information

Article

Dual-Hybrid Modeling for Option Pricing of CSI 300ETF

Kejing Zhao, Jinliang Zhang * and Qing Liu

School of Mathematics and Statistics, Henan University of Science and Technology, Luoyang 471000, China; 190319100412@stu.haust.edu.cn (K.Z.); 200320100647@stu.haust.edu.cn (Q.L.)
* Correspondence: 9901914@haust.edu.cn

Abstract: The reasonable pricing of options can effectively help investors avoid risks and obtain benefits, which plays a very important role in the stability of the financial market. The traditional single option pricing model often fails to meet the ideal expectations due to its limited conditions. Combining an economic model with a deep learning model to establish a hybrid model provides a new method to improve the prediction accuracy of the pricing model. This includes the usage of real historical data of about 10,000 sets of CSI 300 ETF options from January to December 2020 for experimental analysis. Aiming at the prediction problem of CSI 300ETF option pricing, based on the importance of random forest features, the Convolutional Neural Network and Long Short-Term Memory model (CNN-LSTM) in deep learning is combined with a typical stochastic volatility Heston model and stochastic interests CIR model in parameter models. The dual hybrid pricing model of the call option and the put option of CSI 300ETF is established. The dual-hybrid model and the reference model are integrated with ridge regression to further improve the forecasting effect. The results show that the dual-hybrid pricing model proposed in this paper has high accuracy, and the prediction accuracy is tens to hundreds of times higher than the reference model; moreover, MSE can be as low as 0.0003. The article provides an alternative method for the pricing of financial derivatives.

Keywords: CSI 300ETF options; random forest; CNN-LSTM; double hybrid model; ridge regression; model robustness

Citation: Zhao, K.; Zhang, J.; Liu, Q. Dual-Hybrid Modeling for Option Pricing of CSI 300ETF. *Information* **2022**, *13*, 36. https://doi.org/10.3390/info13010036

Academic Editor: Ognjen Arandjelović

Received: 2 December 2021
Accepted: 10 January 2022
Published: 13 January 2022

1. Introduction

Option is an important tool for investors to obtain benefits and avoid risks in financial derivatives, and option pricing has always been a hot research objective of scholars. The effective pricing of options contributes to the stability of the financial market and the sustained development of the national economy. In order to improve the accuracy of option pricing, scholars have performed a variety of research. Currently, option pricing models are mainly divided into parametric models and non-parametric models. In 1973, Black and Scholes [1] studied and obtained the famous Black–Scholes (B-S) pricing formula and established the classical parametric pricing model. The B-S formula can lead all investors to a risk-neutral world with risk-free interest rate as the rate of return, and it can predict the price of options better, regardless of their preferences. However, the formula has made many assumptions in advance, for example, the volatility and interest rate of options are assumed to be a constant; the underlying asset follow geometric Brownian motion, etc., is not completely consistent with the actual market situation; thus, the option price calculated by the formula is far from the actual situation. Later, many scholars made corresponding improvements to the model, such as adjusting the time course of the evolution of the underlying asset price, and the interest rate and volatility were subject to a random process [2–15] so as to establish option pricing models closer to the actual situation, such as the CIR model [16,17] and Heston model [18–22].

The above parametric pricing model follows strict assumptions. Once there is a discrepancy between the actual situation and assumed conditions, the pricing result will have a large error with the real price, which greatly affects the accuracy and stability of the

model. In order to estimate option prices more accurately, in recent years, some scholars have begun to try to apply machine learning methods to option pricing problems, such as the Non-parametric modular Neural Network [23] (NNN), Support Vector Machine [24] (SVM), Decision Tree [25] (DT), Artificial Neural Network [26] (ANN), etc. In this method, the model only needs to focus on the relationship between the features in the option data, without considering the complex economic principles and rigorous mathematical derivation, which provides a new modeling idea for option pricing. However, machine learning-based methods have a common shortcoming, which is the manual extraction of data features. This process is complicated and tedious, and the nonlinear fitting ability of some methods is insufficient, which results in a lack of implicit information in the extracted features. In addition, financial option data often have high dimensional and nonlinear characteristics, which results in unsatisfactory classification effects, thus affecting the overall performance of its pricing model.

It is worth noting that deep learning has received great attention in the field of financial time series analysis due to its strong nonlinear fitting ability and feature capture ability. As an important branch of neural networks in machine learning, deep learning can conduct in-depth feature screening and learning of data, desalting irrelevant factors and strengthening relevant factors while learning heterogeneity information. In particular, CNN in deep learning has excellent performance in this aspect. It can automatically extract features. To a certain extent, the more layers set, the more advanced features extracted, and the more information contained, and the better the classification effect. CNN is widely used in graphic and image processing because it requires less hyperparameters and less computation [27,28]. LSTM is a temporal recursive neural network, which was first proposed by Hochreiter and Schmidhuber [29] in 1997. Originating from recurrent neural networks (RNNs), LSTM overcomes the problem of gradient disappearance in the training process of RNN and is an effective tool for long-term prediction. In recent years, the application of LSTM to time-series-dependent data prediction has gradually become popular, and LSTM has relatively mature studies in options pricing [30], volatility prediction in option market [31], stock price prediction [32,33] and so on. Some scholars try to combine CNN and LSTM to establish CNN-LSTM hybrid models for medical [34] and stock price prediction [35], etc. A large number of empirical results show that the single neural network prediction model is often difficult for accurately predicting option prices due to the impact of the discontinuity of trading. Due to complex and random evolution paths of interest rate and volatility that affect option prices in the real market, CIR and Heston models effectively break through the confinement of constant interest rate and constant volatility assumed by the traditional B-S pricing model and can simultaneously meet dual requirements of the market for volatility and interest rate and better adapt to the real option market. Therefore, combining the classical option pricing parameter model with the deep learning non-parametric model while establishing a hybrid pricing model has become a new research direction The dual-hybrid model can effectively overcome problems existing in parametric model and non-parametric model, such as poor nonlinear fitting ability and poor interpretation.

With the advent of the era of Big Data, a large amount of data is generated in the process of option trading, as well as characteristic data that may affect the final price of options. In machine learning experiments, especially the pricing process of financial derivatives, feature engineering has always been the focus of scholars' research. Effective analysis of option pricing feature data can reduce the complexity of the pricing model and can greatly improve the prediction accuracy and stability of the model. Some scholars screen features in order to reduce their dimensionality to study and discuss important features. The main methods adopted are principal component analysis [36] (PCA), random forest [37,38] (RF), factor analysis [39] (FA), etc. Random forest is a data mining classification algorithm based on statistical sampling theory (Bootstrap), which is subordinate to ensemble learning. There are two types of random forest regression and random forest classification. Another group of scholars focused on the internal relations between features and adopts, for example, the

least squares method (LS) performs feature correlation analysis. Feature-processed data input not only greatly reduces the running time of the model but also effectively improves prediction accuracy, which is more suitable for the option pricing in the real market.

In summary, in order to price options more effectively, improve prediction accuracy and model performance and maintain the sustainable and stable development of the financial market, this paper combines the CIR-Heston stochastic mixed parameter model and the CNN-LSTM neural network mixed non-parametric model to propose a new dual-mixed model for option pricing. This is conducted to better explore the performance of the dual-hybrid model in option price prediction; effectively predict the price trend; avoid losses; provide investors with reference; perform random forest feature importance sorting on the original input feature variables of CSI 300ETF option historical data; filter out the main features; use historical data processed by feature engineering as experimental data to train and test the above-mentioned dual-mixed model; to compare with the prediction results of reference models; and to integrate ridge regression, which completes empirical analysis. The dual-hybrid model and the reference model were tested for model robustness to verify the overall performance of the model. The article provides a new method of thinking for pricing financial derivatives by considering the establishment of a dual-mixed model and adopting different economic models based on the characteristics of experimental data to more closely fit the real financial market.

2. Model Establishment

In view of the high volatility and obvious lag of the option market based on the importance of random forest features and combined with the respective advantages of the deep neural network model and the random parameter pricing model, a dual-mixed pricing model is established, and ridge regression integration is used perform predictive analysis. The structure diagram of the hybrid model is shown in Figure 1.

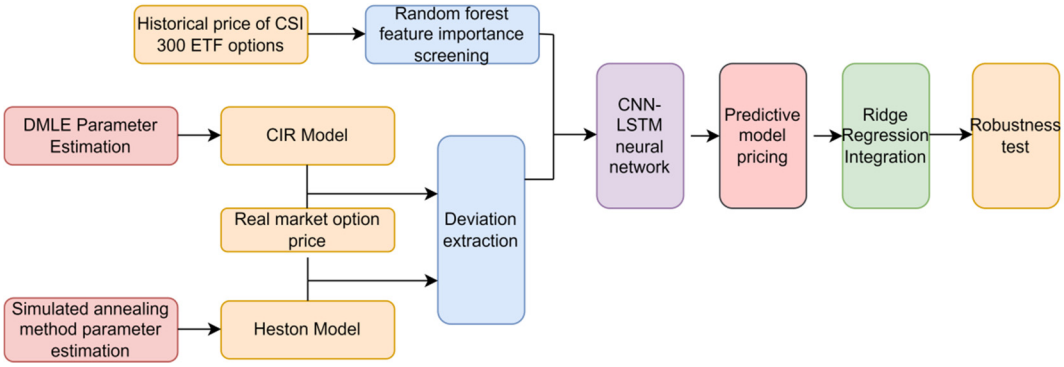


Figure 1. Overall architecture.

2.1. Random Forest Method

Random forest [37] feature importance ranking is performed on the original input feature variables of the historical data of CSI 300ETF options. According to the actual option market transaction data, three groups of sixteen candidate characteristic variables are listed, as shown in Table 1.

Table 1. To select a list of features.

B-S Feature	Transaction Characteristics	Greek Alphabet Characteristics
S	Positions	Delta
r	Turnover	Gamma
K	ETF price	Vega
T	Volume	Theta
Sigma	High	Rho
	Low	

2.2. The CIR-Heston Pricing Model

Cox, Ingersoll and Ross et al. proposed a generalized equilibrium single-factor model-CIR stochastic interest rate model in 1985. The model assumed that under the risk-neutral measure, the evolution process of interest rate is as follows:

$$dR(t) = (a - bR(t))dt + \sigma_1\sqrt{R(t)}dW(t) \tag{1}$$

where *a/b* is the long-term mean, *b* is response rate, σ_1 is the annualized volatility, $R(t)$ is the random interest rate and $W(t)$ denotes Brownian motion. Under the CIR interest rate model, the pricing of zero-coupon bonds $B(t,T)$ is provided by the following.

$$B(t,T) = \exp\{-R(t)C(t,T) - A(t,T)\} \tag{2}$$

In the above CIR pricing model, in addition to the four parameters (S, T, K and r) already included in B-S, there are still three parameters (a, b, σ_1) to be estimated.

Heston proposed the Heston stochastic volatility model in 1993, and the explicit solution formula for European options under the Heston model is given by the following:

$$C(S(t),V(t),t) = S(t)P_1 - Ke^{-r(T-t)}P_2 \tag{3}$$

where C is the market price to be charged for the option. K is the strike option, T is the annualized option maturity, r is the risk free rates and P_1 and P_2 are two probability distribution functions.

In the above formula, in addition to the four parameters (S, K, T and r) contained in B-S, there are six parameters $(\kappa, \theta, \sigma, \rho, V(t), \lambda)$ to be estimated. However, in a risk-neutral world where the volatility risk premium is $\lambda = 0$, the parameters to be estimated are reduced to five, namely $\kappa, \theta, \sigma, \rho, V(t)$.

2.3. CNN-LSTM Deep Neural Network Model

CNN, as a kind of feedforward neural network, uses convolution calculations instead of general matrix multiplication operations to form a neural network specifically designed to process data with a similar grid structure. In convolutional neural networks, each of the convolutional layer is connected to the pooling layer, and the alternating effect of the convolutional layer and the pooling layer can dig out deep-level features with discrimination from a large amount of data. The fully connected layer is preceded by the last pooling layer, which is used to integrate features taken by the alternate convolution and pooling so as to obtain more discriminative features. LSTM is an improved RNN. By adding memory units in the hidden layer, each neural unit is transmitted through "gating." "Gating" determines the degree of memory and forgetting of past and instant information, rendering LSTM to possess a long-term memory function. Option data studied in this paper comprise one-dimensional time-series data for which CNN has a strong feature extraction ability, and LSTM then trained the data extracted by CNN features in order to obtain the predictive model of the deep neural network shown in Figure 2.

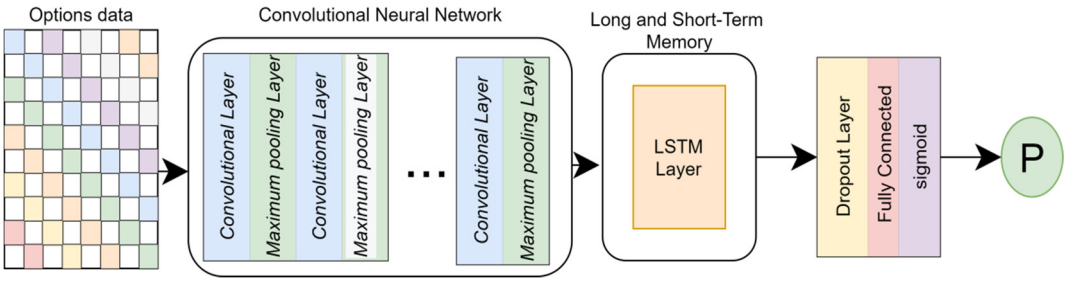


Figure 2. Overall architecture of the CNN-LSTM deep neural network model.

2.4. Double-Hybrid Modeling

From the perspective of financial time series volatility, the key to option pricing is to study its random jump, dynamics, etc., and use scientific theoretical derivations to quantitatively analyze option prices. However, the classic models of option pricing, such as the single B-S, CIR and Heston models, have difficulty meeting the expectations of investors in the actual market. The fundamental reason is that the model itself has many bottlenecks. For example, the actual option price distribution does not match the assumptions. Models established for a specific sequence are often not robust, especially when the underlying asset price has a discontinuous jump or sudden change, the performance of the traditional option pricing model drops significantly, and it cannot adapt well to frequent and drastic fluctuations. CIR model and Heston model can effectively improve the assumptions of constant volatility and constant interest rate in B-S. The mixed parameter model combining the two models help in improving the accuracy and generalization of prediction.

Aiming at the shortcomings of traditional option pricing parameter models, this paper combines the deep neural network model with the parameter pricing model to establish a dual-hybrid pricing model. It not only takes advantage of the strict logic and clear structure of the traditional parameter model, but it also includes the advantages of nonlinear fitting and strong extension of the neural network model. The specific framework of the dual-mixed model is shown in Figure 3.

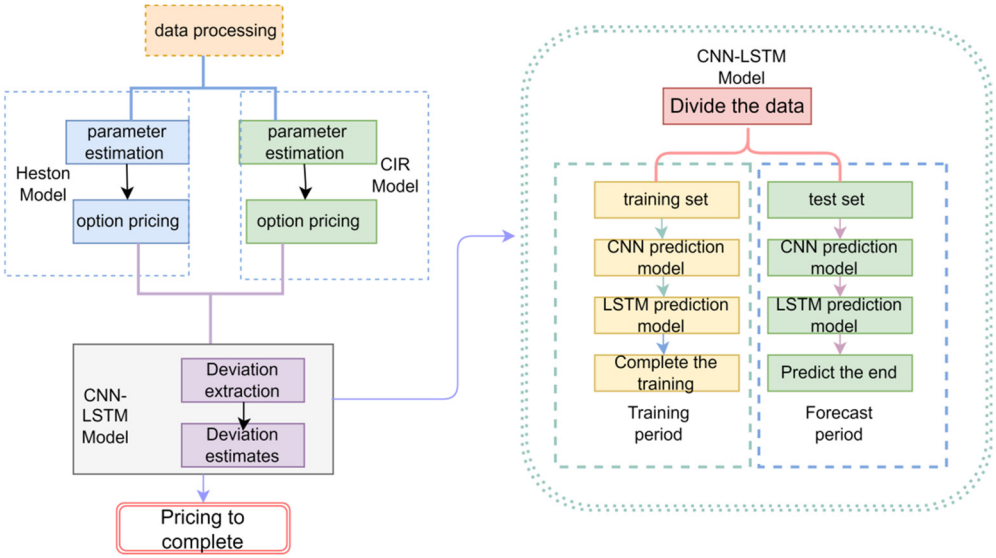


Figure 3. CSI 300ETF option pricing dual-hybrid modeling framework.

The modeling framework of the dual-hybrid pricing model proposed in this article mainly has the following steps:

1. Parameter estimation of CIR model. Parameter estimation is based on the minimum mean square error between the price estimated by CIR and the historical real

price. Discrete Maximum Likelihood Estimation (DMLE) is mainly used for actual programming to estimate three parameters.

2. Parameter estimation of Heston model. Based on the minimum percentage error between Heston's estimated price and the historical real price, parameter estimation is carried out. The objective function is solved by combining the simulated annealing method, and five parameters are estimated.
3. Extraction of deviation sequence. Taking the real option price as the benchmark, find the price deviation sequence between the price estimated by CIR, the price estimated by Heston and the real price and standardize the sequence of pricing deviation.
4. CNN-LSTM model construction. Use the standardized deviation sequence and the original input parameters of the models to establish a training set and test set to train the CNN-LSTM model.
5. Empirical analysis of option pricing. Use the trained CNN-LSTM model to empirically verify test set data and analyze the estimated price of the dual-hybrid model with the real price.

3. Empirical Analysis

3.1. Experimental Data and Performance Indicators

DMLE is used to estimate the parameters of the CIR model. This method approximates the transition probability density with the aid of the discrete form of stochastic differential equation (SDE) simulation and performs Euler discrete approximation of CIR in an interval. This paper uses MATLAB internal functions to generate trajectories, and 200 trajectories are generated with a time interval of 0.01. The average value of 200 experimental results is taken to obtain the estimated values of three parameters to be estimated. Figure 4 shows the results of CIR parameter estimation.

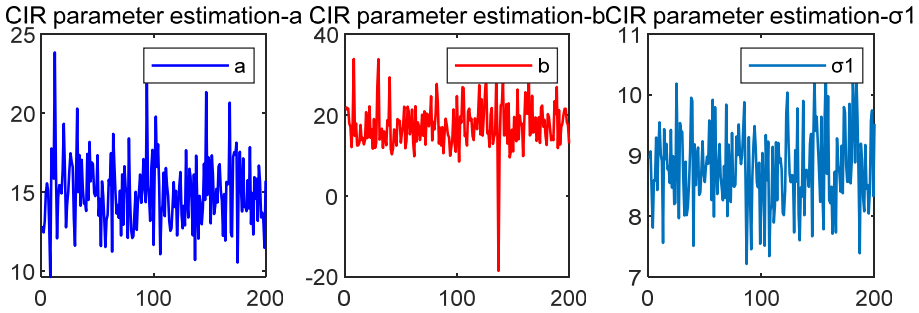


Figure 4. Parameter estimation figure of the CIR model.

In the process of estimating the five estimated parameters of the Heston model with the simulated annealing method, the initial temperature selected in this paper is 100 degrees, the simulated annealing training is performed at an annealing rate of 0.7, and the final temperature is set to 0.00001. The temperature cycle is terminated when the training temperature is lower than the minimum temperature or when the absolute value of the difference between the new and the old optimal value is less than 0.001 for 100 consecutive times.

In order to evaluate the effectiveness and practicability of the proposed dual-hybrid model in option pricing, the dataset used for this research study is CSI 300ETF option data in 2020. It varies from 1 to 2 days before the expiry date, including call options and put options. In order to eliminate the influence of noise in the experiment, this article normalizes original data and eliminates outliers. Some experimental data are shown in Table 2.

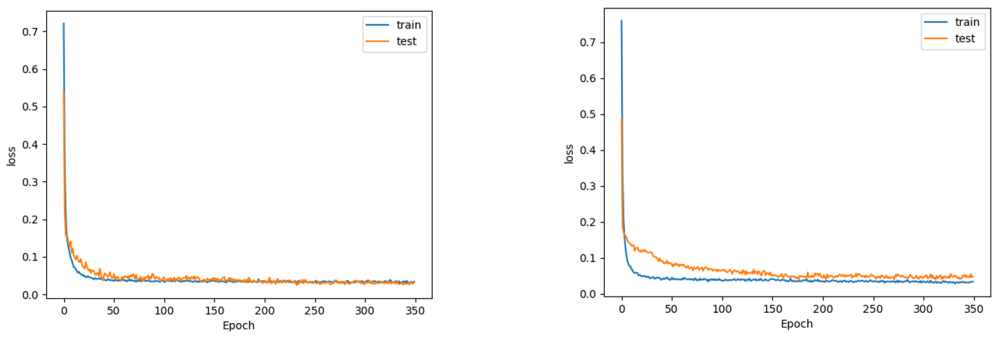
Table 2. Experimental sample data table.

Date	Contract Encoding	At the Opening	Maximum Price	Minimum Price	Settlement	Volume
2 January 2020	510300C2001M03600	0.528	0.589	0.528	0.561	1947
3 January 2020	510300C2001M03600	0.569	0.575	0.546	0.551	1409
6 January 2020	510300C2001M03600	0.535	0.581	0.504	0.534	1970
7 January 2020	510300C2001M03600	0.552	0.569	0.541	0.561	1425
8 January 2020	510300C2001M03600	0.542	0.557	0.505	0.520	2579
9 January 2020	510300C2001M03600	0.544	0.571	0.544	0.564	1331
10 January 2020	510300C2001M03600	0.587	0.587	0.551	0.564	1255

Establish a CNN-LSTM hybrid prediction model in the deep learning environment of Keras. Take 5000 sets of historical data for call options and put options to make pricing predictions, 95% of which are used as training data and 5% as test data. Verification data account for 15% of the training data. The verification set is used in the training process for model parameter selection. Due to the fact that the price of financial options is affected by policy implementation, financial conditions, investor sentiment, etc., it will have obvious characteristics such as jumps and discontinuities. The *sigmoid* activation function selected in this paper has the output value of the neuron limited between 0 and 1, and the gradient is smooth. It can effectively avoid the advantage of jumping output and can improve the discontinuity of option data, but *sigmoid* tends to disappear in gradient, and it performs exponential operation; thus, the running time is longer. When the input of the *RELU* activation function is positive, there is no gradient saturation problem, and because a linear operation is performed, the operation speed is faster, and the *sigmoid* function can effectively overcome the shortcomings of the function. In this paper, the *sigmoid* function and *RELU* function are used alternately, which can not only normalize option data and improve model performance but also shorten program running time. The expression is shown in formula (4). The optimizer in the non-parametric model of the neural network in the article selects Adam and uses a cross-validation method to adjust prediction results inside the program.

$$\begin{cases} sigmoid : \sigma(x) = \frac{1}{1+e^{-x}} \\ RELU : f(x) = \begin{cases} 0, \ if \ x < 0 \\ x, \ if \ x \geq 0 \end{cases} \end{cases}, \quad (4)$$

Figure 5, respectively, shows the loss changes of the dual-hybrid model proposed in this article on CSI 300ETF call options and put options datasets. The abscissa is the Epoch value, and the ordinate is the loss value. It can be observed from Figure 4 that, in the process of training call option data for this model, the loss of the training set gradually becomes flat after Epoch = 50, but the testing set is still in a declining state at this time, and there are obvious fluctuations. When Epoch = 150, the loss of the training set reaches the convergence state, and the test set also remains stable and converges to a lower level. At this time, the loss converges below 0.1, which is close to 0, and there is no underfitting or overfitting. Therefore, the Epoch of the hybrid model training call option data is selected as 150.



(**a**) (**b**)

Figure 5. Loss changes (**a**) of call option; (**b**) of put option.

Figure 5b trains put option data, and the training set has stabilized, but the test set has a significant downward trend when Epoch = 50. When Epoch = 200, both the training set and the testing set reach a stable level. The state of convergence is between 0 and 0.1. Therefore, the Epoch of the hybrid model training put option data is selected as 200. However, it is worth noting that unlike call options, the overall loss of the test set when training put options data is higher than that of the training set, indicating that the neural network model's overall predictive effect on put option pricing is slightly lower than that of call options, but it meets ideal expectations overall. The loss change results of put options are also consistent with previous scholars' research results. The parameter estimates and other parameter settings in the hybrid model are shown in Table 3.

Table 3. Parameter setting table of the model.

Model	Parameter Name	Call Options	Put Options
CIR	a	14.0367	
	b	15.5455	
	σ_1	9.4658	
Heston	κ	1.8638	1.9742
	θ	0.0223	0.1012
	σ	0.1531	0.0511
	ρ	−0.6351	−0.4805
	$V(0)$	0.1302	0.0102
CNN-LSTM	Batchsize/Epoch	64/250	64/150
CIR-Heston-ANN	Batchsize/Epoch	64/200	64/250
CIR-Heston-CNN	Batchsize/Epoch	32/300	32/150
CIR-Heston-LSTM	Batchsize/Epoch	32/250	32/200
CIR-Heston-CNN-LSTM	Batchsize/Epoch	64/150	64/200

The normalized data is the *StandardScaler* function and the formula is shown in Equation (5):

$$\boldsymbol{x} = \frac{\boldsymbol{x_i} - \boldsymbol{x_{mean}}}{\boldsymbol{x_{std}}} \tag{5}$$

This paper uses mean square error (MSE), mean absolute percentage error (MAPE) and coefficient of determination to evaluate the performance of each model. The smaller the error, the higher the prediction accuracy of the model, and the larger the coefficient of determination, the higher the degree of fit. The calculation formula is shown in Equation (6):

$$\begin{cases} \mathrm{MSE} = \frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \left(\mathrm{x_{pred}} - \mathrm{x_{real}} \right)^2 \\ \mathrm{MAPE} = \sum_{\mathrm{i}=1}^{\mathrm{N}} \left| \frac{\mathrm{x_{pred}} - \mathrm{x_{real}}}{\mathrm{x_{real}}} \right| \times \frac{100}{\mathrm{N}} \\ \mathrm{R}^2 = \frac{\sum \left(\mathrm{x_{pred}} - \mathrm{x_{mean}} \right)^2}{\sum \left(\mathrm{x_{real}} - \mathrm{x_{mean}} \right)^2} \end{cases} \tag{6}$$

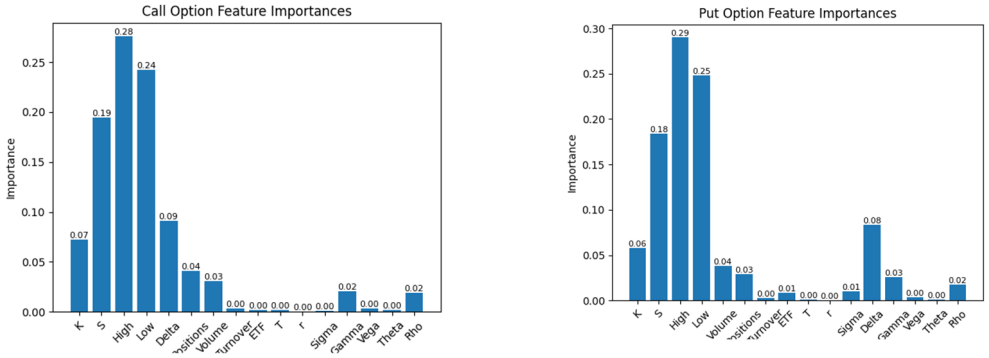
where x_{pred} is the predicted price of the option, x_{real} is the true price of the option, x_{mean} is the average value of the option price and N is the number of samples.

3.2. Random Forest Feature Engineering

This paper uses the random forest regression algorithm under the SK-learn framework in Python to perform feature importance screening. When performing feature importance screening on historical data of call and put options, respectively, according to the experience of previous scholars [38], the number of classifiers is n_estimators = 500. Its random forest

seeds are set to 56 and 217, respectively, and other parameters are set according to the default parameters under the SK-learn framework.

The ranking of importance after the experiment can be observed in Figure 6.



(**a**) (**b**)

Figure 6. Feature importance ranking (**a**) of call option; (**b**) of put option.

In order to maintain model accuracy and moderate complexity, this paper selects features with importance greater than 0.03 for subsequent model experiments. The selection of features can be observed in Table 4, and they are arranged in descending order of importance.

Table 4. Characteristic importance sort table.

Type of Option	Characteristic Importance Sorting (Descending Order)					
Call Option	High	Low	S	Delta	K	Positions
Put Option	High	Low	S	Delta	K	Volume

3.3. Empirical Analysis of Call Options

The predictions of the CSI 300 ETF call options by different prediction models can be observed in Figure 7 and Table 5. It can be observed that the predicted values of the single parameter model CIR and Heston are quite different from true values. The prediction effect of Heston model at peak price is better than CIR, but the prediction effect is very poor in areas where the price fluctuation degree is not large, and it cannot even maintain the same trend. The CIR model has a better prediction effect when the curve fluctuates steadily, but by observing Figure 6, it can be observed that there are a total of about 30 sample points with a prediction value of 0. This is because the exercise price of some of the original data is 0. Heston's predicted value has a clear tendency to underestimate in the early and late stages of the forecast, and there is a significant tendency to overestimate in the middle of the forecast. CIR has underestimated to varying degrees throughout the forecast period. When the option price fluctuates violently, the forecast errors of the two significantly increased, indicating that the parameter model cannot predict market fluctuations well. At the same time, the R^2 of the two are too small, which also indicates that a single parameter model cannot perform well on the real option market. In comparison, the prediction effect of the deep neural network model is significantly better. The performance of both the single neural network model and the hybrid model is distinctly improved compared with the former. This shows from another level that there are some price-influencing factors in the option market that cannot be described by parameter models. As far as the prediction accuracy (MAPE) is concerned, the model proposed in this article is not the best (MAPE is 0.0875), but it is only about 0.06 less than the best accuracy CIR-Heston-CNN, and the MAPE of the single parameter model is several times or even dozens of times lower than that of the hybrid neural network model. The stabilities of the model (MSE is 0.0026) and the coefficient of determination (R^2 is 0.9865) are both the best, which proves the effectiveness

of the model proposed in this paper. It is worth noting that, in addition to the model mentioned in this article, the three performance indicators of the CIR-Heston-CNN model have obvious advantages over other reference models. This shows that the characteristics of automatic extraction of data features by convolutional neural networks can effectively improve the effect of financial sequence prediction.

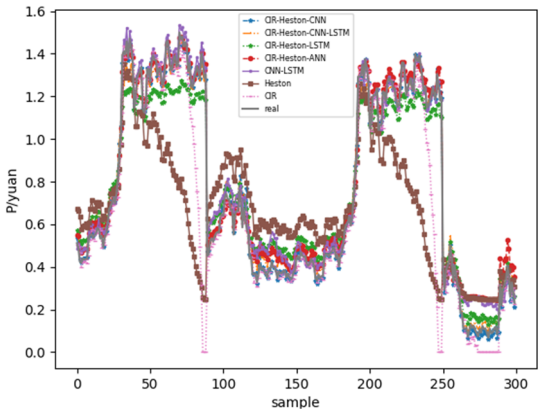


Figure 7. Comparison of real and predicted results of CSI 300ETF call options.

Table 5. Comparison of the prediction accuracy between the different models.

Model	MSE	MAPE	R^2
CIR	0.1193	0.6711	0.0000
Heston	0.2034	5.3295	0.0000
CNN-LSTM	0.0042	0.1873	0.8453
CIR-Heston-ANN	0.0097	0.1688	0.9515
CIR-Heston-CNN	0.0036	0.0249	0.9850
CIR-Heston-LSTM	0.0041	0.2011	0.9794
CIR-Heston-CNN-LSTM	0.0026	0.0875	0.9865

3.4. Empirical Analysis of Put Options

The predictions of the CSI 300 ETF put options by different prediction models can be observed in Figure 8 and Table 6. It can be observed that, compared with call options, the prediction accuracy of the pricing model for put options has decreased overall. Among them, the CIR model has the worst forecasting performance, with obvious overestimation in the first and middle part of the forecast period, and it is very different from the real price trend in the late forecast period. By observing the prediction effect of the Heston model, although there is a clear tendency to underestimate, the general trend is similar to the real price. In the prediction of the last 50 prediction samples, Heston's prediction deviates significantly from the real value. Table 4 shows that the MAPE of all models is not less than 1, and the Heston model has the worst accuracy, which is consistent with the conclusion of call options. It can be observed from the figure that a single parameter model has a clear tendency to underestimate the price of put options, and the overall performance is not as good as the hybrid neural network model. Its model stability is hundreds of times lower than the latter, and its prediction accuracy also dropped several times indicating that the actual option market does not meet the assumptions of the parametric model well. At the peak of the price, there are many models listed in the article that cannot be fully fitted, but the prediction effect of the neural network model is the worst, and when the price is extremely low, the prediction effect is greatly reduced. In the later stage of the forecast, the parameter model has a serious deviation from the actual market price trend. Compared with the CIR-Heston model combined with a single neural network, it can predict price trends overall, but prediction accuracy, model stability and generalization are still somewhat lacking compared to the hybrid model. With respect to comprehensive

model performance evaluation indicators, the dual-hybrid model proposed in this article has the best predictive effect, the MSE and R^2 of the model are the best performances of all reference models, and its MAPE has only dropped by less than 0.5 compared to CIR-Heston-CNN with the highest accuracy.

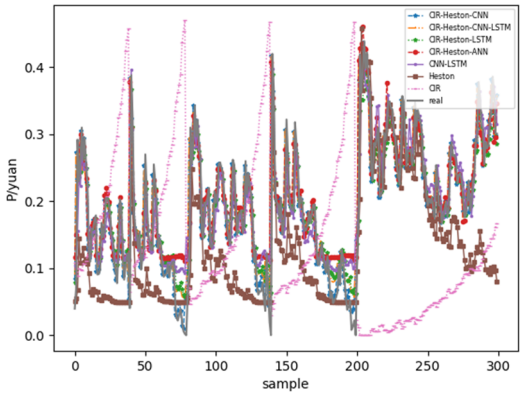


Figure 8. Comparison of real and predicted results of CSI 300ETF put options.

Table 6. Comparison of the prediction accuracy between the different models.

Model	MSE	MAPE	R^2
CIR	0.1288	12.0804	0.0000
Heston	0.1633	15.2168	0.0000
CNN-LSTM	0.0018	10.2812	0.7563
CIR-Heston-ANN	0.0005	5.0868	0.9451
CIR-Heston-CNN	7.0670	4.0608	0.9923
CIR-Heston-LSTM	0.0012	9.2066	0.8732
CIR-Heston-CNN-LSTM	0.0003	4.6382	0.9865

In order to avoid the excessive pursuit of prediction accuracy, ignoring the stability and generalization of the model and increasing the complexity of the model, the dual-hybrid model proposed in this paper takes the above factors into account, thus showing good performance and providing good guiding significance for investors.

3.5. Ridge Regression Integration

In order to improve the prediction accuracy of the model, this paper integrates the dual-hybrid pricing model with the reference model. Taking into account the complexity of the model itself and the characteristics of multicollinearity, the ridge regression [40] integration method is selected. Ridge regression is a linear least squares method with L2 regularization. It was proposed by Hoerl and Kennard in 1970. It is actually a biased estimation regression method. In order to further explore the prediction accuracy of the model, for the same experimental sample, the predicted value of the above model is used as the independent variable; the ridge regression integrated input is performed; the real option price is used as the dependent variable; and the ridge regression integrated output is performed. The experimental results are presented in Table 7:

It can be observed from the Table 7 that, in the ridge regression integration of the call and put option pricing models, the coefficient of the prediction result of the dual-mixed model is the largest, and the coefficient of the prediction result of the single parameter model is negative, showing a negative correlation with the dependent variable. After the above model is integrated by ridge regression, it is consistent with the above conclusion.

Table 7. Ridge regression coefficient table.

Model	Call Option	Put Option
CIR	0.0419	−0.0780
Heston	−0.0255	0.0719
CNN-LSTM	0.1724	0.2295
CIR-Heston-ANN	0.1405	0.0983
CIR-Heston-CNN	0.1970	0.1131
CIR-Heston-LSTM	0.2374	0.1023
CIR-Heston-CNN-LSTM	0.2408	0.3801

3.6. Model Robustness Test

This paper uses the method of adding noise to test the robustness of the dual-hybrid model. On the basis of random forest feature importance screening, the feature "High," which ranks first in importance ranking, is subjected to noise processing, and noise-added data are used as experimental data for model training and prediction again. This paper adds Gaussian noise with a mean value of 0 and a variance of 0.25 for "High" to test the pricing effect of the proposed model on CSI 300 ETF options. The results are presented in Figure 9 and Tables 8 and 9, and there is no significant difference between the prediction error of the dual-hybrid model for call options before and after noise addition. The prediction effect is that CIR-Heston-CNN performs best, which is consistent with the above empirical results. For put options, the performance of the models before adding noise is basically similar, and they are all maintained at a good level; after adding noise, the error distribution of the model begins to change, the degree of discretization of the error distribution of the reference model increases, and the average shifts to the right as a whole. The absolute value of skewness increased to varying degrees compared to before noise enhancement. The mean and variance of the error distribution of the dual-hybrid model remain basically unchanged, skewness has decreased and the coefficient of variation has also been improved, indicating that the dual-hybrid model can adapt to noise interference feature input and still maintain a good prediction effect. The dual-hybrid pricing model is more robust than other reference models.

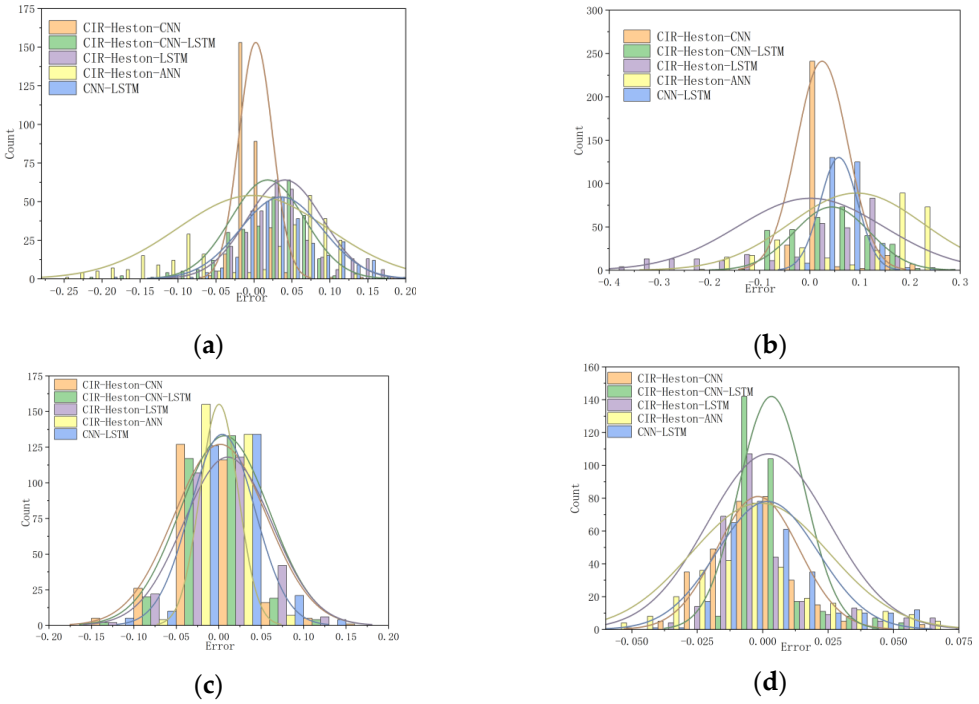


Figure 9. Error distribution (**a**) of unnoisy call option; (**b**) of noisy-up call option; (**c**) of unnoisy put option; and (**d**) of noisy-up put option.

Table 8. Call option model robustness tests for the descriptive statistic.

Model	Whether to Add Noise	Sum	Mean Value	Standard Deviation	Skewness	Kurtosis	Coefficient of Variation	Mode	Minimum	Median	Maximum
CIR-Heston-CNN	No	300	0.00	0.02	3.70	25.99	9.18	0.00	−0.04	0.00	0.19
	Yes	300	0.00	0.02	0.76	2.07	−8.74	−0.01	−0.04	0.00	0.06
CIR-Heston-CNN-LSTM	No	300	0.02	0.05	−1.16	2.10	2.68	−0.02	−0.20	0.03	0.11
	Yes	300	0.00	0.01	2.22	5.10	3.74	0.00	−0.02	0.00	0.06
CIR-Heston-LSTM	No	300	0.04	0.05	0.70	0.30	1.22	−0.04	−0.08	0.04	0.18
	Yes	300	0.00	0.02	1.95	4.53	10.35	−0.01	−0.04	0.00	0.11
CIR-Heston-ANN	No	300	0.00	0.10	−0.67	−0.88	−59.25	−0.12	−0.26	0.05	0.13
	Yes	300	0.00	0.03	0.76	0.47	−29.21	–	−0.06	0.00	0.07
CNN-LSTM	No	300	0.04	0.05	0.11	0.70	1.50	–	−0.19	0.03	0.19
	Yes	300	0.00	0.02	0.95	1.17	10.35	–	−0.06	0.00	0.06

Table 9. Put option model robustness tests for the descriptive statistic.

Model	Whether to Add Noise	Sum	Mean Value	Standard Deviation	Skewness	Kurtosis	Coefficient of Variation	Mode	Minimum	Median	Maximum
CIR-Heston-CNN	No	300	0.00	0.05	2.75	17.10	29.98	−0.01	−0.15	0.00	0.42
	Yes	300	0.02	0.05	2.84	6.64	2.07	0.01	−0.02	0.01	0.23
CIR-Heston-CNN-LSTM	No	300	0.01	0.05	3.03	19.28	7.60	−0.01	−0.15	0.00	0.42
	Yes	300	0.04	0.08	0.08	−1.03	1.70	–	−0.12	0.04	0.21
CIR-Heston-LSTM	No	300	0.01	0.05	2.60	16.27	4.93	−0.01	−0.12	0.01	0.41
	Yes	300	0.00	0.15	−2.93	−0.11	83.80	0.11	−0.38	0.05	0.19
CIR-Heston-ANN	No	300	0.00	0.02	0.21	0.90	55.27	0.00	−0.08	0.00	0.06
	Yes	300	0.09	0.13	−0.70	−1.03	1.42	–	−0.22	0.17	0.26
CNN-LSTM	No	300	0.00	0.04	−0.01	3.27	8.97	–	−0.15	0.00	0.15
	Yes	300	0.06	0.04	1.08	2.65	0.64	–	−0.04	0.05	0.21

4. Conclusions

The article introduces the latest research progress of economic models and deep learning models on option pricing issues and then analyzes applicable scenarios of a single pricing model. By reviewing the characteristics of the main pricing model, it provides a theoretical basis for the establishment of the dual-hybrid pricing model and conducts an empirical analysis based on the real historical data of CSI 300ETF options from January to December 2020. The article combines the stochastic interest rate model, stochastic volatility model, convolutional neural network and long short-term memory model according to the characteristics of ETF option data in the real market, which is a method based on CIR-Heston-CNN-LSTM and is a dual-hybrid model that combines the parameter model and neural network model. This article draws the following conclusions.

Firstly, the article established a dual-hybrid pricing model for CSI 300ETF options. The model can adapt to the characteristics of repeated fluctuations, high intensity and lag in option data. The stability and prediction accuracy of the model are significantly improved compared with the reference models.

Secondly, by using the random forest method, the feature importance ranking of call options and put options is obtained. The selected feature variables mainly include market transaction characteristics and features included in BS; Ridge regression integration can also illustrate the superiority of the proposed dual-hybrid model. The prediction accuracy and robustness outperformed the reference models.

Although the hybrid model proposed in the article has achieved better prediction accuracy, continuous research is still in progress. In order to predict the option price trend more quickly and accurately and to improve the performance of the model, more advanced deep learning models and more practical applications economic model can be applied. The article uses the combination of Heston and CIR to obtain a better prediction effect and proves the feasibility of the hybrid modeling method. The research can also consider the combination of the BS model, the classic binomial model and the classic difference method for different experimental data. The dual-hybrid model proposed in the article has broad application prospects and has huge potential in time series analysis, and it can be applied to portfolio management and asset allocation in the future [41–44].

This study can be further expanded to consider in the money options and out the money options. In the future, we can pay greater attention to the impact of policy imple-

mentation, changes in international financial markets, investor sentiment and other factors on option prices, and we can construct a pricing model that is more in line with the real situation in order to provide a good guiding role for investors.

Author Contributions: Conceptualization, K.Z. and J.Z.; methodology, K.Z.; software, K.Z.; validation, K.Z. and J.Z.; formal analysis, K.Z.; investigation, K.Z.; resources, K.Z.; data curation, K.Z.; writing—original draft preparation, K.Z.; writing—review and editing, K.Z., J.Z. and Q.L.; visualization, K.Z.; supervision, K.Z., J.Z. and Q.L; project administration, J.Z.; funding acquisition, J.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Nature Science Foundation of China, Grant Number 51675161.

Data Availability Statement: The data presented in this study are available on https://www.wind.com.cn (accessed on 7 January 2022).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Black, F.; Scholes, M. The pricing of options and corporate liabilities. *J. Political Econ.* **1973**, *81*, 637–659.
2. Han, Y.C.; Liu, C.Y.; Song, Q.S. Pricing double volatility barriers option under stochastic volatility. *Stochastics* **2020**, *93*, 625–645.
3. Liu, Z.P. Option Pricing Formulas in a new uncertain mean-reverting stock model with floating interest rate. *Discret. Dyn. Nat. Soc.* **2020**, *3*, 1–8.
4. Zhang, L.D.; Sun, Y.; Meng, X. European spread option pricing with the floating interest rate for uncertain financial market. *Math. Probl. Eng.* **2020**, *8*, 1–8.
5. He, X.J.; Lin, S. A fractional Black-Scholes model with stochastic volatility and European option pricing. *Expert Syst. Appl.* **2021**, *178*, 114983.
6. Kim, D.H.; Choi, S.Y.; Yoon, J.H. Pricing of vulnerable options under hybrid stochastic and local volatility. *Chaos Solitons Fractals* **2021**, *146*, 110–118.
7. Wang, L.; Zhang, S. Pricing the Asian option under vasier interest rate. *Acat Math. Appl. Sin.* **2003**, *26*, 15–24.
8. Li, S.; Li, S.H. Exotic options pricing formulae with stochastic interest rates. *Acta Math. Sin.* **2008**, *51*, 299–310.
9. Huang, W.; Tao, X.; Li, S. Pricing formulae for european options under the fractional vasicek interest rate model. *Acta Math. Sin.* **2012**, *55*, 219–230.
10. Geng, Y.J.; Zhou, S.W.; Mathematics, D.O. Pricing Asian option under mixed jump-fraction process. *J. East China Norm. Univ.* **2017**, *3*, 29–38.
11. Zhou, X.; Pan, J. Pricing of power options based on mixed fractional Hull-White interest rate model. *J. Hefei Univ. Technol.* **2017**, *40*, 847–853.
12. Wang, Y.; Xue, H. Vulnerable Option Pricing under Bi-fractional Jump—Diffusion Process. *J. Hangzhou Norm. Univ.* **2018**, *17*, 437–442.
13. Yang, Z. Pricing European Lookback Option in a Special Kind of Mixed Jump-Diffusion Black-Scholes Model. *Acta Math. Sci.* **2019**, *39*, 1514–1531.
14. Yang, Y.; Liu, G.; Wang, Y. Barrier Option Pricing Based on CIR Stochastic Volatility Model. *Nat. Sci. J. Harbin Norm. Univ.* **2019**, *35*, 1–4.
15. Djeutcha, E.; Fono, L.A. Pricing for options in a Hull-White-Vasicek volatility and interest rate model. *Appl. Math. Sci.* **2021**, *15*, 377–384.
16. Cox, J.C.; Ingersolli, J.E.; Ross, A. A Theory of the Term Structure of Interest Rates. *Econometrica* **1985**, *53*, 385–407.
17. Chang, Y.; Wang, Y.; Zhang, S. Option Pricing under Double Heston Jump-Diffusion Model with Approximative Fractional Stochastic Volatility. *Math. Probl. Eng.* **2021**, *9*, 126.
18. Heston, S.L. A closed-form solution for options with stochastic volatility with applications to bonds and currency options. *Rev. Financ. Stud.* **1993**, *6*, 327–343.
19. Heston, S.; Nandi, S.A. Closed-Form GARCH Option Valuation Model. *Rev. Financ. Stud.* **2000**, *13*, 585–625.
20. Zhang, L.J.; Zhang, W.Y.; Economics, S.O. Option pricing model by applying hybrid neural network based on heston model and genetic algorithm. *J. Ind. Eng. Eng. Manag.* **2018**, *32*, 142–149.
21. Sun, Y.; Tian, J.; Chen, Y. Pricing of Arithmetic Average Asian Option under the Fractional Jump Diffusion Heston Model. *J. Hangzhou Norm. Univ.* **2019**, *18*, 629–635.
22. Huang, C. An Empirical Study on the Volatility of CSI 300 Index based on Arch Model. In Proceedings of the 3rd Guangzhou International Forum on Finance (GZIFF), Guangzhou, China, 19 November 2020; pp. 292–299.
23. Gradojevic, N.; Ramazan, G.; Dragan, K. Option pricing with modular neural networks. *IEEE Trans. Neural Netw.* **2009**, *20*, 626–637. [PubMed]

24. Nikon, M.; Mansourfar, G.; Bagherzad, E.H.J. Stock price prediction using DEEP learning algorithm and its comparison with machine learning algorithms. Intelligent Systems in Accounting. *Financ. Manag.* **2019**, *26*, 164–174.
25. Ivacu, C.F. Option pricing using Machine Learning. *Expert Syst. Appl.* **2020**, *163*, 5–12.
26. Tan, D.D. S&P 500 Index Option Pricing Based on the BP Neural Networks. *Stat. Inf. Forum* **2008**, *11*, 40–43.
27. Ikram, A.; Liu, Y. Skeleton Based Dynamic Hand Gesture Recognition using LSTM and CNN. In Proceedings of the 2020 2nd International Conference on Image Processing and Machine Vision (IPMV), Bangkok, Thailand, 5 August 2020; pp. 63–68.
28. Xu, P.C.; Liu, B.Y. Interactive Behavior Recognition based on Image Enhancement and Deep CNN Learning. *Commun. Technol.* **2019**, *52*, 701–706.
29. Li, C.J.; Qu, Z.; Wang, S.Y.; Liu, L. A Method of Cross-layer Fusion Multi-object Detection and Recognition Based on Improved Faster R-CNN Model in Complex Traffic Environment. *Pattern Recognit. Lett.* **2021**, *145*, 127–134.
30. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780.
31. Xie, H.L.; You, T. Research on European Stock Index Options Pricing based on Deep Learning Algorithm:Evidence from 50ETF Options Markets. *Stat. Inf. Forum* **2018**, *33*, 99–106.
32. Jun, R.; Jianhua, W.; Chuanmei, W.; Jianxiang, W. Stock Index Forecast based on Regularized LSTM Model. *Comput. Appl. Softw.* **2018**, *35*, 44–48.
33. Feng, Y.X.; Li, Y.M. A Research on The CSI 300 Index Prediction Model Based on Lstm Neural Network. *Math. Pract. Theory* **2019**, *49*, 308–315.
34. Rai, H.M.; Chatterjee, K. Hybrid CNN-LSTM deep learning model and ensemble technique for automatic detection of myocardial infarction using big ECG data. *Appl. Intell.* **2021**, 1–19. [CrossRef]
35. Anuradha, J. Big data based stock trend prediction using deep CNN with reinforcement-LSTM model. *Int. J. Syst. Assur. Eng. Manag.* **2021**, *1*, 1–11. [CrossRef]
36. Ashok, J.; Rajan, E.G. Principal Component Analysis Based Image Recognition. *Int. J. Comput. Sci. Inf. Technol.* **2010**, *1*, 11–24.
37. Breiman, L. Random forests. *Mach. Learn* **2001**, *45*, 5–32.
38. Abdulkareem, N.M.; Abdulazeez, A.M. Machine Learning Classification Based on Radom Forest Algorithm: A Review. *Int. J. Sci. Bus.* **2021**, *5*, 128–142.
39. Yadav, K.; Sircar, A. Modeling parameters influencing city gas distribution sector based on factor analysis method. *Pet. Res.* **2021**, *1*, 1–11.
40. Alheety, M.I.; Kibria, B.M. A new version of unbiased ridge regression estimator under the stochastic restricted linear regression model. *Commun. Stat. Simul. Comput.* **2021**, *50*, 1–11.
41. Yaman, I.; Dalkili, T.E. A hybrid approach to cardinality constraint portfolio selection problem based on nonlinear neural network and genetic algorithm. *Expert Syst. Appl.* **2020**, *169*, 114517.
42. Spelta, A.; Pecora, N.; Pagnottoni, P. Chaos based portfolio selection: A nonlinear dynamics approach. *Expert Syst. Appl.* **2022**, *188*, 16005.
43. Peter, J.M.; Yuehua, W.; Hong, X. Portfolio Optimization for Binary Options Based on Relative Entropy. *Entropy* **2020**, *22*, 752.
44. Ma, Y.L.; Han, R.Z.; Wang, W.Z. Portfolio optimization with return prediction using deep learning and machine learning. *Expert Syst. Appl.* **2021**, *165*, 113973.