A machine learning hybrid approach to forecasting equity returns volatility: A South African perspective.

by

Student: Mr Tabataba Simon Tloubatla
Student number: TLBTAB001

SUBMITTED TO THE UNIVERSITY OF CAPE TOWN
In partial fulfilment of the requirements for the degree:

M. Com in Finance specialising in Corporate Finance and Valuations

Faculty of Finance and Tax

UNIVERSITY OF CAPE TOWN

Date of submission: 11 February 2024
Supervisor: Prof Francois Toerien

Faculty of Commerce, Department of Finance and Tax, University of Cape Town

Declaration page

You must include a signed and dated declaration in the front of your dissertation/thesis. Please use the standard format shown below:

DECLARATION

I, ***Mr TS Tloubatla***, hereby declare that the work on which this dissertation/thesis is based is my original work (except where acknowledgements indicate otherwise) and that neither the whole work nor any part of it has been, is being, or is to be submitted for another degree in this or any other university.

I empower the university to reproduce for the purpose of study either the whole or any portion of the contents in any manner whatsoever.

Signature: …Signed by candidate…

12/02/2024

Date: …………………………………….

A machine learning hybrid approach to forecasting equity return volatility: a South African perspective.

Abstract

For many years, scholars and professionals in the financial markets have been deeply interested in the forecasting of financial market return volatility. There are many methods for predicting the volatility of financial market returns, and various studies have indicated differing degrees of accuracy in this regard. Research on describing the effectiveness of various approaches under various conditions is still ongoing. This field has moved from simple econometric methodologies like Moving Averages (MA), ARCH-type models and stochastic volatility, to more complex models like LSTM (Long-Short-Term-memory) and SVM (Support Vector Machines) (specifically machine learning algorithms). Machine learning in various forms is currently being explored as an alternative for forecasting the volatility of financial market returns. In this study this exploration is continued by considering a hybrid-based methodology to forecast this volatility, specifically in the South African equities market. There are two guiding principles for this study. The first principle is that specific ARCH-type models that achieve a superior fit to the dataset in question (the JSE All Share Index in this study) can be used in combination with machine learning (ML) models to forecast the volatility in financial market returns. The second principle stems from the work of earlier authors who have demonstrated the suitability and use of LSTM as a ML model that is effective in generating hybrid volatility forecasting models in conjunction with other ARCH-type models. The first guidance is based on the view that accuracy in volatility prediction depends on the ability of a model or group of models to capture volatility stylised facts inherent in a time series dataset used. The approach is based on the idea that various econometric and machine learning forecasting models each have their own advantages and disadvantages, and that combining them results in a stronger forecasting approach. The search for ARCH-type models showing a superior data fit for the Johannesburg Stock Exchange All Share Index (JSE ALSI) revealed the following models as suitable: GARCH(G), EGARCH(E) and TGARCH(T). The study then applies the base econometric models to LSTM and produces seven hybrid models, namely G-LSTM, E-LSTM, T-LSTM, GE-LSTM, GT-LSTM, ET-LSTM and GET-LSTM. Additionally, the averaging of GARCH, EGARCH and TGARCH produces a simple average model. As such, in addition to LSTM and the base econometric models, twelve models are used in this study. This research considered daily prices of the JSE ALSI from January 2004 to December 2022, where 80% of the dataset was used for training purposes and 20% was used for testing purposes. Volatility for this dataset was modeled (both training and testing) using the above models. RMSE, MAE and MAPE were used to evaluate the differential out-of-sample performance of the different models. In addition, the Wilcoxon signed-rank test was used to evaluate the significance of the forecasts from the different models generated. The conclusion made is that well-tuned hybrid models outperform all standalone models, including the average model. Furthermore, based on the results of this study it can be argued that GET-LSTM, GT-LSTM and ET-LSTM are the most effective financial market returns forecasting models considered, at least as it relates to the South African equities market. Furthermore, more complex hybrid models generally dominate the simpler models as well as the traditional ARCH-type models considered.

Keywords/ Key phrases:

Volatility stylised facts, forecasting volatility, conditional volatility, machine learning, hybrid machine learning, loss functions, LSTM, hybrid volatility forecasting models.

Table of Contents

List of figures

List of tables

List of Acronyms

AIC: Akaike information criterion.

A-LSTM: see x-LSTM. This is a x-LSTM where x is the APARCH(p,q) model.

ADF: Augmented Dickey-Fuller test. A test for the presence of a unit root in a time series dataset.

ARCH: Autoregressive conditional heteroskedasticity. This is the model defined by Engle (1982).

ARCH-LM: ARCH Lagrange Multiplier test, a test for ARCH effects in a timeseries dataset.

ARMA: Autoregressive moving average.

BIC: Bayesian information criterion.

BFT: Behavioural Finance Theory.

GARCH: Generalised autoregressive conditional heteroskedasticity of Bollerslev (1986).

EA-LSTM: see x-LSTM. This is a x-LSTM where x is the EGARCH(p,q) and the APARCH(p,q) models.

EGARCH: Exponential GARCH model of Nelson (1991)

E-LSTM: see x-LSTM. This is a x-LSTM where x is the EGARCH(p,q) model.

EMH: Efficient market hypothesis.

EWMA: Exponentially weighted moving average.

G-LSTM: see x-LSTM. This is a x-LSTM where x is the GARCH(p,q).

GA-LSTM: see x-LSTM. This is a x-LSTM where x is the GARCH(p,q) and the APARCH(p,q) models.

GE-LSTM: see x-LSTM. This is a x-LSTM where x is the GARCH(p,q) and the EGARCH(p,q) models.

GEA-LSTM: see x-LSTM. This is a x-LSTM where x is the EGARCH(p,q) and the APARCH(p,q) models.

JSE ALSI: Johannesburg Stock Exchange FTSE All share index, with the code JALSH:IND on Bloomberg and .JALSH on Reuters. It has the index code J203 on the JSE itself. The index tracks all stocks listed on South Africa's Johannesburg Stock Exchange.

LSTM: Long-Short-Term-Memory. A type of a machine learning algorithm.

MA: Moving average.

MAE: Mean absolute error.

ML: Machine Learning.

MAE(U): MAE where a forecast below expectations (target realised volatility in this study) is more penalised.

MSE: Mean squared error.

MAPE: Mean absolute percentage error.

MSSR: The moving average SSR. See SSR.

RMSE: Root mean squared error.

RV: Realised volatility.

SSR: Sum of squared intra-day returns. The original methodology for calculating realised volatility.

SV: Stochastic volatility

SVM: Support Vector Machine

SD: Standard deviation

SL: Supervised learning

SSL: Semi-supervised learning

SVM: Support Vector Machine, a type of a machine learning algorithm.

UL: Unsupervised learning

VARHAC: Vector autoregressive heteroskedastic consistent estimator. A methodology for calculating realised volatility.

WMA: Weighted moving average

x-LSTM: LSTM that includes the output of one or more models as an input. The variable x is usually definitive of the input model used. e.g. G-LSTM. In this study it means a LSTM model where the forecast output of a GARCH(p,q) model is used as an input into the LSTM.

1. Introduction

1.1. Background

The pricing of options, application of managed volatility strategies, portfolio selection and rebalancing, as well as financial risk management, are some of the areas that depend on the ability to make accurate forecasts of the volatility of financial market equity returns (Liu et al., 2022; Poon & Granger, 2005; Pagel et al., 2007; Dreyer & Hubrich, 2019; Petroziello et al., 2022, etc.). Financial market professionals, including investment banks, institutional investors, asset management boutiques, government regulatory bodies and others, are specifically interested in this, because financial market returns volatility is generally used as a measure of risk (Poon & Granger, 2005; Petroziello et al., 2022; Petropoulos et al., 2022). Volatility is also associated with uncertainty (Poon & Granger, 2005), and is normally elevated around periods where there is macroeconomic uncertainty of both internal and external origin. Thus, the volatility of financial market returns can at times be attributed to this macroeconomic uncertainty (Gerlach et al., 2006; Redl, 2018; Binge & Boshof, 2020).

The volatility environment of South African financial markets has been relatively high in the recent past because of a combination of domestic and global economic and political factors. For example, the Johannesburg Stock Exchange (JSE) All Share Index (ALSI) saw a significant decline in its price during the early stages of the Covid-19 pandemic. The index closed at 38,2673.21 points on March 23, 2020, a drop of 35% from its value at the beginning of 2019. This was one of its highest declines in that period,and was a result of the global macroeconomic uncertainty caused by the Covid-19 pandemic.

Beyond this, macroeconomic turbulence has been observed at numerous times in South Africa's post-1994 history. Redl (2018) highlights the following as times of uncertainty in South Africa (and by implication sources of market volatility): the collapse of Long-term Capital Management and the Russian sovereign default around 2000, the excessive volatility of South Africa's currency in 2000 followed by a stagflation in 2002 centered around rand depreciation, the global financial crisis that started around 2008, the forced resignation of former South African president Thabo Mbeki around 2008, the Euro crisis in 2013, Brexit in 2016, uncertainty around the potential downgrade of South Africa's investment grade status around 2018, and the cabinet reshuffle and recall of finance minister Gordhan in 2018. The effect of Covid-19 (as stated above) and the Russia/Ukraine war which started in 2022, have also had an impact on the South African financial markets. These events, amongst others, have characterised the South African equity market as being volatile. This has given credence to the assertion of Masset (2011) that volatility tends to be more pronounced in emerging markets. Therefore, a more precise volatility forecasting methodology would be advantageous to financial market professionals invested in the JSE equity market.

Various methodologies exist that detail how the volatility of financial market equity returns can be forecasted. One of the popular approaches for doing this are Autoregressive Conditional Heteroskedasticity (ARCH)-type models, starting with the ARCH model of Engle (1982). Since 1982, a lot more models have been developed, collectively known as ARCH-type models, as they are in one way or another an innovation from the original ARCH model. These models have received a lot of attention, especially where it relates to the forecasting of volatility of financial markets (Brailsford & Faff, 1996). The industry, however, needs innovation to generate models that can predict volatility better, especially in view of its importance globally and in South Africa where there is a perception of high macroeconomic uncertainty.

In the past few years, there has been a lot of interest in the application of machine learning techniques for forecasting the volatility of financial market equity returns (see, for example, Hyung et al., 2006; Atsalakis & Valavanis, 2009; Monfared & Enke, 2014, Kim & Won, 2018; Filipovic & Khalilzadeh, 2021; Petroziello et al., 2022; Kumbure et al., 2022, etc.**)**. These studies, amongst others, suggest that machine learning techniques have the potential to outperform traditional econometric models (including ARCH-type models) in terms of accuracy and robustness. It is claimed that this is because of their ability to capture complex non-linear relationships and patterns in financial market data.

Recent developments in this area include the application of hybrid approaches, which forecast volatility by combining machine learning techniques with conventional econometric models (Monfared & Enke, 2014; Verma, 2021; Kakade et al., 2022, Petropoulos et al., 2022, etc.). The advocates of this approach do not contest the fact that the individual conventional econometric models and machine learning techniques can serve as efficacious volatility prediction tools for several reasons. Rather, they promote a combinatorial approach that attempts to extract the best elements from various methodologies to produce an enhanced model. The approach is based on the view that each of these models have advantages and disadvantages, and that by combining them it may be possible to produce a model with more advantages than disadvantages (Verma, 2021, Petropoulos et al., 2022).

The specific works of Roh (2007), Kim and Won (2018), Verma (2021), and Kakade et al. (2022) propose a hybrid methodology that combines traditional econometric methodologies with machine learning. Specifically, they propose the combination of LSTM with ARCH-type methodologies. However, the general justification for the econometric methodologies used seems to be haphazard or ambiguous and does not appear to be specifically aimed at establishing relevance to the dataset being used as the basis of their research. The idea of volatility stylised facts - specific characteristics underpinning the behaviour of return volatility in a financial market time series dataset, namely volatility clustering, leptokurtosis, volatility asymmetry and mean reversion (Engle & Patton, 2001; Tsay, 2005; Kirchler & Huber, 2007; Daly, 2008; Masset, 2011) - follow closely when discussing volatility in a financial market time series dataset. Different datasets exhibit different behaviours of these volatility stylised facts. Therefore, it seems reasonable that in modelling the volatility of equity returns, the stylised facts thought to be existent and/or prevalent in a particular time series dataset ought to be targeted by the traditional econometric models used in the hybrid method. This is in line with the assertion of Engle and Patton (2001) that a model that accounts for volatility stylised facts is likely to model volatility better. Therefore, some measure of model fitness to the dataset in question is necessary before the econometric models are used. In this way, the hybrid method is likely to capture the volatility stylised facts better.

In line with all of the above, the purpose of this study is to contribute to the body of research focused on examining the application of LSTM in combination with ARCH-type models in predicting financial market returns volatility. The specific proposition is not different from that of Roh (2007), Kim and Won (2018), Verma (2021), and Kakade et al. (2022), in that ARCH-type models can be enhanced by using machine learning algorithms. In this study an enhancement is that ARCH-type models that are specifically seen to be a fit to the JSE ALSI time series dataset are combined with LSTM and used to forecast volatility. The study is conducted from the perspective of an investor with a JSE ALSI position, whose decisions would be aided by the ability to better forecast the volatility of returns. This concept of using a hybrid approach to volatility forecasting is still under development globally and in its infancy in the South African financial markets.

1.2. The aim of the study

The aim of this study is thus to explore if hybrid financial market return volatility forecasting techniques, specifically the use of LSTM combined with ARCH-type models that are seen as a fit to the JSE ALSI time series dataset, can outperform their standalone counterparts. The perspective in this study is from that of an investor who has an exposure to the JSE ALSI.

1.3. Study questions

The study poses several questions, which are addressed by the experiments specified later. The first is whether there is an improvement in the out-of-sample loss functions of hybrid models fitted to forecast the returns volatility of the JSE ALSI timeseries dataset when compared to base econometric models. If improvements are noted, the follow-up question (second question) is whether an effectiveness order can be deduced from the various models tested. This generates the need to quantify the improvements. The final question is whether the noted improvements are statistically significant.

1.4. Importance of this study

This study is important for investors with an exposure to the JSE ALSI, in that it potentially offers an improvement in forecasting the volatility of financial market equity returns. This is a contribution to the literature by enhancing the prediction of future return volatility, as well as adding insights to the debates surrounding the application of machine learning techniques in financial markets, especially in South Africa. Finally, the study adds to the literature on volatility forecasting and the relative effectiveness of different approaches. While the focus of this study is on South African equity markets, its application can be broadly extended to other financial markets. The novelty of the study is in applying hybrid models to a dataset where the constituent econometric models are specifically tailored for the dataset in question, as opposed to applying a one size fits all approach. At the time of writing, the author is not aware of any other study that applies this approach to the South African listed equity market.

1.5. Document map

The remainder of this document is structured as follows. Chapter 2 discusses the literature on various topics on returns volatility and volatility forecasting. Chapter 3 addresses the data used in this study and how it was specifically managed. Chapter 4 describes the research methodology. Chapter 5 covers the results and discussion of the experiments undertaken in this study, and specifically addresses the research questions. Finally, Chapter 6 (the conclusion section) gives a summary of the research and proposes ideas for further related research.

2. Literature Review

2.1. Introduction

According to Moreira and Muir (2019), a long-term investor who ignores volatility forfeits about 2.4% of returns annually in the financial equity markets. Their examination, which focused on US stock returns data from 1926 to 2015, is the basis of this opinion. If this can to some extent be predicted, then a prudent investor ought to take steps to avoid this. The argument made, which forms the foundation of this study's philosophy, is that volatility in financial markets (and thus volatility forecasting) must not be ignored. This assertion also follows along the empirical evidence that volatility is not constant (Engle, 1982; Bollerslev et al., 1992; Pagel et al., 2007; Daly, 2008) and as such a long-term position cannot be held under the assumption of a one-period volatility forecast. Non-constant volatility is termed heteroskedasticity (Engle, 1993). The fundamental tenet for taking stock return volatility into account is that investors are risk averse and as such the expectation of increased risk should lead to a change in investment behavior (Daly, 2008; Moreira & Muir, 2019). It is therefore not surprising that the prediction and incorporation of equity return volatility forecasts in investment decision making ranks as very important in portfolio management. This has become ever more so since the realisation in the international stock market that large swings in price movements have become more prevalent since the late 1980s (Brailsford & Faff, 1996).

The success of any volatility prediction model (regardless of where the forecasts are applied) lies in its out-of-sample forecasting power (Poon & Granger, 2003; Petropoulos et al., 2022). Thus, the key question in this regard is what the most accurate method is for forecasting out-of-sample volatility of financial market returns. This question is explored in this section through an analysis of existing academic literature to potentially identify the best set of models to combine such that a better prediction methodology can be created. To get to this, the literature review starts with a section defining volatility, followed by an exploration of its attributes and how it is measured. Then, by critically analysing the literature on volatility forecasting models and methods, an attempt will be made to identify the best combination of forecasting methods for application and testing in the context of the listed South African equity market. In line with the studies of Roh (2007), Kim and Won (2018), Verma (2020), and Kakade et al. (2022), the specific focus in this study are ARCH-type models in combination with LSTM. Their studies do not all focus on listed equity, but other asset classes like commodities. However, this study takes the view that the specific approaches from their studies can be replicated regardless of what their focus on the financial markets was.

2.2. What is volatility?

2.2.1. General review

A common definition of volatility is that it is the standard deviation of returns (Poon & Granger, 2003; Wuite, 2009), algebraically described as the second moment (a description of distance from the mean of a dataset) of the distribution of returns (Gerlach et al., 2006, Alagidede & Panagiotidis, 2009). Andersen et al. (2005) and Daly (2008) give an easily understandable definition of volatility: it is the degree to which a variable being studied (in this case, returns of the JSE ALSI) changes over time. The more the variable changes, the greater the volatility measure. This definition does not specifically account for whether the variance is negative or positive. In the context of equity financial market returns, while Pagel et al. (2007) note that volatility can almost be observable using intraday realised volatility, it is important to recognise that volatility is not observable (Andersen et al.; 2005; Tsay, 2005; Daly, 2008). The non-observability of volatility needs a further clarification as this is not a readily appreciated aspect of volatility. As an example, if realised volatility was to be calculated using three hundred days' worth of closing equity index prices, this would be akin to an estimation of volatility, as only summary, end of day prices are used. Intraday prices, especially for a liquid asset, may fluctuate a lot, and they can do so in a one-minute horizon. Over the three-hundred-day period, this could have occurred a lot. The notion that volatility is merely a latent feature, that it is not observable, originates from this viewpoint. Even though generally not observable, volatility has been shown to have various observable characteristics that can be seen in asset return computations (Engle & Patton, 2001; Tsay, 2005; Daly, 2008; Masset, 2011). These are called "volatility stylised facts" and are discussed further in Section 2.4.

While volatility is not the same as risk, it is used as an input into the calculation of risk (Daly, 2008). Thus, investors care about volatility because they care about risk. In line with this, Poon and Granger (2003) highlight that volatility is a proxy for risk. This is contrasted with the definition of Petropoulos et al. (2022), who sees volatility as a measure of risk. Risk is defined as the uncertainty of future returns (Pagel et al., 2007). Thus, the question is: does dispersion (volatility definition above) equate to uncertainty (risk)? Given the perspective that volatility is not observable, any measure made can only be an approximation and therefore a proxy. This is in line with the thinking expressed by Poon and Granger (2005), where the operative word used is "association" - i.e., volatility is associated with risk and uncertainty. It must also be noted that a perfect prediction of volatility does not mean a perfect prediction of risk, but merely a better probability measure of the likelihood of price movements up and down (Engle, 1993).

2.2.2. Standard Deviation (SD)

Standard deviation is the most used measure of volatility in financial markets where the objective is to assess the level of return deviations around a mean (Poon & Granger, 2003; Daly, 2008). In standard deviation, a computation of the sum of the squared errors around a mean for a certain specific time series is made. When used in financial markets as a proxy for risk, a normality assumption to the time series dataset must be made as standard deviation is a metric defined within the confines of a normal distribution (where 66% of observations fall within one standard deviation) (Poon & Granger, 2003).

2.2.3. Realised Volatility (RV)

Pagel et al. (2007) calculate realised volatility using various methods, including a traditional one. The traditional method takes the square root of the sum of the squared intra-day returns that are sampled at very short intervals. This involves the use of high frequency intra-day financial market data (Andersen et al., 2005). For a 24-hour market, daily realised volatility is defined as the sum of 288 intraday squared 5-minute returns, taken day by day. One of the greatest benefits of using realised volatility is that the lag in observations is reduced. This is why this is akin to an observation of volatility, which is normally seen as unobservable (a latent feature) at higher sampling durations like daily and weekly intervals (Andersen et al., 2005).

The normal convention, owing to the easy availability of daily data as opposed to intraday data, especially at extended periods, is to compute realised volatility using daily data. This is akin to averaging out all the intraday price movements and might thus not be as accurate, but in the absence of lower-level data this is usually done. Because of the way realised volatility is calculated, the measure is normally seen as natural benchmarks for the evaluation of other volatility forecasting methodologies (Andersen et al., 2005). In this study, daily data (daily return variation around a return mean) is thus used. This study therefore does not go any further in discussing the intra-day realised volatility estimators.

2.3. Underlying theory of equity volatility in financial markets

It is important to understand the theory underpinning the volatility of financial markets in the greater scheme of multiple theories relevant to concepts used in financial markets and investments. This can help in giving a general direction in terms of how best to predict return volatility. The general theory of financial investments rests amongst others, on the efficient market hypothesis (EMH) of Fama (1965,1970), which assumes investors to be rational and valuing securities rationally (i.e., they are efficient and unbiased processors of relevant information, and their decisions are consistent with utility maximisation) (Lo, 2004; Lawrence et al., 2007; Byrne & Brooks, 2008). Regardless of whether the EMH is of a weak form or a strong form, the assertion is that market prices incorporate all information rationally and instantaneously. Within the EMH, the tradeoff between risk and expected return is the foundation upon which theories like the CAPM and the APT as well other foundational theories of investments (e.g., MPT, PMPT, option pricing theory, etc.) are premised (Daly, 2008). These theories are rational constructs of investments and the investors themselves are expected to be rational. In the EMH, the value of an asset at time t is the discounted value of expected future cashflows (Shiller, 1988; Lawrence et al., 2007). The information relevant to the value calculation arrives (adjust) randomly and hence asset values are also expected to be random. However, the following anomalies have been noted which cannot be explained by the EMH: excessive volatility, high trading volume and the formation and bursting of bubbles (Lawrence et al., 2007). Shiller (1981) and Shiller (1988) specifically focus on excessive volatility. This has sparked what is called the volatility debate, where excessive movements around the expected price are seen as an anomaly that challenges the validity of the EMH. Shiller (2003) narrates that stocks show more volatility than what a theory of efficient markets predicts. Amongst other reasons, the presence of volatility clustering as well as excessive volatility in financial markets serves to question the total validity of this theory as an explanatory premise for financial markets (Shiller, 1981; Shiller, 1988; Lawrence et al., 2007).

Lawrence et al. (2007) try to come up with an update on the EMH to account for these anomalies. In this updated model, investor sentiment is proposed as the rationale for these anomalies, including excessive volatility. It can be argued that this proposal is linked to aspects of Behavioral Finance Theory (BFT) as explained by Byrne and Brooks (2008), in that investor sentiment is a behavioral bias that nevertheless is able to move the market. Indeed, the above noted anomalies in financial markets have led to the Behavioral Finance Theory (BFT) which challenges the concept of rational investors, and which argues that behavioral biases such as optimism, anger, anxiety and others are the explanatory variables of financial market behavior and observed volatility (Byrne & Brooks, 2008). BFT is at odds with EMH, suggesting that financial market volatility and other financial markets metrics are not always predicted accurately as they are due to investor irrationality, which is unpredictable. It can be concluded that if volatility results from irrational behavior, there cannot be any aggregate theory or model that can predict it. The presence of many volatility forecasting models which have been empirically shown to have varying degrees of efficacy is at odds with this. While the presence of the anomalies suggests that either EMH does not hold or that it does not completely hold, a view that financial markets can be completely explained by BFT can probably be rejected. In line with this, Shiller (2003) suggests that a behavioral based approach can be used to profit off anomalies noted as a result of market inefficiency.

An alternative to the behavioral finance theory and the EMH is the Adaptive Market Hypothesis (AMH) of Lo (2004), which in attempting to reconcile the behavioral aspects on investments as espoused by the BFT with the rational elements of the EMH, states that prices reflect as much information as dictated by the combination of the environment and the number of market participants behaving in a common manner. The implication arising from this is that to the extent that it exists, the relationship between risk and return is unstable over time. This implies that the equity risk premium is also time varying. Furthermore, arbitrage opportunities do exist from time to time and as such there is value in active portfolio management and the science of trying to make predictions over various aspects of financial markets as to be the wiser investor. Finally, based on the prevailing economic environment, certain strategies can be fruitful. It might happen that at some point they might not be fruitful, but this does not mean they will not be fruitful again in the future. This calls for an ever-vigilant investor who constantly tracks the economic environment. In this way, market volatility strategies have a place and can be very profitable when correctly timed.

While the markets can be random in nature and perfect agreement on the evolution of volatility over time is yet to be reached, many market practitioners are actively involved in attempts to forecast financial market risk (and volatility). This seems to be in line with the acceptance of the AMH view, where a middle ground is taken between the EMH and BFT, and in this study this is seen as the premise to financial market returns volatility forecasting. A precise articulation of what "middle ground" means is perhaps something that can be investigated further, Shiller (2003) argues that theories around market efficiency are not extremely wrong such that a consistent profiting through the use of a behavioral approach will occur. If the AMH is accepted, then it can be argued that it is a rationale supporting the use of both ARCH-type models and ML models. ARCH-type models can be expected to account rational construct within the volatility prediction, while ML would account for the irrational elements. ML is in principle rational, but the hidden layers make the embedded workings unseen, and the expectation is that this is how it accounts for the irrational elements.

2.4. Volatility stylised facts

2.4.1. Introduction

Volatility forecasting is difficult, with the literature containing conflicting evidence regarding the relative quality of different methodologies for undertaking volatility forecasts (Brailsford & Faff, 1996). However, the search for more suitable forecasting methodologies has been an important issue in the literature. This search continues in practice. The general literature as well as the discussion on some of the econometric models used to discuss volatility and volatility prediction (Tsay, 2005; Daly, 2008; etc.), reveal some foundational elements fundamental to the evolution of volatility over time as observed on a certain equity data series. These are called volatility stylised facts (Engle & Patton, 2001; Tsay, 2005; Daly, 2008; Masset, 2011).

When an attempt is made at predicting volatility these elements need to be accounted for through one or more of the model parameters (Poon & Granger, 2003; Engle & Patton, 2001). To this regard, Filipovic & Khalizadeh (2021) claims that ML techniques can account for these volatility stylised facts. The expectation is that success on this would result in a model with a reasonable forecasting ability. This section therefore addresses these volatility stylised facts, namely volatility clustering, leptokurtosis, volatility asymmetry, mean reversion and long-term memory. In exploring these stylised facts, some volatility prediction methodologies are covered without in only a limited amount of detail.

2.4.2. Volatility clustering

Large and small moves of either sign (return movements) cluster on specific periods such that small changes are followed by small changes and large changes are followed by large changes (thus appearing like a switch between high and low volatility states), indicating that asset returns are not independent across time (Engle, 1993; Cont, 2005; Francq & Zakoian, 2010; Masset, 2011). The volatility process has a positive autocorrelation with a rate of decay described by the values of the parameters alpha and beta, where the decay is slower the closer the sum of these two are to one (Cont, 2005). Put differently, the presence of a long lasting positive auto-correlation of absolute returns defines volatility clustering (Kirchler & Huber, 2007). The clustering effect is an indication of non-independence in financial markets returns (Cont, 2005). Various explanations exist in the literature for why return volatility clusters, but discussing these beyond what has been said in Section 2.3. is not within the scope of this study.

2.4.3. Fat tails/ heavy tails and leptokurtosis

Algebraically, leptokurtosis (an asymmetry/ skewness property) is measured by the third central moment of returns, while fat tails (a tail property) is defined by the fourth central moment (Tsay, 2005). Together with volatility clustering, this is the most cited volatility dynamic (Kirchler & Huber, 2007). Fat tails imply more frequent extreme values and hence a higher volatility measure (Masset, 2011). In recent times financial market return distributions have been found to be more leptokurtic (Masset, 2011). A more formal definition is given by Wuite (2009), which is that a random variable (in this case asset returns) is said to be leptokurtic if it exhibits positive kurtosis, and that the shape of a leptokurtic distribution shows more peak around the mean and a greater mass at the tails. It should be noted that leptokurtosis is not necessarily a term that defines volatility, but specifies non-normality to the return distributions, where the tails of a return distribution are not normally distributed (Poon & Granger, 2003; Daly, 2008). The EGARCH model has been shown to more accurately model the kurtosis phenomenon of the underlying return data (Daly, 2008).

Andersen et al. (2005) highlight a general need in econometrics for allowing for fat tailed distributions. A gravitation towards this would allow for more realistic distributions when dealing with time series datasets that involve financial asset returns. In line with this, a generally accepted view is that the t-distribution is fat tailed and can therefore be used instead of the normal distribution (Samouilhan & Shannon, 2008; Alagidede & Panagiotidis, 2009).

2.4.4. Volatility asymmetry/ leverage effects

Amongst others, Engle & Patton (2001), Gerlach et al. (2006) as well as Masset (2011) describe volatility asymmetry (also called the volatility leverage effect) as the occurrence of more pronounced volatility increases when asset prices drop than when they rise. The implication is that financial markets practitioners need to be more concerned about expected price drops as they come with, or are expected to, lead to higher volatility. This feature is not addressed by the ARCH and GARH models (Daly, 2008), and was arguably the biggest driver for the need of “more robust” volatility forecasting models away from these. This is where the GJR-GARCH and the EGARCH models (amongst others) become necessary as they have asymmetry parameters (Poon & Granger, 2005).

Hyung et al. (2006), however, documents that this volatility dynamic is short lived, fading away in the long term. This fact, however, would be cold comfort for an investor whose portfolio has taken massive losses due to the presence of higher volatility when the market was declining and forced to restructure the portfolio weights. The fact that this dynamic might be short lived does not remove the need to prepare for it. As such a model that incorporates this is important. Another interesting insight with regards to this is that financial market volatility in emerging markets appear to not have a significant asymmetric effect as compared to that in more mature markets (Masset, 2011). This conclusion is made from a review of 34 stock market indices under bull and bear conditions for the periods 1951 to 2010 where South Africa was also included (dataset from 1995). An overall skewness of -0.70 is reported. As shown in table 3, this study reports a skew of -0.35, albeit for a different time period.

2.4.5. Mean reversion

Engle and Patton (2001) describe a “normal” or mean level of volatility as the level to which the current higher or lower than normal volatility will eventually return to. In line with this, Masset (2011) outlines that the mean reversion is a swing around the long-term average volatility value of a data series. This feature is generally associated with long-term volatility and is therefore related to the long-term memory aspect of volatility as discussed below. The implication is that current information which tend to generate volatility spikes has no effect on long-term volatility forecasts. With mean reversion, volatility is expected to drop or rise back to its mean over time. In this way financial market volatility is described as a long memory stationary process with a mean reverting behavior. As a result, volatility peaks that may occur in the short term as a result of a crisis are expected to fade away in the long run.

Volatility clustering and mean reversion are closely related concepts (Masset, 2011) as mean reversion implies the drop of autocorrelation of absolute returns to zero after it was seen to be consistently positive as defined in volatility clustering (note that this does not mean that volatility drops to zero, but only the autocorrelation of past measures). The alpha and beta parameters of a GARCH(p,q) model can indicate the mean reversion factor of the data series: if the sum is less than one, then this means that the process is mean reverting and the autocorrelation will eventually drop to zero (Engle & Patton, 2001). The implication is that current information has no bearing on long-term volatility forecasting. However, this cannot be used to imply that an investor ought not to anticipate spikes in volatility and prepare adequately for it, as stated earlier. A simple example would be a portfolio with a stop loss at a certain level where volatility is likely high. Should this level be reached, the portfolio would be stopped, and the loss will be realised.

A further question raised by Engle and Patton (2001) on this is whether normal volatility (long-term average volatility) is constant over time or what this normal volatility is in the first place. An extension of this could be whether there is a different long-term average volatility for different markets and to the extent that there is, is it constant over time, and are there different decay parameters across markets? However, these are not necessarily the focus points of this study, and thus is not pursued any further in this thesis.

2.4.6. Long term memory/ Volatility persistence

This refers to changes in volatility at a certain point continuing to have longer lasting impact in the subsequent evolution of volatility, stated differently to mean that current volatility estimates, and lagged volatility estimates remain influential to forecasted volatility for longer periods (Cotter, 2005; Samouilhan & Shannon, 2008; Masset, 2011). Poon and Granger (2005) consider a strong volatility persistence as being synonymous with long memory. Hyung et al. (2006) indicate that a long memory series has autocorrelation values that decline slowly at a hyperbolic rate. The hyperbolic rate of decline means that the volatility initially declines rapidly with a subsequent slow decline (Cotter, 2005). The innovation towards ARFIMA (an extension of ARMA) and FIGARCH(an extension of GARCH) were aimed at addressing this issue (Masset, 2011). Samouilhan and Shannon (2008) note significant evidence of volatility persistence in both the long and the short term for the JSE ALSI, the half-life of such a persistence is found to be 5.6 days in the short term and 169 days in the long term. The alpha and beta parameters of a GARCH(p,q) model can indicate the persistence factor of the data series: if the sum is close to one, then this means that the data series has a long memory (Engle

& Patton, 2001). While Poon and Granger (2005) note that all time-series volatility models were designed to capture volatility persistence, GARCH(p,q) and GJR-GARCH(p,q) have memories that are too short to fit the long memory property.

2.5. Traditional methodologies for predicting financial market volatility

2.5.1. Introduction

It is widely claimed in the literature that financial market volatility is predictable (Engle, 1993) and forecastable (Engle, 1993; Poon & Granger, 2003; Poon & Granger, 2005). However, there are some authors such as Christoffersen and Diebold (2000) who argue that volatility is forecastable in the short-term, but not in the long-term. Pederzoli (2006) highlights that the predictability of financial market volatility stems from the fact that the squared returns of financial markets have empirically been shown to be autocorrelated and having a clustering pattern. Daly (2008) argues that volatility may be predictable only when GARCH effects are present. However, the presence of many and competing forecasting methodologies speaks to the quest to find models that generate the greatest forecasting accuracy. These are extensions of the discussions in Section 2.3. The assumption in this study is that the use of the correct models under correct specifications can render the volatility of financial market returns predictable in line with the Adaptive Market Hypothesis (AMH) as detailed in Section 2.3.

This section explores the broader literature to uncover the potential best ways of forecasting financial market volatility. Fundamentally the objective of volatility forecasting models is to capture one or more of the volatility stylised facts discussed above to achieve a more realistic volatility prediction (Engle & Patton, 2001).

There are many traditional econometric methodologies for forecasting the volatility of equity returns. These include AR(q), ARMA(q,p), ARIMA(q,r,p), GARCH(q,p) and stochastic volatility (SV) models (Atsalakis & Valavanis, 2013), random walk, historical mean, simple moving average, EWMA, simple regression, and GARCH(p,q) approaches (Poon & Granger (2003). This list is not exhaustive (e.g. Balaban et al., 2006), and within each methodology there could be many variations. The specific focus of this study is ARCH-type models.

2.5.2. ARCH type models

ARCH models are conditional heteroskedastic models, where the current volatility is a function of past errors and past, non-constant volatility (Tsay, 2005). The first model to be discussed is the ARCH(q) model of Engel (1982) (Equation 1). This computes volatility as a time-varying function of current price information (Andersen et al., 2005). The q term describes the number of lags for the ARCH term used in the model implementation. An ARCH process as expressed in this equation generates an autoregressive structure on the conditional variance which allows volatility clustering to occur (Corhay & Rad, 1994). This model is therefore able to mimic the volatility clustering that is seen in practice in financial markets. Within this model, an increase and decrease in returns have the same effect on volatility, as asymmetry is not considered. Furthermore, the conditional distribution error is assumed to be normally distributed (Corhay & Rad, 1994). While the model can capture volatility clustering, one of its drawbacks is that a long lag is usually required before it adequately fits the dataset in question and as such its flexibility is limited (Bollerslev et al., 1992). This led to the development of the GARCH model of Bollerslev (1987) (Equation 2), where simplicity is obtained through a generalisation of the ARCH model.

The GARCH(p,q) model reduces back to the ARCH(q) model when p is zero, as this causes the lagged conditional variance term to disappear. Just like in the ARCH model (and in the later models to be discussed), p and q relates to the number of lags with regards to the specific model terms, as can be seen in Table 2. In theory, any order of p and q can be used. In simple terms, the GARCH(p,q) model defines a conditional variance which is a linear function of the lagged squared return errors, as well as its own lagged conditional variance (Bollerslev, 1987; Corhay & Rad, 1994). Unfortunately, the GARCH equation as highlighted in Equation 2 does not account for asymmetry. The expectation then, is that negative and positive shocks of the same size have the same effect. The orders of p and q can be determined using the Box and Jenkins technique to autocorrelations (Bollerslev, 1987), even though most empirical implementations of this model have low values (not more than 2) of p and q (Bollerslev et al., 1992). Volatility clustering is incorporated into the model (Bollerslev, 1987). From inception it was unclear whether the model accounts for leptokurtosis, and further works by Corhay & Rad (1994) indicated that it does not fully account for this. The GARCH(p,q) model only depends on the quantity but not the sign of the lagged error term (Bollerslev et al., 1992), which is at odds with practical finance where an asymmetric volatility response occurs as a result of good and bad news. This is a major criticism of the model. Another criticism is that the parameter values are limited to zero and above (Andersen et al., 2005). Given this, as well as the inability to handle asymmetry, other improvements were implemented. At the time of writing, there were more than 30 such (see Bollerslev, 2008). Despite these many developments, a GARCH(1,1) generally does not perform worse (where it is worse) than the more complex models, albeit with simpler layout, and furthermore it is sometimes found to be the same or on par (Pederzoli, 2006). The next few models to be discussed are the most widely used innovations on the GARCH(p,q) approach. These are shown in Table 1. The first is the GARCH-M(p,q) model of Engle, Lilien and Robins (1987) (Equation 3).

The conditional mean of an GARCH-M (p,q) model specifies a tradeoff between risk and expected return (Bollerslev et al., 1992), where an increase in the mean return is associated with an increase in risk and vice versa (Daly, 2008). This model is particularly useful where the relationship between risk and return is considered. The next two related models are the GJR-GARCH of Glosten, Jagannathan and Runkle (1993) and the TGARCH(p,q) of Engle and Ng (1993) (Equations 6 and 7). These are asymmetric models whose specific intent is to cater for the asymmetric shock impact found in financial markets. When modeling asymmetric effects, good and bad shocks of the same magnitude have different volatility impacts. In the event that there is no asymmetry, these would simply default back to the GARCH(p,q). as such, the GJR-GARCH(p,q) is simply the asymmetric version of the GARCH(p,q) (Monfared & Enke, 2014). The EGARCH(p,q) model of Nelson (1991) (Equation 4) is a logarithmic form of the GARCH(p,q) equation, where unlike the earlier models, no restrictions are placed on the GARCH parameters (Bollerslev et al., 1992; Poon & Granger, 2003; Andersen et al., 2005; Daly, 2008). This model specifically attempts to address the limitations of the GARCH(p,q) model with regards to volatility asymmetry (as described in the section on volatility dynamics), parameter limitations imposed by the GARCH model, and volatility persistence (as described in the section on volatility dynamics) (Nelson, 1991). The last model discussed in this section is the APARCH model (Asymmetric Power ARCH) of Ding, Granger & Engle (1993) (Equation 5). This model introduces power transformation on the lagged parameters of a GARCH model while maintaining the asymmetric effects. This model has, amongst others, been shown to be more successful than the simple GARCH(p,q) in modelling the long-term memory of the FTSE 100 stock index for the period 1998 to 1999 where 5-minute time intervals were processed (Cotter, 2005). Cotter (2005) Bollerslev (2008) highlight that this model contains several different ARCH related specifications should the parameters take certain values. The model can thus reduce to the ARCH(p) model of Engle (1982), the GARCH model (p,q) of Bollerslev (1987), the TS-GARCH(p,q) model of Taylor (1986), the log-GARCH(p,q) model of Geweke (1986), the NGARCH model (p,q) of Higgins and Bera (1992), the GJR-GARCH model (p,q) of Glosten et al. (1993), and the TGARCH(p,q) model of Engle and Ng

(1993). This does not mean that when this model is implemented it contains all of these model specifications, but rather that given certain parameters the model can reduce to one of these seven models. The model reduces to a simple GARCH model for a δ of two and a Υ of zero, and the GJR-GARCH model for δ of two and the Υ of between zero and one as well as the TGARCH model for a δ of one and a Υ of zero (Bollerslev, 2008).

In trying to find the best models to incorporate into the hybrid model setup for this study, the above models were considered. Where ARCH type models are used, Andersen et al. (1997) cautions on the need for the correct specification of the p and q parameters (the number of lags). However, Poon and Granger (2003) advise that setting p and q to the value of one is an acceptable practice. Further, Samouilhan and Shannon (2008) claim that a GARCH(2,2) is a better model fit for the JSE ALSI. Based on this and other work, it would seem that it does not automatically follow that a GARCH(1,1) model is appropriate for the JSE ALSI data-set. Therefore, relevant measures need to be computed to evaluate which models fit the JSE ALSI dataset the best. Furthermore, it appears that different methodologies have been empirically shown to model specific volatility stylised facts (e.g., both the EGARCH and GJR-GARCH approaches can model volatility asymmetry). Thus, the question is which model is better suited to a specific dataset. AIC and BIC have been shown to be suitable tools that can evaluate model fit as well as the order of p and q in various ARCH-type models (Tsay, 2005; Petropoulos et al., 2022). This study therefore uses these tools to evaluate the ARCH-type models to be used in the context of this study.

From an empirical perspective, a ten-year review across eleven equity markets of developing and developed countries (South Africa was not included) where out-of-sample comparisons were made, Balaban et al. (2006) found ARCH-type models (specifically the GARCH, GJR-GARCH and EGARCH approaches) to be valid for the forecasting of volatility where there is a greater aversion to under-forecasting. The suggestion is that there are times when ARCH-type models are superior to other econometric models. They however put an interesting qualification on their conclusion, which is that exponential smoothing forecasting methods were found to be superior to ARCH-type methods when the direction of the forecast error is equally important. The implication of volatility asymmetry as explained in Section 2.4.4. is that a downward direction is more important. Furthermore, in a meta-study that reviewed 93 other research papers, Poon and Granger (2003) conclude that ARCH-type models that incorporate asymmetry perform better than the simple GARCH(p,q). A further review by

Brailsford and Faff (1996) of standard GARCH, GJR-GARCH and other volatility forecasting methodologies used on the Australian equity market concluded that there is no clear winner in terms of the models used to forecast volatility. From a South African equity market perspective, Samouilhan and Shannon (2008) conclude that more complex ARCH-type models, especially the asymmetric ones, provide the best volatility forecasts. ARCH-type models have shown effectiveness in various studies. However, it can be argued that the specific parameters of the JSE ALSI dataset are not like those in other equity markets (level of historical persistence volatility persistence, level of the volatility itself, number of exogenous political factors influencing the movement of the markets and a such volatility, etc.) and as such, a broad generalisation cannot be made. It is therefore reasonable that with each dataset measures need to be computed to evaluate the optimal model and model setup to be used.

This section and Tables 1, 2 (see next page), which summarise various aspects of ARCH-models as per the literature, discuss the specific details underpinning various models, but these features by themselves do not necessarily mean a better model fit on the JSE ALSI dataset. Measures like AIC and BIC are therefore important in this respect.

Table 1: ARCH-type models

Model	Model Author	Equation	No
ARCH	Engle (1982)	$h^2{}_t = \omega + \sum_{i=1}^{q} a_i \varepsilon^2{}_{t-\mathrm{i}}$	-1
GARCH(p,q)	Bollerslev (1987)	$h_t^2 = \omega + \sum_{i=1}^{q} a_i \varepsilon_{t-i}^2 + \sum_{\mathrm{j}=1}^{p} \beta_j h_{\mathrm{t-j}}^2$	-2
GARCH-M	Engle, Lilian & Robins (1987)	$h_t^2 = \omega + \sum_{i=1}^{q} a_i \varepsilon_{t-i}^2 + \sum_{\mathrm{j=1}}^{p} \beta_j h_{t-\mathrm{j}}^2$ Where the $conditional$ mean $is\ defined\ by$: $y_t = \delta h_t + \varepsilon_t$	-3
EGARCH(p,q)	Nelson (1991)	$Log(h_t^2) = \omega + \sum_{i=1}^{\mathrm{q}} (a_i * \lvert\frac{\sqrt{\varepsilon^2{}_{t-\mathrm{i}}}}{\sqrt{h^2{}_{t-\mathrm{i}}}}\rvert) + \sum_{j=1}^{\mathrm{q}} (\Upsilon_i * \frac{\sqrt{\varepsilon^2{}_{t-j}}}{\sqrt{h^2{}_{t-j}}}) + \sum_{k=1}^{\mathrm{p}} \beta_k * Log(h_{t-k}^2)$	-4
APARCH	Ding, Granger & Engle (1993)	$h_t^{\delta} = \omega + \sum_{i=1}^{q} \alpha_i (\lvert\varepsilon_{t-i}\rvert + \Upsilon_i \varepsilon_{t-i})^{\delta} + \sum_{j=1}^{p} \beta_j h_{t-j}^{\delta}$	-5
GJR-GARCH	Glosten, Jagannathan & Runkle (1993)	$h_t^2 = \omega + \sum_{i=1}^{q} \left(a_i \varepsilon_{t-i}^2 + \Upsilon_i \varepsilon_{t-1}^2 I(\varepsilon_{t-i} < 0)\right) + \sum_{j=1}^{p} \beta_j h_{t-j}^2$	-6
T-GARCH	Zakoian (1994)	$h_t^2 = \omega + \sum_{i=1}^{q} (\alpha_i \lvert\varepsilon_{t-i}^2\rvert) + \sum_{j=1}^{\mathrm{q}} \left(\Upsilon_j \lvert\varepsilon_{t-j}^2\rvert I\left(\varepsilon_{t-j} < 0\right)\right) + \sum_{k=1}^{\mathrm{p}} \beta_k h_{k-j}^2$	-7

Table 2: Parameters used in ARCH-type models

Model	Description	Implication
a	ARCH coefficient (alpha)	The impact of past squared residuals.
β	GARCH coefficient (beta)	The impact of past conditional volatility.
h_t^2	Conditional volatility at t	The variable being measured, at earlier t, it is a lagged volatility.
ω	Constant term (Omega)	In an x-y grid, it represents the y intercept. Related to the capture of long-term volatility.
Υ	Leverage parameter	The impact of an asymmetric shock.
I	Threshold parameter	It is 1 when return is less than zero, otherwise it is 0.
δ	Power parameter (Gamma)	Degree of impact of the leverage factor. In the GARCH-M equation this variable represents the impact of volatility.
u	Mu/constant	The average long-term volatility.
p and q	The lags in a model	Determines the number of lags on past errors and past conditional volatility to be used in computing current volatility, as well as a lag on asymmetric responses to positive and negative shocks.
y	Variable for the mean equation in a GARCH-M model	This is the conditional part in the GARCH-M equation, where the conditional mean incorporates the conditional variance.

2.6. Predicting financial markets volatility using machine learning

2.6.1. Introduction

In recent times, there has been a move towards machine learning (ML) as a basis for volatility forecasting (Verma, 2021; Filipovic & Khalizadeh, 2021; Kakade et al., 2022; etc.), mainly to move away from the many data assumptions that need to be made when using traditional methodologies. One of the greatest motivations for using ML is that it allows for volatility forecasting without the need to make any assumptions about the underlying data distributions. Thus, ML has been empirically shown to capture volatility stylised facts without having to make any assumptions about the distribution of the stock returns of the underlying data (Filipovic & Khalilzadeh, 2021).

Broadly defined, ML is the art of programming computers to learn from data (Geron, 2019). The method of learning depends on the type of data as well as the algorithm used. This definition is very general and does not specifically speak to the purpose of the learning. Within the context of this study, ML has the specific purpose of teaching a program to learn from data such that it can predict the future - specifically the volatility of financial market returns. There are many algorithms/methodologies that can be used, and even generating a classification of the different methodologies can be a challenge. As per Geron (2019), the broad categories of ML are (1) supervised learning (SL) where a program is given data and the objective output of this data using methods like linear regression, support vector machines, neural networks (including LSTM), (2) unsupervised learning where blank data (data or dataset without a specifically defined target) is given using methods such as clustering and anomaly detection, and (3) semi-supervised learning (SSL), where partially classified data is given. For all of these the objective is that the program must learn patterns from the provided data to generate an output.

SVM and LSTM are amongst the most used methods in financial market forecasting. SVM is a classification tool where classes of data are generated in line with the target output using both linear and nonlinear classifier lines (Geron, 2019; Vasilev et al., 2019). LSTM on the other hand, applies incremental weights to data for use in generating a target value. The next section discusses the relative performance of these two approaches as documented in the literature.

2.6.2. The use of machine learning in financial market return volatility forecasting

Atsalakis & Valavanis (2009) reviewed 100 papers with 43 specifically focusing on financial market volatility. The results were mixed: some studies conclude that machine learning could be an effective financial market volatility forecasting tool when compared to traditional econometric methods, while others are of the opposite view. Specifically, neural networks, fuzzy logic and genetic algorithms have been suggested as the more successful machine learning options. Geron (2019) highlights the most important neural networks as feedforward neural networks, recurrent nets, long-short term memory (LSTM) and others. Another meta-analysis is by Kumbure et al. (2022) where 139 articles between 2000 and 2019 were considered and a conclusion made that LSTM and LSTM derivatives are some of the most effective forms of neural networks. These studies are all general reviews, key individual studies are considered in the section that follows.

Petrozziello et al. (2022) concluded that a LSTM derivative perform better as forecasting tools than GARCH(p,q) and GJR-GARCH(p,q) models in high volatility periods, but does not outperform in a normal market. Fiplovic and Khalizadeh (2021) showed that in predicting volatility LSTM outperforms other machine learning models in the US over the period 1992 to 2016. Liu et al. (2022) also concluded that LSTM and SVR are better at predicting volatility than GARCH(1,1) models. From a South African perspective, Balusik et al. (2021) found that LSTM is the better machine learning option for time series prediction in finance when compared to an ARIMA time series model. With regards to which one is the better model between SVM and LSTM, there is still no consensus, and further studies are warranted. With regards to LSTM, Kumbure et al. (2022) highlights that there have been some issues noted such as the tendency to overfit and the time complexity involved.

The use of machine learning models in isolation to more accurately predict the volatility of financial markets may, however, not be optimal. Several recent studies advocate an ensemble or hybrid approach to volatility forecasting, involving the combination of multiple models to address as much of the volatility stylised facts as possible. Those in support of the use of hybrid methodologies do not dispute that individualised machine learning algorithms can be better than traditional models in predicting financial market volatility, or that at times econometric methodologies can be better. Rather, this approach is premised on the view that all these models have strengths and weaknesses and that combining them might create a methodology that optimises the trade-off between strengths and weaknesses (Verma, 2021). Kim and Won (2018) used hybrid models to forecast volatility – specifically the volatility of the KOSPI 200 equity index was predicted using a hybrid model that considers a combination of GARCH(p,q). EGARCH(p,q), EWMA and LSTM approaches. The study concluded in favor of more complex hybrid setups. A similar study by Verma (2021) investigates crude oil price volatility using a hybrid model consisting of LSTM, GARCH(p,q) and GJR-GARCH(p,q) approaches, and similarly concluded that more complex hybrid setups perform better. In conclusion, Kakade et al. (2022) used GARCH(p,q), EGARCH(p,q) and TGARCH(p,q) models to make volatility predictions in the context of the Indian commodities. In all these studies, the choice of the base econometric models, as well as the order of p and q, seem to be arbitrary. However, the above studies all provide empirical evidence of the superior performance of hybrid volatility prediction models.

2.7. The volatility of the South African equity financial market

Figure 1 (below) shows the returns and volatility of the All Share Index (ALSI) of the Johannesburg Stock Exchange (JSE) from January 2004 to December 2022.

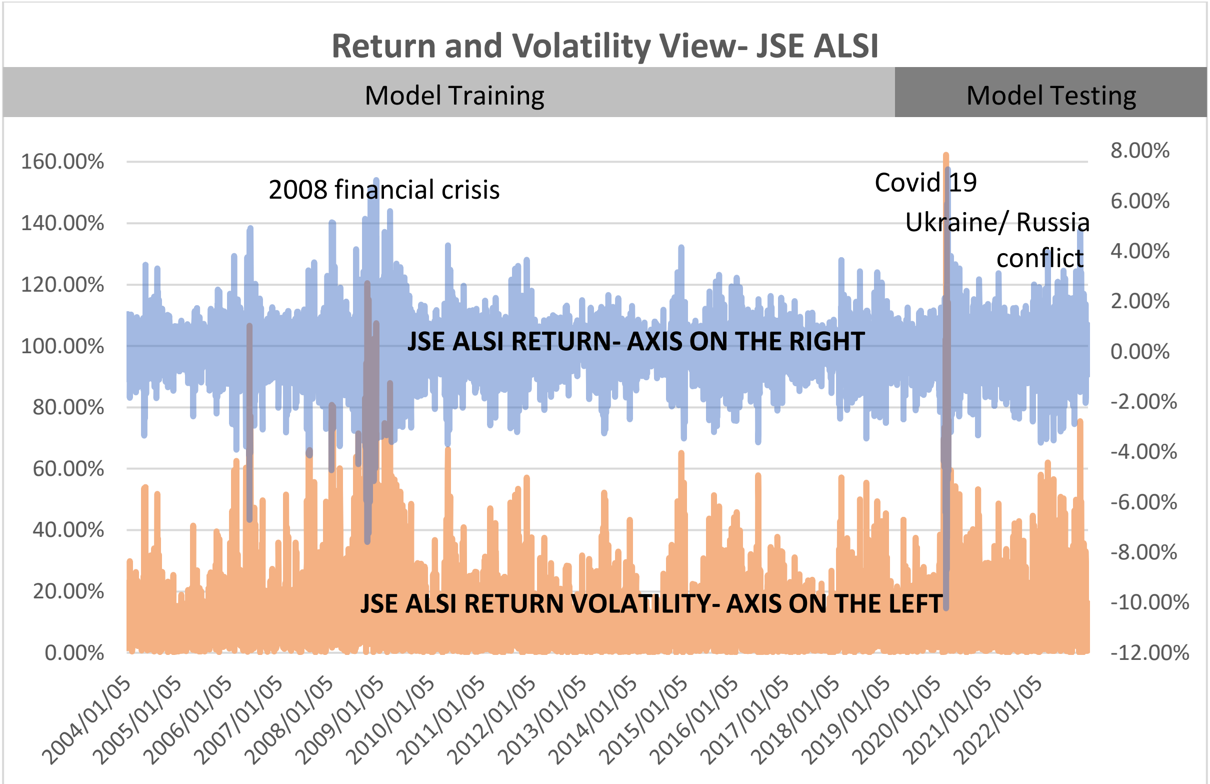


Figure 1: Volatility view of the JSE ALSI returns

Various stylised facts have been documented with regards to the volatility of equity returns on the JSE. Pagel et al. (2007), Samouilhan and Shannon (2008) and Alagidede and Panagiotidis (2009) have documented volatility clustering and leptokurtosis. Samouilhan and Shannon

(2008) and Makoko and Muzindutsi (2018) have documented volatility asymmetry while Samouilhan and Shannon (2008) also finds evidence of volatility persistence on the JSE ALSI. The Jarque-Bera test statistic of the JSE ALSI compared to the critical values at both the 95% and 99% confidence levels showed that the index returns are not normally distributed, and that it is leptokurtic (kurtosis=5.47) with a negative skew (skew = -0.35) (Table 4). The visualisation of the returns (Figure 1) hint at the increased volatility environment as well as volatility clustering. More distribution statistics of the returns on the JSE ALSI over the sample period are summarised in Table 3.

Table 3: Returns summary statistics of the JSE ALSI

Data Range	02/01/2004 – 30/12/2022
Data Frequency	Daily
Mean return	0.04%
Standard deviation	1.18%
Skewness	-0.35
Kurtosis	5.47
Minimum return	-10.23%
Maximum return	7.26%
Jacques Bera statistic	1377.16
Critical value- Chisq (95%)	5.99
Critical value- Chisq (99%)	9.21

Table created by author using the JSE ALSI data specified in the data range within this table.

2.8. Gap in the literature

The use of single machine learning models to more accurately predict financial market volatility seems to be outdated. Several more recent studies favour an ensemble or hybrid approach to volatility forecasting, where multiple models are combined to try and cover gaps left by the other models. Furthermore, the choice of ARCH-type models to be incorporated into the hybrid model output seems to be arbitrary. There is therefore a need to incorporate logic-based metrics whose objective is to capture the volatility stylised facts observed in a specific dataset. As datasets exhibit different levels of volatility stylised facts, use of only one model is akin to a one size fits all approach. In line with these observations, this study investigates relevant ARCH-type models for the JSE ALSI for incorporation into a hybrid model build in conjunction with an LSTM approach. Empirical evidence on the South African financial markets in this regard is very limited, and this study therefore adds to the existing knowledge on this topic, specifically guided by the points highlighted above.

3. Data

3.1. The JSE ALSI

The JSE ALSI daily end of day closing price data from 05/01/2004 to 30/12/2022 was attained from Wall Street Journal (WSJ) Market Data. At the time of writing, the JSE ALSI (or J203, or ALSH, or FTSE/JSE All Share Index) represented 99% of the full market cap of securities listed on the main board of the JSE. It can therefore be argued that this index represents the entire South African listed financial equities market. The index is constituted using market capitalisation weights.

3.2. Data frequency

With regards to volatility forecasting the key question is what the relevant horizon for risk management is (Christoffersen & Diebold, 2000; Poon & Granger, 2005). Intraday returns can be used to measure volatility. However, as stated in Section 2.2.3., this data frequency is not feasible for this study. Even longer time horizons may be used, e.g. a month. However, there is too much activity occurring in a month, and one value representing a volatility measure over such a long period is not sufficient to use in managing a portfolio as it misses the variability at shorter intervals. Christoffersen and Diebold (2000) find that a horizon greater than 20 days (regardless of the frequency of the data used) does not generate much useful information. This is also the view of Petropoulos et al. (2022), who indicate that both too short- and too long-time intervals are not advisable. Thus, Poon and Granger (2005) claim that a forecast that uses one to twenty days as the forecasted duration of volatility delivers 50% to 58% accuracy. The study of Samouilhan and Shannon (2008) compares one-week ahead forecasting with one-day ahead forecasting on the JSE and concludes in favour of the one-week ahead data frequency.

An important consideration identified by Brailsford and Faff (1996) is that the horizon used needs to be based on the needs of those using the model's information in practice, and that doing this must not be at the expense of the volatility arguments given above. However, volatility estimates beyond one day can be computed using one-day forecasts as indicated in Section 2.2.3. Therefore, using one-day ahead forecasts should cover the needs of longer-term predictions, with the added advantage that the variability within the longer time horizon would not be hidden, provided that the forecasts are accurate. Using less than a day's worth of data could likewise be useful, albeit not that easy with the data itself being not that easily accessible.

As such it would seem reasonable to use one-day ahead forecasts, which is the approach adopted in this study.

3.3. Data segmentation

In terms of the allocation of the data between training and in sample testing, an 80/20 split seems to be the norm, even though any split can be made. The only important consideration is that there should be sufficient history for the training stage of the model. As such, the time series dataset for this study is from January 2004 to December 2022, and includes 4894 data points. This time series dataset is subdivided into a learning frame and a testing frame with the cut-off being the 22nd of March 2019. As such, there are 3916 datapoints in the training data frame spanning over 15 years, and just under a 1000 data points in the testing frame spanning just over 3 years.

3.4. Unit Root test

Just as per Roh (2007) and Verma (2021) amongst others, the stationarity (and autocorrelation status) of the time series dataset has to be tested and/or confirmed. One of the most widely used tests to assess this is the Augmented Dickey-Fuller (ADF) test. Table 4 highlights the results of this test for the sample used in this study.

Table 4: The Augmented Dickey-Fuller test output

ADF Test

	Return Lag	Drift	Trend	Lag 5	Lag 4	Lag 3	Lag 2	Lag 1
coefficient	-1.133396	-0.000000	0.042639	0.068345	0.097094	0.123262	0.137798	0.000905
standard error	0.036032	0.000000	0.014301	0.020163	0.024768	0.028813	0.032533	0.000342

t-stat	-31.4552

A workable time series dataset for statistical evaluations and procedures should not have a unit root (i.e., it needs to be stationary). Unitary root suggests that variables like mean, and variance are not constant over time. To evaluate this an Augmented Dickey-Fuller test needs to be run on the time series dataset. The 1% critical value is -3.4401. The ADF statistic is found to be well below the critical value, and the null hypothesis of the presence of a unit root is therefore rejected. A conclusion of covariance stationarity can therefore be made and as such, no further adjustments on the data need to be done in this regard.

3.5. Heteroskedasticity effects

Table 5: The ARCH LM test output

Arch LM Test						
	Lag 5	Lag4	Lag 3	Lag 2	Lag 1	Constant
Coeff	0.16903	0.12816	0.16444	0.23573	0.02492	0.00008
SE	0.01410	0.01399	0.01420	0.01399	0.01410	0.00001
r^2	0.24827	0.00062	ySE			
F Stat	322.53	4 883.00	DOF			
RegSS	0.00063	0.00190	ResSS			

F Stat	322.53	**ChiSq Stat**	1 215	
p-value	2.65E-132	**p-value**	1.64E-260	

The Arch-LM test is used in this study to assess the presence of heteroskedastic effects in the time series dataset used, as suggested by Tsay (2005) and Roh (2007). Table 5 shows the output of this assessment of autocorrelation at five lags. AIC and BIC were used to indicate the appropriate number of lags. The p values using both the F-distribution and the Chi-Square distributions in the Arch-LM test are very close to zero, and as such the null hypothesis of no GARCH effects is rejected at the 1% significance level. There is strong evidence of autocorrelation and heteroskedastic features in the time series dataset. The data thus exhibits heteroskedastic features. The use of ARCH-type models on this data series is therefore valid.

3.6. Test for normality

The results of the Jacques Bera (JB) test for normality as applied to the sample are shown in Table 3. The conclusion made was that the JSE ALSI time series dataset is not normally distributed (the Jacques Bera statistic is significantly higher than the critical value of 9.21 that was attained at a 99% confidence interval). This gives credence to the use of non-linear models, as well as models where assumptions about normality do not affect the conclusions to be made about the results. The extremely high excess kurtosis means that the JSE ALSI distribution has heavy tails, while the small negative skew value indicates a negatively skewed distribution. This is in line with the general finding in the literature that equity market returns are not normally distributed but exhibit leptokurtosis and skewness (Corhay & Rad, 1994).

4. Methodology

4.1. Introduction

Annexure 5 illustrates the entire study process flow, which is explained in this section. The study uses JSE ALSI data starting from 2004. Various ARCH-type models and LSTM are combined (collectively called hybrid models or LSTM+) to make JSE ALSI returns one-day ahead volatility forecasts. The forecasted volatility is compared to out-of-sample realised volatility to measure accuracy within a specified model, as well as make a relative accuracy assessment across models.

As per Engle and Patton (2001), the index price at time t (P_t) has a continuously compounded return at time t (r_t) defined as ln ($\frac{P_t}{P_{t-1}}$). These are used as a base input for all the models in the calculation of realised volatility as per the definitions given in Sections 2.2.3 and 2.5.2. Within this study, twelve methodologies are used: four as standalone models, seven as a combination of LSTM with the standalone models, and one as a simple average of the standalone models.

The study first determines the three ARCH-type econometric models to be used, as well as the order of these models from a list of six (the ARCH(p) model was excluded given the limitations that were noted) popular models that were discussed in Section 2.5.2. of the literature review. The best two ARCH-type models are sought based on the AIC and BIC scores generated. A GARCH(p,q) model was included as the third econometric model, albeit the order of this model was also determined using AIC and BIC scores. A GARCH (p,q) model was included to maintain comparative relevance with other studies (e.g. Roh (2007)) that seem to have used it as a base model as well as to evaluate the assertion of the likes of Andersen and Bollerslev (1997) who hold a view that more complex models are not better than a GARCH model. Ideally, as many models as possible should be considered. However, the complexity of the experiments increases with more models (e.g. the use of two econometric models in the study would require a consideration of 5 model experiments, three would require twelve, four would require 20, and so on…). As such, three is considered as an optimal number for the purposes of this study. The search for these three models is done by computing the AIC and BIC values from the six ARCH-type models over 80% of the JSE ALSI daily returns dataset (the training sample). The objective is to discover the most suitable ARCH type models for the JSE ALSI dataset, as well as the order of efficiency of these models. From these three models, one-day ahead volatility forecasts for the entire dataset are made, but with parameters optimised for

80% of the dataset only. The full forecasted volatility series is fed to the LSTM model to generate what are termed LSTM+ models.

The study therefore considers seven LSTM+ models, namely G-LSTM, E-LSTM, T-LSTM, GE-LSTM, GT-LSTM, ET-LSTM and GET-LSTM. The initial iteration over the LSTM and LSTM+ models was done to discover the hyperparameter range (hyperparameters are specific elements of a LSTM model setup that define how the learning will occur. A discussion of the specific hyperparameters for a LSTM model is not the specific focus of this paper though. Table 9 gives a glimpse of the hyperparameters and hyperparameter range discovered in this study.) that has an optimal (lowest) Mean Squared Error (MSE). Once again this is done using 80% of the dataset (the actual 80% of the dataset, 80% of the forecasts from the three ARCH-type models that the study first searched for). The hyperparameter range is expected to be different for each of the seven models. After this, multiple iterations are done on the same initial hyperparameter per model set to fine tune the MSE value. The forecast output with the lowest MSE value is then used as the output of the model and further review is done using it. To the extent that the LSTM run generates negative volatility (a possibility noted by Petroziello et al.; 2022), an average volatility without these negative values is used as the default. This needs to be done as volatility cannot be negative. If the generated models have higher propensity tendency to forecast negative volatility, then the ability of the outputs to be in line with the actual volatility would be low. This is discussed further in section 4.4. The same steps are undertaken for a simple LSTM model.

4.2. Econometric models

The econometric methodologies used in the study (EGARCH(2,1), TGARCH(1,1) and GARCH(1,1)) were identified as per the specifications in Section 4.1. The general model was used as per the specifications given in Section 3. In Annexure 2, the basic model definitions for these models are shown. Essentially, 80% of the JSE ALSI data are fitted to the three chosen models to generate optimal model parameters. With these parameters, a forecast of volatility is undertaken on the entire dataset using the very same econometric models.

4.3. Base LSTM model

The LSTM model is discussed with the aid of Figure 2, which illustrates the sequence of the data and transformation flows. From this, Equations 8 to 13 can be visualised.

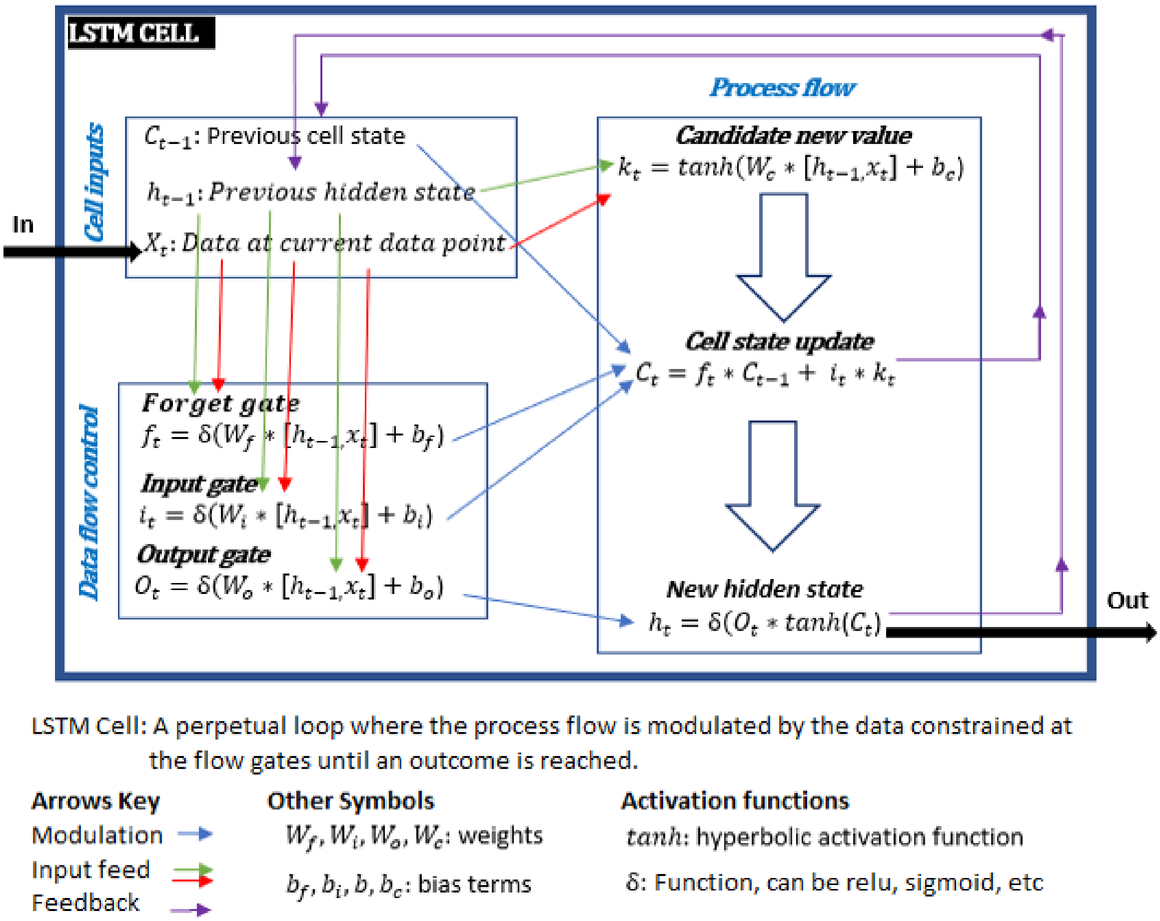


Figure 2: Illustration of a LSTM cell

Source: reconstruction based on Kim and Won (2018), Vasilev et al. (2019), Filipovic and Khalilzadeh (2021), Kakade et al. (2022)

$f_t = \delta(W_t * [h_{t-1,}x_t] + b_f)$ ---8

$i_t = \delta(W_i * [h_{t-1,}x_t] + b_i)$ ---9

$O_t = \delta(W_o * [h_{t-1,}x_t] + b_o)$ ---10

$k_t = tanh(W_c * [h_{t-1,}x_t] + b_c)$ ---11

$C_t = f_t * C_{t-1} + i_t * k_t$ --12

$h_t = \delta(O_t * tanh(C_t)$ ---13

Equations 8 to 13, together with Figure 2, show the repetitive iteration of data in defining the model weights until a volatility output is generated. The strength of this model is in its multiple iterations until optimal weights are found. The LSTM model considers multiple window size (in this study it was fixed at 22) iterations of the data. At each iteration the target output (which is a one-day ahead volatility) and 22 data points preceding this target output, are captured. Various other groupings are made based on the batch size and epoch size, but the window size grouping makes up the foundational cell construct as per Figure 2. The entire dataset is reviewed multiple times as defined by the epoch value (different values of these are searched per model to find an optimal value). Weights are recursively assigned through the multiple iterations of these basic setup for the entire training dataset across the defined epoch value. The detailed specifics are as per the setup in figure 2 through equation 8 to 13.

Specifically, data (volatility with the accompanying data at the date of that volatility) rolls in from the “IN” point in Figure 2, with a volatility output leaving at the “OUT” point. In between these two points, learning occurs as defined by Equations 8 to 13 to discover the optimal weights W_f, W_i and W_o. Bias terms b_f, b_i and b_o are also generated but the main objective is to generate the model weights. This is done multiple times with the weights being updated as the model rolls along. Essentially, the information to be remembered and by what factor, how much of the new information at each iteration needs to be influential to the rolling weights, what hidden connections are assigned in an attempt to fit the provided target volatility, are generated through this Figure 2 flow with Equations 8 to 13 governing how this is done. This is computed independently for each LSTM and LSTM hybrid model. As it can be appreciated, this is not a linear process and is not limited to any data distribution. Furthermore, depending on the specific assumptions used (e.g. window size, epoch value, number of units, number of layers, etc.), this can be time consuming. The specific objective of the model is to generate generalized, optimal weights that align the input data to the defined target data (one day ahead volatility forecasts). The LSTM architecture is setup such that MSE is minimized. These weights are then applied in the testing data set where the learnings are used.

The specific architecture of the LSTM is fixed across iterations. The LSTM is set up to have two layers and is unidirectional. A 22-day time window is assumed (which corresponds to a full month. i.e., assuming there are no holidays, a month has between 20 and 23 trading days. Balaban et al. (2006), Kim & Won (2018) & Kakade et al. (2021) amongst others, uses a days range corresponding to this), meaning the next day's volatility should be informed by information from the last 22 days), with *adam* optimiser, *relu* and *tanh* activation functions. An MSE loss is optimised during the iterations. The choice of which LSTM setup (LSTM architecture) to use is not the specific objective of this study. The following hyperparameter values (this is different from the LSTM architecture) are scoped in this study: number of units for the first layer, number of units for the second layer, number of epochs, batch size, and the learning rate. The window size is actually a hyperparameter, but it is fixed as 22 as explained earlier. The study uses ARCH-type models (x), a stand-alone LSTM model and hybrid LSTM models. In all instances, training is done on 80% of the dataset to discover model hyperparameters.

Training means the model optimises its parameters to generate the one-day ahead volatility using the information it has today. Once the model parameters are discovered, these are used to generate a forecast on the entire time set. The variability in the 80% of the time set is expected to be smaller as the models are not blind to this data. However, all models are blind to the 20% of the dataset though, and they make a forecast on the 20% using the training achieved in the 80% of the time set. ARCH-type models and the stand-alone LSTM are trained using the JSE ALSI data only. Hybrid models are trained using the JSE-ALSI data as well as the initial 80% of the forecasts from the ARCH-type models. All models have the one-day ahead realized volatility as their training target. As an example, hybrid x-LSTM was fed the entire JSE ALSI dataset, as well as the forecasts made by model-x on the entire time frame (remember that model-x was trained on 80% of the JSE ALSI dataset to generate its parameters, and with these parameters forecasts on the entire time frame were made). Hybrid x-LSTM is then trained on 80% of its input (each hybrid has a different ARCH-type forecast series input) to generate its own parameters. After this, Hybrid x-LSTM makes a forecast on the entire dataset. At all steps, the average model is a simple average of the three ARCH-type models. Performance measurement across all the models is computed based on the comparison between the out-of-sample forecasts and the out-of-sample realised volatilities.

4.4. Measuring the performance of the different models

Kumbure et al. (2022) describes four types of model measures, namely accuracy-based measures, error-based measures, return-based measures and statistical tests. The standard practice is to use loss functions (error-based measures) to evaluate the performance of volatility prediction models, and such practice is observed in the works of Poon and Granger (2003), Roh (2007), Samouilhan and Shannon (2008), Kumbure et al., 2022), etc. Some of these are the MSE, MAE, RMSE, RMAE, MAPE, HMSE and HMAE functions. These are classified as error-based evaluation methods as they are used to minimise errors or evaluate if errors were minimised in the forecasting models used. RMSE (Equation 14), MAE (Equation 15) and MAPE (Equation 16) are the three commonly used error-based measures (Kumbure et al., 2022). As such these were also used to evaluate the models in this study. These three loss functions are discussed in the next section.

$$RMSE = sqrt(\frac{1}{N}\sum_{i=1}^{n}(PV_t - RV_t)^2) \text{--} 14$$

$$MAE = \frac{1}{N}\sum_{i=1}^{n}|PV_t - RV_t| \text{--} 15$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{n}|\frac{PV_t - RV_t}{PV_t}| \text{--} 16$$

Where:

- PV_t represents the predicted value of the volatility at time t
- RV_t represents the realised value of the volatility at time t
- N is the number of observations

The MSE simply looks at the average squared error. The RMSE takes the square root of the MSE. The RMSE has a benefit in that it has units that can be easily interpreted, as opposed to the MSE. The MAE output is different from the MSE because the residual is not square; rather an absolute residual is found. In this way, more weight is given to larger errors in MSE because of squaring. Even though large errors are penalised, the MAE still takes a symmetric view of the forecast losses. As an alternative to R-square as a performance metric, MAPE, which computes the mean deviation error, is used to assess model fit. With regards to Equations 14 to 16, lower values are considered better. These loss measures give insights towards the performance of the different models. Furthermore, as per Kim and Won (2018) and Kakade et al. (2022) the statistical significance of the observed forecasts (and by implication, the observed differences in the loss measures across model pairs) should be determined for the error-based measures to be seen as having more merit, even though the meta-analysis of Kumbure et al. (2022) does not highlights statistical measures as the most widely used evaluations of models used. In the present study the Wilcoxon signed-rank test was used. The results of the study are presented and discussed in the next section.

5. Results and discussion

5.1. General model outcomes

5.1.1. ARCH-type models that fit the dataset

The ARCH-type model search was done on 80% of the dataset. AIC (the Akaike Information Criterion) and BIC (the Bayesian Information Criterion) were used as evaluation metrics where a smaller measure was preferred. These are statistical measures that give a model rating based on its loglikelihood (measure of data fit) and the number of parameters used in the computation of the loglikelihood (i.e., model complexity), where a higher ranking is given for a better model fit and a model is penalised for complexity (Geron 2019). If AIC was used, the rating order of models based on the lowest values found was EGARCH, APARCH, TGARCH, GJR-GARCH, GARCH-M, and GARCH. If BIC was used, APARCH and TGARCH switched places in the above sequence. A good question is which information criterion should be used when there are conflicting views. The guidance of Geron (2019) is that when AIC and BIC give conflicting output, the models suggested by BIC tend to be simple as BIC penalises complexity more heavily, and that those selected by AIC tend not to fit the data well. As such, the BIC test was preferred in making the selection in this study. Therefore, the EGARCH(2,1) and TGARCH(1,1) approaches were used. This could be an indication of the volatility persistence and the asymmetry prevalent in the JSE ALSI. If the AIC metric was used, the models chosen were the EGARCH(3,2) and APARCH(2,2) approaches. As per Geron (2019), the AIC based models are more complex. In line with the approach specified in Section 4, the GARCH(1,1) (or GARCH(2,1) if AIC was used) was also included, even though it lies much lower in ranking. Both selection criteria are in contrast with the findings of Samouilhan and Shannon (2008) that a GARCH order where p and q have the value of 2 provides the best in-sample fit for the JSE ALSI as well as the conclusion that the TGARCH(2,2) is the best ARCH-type model that can be used for forecasting domestic volatility. While Kakade et al. (2022) does not articulate the rationale for the specific ARCH-type models they use in relation to their specific dataset, their models are the same as the ones used in this study, even though the order of their models is not specified.

Table 6: EGARCH(2,1) model parameter significance tests

	coef	std err	t	P>|t|	95.0% Conf. Int.
mu	0.0325	1.381e-02	2.351	1.875e-02	[5.395e-03,5.954e-02]

Volatility Model

	coef	std err	t	P>|t|	95.0% Conf. Int.
omega	1.6439e-03	2.240e-03	0.734	0.463	[-2.745e-03,6.033e-03]
alpha[1]	-0.0234	4.092e-02	-0.573	0.567	[-0.104,5.676e-02]
alpha[2]	0.1422	4.068e-02	3.495	4.734e-04	[6.246e-02, 0.222]
gamma[1]	-0.2102	2.337e-02	-8.992	2.425e-19	[-0.256, -0.164]
gamma[2]	0.1169	2.362e-02	4.948	7.493e-07	[7.058e-02, 0.163]
beta[1]	0.9851	2.829e-03	348.218	0.000	[0.980, 0.991]

Source: output of a python implementation of an EGARCH(2,1) model as per Annexure 2

In Table 6, the absolute values of the t-stats are mixed for the EGARCH(2,1) model, even though there are more parameters with statistical significance. The omega and first alpha values do not appear to be significant (the interval horizon includes zero, the absolute value of the t-stats are lower than the critical value of 1.96, and the probabilities for these coefficients if the null hypothesis of zero parameters is accepted is high). Given that four parameters are statistically significant, it can therefore be concluded that there is some merit in using the model in this study.

Table 7: TGARCH(1,1) model parameter significance tests

	coef	std err	t	P>|t|	95.0% Conf. Int.
mu	0.0281	1.401e-02	2.004	4.511e-02	[6.116e-04,5.554e-02]

Volatility Model

	coef	std err	t	P>|t|	95.0% Conf. Int.
omega	0.0174	3.413e-03	5.102	3.356e-07	[1.072e-02,2.410e-02]
alpha[1]	7.8918e-03	7.467e-03	1.057	0.291	[-6.743e-03,2.253e-02]
gamma[1]	0.1084	1.165e-02	9.307	1.318e-20	[8.560e-02, 0.131]
beta[1]	0.9356	7.513e-03	124.527	0.000	[0.921, 0.950]

Source: output of a python implementation for a TGARCH(1,1) as per Annexure 2

Table 7 relates specifically to the TGARCH(1,1) model, which is modelled as a special case of the GARCH(p,q). With the exception of the alpha, the t-stats for the other variables seem statistically significant. All the absolute values of the t-stat are significant. The confidence intervals for most of the point estimates do not include zero, and as such there is significant evidence to reject the null hypothesis of zero parameter values. Something that stands out in this table is the high value of the t-stat for the beta value. This indicates that the previous values of conditional volatility have a huge influence on the calculated current period conditional volatility. This point can also be made for the EGARCH(2,1) model as shown in Table 7, where a t-stat that is greater than 300 (significantly greater than the 95% critical value of 1.96) was noted.

Table 8: GARCH(1,1) model parameter significance tests

```
==============================================================================
                 coef    std err          t      P>|t|      95.0% Conf. Int.
------------------------------------------------------------------------------
mu             0.0669  1.415e-02      4.724  2.307e-06 [3.913e-02,9.462e-02]
                              Volatility Model
==============================================================================
                 coef    std err          t      P>|t|      95.0% Conf. Int.
------------------------------------------------------------------------------
omega          0.0164  4.334e-03      3.785  1.536e-04 [7.910e-03,2.490e-02]
alpha[1]       0.0842  9.823e-03      8.573  1.013e-17   [6.496e-02,  0.103]
beta[1]        0.9033  1.101e-02     82.050      0.000     [  0.882,  0.925]
==============================================================================
```

Source: output of a python implementation for a GARCH(1,1) as per Annexure 2

The GARCH(1,1) model also shows significant parameters (Table 8). All the t-stats are greater than the 1.96 critical value and as such a significance conclusion is made at the 95% confidence interval. This also applies to the beta parameter.

5.1.2. Results from the LSTM and LSTM+ iterations

A large number of iterations across the different values of the hyperparameters (except for the window size) specified in Table 9 were considered. Table 9 contains the output per model where the RMSE was the lowest. The window size was held constant at 22 for this study, as it represents the number of past datapoints that are deemed as necessary to make a forecast one-day ahead, provided that the learnings from all of the available history is used with these datapoints. In this study, a full month was seen as enough data necessary to make a forecast for the next day.

Table 9: Hyperparameter set used for LSTM and hybrid models, outcome of step 3

	Units 1	Units 2	epochs	batch_size	learning_rate
LSTM-Daily	100	125	75	100	0.0010
G_LSTM-Daily	100	100	50	100	0.0100
E_LSTM-Daily	100	100	35	100	0.0010
GE_LSTM-Daily	100	125	100	100	0.0100
A_LSTM-Daily	100	125	35	100	0.0100
GA_LSTM-Daily	100	75	50	100	0.0100
EA_LSTM-Daily	100	100	35	75	0.0100
GEA_LSTM-Daily	75	100	35	75	0.0100

From the different iterations, it became evident that optimality was found at different points for each model. After discovering the values in Table 9, 50 iterations for each model were done and the resultant model outputs within each model were compared to find the most optimal RMSE., i.e. within each model, 50 different datasets corresponding to volatility forecasts for the training and testing horizon were computed and only one was selected for use as an output of this study. The most optimal training horizon dataset per model (the lowest RMSE in the training dataset) was used as the output of the research for that model. An optimal output is defined based on the training stage forecasts only. The testing forecasts are therefore consequential to the training dataset used. Once all the forecasts for each model were attained, comparisons across the different models were made based on how the forecasts in each model compared to the realised volatility dataset.

To assess statistical significance, a Wilcoxon signed-rank test was performed on the model forecasts that were chosen. i.e. all pairs of forecast outputs were tested to determine if their values were different. To assess statistical significance, a 95% confidence level was used. Most model pairs were found to be statistically significant. These significance values appear above the grey diagonal in Table 11.

5.2. Performance discussion

The forecast results can be visualized in Annexures 3 and 4. Performance however, is viewed from the perspective of the values of the loss functions. Lower values of the RMSE, MAE and MAPE in comparison to the other models indicate a superior model. However, different methods can be used to extract, use and make conclusions based on the output from the many iterations made across the twelve models. As an example, an average of each of the fifty iterations could be used for each model, or a forecasted data series that maintains metric stability between the training and testing dataset. Given the objectives of this study, the forecast dataset where RMSE was optimised in the training stage was selected as an output for each model. The discussion that follows is premised on this. The other metrics generated as a result of the optimised RMSE will also be discussed.

Table 10: Training and testing loss metrics

	Training performance			Out of sample performance		
	RMSE	MAE	MAPE	RMSE	MAE	MAPE
GARCH	12.19%	0.307593	246.48%	14.53%	0.334094	239.69%
TGARCH	11.95%	0.303230	244.96%	14.10%	0.329805	233.18%
EGARCH	11.92%	0.302675	245.49%	14.20%	0.330386	233.25%
LSTM	10.89%	0.282766	206.85%	14.26%	0.310507	197.45%
AVERAGES	11.96%	0.303913	245.60%	14.20%	0.330963	235.36%
G-LSTM	10.95%	0.286059	219.00%	13.62%	0.312964	205.25%
E-LSTM	10.66%	0.278534	204.76%	13.33%	0.308011	193.55%
T-LSTM	11.31%	0.288323	217.82%	13.20%	0.304533	198.14%
GT-LSTM	11.18%	0.285036	211.93%	13.10%	0.302958	194.64%
GE-LSTM	10.83%	0.283076	213.12%	13.77%	0.306679	197.56%
ET-LSTM	11.26%	0.284680	208.24%	13.34%	0.302668	188.10%
GET-LSTM	11.17%	0.286585	77.13%	13.26%	0.306072	72.03%

This table was calculated by the author consequent to the experiments made as indicated earlier. Equations 14, 15 and 16 and the associated discussions in section 4.4. specify how the values were calculated.

The most complex of the models generally give better outcomes. This is shown in Tables 10 and 11, as well as Annexures 6 and 7. GET-LSTM is shown to generally dominate the constituent models, and the same is true for ET-LSTM, GT-LSTM and GE-LSTM. In Table 11, the greatest improvement is noted for GET-LSTM as 8.77% in the testing dataset (8.39% in the training set). This is an improvement over GARCH(1,1). The highest improvement over GARCH in the testing date set is by the GT-LSTM model at 9.86%. All models generated (including the averages model) show an improvement over GARCH(1,1). The same also applies to the MAE and MAPE. Figures 4 and 5 illustrate this, as the GARCH(1,1) model is ranked the lowest in both the training and out-of-sample datasets. Kim and Won (2018) and

Verma (2021) have also found that that more complex ARCH-type economic models dominate GARCH. The study of Kim and Won (2008) focused on the KOSPI 200 stock index for the periods 2001 to 2011, while that of Verma (2021) focused on crude oil futures. This finding is therefore in contrast with the view of Andersen and Bollerslev (1997) that ARCH models should be sufficient to accurately model volatility, with the foundational condition being that the models need to be properly specified.

Table 11: Review of RMSE per model in the training and testing datasets

RMSE- Training	GARCH	TGARCH	EGARCH	LSTM	AVERAGES	G-LSTM	E-LSTM	T-LSTM	GT-LSTM	GE-LSTM	ET-LSTM	GET-LSTM
GARCH	0.00%	0.20%	2.08%	0.00%	0.51%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
TGARCH	1.95%		23.60%	0.00%	0.13%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
EGARCH	2.24%	0.30%		0.00%	27.90%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
LSTM	10.70%	8.93%	8.66%		0.00%	0.00%	22.17%	0.04%	0.07%	77.73%	0.00%	35.06%
AVERAGES	1.87%	-0.08%	-0.38%	-9.89%		0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
G-LSTM	10.17%	8.38%	8.11%	-0.60%	8.46%		0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
E-LSTM	12.58%	10.84%	10.58%	2.10%	10.92%	2.69%		0.00%	55.13%	1.04%	0.00%	0.02%
T-LSTM	7.20%	5.35%	5.07%	-3.93%	5.43%	-3.31%	-6.16%		0.00%	0.00%	0.00%	0.00%
GT-LSTM	8.27%	6.44%	6.16%	-2.73%	6.52%	-2.12%	-4.94%	1.15%		0.03%	0.00%	0.00%
GE-LSTM	11.17%	9.40%	9.13%	0.52%	9.47%	1.11%	-1.62%	4.28%	3.16%		0.00%	0.45%
ET-LSTM	7.65%	5.81%	5.53%	-3.43%	5.88%	-2.81%	-5.65%	0.48%	-0.68%	-3.97%		0.00%
GET-LSTM	8.39%	6.57%	6.29%	-2.59%	6.64%	-1.98%	-4.80%	1.28%	0.13%	-3.13%	0.80%	

RMSE- Testing	GARCH	TGARCH	EGARCH	LSTM	AVERAGES	G-LSTM	E-LSTM	T-LSTM	GT-LSTM	GE-LSTM	ET-LSTM	GET-LSTM
GARCH		0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
TGARCH	2.96%		0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
EGARCH	2.31%	-0.67%		0.00%	31.24%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
LSTM	1.86%	-1.14%	-0.46%		0.00%	0.00%	0.00%	0.00%	53.91%	0.09%	0.00%	0.00%
AVERAGES	2.28%	-0.70%	-0.03%	0.43%		0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
G-LSTM	6.29%	3.43%	4.08%	4.51%	4.10%		0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
E-LSTM	8.31%	5.51%	6.15%	6.57%	6.17%	2.16%		0.00%	0.00%	0.00%	0.00%	0.00%
T-LSTM	9.18%	6.40%	7.03%	7.46%	7.06%	3.08%	0.94%		0.00%	0.00%	0.00%	0.00%
GT-LSTM	9.86%	7.11%	7.73%	8.16%	7.76%	3.81%	1.69%	0.76%		0.00%	0.00%	0.00%
GE-LSTM	5.27%	2.38%	3.03%	3.48%	3.06%	-1.09%	-3.32%	-4.30%	-5.09%		0.00%	0.00%
ET-LSTM	8.21%	5.41%	6.04%	6.47%	6.07%	2.05%	-0.11%	-1.06%	-1.83%	3.11%		0.00%
GET-LSTM	8.77%	5.99%	6.62%	7.04%	6.64%	2.65%	0.50%	-0.44%	-1.21%	3.70%	0.61%	

This table shows values above and below the grey diagonal. The values below the diagonal are the percentage change between the two models overlapping on the horizontal and vertical axes. The values above the diagonal are the p-values generated from the Wilcoxon signed-rank test for between the two models overlapping on the horizontal and vertical aces. The values below the diagonal are read from the perspective of the left field. e.g. GET-LSTM dominates the GARCH model in the testing dataset by 8.77%. The Wilcoxon signed-rank measure in the testing table for this is shown above the diagonal. At a 95% confidence level, statistical significance between the forecasts of the GET-LSTM and the GARCH model can be made as the p-value for the Wilcoxon signed-rank test is zero. The two forecasts are therefore different, and the 8.77% differential is valid.

All stand-alone models and the average model are generally dominated by LSTM (an exception was found when considering LSTM against TGARCH and EGARCH in the testing dataset) and the hybrid models from an RMSE perspective. This is true on both the training and testing datasets. When considering MAPE and MAE, the ARCH-type models are also dominated. LSTM shows a level of dominance over hybrid models in the training dataset, but this performance is not replicated in the testing dataset. LSTM also generally dominates the ARCH-type models across all loss metrics considered. A conclusion can therefore be made that LSTM and LSTM+ models dominate the ARCH type models used in this study, both in-sample and out-of-sample. This dominance leads this study to conclude that there is merit in using hybrid models to forecast equity market return volatility where traditional econometrics models are combined with LSTM. Furthermore, it would seem that in hybrid models where T-GARCH and E-GARCH were included, there generally tends to be a superior performance as compared to where GARCH is added. This could be due to the higher ranking of TGARCH and EGARCH in terms of BIC and AIC for the JSE ALSI dataset. The fact that the TGARCH and EGARCH models dominate GARCH indicates that volatility asymmetry is important for the JSE ALSI dataset. For all the metrics evaluated, the averages model (created by averaging the performance of all ARCH-type models in this study) is dominated by the hybrid models. This means that hybrid models apply better weights to information contained in the ARCH-type models, and that applying a similar weight is not the best approach. Overall, models with more than one ARCH-type model generally dominated across all three metrics (The GE-LSTM is the exception). This conclusion is similar to that made in a study by Kakade et al. (2022) that focused on the commodities markets, in which double hybrid models show improvements over single and ARCH-type models. Kim and Won (2018) drew the same conclusion after considering hybrid models using the EWMA, EGARCH and a GARCH model, as well as LSTM, on the KOSPI 200 (Korea composite stock price index) dataset.

A conclusion that any one model has absolute dominance over all the other models cannot be made, but across all loss metrics as seen in Table 11, Figure 4 and 5 as well as Annexures 6 and 7, the GET-LSTM model showed more dominance over all models in the study. Put differently, there is no model that dominates the performance of GET-LSTM across all loss metrics, i.e., while there are some models such as the T-LSTM that shows dominance over GET-LSTM in terms of RMSE and MAE, the GET-LSTM model dominates the T-LSTM one in terms of MAPE. Therefore, GET-LSTM, the most complex model in the study, dominates all models reviewed in at least one metric. This makes a case for a combinatorial approach where simpler models are combined with LSTM for the JSE ALSI return dataset. This very same case can be made for GT-LSTM and ET-LSTM when out-of-sample data is considered. As mentioned earlier, a conclusion of absolute dominance cannot be made, but these three seem to be overall more dominant, especially when out-of-sample data is considered.

The MAPE of GET-LSTM deserves particular attention. Within both the training and testing dataset this measure is below 100%, which is much higher than the measure in all other models. GET-LSTM dominates all models in terms of this metric in both the training and testing datasets. Also, the highest performance improvements from simpler to more complex models are the greatest on the GET-LSTM improvements for MAPE. It can therefore be concluded that the relative fit (assumed to be captured by MAPE) from all models is generated by a model that combines all the models in the study.

5.3. Information transference

To unpack the noted performance, the review of information transference was investigated. The general logic behind the makeup and training of hybrid models where LSTM is embedded is that information is stored as model parameters, and that this stored information is used when forecasting is performed in the testing data frame. This information is visualised by relative performances such as shown in Table 11 and the associated annexures. Transference is depicted in Figure 3. Transference occurs if the performance based on the information learned in the training timeframe was kept when the models were used in (i.e. transferred to) the out-of-sample timeframe. As the curves in Figure 3 expand outwards, it would seem that there is a transference loss from the testing time series dataset to the out-of-sample time series dataset. This is the case across all the models, albeit more pronounced when reviewing RMSE than MAPE. The interpretation could be that not all information pertaining to the future is contained in the past, and that despite a superior learning algorithm there might always be features not captured. With regards to MAPE, the difference between the two curves seems marginal. However, given the other two metrics, a conclusion of general transference loss can be made.

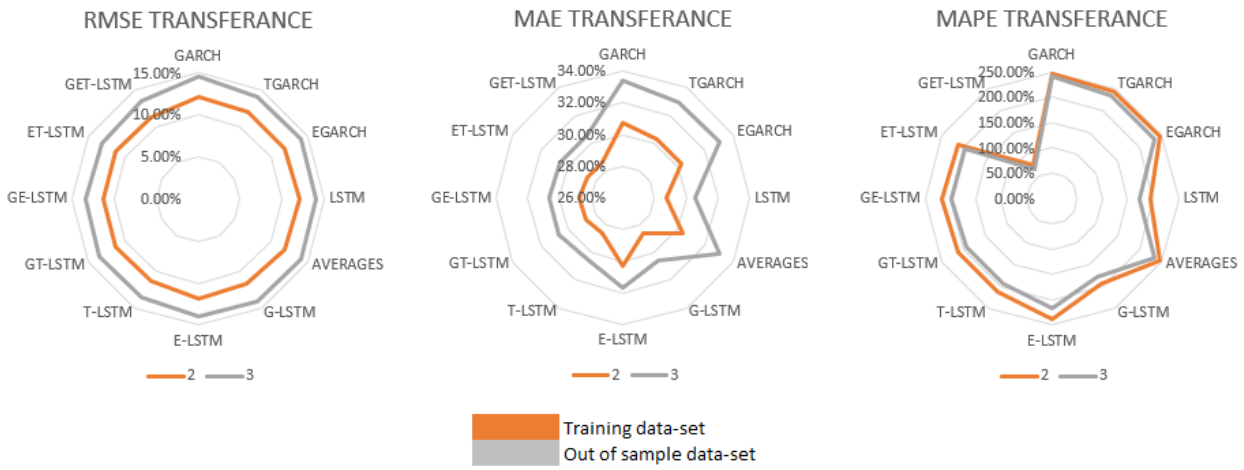
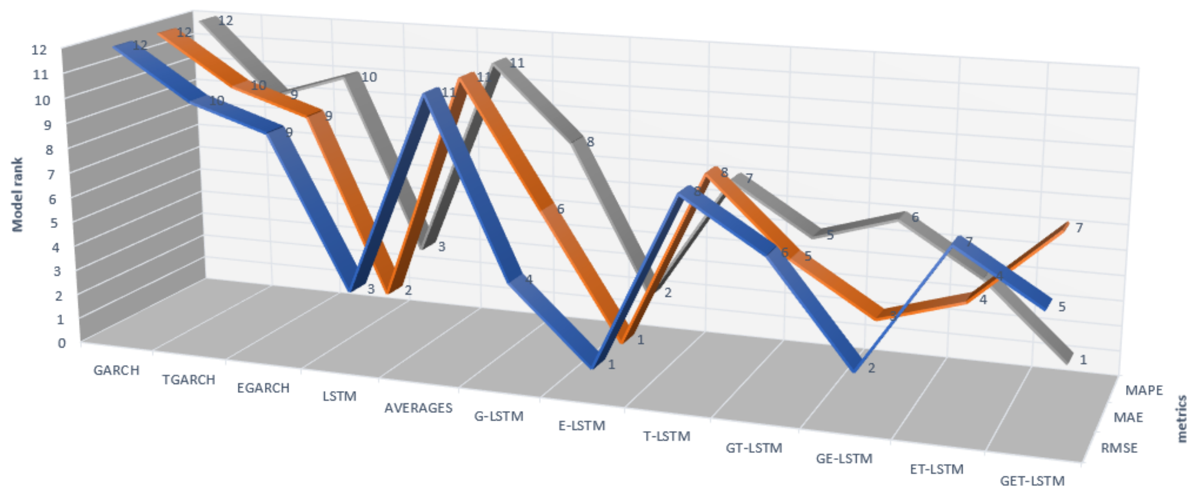


Figure 3: Information transference visualisation

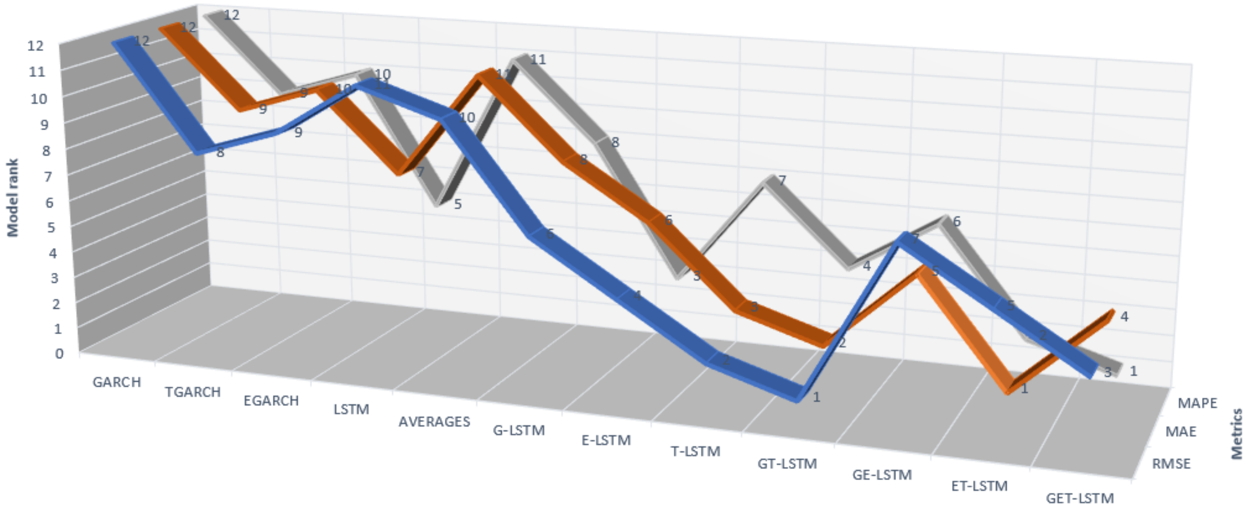
There is a wider information loss seen in MAE when compared to RMSE and MAPE. There seems to be no information loss for MAPE.



	GARCH	TGARCH	EGARCH	LSTM	AVERAGES	G-LSTM	E-LSTM	T-LSTM	GT-LSTM	GE-LSTM	ET-LSTM	GET-LSTM
RMSE	12	10	9	3	11	4	1	8	6	2	7	5
MAE	12	10	9	2	11	6	1	8	5	3	4	7
MAPE	12	9	10	3	11	8	2	7	5	6	4	1

Figure 4: A 3-D view of model ranking in the training dataset

The ranking is ordinal with 1 being the best and 12 being the worst. The ranking is out of 12. ARCH type models are ranked the lowest as compared to ML and hybrid models. The ranking is variable depending on the particular metric used.



	GARCH	TGARCH	EGARCH	LSTM	AVERAGES	G-LSTM	E-LSTM	T-LSTM	GT-LSTM	GE-LSTM	ET-LSTM	GET-LSTM
RMSE	12	8	9	11	10	6	4	2	1	7	5	3
MAE	12	9	10	7	11	8	6	3	2	5	1	4
MAPE	12	9	10	5	11	8	3	7	4	6	2	1

Figure5: A 3-D view of model ranking in the testing dataset

The ranking is ordinal with 1 being the best and 12 being the worst. The ranking is out of 12. ARCH type models are ranked the lowest as compared to ML and hybrid models. The ranking is variable depending on the particular metric used.

5.4. Model ranking

Figures 4 and 5 show the model ranking in the training and testing time series dataset. ARCH-type models, including the averages model, are ranked the lowest in terms of RMSE, MAE and MAPE, both in sample and out-of-sample. It is interesting to note that GARCH(1,1) is ranked the lowest. This is not different from what was shown in Section 5.1.1. where AIC and BIC were used to find the best fitting models for the JSE ALSI dataset, and where GARCH(1,1) was ranked lowest. As noted in Section 2.5.2., the GARCH(1,1) model is simple, and this comparative study shows that its use as a volatility forecasting tool might not be optimal. In this study the EGARCH(2,1) and TGARCH(1,1) models all showed an improvement over GARCH (all improvements are less than 3%) across all datasets.

While a model with an absolute dominance across all models was not found in this study, the highest ranked models are the GT-LSTM, ET-LSTM and GET-LSTM models. In all views, it is noted that training performance does not necessarily become testing performance. The

GARCH(1,1) model is ranked the lowest across both timeframes and across all metrics. LSTM and LSTM+ models are generally ranked higher than all ARCH-type models.

The next section discusses the conclusion, as well as limitations of the study, and also makes suggestions for further research.

6. Conclusion

In this study the use of machine learning, specifically LSTM and hybrid models (termed LSTM+) were investigated as an alternative to predicting return volatility for the JSE ALSI. Specifically, LSTM+ and hybrid models were tested as an alternative to ARCH-type models, that were used as a basis of comparison because of their ability to fit the JSE dataset the best. The academic literature reports many different econometric forecasting methodologies that have been shown to have differing levels of effectiveness in forecasting the volatility of out-of-sample financial market returns. In this study, three specific ARCH-type models that were evaluated to be a better fit to the JSE ALSI dataset were used. Specifically, these are the GARCH(1,1), EGARCH(2,1) and TGARCH(1,1) models.

Using these base models together with LSTM, eight other different derived models were subsequently created and used to forecast the volatility of the JSE ALSI returns dataset on an out-of-sample basis, making twelve models in total. These twelve models were the (i) GARCH(1,1) [G], (ii) TGARCH(1,1) [T], (iii) EGARCH(2,1) [E], (iv) an average of the aforementioned three models, (v) LSTM, (vi) G-LSTM, (vii) T-LSTM, (viii) E-LSTM, (ix) GE-LSTM, (x) GT-LSTM, (xi) ET-LSTM, and (xii) GET-LSTM models.

In this study, LSTM+ models were generally found to perform better than the traditional ARCH-type models in predicting return volatility for the JSE ALSI. Although an absolute conclusion that the most complex construct of these models (the GET-LSTM) dominates all the models tested cannot be made, the study has shown that GET-LSTM, GT-LSTM and ET-LSTM models dominate all the other models in at least one loss metric. All stand-alone models (GARCH(1,1), TGARCH(1,1) and EGARCH(2,1)) and the average model are dominated by hybrid models across all the three loss metrics considered (RMSE, MAE and MAPE). This is true on both the training and testing datasets. When considering LSTM, it showed a dominance over all econometric models considered except for TGARCH(1,1) and EGARCH(2,1) in the testing dataset for RMSE. As such, a general conclusion can be made that LSTM dominates the ARCH type models used in this study, both in-sample and out-of-sample. Generally, models that combine at least two ARCH-type models were seen to be performing better than the others. The use of hybrid models in forecasting JSE return volatility is therefore warranted, but more econometric models that have been rationalised to have a superior fit to the dataset in question should be studied. Furthermore, hybrid models that contain at least one ARCH-type model that was included in this study based on their superior AIC and BIC values dominate GARCH. The relatively poor performance of GARCH and G-LSTM (uses GARCH(1,1) only) as noted in the above figures, comparative tables as well as the annexures suggest that there is validity in pre-selecting ARCH-type models before they are included in hybrid models.

Specific implications of the study findings are as follows:

(1) The study outcomes are transferrable to other markets. The main finding is that ARCH-type models that are embedded into the hybrid makeup need to be assessed for model fit. A one size approach is not optimal.

(2) There is merit for the use of more complex methodologies in forecasting financial market return volatility, especially if the embedded ARCH type models have been rationalised to have a fit to the dataset in question.

(3) There are financial practitioners who have an interest in forecasting volatility beyond one day. However, these practitioners ought not to ignore the variability within each day. For example, it is possible to model a week using a series of one day ahead volatilities. However, if the one day ahead volatilities are premised on an input that contains similar or lower-level data, the variability within each day that makes up the week will not be missed.

(4) The information transference views show that it is likely not possible not to have the information relevant to the future to be fully captured in past dataset and transferrable through smart forecast measures. While every effort ought to be made in making accurate volatility forecasts, it should always be borne in mind that the future has some random elements. It could be said that the transference curves found support the Adaptive Market Hypothesis.

6.1. Study limitations and recommendations for further research

The following are some of the limitations identified for this study. Firstly, no exogenous factors were added. Studies like that of Verma (2021) suggest that exogenous factors (for equity volatility forecasting these could be inputs like the central bank rate, investor sentiment and the inflation figures amongst others) can be used as part of the feed into the used machine learning model. This is not out of sync with normal finance expectations, where macroeconomic factors affect the level of volatility seen in the financial markets. Perhaps the addition of this will increase the usability of the models. However, doing this adds to the complexity of the models used. This is perhaps something to be considered for future studies with a focus on the South African equity financial market.

Secondly, this study is based solely on the JSE ALSI time series dataset. The conclusions made might not be completely transferrable to other time series datasets, especially given that the ARCH-type models used specifically target the known characteristics of the JSE ALSI time series dataset.

A study could be made where a more direct observation of volatility stylised facts in a return series is done. In this study the association between stylised facts captured by a model as it relates to the characteristics of a dataset was implied through the use of AIC and BIC. A different approach could be to directly evaluate the stylised facts on the dataset as well as directly quantify the ability of particular ARCH type models to capture certain stylised facts such that these models are used in the hybrid construction.

Thirdly, the models from which the pair of three ARCH-type models were selected is limited. With more options perhaps a different set could be found. Also, the choice of the base models was based on an attempt to limit complexity. However, great software capacity would be required to handle this complexity.

Fourthly, while LSTM is one of the most widely used machine learning models used in financial modelling, other models could be explored. For example, this same study could be done using SVM, and even a combination of SVM and LSTM with other traditional econometric models.

Finally, the loss metrics used in this study for model evaluation purposes are symmetric in nature and do not consider under- and over-performance differently as suggested by Balaban et al. (2006). Perhaps a different conclusion can be reached if these were added.

References

Alagidede, P. & Panagiotidis, T. 2009. Modelling stock returns in Africa's emerging equity markets. *International Review of Financial Analysis.* 18:1-11.

Andersen, T.G. & Bollerslev, T. 1997. Answering the critics: Yes, ARCH models do provide good volatility forecasts. *NBER Working Paper No 6023.* Available: *https://papers.ssrn.com/sol3/papers.cfm?abstract_id=226433* , [2023, June 15].

Andersen, T.G., Bollerslev, T., Christoffersen, P.F. & Diebold, F.X. 2005. Volatility forecasting. NEBR Working Paper No 11188. Available: https://www.nber.org/system/files/working_papers/w11188/w11188.pdf, [2023,June 15].

Atsalakis, G.S. & Valavanis, K.P. 2009. Surveying stock market forecasting techniques- part II: soft computing methods. *Expert Systems with Applications.* 36:5932-5941.

Atsalakis, G.S. & Valavanis, K.P. 2013. Surveying stock market forecasting techniques- part I: Conventional methods. In *Computational optimization in economics and finance research compendium*. C. Zopounidis. New York: Nova Science Publishers. 49-104.

Balaban, E., Bayar, A. & Faff, R.W. 2006. Forecasting stock volatility: further international evidence. *The European Journal of Finance.* 12(2):171-188.

Balusik, A., De Magalhaes, J. & Mbuvha, R. 2021. Forecasting the JSE Top 40 using long short-term memory networks. Available: https://arxiv.org/abs/2104.09855, [2023,June 15].

Binge, L.H. & Boshoff, W.H. 2020. Economic uncertainty in South Africa. *Economic Modelling.* 88:113-131.

Bollerslev, T. 1987. A conditional heteroskedastic time series model for speculative prices and rates of return. *The Review of Economics and Statistics.* 69(3):542-547.

Bollerslev, T. Chou, R.Y. & Kroner, K.F. 1992. ARCH modelling in finance. A review of the theory and empirical evidence. 1992. *Journal of Econometrics.* 52:5-59.

Bollerslev, T. 2008. Glossary to ARCH(GARCH). *CREATES Research Paper 2008-49.* Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1263250, [2023,June 15].

Brailsford, T.J., Faff, R.W. 1996. An evaluation of volatility forecasting techniques. *Journal of Banking & Finance.* 20: 419-438.

Byrne, A. & Brooks, M. 2008. Behavioral finance: theories and evidence. *Research Foundation of CFA Institute Literature Reviews.* 3(1):1-26.

Christoffersen, P.F. & Diebold, X. 2000. How relevant is volatility forecasting for financial risk management? *The review of economics and statistics.* 82(1):12-22.

Corhay, A. & Rad, T.A. 1994. Statistical properties of daily returns: Evidence from European stock markets. 21(2):271-282.

Cont, R. 2007. Volatility clustering in financial markets: Empirical facts and agent-based models. In *Long Memory in Economics.* G. Teyssiere & A.P. Kirman, Ed. Berlin: Springer. 289-309.

Cotter, J. 2005. Uncovering long memory in high frequency UK futures. *The European Journal of Finance.* 11(4):325-337.

Daly, K. 2008. Financial volatility: Issues and measuring techniques. *Physica A.* 387:2377-2393.

Ding, Z., Granger, C.W.J. & Engle, R.F. 1983. A long memory property of stock market returns and a new model. *Journal of Empirical Finance.* 1:83-106.

Dreyer, A.A. & Hubrich, S. 2019. Tail risk mitigation with managed volatility strategies. *Journal of Investment Strategies.* 8(1):29-56.

Engle, R.F. 1982. Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation. *Econometrica.* 50(4):987-1007.

Engle, R.F. 1993. Statistical models for financial volatility. *Financial Analyst Journal.* 49(1):72-78.

Engle, R.F. & Ng, V.K. 1993. Measuring and testing the impact of news on volatility. *The Journal of Finance.* 48(5):1749-1778.

Engle, R.F. & Patton, A.J. 2001. What good is a volatility model? *Quantitative Finance.* 1(2):237-245.

Engle, R.F., Lilien, D.M. & Robins, R.P. 1987. Estimating time varying risk premia in the term structure: the ARCH-M model. *Econometrica.* 55(2)391-407.

Fama, E.F. 1965. The behavior of stock-market prices. *Journal of Business.* 38(1):34-105.

Fama, E.F. 1970. Efficient capital markets: A review of theory and empirical work. *The journal of Finance.* 25(2):383-417.

Filipovic, D. & Khalilzadeh, A. 2021. Machine learning for predicting stock return volatility. *Swiss Finance Institute.* 21-95. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3995529, [2023, June 15].

Francq, C. & Zakoian, J. 2010. *Garch* m*odels. Structure, statistical inference and financial applications.* Paris: Wiley.

Gerlach, S., Ramaswamy, S. & Scatigna, M. 2006. 150 years of financial market volatility. *BIS Quarterly Review.* September: 77-91.

Geron, A. 2019. *Hands-on machine learning with scikit-learn, keras & tensorflow. Concept, tools and techniques to build intelligent systems*. Sebastopol: O'Reilley.

Glosten, L.R., Jagannathan, R. & Runkle, D.E. 1993. On the relation between the expected value and the volatility of the nominal excess return on stocks. *The Journal of Finance.* 48(5):1779-1801.

Hyung, N., Poon, S. H., & Granger, C.W.J. (2008). A source of long memory in volatility. In *Forecasting in the Presence of Structural Breaks and Model Uncertainty*. D.E. Rapach & M.E. Wohar, Ed. Leeds: Emerald Group Publishing Limited. 3:329-380.

Kakade, K., Mishra, A.K. & Gupta, S. 2022. Forecasting commodity market returns volatility: A hybrid ensemble learning GARCH-LSTM based approach. *Intelligent Systems in Accounting, Finance and Management.* 29(2):103-117.

Kim, H.Y. & Won, C.H. 2018. Forecasting the volatility of stock price index: A hybrid model integrating LSTM with multiple GARCH-type models. *Expert Systems with Applications.* 2018. 103:25-37.

Kirchler, M. & Huber, J. 2007. Fat tails and volatility clustering in experimental asset markets. 2007. *Journal of Economic Dynamics & Control.* 31:1844-1874.

Kumbure, M.M., Lohmann, C., Luukka, P. & Porras, J. 2022. Machine learning techniques and data for stock market forecasting: A literature review. *Expert Systems with Applications.*

197(2). Available: https://www.sciencedirect.com/science/article/pii/S0957417422001452, [2023,June 15].

Lawrence, E.R., McCabe, G. & Prakash, A.J. 2007. Answering financial anomalies: sentiment-based stock pricing. *The journal of Behavioral Finance.* 8(3):161-171.

Liu, R., Jiang, Y. & Lin, J. 2022. Forecasting the volatility of specific risk for the stocks with LSTM. *Procedia Computer Science.* 202:111-114.

Lo, A.W. 2004. The adaptive market hypothesis: market efficiency from an evolutionary perspective. *Journal of Portfolio Management, Forthcoming.* 30(5):15-29.

Makoko, K. & Muzindutsi, P. 2018. Modelling return volatility in the main board and alternative exchange of the Johannesburg Stock Exchange: Application of GARCH models. *EuroEconomica.* 37(3):66-75.

Masset, P. 2011. Volatility stylized facts. *SSRN Electronic Journal.* Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1804070, [2023, June 15].

Monfared, S.A. & Enke, D. 2014. Volatility forecasting using a hybrid GJR-GARCH neural network model. *Procedia Computer Science.* 36: 246-253.

Moreira, A. & Muir, T. 2019. Should long-term investors time volatility? *Journal of Financial Economics.* 131:507-527.

Nelson, D.B. 1991. Conditional heteroskedasticity in asset returns: A new approach. *Econometrica.* 59(2):347-370.

Pagel, I.M., De Jongh, P.J. & Venter, J.H. 2007. An introduction to realised volatility. *Investment Analyst Journal.* 36:47-57.

Pederzoli, C. 2006. Stochastic volatility and GARCH: A comparison based on UK stock data. *European Journal of Finance.* 12(1):41-59.

Petropoulos, F., Apiletti, D., Assimakopoulos, V., Babai, M.Z., Barrow, D.K., Taieb, S.B., Bergmeir, C., Bessa, R.J., et al. 2022. Forecasting: Theory and practice. *International Journal of forecasting.* 38(3):705-871.

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

Poon, S. & Granger C.W.J. 2003. Forecasting volatility in financial markets: A review. *Journal of economic literature.* XLI(June):478-839

Poon, S. & Granger, C. 2005. Practical issues in forecasting volatility. *Financial Analysts Journal.* 61(1):45-56.

Redl, C. 2018. Macroeconomic uncertainty in South Africa. *South African Journal of Economics.* 86(3):326-380.

Roh, T.H. 2007. Forecasting the volatility of stock price index. *Expert systems with applications.* 33:916-922.

Samouilhan, N.L. & Shannon, G. 2008. Forecasting volatility on the JSE. *Investment Analyst Journal.* 37:19-28.

Shiller, R.J. 1981. Do stock prices move too much to be justified by subsequent changes in dividends? *The American Economic Review.* 71(3):421-436.

Shiller, R.J. 1988. The dividend-price ratio and expectations of future dividends and discount factors. *The Review of Financial Studies.* 1(3):195-228.

Shiller, R.J. 2003. From efficient markets theory to behavioural finance. *Journal of Economic Perspectives.* 17(1):83-104.

Tsay, R.S. 2005. Analysis of financial time series- 2nd edition. John Wiley & Sons. New Jersey: John Wiley & Sons.

Vasilev, I., Slater, D., Spacagna, G., Roelants, P. & Zocca, Valentino. 2019. *Python deep learning.* 2nd ed. Packt Publising: Mumbai.

Verma, S. 2021. Forecasting volatility of crude oil futures using a GARCH-RNN hybrid approach. *Intelligent Systems in Accounting, Finance and Management.* 28(2):130-142.

Wuite, R.2009. *The South African dictionary of finance*. Cape Town: Lebone Publishing Services.

Zakoian, J. 1994. Threshold heteroskedastic models. *Journal of Economic Dynamics and Control.* 18:931-955.

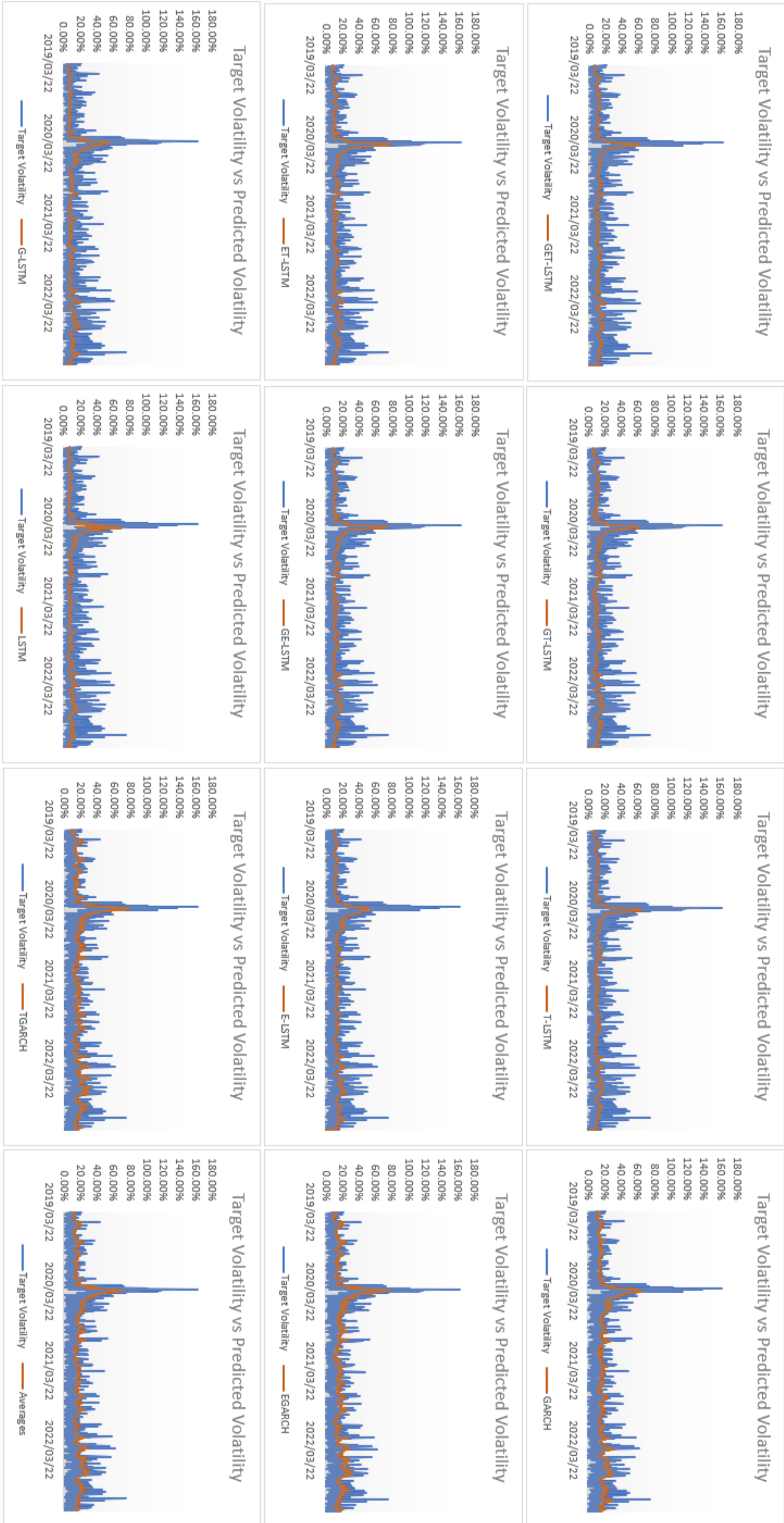
Annexure 1: Excerpt from the code setting the training ARCH-type model parameters onto the testing parameters before a testing stage forecast run

```
print(fit_model_train.params,fit_model_test.params)
fit_model_test.params=fit_model_train.params
print(fit_model_train.params,fit_model_test.params)
```

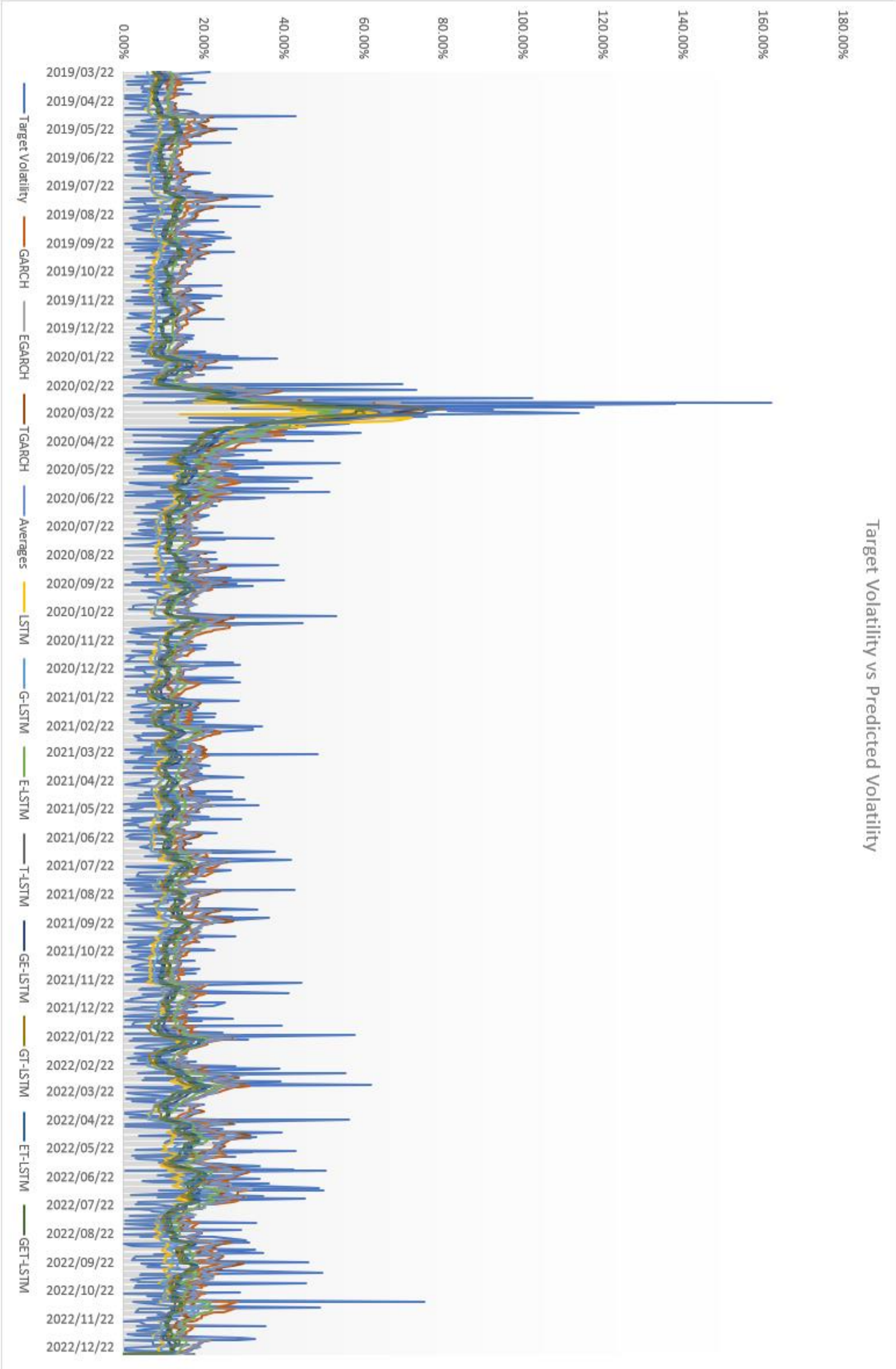
Annexure 2: Excerpt from the Model definition code, showing the definition of the three ARCH-type model used in this study

```
if model_type in ['Garch']:
    model = arch.arch_model(data['Log_Return'], vol='Garch', p=p, q=q)

elif model_type in ['EGarch']:
    model = arch.arch_model(data['Log_Return'], vol='Egarch', p=p, q=q,o=o)

elif model_type in ['TGarch']:
    model = arch.arch_model(data['Log_Return'], vol='Garch', p=p, q=q,o=o, power=1.0)
```

Annexure 3: View of volatility from all models used vs realised volatility



Annexure 4: Target volatility vs Realised volatility



Annexure 5: Research Technical process flow

JSE ALSI data

- Data access
- Data clean-up

Data diagnostics

- Unit root test
- Garch effects
- Normality test

ARCH-type model search that fit the JSE-ALSI dataset

- AIC and BIC evaluation on 6 ARCH-type models
- Best 2 models selected
- Best ordre of a GARCH (p,q) sought and added
- 3 ARCH-type models used

ARCH-type run

- Setup ARCH-type models based on 80% of the data
- Generate predictions on the entire timeframe
- Feed predictions into the next step
- Create the average model

LSTM and LSTM+ hyperparameter search

- Test run multiple hyperparameters to evaluate the RMSE generated
- Test based on 80% of the dataset with ARCH-type models embedded
- Select the hyperparameters per model coinciding with the lowest RMSE

LSTM and LSTM+ run

- Run 50 iterations on each model using the same hyperparameter set per model

Interprete the forecasts

- Discuss a selection criteria from the 50 forecast datasets per model
- Discuss the results of the output selected

Results and discussion

- RMSE, MAE, MAPE
- Model performance, improvements, transference, significance, model ranking

Annexure 6: Review of MAE in the training and testing stages

MAE- Training	GARCH	TGARCH	EGARCH	LSTM	AVERAGES	G-LSTM	E-LSTM	T-LSTM	GT-LSTM	GE-LSTM	ET-LSTM	GET-LSTM
GARCH		0.20%	2.08%	0.00%	0.51%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
TGARCH	1.42%		23.60%	0.00%	0.13%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
EGARCH	1.60%	0.18%		0.00%	27.90%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
LSTM	8.07%	6.75%	6.58%		0.00%	0.00%	22.17%	0.04%	0.07%	77.73%	0.00%	35.06%
AVERAGES	1.20%	-0.23%	-0.41%	-7.48%		0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
G-LSTM	7.00%	5.66%	5.49%	-1.16%	5.87%		0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
E-LSTM	9.45%	8.14%	7.98%	1.50%	8.35%	2.63%		0.00%	55.13%	1.04%	0.00%	0.02%
T-LSTM	6.26%	4.92%	4.74%	-1.97%	5.13%	-0.79%	-3.51%		0.00%	0.00%	0.00%	0.00%
GT-LSTM	7.33%	6.00%	5.83%	-0.80%	6.21%	0.36%	-2.33%	1.14%		0.03%	0.00%	0.00%
GE-LSTM	7.97%	6.65%	6.48%	-0.11%	6.86%	1.04%	-1.63%	1.82%	0.69%		0.00%	0.45%
ET-LSTM	7.45%	6.12%	5.95%	-0.68%	6.33%	0.48%	-2.21%	1.26%	0.12%	-0.57%		0.00%
GET-LSTM	6.83%	5.49%	5.32%	-1.35%	5.70%	-0.18%	-2.89%	0.60%	-0.54%	-1.24%	-0.67%	

MAE- Testing	GARCH	TGARCH	EGARCH	LSTM	AVERAGES	G-LSTM	E-LSTM	T-LSTM	GT-LSTM	GE-LSTM	ET-LSTM	GET-LSTM
GARCH		0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
TGARCH	0.43%		0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
EGARCH	0.37%	-0.06%		0.00%	31.24%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
LSTM	2.36%	1.93%	1.99%		0.00%	0.00%	0.00%	0.00%	53.91%	0.09%	0.00%	0.00%
AVERAGES	0.31%	-0.12%	-0.06%	-2.05%		0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
G-LSTM	2.11%	1.68%	1.74%	-0.25%	1.80%		0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
E-LSTM	2.61%	2.18%	2.24%	0.25%	2.30%	0.50%		0.00%	0.00%	0.00%	0.00%	0.00%
T-LSTM	2.96%	2.53%	2.59%	0.60%	2.64%	0.84%	0.35%		0.00%	0.00%	0.00%	0.00%
GT-LSTM	3.11%	2.68%	2.74%	0.75%	2.80%	1.00%	0.51%	0.16%		0.00%	0.00%	0.00%
GE-LSTM	2.74%	2.31%	2.37%	0.38%	2.43%	0.63%	0.13%	-0.21%	-0.37%		0.00%	0.00%
ET-LSTM	3.14%	2.71%	2.77%	0.78%	2.83%	1.03%	0.53%	0.19%	0.03%	0.40%		0.00%
GET-LSTM	2.80%	2.37%	2.43%	0.44%	2.49%	0.69%	0.19%	-0.15%	-0.31%	0.06%	-0.34%	

This table shows values above and below the grey diagonal. The values below the diagonal are read from the perspective of the left field. E.g. GET-LSTM dominates the GARCH model in the testing dataset by 2.80%. the Wilcoxon signed-rank measure for this is shown above the diagonal. At a 95% confidence level, statistical significance between the forecasts of the GET-LSTM and the GARCH model can be made. The two forecasts are therefore different, and the 2.80% differential is valid.

Annexure 7: Review of MAPE in the training and testing stages

MAPE- Training	GARCH	TGARCH	EGARCH	LSTM	AVERAGES	G-LSTM	E-LSTM	T-LSTM	GT-LSTM	GE-LSTM	ET-LSTM	GET-LSTM
GARCH		0.20%	2.08%	0.00%	0.51%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
TGARCH	0.62%		23.60%	0.00%	0.13%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
EGARCH	0.40%	-0.22%		0.00%	27.90%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
LSTM	16.08%	15.56%	15.74%		0.00%	0.00%	22.17%	0.04%	0.07%	77.73%	0.00%	35.06%
AVERAGES	0.36%	-0.26%	-0.04%	-18.74%		0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
G-LSTM	11.15%	10.60%	10.79%	-5.88%	10.83%		0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
E-LSTM	16.93%	16.41%	16.59%	1.01%	16.63%	6.50%		0.00%	55.13%	1.04%	0.00%	0.02%
T-LSTM	11.63%	11.08%	11.27%	-5.31%	11.31%	0.54%	-6.38%		0.00%	0.00%	0.00%	0.00%
GT-LSTM	14.02%	13.49%	13.67%	-2.46%	13.71%	3.23%	-3.50%	2.71%		0.03%	0.00%	0.00%
GE-LSTM	13.54%	13.00%	13.19%	-3.03%	13.23%	2.69%	-4.08%	2.16%	-0.56%		0.00%	0.45%
ET-LSTM	15.51%	14.99%	15.17%	-0.68%	15.21%	4.91%	-1.70%	4.40%	1.74%	2.29%		0.00%
GET-LSTM	68.71%	68.51%	68.58%	62.71%	68.59%	64.78%	62.33%	64.59%	63.60%	63.81%	62.96%	

MAPE- Testing	GARCH	TGARCH	EGARCH	LSTM	AVERAGES	G-LSTM	E-LSTM	T-LSTM	GT-LSTM	GE-LSTM	ET-LSTM	GET-LSTM
GARCH		0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
TGARCH	6.51%		0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
EGARCH	6.44%	-0.07%		0.00%	31.24%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
LSTM	42.23%	35.73%	35.80%		0.00%	0.00%	0.00%	0.00%	53.91%	0.09%	0.00%	0.00%
AVERAGES	4.33%	-2.18%	-2.11%	-37.91%		0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
G-LSTM	34.44%	27.93%	28.00%	-7.80%	30.11%		0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
E-LSTM	46.14%	39.63%	39.70%	3.90%	41.81%	11.70%		0.00%	0.00%	0.00%	0.00%	0.00%
T-LSTM	41.55%	35.04%	35.11%	-0.69%	37.22%	7.11%	-4.59%		0.00%	0.00%	0.00%	0.00%
GT-LSTM	45.05%	38.55%	38.62%	2.82%	40.73%	10.62%	-1.08%	3.51%		0.00%	0.00%	0.00%
GE-LSTM	42.13%	35.62%	35.69%	-0.11%	37.80%	7.69%	-4.01%	0.58%	-2.93%		0.00%	0.00%
ET-LSTM	51.59%	45.08%	45.15%	9.36%	47.26%	17.16%	5.45%	10.05%	6.54%	9.46%		0.00%
GET-LSTM	167.66%	161.15%	161.22%	125.43%	163.33%	133.23%	121.52%	126.11%	122.61%	125.53%	116.07%	

This table shows values above and below the grey diagonal. The values below the diagonal are read from the perspective of the left field. E.g. E-LSTM dominates the LSTM model in the testing dataset by 1.01%. the Wilcoxon signed-rank measure for this is shown above the diagonal. At a 95% confidence level, statistical significance between the forecasts of the E-LSTM and LSTM cannot be accepted. The two forecasts are therefore seen as similar.