UCLouvain

LOUVAIN
School of Management

Louvain School of Management

Comparative analysis of option pricing models: deep neural network versus traditional approaches

Author: Islin Kamdem
Supervisor: Frédéric Vrins
Academic year 2025-2026
Master 120 - Management Sciences

UCLouvain

LOUVAIN
School of Management

DECLARATION

The declaration below is **mandatory and must appear on the first page (after the front page) of your master's thesis**.

STATEMENTS ON YOUR USE OF GENERATIVE AI

Confirm the following statements:

Statements on responsible use of generative AI	Your answer
The content of my/our master's thesis is my/our original output, even if I/we used generative AI to help prepare it.	☒Yes ☐No
I master the theories, tools and methods that I use for the thesis.	☒Yes ☐No
I verified that the output presented in the thesis is accurate and valid, and that it contains no hallucination, no false reference, etc.	☒Yes ☐No
I acknowledge that I master the final output and that I am able to explain it and answer questions about it.	☒Yes ☐No
I made sure not to provide generative AI tools with access to confidential data or information	☒Yes ☐No

During the preparation of this Master's thesis, the author used Claude (Anthropic) as a supporting tool for the following tasks:

1. **Python and LaTeX support**: assistance with debugging, table formatting, and document layout. The design of the models, the choice of parameters, and the validation procedures remained entirely the responsibility of the author.
2. **English language editing**: assistance with grammar, style, wording, and readability of text drafted by the author, including suggested reformulations of individual sentences and paragraphs, which the author reviewed and rewrote in his own words.

AI tools were not used to generate original scientific content, analyses, or research findings. Throughout the process, this tool supported the writing, coding, and revision process and did not replace the author's own analysis, judgement, or scientific responsibility. The author reviewed, revised, and approved all final content and assumes full responsibility for this thesis.

SIGNATURE

Sign your declaration:

	Author last name, first name(s)	Date	Signature
1	Kamdem Islin	04/08/2026	I.K

Acknowledgements

First and foremost, I would like to express my gratitude to Prof. Frédéric Vrins for his trust and his valuable guidance, which contributed to the completion of this work.

I would also like to thank my parents, without whom none of this would have been possible. Their unconditional support will remain forever engraved in my mind.

A heartfelt thank you to Selma Ersoy, who supported me through the difficult days and the sleepless nights, all along this journey.

Abstract

This thesis compares the ability of three option pricing approaches, the Black-Scholes model, the Heston stochastic volatility model and a neural network, to approximate the implied volatility surface of S&P 500 European call options over the period 2007-2026. The models are evaluated in implied volatility space on several million observations, across market regimes and moneyness segments, with performance rankings validated by a statistical test, the Model Confidence Set. The results establish the dominance of Heston: the model records the lowest errors in both calm and stress regimes. The comparison between Black-Scholes and the neural network reveals no overall winner but a clear division across moneyness: the parametric model prevails near-the-money, while the network dominates at the extremes, most clearly under stress, where the misspecification bias of a constant volatility is most pronounced.

Keywords: Option Pricing, Black-Scholes, Heston, Stochastic Volatility, Neural Network, Implied Volatility, Volatility Smirk, Model Confidence Set, Market Regimes, S&P 500.

Contents

List of Symbols v

List of Tables vii

List of Figures ix

1 Introduction 1

2 Literature Review 4

2.1 Parametric models . . . 4

2.2 Nonparametric models . . . 5

3 Theoretical Framework 9

3.1 Machine Learning . . . 9

3.1.1 Supervised learning . . . 9

3.1.2 Neural Network . . . 11

3.2 Option Pricing . . . 14

3.2.1 Derivatives Payoff . . . 14

3.2.2 Arbitrage-Free Pricing and Risk Neutral Valuation . . . 14

3.2.3 Traditional Models . . . 15

4 Methodology 18

4.1 Data . . . 18

4.2 Data cleaning . . . 19

4.3 Splitting set . . . 19

4.4 Homogeneity hint . . . 21

4.5 Moneyness segmentation . . . 21

4.6 Parametric models . . . 21

4.6.1 Black-Scholes . . . 21

4.6.2 Heston . . . 22

4.7 Neural network . . . 23

4.7.1 Architecture . . . 23

4.7.2 Training phase . . . 23

4.7.3 Test phase . . . 24

4.8 Calibration design . . . 25

4.9 Implied volatility perimeter . . . 25

4.10 Performance metrics . . . 25

4.10.1 Root mean square error . . . 26
4.10.2 Mean error . . . 26
4.10.3 The coefficient of determination . . . 26
4.11 Model Confidence Set . . . 27

5 Empirical Results and Analysis **28**
5.1 Global results . . . 28
5.2 Results by market regime . . . 29
5.2.1 Calm periods . . . 29
5.2.2 Stress periods . . . 30
5.3 Statistical tests . . . 30
5.4 Robustness Analysis . . . 31
5.4.1 Day-before test . . . 31
5.4.2 VIX threshold sensitivity . . . 32

6 Discussion **33**
6.1 Global patterns . . . 33
6.1.1 Black-Scholes . . . 33
6.1.2 Heston . . . 33
6.1.3 Neural network . . . 34
6.2 Calm regimes . . . 35
6.3 Stress regimes . . . 35
6.4 Special Cases . . . 36
6.5 Hypotheses . . . 37
6.6 Limitations and directions for future research . . . 38

7 Conclusion **40**

A Detailed Results **41**

B MCS Results **46**
B.1 Three models . . . 47
B.2 Two models: Neural Network and Black-Scholes . . . 49

C Bias-variance decomposition **51**

D Day-before test **52**

E VIX threshold sensitivity **53**

List of Symbols

Machine learning

$\mathcal{D}$	Training set
$\mathbf{X} \in \mathbb{R}^{N \times p}$	Design matrix
$\mathbf{x}_i$	Input vector of observation i (features)
y_i	Output of observation i (target)
N	Number of observations
p	Number of features
f, $\hat{f}$	Prediction function and its estimator
H	Number of hidden layers
K	Number of neurons per hidden layer
η	Learning rate
$\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}$	Parameter vector and its estimator

Option pricing

S_t	Price of the underlying asset at time t
K	Strike price
$\tau_i = T_i - t$	Time to maturity of option i
V_t	Price of a derivative at time t
C_t	Price of a European call option at time t
r	Risk-free interest rate
σ	Volatility
v_0	Initial variance (Heston)
κ	Mean reversion speed (Heston)
$\bar{v}$	Long-term variance (Heston)
ξ	Volatility of volatility (Heston)
ρ	Correlation between price and variance shocks (Heston)

Calibration and evaluation

S/K	Moneyness ratio (homogeneity hint)
$\hat{\sigma}_s$	Black-Scholes volatility calibrated at date s
σ_{ivo}	Implied volatility provided by IVolatility
$C_{i,t}^{BS}$, C_i^H	Call prices generated by Black-Scholes and Heston
$\hat{\mathcal{M}}_{1-\alpha}^*$	Set of superior models of the MCS test
α	Significance level of the MCS test

List of Tables

4.1 Market regimes defined by the VIX threshold 20
4.2 Moneyness segmentation 21
4.3 Heston calibration bounds 22
4.4 Grid search hyperparameters 23
4.5 Neural network inputs and target 24

5.1 Performance metrics in-sample (2007–2016) and out-of-sample (2020–2026) 28
5.2 MCS results pooled by market regime ($\alpha = 5\%$). A checkmark indicates that the model is retained in the superior set $\hat{\mathcal{M}}^*_{1-\alpha}$. 31

A.1 Performance metrics by moneyness segment – Stress COVID 42
A.2 Performance metrics by moneyness segment – Stress 2022 42
A.3 Performance metrics by moneyness segment – Stress 2025 43
A.4 Performance metrics by moneyness segment – Pre COVID 43
A.5 Performance metrics by moneyness segment – Post COVID 44
A.6 Performance metrics by moneyness segment – Normal 2 44
A.7 Performance metrics by moneyness segment – Normal 3 45

B.1 Model Confidence Set results by period and moneyness segment ($\alpha = 5\%$) . 46
B.2 MCS p-values by period (three models) 47
B.3 MCS p-values by moneyness (three models) 47
B.4 MCS p-values by segment – Pre COVID (three models) 47
B.5 MCS p-values by segment – Stress COVID (three models) 47
B.6 MCS p-values by segment – Post COVID (three models) 48
B.7 MCS p-values by segment – Stress 2022 (three models) 48
B.8 MCS p-values by segment – Normal 2 (three models) 48
B.9 MCS p-values by segment – Normal 3 (three models) 48
B.10 MCS p-values by period (NN vs BS) 49
B.11 MCS p-values by moneyness (NN vs BS) 49
B.12 MCS p-values by segment – Pre COVID (NN vs BS) 49
B.13 MCS p-values by segment – Stress COVID (NN vs BS) 49
B.14 MCS p-values by segment – Post COVID (NN vs BS) 50
B.15 MCS p-values by segment – Stress 2022 (NN vs BS) 50
B.16 MCS p-values by segment – Normal 2 (NN vs BS) 50
B.17 MCS p-values by segment – Normal 3 (NN vs BS) 50

D.1 Day-before test results 52

E.1 Market regimes defined by the alternative VIX thresholds 54
E.2 VIX threshold sensitivity – Pre COVID 55
E.3 VIX threshold sensitivity – Stress COVID 55
E.4 VIX threshold sensitivity – Post COVID 56
E.5 VIX threshold sensitivity – Stress 2022 56
E.6 VIX threshold sensitivity – Normal 2 57
E.7 VIX threshold sensitivity – Stress 2025 57
E.8 VIX threshold sensitivity – Normal 3 58

List of Figures

3.1 Feedforward neural network architecture. 11

4.1 S&P 500 closing price from January 2007 to April 2026. 18
4.2 VIX daily closing values from January 2007 to April 2026. 20

5.1 Out-of-sample RMSE (left) and ME (right) of implied volatility. Moneyness is expressed in percent: negative values correspond to OTM calls, positive values to ITM calls. Positive ME indicates an overestimation of market IV, negative values indicates underestimation. 28
5.2 RMSE (left) and ME (right) of implied volatility in calm conditions. Conventions as in Figure 5.1. 29
5.3 RMSE (left) and ME (right) of implied volatility in stressed conditions. Conventions as in Figure 5.1. 30

A.1 RMSE and ME of implied volatility by moneyness segment across calm periods. 41
A.2 RMSE and ME of implied volatility by moneyness segment across stressed periods. 41

Chapter 1

Introduction

Financial markets are complex systems. Modelling them to isolate analysable mechanisms requires abstracting away from certain dimensions of reality. As Friedman (1953) argued, one must accept a loss of realism to gain understanding and to establish a tractable theoretical framework. In asset pricing, this approach has led to the development of financial models that are now widely used in practice.

Among financial instruments, derivatives play a central role in modern markets. Their value is derived from an underlying asset, such as a stock or a commodity. In June 2025, the market for these assets had a gross market value of $21.8 trillion (Bank for International Settlements, 2025).

This study will focus on options. It is a derivative contract that grants the holder the right, but not the obligation, to buy or sell an underlying asset at a fixed price, known as the strike price, within a specified period of time. Specifically, a call option grants the right to buy, a put option grants the right to sell, and the amount paid to get an option is the premium. This type of contract is used as a tool for hedging risk and for taking speculative positions (Hull, 2022).

Options provide a rigorous testing framework for researchers. Unlike an underlying asset, whose price results directly from the interaction of supply and demand, the price of an option depends on a functional relationship between several variables (Lindo, 2018). The difference between the market price of an option and the price derived from a model, i.e., the pricing error, gives a direct indication of the model's misspecification, reflecting the model inability to fully capture the pricing structure of the market (Bakshi et al., 1997). Hence, the aim is to have an accurate approximation consistent with the no-arbitrage argument (Hull, 2022).

For a single underlying asset, there exists a wide range of contracts differing in strike price and expiry date, each trading at its own market price. Models serve as a common currency for estimating and comparing these option prices (Lindo, 2018). Moreover, if they fail to satisfy put-call parity, arbitrage opportunities arise (Hull, 2022). The price of an option depends on unobservable variables such as future volatility; models are hence needed to approximate option prices. However, the relationship between option prices and market parameters used by models is nonlinear, which increases the difficulty of the approximation. The reason for this nonlinearity is the asymmetric payoff structure of an option, characterised by potential high gains with loss limited to the premium paid (Hull, 2022; Lindo, 2018).

By purchasing an option, one can take a position on the price movements of an underlying asset without having to hold it directly, allowing to capitalise on market fluctuations to generate profit. In the case of commodities, options also eliminate physical constraints, since a position can be closed out without ever requiring delivery of the underlying asset. This enables faster execution and supports larger trading volumes than those observed in the spot market (Lindo, 2018).

The underestimation of risks in the valuation of complex derivatives played a central role in the subprime crises, whose effects spread far beyond the United States (Ilalan, 2015; Wen et al., 2012). This episode highlights how devastating the consequence of an inadequate pricing model can be for national stability and raises a fundamental question: Are valuation models robust during periods of market stress?

Traditional pricing models rest on behavioural and statistical assumptions. von Neumann and Morgenstern (1944) formalised agent rationality: individuals maximise expected utility by making the best possible decision. Merton (1973) assumes that statistical market properties such as volatility and correlation remain stable. Fama (1970) efficient market hypothesis holds that prices reflect all available information instantaneously. These assumptions do not hold unconditionally: economic data is inherently non-stationary (Hendry, 2009), and agents' rationality can be questioned even outside periods of stress. Market shocks simply make the resulting breakdown of traditional pricing models most visible.

In response to the limitation of traditional models, data-driven approaches have emerged. Engsted (2009) explains that the quality of a model lies not in its theoretical elegance, but in its ability to capture the properties of observed data, a criterion that forms the core logic of these new approaches. Rather than adhering to an explicit theoretical structure, these methods empirically identify the relationships relevant to asset pricing from the data (Kelly & Xiu, 2023). In practice, machine learning methods consistently outperform traditional linear models in terms of out-of-sample R^2; the former generalise well to new data (Gu et al., 2020). Moreover, Gu et al. (2020) show that investment strategies based on data-driven models achieve an out-of-sample Sharpe ratio of 0.77, compared with 0.51 for a passive strategy.

Neural networks have attracted particular attention from researchers and practitioners. Their ability to approximate complex and nonlinear functions, without being restricted to a predefined functional structure, makes them natural candidates for option pricing (Hutchinson et al., 1994). Compared to other machine learning models, the nonparametric model consistently ranks among the best performing for option pricing (Amilon, 2003; Ghaziri et al., 2000; Mitra, 2012). However, their superiority remains nuanced. Malliaris and Salchenberger (1993) prove that neural networks do not always outperform parametric models. Although models have evolved considerably since then, İltüzer (2022) shows that this remains the case. The question of robustness remains open, particularly during

periods of market stress.

This study, entitled 'Comparative analysis of option pricing models: deep neural network versus traditional approaches', addresses the following research question: Which of the Black-Scholes, Heston and neural network models provides the most accurate estimates of European call options prices on the S&P 500 index, and how does their relative performance vary across market regimes and moneyness segments?

The main contribution to the literature lies primarily in the scale and scope of its empirical environment. It covers 19 years of S&P 500 options data, comprising millions of observations spanning multiple market regimes, whereas most comparisons in the literature between parametric and nonparametric models rely on shorter sample periods, fewer observations, or a single market environment. This broad setting allows the models' ability to recover current fair prices to be assessed under a wide range of market conditions and makes it possible to identify whether specific market contexts favour one model over the others.

Chapter 2 reviews the literature on parametric and nonparametric option pricing models, highlighting their key findings and limitations.

Chapter 3 establishes the theoretical framework, covering supervised learning and neural networks as well as the pricing models used in this study.

Chapter 4 presents the data and the empirical procedure.

Chapter 5 reports the results across market regimes and moneyness segments and assesses their statistical significance.

Chapter 6 interprets these results, assesses the study's hypotheses, and discusses its limitations and directions for future research.

Chapter 7 summarises the key findings.

Chapter 2

Literature Review

2.1 Parametric models

Numerous models have been developed to address the complex challenge of estimating the fair value of financial assets. The most widely known is the Black-Scholes model (Black & Scholes, 1973). Originally, the formula provided a theoretical estimate of the price of a European option; it was subsequently generalised by Merton (1973), who established a no-arbitrage theoretical framework, incorporated dividends, and extended the analysis to American-style options.

Several studies have analysed the model's performance. Black (1975) supports the use of the option value derived from the formula as a solid benchmark for gauging the degree of an option's overpricing. He establishes that the values produced for deep OTM and deep ITM options diverge from market prices. MacBeth and Merville (1979) reach the same conclusion. These pricing errors at the extremes of moneyness correspond precisely to cases in which the model suffers from a specification bias; its functional structure differs from the true data-generating function of option prices.

One key assumption is primarily responsible for this: the constant volatility. An empirical phenomenon known as the volatility smile shows that the market's estimate of expected volatility i.e., the implied volatility, varies with moneyness (Rubinstein, 1994). On equity indices, this smile takes an asymmetric form known as the smirk, with higher implied volatilities on the OTM put side (Rubinstein, 1994). The aforementioned of Black-Scholes is mainly attributable to the model's inability to capture this mechanism. Black and Scholes (1972) had already shown that a model using an accurate estimate of volatility over the derivative period reduces pricing errors, suggesting that the model's errors stem primarily from volatility misestimation.

Several extensions have been proposed. Merton (1976) incorporates a component of discontinuous jumps following a Poisson process to capture sudden market movements. Scott (1987) attempted to relax the constant-volatility assumption by allowing volatility to follow a stochastic process but obtained a solution that still required numerical methods. Stein and Stein (1991) extend this approach using Fourier transform methods, yielding a semi-closed-form solution that omits the negative correlation between implied volatility and the spot price.

Heston (1993) follows a similar approach and also derives a semi-closed-form solution, so that computationally intensive numerical methods such as Monte Carlo simulation or

finite-difference schemes are no longer strictly necessary, although numerical integration remains required. Unlike Stein and Stein (1991), however, the model explicitly incorporates a correlation between shocks to volatility and to the spot price, allowing it to successfully capture the volatility smile. Bakshi et al. (1997) consider the incorporation of stochastic volatility as essential to improving the Black-Scholes model, reporting pricing error reduction of 25% to 60%. Sidhu and Saxena (2026) reach the same conclusion. Heston (1993) himself notes that the Black-Scholes model produces prices closely similar to those of stochastic volatility models for ATM options, which are the most actively traded contracts in the market. This observation partially explains how the Black-Scholes model was able to hide its limitations for a long time and became so widely adopted in practice.

Heston's volatility process is governed by parameters calibrated to observed option prices. These parameters are frequently recalibrated to reflect the most recent market conditions. However, given their important fluctuation, this lack of stability leads to a loss of economic interpretability of the parameters (Le Floc'h, 2025). According to Gatheral (2006), the main objective of stochastic volatility models is not to perfectly represent reality, but rather to reproduce market prices effectively while avoiding arbitrage opportunities; as market conditions evolve continuously, the parameters must too. In addition, Sidhu and Saxena (2026) show that the accuracy of the estimated parameters also depends on the choice of the optimiser.

While Heston constitutes an improvement over Black-Scholes, it remains misspecified as it still exhibits systematic pricing errors despite its richer dynamic (Bakshi et al., 1997). Fiorentini et al. (2002) locate this residual bias along moneyness on the Spanish index market: Heston tends to overprice OTM calls and underprice ITM calls. They also find that, unlike Black-Scholes, Heston's pricing errors are unrelated to the level of market volatility. Benzoni (2002) reports a comparable moneyness-dependent error pattern on S&P 500 options. Bates (2003) argues that the choice of the optimal parametric model depends heavily on the prevailing market context; no single parametric model can capture all the specific features of real market conditions.

2.2 Nonparametric models

Given the structural limitations of parametric models described above, nonparametric approaches have emerged as an alternative. In this study, we focus on one nonparametric model: neural networks.

Grounded in data and the universal approximation theorem demonstrated by Cybenko (1989) and Hornik et al. (1989), neural networks are not restricted to a predefined axiomatic framework. Numerous studies have reported the general superiority of neural networks over Black-Scholes in empirical finance. Gu et al. (2020) compare several nonparametric models for stock return forecasting and identify neural networks among the

best-performing approaches. Gross et al. (2025) show that neural network forecasts are statistically superior to those of the parametric benchmark. This superiority is also observed in the specific task of option pricing approximation, as discussed by Ghaziri et al. (2000) and Saito and Jun (2000) on SPX call options. Amilon (2003) reaches the same conclusion for the Sweden market, Mitra (2012) too for the Indian one.

Beyond merely confirming previous findings, Garcia and Gençay (2000) show that a homogeneity hint can be incorporated into the neural network model, such that the spot price S and the strike price K are no longer considered as independent variables, but rather as the moneyness ratio S/K. This embedded financial knowledge reduces the model's function space and significantly lowers the out-of-sample error.

However, this superiority is not universal. Malliaris and Salchenberger (1993), using options on the S&P 100, obtain results in which neither model consistently dominates the other: Black-Scholes outperforms in ITM, while the network outperforms in OTM, with larger errors observed in this latter segment for both models. İltüzer (2022) similarly notes an absence of total systematic dominance on the Turkish market: for calls, the nonparametric model is superior during stable periods, and the trend reverses during periods of stress. She also notes that the models find calls on the ITM segment more difficult to price, in contrast to the two previous researchers. Lajbcygier (2004) reaches the same general conclusion on the Australian market. Evidence against Heston is scarcer: using options written on Apple stock, Milyushkov (2025) finds that neural networks fit the option pricing function better than Heston. But his study relies on a single trading day, and the Heston parameters were taken from empirical literature, leaving open the possibility that a calibrated Heston, as used in practice, would narrow this gap.

Hutchinson et al. (1994) provide a way to organise these mixed results: whenever the parametric model suffers from a specification bias, neural networks outperform the Black-Scholes model; in the opposite case, the traditional model prevails.

The market regime adds a further dimension to this comparison. A crisis represents a regime shift characterised by price jumps, higher volatility, and the emergence of atypical market dynamics. In this context, option prices become more sensitive to open interest, reflecting an increased market activity and changes in investor behaviour, while the relative importance of traditional determinants such as moneyness, implied volatility, and time-to-maturity is altered, though the first one remains the most dominant determinant (Gradojevic & Kukolj, 2022).

These periods of market turbulence lead to extreme behaviour, worsening the misspecification of the Black-Scholes model (Bakshi et al., 1997). Its pricing accuracy nevertheless becomes comparable to that of nonparametric models (Gradojevic & Kukolj, 2022). This convergence does not reflect an improvement of the parametric model, Gradojevic and Kukolj (2022) report that market volatility degrades the accuracy of all competing models, and the gap closes because regime shifts limit the generalisation ability of non-

parametric models. The patterns learned from historical data no longer apply in such contexts, which substantially impairs the performance of machine learning approaches (Liu & Huang, 2016). Gradojevic et al. (2009) provide an instructive example: their modular neural network[1], trained on the bull market of early 1987 and tested on the fourth quarter, which includes the October 1987 crash, was outperformed by Black-Scholes in some configurations, a direct consequence of the mismatch between training and testing regimes. This is consistent with Liu and Huang (2016), who argue that a network trained on a bull market period may perform poorly on a crisis regime. Indeed, the fixed structure of the Black-Scholes model makes it less sensitive to regime shifts: it remains biased, but relatively stable in its error (Gradojevic & Kukolj, 2022).

The Heston model also exhibits significant limitations during periods of market stress, despite its stochastic volatility. Jiao et al. (2018) note that the model's volatility process is unable to generate levels as high as those observed in this context. This is confirmed by Rømer (2022), who finds that Heston's fit to the IV surface of SPX options deteriorates sharply around the subprime crisis.

Beyond crisis periods, neural networks also face structural limitations. The use of nonparametric pricing methods requires large amounts of data to achieve an efficient network, limiting their use to derivatives with historical data. Hutchinson et al. (1994) nevertheless note that it is possible to replicate any newly created one using a combination of existing ones.

Furthermore, Ruf and Wang (2020) argue that the choice of network architecture and model parameters is essential to achieving optimal performance. In this regard, Garcia and Gençay (2000), and Hutchinson et al. (1994) themselves, highlight the absence of statistically-based architecture selection criteria in the 1994 study: an overly complex architecture leads to overfitting; the model memorises the relationships in the training data without generalising to new data. Conversely, an architecture that is too simple leads to underfitting; the model fails to sufficiently capture the underlying relationships in the training data (Goodfellow et al., 2016). Selecting an architecture that minimises out-of-sample generalisation error is hence necessary to mitigate the risks of both overfitting and underfitting, and to ensure a fair comparison between parametric and nonparametric models, a methodological rigour that remains absent from many studies (Garcia & Gençay, 2000; Hutchinson et al., 1994).

In addition to these technical challenges, Gradojevic and Kukolj (2022) and Rudin (2019) invoke the term 'black box' to illustrate the interpretability limit inherent to nonparametric models. The internal mechanisms driving the model's output are difficult to interpret and to relate to meaningful economic intuition, in contrast to parametric models, whose parameters carry a generally clear economic interpretation (Rudin, 2019).

[1] A neural network divided into independent sub-networks, called modules, each assigned to a specific task.

Neural networks are thus not perfect. Hutchinson et al. (1994) noted that nonparametric models constitute an alternative to parametric approaches, rather than a definitive solution. The literature suggests that their superiority depends on the market regime and the moneyness segment under consideration. It is against this context that the following hypotheses are examined in the subsequent chapters:

Hypothesis 1: During normal market conditions, neural networks demonstrate superior pricing performance relative to the Black-Scholes and Heston models across all moneyness segments.

Hypothesis 2: During periods of market stress, the relative performance of the three models converges. Neural network performance deteriorates, while parametric model performance becomes comparable to that of neural networks.

Hypothesis 3: The relative superiority of neural networks over parametric models is more pronounced for OTM and ITM options than for ATM options.

Chapter 3

Theoretical Framework

3.1 Machine Learning

Machine learning refers to methods capable of detecting various structures, often complex or difficult to perceive, within large amounts of data to achieve tasks such as prediction, classification, or decision support (Bishop, 2006). While other approaches exist, this study focuses solely on one: supervised learning.

3.1.1 Supervised learning

In supervised learning, the objective is to learn a prediction function $f : \mathbb{R}^p \to \mathbb{R}$, where p denotes the number of features, that maps an input $\mathbf{x}$ to an output y given a training set $\mathcal{D} = (\mathbf{X}, \mathbf{Y}) = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$, where N represents the number of observations (Hastie et al., 2009). Inputs are called 'features' or 'predictors' and outputs are called 'responses' or 'targets'. Supervised learning models fall into two classes: classification, where outputs are categorical, and regression, where outputs are quantitative. This study addresses the second case (Hastie et al., 2009). In practice, the N observations are organised into a design matrix $\mathbf{X} \in \mathbb{R}^{N \times p}$, where each row corresponds to an input vector $\mathbf{x}_i \in \mathbb{R}^p$ and each column to a feature; and into a column vector $\mathbf{Y} \in \mathbb{R}^N$ containing the associated output values y_i.

Considering the following statistical model:

$$\mathbf{Y} = f(\mathbf{X}) + \boldsymbol{\varepsilon} \tag{3.1}$$

where $\boldsymbol{\varepsilon}$ is a random error term such that $E(\boldsymbol{\varepsilon}) = 0$ and $\boldsymbol{\varepsilon} \perp \mathbf{X}$. The function f is unknown; the objective is to estimate it from the observed data. To this end, a loss function is defined to quantify the model's prediction errors. The most commonly used is the squared error loss (Hastie et al., 2009):

$$L(y_i, f(\mathbf{x}_i)) = (y_i - f(\mathbf{x}_i))^2, \quad \forall i = 1, \ldots, N \tag{3.2}$$

The objective is to find a function f that minimises the squared error between the actual output $\mathbf{Y}$ and the prediction $f(\mathbf{X})$, thereby reducing the model's prediction error. This process is referred to as empirical risk minimisation (Hastie et al., 2009). The

expectation of the loss function, which can be interpreted as the risk, is given by:

$$E(L(Y, f(\mathbf{X}))) = E\left((Y - f(\mathbf{X}))^2\right) \tag{3.3}$$

The solution that minimises the expected loss is thus:

$$f^*(\mathbf{x}) = E(Y \mid \mathbf{X} = \mathbf{x}) \tag{3.4}$$

In other words, the best prediction under the squared error loss is the conditional expectation of Y given $\mathbf{X}$. To move from theory to practice, since $f^*(\mathbf{x})$ is unknown, an estimator $\hat{f}(\mathbf{x})$ must be constructed (Hastie et al., 2009) such that:

$$\hat{f}(\mathbf{x}) \to f^*(\mathbf{x}) \quad \text{as } N \to \infty \tag{3.5}$$

To obtain the function to be minimised, the expected loss is replaced by its empirical equivalent, giving the mean squared error (MSE):

$$\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{f}(\mathbf{x}_i))^2 \tag{3.6}$$

Minimising the residual sum of squares (RSS) across all possible functions f:

$$RSS(f) = \sum_{i=1}^{N}(y_i - f(\mathbf{x}_i))^2 \tag{3.7}$$

any function interpolating the training data gives $\text{RSS}(f) = 0$, so the minimisation problem admits infinitely many solutions. To identify a unique one, the space of admissible functions is restricted to a parametric class $f(\mathbf{x}_i; \boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \mathbb{R}^q$ is a finite-dimensional parameter vector.

The empirical loss can then be expressed as a function of the parameter $\boldsymbol{\theta}$:

$$L(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N}(y_i - f(\mathbf{x}_i; \boldsymbol{\theta}))^2 \tag{3.8}$$

The objective is to find the optimal parameter vector $\hat{\boldsymbol{\theta}}$ that minimises the RSS:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) \tag{3.9}$$

Once this optimal parameter vector has been identified, the function estimator is given by:

$$\hat{f}(\mathbf{x}) = f(\mathbf{x}; \hat{\boldsymbol{\theta}}) \tag{3.10}$$

In this study, the parametric family is defined by the architecture of feedforward neural networks:

$$\mathcal{F}_{NN} = \{f(\mathbf{x}; \boldsymbol{\theta}) \mid \boldsymbol{\theta} \in \mathbb{R}^q\} \tag{3.11}$$

3.1.2 Neural Network

This model was originally developed with the aim of replicating the neural connections of the human brain, from which both their name and structural form are derived.

A feedforward neural network is a nonlinear extension of classical linear regression models, consisting of several parametric transformations that link an input layer to an output layer through one or more hidden layers (see Figure 3.1). Model training relies on iteratively adjusting parameters to improve predictive accuracy (Bishop, 2006).

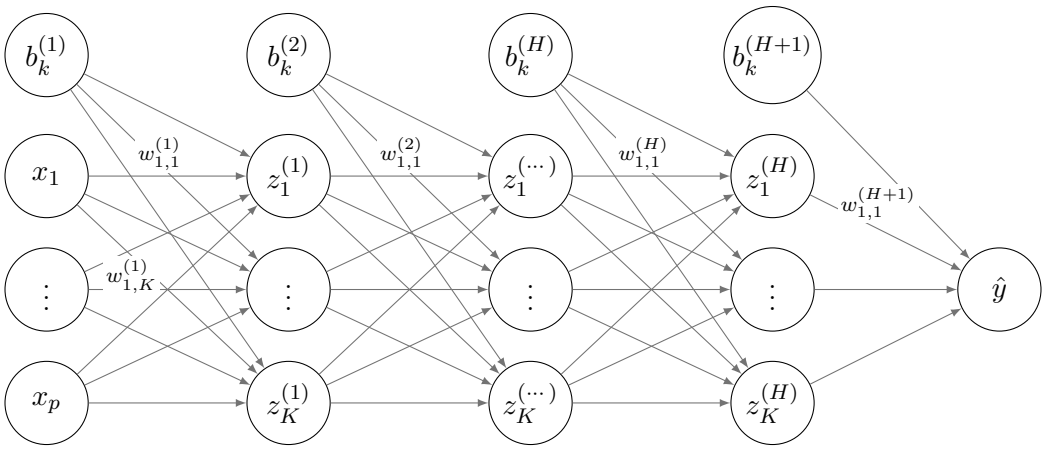


Figure 3.1: Feedforward neural network architecture.

The input layer is the vector $\mathbf{x}_i \in \mathbb{R}^p$, where each neuron represents one input x_{ij}, with i denoting the observation and $j = 1, \ldots, p$ the dimension. A total of H hidden layers are connected, each consisting of K neurons $z_k^{(h)}$, where the superscript (h) indicates the corresponding layer. In practice, K_h may vary from one layer to another, although this study uses a common size K across layers. When the network has only one hidden layer ($H = 1$), it is referred to as a shallow neural network; when it has several ($H > 1$), it is referred to as a deep neural network (Goodfellow et al., 2016).

Each hidden neuron $z_k^{(h)}$ receives all outputs from the previous layer. These are first processed through a linear combination, producing the input unit activation:

$$a_k^{(h)} = \sum_{j=1}^{p} w_{jk}^{(h)} x_j + b_k^{(h)} \tag{3.12}$$

where $w_{jk}^{(h)}$ denotes the weights and $b_k^{(h)}$ the biases, for $k = 1, \ldots, K$ and $h = 1, \ldots, H$. The result is then passed through a nonlinear activation function $g(\cdot)$, here chosen as the

ReLU function (Glorot et al., 2011):

$$g(\alpha) = \max(0, \alpha) \tag{3.13}$$

where $\max(\cdot)$ returns the greater of its two arguments, in order to get the output unit activation:

$$z_k^{(h)} = g(a_k^{(h)}) \tag{3.14}$$

The procedure is repeated until the final layer of the model is reached. Since this study addresses a regression problem, the final activation function is the identity function $Id(x) = x$, leading to an unbounded real-valued output suitable for regression:

$$\hat{y} = Id(a_k^{(H+1)}) \tag{3.15}$$

Therefore, for each neuron unit in the network, the computation is as follows:

$$z_k^{(h)} = g\left(\sum_{j=1}^{K_{h-1}} w_{jk}^{(h)} z_j^{(h-1)} + b_k^{(h)}\right) \tag{3.16}$$

This phase is referred to as forward propagation: the information from the inputs $\mathbf{x}_i$ is propagated through hidden layers to produce a prediction $\hat{y}$ (Goodfellow et al., 2016).

Consider a set of model parameters $\boldsymbol{\theta} = \{\mathbf{W}^{(h)}, \mathbf{b}^{(h)}\}$ for each layer h; the resulting prediction function takes the form:

$$\hat{y}(\mathbf{x}^{(0)}) = f(\mathbf{x}^{(0)}; \boldsymbol{\theta}) = f^{(H+1)} \circ f^{(H)} \circ \cdots \circ f^{(1)}(\mathbf{x}^{(0)}) \tag{3.17}$$

Network training consists of minimising the loss function $L(\boldsymbol{\theta})$ (3.8) by adjusting model parameters. To this end, the gradient of $L(\boldsymbol{\theta})$ with respect to the weights and biases, denoted $\nabla L(\boldsymbol{\theta})$, is computed using the backpropagation algorithm. This algorithm propagates the error from the output layer back through the preceding layers, and the resulting gradient indicates the direction in which the parameters should be adjusted to reduce the error (Bishop, 2006).

The difference between the model's prediction and the actual value is first evaluated at the output layer $H + 1$, giving the error signal $\delta_k^{(H+1)}$. This is defined as the partial derivative of the loss function with respect to the pre-activation of each output unit:

$$\frac{\partial L}{\partial a_k^{(H+1)}} = \delta_k^{(H+1)} \tag{3.18}$$

Since this study addresses a regression problem, the activation function of the output layer is the identity function $Id(\cdot)$, so that $\hat{y}_k = a_k^{(H+1)}$. Under the squared error loss (3.2),

we obtain:

$$\delta_k^{(H+1)} = \hat{y}_k - y_k \tag{3.19}$$

This information is then backpropagated through the network to compute the error signal $\delta_k^{(h)}$ of the units in the preceding layers:

$$\delta_k^{(h)} = g'(a_k^{(h)}) \sum_j w_{jk}^{(h+1)} \delta_j^{(h+1)} \tag{3.20}$$

Thus, by applying the chain rule, the gradients of the parameters constituting each layer h are computed as follows:

$$\frac{\partial L}{\partial w_{ik}^{(h)}} = \frac{\partial L}{\partial a_k^{(h)}} \cdot \frac{\partial a_k^{(h)}}{\partial w_{ik}^{(h)}} = \delta_k^{(h)} z_i^{(h-1)} \tag{3.21}$$

$$\frac{\partial L}{\partial b_k^{(h)}} = \frac{\partial L}{\partial a_k^{(h)}} \cdot \frac{\partial a_k^{(h)}}{\partial b_k^{(h)}} = \delta_k^{(h)} \tag{3.22}$$

The parameters are then updated using stochastic gradient descent. This approach approximates the gradient of the loss function from a mini-batch drawn from $\mathcal{D}$, rather than the entire dataset. The parameters are iteratively adjusted in the direction opposite to the gradient. While several gradient descent methods exist, the Adam optimizer (Kingma & Ba, 2015) is one of the most widely used in deep learning. It adapts the learning rate for each parameter individually using estimates of the first and second moments of the gradient. The update rule is:

$$\boldsymbol{\theta}_{new} = \boldsymbol{\theta}_{old} - \eta \frac{\hat{\mathbf{m}}_t}{\sqrt{\hat{\mathbf{v}}_t} + \epsilon} \tag{3.23}$$

where η is the learning rate, $\hat{\mathbf{m}}_t$ and $\hat{\mathbf{v}}_t$ are the bias-corrected estimates of the first and second moment of the gradient respectively, and $\epsilon > 0$ is a numerical stability term.

Cybenko (1989) and Hornik et al. (1989) show that a shallow neural network with enough neurons can approximate any continuous function defined on a compact domain to arbitrary precision. However, for complex functions, a shallow network may require an excessive number of neurons. Rolnick and Tegmark (2018) demonstrate that a deep network is considerably more efficient in such cases. This property is directly relevant to option pricing. The pricing function $V(S, K, T, \sigma, r)$ is unknown in its general form, and traditional models offer approximations under restrictive assumptions that are often violated in practice. By setting $\mathbf{x} = (S, K, T, \sigma, r)$ and $y = V$, the network learns the estimator $\hat{f}(\mathbf{x}; \hat{\boldsymbol{\theta}}) = \arg\min L(\boldsymbol{\theta})$ directly from the observed data, without imposing any predefined functional structure.

Model complexity introduces a fundamental trade-off between bias and variance, which governs the model's generalisation ability. This is discussed in detail in Appendix C.

3.2 Option Pricing

The fundamental idea of derivatives pricing theory is to construct a portfolio whose payoff replicates that of the derivative under consideration. In the absence of arbitrage, the cost of the replicating portfolio equals the no-arbitrage price of the derivative. This reasoning is developed in detail in the sections that follow.

3.2.1 Derivatives Payoff

An option is a derivative contract that grants its holder the right, but not the obligation, to buy or sell an underlying asset at a fixed price, the strike price K, up to maturity T. A call option grants the right to buy; a put option grants the right to sell. This study focuses solely on European-style options, which can only be exercised at T (Hull, 2022).

At maturity T, the payoff function ψ_T of a call option is:

$$\psi_T = \max(S_T - K, 0) \tag{3.24}$$

where S_T denotes the price of the underlying asset at maturity. The option is only exercised when it is in the money, i.e., when $S_T > K$. Otherwise, the option will not be exercised, as doing so would result in a loss for its holder (Hull, 2022). By contrast, the payoff of a put option is:

$$\psi_T = \max(K - S_T, 0) \tag{3.25}$$

and the option will be exercised only if $S_T < K$. A call is considered in-the-money (ITM) when $S_T > K$ and out-of-the-money (OTM) when $S_T < K$; a put is ITM when $S_T < K$ and OTM when $S_T > K$. Both are considered at-the-money (ATM) when $S_T = K$ (Hull, 2022).

3.2.2 Arbitrage-Free Pricing and Risk Neutral Valuation

Arbitrage refers to the possibility of simultaneously transacting across two or more markets to generate a riskless profit. A market in which no such opportunity exists is said to be arbitrage-free; defining the fair value of a derivative requires consistency with this condition. When it holds, the law of one price (LOP) states that two products with identical payoffs must have identical prices. It follows that the fair price of an option can be obtained by computing the cost of a strategy that replicates its payoff (Vrins, 2025). Under the no-arbitrage condition, there exists a probability measure $\mathbb{P}^*$, equivalent to the real-world measure $\mathbb{P}$, referred to as the risk-neutral measure. In this framework, investors are assumed to be risk-neutral, i.e., they require no additional return for bearing additional risk. Consequently, every asset earns the risk-free rate r, which also serves as the discount rate for any payoff (Hull, 2022). The price of an option at time t can therefore

be rewritten as the discounted expectation of its future payoff:

$$V_t = e^{-r(T-t)} \cdot \mathbb{E}^{\mathbb{P}^*}[\psi_T \mid \mathcal{F}_t] \tag{3.26}$$

This no-arbitrage price is consistent with the Fundamental Theorem of Asset Pricing (Vrins, 2025): a model is arbitrage-free if and only if it admits a risk-neutral measure $\mathbb{P}^*$.

3.2.3 Traditional Models

Black-Scholes-Merton Model

Introduced in 1973 by the three researchers after whom it is named, the Black-Scholes-Merton model's key assumptions are as follows (Black & Scholes, 1973):

1. The risk-free rate r and volatility σ are constant.
2. There are no transaction costs.
3. The stock pays no dividends.
4. Fractional share purchases are permitted.
5. The stock price is lognormally distributed.

The stock price S_t is assumed to follow a geometric Brownian motion (GBM):

$$dS_t = \mu S_t \, dt + \sigma S_t \, dW_t \tag{3.27}$$

where $\mu S_t \, dt$ represents the drift component and $\sigma S_t \, dW_t$ the diffusion component, with W_t denoting a Brownian motion. Black and Scholes (1973) derive an analytical solution for the price of a European call:

$$C_t = S_t \mathcal{N}(d_1) - K e^{-r(T-t)} \mathcal{N}(d_2) \tag{3.28}$$

where

$$d_1 = \frac{\ln(S_t/K) + \left(r + \frac{\sigma^2}{2}\right)(T-t)}{\sigma\sqrt{T-t}}, \qquad d_2 = d_1 - \sigma\sqrt{T-t} \tag{3.29}$$

Heston Model

The Black-Scholes model assumes a constant volatility σ, yet the volatility smile shows that, in practice, implied volatility varies with strike price K and maturity T: $\sigma(K, T)$. Heston (1993) proposes a stochastic volatility model to relax this assumption, while building on the same arbitrage-free and risk-neutral foundations as Black-Scholes. Its key assumptions are as follows (Heston, 1993):

1. The risk-free rate r is constant.
2. The instantaneous variance v_t follows a stochastic process.
3. There exists an arbitrary correlation ρ between shocks to S_t and shocks to v_t: $\rho\, dt = dW_{1,t}\, dW_{2,t}$.

The asset price S_t follows a geometric Brownian motion analogous to (3.27):

$$dS_t = \mu S_t\, dt + \sqrt{v_t}\, S_t\, dW_{1,t} \tag{3.30}$$

where the constant volatility σ is replaced by the stochastic term $\sqrt{v_t}$, and $W_{1,t}$ denotes a Brownian motion. In addition, the instantaneous variance v_t follows a Cox-Ingersoll-Ross (CIR) stochastic process:

$$dv_t = \kappa[\bar{v} - v_t]\, dt + \xi\sqrt{v_t}\, dW_{2,t} \tag{3.31}$$

where $\kappa[\bar{v} - v_t]\, dt$ is the drift component, incorporating a mean-reversion term, and $\xi\sqrt{v_t}\, dW_{2,t}$ is the diffusion component, with $W_{2,t}$ denoting a second Brownian motion. Applying the risk-neutral valuation framework, the price of a European call option is expressed as:

$$C_t = S_t Q_1 - K e^{-r(T-t)} Q_2 \tag{3.32}$$

$$Q_m = \frac{1}{2} + \frac{1}{\pi}\int_0^{+\infty} \Re\left(\frac{e^{-i\omega \ln(K)}\phi_m(\omega)}{i\omega}\right) d\omega, \qquad m \in \{1, 2\} \tag{3.33}$$

where $i = \sqrt{-1}$ denotes the imaginary unit and $\phi_m(\omega)$ is the characteristic function of the log-price under the risk-neutral measure, whose analytical form is derived by Heston (1993).

Taken together, these models can be viewed as competing approximations of the option price surface. The dependence of implied volatility on the strike price and the maturity gives rise to the IV surface, which maps each pair (K, T) to its market-implied volatility. Since option prices and implied volatilities are linked one-to-one through the Black-Scholes formula (3.28), approximating the option price surface amounts to approximating the IV surface (Gatheral, 2006). This is the perspective adopted in this study: each model, whether parametric or nonparametric, is assessed on its ability to reproduce the empirical surface of SPX options.

Pricing Bounds

The no-arbitrage condition inherent in parametric models ensures that any call option price generated by these models satisfies the following bounds (Yang et al., 2017):

$$\max\left(S_t - K e^{-r(T-t)}, 0\right) \leq C \leq S_t \tag{3.34}$$

The lower bound ensures that the option is at least equal to its intrinsic value; otherwise, one could buy the option, exercise it immediately and realise a risk-free profit. The upper bound limits the call price to a value that does not exceed the price of the underlying asset. Conversely, as the neural network is not constrained by a predefined functional structure, the prices it generates are not subject to any such restrictions.

Chapter 4

Methodology

4.1 Data

This analysis focuses on European options on the S&P 500 listed on the Chicago Board Options Exchange (CBOE). This index includes the largest publicly traded U.S. companies by market capitalisation and is updated throughout the year, with companies that no longer meet the required criteria being replaced by new ones. It is one of the most widely studied underlying assets in the option pricing literature, notably due to its liquidity and the depth of its options market. Data were obtained from IVolatility, a company specialising in the provision of data and analytics on the derivative market, except for the interest rate which was sourced from Bloomberg.

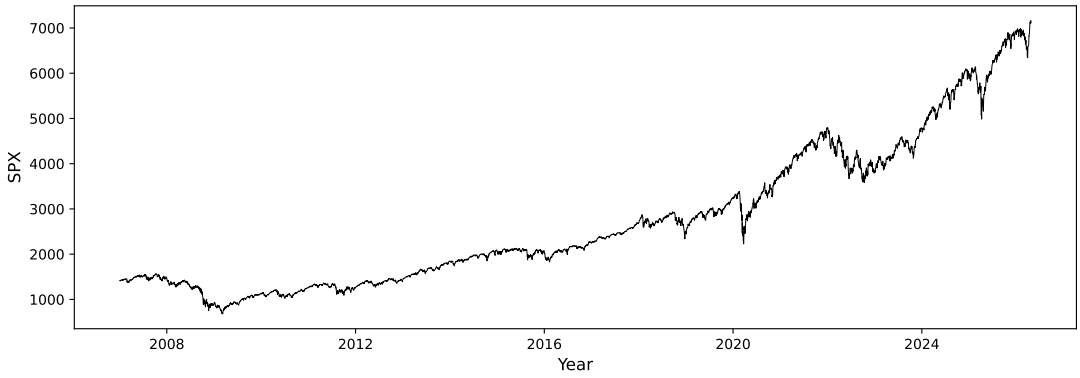


Figure 4.1: S&P 500 closing price from January 2007 to April 2026.

The period covers 3rd January 2007 to 24th April 2026 (see Figure 4.1). The initial dataset contained 47,054,172 observations, consisting of call and put options priced at the daily mid bid-ask spread. For each option, the dataset includes the underlying asset price S, the strike price K, the time to maturity $T - t$ and the volatility σ_{ivo} provided by the data platform. The volatility parameter σ used in the models is estimated directly from option prices using a method detailed below. The ratio of the underlying asset price to the strike price was also added; this is detailed in Section 4.4. Finally, the 3-month U.S. T-Bill is used as a proxy for the risk-free rate r, and dividends are assumed to be zero.

4.2 Data cleaning

Following Bakshi et al. (1997), several filters were applied to exclude observations that may bias the analysis. First, to simplify both the computational and analytical processes, the analysis is restricted to call options. Then, options expiring in less than 11 days were removed to avoid liquidity biases. Also, options priced below \$0.50 were excluded to eliminate non-informative prices. Additionally, only options satisfying the no-arbitrage condition are retained. Finally, options with the data provider's volatility σ_{ivo} greater than 0.8 are excluded too, as they may give rise to numerical instabilities in the calibration of the Heston model. After applying these filters, the dataset contains 14,292,393 observations, representing 30.4% of the initial sample.

4.3 Splitting set

The standard split for machine learning models is 60% of the dataset for training, 20% for validation and 20% for testing. In our case, given that the volume of options grows substantially over the sample period, such a split would not allow the most recent stress periods to be fully included in the test set. A chronological split was therefore adopted.

The period from 2007 to 2016 is used to train the nonparametric model, allowing it to be exposed to stressed market conditions from the training stage onward, with the subprime crisis, consistent with the evidence of Liu and Huang (2016). The validation period, from 2017 to 2019, is then used to select the optimal architecture and determine the optimal number of iterations, that is epochs, via early stopping. Finally, the test set covers 2020 to 2026, including three episodes of market stress: the COVID crisis in 2020, the 2022 monetary tightening episode, and the tariff-driven turbulence of April 2025.

To identify periods of market stress, a threshold is defined based on the Volatility Index (VIX). Developed by the CBOE, it measures the expected short-term volatility of SPX options and is widely seen as a gauge of market fear. This index was obtained from the Federal Reserve Economic Data platform. The threshold is set at the 75th percentile of the VIX distribution, corresponding to a level of 22.78. Any day on which the VIX exceeds this value is classified as a stress day (see Figure 4.2). Consecutive stress days are then grouped into episodes; where multiple episodes occur within the same calendar year, they are merged into a single regime spanning from the first to the last stress day of that year (see Table 4.1), since brief calm periods within a broader crisis do not represent a genuine return to normal conditions. This choice is arbitrary; Section 5.4.2 therefore tests the robustness of the main conclusions to alternative thresholds.

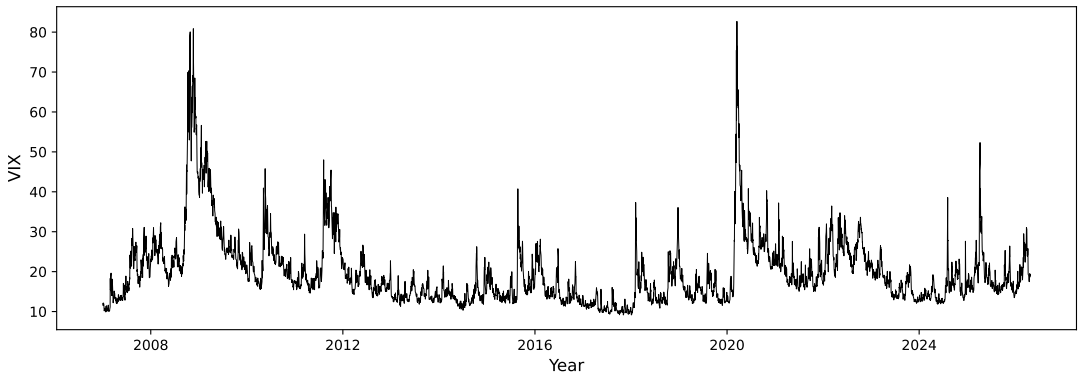


Figure 4.2: VIX daily closing values from January 2007 to April 2026.

Table 4.1: Market regimes defined by the VIX threshold

Period	Start	End	Trading days	Observations
Pre-COVID	2020-01-02	2020-02-21	35	89427
Stress COVID	2020-02-24	2020-11-13	186	936305
Post-COVID	2020-11-16	2022-01-14	294	1709261
Stress 2022	2022-01-18	2022-11-10	207	1083631
Normal 2	2022-11-11	2025-04-02	598	1606747
Stress 2025	2025-04-03	2025-05-01	20	105014
Normal 3[1]	2025-05-02	2026-04-24	239	991994

Note: Stress 2022 corresponds to the monetary tightening episode; Stress 2025 to the April 2025 tariff shock. Appendix E shows that this choice does not affect the results.

Throughout this study, the terms 'in-sample' and 'out-of-sample' denote the 2007-2016 and 2020-2026 windows respectively. This terminology refers to the training data of the neural network. Following Bakshi et al. (1997), parametric models are recalibrated daily over the entire study period; when applied to Black-Scholes and Heston, these terms therefore serve solely as time windows rather than as indicators of generalisation ability. Indeed, the computational cost of recalibrating multiple parameters for a parametric model is negligible, whereas retraining a neural network daily is far more expensive and potentially unstable, so that networks are usually trained once and deployed (Gu et al., 2020; Liu & Huang, 2016). Consequently, any performance gap between the parametric models and the network reflects both differences in functional flexibility and differences in the recency of the information. Section 5.4.1 analyses these two effects by re-pricing options at date t using Heston parameters calibrated at $t-1$.

[1] A short 14-day episode within the Normal 3 period exceeds the VIX threshold. However, it is not treated as a distinct regime given its low magnitude. Appendix E shows that this choice does not affect the results.

4.4 Homogeneity hint

As mentioned previously, we incorporate the homogeneity hint proposed by Garcia and Gençay (2000) by considering the underlying asset price S and the strike price K as a single variable: S/K. The relationship between the two features is proportional to the option price: if S and K are both multiplied by any scalar λ, the option price is multiplied by λ as well. Black-Scholes and Heston satisfy this economic property by construction. To ensure a fair comparison between the parametric models and the neural network, the latter is explicitly provided with this information through the new feature.

4.5 Moneyness segmentation

Also, options are classified into five segments based on the ratio S/K (Bakshi et al., 1997) (see Table 4.2):

Table 4.2: Moneyness segmentation

Segment	Bounds
Deep ITM	$S/K \geq 1.111$
ITM	$1.031 \leq S/K < 1.111$
ATM	$0.971 \leq S/K \leq 1.031$
OTM	$0.909 < S/K < 0.971$
Deep OTM	$S/K \leq 0.909$

4.6 Parametric models

Both models are calibrated daily over the entire study period, representing 4,850 trading days.

4.6.1 Black-Scholes

To obtain option prices under the Black-Scholes model, a single volatility σ_s is estimated for all options observed at date s, calibrated over a rolling window $[s-20, s]$. For each date s, through minimisation of the sum of squared errors between $C_{i,t}$ the market prices, and $C_{i,t}^{BS}$ (Bakshi et al., 1997), we obtain:

$$\hat{\sigma}_s = \arg\min_{\sigma} \sum_{t=s-20}^{s} \sum_{i=1}^{N_t} \left(C_{i,t} - C_{i,t}^{BS}(S_t, K_i, \tau_i, r_t, \sigma)\right)^2 \tag{4.1}$$

where $\tau_i = T_i - t$, N_t is the number of options observed at date t, and i indexes the observations.

This minimisation is computationally straightforward, as it involves performing a scalar search within a bounded range $[0.01, 2]$; only a single parameter needs to be optimised. This bound was established to ensure economic consistency and stability of the Black-Scholes formula. The estimated volatility $\hat{\sigma}_s$ is then used to compute option prices under the Black-Scholes formula (3.28), $C_{i,t}^{BS}$, for all options observed at date s. These generated prices constitute the output of the model.

4.6.2 Heston

The Heston model is more complex than the preceding one. Here, the calibration involves a multidimensional minimisation of five bounded parameters (see Table 4.3): $v_0, \kappa, \bar{v}, \xi, \rho$. For each date t, the sum of squared errors is minimised (Bakshi et al., 1997):

$$(\hat{v}_0, \hat{\kappa}, \hat{\bar{v}}, \hat{\xi}, \hat{\rho})_t = \arg \min_{v_0,\kappa,\bar{v},\xi,\rho} \sum_{i=1}^{N_t} \left(C_i - C_i^H(S_t, K_i, \tau_i, r_t, v_0, \kappa, \bar{v}, \xi, \rho)\right)^2 \tag{4.2}$$

where C_i^H is the call option price generated by Heston.

Table 4.3: Heston calibration bounds

Parameter	Description	Bounds
v_0	Initial variance	$[10^{-6}, 1.0]$
κ	Mean reversion speed	$[10^{-4}, 20]$
$\bar{v}$	Long-term variance	$[10^{-6}, 1.0]$
ξ	Volatility of volatility	$[10^{-4}, 4.0]$
ρ	Correlation	$[-0.999, -0.05]$

Preliminary calibrations conducted with the correlation parameter $\rho \in [-0.999, 0.999]$ systematically converged to values very close to the lower bound. Therefore, ρ is restricted to negative values to reflect the leverage effect documented by Black (1976). A decline in the underlying asset price S_t is often associated with an increase in volatility v_t, implying a negative correlation between shocks to the price dW_1 and shocks to the variance process dW_2. As shown by Heston (1993), this generates a negatively skewed distribution of asset returns which is consistent with the volatility smile observed empirically on the SPX (Rubinstein, 1994).

Since the minimisation (4.2) is a non-convex problem with multiple parameters, numerous local minima exist, creating a risk that the solution may not be globally optimal. To address this, a hybrid optimisation approach is adopted. First, the Differential Evolution algorithm explores the parameter space in search of a promising region near the global minimum. The L-BFGS-B algorithm then locally refines the obtained solution to converge to a precise minimum (Byrd et al., 1995; Storn & Price, 1997). Faster methods

exist for calibrating the Heston model, notably the one proposed by Cui et al. (2016), which relies on the analytical gradient of the objective function. The computational cost of this approach is low, but its implementation is significantly more complex. Thus, the hybrid procedure used yields a set of optimal parameters for each date t, which are subsequently used to compute option prices via the Heston pricing formula (3.32) through numerical integration.

4.7 Neural network

4.7.1 Architecture

We performed a method known as grid search to find the optimal architecture. It consists of training the network across multiple combinations of hyperparameters (see Table 4.4) and selecting the architecture that minimises the MSE on the validation set. Each architecture is trained using a regularisation[2] method to ensure a fair comparison of performance: early stopping.

Table 4.4: Grid search hyperparameters

Hyperparameter	Tested values	Selected value
Hidden layers H	$\{2, 3, 4, 5\}$	4
Neurons per layer K	$\{32, 64, 128, 256\}$	128
Learning rate η	$\{0.001, 0.0001\}$	0.001

Note: Total of 32 architectures tested in the grid search.

The selected architecture is a deep neural network with four hidden layers of 128 neurons each, trained with a learning rate of 0.001 (see Table 4.4).

4.7.2 Training phase

The inputs and target provided to the network (see Table 4.5) are normalised using Z-score standardisation based on the statistical properties of the training set, computed for each feature:

$$z_{ij} = \frac{x_{ij} - \mu_{j,train}}{\sigma_{j,train}} \tag{4.3}$$

This step prepares the features for optimisation by placing them all on the same scale, enabling faster and more stable convergence of gradient descent. Additionally, the weights are initialised according to the distribution proposed by Glorot and Bengio (2010), which assumes inputs centred around zero with constant variance.

[2]Regularisation refers to methods used to reduce overfitting.

Table 4.5: Neural network inputs and target

Variable	Description
Inputs	
S_t/K_i	Moneyness ratio
τ_i	Time-to-maturity
r_t	Risk-free rate
$\hat{\sigma}_s$	Black-Scholes volatility
Target	
C_i/K_i	Scaled call price

In accordance with the homogeneity hint of Garcia and Gençay (2000), the option price C and the underlying asset price S enter the model as ratios, divided by the strike price K before normalisation. Consequently, the absolute price level of the SPX has no bearing on the learning process of the neural network.

$$\frac{C_i}{K_i} = f\left(\frac{S_t}{K_i}, \tau_i, r_t, \hat{\sigma}_s\right) \tag{4.4}$$

The network is then trained by minimising the MSE (3.6) using the Adam optimiser (Kingma & Ba, 2015), with a mini-batch size of 4096, chosen to reduce computation time without significantly affecting model performance, given the large volume of observations.

As stated above, the regularisation method used is early stopping: training is interrupted when the validation loss J, i.e., the MSE computed on the validation set, ceases to improve, allowing the optimal number of epochs t^* to be determined. A patience p of 10 is set, meaning training stops after 10 consecutive epochs without improvement on the validation set, covering the period 2017-2019. The weights θ^* corresponding to the lowest validation loss are then restored. This method was chosen for its widespread use in deep learning and its very low computational cost (Goodfellow et al., 2016).

Let t denote the current epoch, we retain:

$$t^* = \arg\min_t J(t, \mathbf{X}_{valid}, \mathbf{Y}_{valid}) \tag{4.5}$$

Thus, the optimal weights are given by:

$$\theta^* = \theta_{t^*} \tag{4.6}$$

4.7.3 Test phase

Once the training is done, the optimal weights θ^* are fixed and applied to the test set to generate option price approximations at date t; the test data is not involved in any of the preceding phases. Performance metrics are then computed to compare the models and

assess the study's hypotheses.

As discussed in Section 3.2.3, the neural network does not guarantee compliance with the no-arbitrage bounds (3.34); violations may occur.

4.8 Calibration design

This volatility input of Black-Scholes is estimated over a rolling window of one month. This follows the principle adopted by Hutchinson et al. (1994), whose Black-Scholes benchmark receives a volatility estimated over a recent historical window. The three models consequently differ in their degree of adaptation: Heston refits five parameters on the day's cross-section, Black-Scholes estimates a single parameter over the preceding month, and the neural network keeps its weights fixed while receiving current inputs. The comparison thus contrasts approximation ability under each model's own calibration scheme; Heston's advantage is discussed in Section 5.4.1.

4.9 Implied volatility perimeter

We use the IV to evaluate the relative performance of the models as it allows us to isolate the main source of mispricing (Black & Scholes, 1972). For each model and for the market, IV are obtained by numerically inverting the Black-Scholes formula (3.28) against the prices generated by the concerned one; the IV of Black-Scholes is by construction the calibrated parameter $\hat{\sigma}_s$ itself (4.1), so the model requires no inversion. It is possible that the inversion fails to converge, particularly where the price is close to the intrinsic value or is an outlier. Thus, the analysis is restricted to observations in the perimeter for which the market, the Heston model and the neural network jointly converge: 9,813,894 observations, i.e., 68.7% of the filtered sample. This subset excludes cases where the neural network violates the no-arbitrage bounds (3.34).

4.10 Performance metrics

To evaluate and compare the quality of the models, three metrics were selected: the root mean square error (RMSE), the mean error (ME), and the coefficient of determination (R^2).

Let y_i denote the market IV, $\hat{y}_i$ the IV generated by the model, and $i = 1, \ldots, N$ the number of observations.

4.10.1 Root mean square error

RMSE measures the square root of the average of the squares of the differences between the market IV and the IV generated by the model:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{4.7}$$

This is one of the dominant metrics in the option pricing literature. It provides an interpretable measure of error magnitude, with the squaring operation placing greater emphasis on large deviations.

4.10.2 Mean error

ME calculates the average of the individual errors:

$$ME = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i) \tag{4.8}$$

This is a simple average with no absolute value. It complements the RMSE by providing an indication of the direction of the model's error: a positive ME indicates that the model overestimates the market IV, a negative ME indicates that the model underestimates the market IV. Mitra (2012) uses it to describe the systematic bias of the Black-Scholes model.

4.10.3 The coefficient of determination

R^2 measures the proportion of the variance in market IV that the model is able to explain:

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{4.9}$$

where TSS is the total sum of squares and RSS is the residual sum of squares, and $\bar{y}$ denotes the average market IV. R^2 can be negative if the model performs worse than $\bar{y}$.

Each metric covers an aspect not addressed by the others. RMSE captures the magnitude of errors. ME complements it by indicating the direction of the errors, therefore allowing systematic biases to be identified. R^2 adds to this by describing the model's overall ability to capture the structure of the IV. This combination of metrics enables the hypotheses formulated in this study to be assessed.

4.11 Model Confidence Set

A test is conducted to assess the stastistical validity of the results. To this end, the Model Confidence Set (MCS) developed by Hansen et al. (2011) is employed.

It identifies the set of models that cannot be statistically distinguished from the best-performing model or models, denoted $\hat{\mathcal{M}}^*_{1-\alpha}$. The approach is based on sequential elimination: at each iteration, the null hypothesis of equal performance is tested on the remaining models; if it is rejected, the worst-performing model is eliminated. The procedure stops when no statistically superior model can be identified within the remaining group. Each model receives an MCS p-value and a model is eliminated if its p-value falls below the threshold $\alpha = 0.05$, indicating statistical inferiority relative to the remaining models in $\hat{\mathcal{M}}^*_{1-\alpha}$.

The procedure is applied to daily loss series, defined for each model as the cross-sectional RMSE of IV errors per trading day. Inference relies on the stationary bootstrap of Politis and Romano (1994) with an average block length of 21 days, consistent with the Black-Scholes calibration window, and 5,000 bootstrap replications. The test is conducted on the out-of-sample period only, and implemented with the Python `arch` library.

As the procedure is performed multiple times, across pooled regimes as well as individual periods and moneyness segments, it increases the likelihood of false positives. Indeed, while the error rate is controlled within each procedure, a property established by Hansen et al. (2011), this control does not extend across configurations. Consequently, isolated deviations from a dominant pattern (see Section 5.3 and Appendix B) should be interpreted with caution.

Also, configurations with fewer than 30 trading days are excluded from testing. Therefore, no period-level test is conducted on the market stress of 2025, whose 20 trading days are insufficient for reliable block-bootstrap inference with a block size consistent with the rest of the analysis, its observations are nevertheless included when the stress periods are tested as a whole.

Chapter 5

Empirical Results and Analysis

5.1 Global results

Table 5.1: Performance metrics in-sample (2007–2016) and out-of-sample (2020–2026)

	In-sample			Out-of-sample		
Model	RMSE	ME	R^2	RMSE	ME	R^2
NN	0.0319	0.0018	0.8285	0.0529	0.0005	0.7448
BS	0.0549	-0.0100	0.4907	0.0800	-0.0286	0.4149
H	0.0241	-0.0001	0.9018	0.0220	0.0007	0.9556

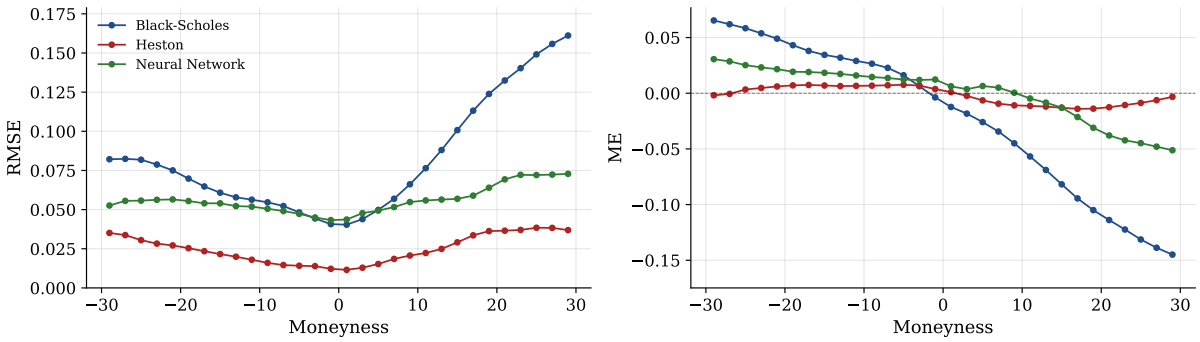


Figure 5.1: Out-of-sample RMSE (left) and ME (right) of implied volatility. Moneyness is expressed in percent: negative values correspond to OTM calls, positive values to ITM calls. Positive ME indicates an overestimation of market IV, negative values indicates underestimation.

Heston stands out as the best performing model over the entire period studied[1] (see Table 5.1). The neural network ranks second overall, followed by Black-Scholes.

During the testing period (Figure 5.1), Heston's RMSE remains the lowest at every level of moneyness, reaching its minimum near-the-money, while its bias stays close to zero, with a slight overestimation on OTM options and underestimation on ITM options. The Black-Scholes error is strongly asymmetric, moderate on OTM options but rising sharply on the ITM side, where it far exceeds the other two models. Its bias follows the same directional pattern as Heston at a much larger magnitude, declining monotonically

[1] Recall that the in-sample/out-of-sample terminology refers to the neural network's training data; for the parametric models, the two windows are simple time periods (see Section 4.3).

from a clear overestimation to a pronounced underestimation. The neural network occupies an intermediate position, with a flatter error profile across moneyness and the same directional bias at a magnitude between the two.

5.2 Results by market regime

5.2.1 Calm periods

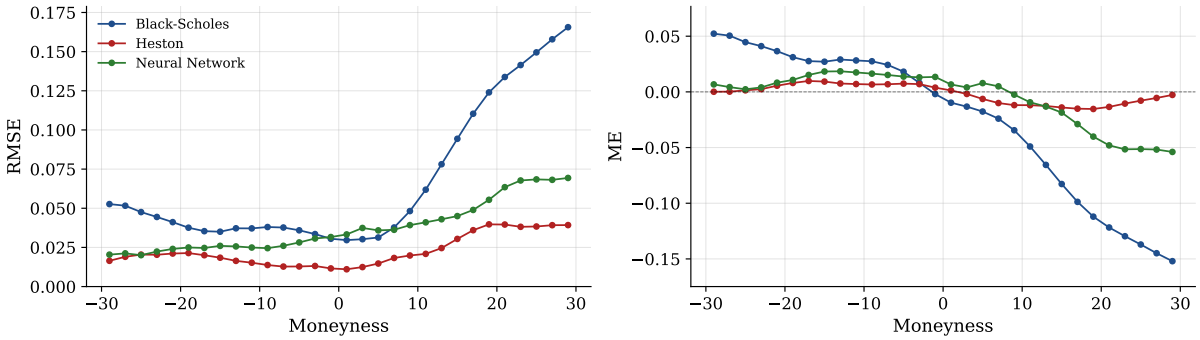


Figure 5.2: RMSE (left) and ME (right) of implied volatility in calm conditions. Conventions as in Figure 5.1.

Taken together, the four calm periods reproduce the global pattern of Section 5.1 (Figure 5.2): the ranking and the direction of the biases are unchanged. Two additional features emerge. Heston consistently reaches its highest R^2 on ATM options (Tables A.4–A.7 in Appendix A). The duel between the two other models is organised by moneyness: Black-Scholes prevails only around the money, deteriorating sharply towards deep ITM, while the neural network dominates over the rest of the range.

At the period level (Figure A.1 in Appendix A, where detailed metrics are also reported), a few configurations deviate from this picture. In Normal 2, Heston is outperformed by Black-Scholes on deep OTM options, an extreme segment where the performance gaps between the three models narrow. The shape of the Black-Scholes bias also changes there, overestimating at the extremes of moneyness and underestimating near-the-money. In the post-Covid period, the hierarchy between the two weakest models reverses: the neural network performs better across all moneyness segments, most visibly on deep ITM options with an error magnitude of 0.066 against 0.157 for Black-Scholes. In Normal 3, the network overestimates IV across all segments, while Heston's bias, though following its usual direction, reaches a magnitude comparable to that of Black-Scholes on the OTM segments. Finally, on deep ITM options in Normal 2 and deep OTM options in the pre-Covid period, the R^2 values are uninformative, as the cross-sectional variance of market IV within these segments is close to zero.

5.2.2 Stress periods

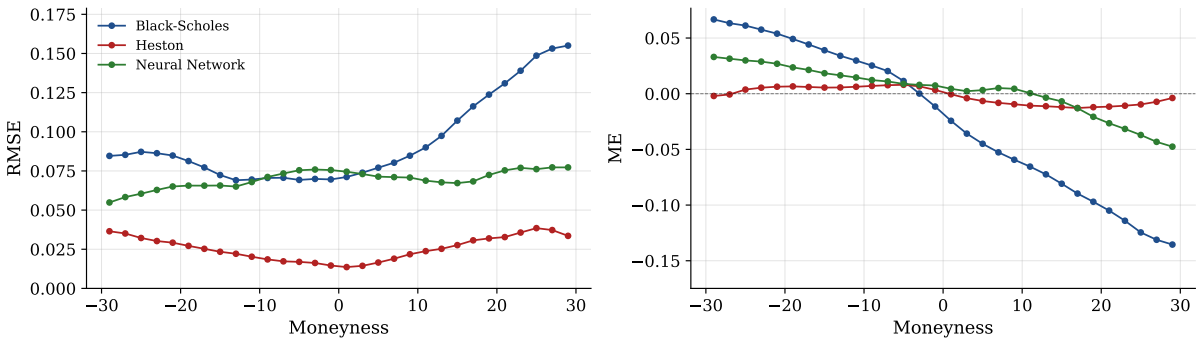


Figure 5.3: RMSE (left) and ME (right) of implied volatility in stressed conditions. Conventions as in Figure 5.1.

The three episodes of stress gathered together preserve the pattern of Section 5.1 while widening its gaps (Figure 5.3). Two differences from calm conditions stand out. Heston's explanatory power, still at its highest on ATM options, now remains positive on every segment. Concerning the two other models, the same structure as in calm conditions: Black-Scholes prevails near-the-money, the neural network everywhere else.

At the period level (Figure A.2, Appendix A), the Covid crisis and the 2022 stress display the same configuration as the aggregate results, with the network's advantage over Black-Scholes most pronounced on deep ITM options. The 2025 episode deviates from it on two points: the neural network outperforms Black-Scholes on all five segments, and the Black-Scholes bias loses the monotonic shape identified in Section 5.1, underestimating IV over the whole moneyness range, most markedly near-the-money and fading towards deep OTM. The results should nevertheless be treated with caution, given that the 2025 episode consists of only 20 trading days.

5.3 Statistical tests

Heston is retained in all ten configurations and is the sole survivor in nine of them (see Table 5.2). In the remaining one, deep OTM options in calm periods, Black-Scholes is retained alongside it. Since Heston dominates throughout, the test including the three models says little about the remaining two, additional tests are therefore conducted on Black-Scholes and the neural network alone.

Black-Scholes is statistically superior on the ITM and ATM segments in calm periods and on the ATM segment in stress periods, while it is the network at the extremes, on deep ITM options in both regimes and on deep OTM options in stress periods. In the remaining cases, both models are retained, meaning that the data cannot separate them at the 5% level.

Table 5.2: MCS results pooled by market regime ($\alpha = 5\%$). A checkmark indicates that the model is retained in the superior set $\hat{\mathcal{M}}^*_{1-\alpha}$.

	Three models			BS vs. NN	
Segment	BS	H	NN	BS	NN
Calm periods					
Deep ITM		✓			✓
ITM		✓		✓	
ATM		✓		✓	
OTM		✓		✓	✓
Deep OTM	✓	✓		✓	✓
Stress periods					
Deep ITM		✓			✓
ITM		✓		✓	✓
ATM		✓		✓	
OTM		✓		✓	✓
Deep OTM		✓			✓

At the period level, the three-model test confirms the dominance of Heston in nearly every configuration. The differences are limited to Normal 2, where Heston is excluded on deep OTM options, and Normal 3, where the test fails to separate it from Black-Scholes on three segments (Appendix B). In the duel, Black-Scholes dominates the network on most calm-period segments. The post-COVID period stands out, as the network dominates on every segment.

5.4 Robustness Analysis

5.4.1 Day-before test

Bakshi et al. (1997) explain that models with a larger number of structural parameters naturally achieve a better in-sample fit, making an out-of-sample test necessary to distinguish genuine structural performance from mere parameter flexibility.

Therefore, similarly to Bakshi et al. (1997) and Benzoni (2002), this test aims to observe the extent to which the model's performance degrades when using a parameter set calibrated at $t-1$ to price options at date t, a means of assessing the weight of the information freshness provided by daily calibration and of testing the robustness of the Heston model.

Using five parameter sets calibrated one day earlier to price options, Heston obtains a poorer fit to the IV surface (see Table D.1). Benzoni (2002) documents a comparable degradation. Nevertheless, although it is the only model that holds no information at date t, Heston$_{t-1}$ retains its advantage over Black-Scholes and the neural network. This

test does not allow recalibration and the model's stochastic structure to be isolated from one another, but by shifting the calibration by one day, it suggests that, at this scale, structure weighs more heavily than freshness.

A word of caution is warranted here. The relative performance of the models is assessed in the presence of Heston, which largely outperforms the other two models, as evidenced by this test as well as the statistical tests. Thus, the performance of Black-Scholes and the neural network may appear weak by comparison, when it could be judged sufficient if Heston were not included in the analysis.

5.4.2 VIX threshold sensitivity

To test the robustness of the 75th percentile, the results were recalculated using two other percentiles: the 70th and the 80th. The periods retain comparable durations when the VIX threshold varies, with the boundaries shifting only moderately. Moreover, the hierarchy among the models is preserved: the pattern described in the global results (Section 5.1) holds under both alternative thresholds. Therefore, the choice of the 75th percentile does not substantially influence the relative performance of the models (see Appendix E for more details).

Chapter 6

Discussion

6.1 Global patterns

6.1.1 Black-Scholes

Black-Scholes' assumption of constant volatility is its main structural limitation: it prevents the model from reproducing the volatility smile observed in the markets. This limitation is more pronounced at the extremes of moneyness given that it is where the smile is maximal.

Indeed, whatever the regime considered, the model struggles in this context, essentially in deep ITM where it generally underestimates IV. Rubinstein (1994) shows that deep OTM puts on the SPX, which are equivalent to deep ITM calls by put-call parity, carry structurally higher IV than those predicted by Black-Scholes, as investors value them as hedging instruments against extreme downside risk. Conversely, as deep OTM call options do not face comparable demand pressure, their IV deviates less from Black-Scholes' constant volatility. Nevertheless, results show that the model typically overestimates IV on this segment, consistent with the findings of Black (1975), MacBeth and Merville (1979) and Malliaris and Salchenberger (1993). Concerning near-the-money options, the error of the model generally decreases, explained by the volatility smile being at its lowest level.

Therefore, the results show that the parametric model exhibits a systematic pricing pattern reflecting its inability to reproduce the smirk inherent to the IV of SPX options.

6.1.2 Heston

By modelling volatility as a stochastic process, Heston effectively captures the volatility smile that Black-Scholes cannot reproduce (Bakshi et al., 1997). Yet the same authors also report that Heston remains misspecified and continues to exhibit systematic pricing errors. Our results are consistent with this observation: a slight overestimation of IV for OTM options and a slight underestimation for ITM options, also in accordance with the evidence reported by Fiorentini et al. (2002) and Benzoni (2002), who both demonstrate that the model's error is, to some extent, dependent on moneyness. Heston's mean error stays, however, close to zero across nearly all periods and segments, suggesting a relatively limited residual bias. In addition, the pooled MCS retains Heston as the sole survivor in nine of the ten (Table 5.2). The period-level results are similar, the only exclusion is

discussed in Section 6.4.

Parameter instability related to the daily calibration (Le Floc'h, 2025) does not prevent the model from efficiently capturing the IV surface. Gatheral (2006) explains that the main objective of a model and its parameters is not to perfectly represent reality but to reproduce market prices; as market conditions evolve continuously, so must the parameters.

The absence of the limit concerning the extreme path of volatility, as extreme as those observed during crises (Jiao et al., 2018), may be, indeed, explained by the daily recalibration of the initial variance parameter v_0 to the observed surface, which eliminates the need for the process to generate these high levels itself.

6.1.3 Neural network

Unlike the two parametric models, the neural network does not have a mathematical structure built to fit the IV surface of a financial asset so that no systematic error pattern should logically be identified. Although the model tends to make errors in the same direction within a given period, the direction varies from one episode to another.

The model generally makes smaller errors on the IV of deep OTM than on that of deep ITM options. This result should, however, be interpreted with caution, as it may reflect a selection effect rather than a true difference in performance. During numerical inversion, the network produces a negative or zero price for 22.5% of OTM observations (deep OTM included), against 1.3% for Heston. Liu and Huang (2016) explain that in the absence of imposed constraints, the nonparametric model is prone to generating negative prices in the OTM region. These prices subsequently cause the IV conversion to fail, so that the network's relative performance on this segment is measured only on the remaining 77.5% of observations, biasing comparisons in its favour.

A further concern is its inability to adapt to market regime shifts. Trained on 2007-2016 data, its weights remain fixed due to the computational cost of frequent retraining. Liu and Huang (2016) demonstrate that a network trained under one regime struggles to generalise to an unseen one. In our results, this risk appears to be softened by two factors. First, the presence of the subprime crisis within the training window may explain why the model does not collapse when confronted with the stress episodes of the test period. Second, the appreciation of the SPX index over the out-of-sample period does not pose a challenge due to the homogeneity hint of Garcia and Gençay (2000), as the absolute level of the index carries no information for the model.

6.2 Calm regimes

Across the four calm periods, Heston dominates both alternatives on every moneyness segment. Calm periods are precisely the environment where a constant-volatility model should be least penalised, as the surface is comparatively flat and evolves slowly. The persistence of a substantial gap between the two parametric models therefore indicates that the residual smirk, even at its lowest level, remains beyond the reach of a single volatility parameter, while five daily recalibrated parameters absorb it effectively.

The comparison between Black-Scholes and the network follows a different logic, organised by moneyness rather than by an overall ranking. Near-the-money, the parametric model prevails. Heston (1993) points out that this is precisely the segment where Black-Scholes is equivalent to a stochastic volatility model, so that the flexibility of the network offers nothing to exploit in this region, while still carrying its cost in estimation variance, in line with the bias-variance trade-off presented in Appendix C. The network still maintains low error there, likely because this is the most heavily represented segment: it has enough data to train effectively on it, consistent with the data representation argument of Malliaris and Salchenberger (1993) and Lajbcygier (2004). At the extremes, the relation reverses on deep ITM options, where the rigidity of a single volatility parameter is most penalising and the network is the sole model retained by the MCS. On the OTM segments, however, the data cannot discriminate between the two. These aggregate results nevertheless mask a strong heterogeneity across periods. In Normal 2 and Normal 3, Black-Scholes statistically dominates the network on every segment, and the network's advantage on deep ITM options rests on the post-COVID period alone (Appendix B). Hence, in genuinely quiet conditions, the flexibility of the network loses on every front, not only near-the-money.

The stability of market conditions also neutralises the main weakness of the Black-Scholes. When the surface moves slowly, the volatility carried by the rolling window remains representative of current conditions, so that the lag of its estimation is essentially inactive. The corresponding cost of this lag under rapidly changing conditions is examined in the next section.

6.3 Stress regimes

Stress episodes widen the gaps without altering the hierarchy. Heston remains the sole model retained by the MCS on nearly every segment, while the errors of its two competitors grow markedly. The mechanism behind this divergence lies in the calibration schemes. Through the daily recalibration of its initial variance v_0, Heston absorbs each day's level of volatility and only needs its remaining parameters to shape the surface around it. Black-Scholes, by contrast, carries a volatility averaged over the preceding

month. When the market shifts abruptly, this estimate lags behind current conditions, and the model's error then combines its structural inability to reproduce the smirk with a delayed volatility level. The 2025 episode isolates this second component. Over its 20 trading days of continuously rising volatility, Black-Scholes underestimates market IV even at-the-money, where its structural bias is close to zero in every other period; the deficit observed there seems attributable to the lag of the rolling window rather than to the constant-volatility assumption.

Under stress, the neural network inherits part of this lag as the Black-Scholes volatility is one of its input, which explains why its bias follows the same direction to that of the parametric model. Its errors nevertheless remain substantially smaller, indicating that the network partially compensates with its other inputs. Its position in the duel improves accordingly. The division observed in calm conditions sharpens, and the network becomes the sole model retained on the deep segments of both wings, where the surface deforms most under stress. This result contrasts with Gradojevic and Kukolj (2022), who find that, during stressed conditions, Black-Scholes cbecomes comparable to that of nonparametric models, including the network.

Heston's dominance under stress could reflect an informational advantage, as it is the only model calibrated on the day's surface. But the day-before test weakens this interpretation. As shown by the day-before (see Section 5.4.1), calibrating the model on the previous day's surface leaves its superiority over both competitors essentially intact, indicating that the advantage does not reduce to the freshness of its parameters, and reflects the adequacy of the stochastic volatility structure itself.

6.4 Special Cases

Three configurations depart from the general pattern established above, each in a limited and interpretable way.

The first concerns deep OTM options during the last two calm periods, where Heston's advantage narrows. In Normal 2, the MCS finds it significantly outperformed by Black-Scholes, the only configuration of the study where this occurs, and in Normal 3 the test does not manage to separate the two models. This localised weakening is consistent with the demand-pressure argument discussed in Section 6.1.1. The magnitude of the reversal is nevertheless minor. The RMSE gap between the two models is about 0.001, a difference that the length of the period renders statistically detectable without carrying much economic significance. When the calm periods are considered together, the test no longer separates the two parametric model (Table 5.2).

The second involves the post-COVID period, the only calm period in which the network statistically outperforms Black-Scholes, and it does so over all moneyness segments (Appendix B). Gradojevic and Kukolj (2022) note that the recovery phase following the

COVID crash degraded model accuracy more than the crash itself. This suggests the period behaves as a perturbed regime rather than a return to calm, despite falling below the VIX threshold. In this situation, the mechanism of Section 6.3 applies. The errors of Black-Scholes rise while those of the network remain stable, giving the latter its edge on every segment.

The third arises in the 2025 stress episode, during which Black-Scholes loses the monotonic error pattern it displays everywhere else and underestimates IV across the whole moneyness range. The mechanism is again that of Section 6.3, in an acute form, as most of the calibration window still consists of calm days over this short episode. The network, although it receives the same lagging input, partially compensates with its other inputs, and Heston shows no such bias, as its parameters are calibrated on the day itself. Given the low number of observations, this episode should nevertheless be read with caution.

6.5 Hypotheses

Hypothesis 1, which proposed, in line with the applied literature, a superior pricing performance of the nonparametric model under normal market conditions, is refuted. The network is outperformed not only by Heston, in every context, but also by Black-Scholes near-the-money, where the MCS retains only the parametric model. Its only clear advantage in calm periods is confined to deep ITM options, and rests on the post-COVID period alone.

Hypothesis 2 proposed a convergence of the three models' performance under stress. It is partially verified. Black-Scholes and the neural network do exhibit performance gaps that narrow across several segments, the MCS failing to separate the two models on the ITM and OTM configurations of the pooled stress regime. This convergence does not, however, operate through the anticipated mechanism: rather than a deterioration of the network, it is driven by a sharp increase in Black-Scholes error, while the network's performance remains comparatively stable. Moreover, the hypothesis is refuted once Heston enters the comparison, as the model diverges further from its two competitors under stress than in calm periods.

Hypothesis 3 proposed a superiority of the network over the parametric models at the extremes of moneyness, more pronounced than near-the-money. Against Black-Scholes, the pooled results support this assumption. The hypothesis nevertheless fails in calm periods on the OTM side, where the test cannot separate the two models, and is refuted altogether once Heston is considered, given that the network never outperforms it on any segment.

Overall, neither Black-Scholes nor the neural network systematically dominates the other, a conclusion in line with Malliaris and Salchenberger (1993) and İltüzer (2022). The dominant fact of the study lies elsewhere. The stochastic volatility model prevails

across regimes and moneyness, in calm conditions where its structural advantage should be smallest, and under stress despite the limits of its volatility process documented in the literature.

6.6 Limitations and directions for future research

First, the perimeter retained for the IV analysis introduces a selection bias. As discussed in Sections 4.9 and 6.1.3, the observations where the network fails to produce an invertible price are mechanically excluded from the comparison, biasing its measured performance on the OTM segments upwards.

Second, part of the results reflects the calibration design rather than the models alone. The three models deliberately differ in their degree of adaptation to the pricing date, as this is the object of the comparison. The length of the Black-Scholes calibration window, however, is a design choice. The 21-day estimate lags behind market conditions when volatility shifts abruptly, inflating the stress-period errors of Black-Scholes and, through its inputs, of the network. A single volatility fitted on the day's cross-section would remove this lag entirely and show how much of the Black-Scholes error is structural, which this study does not explore. Similarly, the network is trained once, its weights remaining fixed over the evaluation period due to the computational cost of frequent retraining.

In addition, the Heston results are, in turn, conditional on the quality of its calibration. As the accuracy of the estimated parameters depends on the choice of the optimiser (Sidhu & Saxena, 2026), the hybrid procedure adopted here places the model close to its best attainable fit, and a weaker optimisation scheme could narrow its advantage. The evidence of Milyushkov (2025), where a Heston model with fixed parameters from the empirical literature is outperformed by neural networks, illustrates this dependence even though the underlying assets differ.

Finally, the classification of market regimes relies on a VIX threshold computed ex post over the 2007-2026 sample, so that the regime labels incorporate information unavailable in real time. This choice belongs to a descriptive perspective, in which the objective is to evaluate model performance within past regimes. The classification also rests on a fixed threshold rather than on a formal regime-switching model, even though the robustness test in Section 5.4.2 indicates that the model hierarchy is insensitive to its exact position.

These limitations open several directions for future research. Black-Scholes could be calibrated on the day's cross-section, the network could receive this refreshed volatility as an input, and its weights could be retrained periodically, to measure how much of the documented gaps stems from the calibration design. The comparison could also be extended to put options and other underlyings. Beyond these extensions, two questions follow naturally from the present results: whether the hierarchy established in approximation carries over to forecasting, by evaluating the models in a genuinely predictive setting

across the same regimes; and whether providing the network with a stochastic volatility input would allow it to match, or surpass, Heston.

Chapter 7

Conclusion

This thesis compared the ability of three option pricing models, Black-Scholes, Heston and a neural network, to approximate the implied volatility surface of S&P 500 European call options. The comparison covers the period 2007-2026 and several million observations, across seven market regimes and five moneyness segments, with the performance rankings validated by the Model Confidence Set.

The central finding is the dominance of Heston. The model records the lowest errors in every regime and on nearly every segment. The analysis show that this advantage does not come from the freshness of its daily calibration alone. The stochastic volatility structure itself carries most of it.

The three hypotheses receive nuanced answers. Hypothesis 1 is refuted. The network does not dominate in normal conditions. It is outperformed by Heston everywhere and by Black-Scholes near-the-money, and against the latter, it loses on every segment in genuinely quiet periods. Hypothesis 2 is partially verified. Black-Scholes and the network do converge under stress, but through a rise of the Black-Scholes error rather than a deterioration of the network, and Heston moves further ahead instead of converging. Hypothesis 3 is partially verified as well. The network beats Black-Scholes at the extremes of moneyness, most clearly under stress, but it never outperforms Heston on any segment.

The main lesson is that nonparametric flexibility only pays where the parametric bias is large. Near-the-money, where a constant volatility is almost sufficient, Black-Scholes wins the duel. At the extremes and under stress, where this assumption fails most, the network takes over. Above both, a stochastic volatility model recalibrated daily captures the surface at low cost and remains the reference.

Appendix A

Detailed Results

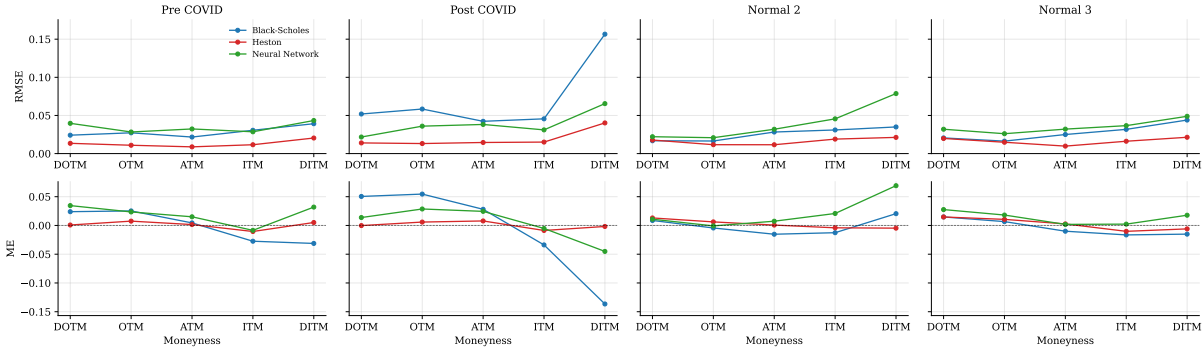


Figure A.1: RMSE and ME of implied volatility by moneyness segment across calm periods.

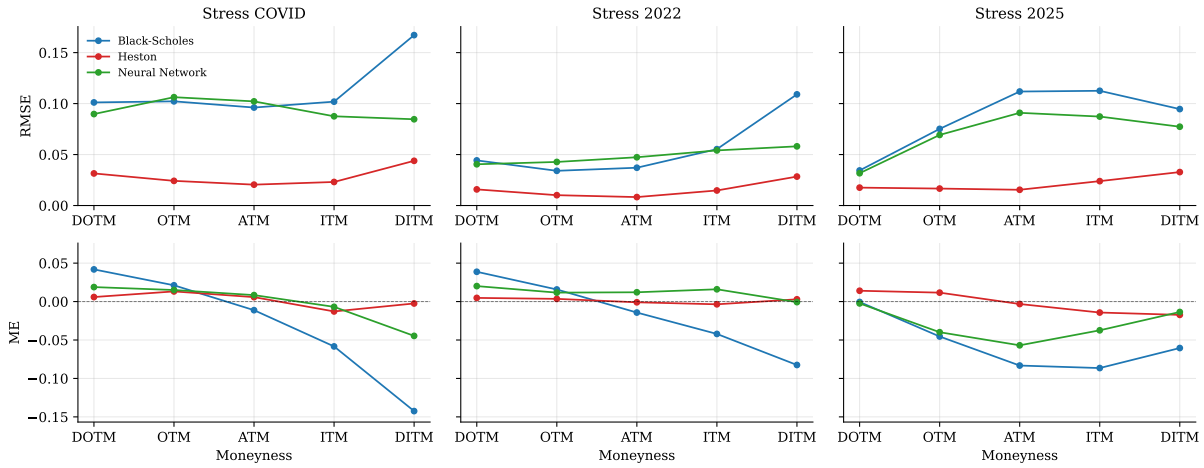


Figure A.2: RMSE and ME of implied volatility by moneyness segment across stressed periods.

Table A.1: Performance metrics by moneyness segment – Stress COVID

Segment	Model	RMSE	ME	R^2
Deep ITM	BS	0.1672	-0.1425	-2.1971
	H	0.0439	-0.0025	0.7800
	NN	0.0846	-0.0446	0.1807
ITM	BS	0.1019	-0.0583	-0.3093
	H	0.0231	-0.0128	0.9327
	NN	0.0875	-0.0069	0.0334
ATM	BS	0.0961	-0.0112	-0.0087
	H	0.0205	0.0058	0.9542
	NN	0.1022	0.0084	-0.1395
OTM	BS	0.1022	0.0212	-0.0492
	H	0.0242	0.0132	0.9412
	NN	0.1063	0.0150	-0.1344
Deep OTM	BS	0.1011	0.0419	-0.5166
	H	0.0315	0.0059	0.8530
	NN	0.0897	0.0188	-0.1938

Table A.2: Performance metrics by moneyness segment – Stress 2022

Segment	Model	RMSE	ME	R^2
Deep ITM	BS	0.1090	-0.0825	-1.2976
	H	0.0284	0.0029	0.8442
	NN	0.0581	-0.0007	0.3485
ITM	BS	0.0553	-0.0421	-1.1255
	H	0.0148	-0.0035	0.8485
	NN	0.0541	0.0160	-1.0317
ATM	BS	0.0371	-0.0143	-0.0823
	H	0.0083	-0.0009	0.9459
	NN	0.0474	0.0121	-0.7659
OTM	BS	0.0341	0.0157	-0.1432
	H	0.0102	0.0035	0.8984
	NN	0.0427	0.0117	-0.8003
Deep OTM	BS	0.0443	0.0388	-3.1947
	H	0.0158	0.0048	0.4636
	NN	0.0405	0.0201	-2.5073

Table A.3: Performance metrics by moneyness segment – Stress 2025

Segment	Model	RMSE	ME	R^2
Deep ITM	BS	0.0946	-0.0604	-1.0599
	H	0.0328	-0.0173	0.7523
	NN	0.0773	-0.0136	-0.3754
ITM	BS	0.1125	-0.0865	-2.0463
	H	0.0239	-0.0143	0.8623
	NN	0.0873	-0.0374	-0.8315
ATM	BS	0.1118	-0.0832	-1.8768
	H	0.0155	-0.0032	0.9451
	NN	0.0909	-0.0569	-0.9022
OTM	BS	0.0752	-0.0454	-1.1916
	H	0.0166	0.0116	0.8927
	NN	0.0692	-0.0399	-0.8588
Deep OTM	BS	0.0344	-0.0005	-0.8624
	H	0.0176	0.0141	0.5146
	NN	0.0317	-0.0024	-0.5816

Table A.4: Performance metrics by moneyness segment – Pre COVID

Segment	Model	RMSE	ME	R^2
Deep ITM	BS	0.0392	-0.0313	-1.7376
	H	0.0204	0.0051	0.2567
	NN	0.0434	0.0319	-2.3535
ITM	BS	0.0306	-0.0274	-3.9900
	H	0.0116	-0.0105	0.2796
	NN	0.0286	-0.0085	-3.3645
ATM	BS	0.0217	0.0044	-0.0457
	H	0.0088	0.0016	0.8264
	NN	0.0323	0.0151	-1.3202
OTM	BS	0.0272	0.0253	-6.1960
	H	0.0109	0.0075	-0.1599
	NN	0.0283	0.0235	-6.7833
Deep OTM	BS	0.0241	0.0239	-51.4652
	H	0.0135	0.0008	-15.3372
	NN	0.0397	0.0345	-140.8681

Table A.5: Performance metrics by moneyness segment – Post COVID

Segment	Model	RMSE	ME	R^2
Deep ITM	BS	0.1566	-0.1365	-2.9217
	H	0.0402	-0.0018	0.7415
	NN	0.0655	-0.0451	0.3137
ITM	BS	0.0456	-0.0338	-0.7448
	H	0.0151	-0.0087	0.8088
	NN	0.0311	-0.0052	0.1886
ATM	BS	0.0423	0.0281	-0.3579
	H	0.0146	0.0078	0.8393
	NN	0.0382	0.0245	-0.1064
OTM	BS	0.0584	0.0545	-4.2092
	H	0.0132	0.0059	0.7354
	NN	0.0360	0.0286	-0.9747
Deep OTM	BS	0.0519	0.0504	-8.0187
	H	0.0140	-0.0002	0.3446
	NN	0.0217	0.0138	-0.5806

Table A.6: Performance metrics by moneyness segment – Normal 2

Segment	Model	RMSE	ME	R^2
Deep ITM	BS	0.0349	0.0205	-1.5830
	H	0.0212	-0.0047	0.0445
	NN	0.0787	0.0691	-12.1074
ITM	BS	0.0310	-0.0126	0.3494
	H	0.0190	-0.0041	0.7556
	NN	0.0456	0.0208	-0.4062
ATM	BS	0.0283	-0.0153	0.3674
	H	0.0116	0.0006	0.8934
	NN	0.0319	0.0074	0.1936
OTM	BS	0.0165	-0.0044	0.6835
	H	0.0117	0.0062	0.8417
	NN	0.0208	-0.0008	0.4955
Deep OTM	BS	0.0168	0.0087	0.5836
	H	0.0178	0.0130	0.5346
	NN	0.0221	0.0108	0.2793

Table A.7: Performance metrics by moneyness segment – Normal 3

Segment	Model	RMSE	ME	R^2
Deep ITM	BS	0.0440	-0.0151	-0.0926
	H	0.0214	-0.0060	0.7412
	NN	0.0489	0.0178	-0.3512
ITM	BS	0.0318	-0.0165	0.1607
	H	0.0162	-0.0102	0.7819
	NN	0.0366	0.0023	-0.1167
ATM	BS	0.0250	-0.0100	0.2810
	H	0.0098	0.0027	0.8894
	NN	0.0321	0.0018	-0.1837
OTM	BS	0.0163	0.0067	0.2685
	H	0.0149	0.0106	0.3946
	NN	0.0261	0.0181	-0.8610
Deep OTM	BS	0.0206	0.0148	-2.4280
	H	0.0198	0.0147	-2.1929
	NN	0.0319	0.0276	-7.2819

Appendix B

MCS Results

Table B.1: Model Confidence Set results by period and moneyness segment ($\alpha = 5\%$)

		Three-model comparison			BS vs. NN duel	
Period	Segment	NN	BS	H	NN	BS
Pre COVID	Deep ITM			✓		✓
	ITM			✓	✓	✓
	ATM			✓		✓
	OTM			✓	✓	✓
	Deep OTM			✓		✓
Stress COVID	Deep ITM			✓	✓	
	ITM			✓	✓	✓
	ATM			✓		✓
	OTM			✓	✓	✓
	Deep OTM			✓	✓	
Post COVID	Deep ITM			✓	✓	
	ITM			✓	✓	
	ATM			✓	✓	
	OTM			✓	✓	
	Deep OTM			✓	✓	
Stress 2022	Deep ITM			✓	✓	✓
	ITM			✓	✓	✓
	ATM			✓	✓	✓
	OTM			✓		✓
	Deep OTM			✓	✓	✓
Normal 2	Deep ITM			✓		✓
	ITM			✓		✓
	ATM			✓		✓
	OTM			✓		✓
	Deep OTM		✓			✓
Normal 3	Deep ITM			✓		✓
	ITM		✓	✓		✓
	ATM			✓		✓
	OTM		✓	✓		✓
	Deep OTM		✓	✓		✓

B.1 Three models

Table B.2: MCS p-values by period (three models)

	NN	BS	H
Pre COVID	0.0000	0.0000	1.0000
Stress COVID	0.0018	0.0000	1.0000
Post COVID	0.0000	0.0000	1.0000
Stress 2022	0.0010	0.0010	1.0000
Normal 2	0.0000	0.0000	1.0000
Normal 3	0.0000	0.0192	1.0000

Table B.3: MCS p-values by moneyness (three models)

	BS	NN	H
Deep ITM	0.0000	0.0000	1.0000
ITM	0.0000	0.0000	1.0000
ATM	0.0000	0.0000	1.0000
OTM	0.0000	0.0000	1.0000
Deep OTM	0.0008	0.0008	1.0000

Table B.4: MCS p-values by segment – Pre COVID (three models)

	NN	BS	H
Deep ITM	0.0000	0.0000	1.0000
ITM	0.0000	0.0000	1.0000
ATM	0.0000	0.0000	1.0000
OTM	0.0000	0.0000	1.0000
Deep OTM	0.0000	0.0000	1.0000

Table B.5: MCS p-values by segment – Stress COVID (three models)

	NN	BS	H
Deep ITM	0.0090	0.0000	1.0000
ITM	0.0130	0.0130	1.0000
ATM	0.0000	0.0130	1.0000
OTM	0.0000	0.0000	1.0000
Deep OTM	0.0000	0.0000	1.0000

Table B.6: MCS p-values by segment – Post COVID (three models)

	NN	BS	H
Deep ITM	0.0000	0.0000	1.0000
ITM	0.0000	0.0000	1.0000
ATM	0.0000	0.0000	1.0000
OTM	0.0000	0.0000	1.0000
Deep OTM	0.0000	0.0000	1.0000

Table B.7: MCS p-values by segment – Stress 2022 (three models)

	NN	BS	H
Deep ITM	0.0132	0.0132	1.0000
ITM	0.0018	0.0018	1.0000
ATM	0.0000	0.0000	1.0000
OTM	0.0000	0.0000	1.0000
Deep OTM	0.0000	0.0000	1.0000

Table B.8: MCS p-values by segment – Normal 2 (three models)

	NN	BS	H
Deep ITM	0.0000	0.0000	1.0000
ITM	0.0000	0.0000	1.0000
ATM	0.0000	0.0000	1.0000
OTM	0.0000	0.0014	1.0000
Deep OTM	0.0000	1.0000	0.0050

Table B.9: MCS p-values by segment – Normal 3 (three models)

	NN	BS	H
Deep ITM	0.0000	0.0026	1.0000
ITM	0.0000	0.0560	1.0000
ATM	0.0000	0.0026	1.0000
OTM	0.0000	1.0000	0.7350
Deep OTM	0.0000	1.0000	0.1116

B.2 Two models: Neural Network and Black-Scholes

Table B.10: MCS p-values by period (NN vs BS)

	NN	BS
Pre COVID	0.0000	1.0000
Stress COVID	1.0000	0.0000
Post COVID	1.0000	0.0000
Stress 2022	1.0000	0.2746
Normal 2	0.0000	1.0000
Normal 3	0.0000	1.0000

Table B.11: MCS p-values by moneyness (NN vs BS)

	BS	NN
Deep ITM	0.0000	1.0000
ITM	1.0000	0.4942
ATM	1.0000	0.0000
OTM	1.0000	0.5570
Deep OTM	0.5296	1.0000

Table B.12: MCS p-values by segment – Pre COVID (NN vs BS)

	NN	BS
Deep ITM	0.0076	1.0000
ITM	1.0000	0.1288
ATM	0.0000	1.0000
OTM	0.0628	1.0000
Deep OTM	0.0000	1.0000

Table B.13: MCS p-values by segment – Stress COVID (NN vs BS)

	NN	BS
Deep ITM	1.0000	0.0000
ITM	1.0000	0.0554
ATM	0.0028	1.0000
OTM	0.8978	1.0000
Deep OTM	1.0000	0.0000

Table B.14: MCS p-values by segment – Post COVID (NN vs BS)

	NN	BS
Deep ITM	1.0000	0.0000
ITM	1.0000	0.0000
ATM	1.0000	0.0000
OTM	1.0000	0.0000
Deep OTM	1.0000	0.0000

Table B.15: MCS p-values by segment – Stress 2022 (NN vs BS)

	NN	BS
Deep ITM	1.0000	0.1232
ITM	1.0000	0.6314
ATM	0.1416	1.0000
OTM	0.0004	1.0000
Deep OTM	1.0000	0.2600

Table B.16: MCS p-values by segment – Normal 2 (NN vs BS)

	NN	BS
Deep ITM	0.0000	1.0000
ITM	0.0000	1.0000
ATM	0.0002	1.0000
OTM	0.0000	1.0000
Deep OTM	0.0000	1.0000

Table B.17: MCS p-values by segment – Normal 3 (NN vs BS)

	NN	BS
Deep ITM	0.0002	1.0000
ITM	0.0002	1.0000
ATM	0.0000	1.0000
OTM	0.0000	1.0000
Deep OTM	0.0000	1.0000

Appendix C

Bias-variance decomposition

This appendix formalises the bias-variance trade-off governing the choice of the function class introduced in Section 3.1.1.

Nonparametric model complexity introduces a fundamental trade-off. Minimising in-sample error does not guarantee low out-of-sample error: an overly complex model memorises the training data without capturing the true underlying relationship, while an overly simple model fails to approximate it sufficiently. This trade-off is managed by controlling the model's capacity, that is, the richness of the function class it can represent (Goodfellow et al., 2016).

The estimator $\hat{\boldsymbol{\theta}}$ is characterised by a bias $B(\hat{\boldsymbol{\theta}})$ and a variance $V(\hat{\boldsymbol{\theta}})$. An increase in model capacity tends to reduce bias while increasing variance, and vice versa. These two terms constitute the components of the mean squared error (MSE), known as the bias-variance decomposition (Hastie et al., 2009):

$$MSE = \mathbb{E}\Big[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^2\Big] = B(\hat{\boldsymbol{\theta}})^2 + V(\hat{\boldsymbol{\theta}}) \tag{C.1}$$

where $V(\hat{\boldsymbol{\theta}}) = \mathbb{E}\Big[(\hat{\boldsymbol{\theta}} - \mathbb{E}(\hat{\boldsymbol{\theta}}))^2\Big]$ and $B(\hat{\boldsymbol{\theta}}) = \mathbb{E}(\hat{\boldsymbol{\theta}}) - \boldsymbol{\theta}$.

A trade-off is therefore necessary to strike the right balance between bias and variance, to control the model's generalisation ability. Note that the MSE here refers to the estimation error of $\hat{\theta}$, distinct from the training loss of the neural network (see Section 4.7.2).

Appendix D

Day-before test

Table D.1: Day-before test results

	In-sample			Out-of-sample		
Model	RMSE	ME	R^2	RMSE	ME	R^2
NN	0.0317	0.0017	0.8260	0.0529	0.0003	0.7449
BS	0.0535	-0.0099	0.5050	0.0801	-0.0289	0.4138
H	0.0227	0.0001	0.9111	0.0219	0.0008	0.9563
H_{t-1}	0.0271	-0.0002	0.8728	0.0276	0.0009	0.9305

Note: The metrics differ from those in Table 5.1 because the test is based on the common perimeter of the four models (9,755,486 observations, i.e., 99.4% of the full perimeter).

Appendix E

VIX threshold sensitivity

This appendix assesses the robustness of the market regime classification to the choice of the VIX threshold. The 75th percentile is the main specification used throughout this study; the 70th and 80th percentiles are the alternative thresholds used to test the robustness of the conclusions. Table E.1 reports the regime boundaries obtained under each threshold, and the subsequent tables report the corresponding performance metrics.

Across the three thresholds, the performance metrics vary only marginally, typically at the third decimal place, and the ranking of the models is unchanged in nearly every period and segment. The configurations deviating from the general pattern in the main analysis, notably deep OTM options in Normal 2, persist under all three thresholds. The only visible sensitivity concerns Stress 2025, whose duration ranges from 17 to 49 trading days depending on the threshold, so that its metric levels shift while the ranking remains the same.

Table E.1: Market regimes defined by the alternative VIX thresholds

Period	Start	End	Trading days	Observations
Q70				
Pre COVID	2020-01-02	2020-02-21	35	89427
Stress COVID	2020-02-24	2021-02-04	241	1220449
Post COVID	2021-02-05	2022-01-14	239	1425117
Stress 2022	2022-01-18	2022-11-21	214	1108752
Normal 2	2022-11-22	2025-02-28	568	1485781
Stress 2025	2025-03-03	2025-05-09	49	229215
Normal 3	2025-05-12	2026-04-24	233	963638
Q75				
Pre COVID	2020-01-02	2020-02-21	35	89427
Stress COVID	2020-02-24	2020-11-13	186	936305
Post COVID	2020-11-16	2022-01-14	294	1709261
Stress 2022	2022-01-18	2022-11-10	207	1083631
Normal 2	2022-11-11	2025-04-02	598	1606747
Stress 2025	2025-04-03	2025-05-01	20	105014
Normal 3	2025-05-02	2026-04-24	239	991994
Q80				
Pre COVID	2020-01-02	2020-02-21	35	89427
Stress COVID	2020-02-24	2020-11-10	183	920386
Post COVID	2020-11-11	2022-02-16	319	1868825
Stress 2022	2022-02-17	2022-11-04	181	925152
Normal 2	2022-11-07	2025-04-02	602	1621581
Stress 2025	2025-04-03	2025-04-28	17	89948
Normal 3	2025-04-29	2026-04-24	242	1007060

Table E.2: VIX threshold sensitivity – Pre COVID

		Q70		Q75		Q80	
Segment	Model	RMSE	R^2	RMSE	R^2	RMSE	R^2
	BS	0.0392	-1.7376	0.0392	-1.7376	0.0392	-1.7376
Deep ITM	H	0.0204	0.2567	0.0204	0.2567	0.0204	0.2567
	NN	0.0434	-2.3535	0.0434	-2.3535	0.0434	-2.3535
	BS	0.0306	-3.9900	0.0306	-3.9900	0.0306	-3.9900
ITM	H	0.0116	0.2796	0.0116	0.2796	0.0116	0.2796
	NN	0.0286	-3.3645	0.0286	-3.3645	0.0286	-3.3645
	BS	0.0217	-0.0457	0.0217	-0.0457	0.0217	-0.0457
ATM	H	0.0088	0.8264	0.0088	0.8264	0.0088	0.8264
	NN	0.0323	-1.3202	0.0323	-1.3202	0.0323	-1.3202
	BS	0.0272	-6.1960	0.0272	-6.1960	0.0272	-6.1960
OTM	H	0.0109	-0.1599	0.0109	-0.1599	0.0109	-0.1599
	NN	0.0283	-6.7833	0.0283	-6.7833	0.0283	-6.7833
	BS	0.0241	-51.4652	0.0241	-51.4652	0.0241	-51.4652
Deep OTM	H	0.0135	-15.3372	0.0135	-15.3372	0.0135	-15.3372
	NN	0.0397	-140.8681	0.0397	-140.8681	0.0397	-140.8681

Table E.3: VIX threshold sensitivity – Stress COVID

		Q70		Q75		Q80	
Segment	Model	RMSE	R^2	RMSE	R^2	RMSE	R^2
	BS	0.1646	-2.1154	0.1672	-2.1971	0.1674	-2.2051
Deep ITM	H	0.0432	0.7859	0.0439	0.7800	0.0438	0.7802
	NN	0.0807	0.2505	0.0846	0.1807	0.0849	0.1750
	BS	0.0908	-0.1523	0.1019	-0.3093	0.1027	-0.3192
ITM	H	0.0213	0.9365	0.0231	0.9327	0.0231	0.9331
	NN	0.0773	0.1656	0.0875	0.0334	0.0882	0.0268
	BS	0.0848	0.1272	0.0961	-0.0087	0.0968	-0.0145
ATM	H	0.0196	0.9534	0.0205	0.9542	0.0206	0.9542
	NN	0.0901	0.0159	0.1022	-0.1395	0.1027	-0.1440
	BS	0.0937	0.0140	0.1022	-0.0492	0.1026	-0.0483
OTM	H	0.0226	0.9425	0.0242	0.9412	0.0243	0.9410
	NN	0.0944	-0.0008	0.1063	-0.1344	0.1069	-0.1386
	BS	0.0975	-0.4380	0.1011	-0.5166	0.1013	-0.5174
Deep OTM	H	0.0304	0.8601	0.0315	0.8530	0.0316	0.8526
	NN	0.0853	-0.1024	0.0897	-0.1938	0.0900	-0.1972

Table E.4: VIX threshold sensitivity – Post COVID

		Q70		Q75		Q80	
Segment	Model	RMSE	R^2	RMSE	R^2	RMSE	R^2
	BS	0.1566	-2.9789	0.1566	-2.9217	0.1565	-2.9529
Deep ITM	H	0.0401	0.7398	0.0402	0.7415	0.0398	0.7437
	NN	0.0652	0.3110	0.0655	0.3137	0.0657	0.3025
	BS	0.0454	-0.8944	0.0456	-0.7448	0.0479	-0.7742
ITM	H	0.0151	0.7911	0.0151	0.8088	0.0153	0.8189
	NN	0.0305	0.1439	0.0311	0.1886	0.0327	0.1716
	BS	0.0429	-0.5064	0.0423	-0.3579	0.0421	-0.1945
ATM	H	0.0140	0.8383	0.0146	0.8393	0.0143	0.8623
	NN	0.0376	-0.1580	0.0382	-0.1064	0.0383	0.0100
	BS	0.0584	-5.0754	0.0584	-4.2092	0.0571	-3.3296
OTM	H	0.0124	0.7260	0.0132	0.7354	0.0129	0.7794
	NN	0.0360	-1.3144	0.0360	-0.9747	0.0357	-0.6908
	BS	0.0511	-10.3730	0.0519	-8.0187	0.0511	-6.6708
Deep OTM	H	0.0130	0.2649	0.0140	0.3446	0.0134	0.4716
	NN	0.0212	-0.9643	0.0217	-0.5806	0.0240	-0.6994

Table E.5: VIX threshold sensitivity – Stress 2022

		Q70		Q75		Q80	
Segment	Model	RMSE	R^2	RMSE	R^2	RMSE	R^2
	BS	0.1087	-1.2426	0.1090	-1.2976	0.0976	-0.9820
Deep ITM	H	0.0283	0.8475	0.0284	0.8442	0.0276	0.8419
	NN	0.0584	0.3530	0.0581	0.3485	0.0552	0.3650
	BS	0.0545	-0.9217	0.0553	-1.1255	0.0532	-0.9531
ITM	H	0.0152	0.8501	0.0148	0.8485	0.0142	0.8601
	NN	0.0553	-0.9760	0.0541	-1.0317	0.0539	-1.0055
	BS	0.0365	-0.0425	0.0371	-0.0823	0.0373	-0.0641
ATM	H	0.0085	0.9431	0.0083	0.9459	0.0080	0.9507
	NN	0.0472	-0.7445	0.0474	-0.7659	0.0489	-0.8325
	BS	0.0338	-0.1419	0.0341	-0.1432	0.0335	-0.1230
OTM	H	0.0102	0.8971	0.0102	0.8984	0.0103	0.8945
	NN	0.0424	-0.7938	0.0427	-0.8003	0.0446	-0.9836
	BS	0.0441	-3.2062	0.0443	-3.1947	0.0440	-3.3496
Deep OTM	H	0.0158	0.4594	0.0158	0.4636	0.0166	0.3793
	NN	0.0403	-2.5043	0.0405	-2.5073	0.0420	-2.9639

Table E.6: VIX threshold sensitivity – Normal 2

		Q70		Q75		Q80	
Segment	Model	RMSE	R^2	RMSE	R^2	RMSE	R^2
	BS	0.0330	-1.4317	0.0349	-1.5830	0.0355	-1.4228
Deep ITM	H	0.0216	-0.0441	0.0212	0.0445	0.0212	0.1381
	NN	0.0755	-11.7224	0.0787	-12.1074	0.0817	-11.8429
	BS	0.0304	0.3783	0.0310	0.3494	0.0308	0.3874
ITM	H	0.0186	0.7670	0.0190	0.7556	0.0190	0.7658
	NN	0.0438	-0.2910	0.0456	-0.4062	0.0465	-0.3946
	BS	0.0262	0.3910	0.0283	0.3674	0.0282	0.3909
ATM	H	0.0111	0.8900	0.0116	0.8934	0.0116	0.8974
	NN	0.0305	0.1730	0.0319	0.1936	0.0319	0.2220
	BS	0.0151	0.7219	0.0165	0.6835	0.0165	0.7080
OTM	H	0.0114	0.8415	0.0117	0.8417	0.0116	0.8543
	NN	0.0201	0.5033	0.0208	0.4955	0.0210	0.5255
	BS	0.0161	0.6229	0.0168	0.5836	0.0170	0.6201
Deep OTM	H	0.0178	0.5420	0.0178	0.5346	0.0176	0.5930
	NN	0.0216	0.3248	0.0221	0.2793	0.0223	0.3493

Table E.7: VIX threshold sensitivity – Stress 2025

		Q70		Q75		Q80	
Segment	Model	RMSE	R^2	RMSE	R^2	RMSE	R^2
	BS	0.0877	-0.7361	0.0946	-1.0599	0.0974	-1.1257
Deep ITM	H	0.0312	0.7803	0.0328	0.7523	0.0311	0.7830
	NN	0.0778	-0.3671	0.0773	-0.3754	0.0779	-0.3583
	BS	0.0842	-0.5671	0.1125	-2.0463	0.1231	-2.3906
ITM	H	0.0218	0.8944	0.0239	0.8623	0.0242	0.8692
	NN	0.0701	-0.0860	0.0873	-0.8315	0.0946	-1.0026
	BS	0.0801	-0.8558	0.1118	-1.8768	0.1209	-2.2256
ATM	H	0.0167	0.9190	0.0155	0.9451	0.0160	0.9436
	NN	0.0677	-0.3231	0.0909	-0.9022	0.0984	-1.1341
	BS	0.0537	-0.1990	0.0752	-1.1916	0.0795	-1.4408
OTM	H	0.0156	0.8994	0.0166	0.8927	0.0156	0.9063
	NN	0.0500	-0.0402	0.0692	-0.8588	0.0731	-1.0673
	BS	0.0312	-0.1924	0.0344	-0.8624	0.0329	-0.6929
Deep OTM	H	0.0175	0.6272	0.0176	0.5146	0.0159	0.6017
	NN	0.0306	-0.1432	0.0317	-0.5816	0.0312	-0.5274

Table E.8: VIX threshold sensitivity – Normal 3

		Q70		Q75		Q80	
Segment	Model	RMSE	R^2	RMSE	R^2	RMSE	R^2
Deep ITM	BS	0.0444	-0.0704	0.0440	-0.0926	0.0436	-0.0870
	H	0.0214	0.7525	0.0214	0.7412	0.0229	0.7000
	NN	0.0439	-0.0481	0.0489	-0.3512	0.0496	-0.4062
ITM	BS	0.0320	0.1131	0.0318	0.1607	0.0320	0.1980
	H	0.0161	0.7758	0.0162	0.7819	0.0164	0.7902
	NN	0.0357	-0.1037	0.0366	-0.1167	0.0368	-0.0571
ATM	BS	0.0252	0.2490	0.0250	0.2810	0.0250	0.3039
	H	0.0098	0.8861	0.0098	0.8894	0.0098	0.8923
	NN	0.0321	-0.2247	0.0321	-0.1837	0.0320	-0.1389
OTM	BS	0.0158	0.2910	0.0163	0.2685	0.0165	0.2779
	H	0.0149	0.3704	0.0149	0.3946	0.0150	0.4029
	NN	0.0260	-0.9175	0.0261	-0.8610	0.0260	-0.8012
Deep OTM	BS	0.0183	-1.9090	0.0206	-2.4280	0.0216	-2.5793
	H	0.0201	-2.4964	0.0198	-2.1929	0.0201	-2.0959
	NN	0.0317	-7.6840	0.0319	-7.2819	0.0320	-6.8921

Bibliography

Amilon, H. (2003). *A neural network versus black-scholes: A comparison of pricing and hedging performances* (Working Paper). Lund University.

Bakshi, G., Cao, C., & Chen, Z. (1997). Empirical performance of alternative option pricing models. *The Journal of Finance*, *52*(5), 2003–2049.

Bank for International Settlements. (2025). Otc derivatives statistics at end-june 2025.

Bates, D. S. (2003). Empirical option pricing: A retrospection. *Journal of Econometrics*, *116*(1–2), 387–404.

Benzoni, L. (2002). *Pricing options under stochastic volatility: An empirical investigation* (Working Paper). Carlson School of Management, University of Minnesota.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Black, F. (1975). Fact and fantasy in the use of options. *Financial Analysts Journal*, *31*(4), 36–41, 61–72.

Black, F. (1976). Studies of stock price volatility changes. *Proceedings of the 1976 Meetings of the American Statistical Association, Business and Economics Statistics Section*, 177–181.

Black, F., & Scholes, M. (1972). The valuation of option contracts and a test of market efficiency. *The Journal of Finance*, *27*(2), 399–417.

Black, F., & Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy*, *81*(3), 637–654.

Byrd, R. H., Lu, P., Nocedal, J., & Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, *16*(5), 1190–1208.

Cui, Y., del Baño Rollin, S., & Germano, G. (2016). Full and fast calibration of the heston stochastic volatility model. *European Journal of Operational Research*, *263*(2), 625–638.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, *2*(4), 303–314.

Engsted, T. (2009). Statistical vs. economic significance in economics and econometrics: Further comments on mccloskey and ziliak. *Journal of Economic Methodology*, *16*(4), 393–408.

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

Fiorentini, G., Calzolari, G., & Panattoni, L. (2002). Estimation and empirical performance of heston's stochastic volatility model: The case of a thinly traded market. *Journal of Empirical Finance*, *9*(2), 149–171.

Friedman, M. (1953). The methodology of positive economics. In *Essays in positive economics* (pp. 3–43). University of Chicago Press.

Garcia, R., & Gençay, R. (2000). Pricing and hedging derivative securities with neural networks and a homogeneity hint. *Journal of Econometrics*, *94*(1-2), 93–115.

Gatheral, J. (2006). *The volatility surface: A practitioner's guide*. Wiley.

Ghaziri, H., Elfakhani, S., & Assi, J. (2000). Neural networks approach to pricing options. *Neural Network World*, *10*(1–2), 271–277.

Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 249–256.

Glorot, X., Bordes, A., & Bengio, Y. (2011). Deep sparse rectifier neural networks. *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Gradojevic, N., Gençay, R., & Kukolj, D. (2009). Option pricing with modular neural networks. *IEEE Transactions on Neural Networks*, *20*(4), 626–637.

Gradojevic, N., & Kukolj, D. (2022). Unlocking the black box: Non-parametric option pricing before and during covid-19. *Annals of Operations Research*.

Gross, E., Kruger, R., & Toerien, F. (2025). A comparative analysis of option pricing models: Black–scholes, bachelier, and artificial neural networks. *Risk Management*, *27*, 8.

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, *33*(5), 2223–2273.

Hansen, P. R., Lunde, A., & Nason, J. M. (2011). The model confidence set. *Econometrica*, *79*(2), 453–497.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning* (2nd ed.). Springer.

Hendry, D. F. (2009). The methodology of empirical econometric modeling: Applied econometrics through the looking-glass. *Economic Modelling*, *26*(2), 287–288.

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), 327–343.

Hornik, K., Stinchcombe, M. B., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*(5), 359–366.

Hull, J. C. (2022). *Options, futures, and other derivatives* (11th ed.). Pearson.

Hutchinson, J. M., Lo, A. W., & Poggio, T. (1994). A nonparametric approach to pricing and hedging derivative securities via learning networks. *The Journal of Finance*, *49*(3), 851–889.

Ilalan, D. (2015). Modeling correlation structure for collateralized debt obligations. *International Journal of Financial Research*, *6*(2), 72–85.

İltüzer, Z. (2022). Option pricing with neural networks vs. black-scholes under different volatility forecasting approaches for bist 30 index options. *Borsa Istanbul Review*, *22*(4), 725–742.

Jiao, Y., Ma, C., Scotti, S., & Zhou, C. (2018). The alpha-heston stochastic volatility model. *Mathematical Finance*.

Kelly, B., & Xiu, D. (2023). *Financial machine learning*. Foundations; Trends in Finance.

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. *International Conference on Learning Representations*.

Lajbcygier, P. (2004). Improving option pricing with the product constrained hybrid neural network. *IEEE Transactions on Neural Networks*.

Le Floc'h, F. (2025). How to choose my stochastic volatility parameters? a review [arXiv preprint arXiv:2512.19821].

Lindo, D. (2018). Why derivatives need models: The political economy of derivative valuation models. *Economy and Society*, *47*(4), 638–665.

Liu, X., & Huang, H. (2016). The performance of hybrid artificial neural network models for option pricing during financial crises. *Expert Systems with Applications*.

MacBeth, J. D., & Merville, L. J. (1979). An empirical examination of the black-scholes call option pricing model. *The Journal of Finance*, *34*(5), 1173–1186.

Malliaris, M., & Salchenberger, L. (1993). A neural network model for estimating option prices. *Applied Intelligence*, *3*(3), 193–206.

Merton, R. C. (1973). Theory of rational option pricing. *The Bell Journal of Economics and Management Science*, *4*(1), 141–183.

Merton, R. C. (1976). Option pricing when underlying stock returns are discontinuous. *Journal of Financial Economics*, *3*(1–2), 125–144.

Milyushkov, G. (2025). Can machine learning algorithms outperform traditional models for option pricing? [arXiv preprint arXiv:2510.01446].

Mitra, S. K. (2012). An option pricing model that combines neural network approach and black scholes formula. *Global Journal of Computer Science and Technology*, *12*(4).

Politis, D. N., & Romano, J. P. (1994). The stationary bootstrap. *Journal of the American Statistical Association*, *89*(428), 1303–1313.

Rolnick, D., & Tegmark, M. (2018). The power of deeper networks for expressing natural functions. *International Conference on Learning Representations (ICLR)*.

Rømer, S. E. (2022). *Empirical analysis of rough and classical stochastic volatility models to the spx and vix smiles*. Quantitative Finance.

Rubinstein, M. (1994). Implied binomial trees. *The Journal of Finance*, *49*(3), 771–818.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215.

Ruf, J., & Wang, W. (2020). Neural networks for option pricing and hedging: A literature review. *Journal of Computational Finance*, *24*(1), 1–46.

Saito, K., & Jun, A. (2000). Neural network option pricing in connection with the black-scholes model. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Scott, L. O. (1987). Option pricing when the variance changes randomly: Theory, estimation, and an application. *Journal of Financial and Quantitative Analysis*, *22*(4), 419–438.

Sidhu, K. S., & Saxena, P. (2026). Beyond black-scholes: A computational framework for option pricing using heston, garch, and jump diffusion models [arXiv preprint arXiv:2604.06068].

Stein, E. M., & Stein, J. C. (1991). Stock price distributions with stochastic volatility: An analytic approach. *The Review of Financial Studies*, *4*(4), 727–752.

Storn, R., & Price, K. (1997). Differential evolution: A simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization*, *11*(4), 341–359.

von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton University Press.

Vrins, F. (2025). Derivatives pricing [Lecture notes, UCLouvain].

Wen, X., et al. (2012). Impact of us financial crisis on different countries: Based on the method of functional analysis of variance. *Physica A: Statistical Mechanics and its Applications*, *391*(4), 1438–1448.

Yang, L., et al. (2017). Gated neural networks for option pricing: Rationality by design. *Proceedings of the AAAI Conference on Artificial Intelligence*.