HAR_{HQC}		—	-3.58***

Panel B: MLP

$q = 0$

	MLP_{AIC}	MLP_{HQC}	MLP_{BIC}
MLP	21.97***	8.68***	8.79***
MLP_{AIC}	—	5.73***	6.23***
MLP_{HQC}		—	4.93***

$q = 10$

	MLP_{AIC}	MLP_{HQC}	MLP_{BIC}
MLP	7.03***	26.39***	26.73***
MLP_{AIC}	—	18.45***	20.01***
MLP_{HQC}		—	15.90***

Panel C: LSTM

$q = 0$

	LSTM_{AIC}	LSTM_{HQC}	LSTM_{BIC}
LSTM	3.77***	4.22***	4.19***
LSTM_{AIC}	—	16.65***	14.79***
LSTM_{HQC}		—	-3.63***

$q = 10$

	LSTM_{AIC}	LSTM_{HQC}	LSTM_{BIC}
LSTM	1.26	1.41*	1.40*
LSTM_{AIC}	—	5.21***	4.61***
LSTM_{HQC}		—	-1.1381

Table 6.5: Diebold-Mariano one-sided pairwise tests of predictive accuracy for SPY realised volatility, measured by QLIKE loss differential. Panels A, B, and C show results for HAR, MLP, and LSTM respectively. Left (right) columns use standard $q = 0$ (Newey-West $q = 10$) bandwidth. Positive statistics indicate column model outperforms row model. Asterisks (*, **, ***) denote significance at 10%, 5%, and 1% levels.

for the LSTM are statistically indistinguishable from each other, they are statistically superior to the standard LSTM and the AIC variant. Again, this demonstrates that BIC and HQC's stronger penalties allow a statistically disciplined method of volatility forecasting across all model candidates, and even meaningful for the already gated, long memory design of the LSTM.

Chapter 7

Conclusion

This study set out to answer a simple, but consequential question for risk management and empirical asset pricing: can a statistically disciplined choice of forecasting horizons improve one day ahead realised volatility forecasts in both linear, and nonlinear model candidates? By tying each model's "memory" to lags selected by a Bayesian Information Criterion (BIC) on an expanding window, and then mapping that choice into inputs for the model candidates of this study, the Heteogeneous Autoregressive (HAR) model, the Multilayer Perceptron (MLP) neural network, and the Long Short-Term Memory (LSTM) neural network, I find that out of sample (OOS) accuracy sharpley increases when forecasts errors are most costly, while keeping the procedure transparent and free of test leakage.

Using SPY realised volatility (RV) constructed from high frequency trades over 1996 to 2024, I take fixed horizon designs to each model candidate with a daily, weekly, and monthly component based on the heterogeneous markets hypothesis, and compare it against model candidates whose horizons were chosen to an $AR(p)$ BIC minimising process. Across these model candidates, BIC variants of each model reduce loss relative to otherwise identical specifications. On SPY, MSE falls by 3.1% for HAR, 15.8% for MLP, and 3.4% for LSTM, where QLIKE falls by 2.7%, 4.2%, and 0.5% respectively. The LSTM with BIC selected memory achieves the lowest average losses overall (0.028924 QLIKE, 0.001424 MSE). Diebold-

Mariano (DM) tests confirm the statistical significance of these gains within each model candidate, all with significance at the 1% level. The time profile of improvements explains why statistical discipline matters. The cumulative error differentials step up notably during the Global Financial Crisis and COVID-19 for all three model candidates, indicating that the criterion expands effective memory when long range dependence intensifies during crisis periods. Between crises, the curves are largely flat, consistent with the idea that fixed horizons suffice in placid periods and that disciplined parsimony prevents overfitting. These patterns align with the study's hypothesis (H_3), that the advantage of information criterion disciplines is largest in high volatility regimes.

Importantly, the gains translate into economic value under a realised utility (RU) framework for a mean variance investor who sizes exposure with model implied volatility. On SPY, BIC variants raise certainty equivalent utility by 1.70 bps for HAR, 1.47 bps for MLP and 0.27 bps for LSTM, closing 5.96%, 5.17%, and 1.01% of the residual gap to the theoretical upper bound, respectively. Because the portfolio rule penalises under prediction of risk more heavily than overprediction, these utility improvements are driven by the reduction of outsized errors in turbulent markets, rather than by small average gains in calm periods.

External robustness checks point in the same direction. Repeating the exercise for liquid sector ETFs for Technology, Consumer Discretionary, and Health Care, shows that BIC guided horizons generally lower MSE and QLIKE within each model candidate, with the largest gains in the Consumer Discretionary where persistence lengthens the most in stress. Modified DM tests across sectors again favour the BIC designs with strong statistical significance. These results support the study's hypothesis (H_1), in that nonlinear learners can outperform linear benchmarks on average, but the advantage is more robust when their memory is disciplined. FNNs particularly benefit from statistical disciplining, while RNNs, whose gating already provides adaptive memory, gains more modestly from tying the lookback length to the data. Finally, robustness to the choice of penalty confirms the mechanism rather than a specific criterion. Substituting the Akaike Information Criterion (AIC) and

Hannan-Quinn Information Criterion (HQC) preserves the main message, that statistically disciplined horizons statistically outperforms fixed horizons, with the strongest models being HQC or BIC: the two strongest penalisers for complexity, preferring most parsimonious models.

The central contribution of this study is to show that a light touch, information criterion rule implementation, can make both traditional econometric models and neural networks more reliable forecasters of realised volatility. By letting the data choose how far back to look, the procedure lowers loss functions, raises investor utility, and does so most when it matters most. Furthermore, to the best of my knowledge, it is the first systematic application of BIC guided memory selection to neural network volatility forecasting, and overall, demonstrates that statistical discipline can "open to black box" without sacrificing performance. In a field where flexibility often comes at the cost of interpretability, this design restores it without sacrificing performance, offering a practical blueprint for risk modellers and an empirically grounded result for the literature on volatility forecasting with machine learning.

Limitations. While this study provides evidence that statistically disciplined memory selection improves volatility forecasting across multiple model architectures and robust to alternative assets, statistical inferences, and information criterion, several limitations warrant discussion and suggest directions for future refinement.

Firstly, the empirical choices on selecting a mapping, while theoretically justified to an extent, have limitations. Firstly, the upper bound p_{max} trading days (approximately two months) reflects a judgement about the maximum plausible memory length in daily volatility, but lacks formal justification. While extending p_{max} further would allow BIC to detect even longer dependencies, it would also increase the risk of spurious lag selection in finite samples and impose computational costs that grow quadratically with the grid size. Similarly, the pivot at $p^{*} = 10$ for mapping AR orders to heterogeneous horizons, while motivated by the distinction between medium and long term scales, remains empirically unjustified. A data driven selection of this threshold, perhaps via structural break tests or cross validation over the pivot choice, could provide a more principled approach, though at the cost of additional complexity and computational costs that are unfeasible at this level.

Similarly, models are re-estimated annually rather than more frequently (e.g., monthly or quarterly) due to computational constraints. This design choice follows similar volatility forecasting and asset pricing literature (Zhang et al., 2023, Gu et al., 2020), but implies that the selected lag structure p^* remains fixed for an entire year, even as market conditions or regimes change within that window During periods of rapid regime change, the annual update may lag behind shifts in true dependence structure, causing the model to under-adapt in the early phase of a volatility spike, or over-persist after conditions normalise. This was, at first glance, evident in the cumulative sum of squared error (CSSE) plotting of the fully linear HAR model, which was shown being unable to normalise post-GFC. More frequent re-estimation would allow the BIC mechanism to track persistence changes more closely, but would require greater computational resources appropriate for future research.

The empirical analysis also focuses on four highly liquid US equity ETFs, and it is untested whether or not these results may generalise to other asset classes (such as fixed income, foreign exchange, or commodities), or other geographies (emerging markets, Europe, Asia, or Oceania). Furthermore, the study focuses exclusively on one day ahead forecasts of realised daily volatility, constructed from high frequency transaction data. This design choice reflects the availability of clean, microstructure robust QMLE estimates for liquid ETFs, but limits the scope of application. Many institutional risk management and portfolio allocation problems operate at weekly or monthly horizons, and many assets lack the high frequency trade data necessary to construct reliable volatility measures. Taking into account that many forecasting research, such as Zhang et al. (2023)'s studies, exploit cross sectional information or cross sectional volatility or a high dimensional panel extension with macroeconomic predictors, a natural extension of this study arises that mitigates its limitations.

Future Research Directions. Therefore, a natural next step to this study is to evaluate the memory/lag selection discipline into a true model selection discipline, especially in nonlinear NNs with an effective degrees of freedom (edf) measure that reflects a learner's sensitivity to the data. Such research has been recently evolving in the computer science and mathematics literature (Gao and Jojic, 2016), however to the best of my knowledge, has currently not had any implementations for high dimensional volatility forecasting using machine learning and edf. Embedding such an edf based penalty into a "BIC style" score would allow dynamic selection not only of horizons but also of the network architecture itself (depth, width, regularisations and hyperparameters), in a unified, time series validation framework. Paired with panel level extensions to high dimensional forecasting with a large amount of predictors, this would move the discipline from a memory lengthening/shortening exercise to a comprehensive, statistically coherent selection rule and regularisation technique for modern nonlinear forecasting systems that has yet to have been explored in the financial economics literature.

Bibliography

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Andersen, T. G., & Bollerslev, T. (1998). Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. *International Economic Review*, *39*(4), 885. https://doi.org/10.2307/2527343

Andersen, T. G., Bollerslev, T., Diebold, F. X., & Labys, P. (2003). Modeling and forecasting realized volatility. *Econometrica*, *71*(2), 579–625. https://doi.org/10.1111/1468-0262.00418

Bee, M., Dupuis, D. J., & Trapin, L. (2019). Realized peaks over threshold: A time-varying extreme value approach with high-frequency-based measures. *Journal of Financial Econometrics*, *17*(2), 254–283. https://doi.org/10.1093/jjfinec/nbz003

Bellman, R. E. (1961, December 31). *Adaptive control processes.* https://doi.org/10.1515/9781400874668

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, *31*(3), 307–327. https://doi.org/10.1016/0304-4076(86)90063-1

Bollerslev, T., Patton, A., & Quaedvlieg, R. (2018). Modeling and forecasting (un)reliable realized covariances for more reliable financial decisions. *Journal of Econometrics*, *207*(1), 71–91. https://doi.org/10.1016/j.jeconom.2018.05.004

Bu, R., Hizmeri, R., Izzeldin, M., Murphy, A., & Tsionas, M. (2022). The contribution of jump signs and activity to forecasting stock price volatility. *Journal of Empirical Finance*, *70*, 144–164. https://doi.org/10.1016/j.jempfin.2022.12.001

Bucci, A. (2020). Realized volatility forecasting with neural networks. *Journal of Financial Econometrics*, *18*(3), 502–531. https://doi.org/10.1093/jjfinec/nbaa008

Christensen, K., Siggaard, M., & Veliyev, B. (2022). A machine learning approach to volatility forecasting. *Journal of Financial Econometrics*, *21*(5), 1680–1727. https://doi.org/10.1093/jjfinec/nbac020

Corsi, F. (2008). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, *7*(2), 174–196. https://doi.org/10.1093/jjfinec/nbp001

Da, R., & Xiu, D. (2021). When moving-average models meet high-frequency data: Uniform inference on volatility. *Econometrica*, *89*(6), 2787–2825. https://doi.org/10.3982/ecta15593

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics*, *13*(3), 253. https://doi.org/10.2307/1392185

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, *50*(4), 987. https://doi.org/10.2307/1912773

Engle, R. F., & Lee, G. G. J. (1999, October 7). A long-run and short-run component model of stock return volatility. In *Cointegration, causality, and forecasting: A festschrift in honour of clive w. j. granger* (pp. 475–497). Oxford University Press. https://doi.org/10.1093/oso/9780198296836.003.0020

Feng, G., Giglio, S., & Xiu, D. (2020). Taming the factor zoo: A test of new factors. *The Journal of Finance*, *75*(3), 1327–1370. https://doi.org/10.1111/jofi.12883

Gao, T., & Jojic, V. (2016). Degrees of freedom in deep neural networks. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1603.09260

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

Hannan, E. J., & Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, *41*(2), 190–195. https://doi.org/10.1111/j.2517-6161.1979.tb01072.x

Hansen, P. R., & Lunde, A. (2006). Realized variance and market microstructure noise. *Journal of Business and Economic Statistics*, *24*(2), 127–161. https://doi.org/10.1198/073500106000000071

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

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*(5), 359–366. https://doi.org/10.1016/0893-6080(89)90020-8

Leites, J., Cerqueira, V., & Soares, C. (2024). Lag selection for univariate time series forecasting using deep learning: An empirical study. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2405.11237

Lin, Z., Tian, F., & Zhang, W. (2022, January 1). Evaluation and analysis of an LSTM and GRU based stock investment strategy. In *Advances in economics, business and management research* (pp. 1615–1626). https://doi.org/10.2991/978-94-6463-052-7_179

Liu, C. (2025). Financial data prediction based on LSTM model. *Procedia Computer Science*, *262*, 1173–1179. https://doi.org/10.1016/j.procs.2025.05.157

Lütkepohl, H. (2005, January 1). *New introduction to multiple time series analysis*. https://doi.org/10.1007/3-540-27752-8

Marshall, B. R., Nguyen, N. H., & Visaltanachoti, N. (2013). ETF arbitrage: Intraday evidence. *Journal of Banking & Finance*, *37*(9), 3486–3498. https://doi.org/10.1016/j.jbankfin.2013.05.014

Matsuda, T., Uehara, M., & Hyvarinen, A. (2022). Information criteria for non-normalized models. *Journal of Machine Learning Research*. https://doi.org/10.48550/arxiv.1905.05976

McInerney, A., & Burke, K. (2024). A statistical modelling approach to feedforward neural network model selection. *Statistical Modelling*, *25*(4), 323–342. https://doi.org/10.1177/1471082x241258261

MSCI. (2018, November). *Cyclical and defensive sectors indexes methodology*. MSCI. Retrieved November 8, 2025, from https://www.msci.com/eqb/methodology/meth_docs/MSCI_Cyclical_and_Defensive_Sectors_Indexes_Methodology_Nov18.pdf

Müller, U. A., Dacorogna, M. M., Davé, R. D., Olsen, R. B., Pictet, O. V., & von Weizsäcker, J. E. (1997). Volatilities of different time resolutions—analyzing the dynamics of market components. *Journal of Empirical Finance*, *4*(2-3), 213–239. https://doi.org/10.1016/S0927-5398(97)00007-8

Pástor, Ľ., & Veronesi, P. (2009). Technological revolutions and stock prices. *American Economic Review*, *99*(4), 1451–1483. https://doi.org/10.1257/aer.99.4.1451

Patton, A. (2010). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics*, *160*(1), 246–256. https://doi.org/10.1016/j.jeconom.2010.03.034

Patton, A. J., & Sheppard, K. (2009, January 1). Evaluating volatility and correlation forecasts. In *Handbook of financial time series* (pp. 801–838). https://doi.org/10.1007/978-3-540-71297-8_36

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 15–18. https://doi.org/10.1214/aos/1176344136

Souto, H. G., & Moradi, A. (2023). Forecasting realized volatility through financial turbulence and neural networks. *Economics and Business Review*, *9*(2). https://doi.org/10.18559/ebr.2023.2.737

Taylor, N. (2022). The determinants of volatility timing performance. *Journal of Financial Econometrics*, *21*(4), 1228–1257. https://doi.org/10.1093/jjfinec/nbac002

Xiu, D. (2010). Quasi-maximum likelihood estimation of volatility with high frequency data. *Journal of Econometrics*, *159*(1), 235–250. https://doi.org/10.1016/j.jeconom.2010.07.002

Zeng, M., Liao, Y., Li, R., & Sudjianto, A. (2022). Local linear approximation algorithm for neural network. *Mathematics*, *10*(3), 494. https://doi.org/10.3390/math10030494

Zhang, C., Zhang, Y., Cucuringu, M., & Qian, Z. (2023). Volatility forecasting with machine learning and intraday commonality. *Journal of Financial Econometrics*, *22*(2), 492–530. https://doi.org/10.1093/jjfinec/nbad005

Appendix A

Validity of Data Source

In order to assess the validity of the selected dataset, compare QMLE SPY volatility and the S&P 500 Volatility from the Realised Library by the Oxford-Man Institute of Quantitative Finance, a now discontinued dataset that was widely used for market microstructure robust volatility measurement (Bee et al., 2019; Taylor, 2022). After aligning the earliest and maximum dates, dropping missing values, and annualising the Realised Library, I test for the correlation between the two datasets (Table A.1).

	r	95% CI	R^2
Values	0.903	[0.898, 0.908]	0.82

Table A.1: Correlation between Dacheng Xiu's SPY QMLE series and the Oxford-Man SPX realised volatility measure (2000 to 2022). The high correlation confirms that the datasets capture nearly identical volatility dynamics despite methodological differences.

Across 4,866 trading days [January 3rd 2000, February 25th, 2022], the two datasets exhibit an extremely high linear association ($r = 0.903$ with CI bounds [0.898, 0.908], $R^2 = 0.82$), indicating they capture essentially the same signal up to a modest scaling or offset attributable to methodological differences.

Appendix B

Hyperparameter Selection

For the HAR model, there is no hyperparameter that needs to be tuned. The hyperparameters for the MLP and LSTM NNs are summarised as follows:

	MLP	**LSTM**
Learning rate	0.001	0.001
Early stopping rounds	10	10
Ensemble	10	10
Batch size	512	512
Epochs	100	100
No. of hidden layers	3	2
Batch normalization	✓	✓

Table B.1: Hyperparameters for MLP and LSTM

The hyperparameters were selected following Zhang et al. (2023), with the exception of batch size, consistent with the fact that they "pool" data from multiple assets to generate a forecast, allowing for a larger batch size of 1024 per epoch. Reflecting this study's approach, where only an asset's past RVs are used as predictors, a more appropriate batch size of 512 is used.