Publication Date: 31 July 2024
Archs Sci. (2024) Volume 74, Issue 4 Pages 170-177, Paper ID 2024422.
https://doi.org/10.62227/as/74422

Financial Option Volatility Prediction Based on Machine Learning Algorithm

Bo Yu[1,*]
[1]School of Economics and Management, Guangdong Polytechnic Insititute, Guangzhou, Guangdong, 510091, China.
Corresponding authors: (e-mail: byu@gdpi.edu.cn).

Abstract Financial risks are increasing worldwide, thus finding suitable tools for financial risk management becomes particularly important. In this paper, the realized volatility and implied volatility of financial option data are calculated through the Black-Scholes option pricing model. Through the support vector machine in machine algorithm, the GARCH model is improved, the two-stage prediction method of GARCH-SVR model is constructed, and the volatility metric is set to examine the prediction effect of the model. Finally, a machine learning model is used to predict the implied volatility of CSI 300 ETF options to realize the prediction of volatility. The results show that the return of the arbitrage strategy constructed by the GARCH-SVR model, the maximum of which is 0.1858. Comparison with the other four models reveals that the machine learning predicts financial option volatility, which can bring better economic returns in the real market.

Index Terms black-scholes, financial options, implied volatility, machine algorithms, GARCH-SVR models

I. Introduction

Options are important derivative instruments in financial engineering. As an institutional trader, the prominent trading strategy in designing strategies for trading options is the sell category [1]–[3]. However, one-way selling options carry significant trading risks as do one-way buying options. Dynamic hedging is required to capture the gains of robust buy and sell class option strategies. How to consider the dynamic continuity and forward-looking nature of hedging becomes the focus of risk management [4]–[6]. Market volatility is an important variable in determining option prices, however, facts and studies have shown that option volatility is not static and it is stochastic [7]–[9]. The unpredictability of volatility means that it is difficult to find the right volatility to price an option. As a result, the forecasting of option volatility becomes a very important task in order to capture the trend of option price changes and the dynamics and foresight of hedging [10-11]. For example, a large portion of the risk in the strategy of selling options comes from a large increase in implied volatility, so if we can predict the increase in implied volatility in advance, we can reduce or avoid the risk of volatility increase by adjusting the hedging position.

There is an urgent need for new methodologies and models for volatility forecasting. In recent years, with the increasing maturity of big data, artificial intelligence, and machine learning technologies, new technologies can be utilized to achieve the prediction of volatility [12]–[14]. Big data is a new technology processing mode, with stronger decision-making power, insight and process optimization capabilities of the massive, high growth rate and diversified information assets, T + 0 trading options in the annual, monthly, weekly, daily, second-degree data at different levels, different depths of data, to meet the data "big" standard. The "artificial intelligence" has been 60 years since its introduction, is a research, development for simulation, extension and expansion of human intelligence theory, methodology, technology and application systems of technical sciences, specific research, including robotics, language recognition, image recognition, natural language processing and expert systems, etc., the core of the research is machine learning [15]. Machine learning design and analysis allows computers to automatically "learn" algorithms, which is exactly what can be applied to volatility forecasting in option strategies.

A two-stage superimposed integrated stock market direction prediction model based on machine learning, empirical modal decomposition and XAI was proposed in literature [16]. It is concluded that the prediction model with locally interpretable model agnostic interpretation support achieves the highest accuracy of 0.9913 when only the most plausible predictions are considered on the KOSPI dataset. Literature [17] enables investors and regulators to expect to predict future bank failures and other financial variables of interest by using stable statistical forecasts. The study reports two successful machine learning methods for predicting bank size one fiscal year prior to the current date and demonstrates that these models are successful. Literature [18] uses three

typical test functions to compare the performance of MLIA prediction algorithms with logistic prediction algorithms. The study shows that machine learning has a good predictive effect on MLIA financial credit risk prediction, which can provide theoretical references for subsequent related research. Literature [19] developed an innovative complex network approach to model interbank networks with systemic risk contagion, using machine learning techniques to identify comprehensive features of the network. The results show that the market factors of interbank networks have a significant impact on risk diffusion, providing a scientific approach for policy makers to choose a response to systemic risk. Literature [20] developed a smart return prediction method for blockchain financial products using deep learning, including the design of a long and short-term memory model for return prediction analysis. The simulation results highlight the advantages of the financial product smart return prediction methodology technique over the current state-of-the-art in terms of various evaluation parameters. Literature [21] compares the forecasting results of two proposed functionally linked artificial neural networks based on low computational complexity with traditional intelligent methods. Simulation-based experimental results show that the predictive performance of the proposed distributed predictor is similar or improved compared to the traditional distributed predictor. In addition, the method saves bandwidth, memory and power consumption. Literature [22] mainly emphasizes on accurate prediction of financial markets, where the main motivation for stock price prediction is to minimize the significant losses faced by investors and to analyze the profitability through the amount of buying and selling. Simulation results show that the proposed deep learning based technique outperforms other models in future prediction irrespective of the financial market. Literature [23] provides a new financial trading strategy system that improves the prediction of stock prices by introducing the optical gradient enhancer algorithm into stock price prediction and constructing the minimum variance portfolio of mean-variance model with conditional value-at-risk constraints. Literature [24] improves the deep neural network algorithm to predict the price of Bitcoin so as to achieve the main objective of reducing the financial risk for e-commerce, which opens up horizons for the development of e-businesses using digital currencies. The method achieved good results in terms of accuracy (53.4%) and correctness of prediction (MSE 1.02), offering prospects for other research in this field.

In this paper, firstly, using the Black-Scholes option pricing model and market data, historical, realized and implied volatilities are obtained by backpropagation and input as three eigenvalues into a machine learning algorithm to be used as model training. Secondly, the GARCH model is improved by using the support vector machine in the machine learning algorithm, and a GARCH-SVR model combining the SVR model and the GARCH model is proposed to predict the return volatility of financial options. Finally, the closing price data of CSI 300 ETF options is selected as the option price history data. Another decision tree, gradient boosting tree, support vector machine, and convolutional neural network are selected to compare the performance and return of option implied volatility prediction under four models, which are used to explore the prediction effect of GARCH-SVR model.

II. Modeling of Financial Option Volatility Forecasts

A. Financial Options and Pricing Models

1) Financial options

A financial option is an option that gives the purchaser the right to buy or sell an asset at a fixed price in a future period. The initial payment made by the purchaser to the seller is called the royalty, which is the price of the option, and consists of two components, the intrinsic value of the option and the time value. The intrinsic value of a financial option refers to the profit that the seller makes by selling the option immediately, while the time value is mainly measured by the change in volatility of the price of the underlying option over time. The price at which the asset is purchased or sold is called, the exercise price of the option.

Options are categorized from the perspective of purchasing or selling the underlying asset of a financial option, and are divided into call and put options. A call option will allow the purchaser to buy the underlying asset in the future, while a put option will allow the purchaser to sell the underlying asset in the future, but of course these rights do not have to be executed at the expiration date. If options are categorized by when they can be executed, they can be divided into European-style options and American-style options. European style options require the purchaser to execute the right only at a specific point in time in the future, whereas American style options can be executed at any point in time within a specified period of time.

2) The Black-Scholes option pricing model

1) Assume that the subject of the contract is tradable.
2) Assume that the underlying contract does not pay dividends or has no storage costs.
3) Assume the underlying contract is shortable.
4) Assume a single constant interest rate.
5) Assume that there are no taxes.
6) Any quantity of the underlying contract can be traded and the change in the price of the underlying contract is continuous.
7) There are no fees for trading the underlying contracts.
8) Volatility is constant and volatility is the only parameter that describes the distribution of returns on the underlying contract. The B-S pricing formula for European call and put options on non-dividend paying stocks can be expressed as follows:

$$C = S_0 N(d_1) - X e^{-rt} N(d_2) \tag{1}$$

$$P = X e^{-rt} N(-d_2) - S_0 N(-d_1), \tag{2}$$

ARCHIVES DES SCIENCES
A Multidisciplinary Journal

where d_1 and d_2 are respectively:

$$d_1 = \frac{\ln\left(\frac{S_0}{K}\right) + \left(r + \frac{1}{2}\sigma^2\right)t}{\sigma\sqrt{t}} \tag{3}$$

and

$$d_2 = d_1 - \sigma\sqrt{t}, \tag{4}$$

where C and P are the prices of call and put options, respectively, $N(x)$ is the cumulative probability distribution function of a normally distributed variable, and S_0 is the initial stock price. X is the strike price, r is the risk-free interest rate, σ is the stock price volatility expressed as an annual standard deviation, and t is the remaining term to expiration of the option contract expressed as an annual standard deviation.

The B-S model does not include dividends, which usually have the effect of lowering the call option price, and can be extended to account for dividend payouts by using instead S_0. In S_0e^{-qt}, q is the continuous dividend payout output:

$$C = S_0e^{-qt}N\left(d_1\right) - Xe^{-rt}N\left(d_2\right). \tag{5}$$

$$P = Xe^{-rt}N\left(-d_2\right) - S_0e^{-qt}N\left(-d_1\right). \tag{6}$$

Similarly, it is possible to lay out an equation that would relate the price of a put option to the price of a call option:

$$C + Xe^{-n} = P + S. \tag{7}$$

The relational equation with dividend payout is:

$$C + Xe^{-rt} = P + S_0e^{-qt}. \tag{8}$$

The above equation is known as the parity relationship between the buy and sell options, and deviations from the parity of the buy and sell options create arbitrage opportunities that usually disappear quickly.

From the above analysis, it is known that every parameter used to price an option has an exact known value except for volatility. Since the value of an option contract is directly dependent on volatility, accurately estimating volatility is a critical skill for option traders as well as financial market regulators. In addition, changes in volatility can be significantly amplified in the price of an option contract, and a doubling of volatility can result in a multi-fold increase in the price of an option.

B. Classification of financial volatility

From the perspective of volatility categorization, this section introduces several common types of volatility and briefly outlines the definitions and calculation methods of different types of volatility. The comparison of different categories of volatility helps us to understand the similarities and differences between different categories of volatility, thus helping us to deepen our understanding of the concept of volatility, which in turn facilitates our volatility modeling and forecasting.

1) Historical volatility

Historical volatility is a measure of the volatility of an underlying asset over a certain period of time. It is usually expressed as the standard deviation of historical log returns:

$$R_i = \ln\left(P_i - P_{i-1}\right), \tag{9}$$

$$\sigma_t^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(R_i - \bar{R}\right), \tag{10}$$

where P_i is the price on the ind trading day, R_i is the historical return, N is the observation period given before the tth day, and $\bar{R}$ is the average of the historical returns on the Nth day. Historical volatility portrays the volatility of returns over a certain period of time, but does not take into account the intraday information of the price series.

2) Realized volatility

Financial high-frequency data obtained at shorter time intervals in accordance with frequency is a hot topic of research in recent years compared to traditional low-frequency observations. One of the commonly used metrics is the realized volatility, i.e:

$$RV_t = \sum_{i=1}^{N(\Delta)}\left(\ln P_{t,i+1} - \ln P_{t,i}\right)^2, \tag{11}$$

where $N(\Delta)$ is the number of high-frequency trading prices at a sampling interval of Δ a day, and $P_{t,i}$ is the trading price of the financial asset at the $i\Delta$th moment of the tth day. Under certain conditions, when the sampling interval Δ tends to 0, RV will converge to the true integral volatility, which portrays the overall situation of price changes on that day. Classical time series models need to utilize historical data when estimating day t volatility, and also require the assumption that the data is smooth over time. In contrast, the measure of volatility used in RV utilizes only intraday high-frequency data for that day and does not require the assumption that the intraday series data is smooth. In a sense, RV gives visibility to unobservable volatility. For empirical analysis, realized volatility is typically calculated using a 5-minute sampling interval.

3) Implied volatility

Implied volatility is obtained by inverting the Black-Scholes option pricing formula. Implied volatility reflects market participants' expectations of current market volatility. The higher the implied volatility, the higher the expected market volatility increases, and the higher the price of the option, provided that the other parameters of the option pricing formula remain unchanged. Conversely, the lower the implied volatility, the lower the expected market volatility decreases, and the lower the price of the option.The BS option pricing formula is:

$$c = S_tN\left(d_1\right) - Ke^{-r(T-t)}N\left(d_2\right), \tag{12}$$

where

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

ARCHIVES DES SCIENCES
A Multidisciplinary Journal

and

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

denotes the option price, S_t denotes the price of the underlying asset at t, K is the strike price of the option, T is the time to expiration, t is the current moment, and r is the risk-free rate of return. The implied volatility can be calculated by Newton's iterative method or interpolation given the option price and other known parameters.

Since there may be options with different strike prices and different maturity dates for an underlying asset, the liquidity of each option is different. As a result, this paper uses the implied volatility index (VIX) to characterize the volatility of the two markets.The difference between the VIX index and the implied volatility model for individual options is that it does not use the Black-Scholes model. Instead, it is obtained based on the variance swap, and in this paper, the VIX is used as a measure of implied volatility, and the specific calculation steps are as follows:

1) Calculate near-month and next-month volatilities with an option volatility index rollover of 7 days. Meet the remaining days to expiration of more than 7 days of the most recent expiration contract for the near-month contract, the next closest expiration contract for the second near-month contract, the two implied volatility for the near-month and the second near-month volatility, respectively. Near-month volatility is calculated as:

$$\sigma_1^2 = \frac{2}{T}\sum_i \frac{\Delta K_i}{K_i^2} e^{RT} P(K_i) - \frac{1}{T}\left[\frac{F}{K_0} - 1\right]^2. \quad (13)$$

Among them, T represents the remaining expiration time of the contract, and R represents the risk-free rate adopted by the SSE. K represents the strike price with the smallest difference between the call option and put option price, F represents the price of the theoretical forward, and $F = K + e^{RT}[P^c - P^p]$. P^c is the call option price, P^p is the put option price, and K_0 represents the option strike price that is less than F and close to F. K_i represents the strike price sorted from small to large $(i = 1, 2, 3, \ldots)$, and ΔK_i represents the strike price interval corresponding to the ith strike price, which is generally $(K_{i+1} - K_{i-1})/2$. $P(K_i)$ means that if K_i is less than K_0, then the put option price corresponding to K_i is taken. If K_i is greater than K_0, take K_0 as the call option price corresponding to K_i. If K_i equals K_0, then the average of the call and put options corresponding to K_i is taken.

2) After completing the calculation of the near-month volatility σ_1 and the next-nearest-month volatility σ_2, the following formula is used to calculate the volatility index of the option:

$$iVIX = 100\times \sqrt{\left\{T_1\sigma_1^2\left[\frac{NT_2 - NT_{30}}{NT_2 - NT_1}\right] + T_2{\sigma_2}^2\left[\frac{NT_{30} - NT_1}{NT_2 - NT_1}\right]\right\}\times\frac{N_{365}}{N_{30}}}, \quad (14)$$

where NT_1 and NT_2 denote the number of days to expiration for near-month and sub-near-month, and NT_{30} denotes the number of days in a month. If the number of days to expiration of the near-month contract is not less than 30 days, the sub-near-month volatility is not used, and $iVIX$ is the near-month volatility multiplied by 100.

C. GARCH- SVR Based Volatility Forecasting Models

1) GARCH model

The GARCH model is an improved model based on the autoregressive conditional heteroskedasticity (ARCH) model, which embodies asymmetry and facilitates the description of financial price fluctuations.

The GARCH model is defined as follows:

$$\begin{cases} y_t = \mu_t + \varepsilon_t, \\ \varepsilon_t | I_{t-1} \sim N\left(0, \sigma_t^2\right), \\ \sigma_t^2 = \alpha_0 + \sum_{i=1}^{q}\alpha_i\varepsilon_{t-i}^2 + \sum_{j=1}^{p}\beta_j\sigma_{t-j}^2 \\ = \alpha_0 + \alpha(L)\varepsilon_t^2 + \beta(L)\sigma_t^2, \end{cases} \quad (15)$$

Of these, $p{\sim}0, q{\sim}0, \alpha_0 > 0, \alpha_i{\sim}0(i = 1, \cdots, q), \beta_j{\sim}0(j = 1, \cdots, p)$.

Since the GARCH model has some limitations and drawbacks, it needs some improvements.The shortcomings of the GARCH model are as follows:

1) The model ignores non-negativity conditions in its estimation.
2) Although it accounts for thick-tail effects and volatility clustering, it does not account for leverage effects.
3) It fails to establish a direct relationship between conditional mean and conditional variance.
4) It does not consider the asymmetry of volatility.

In this paper, through the support vector machine (SVR) in machine algorithms, the GARCH model is improved even further, and a stage prediction method combining the SVR algorithm and the GARCH-like model is established to enhance the prediction accuracy of the original GARCH model, and the volatility metric is set up to examine the prediction effect of the model.

2) GARCH- SVR models

The $GARCH(1,1)$ model provides a simple representation of the main statistical features of the return series y_t of various assets, and therefore it is widely used to model real financial time series. If y_t follows the $GARCH(1,1)$ model, then:

$$\begin{cases} y_t = \mu + u_t, \\ \sigma_t^2 = \omega + \alpha u_{t-1}^2 + \beta\sigma_{t-1}^2, \end{cases} \quad (16)$$

where $u_t = \sigma_t\varepsilon_t$. As defined above, the conditional variance σ_t is a stochastic process assumed to be a constant plus a weighted average of the predicted value σ_{t-1}^2 for one period and the squared value u_{t-1}^2 of the observations for the previous period. Parameters ω, α and β must satisfy $\omega > 0$, $\alpha \geq 0, \beta \geq 0$ to ensure that the conditional variance is

ARCHIVES DES SCIENCES
A Multidisciplinary Journal

positive. In order for process y_t not to degenerate, parameter ω must be strictly positive. When $\alpha + \beta < 1$, process y_t is smooth.

Taking advantage of nonlinear regression estimation, the GARCH model parameters (GARCH- SVR) are estimated by SVR model instead of ML method. The specific framework is as follows:

$$y_t = f\left(y_{t-1}\right) + u_t, \tag{17}$$

where f is the decision function of the mean equation estimated from SVR. The squared residuals u_t are obtained from the conditional mean estimation of the SVR-GARCH model and then the conditional variance equation is estimated with the formula:

$$\sigma_t^2 = g\left(\sigma_{t-1}^2, u_{t-1}^2\right), \tag{18}$$

where g is the decision function of the conditional variance equation estimated by SVR.

The GARCH-based SVR method to model the correlation between information volume and trading volume volatility uses a regression function of:

$$\sigma_t^2 = f_2\left(\sigma_{t-1}^2, y_{t-1}^2, W_{t-1}^2\right), \tag{19}$$

where W_t is the amount of online financial information available on that day.

3) Volatility forecasting steps

In this paper, a GARCH-SVR model combining a two-stage approach of SVR model and GARCH model is proposed to predict the return volatility of financial options by using SVR model instead of ML method to estimate the GARCH parameters.

For real gain data y_t, σ_t^2 is unobservable. Therefore, it is necessary to set σ_t^2 the value of the metric, and a feasible metric $\sigma_t^2 = \frac{1}{5}\sum_{i=0}^{4} y_{t-i}^2$ was chosen as the value of the σ_t^2 metric for its conduct of the study.

The specific steps are as follows:

1) Estimate the parameters of the GARCH model using the maximum likelihood method to obtain the conditional variance series $\hat{\sigma}_t^2$.
2) Nonlinear regression using SVR method with the following regression function:

$$Z_t = f\left(Z_{t-1}, \hat{\sigma}_{t-1}^2, u_{t-1}^2, \sigma_{t-1}^2\right), \tag{20}$$

where $Z_t = \sigma_t^2 - \hat{\sigma}_t^2$.
3) Combining the linear GARCH model and the nonlinear SVR model, the predicted values are calculated. Similarly, the conditional variance equation in the GJR-SVR model with the regression function can be written as:

$$Z_t = f\left(Z_{t-1}, \hat{\sigma}_{t-1}^2, u_{t-1}^2, S_{t-1}^- u_{t-1}^2, \sigma_{t-1}^2\right), \tag{21}$$

where f is the decision function estimated by SVR. If $u_{t-1} < 0$, then $S_{t-1}^- = 1$, otherwise $S_{t-1}^- = 0$.

Before applying GARCH-SVR to predict financial option volatility, the kernel parameters, regularization parameters, and loss function parameters need to be selected using lattice-based search and sensitivity analysis. The data is categorized into three mutually exclusive sets, i.e., training, validation, and testing. The training set is used to estimate the model parameters and then the performance of various parameter values is evaluated in the validation set. Sensitivity analysis is done in order to assess the impact of parameter variations on the MAE of volatility prediction in the validation set. Thus, a grid search will be performed for each parameter and other parameters will be kept fixed. For each parameter, the prediction is performed in the validation set and then the MAE is calculated.

Contract code	A	B	C	D	E	…
2022-1-3	4.5045	13.0501	7.6201	9.5403	2.3132	…
2022-1-4	4.4351	12.9843	7.5808	9.4575	2.2437	…
2022-1-5	4.4782	13.0409	7.5814	9.4912	2.2761	…
2022-1-6	4.5683	13.0686	7.6642	9.5814	2.3439	…
2022-1-7	4.4812	13.0984	7.6585	9.5563	2.4038	…
…	…	…	…	…	…	…

Table 1: Partial data of the CSI 300ETF and its options

III. Empirical Results and Analysis of Option Volatility Forecasts

A. Data and Preprocessing

In this paper, the closing price data of CSI 300 ETF options during the period from January 1, 2022 to December 31, 2023 is selected as the option price history data. The closing price data of CSI 300 ETF during the same period is selected as the underlying price history data, and all the data there are from Flush IFIND.

In order to ensure the validity of the data, this paper removes the data of options with daily trading volume less than 10, listing time less than 5 trading days, and remaining expiration date less than 5 trading days. In addition, this paper also removes the options data of newly listed non-standard contracts, as their position and trading volume are small.

Some of the basic information of the obtained data is shown in Table 1, where the data are the underlying prices of the financial options in dollars. The start date of the data is the first trading day of 2022, i.e., 2022-1-3. The end date is the last business day of 2023, i.e., 2023-12-29.

For the implied volatility of a call option, we consider that it is influenced by the underlying price of the financial option S, the remaining maturity T, the risk-free rate r, and the strike price K. Therefore, we select the above four data as the eigenvalues.

For a call option contract, the price of the contract must satisfy the following conditions at any moment:

$$\max\left\{0, S - Ke^{-rt} < Call\ Price < S\right\}. \tag{22}$$

But there is still a small percentage of deep real, or imaginary call options in the actual market that do not satisfy the formula. Such call options can be considered as alternative

ARCHIVES DES SCIENCES
A Multidisciplinary Journal

Contract code	0/N	1 Week	2 Week	1 Month	3 Month	6 Month	9 Month	1 Year
2022-1-3	2.0149	2.3242	2.3325	2.3959	2.4665	2.5825	2.8418	3.0110
2022-1-4	2.0166	2.3017	2.3394	2.3883	2.4396	2.6097	2.8311	3.0257
2022-1-5	2.0405	2.3017	2.3130	2.4287	2.4395	2.6000	2.8047	3.0994
2022-1-6	2.0129	2.4632	2.3386	2.3569	2.4563	2.5982	2.8230	3.0936
2022-1-7	2.2007	2.3032	2.3649	2.3992	2.4576	2.5696	2.8105	3.0104
…	…	…	…	…	…	…	…	…

Table 2: Interbank offered rate

In value state	short-term	Metaphase	Long-term	Total	Proportion
Depth real value	675	692	205	1572	0.09924
Shallow real value	2314	1553	433	4300	0.27147
Flat value	1750	909	251	2910	0.18371
Shallow virtual value	1819	1231	319	3369	0.21269
Depth virtual value	1903	1587	199	3689	0.23289
Total	8461	5972	1407	15840	1
Proportion	0.53415	0.37702	0.08883	1	-

Table 3: Results of financial options sample classification

options, with higher risk, and are categorized as 0. For other normal options, they are categorized as 1.

For the selection of the risk-free rate, three kinds of interest rates are commonly used as the risk-free rate: treasury bond rate, interbank offered rate, and interbank bond repo rate. In this paper, we choose to use X interbank offered rate as the risk-free rate because its corresponding term matches the maturity term of the option. For the option whose remaining maturity date does not match the term of the interbank offered rate, the risk-free interest rate is estimated for the option using the linear interpolation method, and the basic information of the resulting interbank offered rate data is shown in Table 2. The data in the table is the risk-free rate in %. It can be seen that the 1-year risk-free rate for the first trading week of 2022 has reached over 3%.

B. Analysis of implied volatility of financial options

Next, options are categorized and analyzed as an example of a call option, first consider the option's in-value state, let $h = K/\ln(Se^{rT})$ be the in-value state of the call option, where r is the risk-free rate. When $h \leq -0.15$, the call option is a deep real option. When $-0.15 < h \leq -0.05$, the call option is a shallow real option. When $-0.05 < h \leq 0.05$, the call option is a flat option. When $0.05 < h \leq 0.15$, the call option is a shallow dummy option. When $h > 0.15$, the call option is a deep-dummy option. In addition to the in-value state, options can also be categorized according to the remaining expiration date of the period, let T is the remaining expiration date of the call option. When $T < 60$, the call option is a short-term option. On the $60 < T \leq 180$th day, the call option is a medium-term option. On the $T > 180$th day, the call option is a long term option.

The results after obtaining the classification of CSI 300 ETF financial options are shown in Table 3, from the perspective of time dimension, the number of short-term and medium-term options is the most, accounting for 0.53415 and 0.37702, respectively, while the number of long-term options only accounts for about 9%. In terms of in-value status, shallow real, shallow imaginary, and deep imaginary account for more and more evenly, flat options account for about 18%, and deep real options are less, only about 10%.

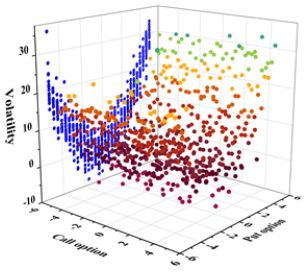

Figure 1: Implied volatility surface of CSI 300ETF options

Then, this paper takes a preliminary view of the implied volatility surface and obtains the results of the implied volatility surface of CSI 300 ETF options as shown in Figure 1. It can be seen that the volatility curve of this call option is relatively smooth and stable in the range of -10% to 35%. There are only a few singular values, and the singular values will be used later as labels for outliers for machine learning. The volatility surface reflects the volatility smile phenomenon, and the shorter the term, the more pronounced the phenomenon. The longer the horizon, the flatter the surface.

C. Model Prediction Results and Analysis

1) Analysis of empirical results on predictability

This paper compares the model of the volatility prediction model based on garch-svr in financial data analysis. This subsection presents the out-of-sample prediction accuracy of different machine learning models under the out-of-sample R^2 metric. Monthly time rolling is used to move the window during the financial option volatility prediction process for a total of 24 months (2022.1∼2023.12). Where Mean refers to the R^2 average of the out-of-sample 24 months, and Min and Max refer to the maximum of each model when predicting the out-of-sample data R^2 in %, respectively. The descriptive statistics of the out-of-sample R^2 results of each machine learning model are shown in Table 4, which show that the out-of-sample R^2 mean values of decision trees (CART), gradient boosted trees (GBDT), support vector machines (SVR), convolutional neural networks (CNN), and the GARCH-SVR proposed in this paper, for the five machine learning models, are greater than 0, indicating that the five machine learning models have the effect of prediction. Even the linear regression model with stochastic gradient descent added has some prediction effect. Specifically, the GARCH-SVR model has the best prediction result for financial option volatility, and its R^2 mean value reaches 0.6349%, followed by the SVR and CNN models, which proves the high efficiency of the GARCH-SVR model constructed in this paper.

The R^2 of the different machine learning models for each trading day is shown in Figure 2, where it can be seen that although the SVR model has a higher average R^2, the R^2 fluctuations for each trading day are higher and less robust. Although the CNN, GBDT, and CART models are less volatile, none of them R^2 are large overall. the overall volatility and R^2

ARCHIVES DES SCIENCES
A Multidisciplinary Journal

Model	Mean	Min	Confidence interval			Max
			0.25	0.5	0.75	
CART	0.1354	0.0163	0.0198	0.0307	0.1952	0.6326
GBDT	0.1896	0.0204	0.0186	0.0230	0.4255	0.7417
CNN	0.4534	-1.2334	-0.1941	0.0605	0.2941	1.2737
SVR	0.5166	-4.2585	-1.7641	0.0275	1.1641	9.3817
GARCH-SVR	**0.6349**	**-1.1517**	**0.6581**	**0.1828**	**1.6581**	**4.7986**

Table 4: Out of sample R^2 Descriptive statistics for each machine learning model (%)

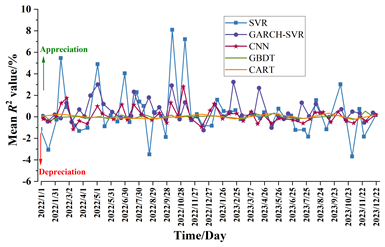

Figure 2: R^2-value results of different machine learning models for each trading day

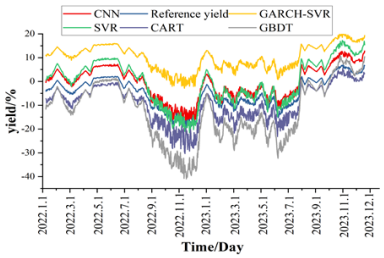

Figure 3: Cumulative returns of different machine learning portfolios

Index	CART	GBDT	CNN	SVR	**GARCH-SVR**
Accumulated income	22.15%	4.31%	32.08%	41.29%	46.33%
Annual income	4.51%	0.97%	8.35%	6.23%	9.06%
Excess income	5.91%	-5.37%	23.15%	15.41%	27.34%
Sharpe ratio	0.0038	0.056	0.1418	0.2204	0.3208
Winning percentage	0.4680	0.4705	0.4756	0.5047	0.5251
Maximum retest	32.45%	34.31%	32.86%	34.07%	28.39%
Volatility	0.1848	0.1952	0.1736	0.1715	0.1608
Profit and loss ratio	0.1233	0.1015	1.2507	1.3721	1.5238

Table 5: Details of portfolio strategies for different machine learning

mean of the GARCH-SVR model are more desirable, with the difference between the maximum and minimum values being 3.9726%. It is also worth noting that the GARCH-SVR model has more consistent peaks and troughs in the volatility process, and is basically isotropic in terms of predictive ability.

Overall the GARCH-SVR model does improve the out-of-sample prediction results of the model, even compared to stochastic gradient descent linear regression, proving the superiority of this machine learning method in capturing the complex interactions between the predictor variables, and verifying its validity in the prediction of the financial options market.

2) Investment Performance Analysis under Machine Learning Predictions

This subsection results in the construction of a simple financial options investment strategy based on the five machine learning forecasts above. A financial options portfolio for each month is constructed by sorting the stock return forecasts for the following month obtained from the monthly rolling forecasting model from high to low. The test sample interval is from January 2022, to December 2023. At the same time, in order to simulate the investment requirements of risk diversification in the real market, each model constructs a portfolio by screening the top twenty financial options in terms of yield based on the rolling monthly forecasts. In this paper, the slippage point is set to 0.0025, which means that when a buy order is placed, the price of the transaction is equal to the average price at that moment plus half of the spread. The spread in this paper is 0.0025 of the price at that moment, the same when selling.

The cumulative returns of the different machine learning out-of-sample asset portfolios are shown in Figure 3, and the benchmark return curve is the return of the Shanghai Composite Index during the out-of-sample test period used to complete the comparison study. It can be seen that the different models have some synergy in the trend of constructing the portfolio strategy, and in general with the general market trend of the options market is almost the same. If we only look at the portfolio returns out of sample, the portfolio strategy constructed based on the GARCH-SVR model has the largest cumulative return, with a maximum of 18.58%. It is worth noting that the cumulative returns of the portfolio strategies constructed based on the CART and GBDT models are low, even lower than the benchmark returns, which may be related to the importance analysis of the characteristics above.

In order to better measure the risk and further compare the predictive ability of each model, this paper will use the evaluation metrics of trading strategies such as strategy cumulative return, strategy annualized return, excess return, Sharpe ratio, win rate, maximum retracement, volatility, and profit/loss ratio. The details of each indicator under different machine learning models are shown in Table 5. In terms of the traditional Sharpe ratio, the portfolio strategy constructed based on the GARCH-SVR model is the highest at 0.3208, followed by the SVR (0.2204) model and the CNN (0.1418) model. In terms of win rate, the portfolio strategy constructed based on the GARCH-SVR model has a win rate of 0.5251, which is significantly higher than the other four models. In terms of maximum retracement and volatility, the portfolio strategies constructed based on the GARCH-SVR model are both optimal, presenting small retracement and high return among the five models. The portfolio strategies constructed based on other machine learning models, on the other hand, are characterized by high volatility and low returns.

In summary, compared with several other machine learning algorithms, the GARCH-SVR model constructed in this paper is more advantageous for volatility prediction in China's financial options market, which has economic and investment significance.

IV. Conclusion

As an important attribute of financial derivatives, financial volatility can be effectively used in the pricing of financial derivatives, allocation of financial assets and risk management. This paper integrates multi-source data, constructs a financial option volatility prediction model based on the GARCH-SVR algorithm, and verifies its effectiveness through empirical analysis. The basic conclusions are as follows:

1) Machine learning is effective in predicting financial option volatility scenarios. The garter SVR model has the best prediction of the volatility of financial options, and its R^2 average is 0.6349%. For the factor set of inputs constructed in this paper, the GARCH-SVR model achieves an efficient return of 0.3208 and a win rate of 0.5251 for the portfolio strategy. Based on the income of the portfolio, the cumulative yield of the portfolio strategy built based on the garter SVR model is the largest, maximum to 18.58%.
2) Machine learning that integrates information from multiple sources of data has advantages in the field of prediction, and the machine learning GARCH-SVR model constructed in this paper also provides a new management tool for the financial sector, which broadens the application of machine learning theory in the field of financial risk management.

Finally, the shortcoming of this paper is that the grid search method is used in the setting of hyperparameters for machine learning. Further the method of manual parameter tuning can be used in order to optimize the model and improve the effectiveness of the machine learning algorithm model.

Acknowledgments

2021 in the key areas of ordinary universities in Guangdong Province (science and technology services for rural revitalization): Study on the development model of agricultural productive service industry in guangdong-hong Kong-macao Bay Area (2021 ZDZX4030).

References

[1] Sankhwar, S., Gupta, D., Ramya, K. C., Sheeba Rani, S., Shankar, K., & Lakshmanaprabu, S. K. (2020). Improved grey wolf optimization-based feature subset selection with fuzzy neural classifier for financial crisis prediction. *Soft Computing, 24*(1), 101-110.
[2] Lin, W. C., Tsai, C. F., & Chen, H. (2022). Factors affecting text mining based stock prediction: Text feature representations, machine learning models, and news platforms. *Applied Soft Computing, 130*, 109673.
[3] Huang, Y. P., & Yen, M. F. (2019). A new perspective of performance comparison among machine learning algorithms for financial distress prediction. *Applied Soft Computing, 83*, 105663.
[4] Zhang, F. (2023). RETRACTED: Extreme learning machine for stock price prediction. *International Journal of Electrical Engineering & Education, 60*(1_suppl), 3972-3985.
[5] Ntakaris, A., Kanniainen, J., Gabbouj, M., & Iosifidis, A. (2020). Mid-price prediction based on machine learning methods with technical and quantitative indicators. *Plos one, 15*(6), e0234107.
[6] Jang, H., & Lee, J. (2019). Generative Bayesian neural network model for risk-neutral pricing of American index options. *Quantitative Finance, 19*(4), 587-603.
[7] Das, S. P., & Padhy, S. (2017). Unsupervised extreme learning machine and support vector regression hybrid model for predicting energy commodity futures index. *Memetic Computing, 9*, 333-346.
[8] Wei, W., & Zhang, Q. (2022). Evaluation of rural financial ecological environment based on machine learning and improved neural network. *Neural Computing and Applications*, 1-18.
[9] Du, X., Li, W., Ruan, S., & Li, L. (2020). CUS-heterogeneous ensemble-based financial distress prediction for imbalanced dataset with ensemble feature selection. *Applied Soft Computing, 97*, 106758.
[10] Chou, J. S., Nguyen, N. M., & Chang, C. P. (2022). Intelligent candlestick forecast system for financial time-series analysis using metaheuristics-optimized multi-output machine learning. *Applied Soft Computing, 130*, 109642.
[11] Liang, M. (2021). Optimization of quantitative financial data analysis system based on deep learning. *Complexity, 2021*(1), 1-11.
[12] He, H., Gao, S., Jin, T., Sato, S., & Zhang, X. (2021). A seasonal-trend decomposition-based dendritic neuron model for financial time series prediction. *Applied Soft Computing, 108*, 107488.
[13] Bumin, M., & Zalici, M. (2022). Predicting the direction of financial dollarization movement with genetic algorithm and machine learning algorithms: the case of turkey. *Expert Syst. Appl., 213*, 119301.
[14] Qian, H., Wang, B., Yuan, M., Gao, S., & Song, Y. (2022). Financial distress prediction using a corrected feature selection measure and gradient boosted decision tree. *Expert Systems with Applications, 190*, 116202.
[15] Gu, N. (2021). Digital financial inclusion risk prevention based on machine learning and neural network algorithms. *Journal of Intelligent & Fuzzy Systems,* (Preprint), 1-16.
[16] Çelik, T. B., İcan, Ö., & Bulut, E. (2023). Extending machine learning prediction capabilities by explainable AI in financial time series prediction. *Applied Soft Computing, 132*, 109876.
[17] Booth, D. E., Alam, P., & Ozgur, C. (2019). Machine learning methods for financial prediction. *Expert Systems, 10*(1), 86-92.
[18] Ma, X., & Lv, S. (2019). Financial credit risk prediction in internet finance driven by machine learning. *Neural Computing and Applications, 31*(12), 8359-8367.
[19] Yu, J., & Zhao, J. (2020). Prediction of systemic risk contagion based on a dynamic complex network model using machine learning algorithm. *Complexity, 2020*(1), 6035372.
[20] Metawa, N., Alghamdi, M. I., El-Hasnony, I. M., & Elhoseny, M. (2021). Return rate prediction in blockchain financial products using deep learning. *Sustainability, 13*(21), 11901.
[21] Mohapatra, U. M., Majhi, B., & Satapathy, S. C. (2019). Financial time series prediction using distributed machine learning techniques. *Neural Computing and Applications, 31*, 3369-3384.
[22] Mohanty, D. K., Parida, A. K., & Khuntia, S. S. (2021). Financial market prediction under deep learning framework using auto encoder and kernel extreme learning machine. *Applied Soft Computing, 99*, 106898.
[23] Chen, Y., Liu, K., Xie, Y., & Hu, M. (2020). Financial trading strategy system based on machine learning. *Mathematical Problems in Engineering, 2020*(1), 3589198.
[24] Chen, S. (2022). Cryptocurrency financial risk analysis based on deep machine learning. *Complexity, 2022*(1), 2611063.