Forecasting VIX Futures Using Machine Learning and Volatility Surfaces

MAX HARPER

SID: 510471039

Finance Supervisor: Dr Richard Philip
Engineering Supervisor: Dr Clément Canonne

This thesis is submitted in partial fulfillment of
the requirements for the degree of
Bachelor of Advanced Computing (Honours)

School of Engineering
The University of Sydney
Australia

5 January 2026

Student Plagiarism: Compliance Statement

I Max Harper certify that:

I have read and understood the University of Sydney Student Plagiarism: Coursework Policy and Procedure;

I understand that failure to comply with the Student Plagiarism: Coursework Policy and Procedure can lead to the University commencing proceedings against me for potential student misconduct under Chapter 8 of the University of Sydney By-Law 1999 (as amended);

This work is substantially my own, and to the extent that any part of this work is not my own I have indicated that it is not my own by acknowledging the source of that part or those parts of the work.

Name: Max Harper

Signature: **Date**: January 5, 2026

Abstract

This study examines whether the S&P 500 implied volatility (IV) surface contains predictive information for VIX futures, motivated by the construction of the VIX itself as a weighted average of option-implied volatility. A range of dimensionality reduction techniques are employed to condense the high-dimensional IV surface into features suitable for forecasting, with particular attention given to principal component analysis (PCA) for its interpretability, and autoencoders as a modern non-linear alternative. These features are evaluated across several machine learning models to assess predictive accuracy, with economic significance measured through trading simulations on VIX futures. The findings show that the IV surface provides valuable predictive signals, with PCA yielding parsimonious and interpretable predictors that achieve strong performance. Autoencoders deliver competitive results but present challenges in interpretability. The analysis highlights that machine learning models leveraging IV-surface features can generate economically meaningful trading profits. The paper contributes by introducing the IV surface as a predictor of VIX futures, demonstrating the utility of dimensionality reduction methods as forecasting inputs, and providing a comparison of dimensionality reduction techniques new to this domain.

Acknowledgements

Firstly, I would like to thank Richard Philip as the main overseer of my project and Clément Canonne as my co-supervisor. I would also like to thank my parents Edwina and Wade, and girlfriend Hannah for support throughout the year.

Contents

CONTENTS

List of Figures

List of Tables

1. Introduction

Volatility in financial markets has attracted significant attention from investors and academics aiming to mitigate losses from large price deviations [1]. Successful forecasting of volatility would allow fund managers to sell or hedge their positions in advance to reduce risk. The introduction of volatility derivative products has also allowed investors to hedge their portfolios with "long volatility" products and speculators to trade the markets future expectations of volatility [2]. The inclusion of this hedge in a portfolio has been shown to improve risk adjusted returns [3]. However, research has consistently shown that volatility is a complex phenomenon that cannot be easily forecasted [4].

The Chicago Board of Options Exchange (CBOE) created the volatility index (VIX) which measures the S&P500 option-implied volatility, acting as a "fear gauge" for future volatility [5]. Since 2004, monthly futures are also traded against the VIX, which create a highly liquid product to trade volatility expectations [6]. This paper investigates whether the S&P 500 implied volatility (IV) surface contains predictive information for VIX futures and how such information can be extracted and interpreted. This question is motivated by the construction of the VIX itself, which is a weighted average of implied volatility from across the surface but has yet to be used as a predictor.

Earlier research models VIX futures using traditional econometric models which are limited by their parametric form and assumptions [1]. The explosion of machine learning and artificial intelligence has led to new attempts to forecast VIX futures using these more modern techniques [7]. Existing predictors of VIX futures typically rely on exogenous macroeconomic variables or VIX derivatives, with mixed empirical success. In contrast, little attention has been given to the S&P 500 options data as a predictor, despite the VIX being derived from these instruments. The implied volatility surface, obtained by comparing option-implied volatility across strikes and maturities, contains rich information, yet it has rarely been applied beyond a limited high-frequency study with little practical relevance. This lack of research likely reflects the surface's high dimensionality and the computational challenges involved in processing such data.

Given the surface's high dimensionality, the analysis considers whether dimensionality reduction methods can produce more parsimonious representations suitable for forecasting. Particular attention is given to principal component analysis (PCA) for its interpretability, with further interest in the comparative performance and explainability of features derived from modern non-linear approaches such as autoencoders. In addition, the study evaluates which machine learning models best exploit these features for predictive accuracy and economic relevance.

The analysis is conducted from a machine learning perspective, evaluating the predictive fit of volatility-surface features across a range of models and benchmarking against existing studies. Economic significance is assessed through a trading simulation that gauges the practical value of these predictors. Finally,

the study compares the interpretability of dimensionality reduction techniques, highlighting differences between PCA and autoencoders.

The results show that the implied volatility surface does contain predictive information for VIX futures. PCA provides a parsimonious and interpretable feature set that achieves strong predictive performance and allows attribution back to the original surface. Autoencoders offer competitive accuracy, but their features are more difficult to interpret. From an economic perspective, the forecasts translate into profitable trading simulations, supporting their practical applicability.

These findings contribute to the existing literature by directly addressing several gaps identified in prior research. First, whereas most studies rely on macroeconomic or derivative-based predictors of VIX futures, this paper introduces the implied volatility surface itself as a new and accurate predictor. Second, it extends the methodological literature by showing that dimensionality reduction techniques can extract informative and interpretable factors from high-dimensional option data. Third, by comparing PCA with non-linear autoencoders, the study provides the first evaluation of traditional versus modern feature extraction approaches for creating volatility forecasting inputs, contributing to both the econometric and machine learning strands of the volatility literature.

2. Literature Review

2.1. Volatility, VIX and VIX Futures

Volatility is defined as a measure of the uncertainty of the return realised on an asset [8], however the specifics of how this is calculated and measured has multiple forms. While volatility is often modeled latently using conditional variance, it is measured ex post as "realised volatility" using the standard deviation of returns. Realised volatility is however variable and error prone due to market microstructure fluctuations and varying sampling frequency [9]. There is also "implied volatility", a forward-looking estimate of realised volatility. This is the only free parameter in the Black-Scholes options pricing model [10], which reflects the market's expectation of volatility and is calculated by inverting the pricing formula with an option's market price. The assumptions underlying the calculation of implied volatility have several problematic elements such as the log-normal return distribution and the use of a constant value for volatility, however, it still serves as a useful proxy for risk and uncertainty [1].

The VIX index is a weighted average of implied volatility. It is weighted using the inverse square of an option's strike price to create a payoff independent of underlying index price and proportional to volatility [11]. When markets experience uncertainty, investors pay more for the hedging insurance provided by options, thus increasing the value of the VIX index [12]. This again is why it serves as a prominent "fear gauge" from within the options market [13].

The VIX is calculated using the following formula:

$$\text{VIX} = 100 \times \sqrt{\frac{2}{T}\sum_i \frac{\Delta K_i}{K_i^2} e^{RT} Q(K_i) - \frac{1}{T}\left(\frac{F}{K_0} - 1\right)^2} \tag{1}$$

Where:

- T is the time to expiration (in years) of the option.
- K_i is the i-th strike price.
- ΔK_i is the interval between strike prices.
- R is the risk-free interest rate to the option's expiration.
- $Q(K_i)$ is the midpoint of the bid-ask spread for the option with strike K_i.
- F is the forward index level, derived from option prices.
- K_0 is the first strike below the forward index level F.

There have been concerns raised with this calculation as the Chicago Board Option Exchange's (CBOE) applies a cutoff rule once two consecutive option strikes with no bids and offers are encountered; which can exclude options and produce erratic jumps [14].

The VIX is composed of thousands of S&P500 puts and calls which constantly change price and weighting. Thus, maintaining a replication of the VIX would involve an impractical amount of transactions in options markets affected by illiquidity. The VIX is therefore not directly tradable, but from 2004, monthly cash-settled VIX futures have been traded on the CBOE futures exchange [6]. This serves as a highly-liquid marketplace for speculators to trade their expectations of volatility, theoretically free from arbitrage with a replicated VIX. It also provides a product for the "long volatility hedge" discussed earlier, with its negative correlation improving portfolio protection in market crashes [15]. It is worth noting that this is not a completely pure market of hedgers and speculators, with significant evidence of attempted manipulation of VIX futures at settlement time via deep out-of-the-money (OTM) put options [16].

Comparing the prices of VIX futures contracts across different maturities yields the futures term structure. Traditional theory suggested that the futures term structure reflected the market's expectations of future prices (the Expectations Hypothesis) [17]. While the shape of this curve fluctuates over time, its long-run average tends to be upward sloping, centered around a long term mean level [18]. This structure is known as *contango*, where futures prices exceed the spot VIX, and implying that holding a long position in VIX futures typically results losses over time as the futures price converges to spot [19].

More recent literature strongly rejects the Expectations Hypothesis, instead suggesting the existence of a time varying investor risk premium. This is calculated as the difference between realised and implied

volatility, and quantifies the premium that sellers charge to compensate for taking downside risk in the event of a market crash [20]. Dew-Becker [21] finds this on average to be negative, implying that there is a premium, and it is largest for near-term futures. Bollerslev, Johnston and Nossman all confirm the existence of this phenomenon and show its predictive ability on index returns, volatility product returns and the VIX index respectively [22][18][23]. Risk premiums have been demonstrated to display counter-intuitive behaviour such as staying constant during risky periods [24].

2.2. Forecasting VIX Futures

Previous research approaches forecasting VIX futures through two techniques: volatility models and machine learning. Traditional volatility models use assumptions about the process of volatility in order to derive pricing formulas; notably using historical volatility models, generalised autoregressive conditional heteroskedasticity (GARCH) models and stochastic volatility models. These parametric forms are not used by the machine learning approaches which instead use explanatory variables to estimate the futures price. Machine learning methods can be distinguished across higher and lower frequency prediction intervals and further by the types of explanatory variables used.

The simplest methods used in past research regress future volatility using the historical volatility series. These include autoregressive (AR) models, autoregressive integrated moving average processes (ARIMA) and Heterogeneous AR models (HAR). These models are oversimplified, constrained by linearity and have been shown to be inconsistent with observed market behavior [25][1]. These are typically used as baseline forecasts in machine learning papers and do not produce notable forecasting accuracy [26][27][28][29].

GARCH and stochastic volatility models utilise different forms and assumptions to model complex behaviour not captured by historical volatility models. GARCH models assume volatility follows a conditional autoregressive process, which is designed to capture more complex processes like volatility clustering [30]. This was used to price VIX futures with a Heston-Nandi GARCH model and extended using a GJR-GARCH model, with both models aimed at incorporating the asymmetrical responses of implied volatility to returns [31][32]. In contrast, stochastic models assume that underlying volatility follows a stochastic process and forecast futures as the conditional expected mean of this volatility. A prominent example of this process is the stochastic square root mean reverting process first explored by Zhang and Zhu [33][34] and later expanded upon by Dotsis et al [35] by adding jumps. Other processes such as the log Ornstein-Uhlenbeck process (diffusion with a mean reverting drift) improved the fit of VIX futures data with the addition of a central tendency component [25].

These models have two main limitations. Firstly, VIX futures pricing models involve several layers of abstraction that can potentially oversimplify the relationship between the VIX index and VIX futures prices. For instance, Zhang and Zhu [33] propose a linear relationship between the squared VIX and instantaneous variance, leading to a futures price derived from a risk-neutral integral. These assumptions

may not always hold in reality and as discussed earlier, VIX futures are influenced by an interplay of market expectations and risk premia, suggesting a more complex relationship. VIX futures move in the opposite direction to the VIX on 26% of trading days, implying that forecasting volatility, implied volatility and futures on the products are distinct tasks [36]. This highlights the value of machine learning models which can be directly fit to VIX futures price without any abstraction or assumptions.

Furthermore, these models are restricted by the rigidity of their assumptions. Small alterations to the underlying process for volatility constantly yield small improvements between papers however there is no consensus about the underlying dynamics [35][25]. The success of multiple underlying processes suggests that each simplifies a more complex underlying pricing mechanism. Poon surveys 93 papers and found GARCH models don't show significant improvement in volatility forecasts compared to more simplistic historical volatility methods [1]. This inherent rigidity in model structure indicates a clear need for more flexible and adaptable forecasting techniques, such as those offered by machine learning algorithms.

Machine learning's flexibility and generalisation capabilities have led to impressive results in volatility forecasting. Deep learning and ensemble methods, in particular, have demonstrated superior performance over traditional models in predicting realised variance across various studies, with similarly promising outcomes for implied volatility forecasts [7][37].

One focus of current machine learning research into the VIX specifically is high frequency pricing using deep learning. Hirsa and Osterrieder [38][39] use recurrent neural networks and long-short term memory (LSTM) models to process S&P500 options quotes and predict the VIX index on a minute to minute basis, while Hirsa extends this to also price VIX futures. This highlights the predictive efficacy of options, however the minutely time scale has limited applicability for investors looking to hedge. Both papers also don't quantify or elaborate on any trading strategies based on their forecasts and the deep learning approaches offer limited economic insights as to how these S&P500 options affect prices. Beyond forecasting, deep learning has also been used for dimensionality reduction via autoencoders [40]; this approach will be examined later as an alternative to principal components and has not yet been applied in this context.

More research has been devoted to "mid-frequency" prediction intervals, characterised by mixed independent variables and predictive success. Early attempts by Konstantinidi [27][29] used exogenous macro-economic factors such as oil price and bond curve slopes to forecast VIX and VIX futures, concluding that the index and its futures have limited predictability and make a case for strong market efficiency. In contrast to this, macro-economic predictors were however shown to produce profitable returns on a longer month-to-month time scale by Vrontos [41]. The returns and volatility of related financial markets have exhibited more success as predictors however experience significant drawdowns and variability when used in trading strategies [26][42]. Finally, a comparison of the trading strategies

shown in these papers highlights the need for more complexity with option straddles outperforming simplistic trading strategies, such as taking a long position if VIX forecasts are positive [29][26].

In contrast to exogenous factors, VIX derivative products show notable predictive value across multiple studies. Johnston [18] shows that the second principal component of the VIX futures term structure (known as slope) is a statistically and economically significant predictor of VIX futures returns. Hosker [28] also uses spreads between VIX futures and options as predictors to predict 3 and 5 day-ahead VIX futures returns with the best results emerging from deep learning models. These studies broadly highlight the predictive merit of volatility derivative market sentiment and curve structure.

The machine learning attempts at forecasting VIX futures reveal a marked contrast between "black box" deep learning models and more interpretable linear approaches. While deep learning techniques [38][39] often achieve strong predictive performance, they tend to provide limited economic insight. In contrast, explainable models allow for a clearer assessment of variable significance and predictive value which is increasingly valued within the machine learning and finance communities [43][7]. Notably, the most robust studies extend their evaluation beyond traditional error metrics such as RMSE and MAE by incorporating economic performance measures like the Sharpe ratio to quantify trading strategy efficacy [18].

Machine learning offers a viable framework to both forecast VIX futures and assess variable efficacy. Directly forecasting futures prices using machine learning bypasses the multiple layers of abstraction required of traditional modeling approaches. This enables models to more effectively capture the complex interplay of expectations and risk premia present in the VIX futures markets. As discussed, medium-frequency prediction remains largely dominated by models using exogenous variables and VIX derivative data. While the option quote surface has been employed in high-frequency forecasting of VIX futures, there is also a notable gap in the current literature regarding the predictive efficacy of the S&P 500 option-implied volatility surface for forecasting VIX futures in the medium frequency.

2.3. The SPX Implied Volatility Surface

A consistent finding in volatility literature is the strong predictive power of option-implied volatility. Latane [44] found this to be a better predictor of future realised volatility than historical volatility and is corroborated by Poon [1], who observed implied volatility outperforming historical volatility models in a significant majority (76%) of reviewed studies. Interestingly, it has been shown that the VIX index itself, as a measure of implied volatility, forecasted future realised volatility more accurately than other models [45].

By comparing implied volatilities of options across strikes of the same maturity, one observes the well-documented “smirk” or “skew,” wherein deep out-of-the-money (OTM) puts carry higher implied volatilities as compensation for downside insurance [46][8]. The shape and slope of this surface have

been shown to predict equity returns [47][48] and histogram-based measures of skew were found to be significant predictors of the probability of market crashes [49].

Short-dated OTM puts, in particular, provide a sensitive gauge of market risk sentiment. Their convex payoff structure makes them attractive to informed traders ahead of downturns, leading to price and volatility spikes that may act as early warning signals [48][50]. Options markets more broadly have been found to lead equity price discovery, often reflecting non-public information before it becomes evident in spot prices [51]. A striking historical case occurred before the 1987 crash, when S&P 500 OTM puts were priced at a 25% premium over theoretical values, anticipating the subsequent 23% market decline [52].

Given the high dimensionality of the IV surface, Principal Component Analysis (PCA) has previously been used to study the dynamics of implied volatility surfaces [53][54], yielding the traditional 'level and slope' interpretations for the first two components [55]. PCA produces orthogonal linear combinations of predictive features to maximise variance which can thus produce parsimonious summaries of dominant factors within high-dimensional data [56]. These studies largely employ PCA descriptively rather than as forecasting input, however PCA has been shown to provide efficient forecasting inputs [57]. This research addresses that gap by applying PCA to the IV surface specifically for forecasting VIX futures, leveraging dimensionality reduction to extract predictive signals from complex data.

As an alternative to PCA, this research also employs autoencoders for dimensionality reduction. Neural networks are trained to compress the input features into a lower-dimensional latent representation and then reconstruct the original feature space, thereby capturing efficient latent factors with the capacity to model nonlinear structure beyond PCA [40]. In addition, this study applies masked autoencoders (MAE), a recent line of research designed to learn cross-feature dependencies by reconstructing deliberately hidden inputs. This has proved highly effective in text and image representation learning [58][59], and has not yet been applied to volatility surfaces or VIX futures forecasting.

3. Data

3.1. Data Sources

Daily close and last price data for VIX futures was obtained from LSEG's Datascope platform from June 2004 (product inception) to June 2025 [60]. Specifically, this contained rows of quote dates, contract identifiers, last price and universal close price which was used for significance testing. This platform automatically performs contract rolling hence this was not required during pre-processing. The validity of this data was also cross-checked with other data sources such as CBOE for deviations which yielded only small variation.

Daily closing SPX option quote and implied volatility data was collected from OptionMetrics [61]. This contained quote date, expiry date, strike price, price, implied volatility and option greek metrics from

Jan 1996 to Feb 2023 and forms the basis of the implied volatility surface predictive features. Due to data availability limitations, the SPX options dataset was truncated at February 2023.

Finally, daily S&P500 close data was collected from Yahoo Finance which was used to standardise and discretise option strike prices into "percentage moneyness" as is discussed during Data Preparation and Feature Engineering [62]. Spot VIX data was also collected from Yahoo Finance for Exploratory Data Analysis.

3.2. Data Preparation

The VIX futures dataset was truncated at January 2006 to avoid the higher incidence of missing values around the product's introduction. After this cutoff, rows with missing price data were removed, representing 0.75% of trading days.

Compared to the VIX futures, more substantial processing was necessary to construct the implied volatility surface. Option quote dates were first aligned temporally with S&P 500 closing prices to calculate a "moneyness" measure. This was defined as the strike price divided by the index level in order to standardise strike prices across time as used in multiple studies [28][63]. This measure was then discretised into 10% moneyness buckets ranging from 80% to 110% to categorise options into groups ranging from out-of-the-money puts to at-the-money calls. Similarly, time-to-expiry was calculated in days and grouped into four buckets: less than 30 days, 60 days, 90 days, and 180 days. These bins were chosen to aid data completeness as increased granularity created a larger degree of missing values. Finally, the remaining missing values were filled with their rows mean value. By aggregating quotes within each two-dimensional bucket and averaging their implied volatilities, disparate option contracts were transformed into a discretised implied volatility surface, forming the base of the predictive features.

This discretised surface was temporally joined to the VIX futures data which formed a combined daily dataset from January 2006 to February 2023.

3.3. Exploratory Data Analysis

The purpose of this exploratory data analysis is to characterize the statistical properties of the VIX futures and SPX implied volatility surface. This will identify features such as contango in the term structure and the volatility smirk, which will later inform feature engineering and model design.

VIX Futures and VIX

Table 1 shows summary statistics for all the VIX futures contracts obtained, with the mean term structure displayed in Figure 1. Examining the mean term structure of VIX futures reveals the contango structure from the front month mean of 19.517 to the 8th month mean of 22.250 as described by Johnston [18]. This reflects the mean-reverting property of volatility, as investors expect convergence to the longer-term contract's value, representing the markets long-term expected mean of volatility. A higher standard

deviation of 7.836 is also observed in the front month compared to the 4.971 of the 8th contract given its higher sensitivity and reactivity to spot movements. This is also reflected in the inter-quartile range and maximum values, which are higher for near-term contracts compared to longer-term as these contracts don't adjust as much with spikes. It also worth noting the high standard deviation representing 40% of the mean value.

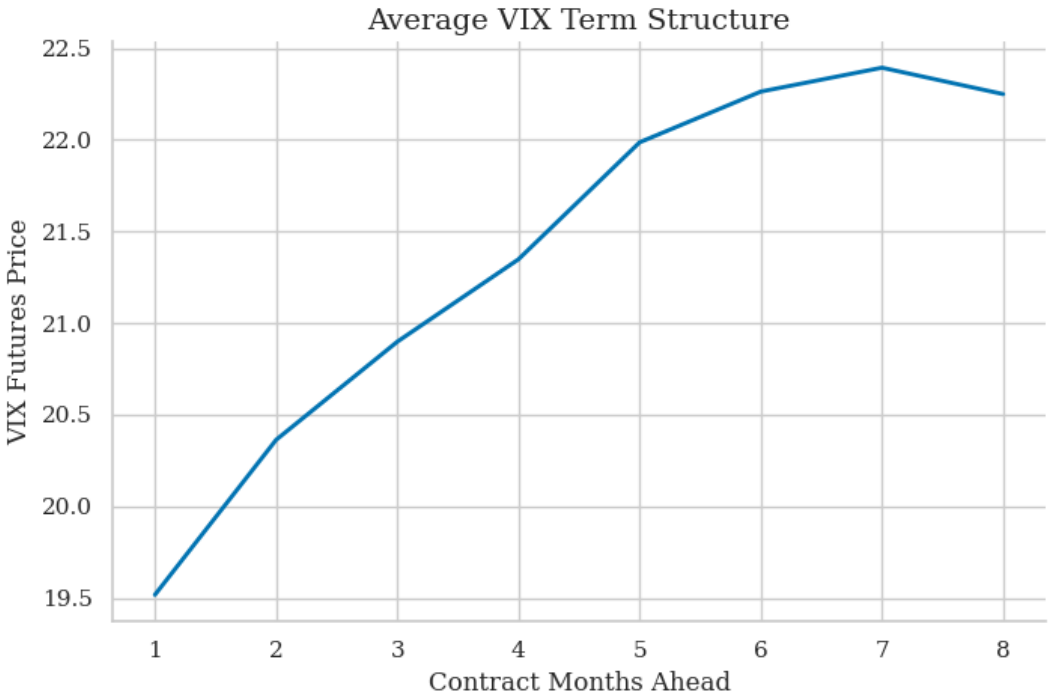

FIGURE 1: Mean VIX Futures Term Structure showing Contango Shape

TABLE 1: Summary statistics of VIX Futures Contracts One to Eight Months Ahead (VXc1–VXc8)

	VXc1	VXc2	VXc3	VXc4	VXc5	VXc6	VXc7	VXc8
Mean	19.517	20.363	20.899	21.348	21.987	22.264	22.394	22.250
Std	7.836	7.057	6.490	6.082	5.851	5.576	5.319	4.971
Min	9.600	11.300	12.200	12.990	13.470	13.900	14.300	14.690
Median	17.205	18.250	18.800	19.300	20.105	20.385	20.700	20.485
Max	81.950	70.800	60.080	51.680	47.760	45.990	44.500	44.000

Table 2 reports the correlation coefficient of the VIX index with each futures contract. Intuitively, as the front month contract (VXc1) is the nearest to expiry it has the highest correlation at 98.1% compared to

77.5% for the 8th contract. The front month contract's high correlation with spot VIX can be seen over time in the overlayed time series in Figure 2. This underscores its role as a "fear index," with pronounced spikes observed during the Global Financial Crisis reaching 67.9 and 81.95 during the COVID-19 pandemic. As a result of this strong correlation with the VIX, combined with its superior liquidity, it is the most suitable instrument for trading directional views on volatility and thus the primary focus of the subsequent analysis.

TABLE 2: Correlation coefficients of VIX futures contracts with the spot VIX

Contract	Correlation (%)
VXc1	0.981
VXc2	0.941
VXc3	0.905
VXc4	0.875
VXc5	0.843
VXc6	0.815
VXc7	0.793
VXc8	0.775

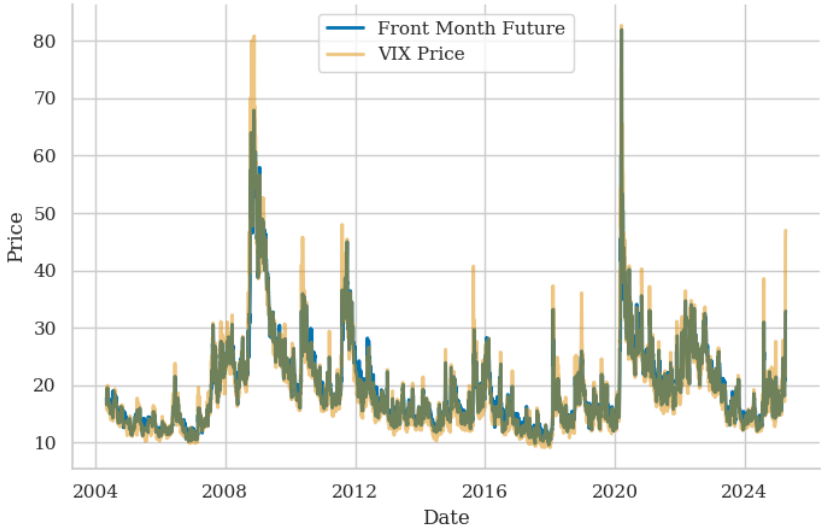

FIGURE 2: High Correlation of VIX Front Month Future Price and VIX over Time

Examining the distribution of the front-month VIX futures contract highlights distinctive properties of volatility. Figure 3 depicts a histogram of the contract. The distribution's skewness of 2.28 and pronounced right tail in the figure indicating that large upward moves in volatility are more frequent

than large downward moves. In addition, the distribution's kurtosis value of 7.65 reflects a strongly leptokurtic distribution, with a much higher probability of extreme outcomes than a Gaussian. These features are consistent with the presence of volatility shocks, a phenomenon often incorporated into econometric models through explicit jump components [25][47][64].

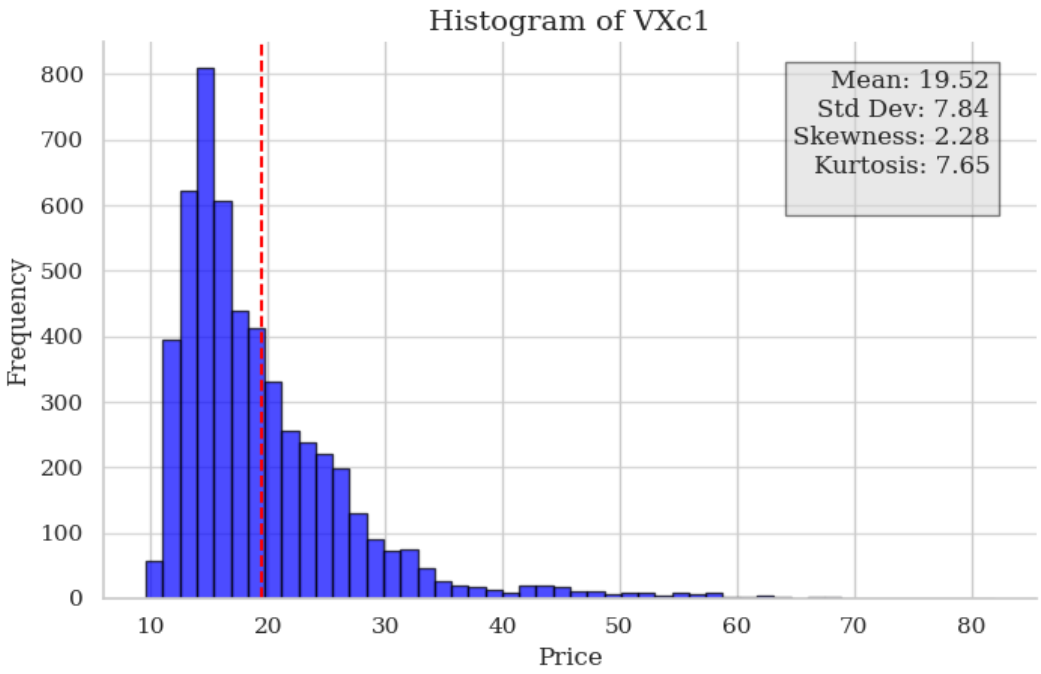

FIGURE 3: VIX Front Month Future Price Histogram

SPX Implied Volatility Surface

Plotting the mean implied volatility against moneyness in Figure 4 reveals the well-documented volatility smirk. This is consistent with previous studies, as option issuers demand higher compensation for highly leveraged risk insurance (tail risk pricing) [46].

Table 3 reports summary statistics for the bucketed dataset used in subsequent analysis. The smirk shape is clearly expressed across all expiry cross-sections, most pronounced in short term (< 30 day) options where mean IV in the 80% moneyness bucket has a mean value of 0.283, compared with 0.154 for ATM options in the 100% bucket. Examining the surface across term structure has less clear patterns, with the IV of 80% moneyness options declining notably at longer maturities, whereas ATM options remain nearly constant around 0.154.

Interestingly, the highest mean, maximum, and variance of implied volatility are observed for the 0-day expiry 80% moneyness options, representing short-dated out-of-the-money puts. This highlights the region of the surface that is most pronounced in risk sentiment signalling [52]. Nevertheless, valuable information is embedded across the entire surface. To capture these broader patterns, the subsequent analysis applies Principal Component Analysis, providing a more holistic and dimension-reduced representation of the volatility surface.

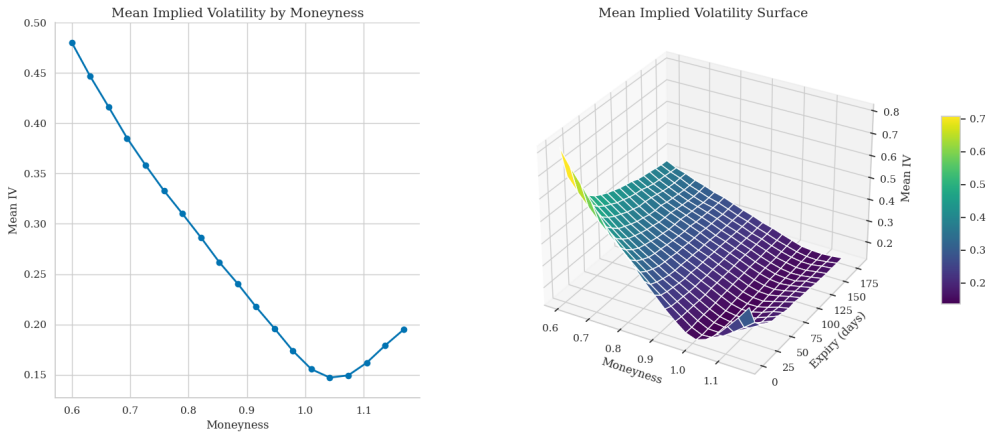

FIGURE 4: Volatility Smirk and Mean SPX Implied Volatility Surface

TABLE 3: Summary statistics of implied volatility by expiry and moneyness

Expiry / Moneyness	Mean	Std. Dev.	Min	Max
30 days, 80%	0.327	0.080	0.184	0.976
30 days, 90%	0.212	0.085	0.107	0.897
30 days, 100%	0.154	0.076	0.066	0.785
30 days, 110%	0.212	0.063	0.109	0.664
60 days, 80%	0.280	0.075	0.156	0.834
60 days, 90%	0.205	0.077	0.109	0.775
60 days, 100%	0.148	0.072	0.070	0.699
60 days, 110%	0.163	0.053	0.095	0.624
90 days, 80%	0.267	0.071	0.149	0.773
90 days, 90%	0.205	0.072	0.113	0.700
90 days, 100%	0.152	0.069	0.070	0.633
90 days, 110%	0.151	0.053	0.080	0.563
180 days, 80%	0.258	0.067	0.156	0.688
180 days, 90%	0.207	0.067	0.119	0.624
180 days, 100%	0.161	0.066	0.078	0.581
180 days, 110%	0.143	0.057	0.075	0.535

4. Methodology

4.1. Feature Engineering

Principal Component Analysis

To extract informative features from the discretised SPX implied volatility surface, Principal Component Analysis (PCA) was applied. PCA generates latent factors that are linear combinations of the original variables, chosen to maximise the variance explained. Each successive component is constructed to be orthogonal to those that precede it. PCA thus reduces dimensionality while capturing the dominant and unique modes of variation. This provides both parsimonious inputs for machine learning models and an alternative perspective on the surface's dynamics [57][56].

Table 4 reports the validation set RMSE across machine learning models for one day ahead VIX futures price using the full IV feature set and the PCA feature set. PCA yielded improvements across all models and thus the components were used as the primary feature set for subsequent analysis.

As this analysis is purely conducted on the discretised IV surface, each principal component can be represented as:

$$PC_n = \alpha_1 IV_{30,0.8} + \alpha_2 IV_{30,0.9} + ... + \alpha_{16} IV_{180,1.1}.$$

TABLE 4: Validation RMSE Comparison: PCA vs. IV Features

Model	IV Columns	PCA columns
Linear Regression	1.356	1.327
Ridge Regression	1.332	1.326
Lasso Regression	1.336	1.326
Random Forest	1.516	1.311
Gradient Boosting	1.443	1.376
Neural Network	2.218	1.281
Nearest Neighbors	1.502	1.435
LSTM	1.660	1.551

The sign and magnitude of the loading coefficients (α_i) can highlight the dominant sources of variation and can be examined across the surface to associate principal components with underlying economic factors. Figure 5 graphs cumulative explained variance across the principle components. The first three principal components accounted for 96.1% of the explained variance, with subsequent components contributing negligibly. Given this, only the first three were retained for analysis.

Furthermore, preliminary statistical tests supported the suitability of applying principal component analysis. Bartlett's Test of Sphericity yielded a p-value of 0.000, indicating significant correlations among features, while the Kaiser–Meyer–Olkin measure returned a value of 0.927, demonstrating a high degree of shared variance.

Auto-Encoders

As an alternative and more modern approach to dimensionality reduction, this study also employs autoencoders on the same IV surface data used for PCA. An autoencoder consists of a neural network encoder f that maps each input vector of implied volatilities $x_t = [IV_{30,0.8}, IV_{30,0.9}, \ldots, IV_{180,1.1}]$ into a k-dimensional latent representation z_t, and a decoder g that reconstructs x_t from z_t, as illustrated in Figure 6. k is a hyper-parameter chosen based on a validation set. Unlike PCA, autoencoders have been shown to capture nonlinear structures in the data and can extract more complex feature interactions [40].

In addition, this research tests masked autoencoders, which use the same architecture but randomly replace a subset of input features in x_t with zeros and compute the reconstruction loss only on the forecasted output of these masked components. This forces the model to learn dependencies among features to recover the hidden values. As discussed, masked autoencoders have recently achieved significant success in representation learning for natural language and computer vision tasks [58][59].

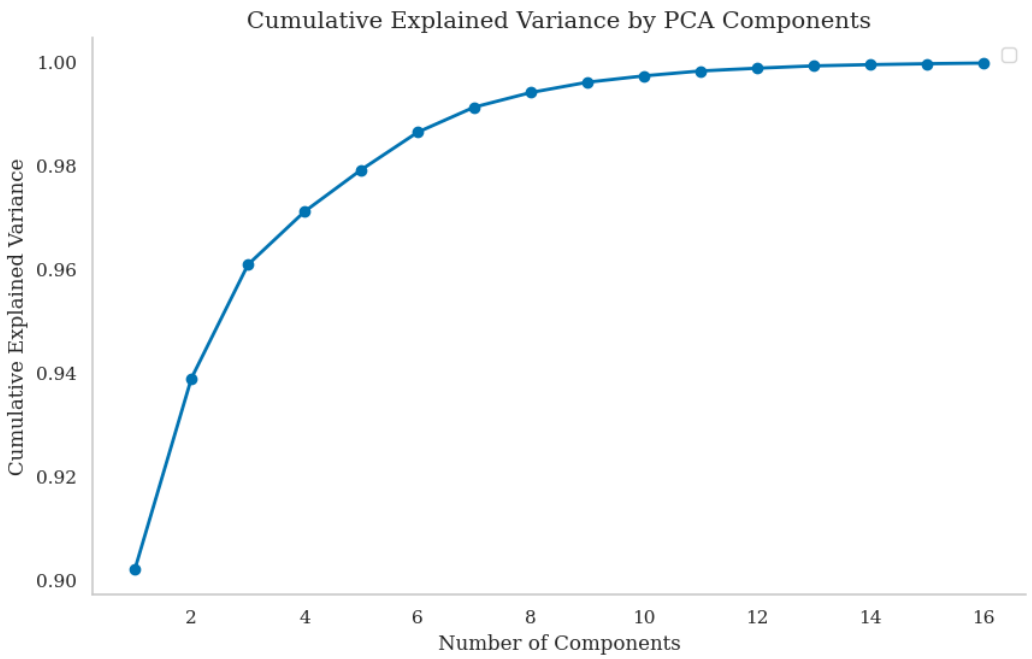

FIGURE 5: Explained Variance of Principal Components showing Diminishing Returns of Components Explanatory Power

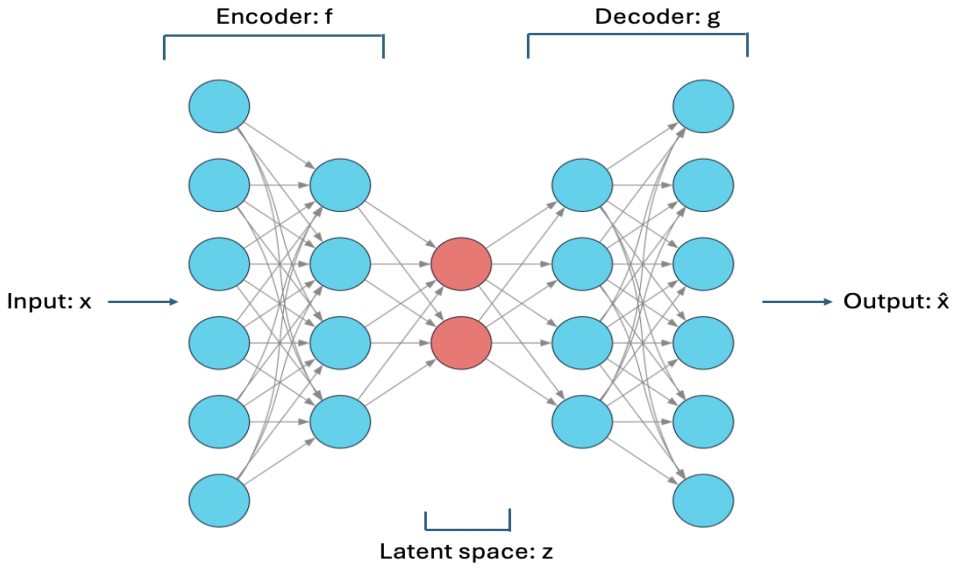

FIGURE 6: Auto-Encoder Architecture

4.2. Modelling

Dependent Variable

This study models VIX futures prices rather than returns for several reasons. Returns are highly volatile and heavy-tailed compared to the price series: for example, a move from \$15 to \$18 represents a 20% change but only a \$3 absolute shift. This is further demonstrated by the front-month return series having a standard deviation around 40% of its mean. Logarithmic transformations can reduce this volatility but at the cost of intuitive meaning and added analytical complexity.

In contrast, modelling raw prices provides a more stable and interpretable framework for assessing predictive performance and facilitates comparison with other studies that mainly use price. Table 5 reports R^2 for various return and price horizons. This demonstrates a significantly higher portion of variance in price can be explained than returns and hence why price was used as the dependent variable.

The front month contract is of most interest due its high correlation with VIX and superior liquidity in practice. In this study, this is denoted by *VXc1*, with *VXc2* representing the second-month-ahead contract, and subsequent contracts labeled similarly. This paper thus focuses on the regression of next day VIX front month futures price from the previous days principal components ($VXc1_{t+1} = f(PC1_t, PC2_t, PC3_t)$).

TABLE 5: Validation Set Explained Variance Across Dependent Variables from a Linear Regression with PC1, PC2 and PC3

Dependent Variable	R^2
One Day Return	0.020
One Week Return	0.065
One Month Return	0.161
$VXc1_{t+1}$	0.963
$VXc1_{t+3}$	0.934
$VXc1_{t+5}$	0.911

Models

A wide variety of machine learning models ranging from linear regressions to random forests and neural networks were tested and considered. Baseline controls were also included using traditional forecasting approaches, such as ARIMA, commonly employed in machine learning studies [26][28][29].

Hyperparameter optimisation was performed using a grid search over key parameters, with a validation subset spanning November 2016 to July 2019. The first three principal components of the implied volatility surface were used as predictors to forecast front-month VIX futures prices at the 1-day horizon.

Final model parameters were selected based on the validation root mean squared error (RMSE), with results reported in Table 6.

Most machine learning models showed comparable performance on the validation set, with ARIMA having the highest validation RMSE. The low shrinkage parameters selected for Ridge and Lasso indicate that regularisation was not strongly required for the regression coefficients. In contrast, the relatively shallow maximum depths chosen for Gradient Boosting and Random Forest suggest that these models have some potential to overfit the data.

TABLE 6: Models considered in the study with Hyperparameters and Validation Performance

Model	**Key Hyperparameters**	**Validation RMSE**
Machine Learning Models		
Linear Regression	–	1.327
Ridge Regression	Shrinkage parameter $\alpha = 0.1$	1.327
Lasso Regression	Shrinkage parameter $\alpha = 0.001$	1.326
Random Forest	Number of trees $n = 100$, max depth $d = 5$	1.306
Gradient Boosting	Number of trees $n = 125$, max depth $d = 1$, learning rate $\ell = 0.5$	1.376
Neural Network (MLP)	Hidden layer size $n = 100$, learning rate $\ell_0 = 1$, max iterations = 50	1.250
Long Short-Term Memory	Hidden layer size $n = 50$, learning rate $\ell_0 = 0.001$, epochs = 250	1.458
K-Nearest Neighbours	Number of neighbours $k = 8$	1.396
Baseline Control		
ARIMA	Orders $(1, 1, 3)$	1.496

Hyperparameter optimisation was also undertaken for the autoencoders using a linear regressions fit on the same validation set using the latent features as predictors. A 64-32-k-32-64 architechure was used for both standard and masked autoencoders, while a grid search was conducted for the amount of epochs, learning rate, latent space dimension (k) and the masking rate. The results can be below.

TABLE 7: Autoencoder Hyperparameters and Validation Performance

Autoencoder	Key Hyperparameters	Validation RMSE
Standard Autoencoder	Epochs $= 50$, learning rate $\ell = 0.01$, latent dimension $k = 4$	1.334
Masked Autoencoder	Epochs = 50, learning rate $\ell = 0.01$, latent dimension $k = 5$, masking rate $m = 30\%$	1.272

4.3. Training and Evaluation Framework

This research will employ a multi-faceted testing approach to evaluate model performance and practical utility.

Machine Learning Evaluation Framework

These forecasts will be for the price of the front month contract: $VXc1_{t+1} = f(PC1_t, PC2_t, PC3_t)$, with tests collecting Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and explained variance (R^2) to evaluate model performance.

First, a standard 80:20 train-test split will be implemented by withholding the final 20% of the time series data for out-of-sample testing, spanning July 2019 to February 2023. To mitigate the risk of overfitting to a specific time period and improve generalisation, this will be complemented by a 5 fold expanding time series cross validation averaging metrics across all folds as per Hosker's paper [28]. See Figure 7 for the dates of each period.

To strengthen the robustness and practical relevance of the results, multiple forecast horizons will be examined. These include a 1-day-ahead forecast, consistent with multiple studies [18][26][27], as well as 3 and 5 day horizons used by Hosker [28].

A modified version of these tests was used for the traditional forecasting baselines as these operate on time series data. ARIMA models were trained on the VXc1 training series, and forecasts were generated one step at a time. After each step, the model's internal state was updated with the new observation, but the model parameters were not retrained. This allows a more accurate comparison with the machine learning models.

Economic Significance Evaluation Framework

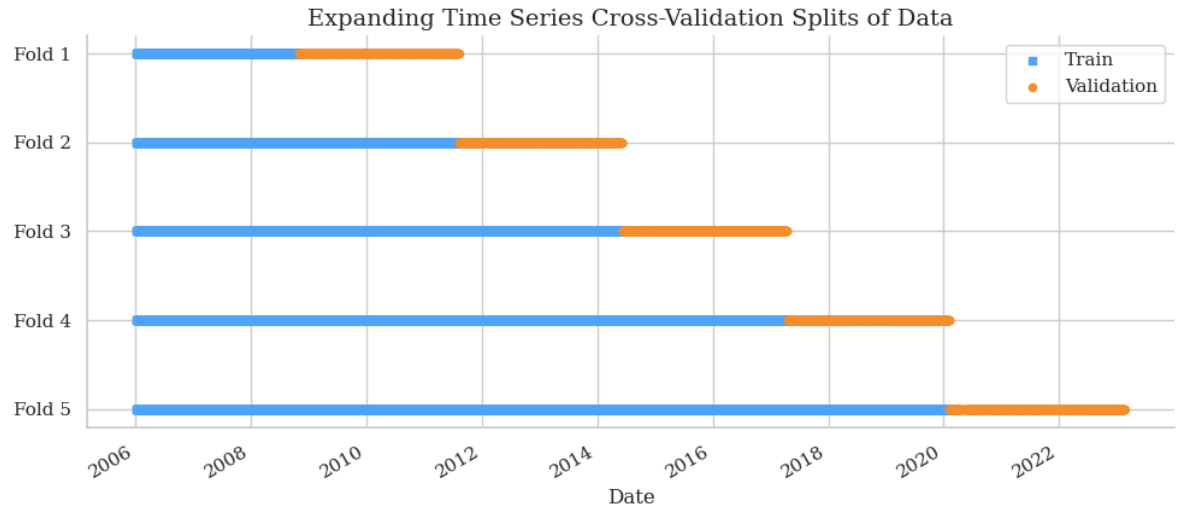

FIGURE 7: Expanding Time Series Cross Validation

Economic significance will be assessed using a trading strategy simulation. The models will be trained on the first 50% of the dataset, with the remaining data divided into three expanding folds for evaluation. This is due to the sensitivity of trading simulations to low initial data size and the comparatively large sample available. Trading costs and the a historical median spread of 5 cents, are incorporated to approximate realistic conditions. The trading simulation will be constructed the following trading rules:

If $\hat{y}_{t+1} - y_t >$ threshold; go long and hold until $\hat{y}_{t+1} - y_t <$ threshold.
If $\hat{y}_{t+1} - y_t < -$threshold; go short and hold until $\hat{y}_{t+1} - y_t > -$threshold.
Where y_t is the current VIX front month price on day t, and $\hat{y}_{t+1}$ is the forecast for tomorrows.

The trading rule acts directionally on the forecasted price change $\hat{y}_{t+1} - y_t$, however positions are only entered when the predicted move has sufficient value relative to the current price, regulated by the threshold. This again uses $\hat{y}_{t+1} = f(PC_1, PC_2, PC_3)$ to forecast, and thus the whole strategy is predicated on on the strength of the forecasts.

Comparisons across studies can then be made using common metrics when the time scales are broadly comparable. Normalised measures, such as the annualised Sharpe ratio and annualised compound return, will be collected and are the preferred metrics for cross-study evaluation. Adjustments to features, feature subsets, and trading strategy parameters will also be explored to enhance performance metrics and provide separate analysis.

Table 8 shows the sharpe ratio performance of a linear regression model using various thresholds for the strategy on a validation set. There is clear robustness to multiple thresholds and based on the performance metric a threshold of 1.0 was chosen for subsequent analysis.

TABLE 8: Linear Regression Validation Set Trading Performance with Various Thresholds

Threshold	Sharpe Ratio
0.00	1.557
0.25	1.307
0.50	1.488
0.75	1.698
1.00	2.357
1.25	2.216
1.50	1.959

Autoencoder Evaluation Framework

Autoencoded latent features z_t simply replace the principal components as predictive features in identical pipelines and training splits for both machine learning and economic evaluation frameworks. The regression is thus $VXc1_{t+1} = f(AE1_t, AE2_t, AE3_t...)$.

5. Results and Discussion

5.1. Principal Components

Principal Component One

The first principal component (PC1) of the implied volatility surface closely tracks the VIX front-month futures price (VXc1). Figure 8 plots the two variables together, showing a strong linear relationship. Further statistical testing showed a regression R^2 of 0.967 between the two variables, and rank one cointegration at a 95% confidence level (see Appendix Tables 19 and 18). These respectively imply high explanatory power and a long term equilibrium relationship between these two variables.

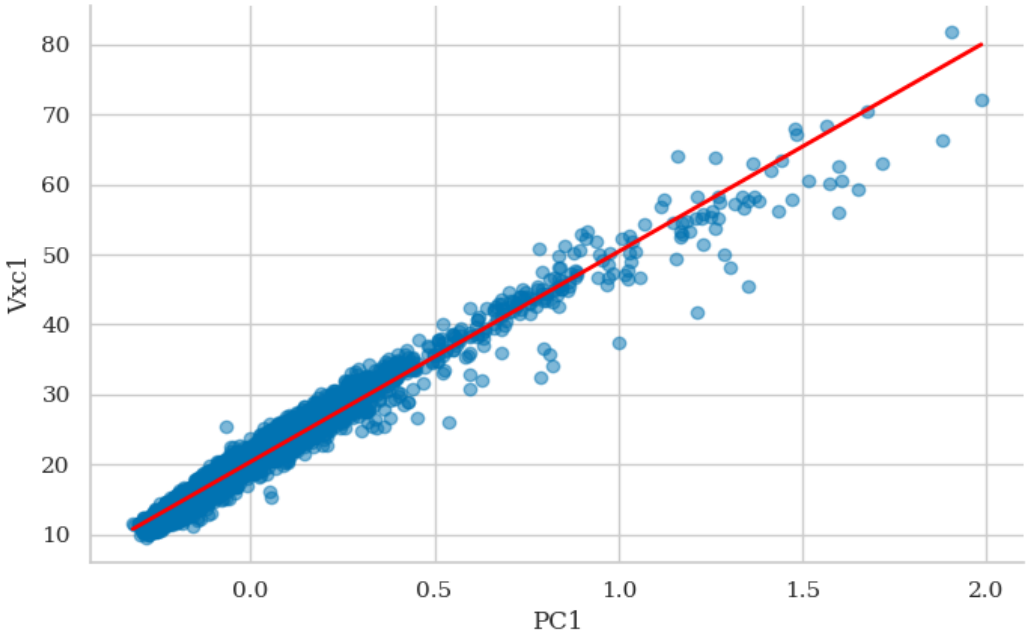

FIGURE 8: PC1 scattered against VIX front month

Figure 9 shows the principle component loading values across expiry and moneyness. This highlights comparatively minimal variation, with all values ranging from 0.181 to 0.309, indicating that PC1 functions as a weighted average of implied volatility (similar to the VIX index). This observation is consistent with prior PCA analyses of financial curves, where the first principal component is typically interpreted as the "level" of the curve [55]. Similarly, cross-sectional PCA of the implied volatility surface has also identified a dominant "level" factor [54].

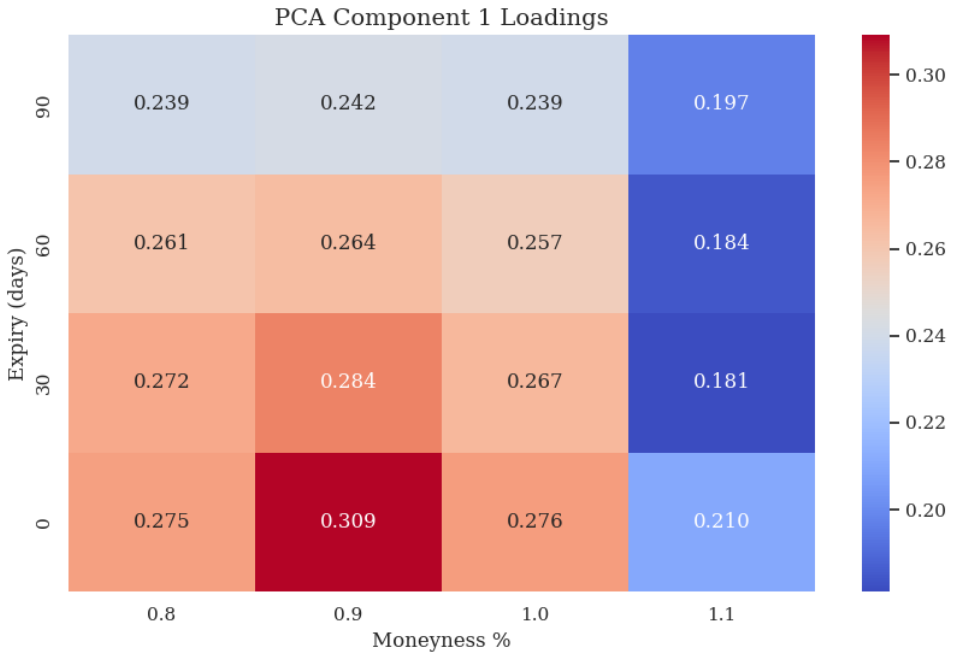

FIGURE 9: Loadings of PC1 by Moneyness and Expiry

Principal Component Two

The second principal component (PC2) is associated the skew of the option smirk. Figure 10 shows the loadings of the principle components, where there is a clear trend of negative loadings across puts and positive loadings on calls. This is consistent with previous work where the second component typically represents skew or slope of the surface [53][55][18].

Table 9 shows the regression results of VIX futures prices at various intervals from PC2. This was to test correlations with VXc1 like PC1, however the near zero values across all horizons demonstrate it has minimal explanatory power for price. This therefore suggests that variations in surface skew have limited impact on VIX futures prices.

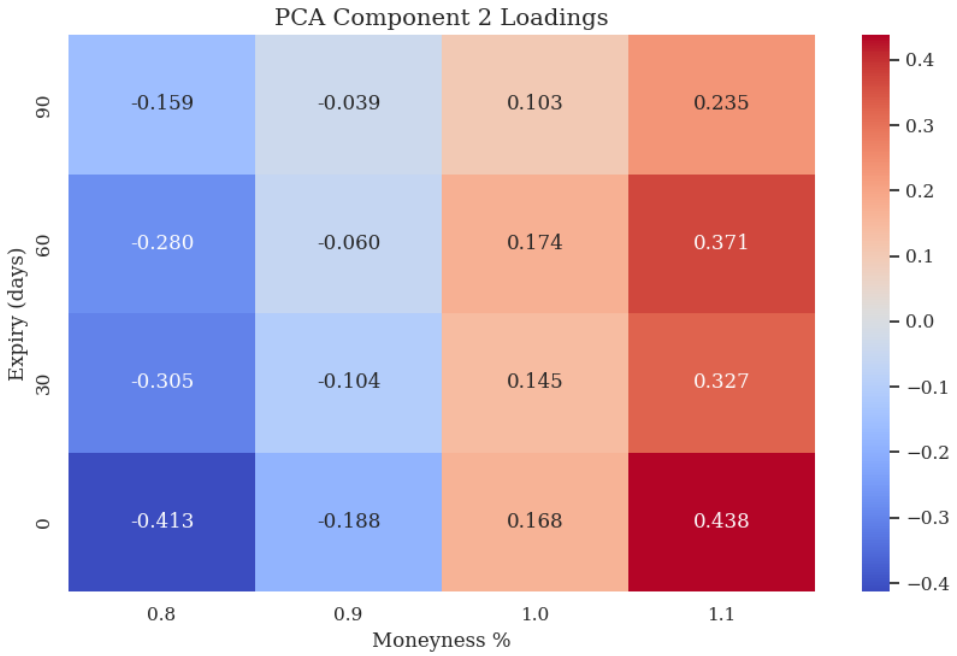

FIGURE 10: Principal Components Two Loadings

TABLE 9: OLS Regression Statistics: PC2 Regressing Front Month Futures

Predictor	R-squared
$VXc1_{t+1}$	0.003
$VXc1_{t+3}$	0.002
$VXc1_{t+5}$	0.002

Principal Component Three

The third principal component (PC3) has less intuitive interpretations than the previous components. Figure 11 depicts the loadings heatmap across expiry and moneyness. It shows negative loadings on longer expiry options and positive loadings on shorter dated options, with notably high loadings on short dated OTM puts and short dated ATM calls. This factor therefore captures term structure dynamics, and is particularly related to the IV of short-dated puts and calls.

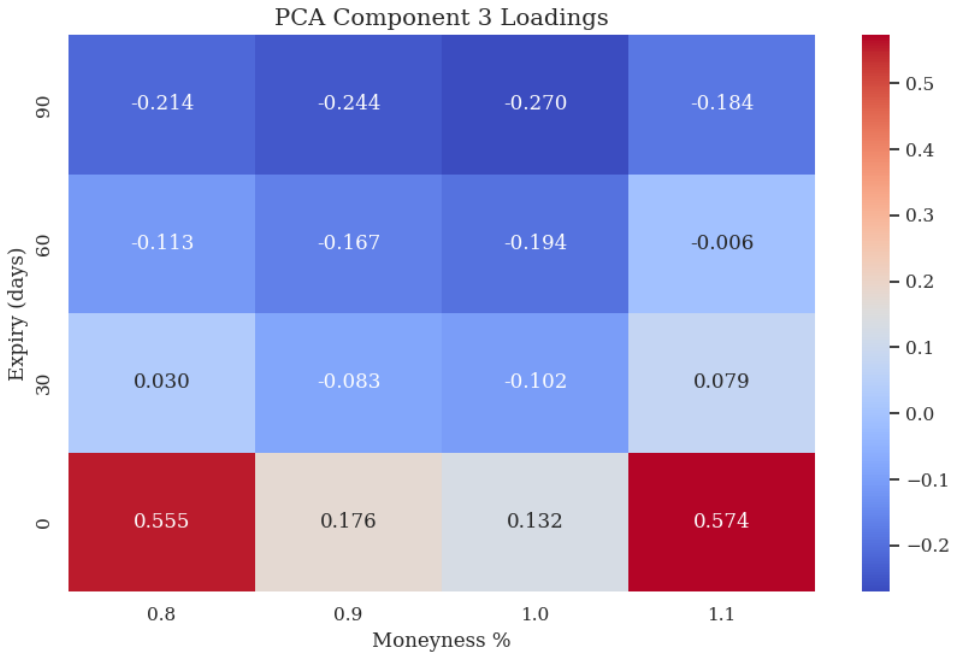

FIGURE 11: Loadings of PC3 by Moneyness and Expiry

PC3 demonstrated low correlations with VIX price similarly to PC2, however it exhibits a clear correlation with VIX returns. Figure 12 illustrates a scatter plot between PC3 and one month VIX returns. While not especially strong, such a relationship is notable given the well-known unpredictability and noise of short-term financial returns. Table 10 reports a regression summary for this relationship. The coefficient is both statistically significant (p-value = 0.000) and economically meaningful with an R^2 of 0.104, double that of PC2.

This return correlation and the heatmap suggest that PC3 may function as a stress signal, consistent with shifts in demand for short-dated options. These options typically become more expensive during economic panic, reflect hedging activity through downside protection and speculative demand for leveraged exposure [52].

[3]This utilises heteroskedasticity-robust p-values based off a residual plot, see Figure 18.

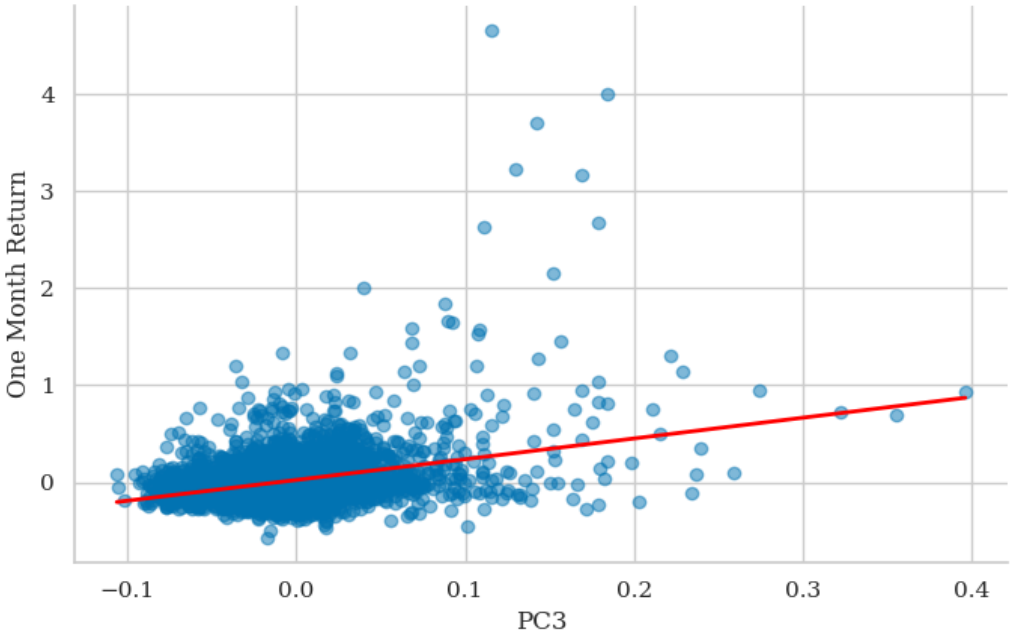

FIGURE 12: PC3 scattered against one month VIX front month Returns

TABLE 10: OLS Regression of One-Month Return on PC3 [3]

Variable	Coefficient	Std. Error	z-statistic	p-value
Intercept	0.024	0.004	6.453	0.000
PC3	2.140	0.235	9.117	0.000
R-squared		0.104		
Adjusted R-squared		0.104		
F-statistic		83.12		
Prob (F-statistic)		1.16×10^{-19}		

5.2. Machine Learning Tests of Significance

TABLE 11: Model Performance Metrics across 80:20 Train Test Split

	One-Day Ahead			Three-Day Ahead			Five-Day Ahead		
Model	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2
Linear Regression	2.118	1.344	0.921	3.140	1.902	0.826	3.910	2.261	0.729
Ridge Regression	2.119	1.345	0.921	3.141	1.903	0.825	3.912	2.262	0.729
Lasso Regression	2.125	1.350	0.920	3.147	1.907	0.825	3.919	2.267	0.728
Random Forest	2.328	1.416	0.904	2.973	1.920	0.844	3.768	2.296	0.748
Gradient Boosting	2.353	1.444	0.902	3.028	1.950	0.838	3.820	2.313	0.741
Neural Network	2.336	1.443	0.904	3.068	1.902	0.833	4.344	2.723	0.666
Nearest Neighbours	2.269	1.476	0.909	3.094	2.075	0.831	3.889	2.518	0.732
LSTM	2.765	1.720	0.865	3.924	2.231	0.727	4.643	2.704	0.617
ARIMA	2.598	1.615	0.881	3.806	2.298	0.744	4.578	2.734	0.628

TABLE 12: Model Performance Metrics across 5 Fold Expanding Cross Validation

	One-Day Ahead			Three-Day Ahead			Five-Day Ahead		
Model	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2
Linear Regression	1.725	1.188	0.898	2.332	1.576	0.814	2.815	1.880	0.727
Ridge Regression	1.728	1.189	0.898	2.334	1.577	0.814	2.815	1.880	0.727
Lasso Regression	1.727	1.190	0.898	2.334	1.577	0.814	2.817	1.881	0.727
Random Forest	2.015	1.372	0.875	2.533	1.750	0.796	3.105	2.090	0.694
Gradient Boosting	2.145	1.471	0.866	2.598	1.828	0.782	3.219	2.181	0.670
Neural Network	1.942	1.401	0.886	2.436	1.681	0.780	3.277	2.408	0.637
Nearest Neighbours	1.960	1.380	0.876	2.542	1.803	0.787	3.093	2.149	0.681
LSTM	2.231	1.514	0.839	2.848	1.865	0.730	3.648	2.337	0.586
ARIMA	1.892	1.245	0.869	2.569	1.728	0.761	3.043	2.052	0.663

Both the train–test split and the five-fold expanding cross-validation produced consistent and informative results. Tables 11 and 12 show machine learning metrics RMSE, MAE and R^2 for forecasting front month VIX futures price at the 1,3 and 5 day horizons for the test set and cross validation respectively. The most striking pattern is the dominance of linear models: OLS, Ridge, and Lasso repeatedly achieved the lowest RMSE and MAE, and the highest R^2 values across all forecast horizons. This suggests a meaningful degree of linearity between the principal components and VIX futures prices, as highlighted in the previous section.

Models with higher complexity and a tendency to overfit generally underperformed across all forecast horizons and evaluation metrics, with the LSTM model showing the poorest performance. This result

is contrasting to other studies where there was an out-performance of more complex models, namely artificial neural network models [7][28].

It is also notable that most models outperformed the traditional ARIMA benchmark, which demonstrated high performance metrics in other studies [28]. As a baseline comparison, this suggests that the principal components capture predictive information beyond what is contained in the VIX futures time series used by ARIMA.

As expected, forecast accuracy deteriorated with longer horizons, reflected in rising errors and falling explained variance. Performance under the simple 80:20 train–test split was also considerably weaker than under cross-validation. This discrepancy can be attributed to the unusually turbulent test period, which encompassed both the COVID-19 shock and the 2022 European bond market crisis.

Direct comparison of error metrics across studies is challenging given differences in test frameworks and time periods. Nevertheless, Hosker's 10-fold cross-validation over 2006–2018 provides a useful benchmark: their best-performing model achieved an RMSE of 4.73 and an R^2 of 0.43 for three-day-ahead VIX front-month futures [28]. By contrast, this study's RMSE of 2.332 and R^2 of 0.814 highlight both the stronger predictive accuracy and explanatory power of principal components in modeling VIX futures prices. It is also lower than Guo, Qiao and Konstantinidi's respective studies reported RMSE however these are on smaller train-test samples, making direct comparison difficult [32][65][29].

Feature Analysis

TABLE 13: Performance Metrics with Various Feature Subsets using Linear Regression and 5 Fold Expanding Cross Validation

	One-Day Ahead			Three-Day Ahead			Five-Day Ahead		
Features	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2
PC1	1.945	1.376	0.879	2.462	1.698	0.801	2.908	1.970	0.716
PC2	8.482	6.647	-1.409	8.480	6.643	-1.409	8.468	6.641	-1.410
PC3	8.619	6.714	-1.407	8.592	6.700	-1.408	8.560	6.686	-1.412
PC1, PC2	1.824	1.281	0.883	2.395	1.635	0.802	2.869	1.928	0.715
PC1, PC3	1.859	1.292	0.889	2.406	1.638	0.810	2.858	1.923	0.725
PC2, PC3	8.442	6.561	-1.373	8.448	6.573	-1.382	8.439	6.579	-1.390
PC1, PC2, PC3	1.725	1.188	0.898	2.332	1.576	0.814	2.815	1.880	0.727

The cross-validated performance of different principal component subsets highlights the dominant predictive power of PC1 for VIX futures prices. This can be seen in Table 13, which reports machine learning error metrics for feature subsets using the 5 fold cross validation with a linear regression model. The dominance of this factor is clear as excluding PC1 and predicting using only on PC2 and PC3 produces negative R^2 values, indicating forecasts were worse than the mean VXc1 value. It is also clearly economically significant, achieving an out-of-sample R^2 of 0.879 for one-day-ahead front-month VIX futures prices with no other factors. Notably however, incorporating PC2, PC3, or both alongside PC1 improves RMSE, MAE, and R^2 across all three forecast horizons, suggesting incremental predictive value in these components.

Figure 13 shows the mean absolute SHAP values for each feature to gauge importance. PC1 is the most influential feature by a wide margin. PC3 also carries meaningful incremental importance, exceeding PC2, which is consistent with the earlier exploratory analysis. Overall, the attribution confirms that most predictive power sits in the "level" factor (PC1), with PC3 providing additional information.

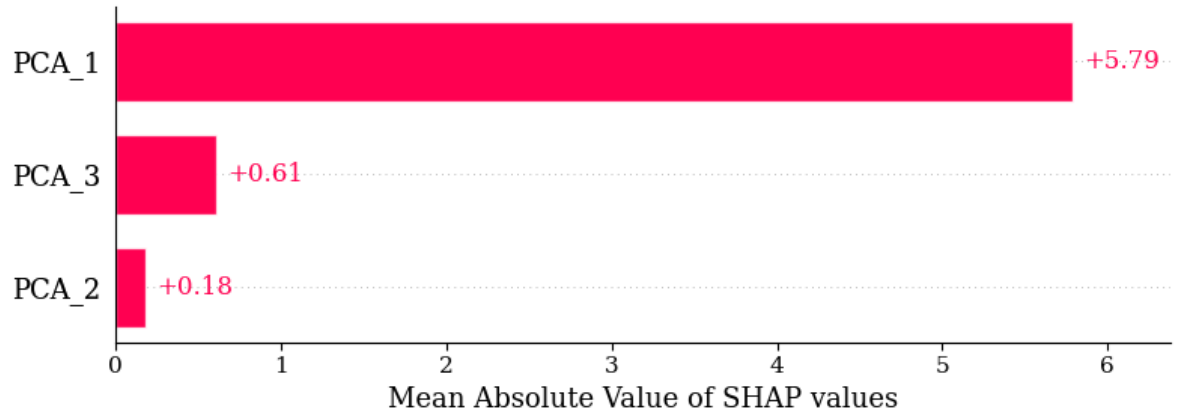

FIGURE 13: Mean SHAP values from Lasso Regression for $VXc1_{t+1}$

5.3. Economic Tests for Significance

TABLE 14: Trading Performance with Various Models

Model	Sharpe Ratio	Sharpe 95% CI	Fixed Trade Return Annualised
Linear Regression	2.120	1.493, 2.690	0.340
Ridge Regression	2.099	1.532, 2.665	0.339
Lasso Regression	2.129	1.528, 2.667	0.341
Random Forest	1.968	1.432, 2.567	0.354
Gradient Boosting	1.811	1.310, 2.433	0.349
Neural Network	1.075	0.411, 1.884	0.286
Nearest Neighbors	1.592	0.983, 2.165	0.315
LSTM	0.888	0.174, 1.621	0.267
ARIMA	-0.155	-0.767, 0.477	-0.139

Table 14 shows sharpe ratio, sharpe confidence interval and annualised return from a fixed trade size using the trading simulation for the machine learning models. Linear models again outperformed other models highlighting the notable linearity in the data.

Network models, such as LSTM and Neural Networks, performed notably poorly in the trading simulations, mirroring their cross validation performance. The tendency of networks to be over-trained, over-fit, or incorrectly tuned highlights that a strong fitting ability does not necessarily translate into profitable trading models [66][67]. Figure 14 shows the cumulative returns of trades over the simulation, highlighting the high variance of neural network simulation and the low performance of the LSTM model.

The Sharpe ratios highlight the economic significance of the principal components as predictors, especially when compared with the negative Sharpe ratio from the ARIMA simulation. This again suggests predictive value beyond that embedded in the VIX futures time series. Furthermore, the Sharpe ratios observed here exceed those previously reported in the literature, such as the 0.085 reported by Konstantinidi [27] and the 1.42 reported by Vrontos (based on an assumption of spot VIX tradability) [41]. It is also important to note the wide confidence intervals implying a degree of variability in returns, however this can be partially attributed to the comparatively large eight year test sample

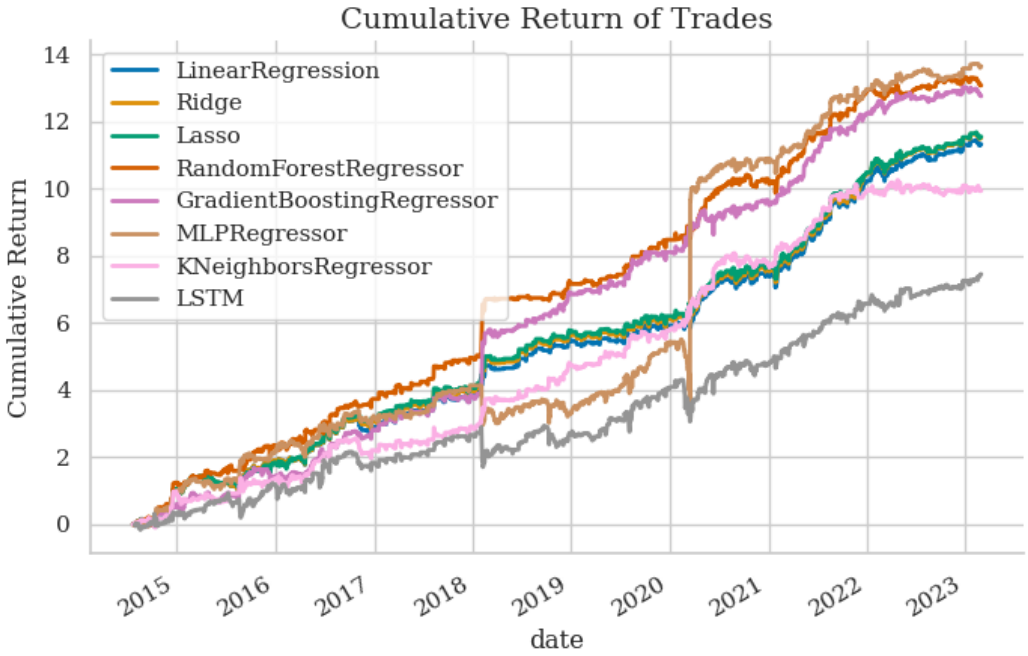

FIGURE 14: Trading Strategy Returns by Models

Feature Analysis

TABLE 15: Sharpe Ratios with Various Principal Component Feature Subsets

Model and Subset	PC1	PC1, PC2	PC1, PC3	PC1, PC2, PC3
Linear Regression	1.806	1.782	1.938	2.120
Ridge Regression	1.807	1.801	1.940	2.099
Lasso Regression	1.807	1.804	1.965	2.129
Random Forest	1.827	1.797	1.905	1.968
Gradient Boosting	1.576	1.719	1.874	1.811
Neural Network	1.565	1.779	2.079	1.075
Nearest Neighbors	1.802	1.617	1.662	1.592
LSTM	1.264	1.035	1.192	0.888

Table 13 provides the simulation sharpe ratio across feature subsets and models to analyse the marginal contribution of individual predictors. The analysis was conducted with PC1 as the baseline, given its emergence as the strongest single predictor, whereas the remaining components in isolation exhibited limited explanatory power for VIX futures. Across the models, the inclusion of PC2 generally reduced performance relative to PC1 alone, a result consistent with earlier exploratory analysis indicating weak correlation between PC2 and VIX futures. By contrast, the addition of PC3 yielded incremental improvements in most cases, again in line with prior findings.

With respect to model class, linear specifications attained their highest performance when all features were included. In contrast, more overfitting-prone models such as LSTMs and neural networks tended to perform better with more parsimonious feature sets. This pattern suggests that the predictive signal is largely linear and concentrated in PC1, while later components primarily introduce noise that complex models tend to overfit. Neural networks, in particular, had notable performance deterioration when all three principal components were used.

5.4. Auto-Encoder Comparisons

There is considerable variation in performance with autoencoded features compared with the principal components. Table 16 reports machine learning metrics RMSE, R^2 and a trading sharpe ratio using standard and masked autoencoded features for various machine learning models. Relative to the principal component features, the machine learning models and trading simulations achieve comparable yet inferior performance when based on autoencoded features. Between the two autoencoder architechures there is no clear standout in machine learning errors, however a clear trend of outperformance of masked autoencoded features over a standard autoencoder in trading performance. Overall, both standard and masked autoencoders provide comparable forecasting inputs to principal components when compressing the IV surface.

TABLE 16: One day Ahead Five Fold Expanding Cross Validation and Trading Simulation Results for autoencoded Features

	RMSE		R^2		Sharpe Ratio	
Model	Standard	Masked	Standard	Masked	Standard	Masked
Linear Regression	1.865	1.964	0.900	0.886	1.597	1.892
Ridge Regression	1.865	1.964	0.900	0.886	1.597	1.894
Lasso Regression	1.864	1.961	0.900	0.887	1.591	1.932
Random Forest	2.254	2.045	0.854	0.883	1.101	1.854
Gradient Boosting	2.270	2.188	0.863	0.869	1.212	1.571
Neural Network	2.050	2.354	0.872	0.837	1.508	1.760
Nearest Neighbors	2.041	1.968	0.879	0.890	2.011	1.774
LSTM	2.267	2.239	0.840	0.847	0.656	0.933

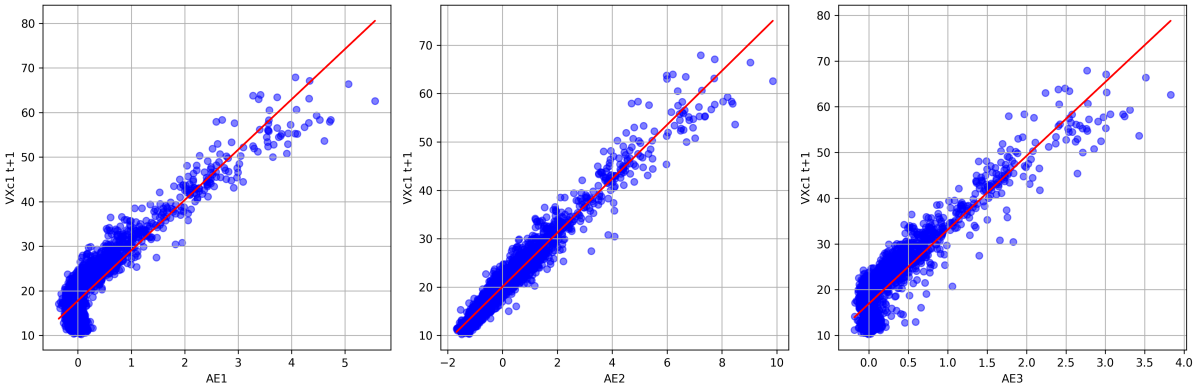

FIGURE 15: Scatter Plots of Masked Autoencoded Features against $VXc1_{t+1}$ in Training Set

Figure 15 shows scatter plots of the first three masked autoencoded features against front month VIX futures price. These plots highlight the analytical advantages of principal component analysis relative to autoencoders. PCA's linear form yields explicit loadings that can facilitate interpretation, while the orthogonality constraint ensures that components capture distinct sources of variation. In contrast, autoencoders function as black-box models, producing latent variables that are both difficult to interpret and can be highly correlated as seen in the plots. Variance inflation factor tests for the masked autoencoded features also confirm substantial multicollinearity (Appendix Table 21), raising concerns for both forecasting stability and interpretability. By contrast, principal components provide more transparent and differentiable outputs, enabling clearer variable significance and interpretations, properties which are gaining popularity in machine learning [43].

Figure 13 shows the mean absolute SHAP values for each of the MAE features to demonstrate importance. The second MAE feature was the most powerful feature, suggesting it potentially captures a

level factor analogous to PC1. Notably, the first and third features exhibit nearly identical importance, consistent with their close resemblance in Figure 15.

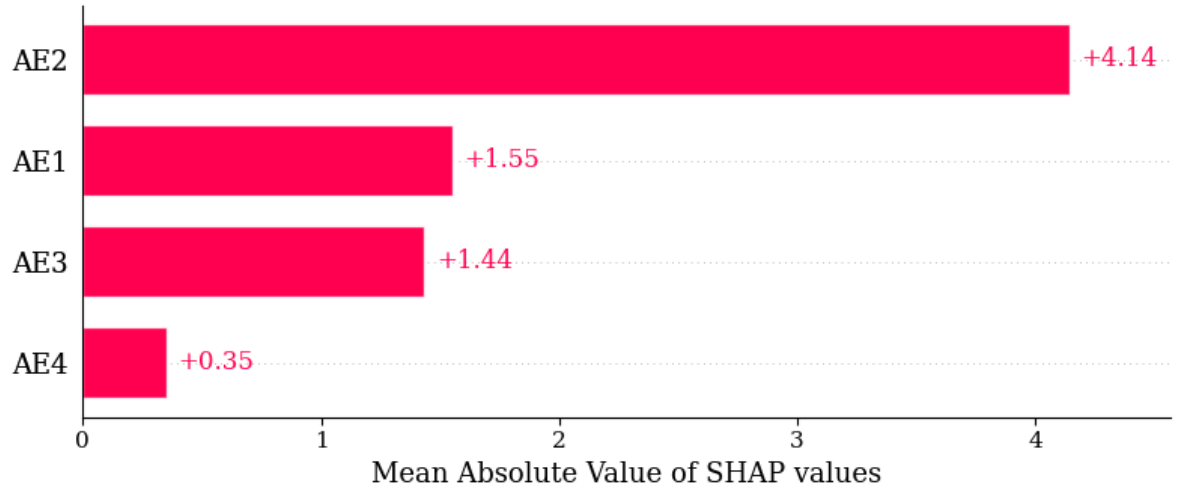

FIGURE 16: Mean SHAP values from Lasso Regression for $VXc1_{t+1}$

Table 17 reports correlation coefficients between the MAE features and the dependent variable. Consistent with the multicollinearity established earlier, all variables exhibit absolute correlation values above 90% with the dependent variable, raising questions about what each feature is actually capturing. This correlation pattern closely mirrors the SHAP analysis, demonstrating that AE2 has the strongest correlation with the dependent variable at 97.7%. Similarly, the closeness of AE1 and AE3 is shown with their correlation of 98.0%. This further illustrates how autoencoders can reduce interpretability, highlighting PCA's advantage in producing more distinct and interpretable factors.

TABLE 17: Correlation between Masked Latent Features and $VXc1_{t+1}$

	AE1	AE2	AE3	AE4	AE5	$VXc1_{t+1}$
AE1	1.000					
AE2	0.950	1.000				
AE3	0.980	0.959	1.000			
AE4	-0.938	-0.995	-0.951	1.000		
AE5	-0.954	-0.995	-0.966	0.996	1.000	
$VXc1_{t+1}$	0.916	0.977	0.919	-0.972	-0.967	1.000

6. Conclusion

This thesis examined the predictive value of the implied volatility surface of S&P 500 index options for forecasting VIX futures using dimensionality reduction and machine learning. PCA was shown to extract interpretable components that capture variation in the IV surface and provide economically meaningful signals. The first principal component (PC1) was particularly important, functioning as a broad level factor strongly linked to VIX futures prices, while the second and third components captured slope and stress-related dynamics respectively. Predictive models built on these components were effective compared to prior literature, producing outperformance in both machine learning fit and economic significance.

Autoencoders were explored as an alternative dimensionality reduction approach. They typically offered worse performance across machine learning and economic tests while lacking economic intuition. In comparison, PCA's variance-maximising structure produced more distinct and interpretable factors.

Several limitations should be acknowledged. First, the dataset focused on a single market, asset class and frequency, hence results may not generalise to other contexts. Second, while a variety of dimensionality reduction and machine learning techniques were used, this scope could be widened. Finally, the trading strategy design was deliberately simple to isolate model performance and this could be extended further.

Future research could extend these findings in several directions. Applying the methodology across different asset classes or volatility indices would provide insight into its robustness. Similarly, a wider exploration of modern machine learning architectures could be used, including more advanced network forecasting models and dimensionality reduction techniques. Finally, extensions to the trading strategy in its complexity and testing its inclusion within an equity portfolio would advance its practical applications.

Overall, this work demonstrates that factor extraction from the IV surface can generate statistically and economically significant signals for forecasting VIX futures. PCA produced parsimonious and interpretable representations of the high-dimensional surface, which, when combined with linear models, yielded strong predictive performance. Autoencoders achieved comparable forecasting results with less interpretable and collinear features which could be problematic in both attribution and forecasting.

Bibliography

[1] S.-H. Poon and C. W. J. Granger, "Forecasting volatility in financial markets: A review," *Journal of Economic Literature*, vol. 41, no. 2, pp. 478–539, 2003.

[2] R. E. Whaley, "Derivatives on market volatility," *The journal of Derivatives*, vol. 1, no. 1, pp. 71–84, 1993.

[3] A. Clements, J. Fuller, et al., "Forecasting increases in the VIX: A time-varying long volatility hedge for equities," *NCER Working Paper Series 88*, 2012.

[4] B. Mandelbrot, "The variation of certain speculative prices," *The Journal of Business*, vol. 36, no. 4, pp. 394–419, Oct. 1963.

[5] P. Carr and L. Wu, "A tale of two indices," *Available at SSRN 871729*, 2005.

[6] "Volatility Index Methodology: Cboe Volatility Index®," Cboe Global Markets, Tech. Rep., 2024. [Online]. Available: `https://cdn.cboe.com/api/global/us_indices/governance/Volatility_Index_Methodology_Cboe_Volatility_Index.pdf`.

[7] E. S. Gunnarsson, H. R. Isern, A. Kaloudis, M. Risstad, B. Vigdel, and S. Westgaard, "Prediction of realized volatility and implied volatility indices using ai and machine learning: A review," *International review of financial analysis*, p. 103 221, 2024.

[8] J. C. Hull and S. Basu, *Options, futures, and other derivatives*. Pearson Education India, 2016.

[9] M. McAleer and M. C. Medeiros, "Realized volatility: A review," *Econometric reviews*, vol. 27, no. 1-3, pp. 10–45, 2008.

[10] F. Black and M. Scholes, "The pricing of options and corporate liabilities," *Journal of political economy*, vol. 81, no. 3, pp. 637–654, 1973.

[11] K. Demeterfi, E. Derman, M. Kamal, and J. Zou, "More than you ever wanted to know about volatility swaps," *Goldman Sachs Quantitative Strategies Research Notes*, Mar. 1999.

[12] R. E. Whaley, "Understanding the VIX," *Journal of Portfolio Management*, vol. 35, no. 3, pp. 98–105, 2009.

[13] R. E. Whaley, "The investor fear gauge," *Journal of portfolio management*, vol. 26, no. 3, p. 12, 2000.

[14] T. G. Andersen, O. Bondarenko, and M. T. Gonzalez-Perez, "Exploring return dynamics via corridor implied volatility," *The Review of Financial Studies*, vol. 28, no. 10, pp. 2902–2945, 2015.

[15] E. Szado, "VIX futures and options: A case study of portfolio diversification during the 2008 financial crisis," *The Journal of Alternative Investments*, vol. 12, no. 2, p. 68, 2009.

[16] J. M. Griffin and A. Shams, "Manipulation in the VIX?" *The Review of Financial Studies*, vol. 31, no. 4, pp. 1377–1417, 2018. DOI: 10.1093/rfs/hhx085.

[17] J. M. Keynes, *A Treatise on Money*. London: Macmillan Press, 1930.

[18] T. L. Johnson, "Risk premia and the VIX term structure," *Journal of Financial and Quantitative Analysis*, vol. 52, no. 6, pp. 2461–2490, 2017.

[19] D. P. Simon and J. Campasano, *The VIX futures basis: Evidence and trading strategies*. SSRN, 2014.

[20] P. Carr and L. Wu, "Variance risk premiums," *The Review of Financial Studies*, vol. 22, no. 3, pp. 1311–1341, 2009.

[21] I. Dew-Becker, S. Giglio, A. Le, and M. Rodriguez, "The price of variance risk," *Journal of Financial Economics*, vol. 123, no. 2, pp. 225–250, 2017.

[22] T. Bollerslev, G. Tauchen, and H. Zhou, "Expected stock returns and variance risk premia," *The Review of Financial Studies*, vol. 22, no. 11, pp. 4463–4492, 2009.

[23] M. Nossman and A. Wilhelmsson, "Is the VIX futures market able to predict the VIX index? a test of the expectation hypothesis," *The Journal of Alternative Investments*, vol. 12, no. 2, p. 54, 2009.

[24] I.-H. Cheng, "The VIX premium," *The Review of Financial Studies*, vol. 32, no. 1, pp. 180–227, 2019.

[25] J. Mencia and E. Sentana, "Valuation of VIX Derivatives," *Journal of Financial Economics*, vol. 108, no. 2, pp. 367–391, 2013. DOI: 10.1016/j.jfineco.2012.11.008.

[26] K. Ahoniemi, "Modeling and forecasting implied volatility: An econometric analysis of the VIX index," *Helsinki: Helsinki Center of Economic Research*, 2006.

[27] E. Konstantinidi, G. Skiadopoulos, and E. Tzagkaraki, "Can the evolution of implied volatility be forecasted? evidence from european and us implied volatility indices," *Journal of Banking & Finance*, vol. 32, no. 11, pp. 2401–2411, 2008.

[28] J. Hosker, S. Djurdjevic, H. Nguyen, and R. Slater, "Improving VIX futures forecasts using machine learning methods," *SMU Data Science Review*, vol. 1, no. 4, 2018, https://scholar.smu.edu/datasciencereview/vol1/iss4/6.

[29] E. Konstantinidi and G. Skiadopoulos, "Are VIX futures prices predictable? an empirical investigation," *International Journal of Forecasting*, vol. 27, no. 2, pp. 543–560, 2011.

[30] T. Bollerslev, "Generalized autoregressive conditional heteroskedasticity," *Journal of econometrics*, vol. 31, no. 3, pp. 307–327, 1986.

[31] T. Wang, Y. Shen, Y. Jiang, and Z. Huang, "Pricing the cboe VIX futures with the heston–nandi garch model," *Journal of Futures Markets*, vol. 37, no. 7, pp. 641–659, 2017.

[32] S. Guo and Q. Liu, "Efficient out-of-sample pricing of VIX futures," *Journal of Derivatives*, vol. 27, no. 3, pp. 126–139, 2020.

[33] J. E. Zhang and Y. Zhu, "VIX futures," *Journal of Futures Markets: Futures, Options, and Other Derivative Products*, vol. 26, no. 6, pp. 521–531, 2006.

[34] Y. Zhu and J. E. Zhang, "Variance term structure and VIX futures pricing," *International Journal of Theoretical and Applied Finance*, vol. 10, no. 01, pp. 111–127, 2007.

[35] G. Dotsis, D. Psychoyios, and G. Skiadopoulos, "An empirical comparison of continuous-time models of implied volatility indices," *Journal of Banking & Finance*, vol. 31, no. 12, pp. 3584–3603, 2007.

[36] S. A. Degiannakis, "Forecasting VIX," *Journal of Money, Investment and Banking*, no. 4, 2008.

[37] K. Christensen, M. Siggaard, and B. Veliyev, "A machine learning approach to volatility forecasting," *Journal of Financial Econometrics*, vol. 21, no. 5, pp. 1680–1727, 2023.

[38] A. Hirsa et al., "The VIX index under scrutiny of machine learning techniques and neural networks," 2021, https://arxiv.org/abs/2102.02119.

[39] J. Osterrieder, D. Kucharczyk, S. Rudolf, et al., "Neural networks and arbitrage in the VIX," *Digital Finance*, vol. 2, pp. 97–115, 2020. DOI: 10.1007/s42521-020-00026-y.

[40] K. Berahmand, F. Daneshfar, E. S. Salehi, Y. Li, and Y. Xu, "Autoencoders and their applications in machine learning: A survey," *Artificial intelligence review*, vol. 57, no. 2, p. 28, 2024.

[41] S. D. Vrontos, J. Galakis, and I. D. Vrontos, "Implied volatility directional forecasting: A machine learning approach," *Quantitative Finance*, vol. 21, no. 10, pp. 1687–1706, 2021. DOI: 10.1080/14697688.2021.1929418.

[42] L. V. Ballestra, A. Guizzardi, and F. Palladini, "Forecasting and trading on the VIX futures market: A neural network approach based on open to close returns and coincident indicators," *International Journal of Forecasting*, vol. 35, no. 4, pp. 1250–1262, 2019.

[43] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," *IEEE access*, vol. 6, pp. 52 138–52 160, 2018.

[44] H. A. Latane and R. J. Rendleman, "Standard deviations of stock price ratios implied in option prices," *The Journal of Finance*, vol. 31, no. 2, pp. 369–381, 1976.

[45] B. J. Blair, S.-H. Poon, and S. J. Taylor, "Forecasting s&p 100 volatility: The incremental information content of implied volatilities and high-frequency index returns," *Journal of econometrics*, vol. 105, no. 1, pp. 5–26, 2001.

[46] E. Derman and I. Kani, "Riding on a smile," *Risk*, vol. 7, no. 2, pp. 32–39, 1994.

[47] S. Yan, "Jump risk, stock returns, and slope of implied volatility smile," *Journal of Financial Economics*, vol. 99, no. 1, pp. 216–233, 2011.

[48] Y. Xing, X. Zhang, and R. Zhao, "What does the individual option volatility smirk tell us about future equity returns?" *Journal of Financial and Quantitative Analysis*, vol. 45, no. 3, pp. 641–662, 2010.

[49] J. S. Doran, D. R. Peterson, and B. C. Tarrant, "Is there information in the volatility skew?" *Journal of Futures Markets: Futures, Options, and Other Derivative Products*, vol. 27, no. 10, pp. 921–959, 2007.

[50] S. Chakravarty, H. Gulen, and S. Mayhew, "Informed trading in stock and option markets," *The Journal of Finance*, vol. 59, no. 3, pp. 1235–1257, 2004.

[51] J. Pan and A. M. Poteshman, "The information in option volume for future stock prices," *The Review of Financial Studies*, vol. 19, no. 3, pp. 871–908, 2006.

[52] D. S. Bates, "The crash of 87: Was it expected? the evidence from options markets," *The journal of finance*, vol. 46, no. 3, pp. 1009–1044, 1991.

[53] G. Skiadopoulos, S. Hodges, and L. Clewlow, "The dynamics of the s&p 500 implied volatility surface," *Review of derivatives research*, vol. 3, no. 3, pp. 263–282, 2000.

[54] P. Christoffersen, M. Fournier, and K. Jacobs, "The factor structure in equity options," *The Review of Financial Studies*, vol. 31, no. 2, pp. 595–637, 2018.

[55] R. Litterman, "Common factors affecting bond returns," *Journal of fixed income*, pp. 54–61, 1991.

[56] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley interdisciplinary reviews: computational statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[57] J. H. Stock and M. W. Watson, "Forecasting using principal components from a large number of predictors," *Journal of the American statistical association*, vol. 97, no. 460, pp. 1167–1179, 2002.

[58] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019, pp. 4171–4186.

[59] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 000–16 009.

[60] LSEG DataScope, *Vix futures data*, Data retrieved from LSEG DataScope Select (formerly Refinitiv DataScope)., 2025. [Online]. Available: `https://select.datascope.refinitiv.com/datascope/`.

[61] OptionMetrics, *S&p 500 (spx) option quotes data*, Data retrieved from OptionMetrics database., 2025. [Online]. Available: `https://optionmetrics.com/`.

[62] Yahoo Finance, *S&p 500 index data*, Data retrieved from Yahoo Finance., 2025. [Online]. Available: `https://au.finance.yahoo.com/quote/%5EGSPC/`.

[63] M. R. Fengler, W. K. Härdle, and C. Villa, "The dynamics of implied volatilities: A common principal components approach," *Review of Derivatives Research*, vol. 6, no. 3, pp. 179–202, 2003.

[64] X. Yang, P. Wang, and J. Chen, "Vix futures pricing with affine jump-garch dynamics and variance-dependent pricing kernels," *Journal of Derivatives*, vol. 27, no. 1, pp. 110–127, 2019.

[65] G. Qiao, G. Jiang, and J. Yang, "Vix term structure forecasting: New evidence based on the realized semi-variances," *International Review of Financial Analysis*, vol. 82, p. 102 199, 2022.

[66] O. B. Sezer, M. U. Gudelek, and A. M. Ozbayoglu, "Financial time series forecasting with deep learning: A systematic literature review: 2005–2019," *Applied soft computing*, vol. 90, p. 106 181, 2020.

[67] J. R. Coakley and C. E. Brown, "Artificial neural networks in accounting and finance: Modeling issues," *International Journal of Intelligent Systems in Accounting, Finance and Management*, vol. 9, no. 2, pp. 119–144, 2000.

7. Additional Figures

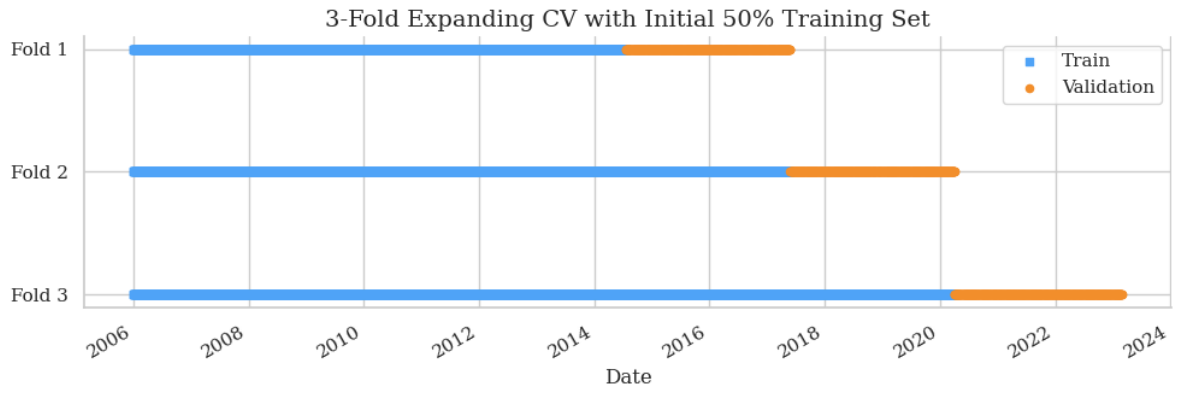

FIGURE 17: Trading Simulation Time Series Validation Framework

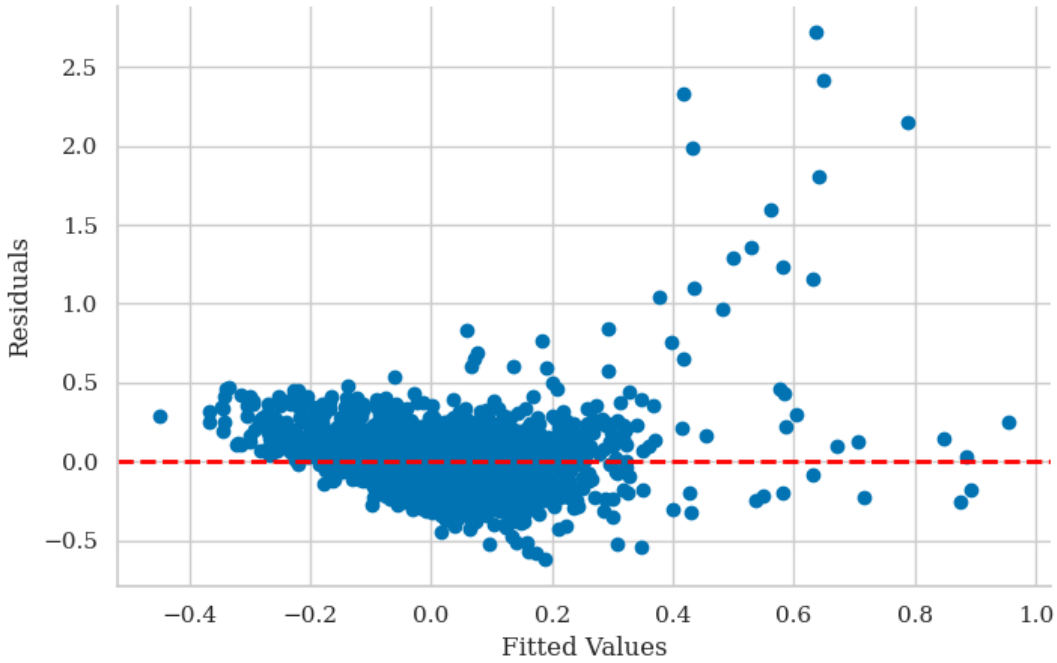

FIGURE 18: Fitted Values and Residuals Plot for PC3 and One Month VXc1 Return

[4]This utilises heteroskedasticity-robust standard errors (HC3), see Residual Plot at Figure 19.

TABLE 18: Johansen Cointegration Test

Rank	Trace Stat	Crit 90%	Crit 95%	Crit 99%
0	32.466	10.474	12.321	16.364
1	1.495	2.976	4.130	6.941

TABLE 19: OLS Regression of VXc1 on PC1

Variable	Coefficient	Std. Error	t-statistic	p-value
Intercept	20.364	0.023	878.570	0.000
PC1	30.055	0.229	131.418	0.000
R-squared		0.967		
Adjusted R-squared		0.967		
F-statistic		1.727e+04		
Prob (F-statistic)		0.000		

TABLE 20: OLS Regression of One-Day ahead VIX Front-Month Futures on Principal Components on Full Sample [4]

Variable	Coefficient	Std. Error	z-statistic	p-value
Intercept	20.369	0.026	770.373	0.000
PC1	29.725	0.249	119.429	0.000
PC2	-10.257	0.784	-13.076	0.000
PC3	-20.188	1.449	-13.936	0.000
R-squared		0.957		
Adjusted R-squared		0.957		
F-statistic		8968		
Prob (F-statistic)		0.000		

TABLE 21: Variance Inflation Factor for Masked Autoencoder Features in Training Set

Latent Feature	VIF standard	VIF masked
AE1	1.651	28.670
AE2	1.606	124.769
AE3	1.271	37.055
AE4	1.675	230.224
AE5		268.707

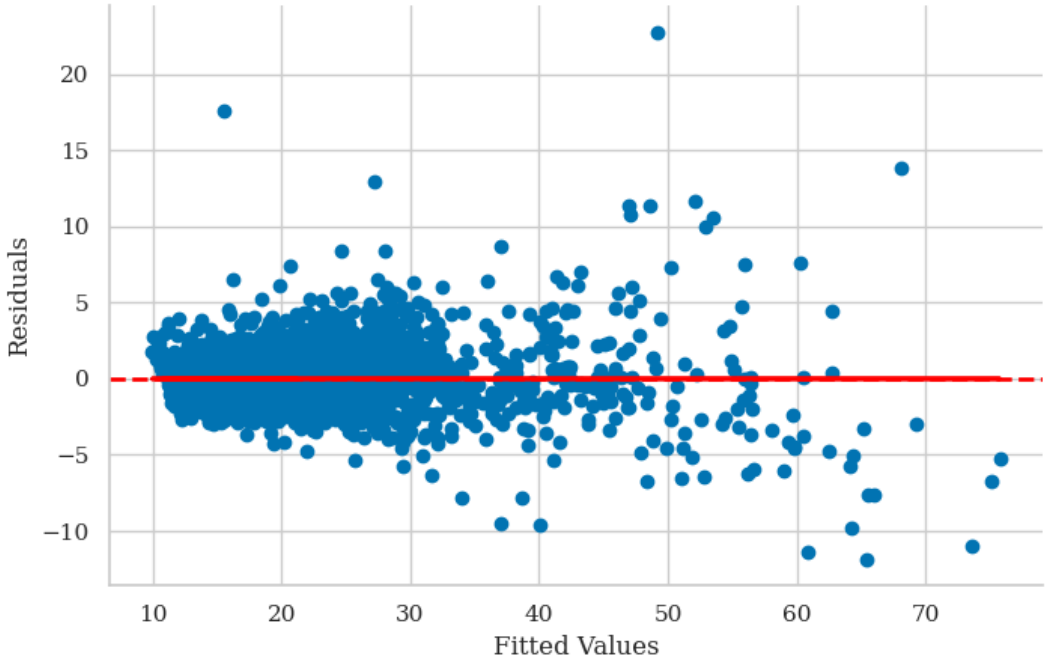

FIGURE 19: Fitted Values and Residuals Plot for PC1, PC2, PC3 and $VXc1_{t+1}$