Thesis Proposal: Machine Learning and Volatility Models

Liam Andrew Beattie[a]

[a]*Stellenbosch University, South Africa*

1. Introduction

Here are the papers I intend to read and give a brief summary of them:

Gunnarsson, Isern, Kaloudis, Risstad, Vigdel & Westgaard (2024: 33) Horvath, Muguruza & Tomas (2021) Petrozziello, Troiano, Serra, Jordanov, Storti, Tagliaferri & Rocca (2022) Wu, Cheng, Jankovic & Kolanovic (2024) Wu, Jankovic, Kaplan & Lee (2025) Zeng & Klabjan (2019) Zhang, Zhang, Cucuringu & Qian (2023)

Zhang *et al.* (2023)

Data

To write a comprehensive, focused quantitative econometric paper on machine learning and volatility models with less emphasis on an extensive literature review, you should prioritize a **clearly defined and narrow research question** and a **rigorous quantitative analysis** directly addressing that question. Here's how you can structure your paper, drawing on the provided sources:

I. Introduction and Focused Research Question:

- Start with a concise introduction that immediately highlights the specific gap you are addressing in the literature or a particular problem you are investigating.
- **Clearly state your focused research question(s)**. For instance, instead of broadly investigating if ML can forecast volatility better than traditional models, you might focus on:

Email address: 22562435@sun.ac.za (Liam Andrew Beattie)

  - The performance of a specific novel ML architecture (e.g., an attention-based deep learning model) in forecasting the implied volatility of a particular asset class (e.g., S&P 500 options) compared to a specific benchmark (e.g., a specific GARCH extension or HAR model).
  - The impact of a specific type of high-frequency intraday data (e.g., order book data, transaction volume) as features in an ML model for short-term realized volatility forecasting for a specific set of stocks.
  - The effectiveness of a particular XAI technique (e.g., SHAP values) in interpreting the predictions of a specific ML model applied to implied volatility surface modeling.
  - The economic value (e.g., through backtesting trading strategies) of volatility forecasts from a specific ML model compared to a standard econometric model for a defined set of assets.
- Briefly mention the contribution of your focused study to the existing literature without an exhaustive review at this stage. You can refer to recent reviews to justify the relevance of your specific angle.
- Outline the structure of your paper.

II. Data:

- Provide a detailed description of the specific dataset(s) you will use.
- Clearly define your volatility measure (e.g., realized volatility calculated from high-frequency returns, implied volatility from option prices, or volatility indices like VIX).
- Specify the asset class(es), time period, and sampling frequency. If using high-frequency data, explain the cleaning and aggregation methods.
- Describe the predictor variables you will use, justifying their selection based on previous findings (cite specific papers from the existing literature concisely) or theoretical arguments. If using exogenous data, clearly state its source and frequency.

III. Methodology:

- **Clearly and concisely present the econometric and machine learning models** you will employ. Focus on the mathematical specification of the models.
  - For ML models, describe the architecture, activation functions, optimization algorithm (e.g., Adam), and hyperparameter tuning process (e.g., cross-validation).

  - For benchmark econometric models (e.g., GARCH, HAR), provide their standard formulations.

- Explain your chosen training scheme (e.g., single asset, pooling data) and out-of-sample testing procedure (e.g., rolling window).
- If applicable, describe any feature engineering techniques or data transformations you will use.
- If your focus includes explainability, detail the XAI methods you will apply.
- If you are investigating computational aspects, describe the hardware or software used (e.g., FPGA technology).

IV. Evaluation Metrics:

- Specify the **quantitative evaluation metrics** you will use to compare the forecasting performance of your models. These could include statistical measures like Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), Quasi-Likelihood (QLIKE), and statistical tests for predictive accuracy like the Diebold-Mariano (DM) test.
- If your research question involves economic significance, clearly outline how you will evaluate the models from an economic point of view (e.g., through backtesting trading strategies, Value-at-Risk (VaR) analysis).

V. Empirical Results and Discussion:

- Present your quantitative results in a clear and organized manner using tables and figures.
- **Focus your discussion directly on answering your stated research question(s)** based on the empirical evidence.
- Compare the performance of your chosen ML model(s) against the benchmark econometric model(s) using the selected evaluation metrics.
- Analyze the statistical significance of any performance differences (e.g., using DM tests).
- Discuss the implications of your findings in the context of the specific niche you are investigating. Briefly connect your results to the broader literature, highlighting how your focused study contributes or contrasts with existing findings. You can be more selective in the literature you discuss here, focusing on the most directly relevant studies.

- If applicable, interpret the results of your XAI analysis, explaining which features are most important for your ML model's predictions.
- If you conducted an economic evaluation, discuss the practical implications of your findings.

VI. Conclusion and Future Work:

- Summarize your main quantitative findings and directly address your research question(s).
- Briefly reiterate the contribution of your focused study.
- Suggest specific avenues for future research that build directly on your findings and the limitations of your study.

Strategies for Reducing Literature Review:

- **Target a Very Specific Gap:** Instead of reviewing the entire landscape of ML in volatility, focus on a very narrow sub-topic where the existing literature is limited or where a specific technique has not been thoroughly explored (e.g., the application of a particular deep learning architecture to a niche market). The literature review can then be more concentrated on the immediate context of your research question.
- **Build Upon a Recent Survey:** Since comprehensive literature reviews exist, you can briefly summarize the state of the art based on these reviews and then immediately narrow down to the specific issue your paper addresses.
- **Focus on Methodological Innovation or Application:** If your paper introduces a novel ML technique or applies an existing one to a new and underexplored area of volatility modeling, the literature review can be more focused on the methodological background and the specific context of your application, rather than a broad overview of all volatility models or all ML techniques.
- **Assume Familiarity with Standard Models:** For well-established econometric models (like GARCH) and common ML algorithms (like basic neural networks), you can often provide a brief description and cite seminal works without an extensive historical review.

By adopting a focused research question and prioritizing a rigorous quantitative analysis, you can write a comprehensive and valuable econometric paper on machine learning and volatility models with a less extensive literature review. The key is to be precise in your question, thorough in your methodology and evaluation, and direct in your discussion of the results in relation to your specific focus.

Gunnarsson, E.S., Isern, H.R., Kaloudis, A., Risstad, M., Vigdel, B. & Westgaard, S. 2024. Prediction of realized volatility and implied volatility indices using AI and machine learning: A review. *International Review of Financial Analysis*. 93:103221.

Horvath, B., Muguruza, A. & Tomas, M. 2021. Deep learning volatility: a deep neural network perspective on pricing and calibration in (rough) volatility models. *Quantitative Finance*. 21(1):11–27.

Petrozziello, A., Troiano, L., Serra, A., Jordanov, I., Storti, G., Tagliaferri, R. & Rocca, M.L. 2022. Deep learning for volatility forecasting in asset management. *Soft Computing*. 26(17):8553–8574.

Wu, E., Cheng, P., Jankovic, L. & Kolanovic, M. 2024. *Investable AI for volatility trading deep learning model for cross asset volatility strategies*. (Research Report). J.P. Morgan Global Quantitative & Derivatives Strategy. [Online], Available: httpswww.jpmorganmarkets.com.

Wu, E., Jankovic, L., Kaplan, B. & Lee, T.S. 2025. *Cross asset volatility: Machine learning based trade recommendations*. (Research Report). J.P. Morgan Global Markets Strategy. [Online], Available: https://www.jpmorganmarkets.com.

Zeng, Y. & Klabjan, D. 2019. Online adaptive machine learning based algorithm for implied volatility surface modeling. *Knowledge-Based Systems*. 163:376–391.

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.