Hans Journal of Data Mining 数据挖掘, 2022, 12(2), 173-181
Published Online April 2022 in Hans. http://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2022.122018

Hans汉斯

比较分析深度学习方法在沪深300指数价格预测的研究

王佳帅，陈立波

长沙理工大学数学与统计学院，湖南 长沙

收稿日期：2022年3月12日；录用日期：2022年4月12日；发布日期：2022年4月20日

摘 要

沪深300指数价格变动反映市场股票价格变动趋势，是投资者最关注的问题之一。如何构建合适的模型拟合价格时间序列变成了解决这一问题的关键之处。本文探究了不同深度学习方法对于价格的预测情况，分析得到几点探索性建议。实证分析中，数据选择沪深300指数2016年3月至2021年3月的价格数据，包括每日开盘价、最高价、最低价、收盘价四个数据特征共1218条数据，并对不同模型预测结果通过评价性指标进行对比分析。结果表明，对于数据信息利用更充分的模型，预测效果更好。

关键词

沪深300指数价格，深度学习方法，评价性指标

Comparative Analysis of Deep Learning Methods in the Price Prediction of the Shanghai and Shenzhen 300 Index

Jiashuai Wang, Libo Cheng

School of Mathematics and Statistics, Changsha University of Technology, Changsha Hunan

Received: Mar. 12th, 2022; accepted: Apr. 12th, 2022; published: Apr. 20th, 2022

Abstract

The price change of the CSI300 price Index reflects the trend of market stock price changes, which is one of the most concerned issues for investors. How to build a suitable model to fit the price time series has become the key to solving this problem. This article explores the price predictions

文章引用: 王佳帅, 陈立波. 比较分析深度学习方法在沪深 300 指数价格预测的研究[J]. 数据挖掘, 2022, 12(2): 173-181. DOI: 10.12677/hjdm.2022.122018

of different deep learning methods, and acquires several exploratory suggestions. In the empirical, this paper takes the CSI300 price index as the research object and the data from March 2016 to March 2021 is selected, including daily opening price, highest price, lowest price and closing price, with a total of 1218 data. Comparing and analyzing the prediction results of different models through evaluative indicators show that Models with more adequate use of data and information provide better prediction results.

Keywords

CSI300, Deep Learning Methods, Evaluative Index

1. 引言

随着经济的发展，我国的股票市场建设真正不断加强。股票市场行情的涨落与国民经济的发展密切相关[1]。沪深300指数价格是衡量股票市场总体价格水平和变化趋势的指标[2]，是投资者分析市场整体价格变化的重要依据和预测市场价格的重要工具。近年来，有许多学者涉及金融时间序列价格预测的研究，对金融资产价格走势的预测本身就是一项极具挑战性的任务，而现有的模型大多效果一般，ARMA模型和ARIMA模型[3] [4]是传统序列预测最普遍的模型。然而，随着真实数据的高复杂性、无规律性、随机性和非线性存在[5]，很难使用传统的复杂模型实现价格预测的准确性。随着机器学习方法[6]的发展，深度学习模型可以获得比传统统计模型更准确的预测效果。

众所周知，深度学习具有不依赖于先验知识、从大量原始数据提取特征这一特点，它对于金融数据的研究具有很大的潜力。RNN神经网络[7]是经常被用来作为价格时间序列最有效的方法，然而随着序列长度的增加，RNN模型容易出现梯度消失问题。LSTM模型[8] [9]是基于RNN模型设计的深度学习模型，并且此模型通过添加门机构，解决了RNN模型的记忆储存和遗忘的问题，具备了长序列价格序列的预测能力。随着社会要求模型预测效果的不断提高，不同深度学习方法也被构建并加入网络模型中，CNN模型[10]中卷积层的作用是捕捉输入数据内在关系一种方法，然而根据卷积核的特性，其对所有输入数据都进行同样的卷积，并没有识别不同输入数据的关联性的差异，Attention mechanisms [11]被构建很好的识别输入数据之间的差异性，有选择地连接相关信息，更好地抽取数据之间的特征信息。概率预测模型[12] [13]是一直根据概率知识构建的预测模型，此模型具有对不确定性建模、分析变量之间的关系、实现因果推理和随机生成样本数据的优点，DeepAR [14]模型是基于LSTM模型基础上添加概率预测的方法，实现更高的预测准确度和较小的预测误差。

为此，本文首先从中国股票市场的实际情况出发，以沪深300指数价格为研究对象，通过构建不同的深度学习模型来预测价格序列；其次，提出模型评价指标，通过对模型的综合评价比较，得出一些模型构建的探索性建议；最后，本文验证了不同深度学习方法用于深度模型对序列预测优劣性。

2. 时间序列分析与模型描述

时间序列问题通过使用历史序列值作为输入数据，区别为回归和分类 2 类基本问题。给定训练序列的滑动窗口特征 $X=\left(X_1,X_2,\cdots,X_T\right)$ 和 $X_t=\left(X_t^1,X_t^2,\cdots,X_t^L\right),X_t\in X$ 序列，定义时间步长度为 L 的间隔长

度，并给定历史值 $y=(y_1,y_2,\cdots,y_{T-1})$。预测序列未来趋势和特征，通常使用历史序列特征 X 和对应的目标值 y 学习非线性映射函数来预测未来值 y_T，对应模型公式：$y_T=f(X,y)$。

2.1. LSTM 模型

把数据 $X_t^{train}=\{X_{t-i},y_{t-i}\}_{i=1}^{p}$ 作为输入数据。如图 1，简单介绍了 LSTM 网络模型的结构，由三个门门结构组成的网络模型，分别为：输入门、遗忘门、输出门。

- 输入门(input gate)，作用确定信息被存放在细胞状态，由下列计算公式构成：

$$i_t=\sigma\left(W_t\times\left[h_{t-1},X_t^{train}\right]+b_i\right) \tag{1}$$

$$c_t'=\tanh\left(W_c\times\left[h_{t-1},X_t^{train}\right]+b_c\right) \tag{2}$$

- 遗忘门(forget gate)，作用是控制记忆信息遗忘和保留：

$$f_t=\sigma\left(W_f\times\left[h_{t-1},X_t^{train}\right]+b_f\right) \tag{3}$$

- 输出门(output gate)，作用是确定模型的输出值：

$$c_t=f_t\times c_{t-1}+i_t\times c_t' \tag{4}$$

$$o_t=\sigma\left(W_o\times\left[h_{t-1},X_t^{train}\right]+b_o\right) \tag{5}$$

$$h_t=o_t\times\tanh(c_t) \tag{6}$$

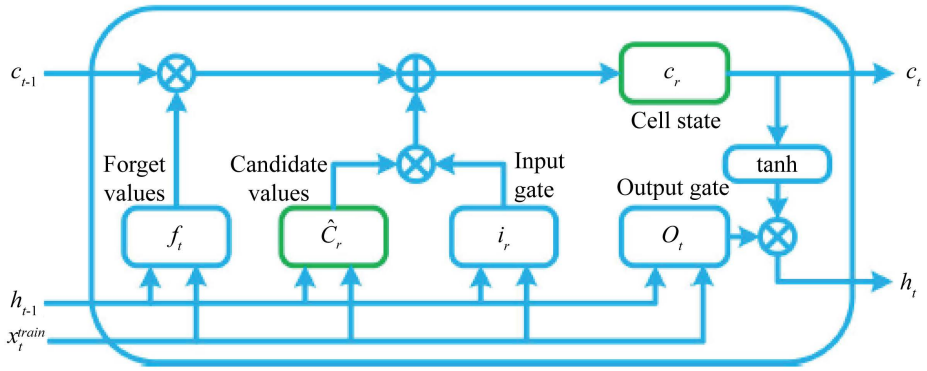

Figure 1. Framework of LSTM
图 1. LSTM 结构

2.2. CNNLSTM 模型

CNN 网络已经在图像分类、人脸识别和时间序列分析领域已经有非常成功的应用。CNN 的结构有三个主要网络层堆积：卷积层(convolution)、池化层(pooling)和全连接层(FC)。而其中卷积层的作用是从输入数据中抽取局部连接信息，识别输入数据不同位置的信息。而池化层目的是对卷积层进行作用，通下采样算子减少特征图的维度和避免模型的过拟合。全连接层一般用于最后几层，目的是组合由卷积层所产生的特征抽取来得到最后的模型输出，其中第 L 层卷积计算公式如下：

$$x_j^L=f\left(\sum_{i\in M_j}x_i^{L-1}*k_{ij}^L+b_j^L\right) \tag{7}$$

CNNLSTM 模型是在原始 LSTM 模型之中加入卷积层(convolution)和池化层(polling)，目的是对输入的时间序列数据先进行特征提取，提取数据之间相互的信息，然后再经过 LSTM 模型进行序列预测，其模型运行过程如下，对其输入数据 x，y：

$$y_j^L = \mathrm{LSTM}\left(\mathrm{maxpool}\left(\mathrm{cov}\left(x,y\right)\right)\right) \tag{8}$$

2.3. 修改的 Transformer 模型

Transformer 模型已经广泛用于序列任务，如机器翻译，时间序列预测等，其模型的最主要的贡献的是提出注意力机制(Attention Mechanism)，如图 2。

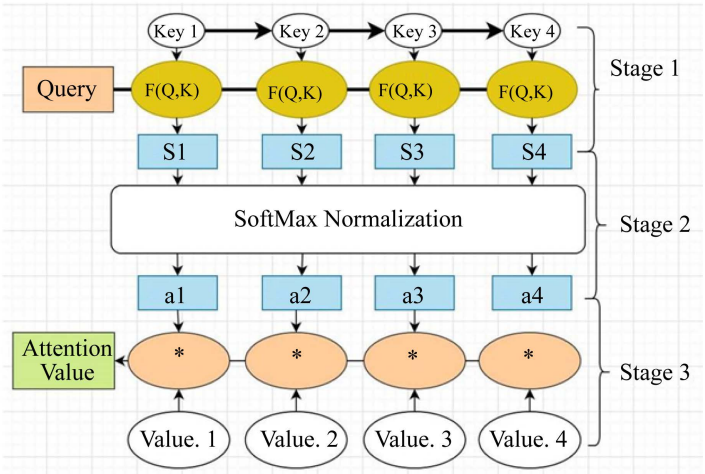

Figure 2. Framework of attention mechanism [15]
图 2. 注意力机制结构图[15]

其作用是捕捉输入序列之间的依赖性，并集中挑选出更有影响的信息，而忽略那些不需要的信息。注意力机制有三部分构成：Query，Key 和 Value，其函数结构是由 Q 到序列对 K-V 的映射。注意力机制分为 3 个步骤执行，对于输入序列 h_t，第一阶段通过作用权重 W_h 得到 Q，K，V，再通过激活函数得到 Q，K 之间的相似性，由计算公式(9)产生注意力分数，第二阶段通过对注意力分数进行标准化得到权重系数，最后能过(11)加权 V 求和得到具有相互依赖的序列 b。

$$s_t = \tanh\left(W_h \times h_t + b_h\right) \tag{9}$$

$$a_t = \frac{\exp\left(s_t\right)}{\sum_t \exp\left(s_t\right)} \tag{10}$$

$$b = \sum_t a_t \times v_t \tag{11}$$

修改的 Transform 模型主要对 Transform 模型的解码器进行改变，预测金融时间序列数据，本文将其编码器变为一层全连接层(FC)，其输出数据的维度对应训练数据标签的维度。

2.4. DeepAR 模型

用 $z_{i,t}$ 表示第 i 个序列在时间步 t 的值，$x_{i,t}$ 表示特征，t_0 表示预测初始时间。DeepAR 模型基于自回归循环神经网络预测 $z_{i,t}$ 的概率分布，用似然函数 $l\left(z_{i,t} \mid \theta_{i,t}\right)$ 表示。

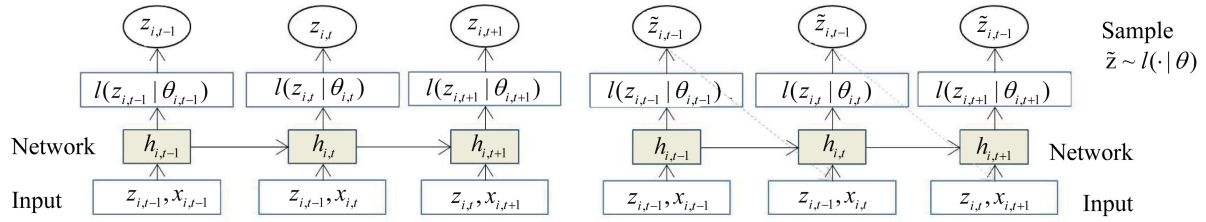

Figure 3. The left is the model training process, the right is the model prediction process
图 3. 左边为模型训练过程，右边为模型数据预测过程

如图 3，模型训练过程，网络输入 $x_{i,t}$、上一个时间步的取值 $z_{i,t-1}$，以及上一个时间步的状态 $h_{i,t-1}$。先计算当前的状态 $h_{i,t}=h\left(h_{i,t-1},z_{i,t-1},x_{i,t}\right)$，再计算似然 $l(z\mid\theta)$ 的参数 θ，最大化对数然函数(12)来学习网络参数：

$$L=\sum_i\sum_t \log l\left(z_{i,t}\mid\theta\left(h_{i,t}\right)\right) \tag{12}$$

模型预测过程，将 $t<t_0$ 的历史数据输入网络，获得初始状态 h_{i,t_0-1}，然后通过抽样得到预测结果，对于 $t_0,t_0+1,\cdots,T$，在每一个时刻随机抽样得到 $z_{i,t}\sim l\left(\cdot\mid\theta_{i,t}\right)$，这个样本值作为下一个输入。反复这个过程，我们可以得到一个序列，然后我们可以计算目标值的中位数、均值等当做模型预测值。$\theta\left(h_{i,t}\right)$ 的具体形式取决似然函数 $l\left(z\mid\theta\right)$，而似然函数的形式取决于数据本身的特征，本文采取 Gaussian 分布，则 $\theta=\left(u,\sigma\right)$，其似然函数(13)、分布均值(14)和方差(15)通过如下计算公式得到：

$$l_G\left(z\mid u,\sigma\right)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(\frac{-\left(z-u\right)^2}{2\sigma^2}\right) \tag{13}$$

$$u\left(h_{i,t}\right)=W_u^{\mathrm{T}}h_{i,t}+b_u \tag{14}$$

$$\sigma\left(h_{i,t}\right)=\log\left(1+\exp\left(W_u^{\mathrm{T}}h_{i,t}+b_u\right)\right) \tag{15}$$

3. 实验及结果

3.1. 数据来源

本文选取 1 只股票：沪深 300 指数(000300)，采用日期 2016 年 3 月 22 日至 2021 年 3 月 22 日的数据，数据特征包括每日开盘价、最高价、最低价、收盘价四个数据特征，共 1218 个样本数据(数据来源：万得数据)。

3.2. 数据预处理

为了让模型训练速度加快，本文对原始股票数据进行归一化处理，加快模型拟合过程。本文采取“min-max”归一化方法(16)对原始数据进行变换[16]，得到变换后数据输入模型进行预测。

$$X_t=\frac{X_t-\min\left(X_t\right)}{\max\left(X_t\right)-\min\left(X_t\right)} \tag{16}$$

3.3. 模型输入数据的构建

本文样本数据为 1218 条数据，首先我们通过数据按照一定的比例分为训练集和测试集数据，训练集样本数为 1178 条样本，测试集数据为 50 条样本。接下来所划分数据构建模型输入的数据结构，假设采

取时间长度 $L=5$，通过滑动窗口方法获取一个数据样本，如表 1。最后设置训练标签和输入数据，本文把一个数据样本的前四个数据作为模型输入，目标值为数据样本最后一个时刻的收盘价格，如图 4。

Table 1. Training sample data
表 1. 训练样本数据

交易日期	开盘价	最高价	最低价	收盘价
2021-03-10	5047.06	5055.28	4981.62	5003.61
2021-03-09	5066.15	5094.31	4917.91	4971
2021-03-08	5299.79	5326.26	5079.80	5080.02
2021-03-05	5191.97	5307.82	5174.25	5262.8
2021-03-04	5388.48	5392.37	5254.78	5280.71

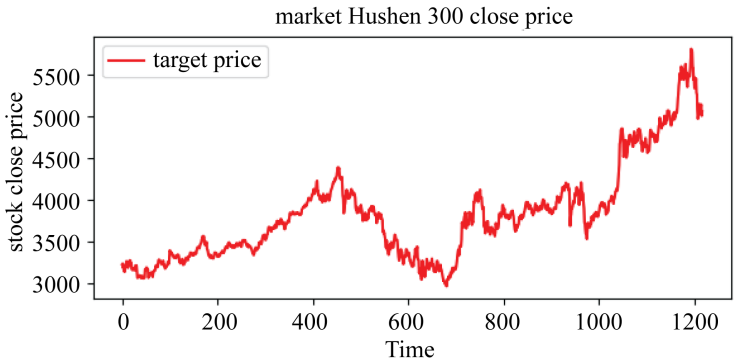

Figure 4. Target value for model training data
图 4. 模型训练数据的目标值

3.4. 训练参数设置和实验细节

本文挑选沪深 300 指数价格数据对每日收盘价进行预测，挑选 4 个模型进行比较，包括 LSTM、CNNLSTM、ModTransformer、DeepAR 探究在同一训练设置下，不同模型对于同一时间序列数据预测效果。对于本文设置的超参数如下：batch size = 12，learning rating = 0.005，loss = MSE，epoch = 150，multi-head = 8，GPU = GeForce GTX 1080 Ti

3.5. 评价指标

对于不同模型预测效果的评价，本文用到四个不同的评价指标：均方误差(MSE)，均方根误差(RMSE)，平均绝对误差(MAE)和 R^2。这四个评价指标[17]，通常是作为评价回归问题的准则，给定模型预测值和真实值，通过下列四个计算公式得到评价值：

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\hat{y_i}\right)^2} \tag{17}$$

$$\mathrm{MSE}=\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\hat{y_i}\right)^2 \tag{18}$$

$$\mathrm{MAE}=\frac{1}{n}\sum_{i=1}^{n}\left|y_i-\hat{y_i}\right| \tag{19}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i^{-} - y_i\right)^2}{\sum_{i=1}^{n}\left(y_i^{\sim} - y_i\right)^2} \tag{20}$$

3.6. 结果分析和探究

本文通过比较多个深度学习模型预测效果进行如下分析：股票数据验证样本为 50，通过对样本进行滑动窗口方法构建模型输入，留下 45 期数据进行模型验证和评估。如图 5，这是本文模型对于沪深 300 指数收盘价格的预测比较图，首先，红色曲线是股票收盘价格的真实价格，其浅蓝色曲线是模型 LSTM 训练完成后对验证数据的预测值，根据结果分析，在同等训练设置的背景下，模型 LSTM 收敛最慢，预测效果最差，因为 LSTM 模型的序列串行预测，训练速度慢且收敛性差。蓝色曲线是模型 DeepAR 模型预测曲线，DeepAR 模型通过前时刻信息预测下一时刻的数据分布，再根据数据分布进行采样，得到样本点，重复多次得到一个集合，本文选取这些集合的中位数点作为当前时刻的预测值，由于此模型也是串行预测，所以训练拟合速度也不是很快，总体而言，取中位数点当预测值降低了预测的偶然性和突变性，从图 5 可以看出，DeepAR 模型预测值与真实预测曲线保持一定误差，但不会有较大差异。黑色虚线是模型 ModTransformer 模型的预测值，本文通过把原始 Transformer 模型的解码器进行修改，得到了股票价格预测模型，通过观测，此模型预测效果与真实结果最接近，效果最好，收敛性快，根据注意力机制的作用，模型的输入数据得到更多有益信息，再经过模型预测得到较好结果。绿色虚线是 CNN + LSTM 模型，根据曲线拟合情况，比起原始 LSTM 模型，预测结果较好。

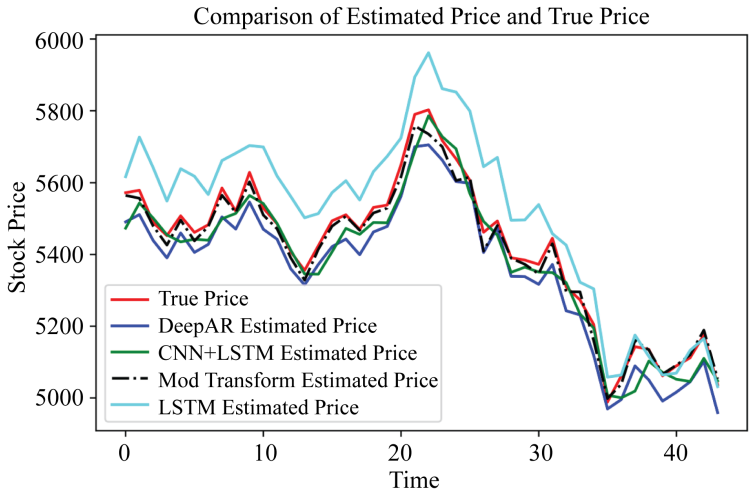

Figure 5. Prediction result of deep learning model for Hushen300 index price
图 5. 深度学习模型对于沪深 300 指数价格的预测图

根据表 2，我们得到模型的评价指标的值，我们了解到 ModTransformer 模型预测误差最小，根据 R2 值也可以肯定此模型拟合优度较高。在同等训练设置下，LSTM 模型在测试样本预测较差，但对模型添加卷积之后，模型得到较大改善，对比 DeepAR 模型，同样是串行预测，利用模型隐藏层数据对输出数据进行分布预测，通过大量采样得到模型预测值集合，再取数据集的中位数当预测值，此方法减少了模型预测数据的偶然性和并且降低预测误差。对此可以给出一些建设性意见：对于时间序列数据，对于输入数据，其实可以抽取数据之间内在信息，以此提高模型预测精度和训练拟合速度。对模型中隐藏层数

据进行再利用，能够降低模型预测误差。

Table 2. The evaluation index score of models
表 2. 模型的评价指标得分

	MSE	RMSE	MAE	R^2
LSTM	128.4601	10.5166	110.5997	0.9310
CNN + LSTM	64.5461	7.2836	53.0502	0.9540
DeepAR	61.6524	7.2677	52.8197	**0.9740**
ModTransformer	**34.9554**	**5.1822**	**26.8547**	**0.9790**

3.7. 消融实验

在本节中我们比较模型LSTM与模型CNN + LSTM进行消融实验探究[18]，根据表2，对于模型LSTM添加卷积层之后，模型得到较大改善。

4. 结论与展望

目前对于时间序列预测模型大致分为 2 种：传统时间序列模型与深度学习预测模型，传统模型对于现代大量级的数据难以建模和推理检验，而深度学习模型的出现对于处理这类问题迎来了发展，尽管许多深度网络在训练过程中往往存在梯度消失或者维数灾难问题，但随着社会发展，各种优异的模型已经被提出并且被应用于各种领域。本文推荐 4 个在时间序列数据预测效果较好的深度学习模型，通过互相比较，得到一些探究性的结果。LSTM 模型对于时间序列数据预测已经成为模型预测的基准，它通过对序列的串行预测，能得到较好的预测结果，而在 LSTM 模型基础上进行一些改进，模型预测效果得到较好的改善，本来通过对 LSTM 模型中添加卷积层，结果分析能够提高模型预测能力。根据 DeepAR 模型的测试结果，通过对模型中隐藏层数据利用也能降低模型预测误差。对比 CNN 提取输入数据信息，注意力机制可以有选择性的抽取与本身相似的数据特征，并成为数据内嵌信息，使模型预测结果较好。由此本文给出展望，对于数据输入，模型应充分利用数据输入信息和模型隐藏层数据信息，所以需研究改善提取输入数据之间的内在关系的方法和有效利用模型内部数据的方法，以此提高模型的预测精度。

参考文献

[1] 马超群, 杨竟澜, 任奕帅. 基于 H-LSTM 模型的沪深 300 指数价格预测[J]. 计量经济学报, 2021, 1(2): 437.

[2] Kong, A. and Zhu, H. (2018) Predicting Trend of High Frequency CSI 300 Index Using Adaptive Input Selection and Machine Learning Techniques. *Journal of Systems Science and Information*, **6**, 120-133. https://doi.org/10.21078/JSSI-2018-120-14

[3] Ariyo, A.A., Adewumi, A.O. and Ayo, C.K. (2014) Stock Price Prediction Using the ARIMA Model. 2014 *UKSim-AMSS 16th International Conference on Computer Modelling and Simulation*, Cambridge, 26-28 March 2014, 106-112. https://doi.org/10.1109/UKSim.2014.67

[4] Weiss, A.A. (1984) ARMA Models with ARCH Errors. *Journal of Time Series Analysis*, **5**, 129-143. https://doi.org/10.1111/j.1467-9892.1984.tb00382.x

[5] Kramer, M.A. (1991) Nonlinear Principal Component Analysis Using Autoassociative Neural Networks. *AIChE Journal*, **37**, 233-243. https://doi.org/10.1002/aic.690370209

[6] Baldi, P. and Brunak, S. (2001) Bioinformatics: The Machine Learning Approach. MIT Press, Cambridge.

[7] Zaremba, W., Sutskever, I. and Vinyals, O. (2014) Recurrent Neural Network Regularization. arXiv preprint arXiv:1409.2329.

[8] Graves, A. (2012) Long Short-Term Memory. Springer, Berlin, 37-45. https://doi.org/10.1007/978-3-642-24797-2_4

[9] Hochreiter, S. and Schmidhuber, J. (1997) Long Short-Term Memory. *Neural Computation*, **9**, 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

[10] Kalchbrenner, N., Grefenstette, E. and Blunsom, P. (2014) A Convolutional Neural Network for Modelling Sentences. arXiv preprint arXiv:1404.2188. https://doi.org/10.3115/v1/P14-1062

[11] Vaswani, A., Shazeer, N., Parmar, N., *et al*. (2017) Attention Is All You Need. *Advances in Neural Information Processing System*, **30**, 5998-6008.

[12] Li, Y., Chen, K.Z. and Wang, J. (2011) Development and Validation of a Clinical Prediction Model to Estimate the Probability of Malignancy in Solitary Pulmonary Nodules in Chinese People. *Clinical Lung Cancer*, **12**, 313-319. https://doi.org/10.1016/j.cllc.2011.06.005

[13] Beal, M.J., Ghahramani, Z. and Rasmussen, C.E. (2002) The Infinite Hidden Markov Model. *Advances in Neural Information Processing Systems*, **1**, 577-584.

[14] Salinas, D., Flunkert, V., Gasthaus, J., *et al*. (2020) DeepAR: Probabilistic Forecasting with Autoregressive Recurrent Networks. *International Journal of Forecasting*, **36**, 1181-1191. https://doi.org/10.1016/j.ijforecast.2019.07.001

[15] Kavianpour, P., Kavianpour, M., Jahani, E., *et al*. (2021) A CNN-BiLSTM Model with Attention Mechanism for Earthquake Prediction. arXiv preprint arXiv:2112.13444.

[16] Eesa, A.S. and Arabo, W.K. (2017) A Normalization Methods for Backpropagation: A Comparative Study. *Science Journal of University of Zakho*, **5**, 319-323. https://doi.org/10.25271/2017.5.4.381

[17] Willmott, C.J. (1982) Some Comments on the Evaluation of Model Performance. Bulletin of the American Meteorological Society, **63**, 1309-1313. https://doi.org/10.1175/1520-0477(1982)063<1309:SCOTEO>2.0.CO;2

[18] Du, L. (2020) How Much Deep Learning Does Neural Style Transfer Really Need? An Ablation Study. 2020 *IEEE Winter Conference on Applications of Computer Vision* (*WACV*), Snowmass, 1-5 March 2020, 3150-3159. https://doi.org/10.1109/WACV45572.2020.9093537