IMPERIAL COLLEGE LONDON

DEPARTMENT OF PHYSICS

Adaptive Option Pricing: Enhancing the Black-Scholes Model with Stochastic Volatility and Machine Learning

Authors:

Mohammed Zain Mughal

David Blair

Laksh Gehani

Nuo Miao

Supervisor:

Prof. Dimitri Vvedensky

Submitted in partial fulfillment of the requirements for the MSci degree in Master of Science of Imperial College London

January 2025

Abstract

The Black-Scholes (BS) model, introduced in 1973, revolutionised the field of option pricing by providing a closed-form solution to price European options. However, one of its fundamental assumptions of constant volatility has long been a point of critique, as it fails to account for the dynamic and fluctuating nature of volatility observed in real financial markets. This project seeks to enhance the BS model by incorporating fluctuating volatility through the use of stochastic volatility models and machine learning (ML) techniques, which can predict future volatility and adapt to changing market conditions.

Traditional volatility models such as Lévy processes (Variance Gamma), GARCH, Heston and SABR are investigated and several ML models (ARIMA, KNN, CNN, RF, LSTM, GAN, GRU and XGBoost) are developed to replace the constant volatility assumption, feeding back into the pricing framework.

Through the combination of these advanced techniques, an enhanced ensemble option pricing model was developed, which leverages the strengths of both SABR (macro influence) and a GAN-LSTM hybrid (micro adjustments). The ensemble model dynamically adjusts to market volatility, offering superior accuracy and precision compared to the traditional BS model. This project also evaluates the effectiveness of the ensemble model using real market data, including a 25-year dataset of the S&P 500 Index, to forecast volatility and stress-test the models under extreme market conditions using Monte Carlo simulations. By comparing the performance of the traditional BS model with the new enhanced framework, the results show that the final ensemble model significantly increases responsiveness to changes in market conditions, particularly during periods of high volatility. The ensemble model produces volatility predictions with a narrower error distribution ($\pm 0.2\sigma$) compared to traditional BS ($\pm 0.4\sigma$) and with no significant outliers.

Contents

1 Introduction 4
1.1 Financial Markets 4
1.2 Derivatives 5

2 Risk & Volatility 6
2.1 Hedging & Leveraging 6
2.2 Arbitrage 7
2.3 Volatility 7

3 Black-Scholes Model 8

4 Review of ML Models 10
4.1 Auto-Regressive Integrated Moving Average (ARIMA) 10
4.2 K-Nearest Neighbours (KNN) 10
4.3 Random Forests (RF) 10
4.4 Extreme Gradient Boosting (XGBoost) 11
4.5 Convolutional Neural Networks (CNN) 11
4.6 Long Short-Term Memory (LSTM) 11
4.7 Gated Recurrent Units (GRU) 12
4.8 Generative Adversarial Networks (GAN) 12

5 Review of the Stochastic Alpha Beta Rho (SABR) Model 13

6 Method 14
6.1 Dataset & Setup 14
6.2 Model Evaluation 14

7 Results 16
7.1 Stochastic Alpha Beta Rho (SABR) Model 16
7.2 Random Forest 17
7.3 CNN 18
7.4 GAN-LSTM Hybrid 19
7.5 Ensemble Model 20
7.6 Monte Carlo Simulations 21
7.7 Option Price Comparison 22

8 Discussion on Computational Efficiencies 24

9 Conclusion 25

10 Appendices 26

1. Introduction

The Black-Scholes (BS) model, introduced in 1973 [1], revolutionised option pricing by providing closed-form solutions for determining the price of European options. Options serve as an important tool for investors to employ strategies to mitigate against risk or utilise risk to amplify potential returns. The BS model remains widely used today due to its simplicity and efficiency, allowing for quick and reliable calculations that underpin these strategies. However, to achieve this, the BS model makes a number of simplifying assumptions, including that volatility remains constant over time. The assumption of constant volatility leaves the BS model unable to adapt to the dynamic volatility empirically observed in markets, reducing the model's accuracy, particularly in highly volatile periods, such as market crash events.

Therefore, this project aims to enhance the traditional BS model by incorporating fluctuating volatility. This will be achieved by introducing adaptive volatility predictions through existing stochastic volatility models along with training and testing of machine learning (ML) models. These models will be compared against each other to determine the final enhanced *ensemble model* - which incorporates the best strengths of both stochastic and ML models. Monte Carlo (MC) simulations will serve as stress tests for the ensemble model, evaluating its effectiveness under extreme conditions, as well as its capacity to capture collective market behaviours. By capturing dynamic market conditions, this approach aims to improve option pricing accuracy and adaptability. The success of the ensemble model will be evaluated by directly comparing it against the existing BS model, focusing minimising pricing errors around high volatility scenarios.

1.1 Financial Markets

A financial market is a system in which financial instruments may be traded between investors, businesses or governments. These instruments range from equities (stocks), bonds, currencies and derivatives, which this paper will discuss in greater detail later. Financial markets form the backbone of an economy, playing a crucial role in the allocation of resources, providing liquidity and determining efficient and appropriate pricing for the aforementioned instruments. The complex nature of financial markets makes it very difficult to predict, as its behaviour is constantly influenced by hundreds of continuously evolving factors. There exists three main approaches to analysing market behaviours, which can be categorised as the following.

Fundamental analysis considers the most important factors in determining a stock or commodity price are intrinsic to the asset. For a stock price, this may be profit to earnings ratios (P/E), earnings reports or dividends for a company. A commodity's value, such as wheat, may be determined by weather conditions impacting the latest crop yield. From these factors, a fair market value can be established and used to determine whether a particular market/instrument is relatively undervalued or overvalued.

Technical analysis studies market behaviour primarily through the use of key indicators and market statistics. It is underpinned by the idea that any factor that can affect the price, will indeed affect the price of that market/instrument [2]. Therefore, analysis of these price movements is sufficient in isolation to understand market behaviours as past price patterns or trends serve as valuable indicators of future movement.

Quantitative analysis utilises mathematical models that can be constructed to replicate the main statistical properties of markets. Parameters can be adapted to mirror the factors influencing stock market behaviour in the real-world. These models are particularly useful when analysing complex financial instruments such as derivatives, which are crucial in modern markets.

1.2 Derivatives

A derivative is a financial instrument whose value is derived from some underlying asset. It is a contract between two or more parties, which can be traded on an exchange or over-the-counter (OTC), through a decentralised broker-dealer network.
Derivatives can be used to hedge, to speculate on the price movement of an underlying asset or to leverage a position. Types of derivatives include futures contracts (standardised contract buying/selling an asset on an exchange at a predetermined price on a specific future date), forwards (similar to futures, however they are not standardised and are traded OTC), swaps (OTC exchange of cash flow or liabilities from two different financial instruments) and options. An option gives the buyer the right, but not obligation, to sell (put option) or buy (call option) an underlying asset at a given quantity and strike price (also known as the exercise price) until the maturity date [3]. Options may be traded on an exchange or OTC, with exchange-traded options being more common due to their higher liquidity and standardised contracts making them more accessible to a larger range of investors. Accurately determining the price of an option, also known as the premium, is increasingly important for investors making informed decisions, especially as the EU equity options market has grown significantly in notional value, from EUR 5.7tn in Q4 2020 to EUR 9.1tn in Q4 2022 [4]. American options allow the buyer to exercise this right at any date up to the maturity date, whereas European options can only be exercised on the maturity date. This makes determining the option price of an American option particularly complex, as it requires consideration of the entire price path and has no fixed boundaries (as the option may be exercised at any time up to maturity). As a result, European options, which are simpler to model, are more commonly analysed and will be the focus of this report.

2. Risk & Volatility

Risk refers to the variation of potential outcomes around an expected (mean) outcome. The larger the range of outcomes, the higher the associated risk. High risk investments, such as equities, carry a significant chance of both large gains and losses. Lower risk investments, with a narrow range of outcomes, typically offer less potential for substantial gains. However, they mitigate against significant losses, providing greater stability for investors. An example of such investments are government bonds - debt securities issued by the government that generally offer low interest rates. We will later discuss the assumption of a portfolio earning a *risk-free* interest rate, which we will take to be the average return of a long term government bond.

2.1 Hedging & Leveraging

Investors employ a range of strategies to mitigate against risk, or use risk to potentially amplify returns. Options serve as a powerful tool for both purposes. Hedging with options is a strategy used to mitigate potential losses from unfavourable price movement in an underlying asset, whilst retaining the possibility of benefitting from more favourable movements. The option contract provides a form of insurance [3]. To illustrate this, consider an investor who owns 100 shares of a company at £50 per share. The investor is concerned that the share price may drop in the near future. To *hedge* their position, the investor purchases put options on the same stock, with a strike price of £48 and maturity time of 3 months. Taking the option premium as £1 per share, a total premium of £100 is paid. Consider the price falls to £40 per share in 3 months time. The investor could exercise the put option, selling their 100 shares at £48 each, for a net loss of £300. (£200 loss in total share value plus £100 premium paid). If they did not have a put option, the value of their shares would have decreased by £1000. Alternatively, the price may have increased to £60 per share, in which case the put option will not be exercised, and a profit of £1000 $-$ £100 $=$ £900 can be taken. The use of the options contract set a cap on the investor's potential loses, but did not limit upside except for the premium paid.

Leveraging with options allows investors to gain greater exposure to price movements of an underlying asset, amplifying profits but also increasing the risk of greater losses. This strategy provides greater market access with relatively smaller capital than would otherwise be required to achieve similar returns. Consider an investor who has £500, and anticipates a company's stock, currently priced at £50 per share, to increase in value in the near future. Instead of purchasing 10 shares, the investor can leverage their capital by purchasing a call option for the same stock, with a strike price of £50 and a premium of £2 per share. The investor now commands a much more significant 250 shares. In the favourable scenario, the share price has risen to £60 by the maturity date. The investor will exercise the call option, purchasing 250 shares at £50 per share, then proceed to sell at the current price of £60 per share. This results in a substantial £2000 net profit (250 shares $\times$ £10 profit per share $-$ £500 premium). Conversely, had the share price fallen below the strike price at the time of maturity, the call option would not be exercised and the entire £500 investment would be lost. This illustrates the potential for substantial returns as well as the mitigating the risk of losing the entire investment when using options to leverage capital.

It is clear that the pricing of the option premium is crucial in determining the effectiveness of the strategies, emphasising the importance of developing and enhancing option pricing models.

2.2 Arbitrage

Investors seeking to earn risk-free profits can achieve this through arbitrage. Arbitrage is the act of exploiting any price discrepancies between different markets for the same instrument [5]. Consider a stock which is traded on two different markets which operate with different currencies. An investor may purchase the stock in one market, then immediately (as far as the market allows) sell it for a higher price on the other market, which hasn't been able to adjust for the fluctuating exchange rate. The act of arbitrage trading itself, drives prices towards an equilibrium between markets and in practice, such opportunities are often short-lived.

2.3 Volatility

Variance is used to measure the dispersion of outcomes around an expected value and is determined as follows,

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{2.1}$$

where we have the sample variance σ^2, the number of outcomes in the sample n, the value of an individual outcome x_i and the mean (expected) value of all outcomes $\bar{x}$. The square root of the variance is the standard deviation, σ. In the context of financial markets, σ is a standard measure of the volatility (of a stock price) which quantifies the dispersion of outcomes and hence represents market risk. Many theoretical option pricing models, including BS, make the simplifying assumption that the volatility is constant. This allows for a closed form solution to be obtained and is based on the premise that stock prices follow a log-normal distribution, reflecting that stock prices cannot be negative.

However, in real-world markets, volatility greatly varies over time, influenced by effectively an infinite number of sources of risk. As a result, empirical evidence shows deviation from the modelled log-normal distribution with *fatter* tails. This is described by the *kurtosis* of a distribution, which is a measure of the *tailedness*. Fatter tails indicate a higher likelihood of extreme events occurring, far from the mean. A normal distribution has a kurtosis value of 3. Distributions with kurtosis greater than 3, have a greater distribution of outcomes at the extremes and are called *leptokurtic*. The BS model therefore fails to predict these extreme events which occur more frequently than any normal distribution would indicate.

In practice, implied volatility - the market's expectation of the future volatility of the underlying asset - is not constant across all strike prices as the BS model assumes. Instead when the implied volatility is plotted against the strike price, we observe a 'U-shaped' curve, called the *volatility smile* [3]. Implied volatility is greater for ITM (in-the-money) and OTM (out-of-the-money) scenarios, reflecting the market's expectation for greater volatility. An option is considered ITM if it would be profitable to exercise it at the current time. For a call option, this would be when the current price of the underlying asset is higher than the strike price and for a put option, when the current price is lower. Conversely, an option is considered OTM if it would not be profitable to exercise it. For a call option, the current price of the underlying asset would be lower than the strike price, and for a put option, the current price would be higher. These observations suggest that we must replace the assumption of constant volatility with a model that incorporates varying volatility, better capturing real-world markets.

3. Black-Scholes Model

In derivatives markets, risk-neutral valuation is crucial because it provides a consistent and arbitrage-free way to price financial instruments by assuming that investors are indifferent to risk. Under this approach, the value of the derivative is determined by discounting its future payoffs at a risk free rate. This approach ensures that all assets are priced in a way that reflects their true economic value under a hypothetical and idealistic risk-neutral world.

The Black-Scholes model, which was developed by Fischer Black and Myron Scholes in 1973 [1], explicitly incorporates risk-neutral valuation. The model revolutionised the field of options pricing. Black and Scholes' work focused on modelling the dynamics of asset prices and provided a mathematical framework for pricing European-style options. The Black-Scholes equation is based on the assumption of continuous price movements following a geometric Brownian motion with constant volatility. The closed-form solution allows for efficient real-time calculation and clear interpretation, making it a preferred model in finance despite assumptions like constant volatility.

Consider a portfolio containing Δ number of shares of a risky underlying asset (long position) and an option $V(S,t)$, which has been sold (short position). The value, Π, of this portfolio is

$$\Pi = \Delta S(t) - V(S,t) \tag{3.1}$$

where S is the share price. We want to consider how the value of this portfolio changes with time, so we differentiate to obtain an expression for $d\Pi$. From this, we obtain dS, the change in the share price, which is expressed as,

$$dS(t) = \mu S(t)dt + \sigma S(t)dW(t) \tag{3.2}$$

This is a stochastic differential equation describing the share price as a random process which is assumed to follow simple geometric Brownian motion. The first term describes the drift of the price, with μ being the percentage drift. The second term characterises the volatility, where σ is the percentage variance (the amount $S(t)$ spreads) and $W(t)$ represents a Wiener process (Brownian motion). Both μ and σ are assumed to be constant.

Since $V(S,t)$ is a function of a stochastic process S, the change in the option price dV, follows Itô's lemma [6]. We can now combine our expressions for dS and dV to obtain,

$$d\Pi(t) = \left(\Delta - \frac{\partial V}{\partial S}\right) dS - \left(\frac{\partial V}{\partial t} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 V}{\partial S^2}\right) dt \tag{3.3}$$

The coefficient of dS corresponds to the volatility and associated risk. This risk can be eliminated by taking $\Delta = \partial V/\partial S$ and the resulting equation (3.3) is now deterministic [1]. Black and Scholes determined that since the portfolio Π is riskless, it must earn the risk-free interest rate, r, which is constant : $d\Pi = r\Pi dt$ We assume no arbitrage exists in this market. By substituting (3.1) into our expression for a risk-free interest earning portfolio, we get a second expression for $d\Pi$ which we can equate to (3.3), obtaining the BS equation.

$$\frac{\partial V}{\partial t} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} + rS\frac{\partial V}{\partial S} - rV = 0 \tag{3.4}$$

We could have equally considered a portfolio containing an option (long position) and a number of shares sold against that option (short position), which would lead us to the same result (3.4).

We may obtain explicit solutions for determining the value of the options by transforming the BS equation (3.4) into the heat equation for which the standard solution is known. This was the original approach adopted by Black and Scholes [1].

$$C(S,T) = SN(d_1) - Ke^{-rT}N(d_2) \tag{3.5}$$

in which $C(S,T)$ is the price of the call option, $N(d_{1,2})$ is the cumulative normal distribution function and where,

$$d_1 = \frac{\ln\left(\frac{S}{K}\right) + \left(r + \frac{\sigma^2}{2}\right)T}{\sigma\sqrt{T}}, \qquad d_2 = d_1 - \sigma\sqrt{T} \tag{3.6}$$

This is the pricing formulae for a European call option. We have defined the time to maturity, $T = t^* - t$, where t^* is the maturity date, t is the current time, the current share (or other underlying asset) price S, the strike price K, the risk-free interest rate r and the volatility σ. Given the same underlying asset, strike price and maturity date; the price for a European put option P, is inherently related to the call option price C, through the put-call parity relationship,

$$C + Ke^{-rT} = P + S_0 \tag{3.7}$$

where S_0 is the current price of the asset [3]. This arises from the principle of no arbitrage. Since the options would have identical values at maturity, they must have identical values today, otherwise arbitrage opportunities would exist through exploiting the price difference between the two options. In the BS model, the volatility is assumed to be constant. This is, of course, not fully representative of real market conditions, therefore we aim to enhance the traditional BSM by incorporating fluctuating stochastic volatility.

4. Review of ML Models

Although the BS model is known for its simplicity and analytical elegance, it has significant limitations, particularly its assumption of constant volatility. This assumption often results in inaccuracies when applied to dynamic real-world financial markets. Therefore, incorporating advanced ML techniques into the BS model offers a promising avenue for enhancement. Through exploring and developing ML models: Auto-Regressive Integrated Moving Average (ARIMA), K-Nearest Neighbours (KNN), Random Forests (RF), Extreme Gradient Boosting (XGBoost), Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), Gated Recurrent Units (GRU) and Generative Adversarial Networks (GAN), the strengths of ML techniques may be leveraged to tackle the assumption of constant volatility by aiming to predict and therefore vary the volatility, feeding back into the BS framework.

4.1 Auto-Regressive Integrated Moving Average (ARIMA)

ARIMA was introduced in the 1970s by Box and Jenkins. It is primarily used for time series forecasting, as the model captures temporal (time) dependencies in data by combining its components together. ARIMA models data points as a function of their (lagged) previous values (AR) and a weighted combination of the associated past errors to account for residual noise [7][8], building the moving average (MA) part. Within the auto-regression (AR), where the term "regression" refers to a statistical technique that is used to understand and model the relationship between variables. Specifically, regression estimates how independent variables influence a dependent variable. The overall goal of regression is to create an equation that predicts the dependent variable, based on the values of the independent variables. The integration (I) comes from ensuring stationarity in the data by applying differencing techniques, removing obscure trends to reveal underlying patterns.

4.2 K-Nearest Neighbours (KNN)

The KNN model was developed in 1951 by Fix and Hodges and is dual-natured. This is because it can perform both classification and regression tasks [9]. It uses an associative mechanism for classification and an averaging mechanism for regression. KNN calculates the distance between data points using either Euclidean distance, Minkowski distance, Manhattan distance or cosine similarity, dependent on problem requirements. They then make predictions based on the values of the closest k neighbours. KNN treats data points independently and does not account for the sequential nature of time series data. Volatility modelling often requires recognising temporal dependencies, such as volatility clustering, which KNN cannot capture. KNN also exhibits limitations as it faces data and dimensional scaling, as higher dimensional distances hold less significance and lead to the *curse of dimensionality*, which can only be mitigated through impractical computationally expensive techniques.

4.3 Random Forests (RF)

Random Forests (RF) was developed in 1995 by Ho and is an ML model which aggregates predictions from multiple decision trees, making them powerful for classification and regression

tasks [13]. These trees are models that divide the data into branches based on feature conditions to make predictions. Multiple trees then combine into a forest, which improves accuracy because predictions from multiple decision trees are combined. Each tree in the ensemble is trained on a different (bootstrapped) sample of data, with splits assigned on random subsets of features, which introduces diversity among the trees and decorrelating their predictions.
RF are extremely powerful for feature importance ranking, however they inherently tend to overfit due to the trees growing to unnecessary depths. They also lack temporal dependencies, as they do not inherently account for the sequential nature of time series data by treating each data point individually, akin to KNN. Volatility modelling often requires capturing patterns over time, so RF is not necessarily most effective fit.

4.4 Extreme Gradient Boosting (XGBoost)

XGBoost is an ML model introduced in 2016 by Chen and is an optimised version of gradient boosting (a ML technique that combines multiple weaker models to create a single, more robust model), known for its accuracy and speed in classification and regression tasks [19]. XGBoost builds decision trees in sequence, optimising each based on the residuals of the previous trees (minimises loss function via gradient descent), which minimises error over iterations. However, XGBoost requires intensive parameter tuning, like grid search or random search; this is not unique to XGBoost but instead, a challenge that applies broadly across ML models. XGBoost also does not handle the temporal structure of time series data in the same way as KNN and RF, due to treating each data point independently, leading to the inability to model clustering or mean reversion.

4.5 Convolutional Neural Networks (CNN)

CNN was developed in 1989 by LeCun and his collaborators [10]. Neural networks are computational models inspired by the human brain, designed to recognise patterns and make predictions by processing data through interconnected layers of *neurons* [11]. In a neural network, each neuron receives inputs, applies weights (trains) and produces an output through an activation function, which is a mathematical function applied to the output of a neuron after the weighted sum of its inputs. The CNN method was initially used in image processing, as CNNs are powerful for feature extraction in structured data. CNNs use convolutional layers to detect spatial hierarchies, reducing dimensionality and highlighting important features in data [12].

4.6 Long Short-Term Memory (LSTM)

LSTM was introduced in 1997 by Hochreiter and Schmidhuber and is a type of recurrent neural network (RNN) [14]. RNN is a type of artificial neural network designed to handle sequential data, where the order of data points matters, such as in time series, natural language or audio processing. LSTM was designed to capture long-term dependencies in sequential data, making it effective for time series forecasting. LSTM networks use gates to control the flow of information, enabling them to pick and chose which information to retain over longer and shorter sequences. This means that LSTM has a memory, which helps in modelling complex, non-linear patterns over time. They are especially good for financial applications, effectively capturing complex patterns such as clustering and mean reversion. LSTMs require large amounts of high-quality data to train effectively, but financial data can sometimes be sparse, noisy or incomplete, which can hinder the models performance, whilst also being prone to overfitting, especially in financial markets where the data can be highly volatile and noisy. LSTMs are computationally intensive due to their sequential nature and large number of parameters, which can present challenges in terms of scalability and efficiency.

4.7 Gated Recurrent Units (GRU)

GRU was introduced by Cho and his collaborators in 2014, it is similar to LSTM but has fewer parameters [17]. It can be viewed as a simplified version of LSTM, as it builds on the same foundational concepts. It is efficient for sequential data, making it popular in time series forecasting. GRUs simplify LSTM's gating mechanism while maintaining effectiveness in capturing sequence dependencies [18]. GRUs have limited expressiveness compared to LSTM and are less suitable for modelling volatility in comparison. They also have fewer parameters and a simpler gate structure, making GRUs less computationally expensive and faster to train compared to LSTMs. The simpler architecture makes them less versatile for complex tasks requiring extensive memory retention, hence they are effective for sequential data but are less capable of capturing long-term dependencies compared to LSTMs. Overall, GRUs are suitable for scenarios requiring faster training and lower computational costs, where long-term dependencies are not critical.

4.8 Generative Adversarial Networks (GAN)

GAN was proposed in 2014 by Goodfellow and his collaborators, they generate synthetic data by training two networks, a generator and an advisor, which learn from each other [15]. The generator tries to create realistic data and the discriminator attempts to distinguish between real and generated data, leading to improved data generation over iterations. The generator minimises the discriminator's ability to distinguish between real and synthetic data, while the discriminator minimises its own loss function based on identifying real vs. synthetic data and this adversarial training forces both networks to iteratively improve. The loss function governs how the generator and discriminator compete and improve during training. The "cop and thief" example is a good analogy for GAN [15]. It explains GANs as operating in a zero sum game, where the thief (generator) tries to create fake money (synthetic data) that looks real, while the cop (discriminator) tries to distinguish fake money from genuine money (real data). Initially, the thief's fakes are poor but as the cop improves on counterfeit detection, the thief is forced to create more convincing fakes. Over time, the thief's counterfeits become so realistic that the cop can no longer distinguish them from real money. This back-and-forth captures the adversarial training dynamic of GANs. GANs are resource-intensive and prone to overfitting. They are better suited for scenarios requiring synthetic data generation or market simulation. Standard GAN architectures do not inherently model temporal dependencies, which limits their use in time-series tasks like volatility forecasting. Hybrid models using RNNs or LSTMs within the generator or discriminator can capture sequential patterns.

5. Review of the Stochastic Alpha Beta Rho (SABR) Model

The SABR model is a widely used stochastic volatility model that was introduced in 2002 by Patrick Hagan and collaborators [25]. It consists of two coupled stochastic differential equations (SDEs): one models the forward price of an asset, which is the price for a specific asset to be delivered at a future date, and the other describes volatility. The model has the parameters β, ρ and ν, where β determines the sensitivity of volatility to the forward price, ρ represents the correlation between forward price and volatility and ν governs the volatility of volatility, which shows how quickly volatility changes over time. These parameters shapes the implied volatility surface, which is critical for capturing the volatility smile - a pattern that is seen in options markets where implied volatility varies with the strike price. A key strength of the model is that it can derive an implied volatility surface with closed-form approximation, making it particularly useful for calibrating to market data. Its flexibility and computational efficiency also makes it well-suited for real-time options pricing. This makes it extensively used in pricing interest rate derivatives, exotic options and managing risk for complex portfolios. SABR is inherently different to ML models, as it incorporates risk-neutral valuation and recalibrates at every point to reflect current market conditions. However, the SABR model assumes a particular shape of the volatility smile. This particular shape is a smooth and continuous volatility smile that the SABR model produces, which may not match the real-world implied volatility profile during periods of extreme market stress.

Details about other stochastic models explored in this investigation, such as Lévy processes (Variance Gamma model), GARCH and Heston can be found in Appendix A.

6. Method

6.1 Dataset & Setup

A diverse dataset capturing a broad range of market conditions is essential for robust training of the ML models. To achieve this, sample data was taken from *SPX* (S&P 500 Index) shares, consisting of daily market open and close share prices and trading volumes. The dataset spans over a period of 14 years, from January 2010 to January 2024, with approximately 250 trading days per year, resulting in a total of 3491 data entries. This long timescale was chosen to capture a wide range of market conditions and different trends in volatility. Daily data was the highest frequency of freely available data accessible within the limitations of this project. *SPX* is an index that measures the performance of the 500 largest publicly traded companies in the US, as measured by market capitalisation. It was chosen specifically as it is one of the best representations of overall market dynamics, tracking a variety of sectors, with high trade volume and a broad range of market participants - from individuals to hedge funds. This diversity is useful for training a versatile model. Additionally, the very high liquidity and market activity of *SPX* stock ensures consistent data, less susceptible to anomalies caused by low trading volume or being over-influenced by isolated large trades.
From this dataset, simple moving and exponential moving averages, taken over a period of 20 days (SMA-20 and EMA-20) were extracted. These averages provide good indicators of recent price movements, with EMA-20 giving greater significance to more recent prices [26]. These were utilised in the development of the stochastic volatility models. For each ML model, 80% of the available data was used to train the model and the remaining 20% was used for validation. This split was chosen to minimise the risk of overfitting. This is when the model performs extremely well on the training dataset, however it fails to generalise effectively to unseen data, due to its parameters becoming too 'tailored' to the training data [27][28].

6.2 Model Evaluation

Each model was trained on the same dataset and then their performance in predicting volatility was evaluated across a range of metrics. Accuracy and precision were assessed through a 'confusion matrix', which included true positives (correctly predicted high volatility), false positives (low volatility predicted high), true negatives (correctly predicted low volatility) and false negatives (high volatility predicted low) [29]. Residuals (the difference between predicted and actual volatility) also provided valuable insights into model performance. Randomly distributed residuals indicated unbiased predictions and effective modelling of the underlying trend. Clustering of residuals suggested the model struggled in a particular scenario, such as a period of high volatility. Additionally, large residuals indicate instances in which the model failed to predict a significant shift in volatility. Monte Carlo price path simulations, of varying extremities, were utilised to stress test models. Since the dataset consisted of daily data, it did not incorporate the frequent fluctuations of volatility that occur on a 'tick' basis - the smallest possible price movement in real-world markets, often occurring multiple times per second. Stress testing was therefore implemented as an additional layer to assess the robustness of models. It is crucial to recognise that, while these metrics provided valuable insights during the development of the models, they alone do not provide a complete picture of model performance. The underlying theory of each model was carefully scrutinised to identify instances where a model appeared to perform well on the metrics but significantly deviated from theoretical expectations, indicating potential overfitting. Additionally, the computational efficiency of each model was considered as it is necessary to strike a balance between model accuracy with the time and computational power required to run it. This balance is particularly important for ensuring the model is capable of being utilised in real world markets. This will be further discussed in Section VIII.

7. Results

7.1 Stochastic Alpha Beta Rho (SABR) Model

The SABR model is widely used in industry for options pricing and was identified to be the most practical and accurate stochastic model to serve as the foundation for the ensemble model. Continued challenges with the implementation of Lévy, GARCH and Heston models, combined with extensive run times (see Table B1) made them impractical for further development and application, leading to the focus on SABR.

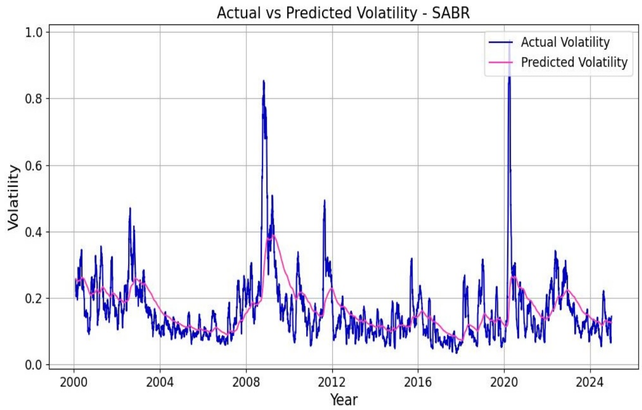

Figure 7.1: Actual volatility (dark blue) versus predicted volatility (magenta) from the SABR model between 2000 and mid-2024. SABR accurately follows the underlying trend, however fails to pickup on short term fluctuations. Sharp spikes in volatility can be observed at 2007-08 (financial crisis) and 2020 (Covid-19 Pandemic).

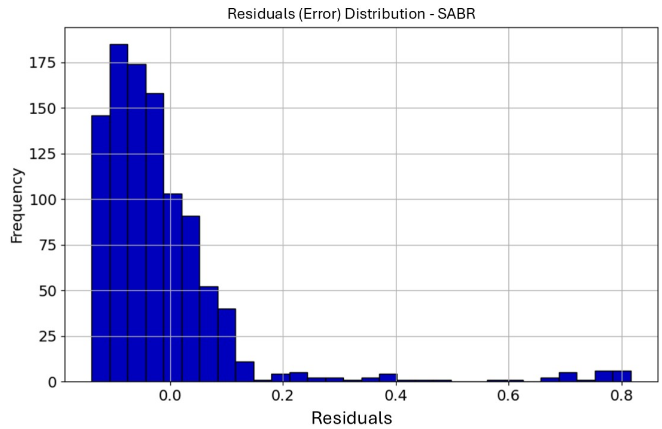

Figure 7.2: Error distribution plot shows the residuals between the SABR model's predicted volatility and the actual observed volatility. Residuals are concentrated at or below zero, indicating the model consistently over predicts volatility but captures the underlying trend. Outliers between 0.6-0.8 can be attributed to spikes in volatility around 2007-08, 2012 and 2020.

The main strength observed in the SABR model is it's ability to accurately capture long-term, underlying trends, as seen over the period of 14 years Fig.(7.1). SABR's mathematical foundation enables the model to not be overly sensitive to sharp changes in volatility. This is evident during sharp spikes in volatility around 2007-2008 (the global financial crisis) and 2020 (the onset of the Covid-19 Pandemic). While the predicted values respond to these events, they quickly revert to the long term trend, demonstrating the model's mean reverting nature. It is clear SABR does not respond to regular, short-term fluctuations in the volatility, however, it's overall stability provides a good foundation for a ML-based model to capture these more intricate features. The residuals (difference between predicted and actual volatility values) are concentrated at and just below 0, as shown in Fig.(7.2). This is consistent with the observation of SABR tending to over-predict volatility (seen in figure 7.1), but always following the long term trend, with slight influences. Outlying residuals are further evidence of SABR's inability to respond to sharp, short-term volatility spikes.

1. Random Forest (RF): Chosen for its simplicity, making it suitable for initial exploratory analysis.

2. Convolutional Neural Network (CNN): A more complex model, typically designed for image processing. This was chosen to explore how models not traditionally used in financial modelling, and inherently not capable of handling time series data, may produce misleading results.

3. GAN-LSTM Hybrid: The most complex and best-performing ML model, which was chosen to be incorporated in the final ensemble model.

7.2 Random Forest

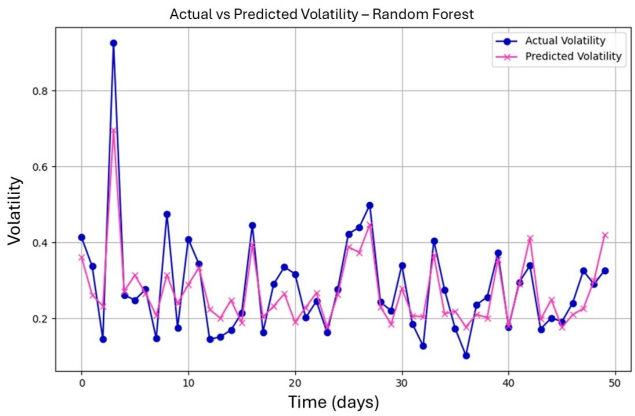

Figure 7.3: Actual volatility (dark blue) versus predicted volatility (magenta) from the RF model over a period of 50 days. The model demonstrates excessive sensitivity to spikes in volatility (day 4, day 17, day 24). The model fails to capture the underlying trend, predicting opposite shifts in volatility (day 6, day 13, day 26).

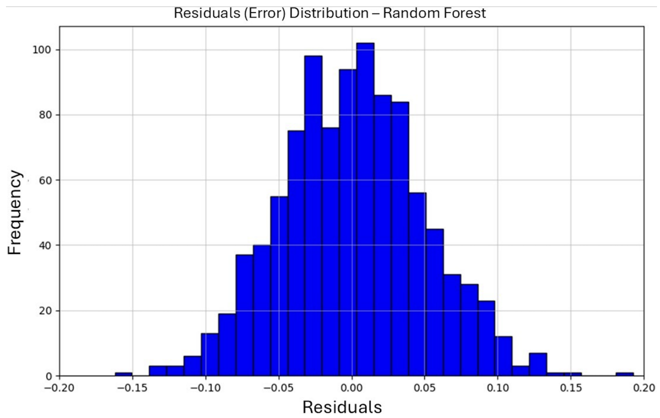

Figure 7.4: Error distribution plot shows the residuals between the RF model's predicted volatility and the actual observed volatility. Distribution is tightly packed and symmetric around 0, with very few outliers. Nearly all the points are within the range of -0.1 to 0.1. This indicates the model is capturing all features of the data, including noise.

The most basic ML model investigated in this project is RF. From an initial look at Fig.7.3, the predicted volatility seems to model the actual volatility well. However, a deeper look at the data suggests that the predictions unrealistically track the true values, suggesting overfitting. This occurs because the model learns from each spike, fluctuation and noise in the training data, leading to the model being overly tailored towards every tiny aspect in the training set and therefore unable to perform and accurately predict on any set of unseen test data. This overfitting can be seen in reigons such as days 12 to 15, in which the model over-predicts the volatility over a relatively stable period, and day 3, where it under-predicts volatility in a more volatile period.

This highlights RF's key downfall in this scenario, overfitting which stems from excessive tree depth, which allows the individual tree depths to exceed the optimal level and grow too complex, eventually capturing every detail of the training data. When this model is then applied to new and unseen data, it will attempt to reproduce the patterns from the original training data, causing the predictions to not fit the test data well; as seen in day 3, where the predicted volatility is much lower than the true volatility. To reduce overfitting, trees can be *pruned*, which effectively limits the max depth of trees, therefore limiting their complexities and over-

fitting potential. However, we can see that our pruning measures may have been insufficient, as there is still clear evidence of overfitting.
This behaviour is further reinforced in the error distribution (Fig.7.4), which resembles a vaguely normal distribution with residuals distributed around 0, with a spread of approximately 0.3. However, while this may suggest a impressive performance, the spread suggests that the model is overly tailored to the training data and therefore it's inability to generalise. The narrow spread can be explained by the RF capturing noise rather than underlying trends as intended. There are also a few outliers outside of the ± 0.15 regions, indicating sensitivity to extreme values. This together with the previous plot shows that RFs are inherently not the best fit for modelling volatility and capturing general trends, as they are prone to overfitting and subsequent subpar generalisation.

7.3 CNN

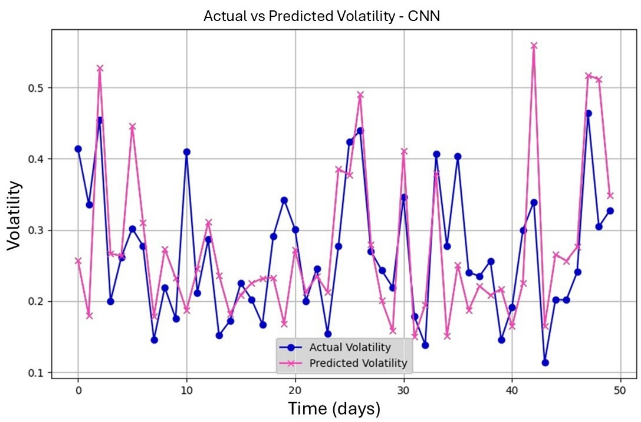

Figure 7.5: Actual volatility (dark blue) versus predicted volatility (magenta) from the CNN model over a period of 50 days. The model captures some local patterns (days 20-24 and 30-33) however, demonstrates extreme sensitivity to spikes in volatility (day 3, day 6, day 42, day 49). Deviations in both magnitude and timing of volatility changes can be seen (day 2, day 17, day 19, day 48).

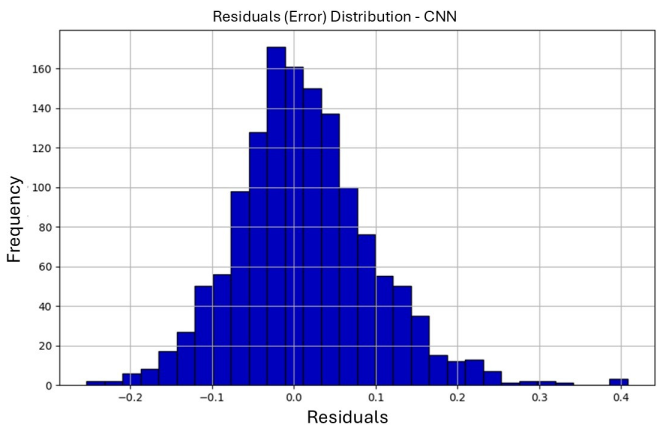

Figure 7.6: Error distribution plot shows the residuals between the CNN model's predicted volatility and the actual observed volatility. Residuals observed to be normally distribution around 0, which indicates an accurate prediction of some local volatility patterns. Outliers observed at 0.3 and 0.4 indicate the model's over-sensitivity to sharp changes in volatility.

As seen in Fig.7.5, CNN does not capture the overall trend and both over-predicts and under-predicts the volatility. This is because CNNs tend excel at extracting spatial features, but financial market volatility data lacks these patterns that the CNN is suited to recognise. During periods of high volatility, as seen in days 2, 26, 30, 42 and 47, the model over-predicts the volatility and in days 10, 19, 25, 33 and 35, it under-predicts. This shows that the model is unable to fully understand the data and instead relies on capturing general fluctuations rather than underlying trends in the data, resulting in predictions that deviate in both magnitude and timing, as seen in the plot. As the plot shows predictions with fluctuations of greater magnitudes in comparison to the plot for the RF model, the clear distinction between a model that does not fit well (CNN) and a model that fits excessively well (RF) can be compared and contrasted, with the magnitude of predicted volatility varying drastically. Although CNN has more

features than RF, it performs significantly worse. This is as the CNN model struggles to capture the broader dynamics of market volatility, leading to higher error rates relative to other models. These deviations shown illustrate CNN's failure to fully understand temporal data. The model creates an illusion of capturing volatility clustering, as the model learns to replicate local patterns in the training data without understanding their sequential or contextual significance, this is further exacerbated by the small dataset size. An example is day 19, in which the predicted volatility is approximately half of the actual volatility. Furthermore, the trend of the volatility is predicted to be in the complete opposite direction to the actual volatility. This again highlights the issue of systematic errors, demonstrating the model's inability to fully understand or adapt to the structure of volatility.
This is further supported by the distribution of residuals, as seen in Fig. 7.6. The spread of the residuals spans approximately 0.6, around double that of RF (overfitted), showing that the CNN model does not fit the data well at all with significant deviations in timings and magnitude. There are also more outliers in comparison to RF, indicating CNN is incapable of generalising in extreme conditions, as well as CNN's sensitivity to volatility surges, such as those observed on days 3, 6, and 42. Overall, the broad spread and the higher number of outliers in residuals showcase CNN's systematic inability to capture the temporal dependency component of volatility, resulting in persistent errors and poor adaptability to market dynamics.

7.4 GAN-LSTM Hybrid

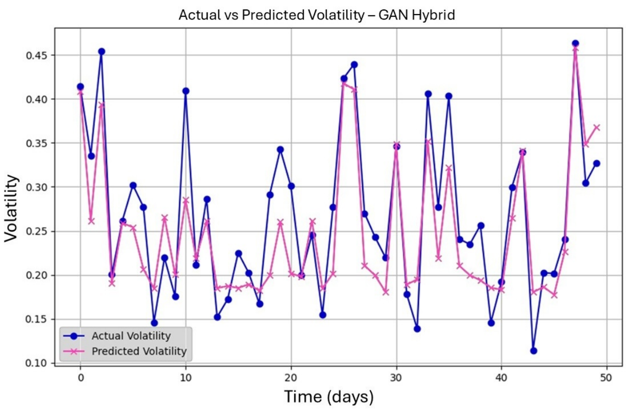

Figure 7.7: Actual volatility (dark blue) versus predicted volatility (magenta) from the GAN-LSTM Hybrid model over a period of 50 days. The model performs well at capturing the overall trend throughout. Notably, it correctly predicts the direction of changes at sharp spikes in volatility, without overreacting to the magnitude of these changes (day 10, day 19, day 35).

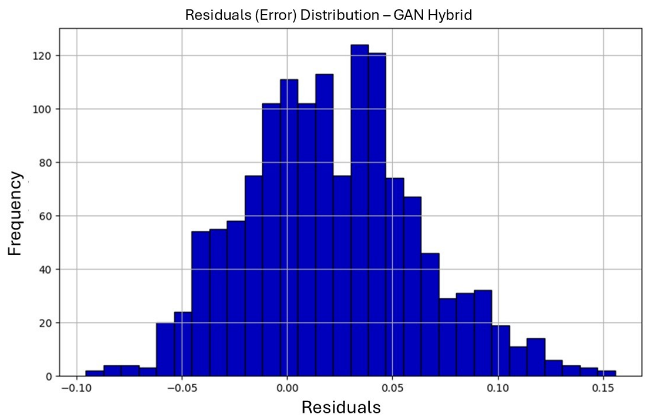

Figure 7.8: Error distribution plot shows the residuals between the GAN-LSTM Hybrid model's predicted volatility and the actual observed volatility. Residuals are contained between -0.10 and 0.15, with no major outliers, indicating consistent precision in predictions. Majority of residuals are contained within (and frequency remains stable) between -0.05 and 0.05, with slight increase around and just above 0.

The GAN-LSTM Hybrid is the most complex ML model explored in this project and performs significantly better than CNN and RF. The GAN component enables augmentation of the training data, whilst memory features of LSTM can be leveraged to improve the model's capability

to capture sequential dependencies. The predicted volatility closely follows the actual volatility throughout, as shown in Fig.7.7). The most notable improvement over CNN and RF is the hybrid model's response to sharp changes in volatility, as seen on days 10, 19 and 35. Due to LSTM's ability to weigh the importance of these changes against previous data points, the hybrid model does not overreact to the magnitude of these changes and subsequent predictions do not become erratic, as seen in Fig.7.5. During short periods of lower and more stable volatility, as seen between days 13 to 17, the model predicts near-constant volatility, a significant improvement over previous models. This stability highlights the model's robust ability to handle both stable and highly volatile periods, without greatly deviating from the overall trend (as observed in CNN Fig.7.5).
The majority of the residuals are contained within a very narrow range of -0.05 to 0.05, this shows a significant improvement upon RF and CNN models in predicting the fluctuations in volatility. A slight increase in the frequency of residuals around 0 indicates the model may be overfitting to the dataset in some cases, which aligns with the observation of predicted values exactly matching actual values in days 0, 3, 25, 30, 42 and 47. (Fig.7.7). An increase in the frequency of residuals just below 0.05 indicates the model often under-predicts the volatility by a small amount, this further demonstrates the advantage of incorporating the LSTM's ability to capture sequential dependencies and long term trends.

7.5 Ensemble Model

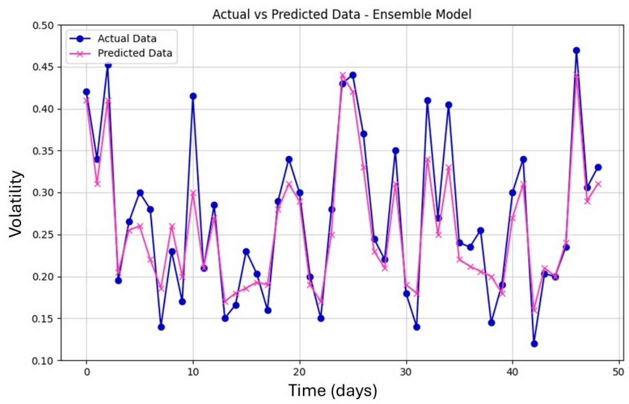

Figure 7.9: Actual volatility (dark blue) versus predicted volatility (magenta) from the ensemble model over a period of 50 days. The predicted values are observed to capture the overall trend, with only minor deviations between predicted and actual values (days 13-17, days 35-39).

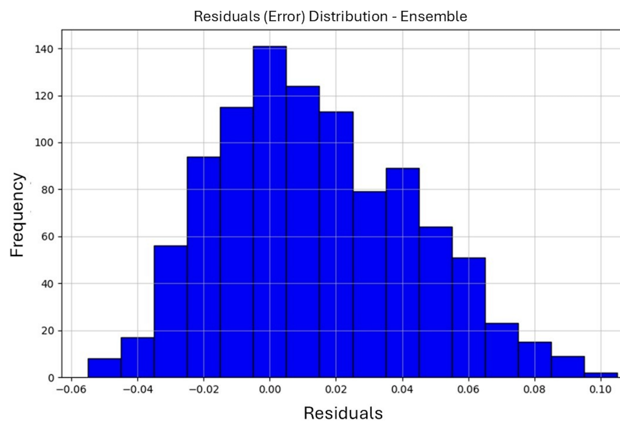

Figure 7.10: Error distribution plot shows the residuals between the ensemble model's predicted volatility and the actual observed volatility. Majority of the residuals are observed between a narrow range of -0.04 to 0.06. There is an even distribution of residuals within this range and only a small increase is observed around 0.

The ensemble model combines the strengths of both SABR and GAN-LSTM Hybrid into one comprehensive model. The precision and accuracy of this model are the highest in comparison to all other models (see Appendix B). SABR provides a solid foundation for capturing large-scale market behaviours and underlying volatility trends while the GAN-LSTM Hybrid can refine volatility predictions on a more granular level, working as a secondary layer on top of SABR. This secondary layer was designed to enhance predictive accuracy by identifying and reacting

to these finer, short-term dynamics, whilst leveraging the temporal memory component of the LSTM.
The actual and predicted values are in agreement throughout. The benefits of the GAN-LSTM component can clearly be seen as the model captures the spikes in volatility, notably around days 24, 32, 34 and 46 (Fig.7.9). The residuals for the Ensemble model predictions are contained entirely within the range of -0.05 to 0.10, with the majority existing between -0.03 and 0.05 (Fig.7.10). This represents the narrowest spread (and lowest values) of residuals out of all the previous models discussed. There are no observed outliers, indicating the ensemble model consistently performs well at capturing volatility fluctuations. The distribution of residuals is observed to very slightly skew towards the negative values, indicating a slight bias towards under-predicting values. This is due to the intrinsic way the GAN component regulates and scales data predictions. In training, predictions exceeding the actual value are *penalised* more harshly compared to predictions that fall below it. This is favourable to investors, as overestimating volatility can lead to excessive hedging costs. Over longer timescales, this introduces a *smoothing* effect on the upper bounds of predictions, as observed.
This combined ensemble model performs very well when observed over long periods of time, such as in figure 2, where we see the model predictions over 6 years. Visually, we are able to see that, in all scenarios of stability and uncertainty, the model is able to successfully adapt and predict close to the true volatility, with slight over-predictions due to reflected features from training scenarios. Therefore, showcasing this optimal combination of the fundamental SABR feeding into the precise GAN-LSTM architecture.

7.6 Monte Carlo Simulations

Monte Carlo stress testing was employed to synthetically simulate scenarios of extreme and stable conditions, in which we tested the reactions of the final ensemble model vs. the traditional Black-Scholes, evaluating their respective performances and errors. In Fig.7.11A (top left), we can see that the true volatility exhibits large, rapid, almost discontinuous spikes/jumps for the extreme market conditions. The Ensemble model does very well in tracking these true movements closely, successfully predicting and reacting to short-term surges and subsequent retractions with minimal lag, with some aspects of minute oscillations about a stable equilibrium, however this slight inaccuracy is expected due to the intrinsic way the model is trained and evaluated, keeping to the idea that over multiple time steps, the moving average of the predicted volatility converges well to the true values. We also notice a slight lag and therefore a slight under-prediction introduced by the ensemble predictions, this is due to the model learning of the sudden jump and working as quickly as possible to adapt to it successfully, as it does. By contrast, we can see the BS model remains at it's constant volatility throughout, evidently failing to capture any market movements, and consistent under-prediction in extreme scenarios. In the stable scenario Fig.7.11B (top right), overall volatility remains low and constant, the ensemble model exhibits a slight under-prediction coupled with variation within a very narrow band, which can be seen as constant, maintaining a very close match to the true stable volatility; this further proves the adaptability of the ensemble model in all scenarios due to it's optimal dynamic weighting, leading to larger influence from the SABR model in these more macro dominated regions. Meanwhile BS performs well by staying constant, however this is often short lived.
The error distribution in Fig.7.11C (bottom) further reinforces the ensemble model out-performing the standard BS model. We can that the BS exhibits broader ($\pm 0.4\sigma$), heavier-tailed error dis-

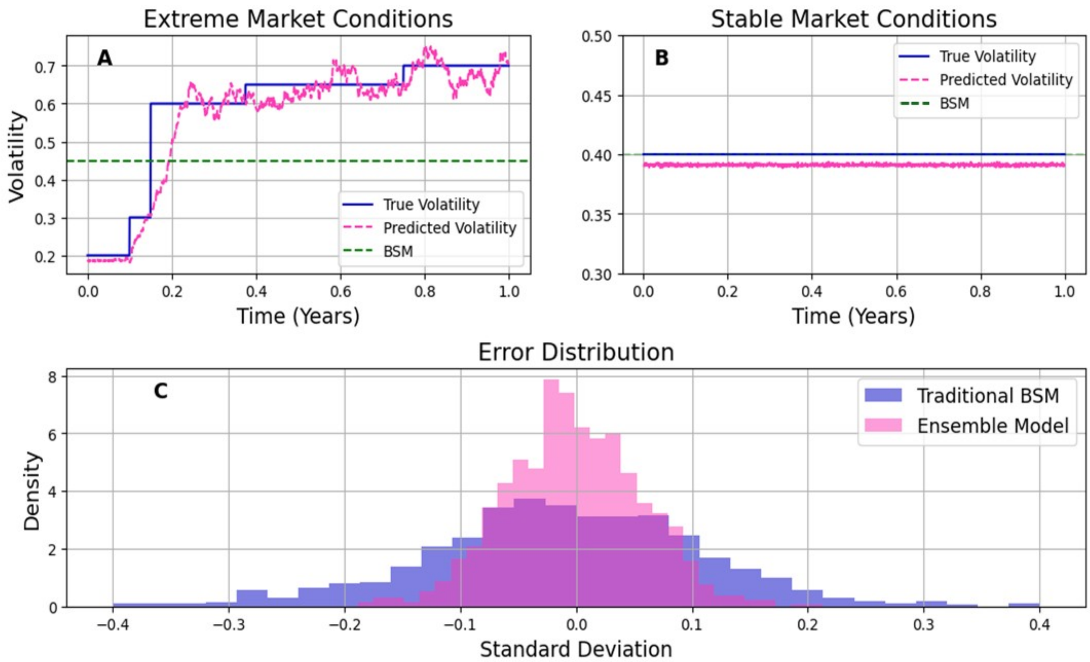

Figure 7.11: (a) Extreme market conditions: true volatility (blue) displays large, rapid spikes, the ensemble model predictions (dashed magenta) align with these abrupt changes closely, whilst the BS constant volatility (dashed green) remains at constant assumed equilibrium, always underestimating the true volatility. (b) Stable market conditions: true volatility constant (blue), the ensemble's prediction (dashed magenta) align closely with the flat trend with a slight shift and narrow variations, whilst the BS model's single volatility value (dashed green) matches the constant true value exactly. (c) (Standard deviation) Error distributions: The BS model's broader ($\pm 0.4\sigma$), heavier-tailed distribution (blue) contrasts with the ensemble model's narrower ($\pm 0.2\sigma$), more centred distribution (pink), illustrating the ensemble's reduced prediction error across both stable and extreme scenarios .

tributions, which are especially pronounced under extreme market conditions. This reflects a large under or over estimation of volatility when faced with shocks or even moderate trends not aligned with its single fixed value. The ensemble model shows, narrower ($\pm 0.2\sigma$), more centred error distributions in both scenarios. Whilst within extreme periods, outliers remain relatively contained, indicating that the hybrid approach combined with regularisation and scaling techniques effectively absorb and adapt to large volatility shocks, whilst minimising any significant error. During the stable periods, the distribution will remain sharply peaked around 0, with minimal variations from slight oscillations, underscoring minimal systematic bias.

7.7 Option Price Comparison

Call option contract prices were determined for *SPX*. The time to maturity was set to one month, which allowed for the clearest comparison between models. With a shorter contract duration, the strengths of the ensemble model wouldn't be as apparent, as shorter contracts have less exposure to asset price fluctuations (and other factors) leading to a reduced implied volatility.

The price of the call option, is expected to decrease during a significant market crash event, as the price of the underlying asset decreases. This is clearly seen during each such event (Fig.3. The volatility increases during these events as previously shown (Fig.7.1). The ensemble model

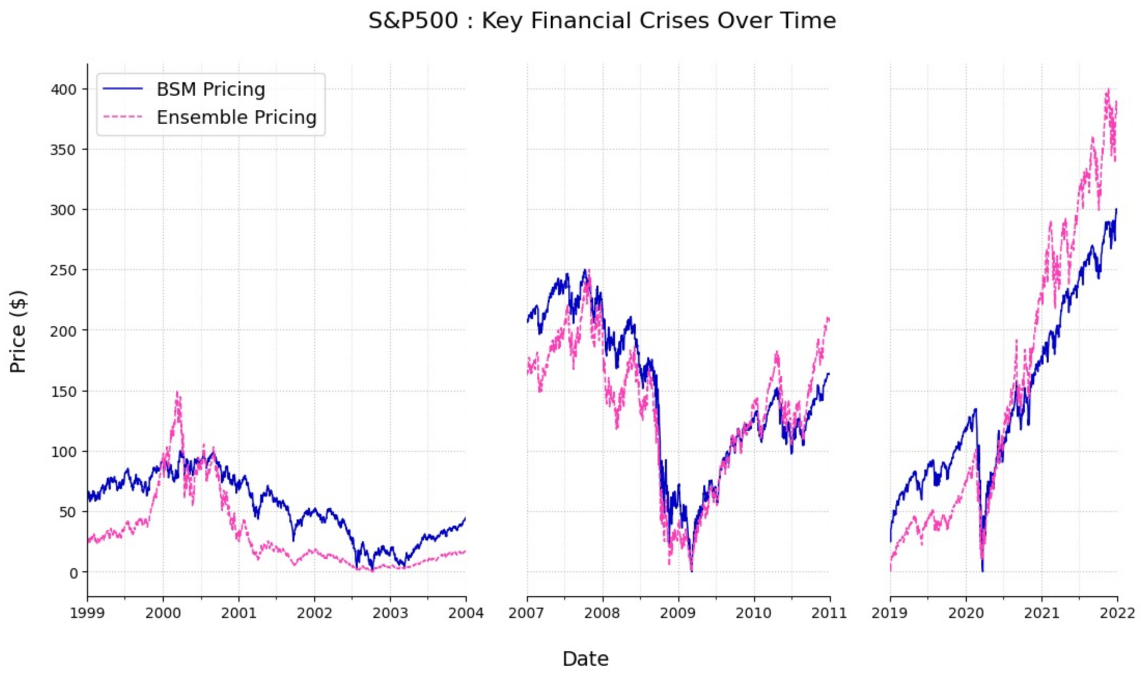

Figure 7.12: Call option contract pricing comparison between the traditional BS model (dark blue) and the ensemble model (magenta). One month option contracts are evaluated daily at a (daily) averaged at-the-money (ATM) strike price. Extreme market events (2000 *Dot-com bubble* crash, 2007-08 financial crisis and Covid-19 market crash in early 2020) are shown to highlight the ensemble model's improved pricing responsiveness to market shocks, compared to the traditional Black Scholes pricing.

reacts quicker to sudden market changes and with more significant price adjustments, as a result of its built-in dynamic volatility predictions, in contrast to the BS constant volatility assumption failing to account for these sudden changes in volatility. This is a clear indication that the GAN-LSTM Hybrid component of the ensemble model successfully captures dynamic market conditions, improving option pricing and adaptability compared to the traditional BS model. During periods in which volatility remains lower and more predictable (see Appendix Fig. B3), the performances of the two models are more closely aligned. This highlights the strength of the SABR component of the ensemble in accurately capturing long-term trends.

8. Discussion on Computational Efficiencies

One of the key challenges in this project was balancing the trade-off between the model's ability to accurately predict volatility and computational efficiency. Computational efficiency is a measure of how effectively an algorithm utilises resources (such as time, memory and processing power) to solve a problem. In this project, we solely consider the time resource, as well as the overall complexity.

ML models rely on large datasets and significant computational resources, often making them impractical for time-sensitive real world application. This is one reason why the BS model remains so dominant in the industry. It's not the most accurate, but its speed and simplicity make it *good enough* for many applications. Larger models with more parameters tend to be more accurate but significantly slower, while smaller models are faster but sacrifice accuracy and precision. Efforts to find this balance influenced decision making throughout the project and were guided by metrics such as RMSE per second and time to convergence.

In stochastic models, very little could be changed to reduce run times. For example, the Heston model is highly accurate, able to capture intricate market behaviours such as volatility smiles and skews. However, it is computationally expensive to the point of impracticality, with the second longest run time of all the models considered (Table B1). Additionally, we tested how these models performed under more extreme market conditions artificially created through MC price path simulations. Stochastic models tended to slow down significantly in highly volatile scenarios due to the increased need for parameter recalibration and additional Monte Carlo paths. ML models benefited from parallelization (in which a large task can be broken down into smaller sub-tasks and performed simultaneously), reducing prediction times. Ultimately, combining SABR and GAN-LSTM Hybrid into an ensemble model, with a runtime of 17 seconds (see Table B1) represents the best balance we could achieve whilst maximising accuracy.

9. Conclusion

This project successfully enhanced the traditional Black-Scholes model by incorporating stochastic volatility models and ML techniques to replace the existing assumption of constant volatility. By combining the SABR and GAN-LSTM Hybrid models, a final ensemble model was created enabling dynamic volatility predictions, across both stable and extreme market volatility regimes. This optimal ensemble combines the macro influence strengths into the memory of the GAN-LSTM to produce more granular short term micro adjustments, all whilst balancing accuracy and efficiency with a final model run time of 17 seconds, placing it well for real world, real time application. The ensemble model consistently aligns with the true volatility more accurately, whilst producing smaller errors, with a 50% improvement on error spread, ($\pm 0.2\sigma$) vs standard BS ($\pm 0.4\sigma$), alongside fewer large deviations and no significant outliers compared to the flawed constant volatility BS approach. This improvement therefore validates our core objective, to incorporate live adaptive volatility to the BS model to reduce its pricing error and enhance its reactions to dynamic markets, by doing this we produce a framework to efficiently evaluate more reliable options premiums that reflect actual market behaviour and dynamics at that time, rather that assuming a static market.

Several challenges were faced throughout the duration of the project. Computational inefficiencies, particularly in stochastic models like Heston and Lévy, limited the prospect of practical implementation. Throughout the development process, models occasionally struggled to adapt to extreme market conditions, highlighting areas for improvement. Limited access to market data restricted the training and evaluation of ML models. Our training dataset consisted of 3491 daily entries across 14 years. In comparison, market volatility is observed to fluctuate on a tick-basis, changing tens of thousands of times per day. Future work could focus on expanding the dataset by integrating synthetic data generation techniques to complement real-world data, allowing for more robust model training. Further refinement of the GAN-LSTM hybrid, especially in capturing rare market events, and exploring alternative ensemble strategies could improve the models applicability to real markets. While not yet optimized for real-time pricing, the developed framework provides a strong foundation for advancing options pricing methods and underscores the complexity of financial markets.

Appendices

Appendix A: Extended Review of Volatility Models

Lévy Process Models

Lévy process models are a class of model based on the work produced by Paul Lévy in the 1930s [20]. These models differ from other stochastic processes by generalising Brownian motion. They allow for discontinuous price movements called 'jumps', which better reflect real market behaviours compared to models assuming continuous price paths [21]. They are used in derivatives pricing to model asset returns with heavy tails (higher chances of extreme events happening than what the normal bell curve predicts) and skewness (asymmetry in the data). They can capture these features which are often observed in real asset returns, this is something that BS fails to capture [20].

Lévy processes contribute to capturing market risk by incorporating stochastic processes that accommodate sudden price movements. This is a valuable aspect for more realistic derivatives pricing. The Variance Gamma (VG) process is a prominent example of a Lévy process and is widely used in financial modelling for its ability to capture the heavy tails and skewness observed in asset returns. Developed by Madan and Seneta in the 1990s, it extends the BS framework by introducing jumps while maintaining analytical tractability, making it particularly useful for pricing derivatives and managing financial risk [22]. VG allows for discontinuity, whereas BS is inherently continuous. Therefore, this allows a non-symmetric distribution, which is more realistic than BS. However, VG is too complex, slow and intensive for practical implementation, because of its parameter estimation, simulation requirements and high computational expense.

Generalised Autoregressive Conditional Heteroskedasticity (GARCH)

The GARCH model is a statistical model developed in 1986 by Bollerslev as an extension of Engle's ARCH model[23]. GARCH models the volatility of time series data by assuming that current volatility depends on past squared observations and past variances. The model is particularly useful for capturing volatility clustering, where large changes in asset prices are often followed by more large changes and small changes are often followed by more small changes. The stationary requirements means that GARCH assumes that the time series data is stationary with a constant mean and variance over time, which may require financial dataset preprocessing. GARCH is quite commonly used in financial applications such as volatility in interest rates, where it excels in volatility clustering. One downside is that the assumption of symmetry in standard GARCH models means that they treat positive and negative shocks of the same magnitude as having an identical impact on future volatility. Consequently, whether the market moves up or down by a certain percentage, the model assumes the resulting change in volatility will be the same. It is also not as computationally expensive, but fitting the model to the data can be challenging due to numerical optimisation requirements.

Heston Model

The Heston model is a stochastic model introduced in 1993 by Steven Heston, it extends the BS model by incorporating stochastic volatility, allowing volatility itself to fluctuate over time [24]. This enables the model to better capture the volatility observed in options markets. The model is defined by a system of stochastic differential equations that represent both the underlying asset price and its volatility. Due to its ability to provide a closed-form solution for European options, the Heston model is commonly used in real-time derivatives pricing and risk management. However, the model is computationally intensive and complex, as it requires inverse Fourier transforms. Additionally, it relies on finite difference methods, various approximations and Monte Carlo simulations, making analytical solutions exceptionally challenging to obtain.
placeins

Appendix B: Model Performance Comparison

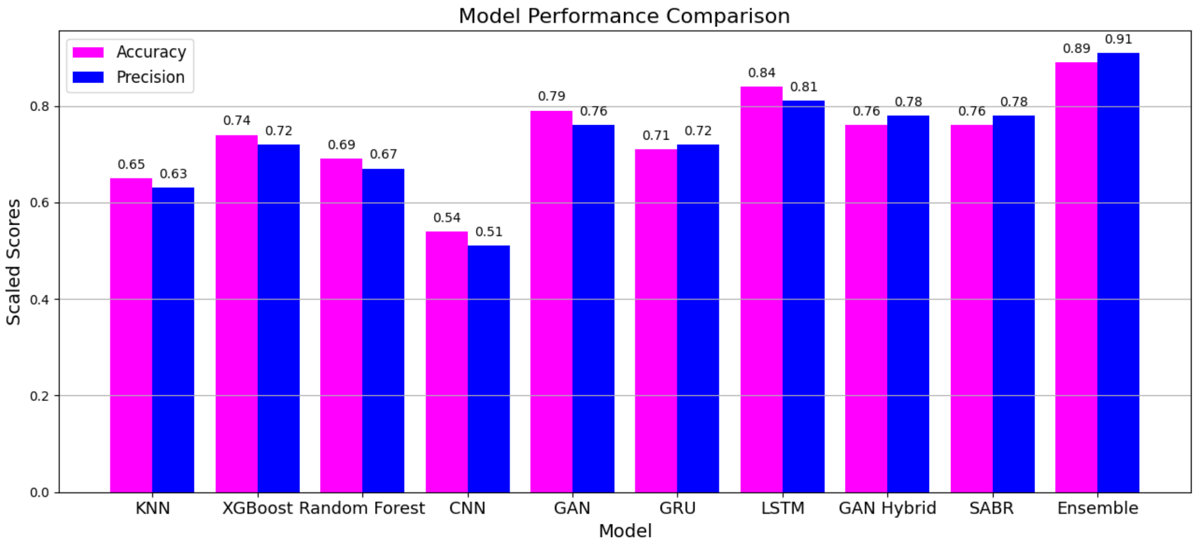

Fig. B1: Model performance comparison presenting the accuracy (magenta) and precision (dark blue) scores (scaled 0-1) for each model, as determined from the confusion matrix [29]. The ensemble model displays the highest performance in both metrics (accuracy: 0.89, precision: 0.91). The CNN model exhibits the lowest performance (accuracy: 0.54, precision: 0.51).

Further Ensemble Model Volatility Predictions

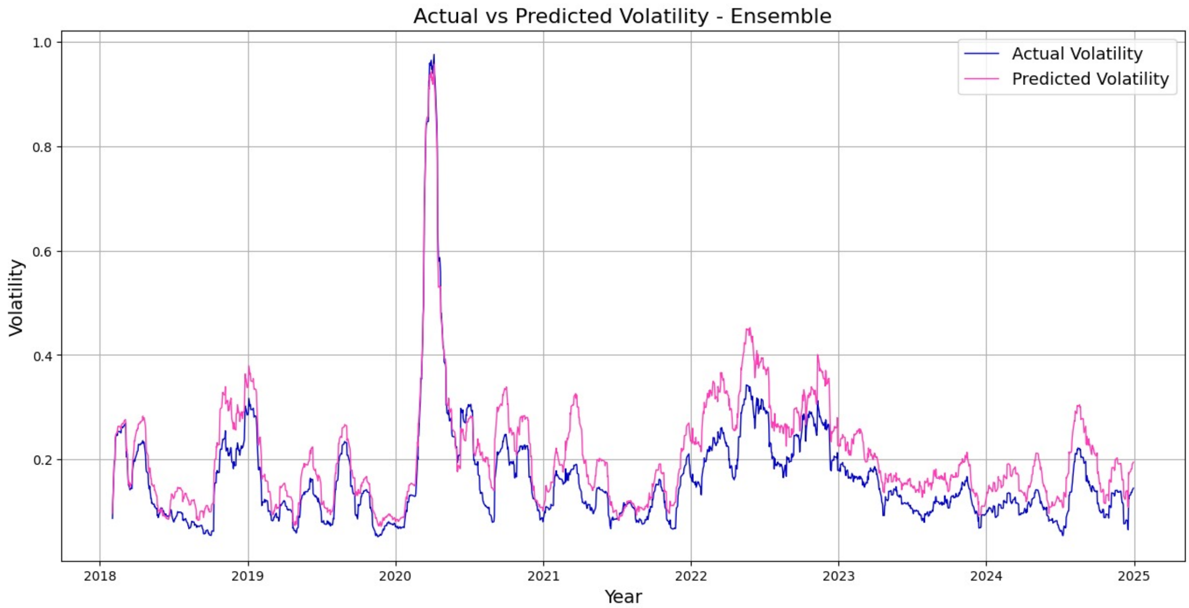

Fig. B2: Volatility predictions of Ensemble model vs True volatility (2018 to 2024): This plot illustrates the ensemble model prediction performance in both macro and micro regimes. The true volatility (blue) and predicted volatility (magenta) align closely, demonstrating models ability to capture long-term structural market behaviors through SABR and short-term fluctuations through the GAN component. Showcases impressive robustness and adaptability across various market scenarios. We notice consistent overprediction due to the GAN intrinsically adapting for more extreme scenarios over time, and therefore reflecting them on all periods.

Option Price Comparison

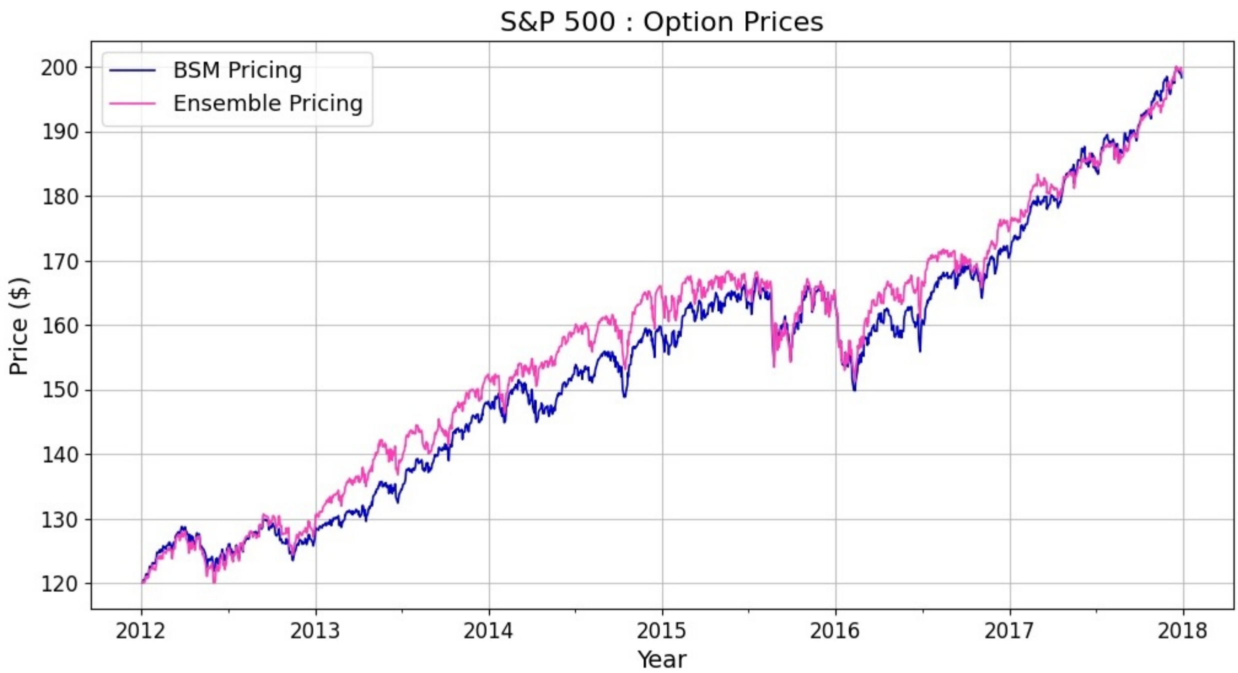

Fig. B3: Call option contract pricing comparison of the traditional BS model (dark blue) and the ensemble model (magenta), between 2012 and 2018. One month option contracts are evaluated daily at a (daily) averaged at-the-money (ATM) strike price.

Model Run Times

TABLE B1: Model Run Times

Model:	Run Time (s):
KNN	0.07
XGBoost	0.37
RF	3.35
CNN	3.63
SABR	4.31
ARIMA	5.76
LSTM	6.92
GAN-HYBRID	7.11
GRU	7.61
GARCH	10.50
Ensemble	17.00
Heston	390.54
Levy	406.18

TABLE B1: Comparison of the run times (in seconds) for each model, executed on a device with an Apple M2 Pro Processor and 16GB RAM.

Bibliography

[1] Black, F. & Scholes, M. (1973) *The Pricing of Options and Corporate Liabilities*, The Journal of Political Economy, 81(3), pp. 637–654. Available at: https://doi.org/10.1086/260062 (Accessed: 4 November 2024). 4, 8, 9

[2] Murphy, J.J. (1999) *Study Guide for Technical Analysis of the Financial Markets: A Comprehensive Guide to Trading Methods and Applications*, New York Institute of Finance, pp. 1–16 (Accessed: 3 November 2024). 4

[3] Hull, J. (2018) *Options, Futures, and Other Derivatives*, Boston: Pearson Education Limited, pp. 23–41, 234–251, 453–465. Available at: https://elibrary.pearson.de/book/99.150005/9781292212920 (Accessed: 28 November 2024). 5, 6, 7, 9

[4] European Securities and Markets Authority (ESMA) (2023) *ESMA Market Report EU Derivatives Markets 2023*, Risk Analysis and Economics Department. Available at: https://www.esma.europa.eu/document/eu-derivatives-markets-2023 (Accessed: 13 November 2024). 5

[5] Mantegna, R.N. & Stanley, H.E. (1999) *Introduction to Econophysics: Correlations and Complexity in Finance*, Cambridge University Press, pp. 1–13. Available at: https://doi.org/10.1017/CBO9780511755767.001 (Accessed: 4 November 2024). 7

[6] Eberlein, E. & Kallsen, J. (2019) *Mathematical Finance*, Springer Finance, pp. 21–22. Available at: https://doi.org/10.1007/978-3-030-26106-1 (Accessed: 5 November 2024). 8

[7] Box, G.E.P. & Jenkins, G.M. (1970) *Time Series Analysis: Forecasting and Control*, San Francisco: Holden-Day, pp. 1–500 (Accessed: 16 October 2024). 10

[8] Brockwell, P.J. & Davis, R.A. (2002) *Introduction to Time Series and Forecasting*, 2nd ed., New York: Springer, pp. 1–450. Available at: https://doi.org/10.1007/978-1-4757-3265-6 (Accessed: 19 October 2024). 10

[9] Fix, E. & Hodges, J.L. (1951) *Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties*, reprinted in *International Statistical Review*, 57(3), pp. 238–247. Available at: https://doi.org/10.2307/1403797 (Accessed: 24 October 2024). 10

[10] LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., & Jackel, L.D. (1989) *Backpropagation Applied to Handwritten Zip Code Recognition*, *Neural Computation*, 1(4), pp. 541–551. Available at: https://doi.org/10.1162/neco.1989.1.4.541 (Accessed: 27 October 2024). 11

[11] Goodfellow, I., Bengio, Y., & Courville, A. (2016) *Deep Learning*, Cambridge, MA: MIT Press (Accessed: 27 October 2024). 11

[12] Krizhevsky, A., Sutskever, I., & Hinton, G. (2012) *ImageNet Classification with Deep Convolutional Neural Networks* in *Advances in Neural Information Processing Systems*, Lake Tahoe, NV, pp. 1097–1105 (Accessed: 27 October 2024). 11

[13] Ho, T.K. (1995) *Random Decision Forests* in *Proceedings of the 3rd International Conference on Document Analysis and Recognition*, Montreal, Canada, pp. 278–282 (Accessed: 29 October 2024). 11

[14] Hochreiter, S. & Schmidhuber, J. (1997) *Long Short-Term Memory*, *Neural Computation*, 9(8), pp. 1735–1780. Available at: `https://doi.org/10.1162/neco.1997.9.8.1735` (Accessed: 2 November 2024). 11

[15] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014) *Generative Adversarial Nets* in *Advances in Neural Information Processing Systems*, Montreal, Canada, pp. 2672–2680 (Accessed: 21 October 2024). 12

[16] LeCun, Y., Bengio, Y., & Hinton, G. (2015) *Deep Learning*, *Nature*, 521(7553), pp. 436–444. Available at: `https://doi.org/10.1038/nature14539` (Accessed: 18 November 2024).

[17] Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). *Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation*, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1724–1734. Available at: `https://doi.org/10.3115/v1/D14-1179` (Accessed: 30 October 2024). 12

[18] Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014) *Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling*, *arXiv preprint*, Available at: `https://arxiv.org/abs/1412.3555` (Accessed: 30 October 2024). 12

[19] Chen, T. & Guestrin, C. (2016) *XGBoost: A Scalable Tree Boosting System* in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, pp. 785–794. Available at: `https://doi.org/10.1145/2939672.2939785` (Accessed: 1 November 2024). 11

[20] Lévy, P. (1948) *Processus Stochastiques et Mouvement Brownien*, Paris: Gauthier-Villars, pp. 1–398 (Accessed: 8 November 2024). 26

[21] Cont, R. & Tankov, P. (2004) *Financial Modelling with Jump Processes*, Boca Raton: CRC Press, pp. 1–527. Available at: `https://doi.org/10.1201/9780203485217` (Accessed: 11 November 2024). 26

[22] Madan, D.B., & Seneta, E. (1990) *The Variance Gamma (VG) Model for Share Market Returns*, *Journal of Business*, 63(4), pp. 511–524. Available at: `https://doi.org/10.1086/296519` (Accessed: 17 November 2024). 26

[23] Engle, R.F. (1982) *Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of United Kingdom Inflation*, *Econometrica*, 50(4), pp. 987–1007. Available at: `https://doi.org/10.2307/1912773` (Accessed: 6 November 2024). 26

[24] Heston, S.L. (1993) *A Closed-Form Solution for Options with Stochastic Volatility with Applications to Bond and Currency Options*, *The Review of Financial Studies*, 6(2), pp. 327–343. Available at: `https://doi.org/10.1093/rfs/6.2.327` (Accessed: 6 November 2024). 27

[25] Hagan, P.S., Kumar, D., Lesniewski, A.S., & Woodward, D.E. (2002) *Managing Smile Risk*, *Wilmott Magazine*, pp. 84–108 (Accessed: 6 November 2024). 13

[26] Carlson, E., Dickson, R., Knudsen, T., & Tatro, Q. (2011) *Technical Analysis Trading Methods and Techniques*, 1st ed., Pearson. Available at: https://learning.oreilly.com/library/view/technical-analysis-trading/9780132938549/book1_ch14.html (Accessed: 4 January 2025). 14

[27] Ying, X. (2019) *An Overview of Overfitting and its Solutions*, *Journal of Physics: Conference Series*, 1168(2), p. 022022. Available at: https://doi.org/10.1088/1742-6596/1168/2/022022 (Accessed: 7 January 2025). 14

[28] Montesinos López, O.A., Montesinos López, A., Crossa, J. (2022) *Overfitting, Model Tuning, and Evaluation of Prediction Performance*, in *Multivariate Statistical Machine Learning Methods for Genomic Prediction*, Springer, Cham. Available at: https://doi.org/10.1007/978-3-030-89010-0_4 (Accessed: 7 January 2025). 14

[29] Sammut, C. & Webb, G.I. (2011) *Encyclopedia of Machine Learning*, Springer Science, p. 209. Available at: https://doi.org/10.1007/978-0-387-30164-8 (Accessed: 8 January 2025). 14, 27

[30] Ahsan, M.Z.M. (n.d.) *Volatility Models: From GARCH to Multi-Horizon Cascades*, Academia.edu. Available at: https://www.academia.edu/70999256/Volatility_Models_From_GARCH_to_Multi_Horizon_Cascades (Accessed: 10 January 2025).