Advances in Applied Mathematics 应用数学进展, 2025, 14(4), 548-557
Published Online April 2025 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2025.144185

Hans汉斯

基于统计模型和机器学习模型的金融时间序列预测

臧　双*，孙德山

辽宁师范大学数学学院，辽宁　大连

收稿日期：2025年3月18日；录用日期：2025年4月11日；发布日期：2025年4月21日

摘　要

金融时间序列的预测一直是投资者关注的重点，针对金融时间序列预测难度大、准确度低、影响因素众多等问题，以深证成指、上证指数和中证500的收盘价为例，分别构建自回归求和移动平均(ARIMA)模型、卷积神经网络(CNN)模型、长短期记忆神经网络(LSTM)模型。CNN-LSTM模型和CNN-LSTM-ARMA模型用于这三种金融时间序列的收盘价预测的研究。通过比较，发现CNN-LSTM-ARMA模型比单一预测模型和CNN-LSTM模型预测收盘价的准确性高，拟合效果好。

关键词

金融时间序列，ARIMA模型，CNN模型，LSTM模型

Financial Time Series Forecasting Based on Statistical Models and Machine Learning Models

Shuang Zang*, Deshan Sun

School of Mathematics, Liaoning Normal University, Dalian Liaoning

Received: Mar. 18th, 2025; accepted: Apr. 11th, 2025; published: Apr. 21st, 2025

Abstract

The prediction of financial time series has always been the focus of investors' attention, in view of the problems of difficult forecasting, low accuracy and many influencing factors of financial time

*通讯作者。

文章引用: 臧双, 孙德山. 基于统计模型和机器学习模型的金融时间序列预测[J]. 应用数学进展, 2025, 14(4): 548-557. DOI: 10.12677/aam.2025.144185

series, taking the closing prices of Shenzhen Stock Exchange Component Index, Shanghai Composite Index and CSI 500 as examples, the Autoregressive Sum Moving Average (ARIMA) model, Convolutional Neural Network (CNN) model, Long Short-Term Memory Neural Network (LSTM) model are constructed respectively. The CNN-LSTM model and the CNN-LSTM-ARMA model are used for the study of closing price forecasting for these three financial time series. Through comparison, it is found that the CNN-LSTM-ARMA model has higher accuracy and better fitting effect than the single prediction model and the CNN-LSTM model in predicting the closing price.

Keywords

Financial Time Series, ARIMA Model, CNN Model, LSTM Model

1. 引言

随着社会经济的发展，人们的生活水平逐步提高，股市市场从无到有，越来越多的人开始进行投资。投资股票会得到回报，同时也伴随着一定的风险。股票价格是股票市场投资风向的直接体现[1]，所以投资者十分关注股票的价格，期望能够从过往的历史股票价格来找到规律，预测股票价格。但是，这并不是一个简单的问题。股票价格是一组非线性的，具有白噪声的，而且影响因素众多的数据。所以，众多学者开始关注这一具有挑战性的问题。

在探索金融数据规律的这条道路上，人们做出了很多的努力。研究方法也经历了一系列的变化，从统计模型到智能模型，于德亮利用自回归球和移动平均(Autoregressive Integrated Moving Average, ARIMA)模型对有滞后特征的金融数据进行预测，对其实用性进行了分析[2]，发现它短期预测的准确性较高，但是没有考虑其他干预因素。为了更深入地挖掘金融时间序列的特征，张贵勇对卷积神经网络(Convolutional Neural Networks, CNN)进行改进，应用到股票预测和汇率预测中[3]。为了解决循环神经网络(Recurrent Neural Network, RNN)模型存在的严重梯度下降快、无法收敛到最优解等问题，孙瑞奇对 RNN 模型进行修正，引入长短时记忆网络(Long Short Term Memory Network, LSTM)模型对中美股市进行预测[4]。不仅如此，还有更多的学者发现，将一些模型组合起来，在拟合并预测金融数据时会有更好的效果，曹超凡等在融合了 CNN 模型和 LSTM 模型基础上引入其他技术对股票因子进行预测[5]，取得了不错的效果。以上研究为本文的写作提供了思路，本文建立单一预测模型 ARIMA 模型、CNN 模型和 LSTM 模型，组合预测模型 CNN-LSTM 模型和 CNN-LSTM-ARMA 模型预测三组金融时间序列，对预测结果进行比较分析，找到其中预测精度最高和拟合效果最好的模型。

2. 模型介绍

2.1. CNN 模型

本 CNN 模型是一种具有局部连接、权重共享的深层前馈神经网络，适用于处理图像信息，能够克服连接网络的参数太多缺点，从而避免出现的过拟合等问题。卷积神经网络一般由卷积层、汇聚层和全连接层交叉堆叠而成。图 1 是一个简单的卷积神经网络。

卷积层是卷积神经网络最核心的特点，通过卷积操作可以对高维数据进行降维和对输入数据进行特征提取。股票价格数据往往蕴含了大量的特征集合，所以对特征的提取和选择尤为重要[6]。但是有时提

取后的维数还是很高，为此，引入下采样技术，也就是池化操作，从而实现对复杂数据更为准确的预测[7]。

Figure 1. Convolutional neural networks
图 1. 卷积神经网络

2.2. LSTM 模型

本 LSTM 模型是一种特殊的 RNN (循环神经网络)模型，能够学习长期依赖关系。LSTM 的计算过程是输入门作用与当前时刻的输入值，遗忘门作用于之前的记忆值，二者加权求和，得到新的汇总信息，最后通过输出们决定输出值。如果将 LSTM 在各个时刻的输出值进行展开，会发现部分最早时刻的输入值能够避免与权重矩阵的累次乘法，从而缓解梯度消失的问题[8]，这也是 LSTM 相较 RNN 的优势。

LSTM 在 RNN 的基础上加入了三个门的控制，分别为“输入门”“遗忘门”和“输出门”。3 个门的计算公式如下：

$$i_t = \sigma\left(W_{xi}x_t + W_{hi}h_{t-1} + b_i\right) \tag{1}$$

$$f_i = \sigma\left(W_{xf} + W_{hf}h_{t-1} + b_f\right) \tag{2}$$

$$o_t = \sigma\left(W_{xo} + W_{h0}h_{t-1} + b_o\right) \tag{3}$$

LSTM 模块中循环层及其更新公式：

$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh\left(W_{xc}x_t + W_{hc}h_{t-1} + b_c\right) \tag{4}$$

$$h_t = o_t \odot \tanh\left(c_t\right) \tag{5}$$

在公式中，i_t 是输入门，控制着当前时刻的输入有多少可以进入记忆单元，f_t 是遗忘门，遗忘门决定了记忆单元上一时刻的值有多少会被传到当前时刻，o_t 为输出门，输出门决定了记忆单元中存储的记忆值有多大比例可以被输出。σ 为 sigmoid 函数，c_{t-1} 是记忆单元在上一时刻的值，记忆值 c_t 是循环层神经元记住的上一个时刻的状态值，随着时间进行加权更新。参数集合 $\{W_{xi}, W_{xf}, W_{xo}\}$ 和 $\{b_i, b_f, b_o\}$ 对应不同门的权重矩阵和和偏置项，h_t 是 t 时刻的循环层状态的输出值。三个门的计算公式都是一样的，分别使用了自己的权重矩阵和偏置项量。LSTM 的计算过程是输入门作用于当前时刻的输入值，遗忘门作用于之前的记忆值，二者加权求和，得到汇总信息，最后通过输出门决定输出值。

3. 实证分析

3.1. 数据选取

数据选取深证成指、上证指数和中证 500 三组金融时间序列，日期从 2017-10-26 到 2023-12-25，一共 1500 条数据，为了评估单一模型和组合模型的性能和泛化能力，将数据划分为训练集和测试集，其中前 1200 个数据作为训练集，后 300 个数据作为测试集，数据包括开盘价、收盘价、最高价、最低价、成交量、成交额、振幅等 10 个指标，综合考虑各个指标对于收盘价的影响。每次预测都是基于前一个时间节点的数据来预测后一个时间节点，时序长度为 1 天。通过观察各个模型对金融时间序列收盘价预测的

结果，找到预测效果最好的模型。

3.2. 数据预处理

首先对数据进行预处理，主要是缺失值和异常值处理，查看数据集的前几行，初步了解数据的结构和内容。计算数据集中每列的空值数量，这里发现三组金融数据不存在空值。将数据集中名为“日期”的列设置为索引。为了避免量纲对数据分析的影响，对数据进行归一化处理，将收盘价的值缩放到 0 到 1 之间，把有量纲的表达式转变成无量纲的表达式，对于数据处理更加快捷迅速。

3.3. 数据分析

首先对股票收盘价趋势进行分析，如图 2 所示。

Figure 2. The closing price of the SZSE component index
图 2. 深证成指收盘价

由图 2 可以看出深证成指从 2018 年到 2019 年呈下降趋势，2019 年到 2021 年呈增长趋势，2021 年到 2024 年呈下降趋势。

接下来对股票每日收益率趋势进行分析，计算对数收益率，绘制对数收益率折线图，如图 3 所示。

由图 3 可以看出深证成指整体呈上下波动趋势，个别时间点波动性较大，没有明显的趋势特征。对对数收益率进行平稳性检验，ADF 检验的 p 值小于 0.05，可见对数收益率为平稳序列。JB 检验的 p 值远小于 0.01，不服从正态分布。

最后进行 ARCH 效应检验，可以考虑建立 GARCH 模型，常用 GARCH 模型有 GARCH(1,1)、GARCH(1,2)、GARCH(2,1)、GARCH(2,2)，对这四个模型分别建模，通过比较 AIC 值和 BIC 值确定 GARCH(1,1)为最佳模型。分析模型结果可以发现，在 2019 年上半年、2020 年 2 到 3 月份和 2022 年 4 月份对数收益率波动性较大，可能与当时的国情和政策有关[9]。

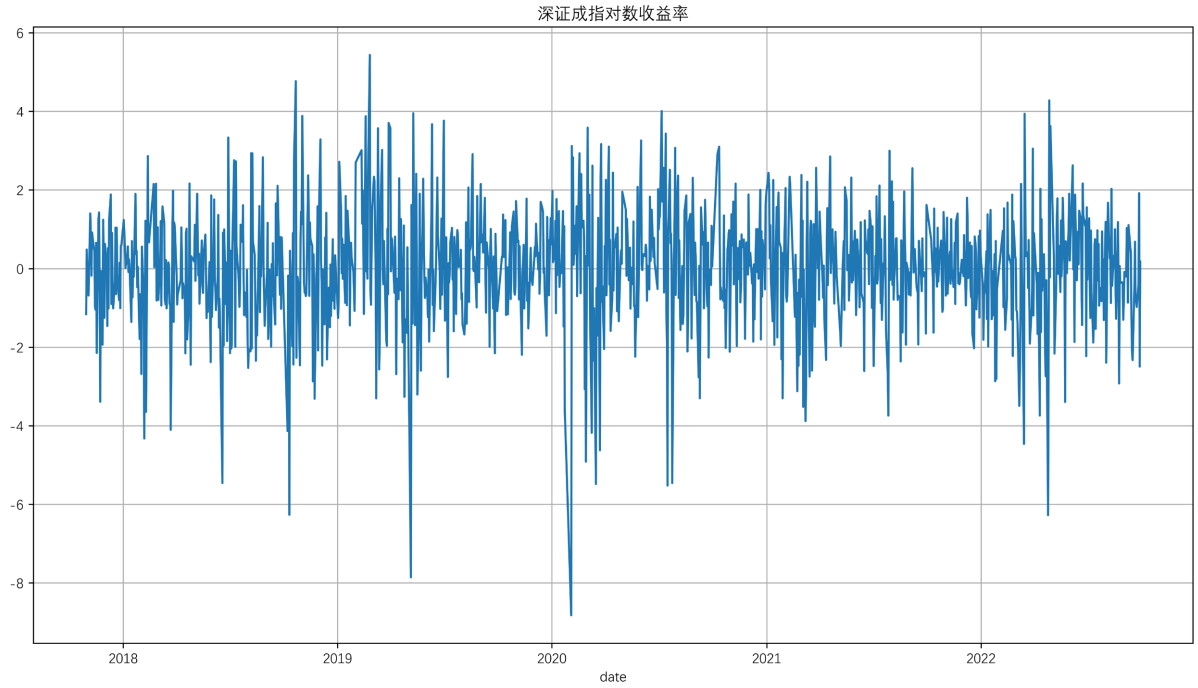

Figure 3. The logarithmic yield of the SZSE component index
图 3. 深证成指对数收益率

3.4. 评价指标

本文采用平均绝对误差(Mean Absolute Error, *MAE*)、均方误差(Mean Squared Error, *MSE*)和决定系数(R-Square, R^2)来评价模型的精准程度，*MAE* 能更好地反映预测值误差的实际情况，可以衡量真实值与预测值之间的误差绝对值的均值。*MSE* 是预测数据和原始数据对应点误差的平方和的均值。R^2 也称判定系数或者拟合优度。它是表征回归方程在多大程度上解释了因变量的变化，或者说方程对观测值的拟合程度如何。*MAE*、*MSE* 和 R^2 的公式如下：

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{y}_i - y_i\right| \tag{6}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2 \tag{7}$$

$$R^2 = \frac{SSR}{SST} \tag{8}$$

$$SSR = \sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2 \tag{9}$$

$$SST = \sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 \tag{10}$$

公式中 n 是样本数量，$\hat{y}_i$ 是预测收盘价，y_i 是实际收盘价，$\bar{y}$ 是收盘价均值。其中，*MAE* 和 *MSE* 的值越小，预测的准确度越高，R^2 越接近 1，拟合程度越好。

在预测未来收盘价后，将预测值反归一化，再将预测值代入公式，以便于原始数据进行比较。利用测试集中的实际收盘价和通过 ARIMA 模型、CNN 模型、LSTM 模型、CNN-LSTM 模型和 CNN-LSTM-

ARIMA 模型预测出来的收盘价来计算 *MAE*、*MSE* 和 R^2。这些评价指标除了判断预测的数值是否正确之外，还能够判断我们的模型是否拟合了足够多的数值之外的信息。

3.5. 本文的模型构建

3.5.1. ARIMA 模型的构建

ARIMA 模型与自回归移动平均(Auto-Regression and Moving Average, ARMA)模型有着紧密的联系，ARIMA 模型是在 ARMA 模型的基础上进行差分运算。ARIMA(*p*,*d*,*q*)进行结构分解可以分成自回归(Autoregressive, AR)模型、*d* 阶差分、移动平均(Moving Average, MA)模型。ARIMA 模型适合拟合非平稳的时间序列，因为通过 *d* 阶差分，非平稳序列会转化为平稳序列。

构建 ARIMA 模型的步骤如下：

步骤一：检验金融时间序列是否平稳，代码中调用 ADF 函数对数据进行单位根检验，如果 *p* 值小于显著性水平(通常为 0.05)，则可以拒绝原假设，即时间序列是平稳的。否则，不能拒绝原假设，时间序列可能是非平稳的。对不平稳的时间序列进行一阶差分后再次进行单位根检验，检查是否平稳，如果不平稳则重复差分操作。在本实验中，金融时间序列不平稳，经过一阶差分后平稳，深证成指数据经过一阶差分后得到的曲线如图 4 所示。

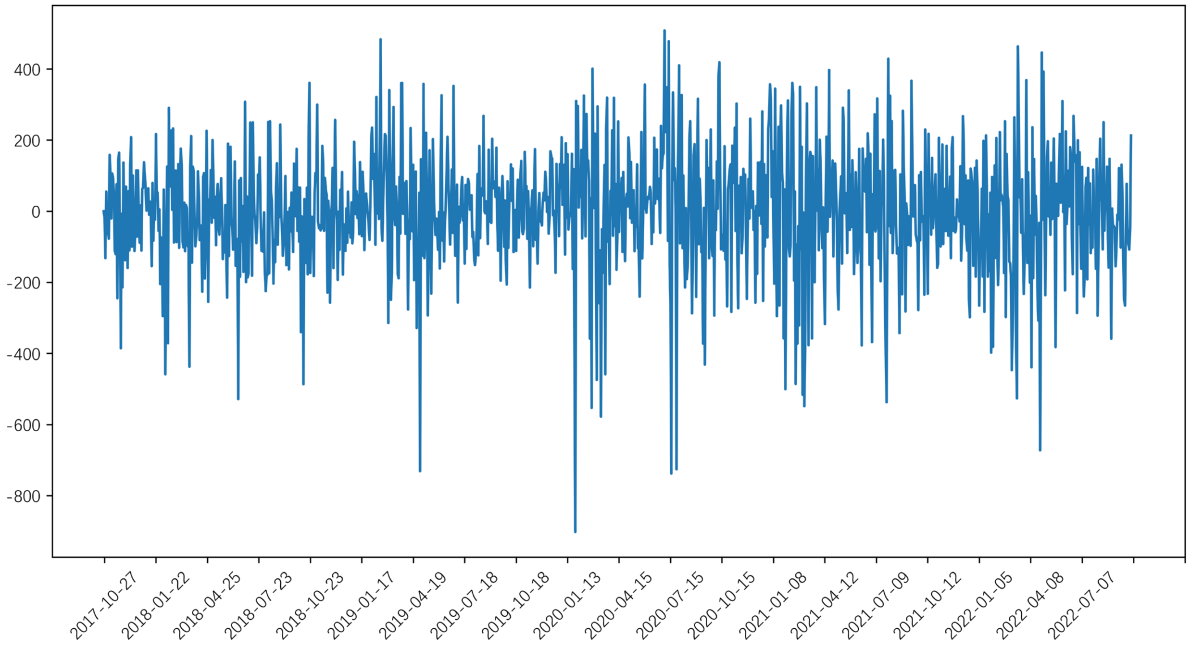

Figure 4. SZSE component index first-order difference data graph
图 4. 深证成指一阶差分后数据图

步骤二：ARIMA 模型参数设置。先确定差分参数 *d*。由步骤一可以得到差分一次后金融数据平稳，确定了 ARIMA 模型中的 *d* 是 1。再确定自回归参数 *p* 和移动平均参数 *q*。可以根据自相关函数和偏自相关函数的趋势特征初步判断模型的阶数。如图 5 所示。

模型可以构建多个参数进行组合，通过比较 AIC 准则和 BIC 准则，来确定 *p* 和 *q* 的值。AIC 函数中包含模型的独立参数个数和模型得出的极大似然估计值，当模型越复杂或者似然函数越小，AIC 值越大，而我们的目标是希望模型简单，并且模型的拟合度高，所以通常选择 AIC 值较小的模型。BIC 准则是对

AIC 准则的改进，考虑到了观测值数量，同 AIC，BIC 值越小越好。所以通常来说，选择较小的 AIC、BIC 的值所对应的参数 p 和 q，最终，选出 ARIMA(1,1,0)为最合适的模型。

步骤三：检验残差是否符合白噪声序列，通过 LB 法，得到 p 值大于 0.05，接受原假设，残差为白噪声序列，说明模型选择良好。

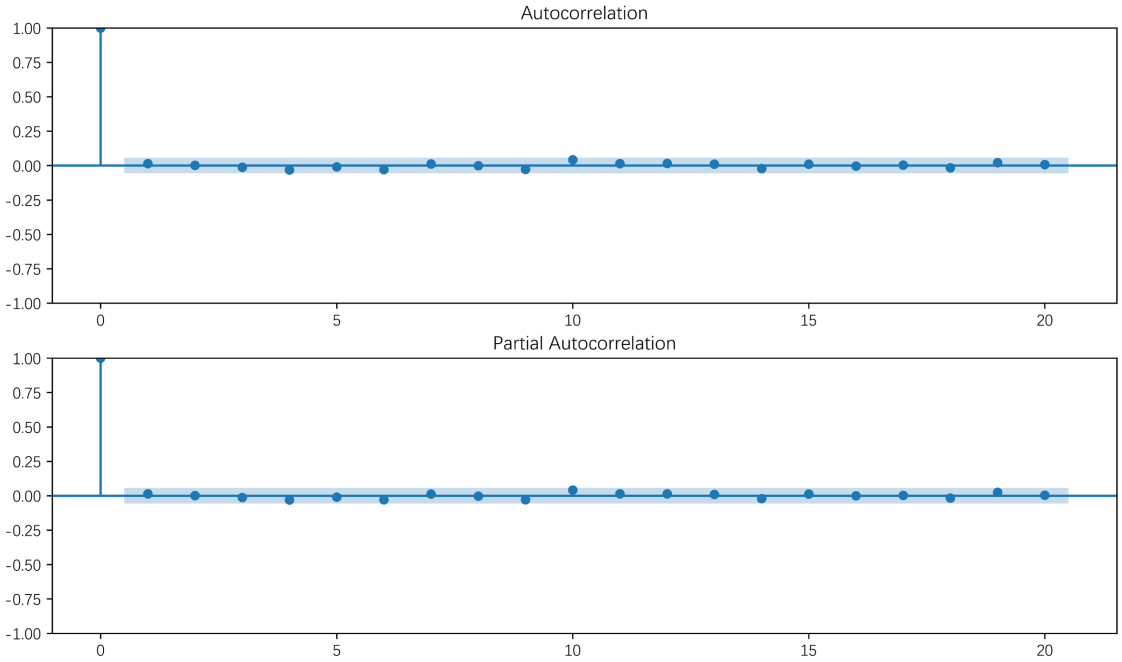

Figure 5. SZSE component refers to autocorrelation and partial autocorrelation results
图 5. 深证成指自相关和偏自相关结果

通过以上操作，建模结束，最后来预测收盘价未来走势，利用前一天的收盘价来预测后一天的收盘价。

3.5.2. CNN、LSTM 和 CNN-LSTM 模型的构建

论近年来，CNN-LSTM 模型已经被广泛应用于各种领域，它兼顾 CNN 模型特征提取的能力和 LSTM 模型处理时间序列数据的优势，展现出了较好的预测性能[10]。CNN-LSTM 模型是单变量时间序列预测的混合模型，主要是由 CNN 和 LSTM 两部分组成。将每个样本划分为进一步的子序列，CNN 模型将解释每一个子序列，LSTM 将子序列的解释拼凑在一起，将整个 CNN 模型包装在 TimeDistributed 包装层中，以便于将其应用样本中的每个子序列。然后在模型输出预测之前，LSTM 层对结果进行解释。该模型可以支持非常长的输入序列。

在本实验中，构建 CNN-LSTM 预测金融数据模型步骤如下：

步骤一：将金融时间序列数据的 10 个特征作为输入数据；

步骤二：创建一种线性堆叠神经网络层模型，添加时间分布的卷积层，这里 TimeDistributed 包装器用于将一个层应用于输入的每个时间步，对输入的数据进行特征提取；

步骤三：添加时间分布的最大池化层，添加时间分布的扁平化层，将多维数据转化为 1D 数据；

步骤四：添加 LSTM 层和全连接层，全连接层将前面 LSTM 层的输出连接到一个单一的神经元，这个神经元的输出将是模型的最终预测结果。

CNN 模型、LSTM 模型和 CNN-LSTM 模型都具有 10 个输入特征、1 个输出标签(收盘价)、Adam 优化器、激活函数 Relu 和均方误差损失函数编译模型，每个模型具体参数如表 1 所示。

Table 1. Model parameter settings
表 1. 模型参数设置

模型	参数设置
CNN	filters = 16, kernel_size = 2, pool_size = 2, batch_size = 20, stride = 1, flatten = 1, epochs = 100
LSTM	lstm units = 50, batch_size = 20, earning rate = 0.001, flatten = 1, epochs = 100
CNN-LSTM	filters = 64, kernel_size = 1, pool_size = 1, lstm units = 50, earning rate = 0.001, stride = 1, epochs = 100

3.5.3. CNN-LSTM-ARMA 模型的构建

论 CNN-LSTM-ARMA 模型在 CNN-LSTM 模型的基础上引入 ARMA 模型对金融时间序列的波动进行捕捉。首先利用数据训练集构建 CNN-LSTM 模型，接着利用训练好的 CNN-LSTM 模型对测试集进行预测，将收盘价预测值和真实的观察值进行比较，计算它们的残差，接下来对残差构建模型：

步骤一：对残差序列进行平稳性检验。经过检验，p 值小于 0.05，残差序列是平稳的，不需要再进行差分。

步骤二：对残差序列进行白噪声检验，经过检验，p 值小于 0.05，残差序列不是白噪声序列，残差中还有未被提取的有价值的信息。

步骤三：分析残差序列的自相关函数图和偏自相关函数图。从图 6 可以看出，自相关函数图和偏自相关函数图都具有拖尾性，可以建立 ARMA 模型。

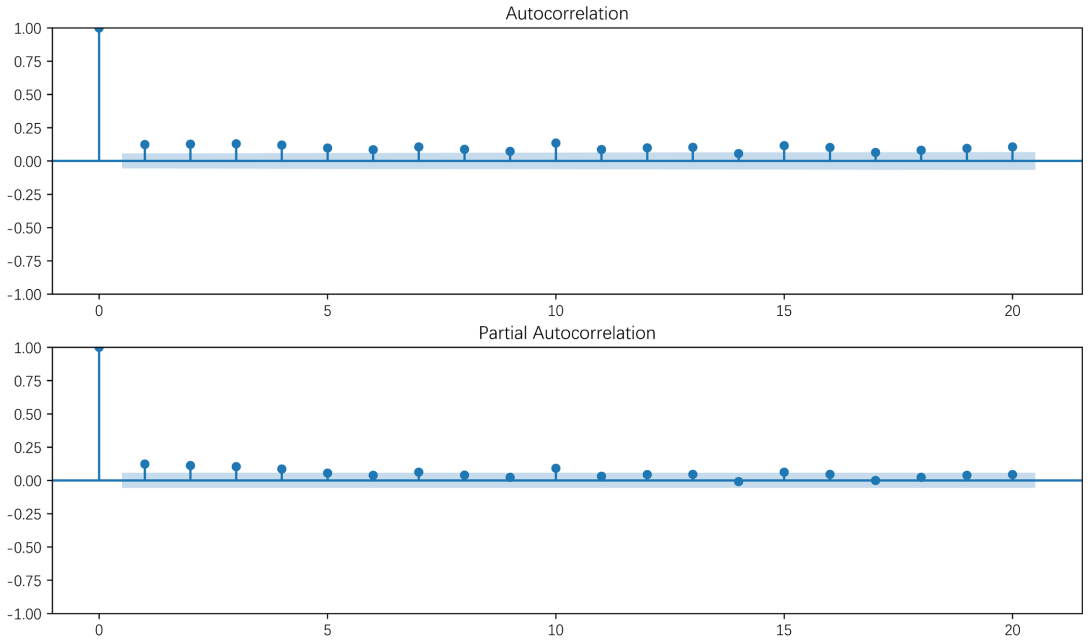

Figure 6. Autocorrelation and partial autocorrelation result plots of residual series
图 6. 残差序列自相关和偏自相关结果图

步骤四：确定 ARMA 模型的阶数。为了确认模型的显著性，建立 ARMA(1,0)模型，ARMA(0,1)模型，

ARMA(1,1)模型，ARMA(1,2)模型，ARMA(2,1)模型和 ARMA(2,2)模型，通过对比 AIC 和 BIC 值的大小，选取最佳模型最为残差的预测模型，预测时，可以将前一时间点的残差带入 ARMA 模型来计算后一时间点的残差值，从而修正 CNN-LSTM 模型预测值的误差。

通过 CNN-LSTM 模型预测出来的未来收盘价加上未来的残差修正，得到最终收盘价预测值。

3.6. 实证分析

3.6.1. 深证成指数据预测结果分析

深证成指即深圳成分股指数，是深圳证券交易所抽取的具有市场代表性的主要股指，综合反映深交所上市股的股价走势[11]，充分反映深圳市场的运行特征。本文选取深证成指，用训练集拟合模型，在测试集上评估模型。计算深证成指数据集在不同模型下的 *MAE*、*MSE* 和 R^2 的值，结果如表 2 所示。

Table 2. SZSE component index evaluation index
表 2. 深证成指评价指标

模型	*MAE*	*MSE*	R^2
ARIMA	127.4996	25074.5938	0.9511
CNN	106.3032	17800.8832	0.9653
LSTM	104.0285	16796.0251	0.9673
CNN-LSTM	101.5164	16503.6913	0.9678
CNN-LSTM-ARIMA	95.2927	15422.7537	0.9699

3.6.2. 上证指数数据预测结果分析

上证指数即上海证券综合指数，包含上海证券交易所全部上市股票，反映了上海证券交易市场的总体走势。上证指数在金融股票市场具有举足轻重的地位，能够全面衡量国内股票市场的稳定与繁荣程度，所以，对上证指数预测问题的研究在金融市场中存在广泛的研究价值[12]。本文选取上证指数，用训练集拟合模型，在测试集上评估模型。计算上证指数数据集在不同模型下的 *MAE*、*MSE* 和 R^2 的值，结果如表 3 所示。

Table 3. The evaluation index of the Shanghai Composite Index
表 3. 上证指数评价指标

模型	*MAE*	*MSE*	R^2
ARIMA	27.8120	1243.0523	0.9029
CNN	22.9092	860.5562	0.9328
LSTM	21.5448	786.6441	0.9386
CNN-LSTM	20.6583	735.7494	0.9426
CNN-LSTM-ARIMA	20.5093	723.6141	0.9435

3.6.3. 中证 500 数据预测结果分析

中证 500 指数是由全部 A 股中剔除沪深 300 指数成份股及总市值前 300 名的股票后，样本股由总市值排名靠前的 500 支股票组成，样本股的选择结构也基本符合整个市场的行业分布[13]，因此中证 500 指数在代表股市变化的同时，也能综合反映中国 A 股市场中一批中小市值公司的股票价格表现。本文选取

中证 500，用训练集拟合模型，在测试集上评估模型。计算上证指数数据集在不同模型下的 *MAE*、*MSE* 和 R^2 的值，结果如表 4 所示。

Table 4. CSI 500 evaluation index
表 4. 中证 500 评价指标

模型	*MAE*	*MSE*	R^2
ARIMA	58.8171	5507.3296	0.9336
CNN	48.7548	3746.0307	0.9549
LSTM	44.7167	3294.5913	0.9603
CNN-LSTM	43.3340	3047.1465	0.9633
CNN-LSTM-ARIMA	42.4737	3020.3120	0.9636

4. 总结

主分析对比 *MAE*、*MSE* 和 R^2 的值，对于深证成指、上证指数和中证 500 收盘价的预测结果，从单一模型来看，神经网络模型要比 ARIMA 模型准确性高，拟合效果好，神经网络模型中 LSTM 模型要比 CNN 模型的准确性高。部分学者在实证中分析得出在分析复杂的金融时间序列时[14]，神经网络模型比传统的统计模型更精确的结论，在本文所选的三组金融数据中也适用。从整体来看，组合模型要比单一模型的拟合效果更好，准确性更高，但是构建的过程要复杂一些。与其他模型相比，神经网络模型和传统预测模型的组合模型 CNN-LSTM-ARIMA 的 *MAE* 和 *MSE* 的值更小，对金融时间序列的预测精度高于其他模型，其决定系数 R^2 最接近 1，拟合效果最好，不过运算时所用的时间较长。

参考文献

[1] 管健. 基于 RNN-CNN 模型股票价格预测方法研究[D]: [硕士学位论文]. 南京: 南京信息工程大学, 2024.

[2] 于德亮. ARIMA 模型在基本建设投资预测中的应用[J]. 江苏统计, 2001(2): 28-29.

[3] 张贵勇. 改进的卷积神经网络在金融预测中的应用研究[D]: [硕士学位论文]. 郑州: 郑州大学, 2016.

[4] 孙瑞奇. 基于 LSTM 神经网络的美股股指价格趋势预测模型的研究[D]: [硕士学位论文]. 北京: 首都经济贸易大学, 2016.

[5] 曹超凡, 罗泽南, 谢佳鑫, 等. MDT-CNN-LSTM 模型的股价预测研究[J]. 计算机工程与应用, 2022, 58(5): 280-286.

[6] 曹玉贵, 谢梦醒. 基于 WD-CNN-LSTM 模型的股票价格预测分析[J]. 华北水利水电大学学报(社会科学版), 2023, 39(5): 15-22.

[7] 雒亚锋, 赵庆生, 梁定康, 王旭平. 基于金融技术指标与 CNN-BiLSTM 网络的短期用电量预测[J/OL]. 计算机仿真: 1-7. https://link.cnki.net/urlid/11.3724.TP.20240328.1153.016, 2024-04-02.

[8] 陈治颖. LSTM 模型优化及其在中国股票指数预测中的对比研究[D]: [硕士学位论文]. 济南: 山东财经大学, 2024.

[9] 陈苍, 赵志琴. 基于 GARCH 族模型的沪深 300 指数波动性研究[J]. 全国流通经济, 2021(15): 110-112.

[10] 王夷龙, 张生润, 唐小卫, 等. 基于 CNN-LSTM 混合模型的航空公司机票价格预测研究[J]. 北京交通大学学报, 2024, 48(5): 21-29.

[11] 王雪. 基于随机波动率模型与 GARCH 模型的资本市场研究[J]. 西部金融, 2021(3): 49-56.

[12] 李铖健, 孙海燕. 基于 CNN 及 LSTM 融合模型的上证指数预测[J]. 计算机仿真, 2024, 41(7): 299-302, 435.

[13] 黄珣. 基于组合模型的股票价格指数预测方法研究[D]: [硕士学位论文]. 上海: 上海财经大学, 2023.

[14] 管学英. 基于 ARIMA-RNN 混合模型的股价预测[J]. 哈尔滨商业大学学报(自然科学版), 2024, 40(2): 250-256.