Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2020, 9(1), 16-23
Published Online February 2020 in Hans. http://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2020.91003

Disease Prediction Models Based on Hybrid Deep Learning Strategy

Min Liang[1], Yuchang Mo[1*], Dong Lin[2], Qian Lu[1], Ningning Li[1]

[1]Fujian Province University Key Laboratory of Computational Science, School of Mathematical Sciences, Huaqiao University, Quanzhou Fujian
[2]College of Acupuncture, Fujian University of Traditional Chinese Medicine, Fuzhou Fujian
Email: *yuchangmo@sina.com

Received: Dec. 30th, 2019; accepted: Jan. 14th, 2020; published: Jan. 21st, 2020

Abstract

Predictive models built using temporal data in electronic health records (EHRs) can potentially play a major role in improving management of diseases. Due to the sequence correlation and large feature space dimensions, traditional methods such as machine learning and non-deep neural networks are difficult to provide accurate predictions of disease. Recent works show that the long short term memory (LSTM) neural network outperforms most of those traditional methods for disease prediction problems. In this study, a hybrid deep learning neural network framework that combines convolutional neural network (CNN) with LSTM is proposed to further improve the prediction accuracy. Empirical studies using the real-world datasets in electronic health records have shown that using the proposed hybrid deep learning neural network for disease prediction significantly improves predictive performance compared to the use of support vector machine (SVM) model, CNN and LSTM alone.

Keywords

Electronic Health Record, Long Short Term Memory Neural Network, Convolutional Neural Network, Hybrid Deep Learning

基于混合深度学习算法的疾病预测模型

梁　敏[1]，莫毓昌[1*]，林　栋[2]，陆　迁[1]，李宁宁[1]

[1]华侨大学数学科学学院，计算科学福建省高校重点实验室，福建　泉州
[2]福建中医药大学针灸学院，福建　福州
Email: *yuchangmo@sina.com

收稿日期：2019年12月30日；录用日期：2020年1月14日；发布日期：2020年1月21日

*通讯作者。

文章引用: 梁敏, 莫毓昌, 林栋, 陆迁, 李宁宁. 基于混合深度学习算法的疾病预测模型[J]. 人工智能与机器人研究, 2020, 9(1): 16-23. DOI: 10.12677/airr.2020.91003

摘　要

利用电子健康档案中时间序列数据建立的预测模型在改善疾病管理方面发挥着重要作用。由于时态数据的序列相关性和特征空间维度大等特点，机器学习和非深度神经网络等传统方法难以提供疾病的准确预测。最新工作表明，长短时记忆(long short term memory, LSTM)神经网络性能优于大多数传统的疾病预测方法。为了进一步提高预测精度，本文提出了一种将卷积神经网络(convolutional neural network, CNN)与LSTM相结合的混合深度学习神经网络框架。使用电子健康档案中真实数据集的研究结果表明，相比传统SVM，CNN和LSTM模型，该算法的预测性能得到显著提高。

关键词

电子健康档案，长短时记忆网络，卷积神经网络，混合深度学习

1. 引言

通过追踪一段时间内对患者状态的重复测量，电子健康档案(electronic health records, EHRs)数据包含有关疾病演变的重要信息，该信息可用于构建潜在帮助预测疾病进展的模型。然而，仅在医疗保健事件期间或患者前往医院接受常规医疗护理时才记录患者数据，导致数据的不规则采样；同时对患者进行不同周期的追踪。因此，EHR 中存储的医学数据对建立预测模型提出了许多技术挑战：异构数据类型的集成和复杂纵向数据的分析。为了解决集成问题，研究者分别对知识层和数据层进行了研究。有些人依靠领域知识，通过定义来自不同数据类型[1]的标准来提取联合患者队列，而另一些人则研究在建模之前或建模后集成异构 EHR 数据的可能性[2]。本文的研究重点是后者：分析复杂的纵向数据。

传统的疾病预测方法将相似模式的患者聚类到同一子组以减少不规则性。此外，单变量数据预测仍然是机器学习领域最具挑战性的问题之一，因为大多数因变量是未知的。经典的单变量预测方法通常适用于其他特征难以度量或需要度量的变量太多的情况，例如股票市场指数预测问题[3]。在不需要额外信息的情况下，单变量预测方法十分灵活，只要 EHR 中有历史数据，所提出的方法就可以应用到其他患者的疾病预测。

近年来，深度学习神经网络(deep learning neural networks, DLNNs)在世界范围内得到了越来越广泛的应用，包括自然语言处理(natural language processing, NLP) [4]、图像目标检测、时间序列分析等领域。对于疾病预测问题，最近的研究工作表明，长短时记忆(long short term memory, LSTM)神经网络在预测[5]上提供极高的精度。实验结果表明，由于在循环神经网络(recurrent neural network, RNN)中引入了贮存长久信息的记忆门，仅使用 LSTM 神经网络，预测精度就超过了大多数传统的统计和机器学习方法，包括自回归综合移动平均(auto-regressive integrated moving average, ARIMA)模型，支持向量机(support vector Machine, SVM) [6]，非深度人工神经网络(non-deep artifificial neural networks, ANN) [7]及其组合。

此外，LSTM 神经网络是 RNN 的一种特殊形式[8]。还有其他类型的 DLNN，例如卷积神经网络(convolution neural networks, CNNs) [9]和深度信念网络(deep belief nets, DBN)。由特殊的一维卷积运算组

成的时态 CNN 对于时间序列预测问题也是十分有效[10]。在自然语言处理领域，有学者建议将时态 CNN 与 RNN 结合使用以获得更精确的分类结果[11]。

2. 相关工作

疾病预测在医疗诊断领域十分重要。传统的预测方法包括支持向量回归(support vector regression, SVR)，时间序列分析方法以及灰色模型(grey models, GMs) [12]。王等人[13]比较了使用 ARIMA 和 GM(1,1)模型进行的中国乙型肝炎月发病率预测的结果。马等人[14]使用季节 ARIMA 和 Holt-Wins 季节模型预测中国梅毒月发病率。张和李等人[15]提出了一种结合 ARIMA 模型和 SVR 模型的日放射科急诊病人流量预测方法。所有的单个基础预测模型都是以非线性的方式集成的，实验结果表明了所提出的混合方法的预测精度和可靠性。

深度学习神经网络是现代流行的处理大数据的机器学习技术，具有较高的分类和预测精度，已广泛应用于多个领域。与传统的人工神经网络 ANNs 相比，由于内部隐藏层和计算量的增多，DLNN 被用于更具挑战性的问题。Kann 等人[16]训练深度学习卷积神经网络来识别淋巴结转移和 ENE，其性能优于人类临床医生在历史上取得的成就。顾等人[17]提出了一种基于 GeoDetector 和 LSTM 进行手足口病预测的新方法，并将该模型扩展到其他传染病的时间序列预测。Chae 等人[18]使用深层神经网络(deep neural network, DNN)和 LSTM 模型预测传染病，结果表明，DNN 和 LSTM 比 ARIMA 具有更好的预测精度。

在本研究中，设计了一种将 LSTM 神经网络与 CNN 相结合的混合深度学习神经网络框架，用于解决疾病预测问题。通过增加使用 CNN 预处理阶段扩展传统的 LSTM 神经网络。预处理阶段从原始数据中提取有用的特征，通过一维卷积将单变量数据转换为多维数据，增强 LSTM 神经网络的预测能力。为了评估所提出框架的性能，使用 EHR 中真实数据集进行实验。实验结果表明，所提出的混合 DLNN 框架优于文献中现有的大多数方法，包括支持向量机(SVM)，单独的 CNN 和 LSTM 模型。本文的贡献包括：1) 引入一维卷积神经网络预处理单变量数据集，并经过两层时间卷积运算后将原始数据转换为多维特征数据；2) 提出的混合深度神经网络模型用于疾病预测。实验结果表明，提出的框架优于大多数现有方法，包括 SVM，CNN 和 LSTM。

3. 材料和方法

长短时记忆网络(LSTM)和卷积神经网络(CNN)是深度学习神经网络的两个热门分支，近年来，它们已引起了全世界的广泛关注。在本文中，针对解决时态数据预测问题的不规则性和序列长期依赖性，我们将 LSTM 和 CNN 结合起来，形成了一种混合式深度学习方法，与传统方法相比，该模型能够提供更准确，更可靠的预测结果。

使用真实世界的数据集，提出的框架用 CNN 对原始数据进行预处理，并利用 CNN 的输出来训练 LSTM 模型。

3.1. 数据描述

实验数据使用 Adadelta [19]。预处理后共有 578 个样本，其中阳性样本数 361 个，阴性样本数 217 个。我们以 0.8:0.1:0.1 的比例将数据集划分为训练、验证和测试集，训练集用于训练提出的 DLNN 框架；验证集是模型训练过程中单独留出的样本集，用于调整模型的超参数和用于对模型的能力进行初步评估；测试集用来评估最终模型的泛化能力。

3.2. 基于长短时记忆的循环神经网络

长短时记忆(LSTM)模型是循环神经网络(recurrent neural network, RNN)的一种特殊形式，可在每个神

经元处提供反馈。RNN 的输出不仅取决于当前神经元的输入和权重，还取决于先前神经元的输入。因此，从理论上讲，RNN 结构通常适用于处理时间序列数据。然而，在处理一系列长期相关的数据样本时，RNN 会出现梯度爆炸和梯度消失问题[20]，这成为后来引入 LSTM 模型的关键点[21]。

为了克服 RNN 模型的梯度消失问题，LSTM 模型包含贮存有用信息和丢弃无用信息的内部循环。LSTM 模型的流程图中有四个重要元素：单元状态，输入门，遗忘门和输出门(图 1)。输入、遗忘和输出门用于控制单元状态中包含信息的更新，维护和删除。前向计算过程可以表示为：

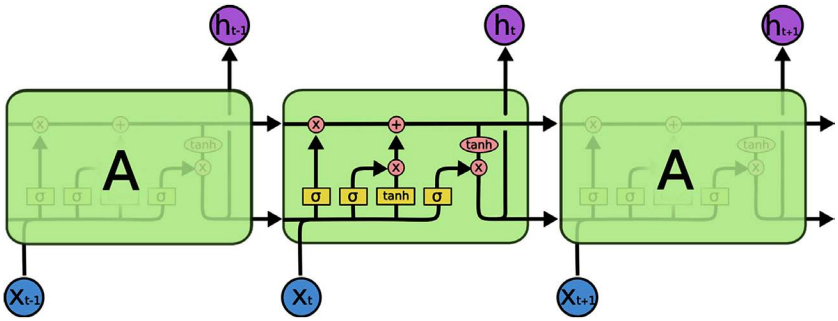

Figure 1. The training process of LSTM model
图 1. LSTM 模型的训练过程

$$f_t = \sigma\left(W_{fh} \cdot h_{t-1} + W_{fx} \cdot x_t + b_f\right) \tag{1}$$

$$i_t = \sigma\left(W_{ih} \cdot h_{t-1} + W_{ix} \cdot x_t + b_i\right) \tag{2}$$

$$\tilde{C}_t = \tanh\left(W_{ch} \cdot h_{t-1} + W_{cx} \cdot x_t + b_c\right) \tag{3}$$

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t \tag{4}$$

$$o_t = \sigma\left(W_{ot} \cdot h_{t-1} + W_{ox} \cdot x_t + b_o\right) \tag{5}$$

$$h_t = o_t \cdot \tanh\left(C_t\right) \tag{6}$$

其中 C_t，C_{t-1} 和 $\tilde{C}_t$ 分别表示当前单元状态值，上一时刻的单元状态值和当前单元状态值的更新。符号 f_t，i_t 和 o_t 分别表示遗忘门，输入门和输出门。在适当的参数设置下，根据等式(4)~(6)，基于 $\tilde{C}_t$ 和 C_t 的值计算输出值 h_t。根据输出值与实际值之间的差值，所有的权重矩阵通过时间反向传播算法(back-propagation through time, BPTT)进行更新[22]。

3.3. 时态卷积神经网络

卷积神经网络(CNN)可能是最常用的深度学习神经网络，目前主要应用于计算机视觉领域的图像识别/分类主题。对于大量原始数据样本，CNN 通常能够有效地提取输入数据的有用子集。一般来说，CNN 仍然是前馈神经网络，由多层神经网络(multi-layer neural network, MLNN)扩展而来。CNN 与传统 MLNN 的主要区别在于 CNN 具有稀疏交互和参数共享的特性[23]。

传统 MLNN 使用全连接策略在输入层和输出层之间建立神经网络，这意味着每个输出神经元都有机会与每个输入神经元进行交互。假设有 m 个输入神经元和 n 个输出神经元，权重矩阵有 $m \times n$ 个参数。CNN 通过设置大小为 $k \times k$ 的卷积核大大减少权重矩阵的参数。CNN 的两个属性提高了参数优化的训练效率；在相同的计算复杂度下，CNN 能够训练具有更多隐藏层的神经网络，即深层神经网络。

时态卷积神经网络引入了特殊的一维卷积，适用于处理单变量时间序列数据。时态 CNN 不像传统

CNN 那样使用 $k \times k$ 卷积核，而是使用大小为 $k \times 1$ 的卷积核。经过时间卷积运算之后，原始的单变量数据集可以扩展为 m 维特征的数据集。这样，时态 CNN 将一维卷积应用于时间序列数据，并将单变量数据集扩展为多维提取的特征(图 2 中的第一阶段)；扩展后的多维特征数据更适合使用 LSTM 进行预测。

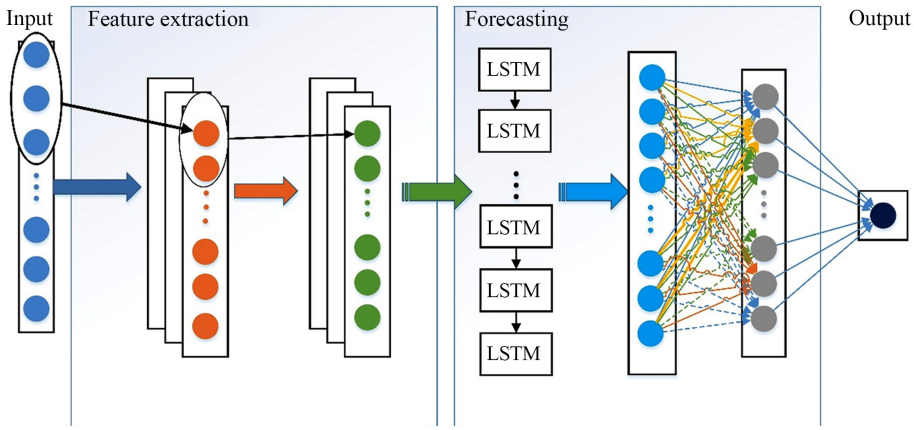

Figure 2. The proposed hybrid DNN disease prediction framework
图 2. 混合深度神经网络疾病预测框架

3.4. CNN-LSTM 预测框架

为了解决序列相关性和单变量数据这两个挑战，本文提出了一种结合CNN和LSTM模型的混合深度神经网络(deep neural network, DNN)。混合DNN框架的结构如图 2 所示。在预处理阶段，CNN从输入数据中提取重要信息，使用卷积将单变量输入数据重组为多维特征数据(图 2)。在第二阶段，将重组后的特征数据输入LSTM单元进行预测。

从图 2 可以看到，使用两个隐藏层的CNN预处理输入数据集。值得注意的是，当隐藏层的个数多于 5 个时，传统时态CNN通常包含池化操作防止过拟合。本文省略了池化操作以最大程度地保留提取的特征信息。

在对输入数据进行预处理之后，设计一个LSTM神经网络来训练和疾病预测。LSTM结构的训练过程如图 1 所示，其中从第一阶段提取的特征被作为训练LSTM模型的输入。为LSTM神经网络添加一个脱落层防止过拟合。预测输出值与实际输出值的差，即损失值用于优化所有LSTM单元的权重。优化过程遵循名为RMSprop的梯度下降优化算法，该算法通常用于深度神经网络的权重优化[24]。

4. 结果

使用Python3.7.3 (64 位)实现提出的混合DNN框架。基于Google提出的开源深度学习工具Tensorflow构建，并使用Keras 2.3.1 版本作为前端接口。

本研究提出的 CNN-LSTM 的预测结果与 SVM 模型，CNN 和 LSTM 等现有方法进行了比较。准确率是最常用的性能评价指标，即预测正确的样本占样本总数的比重。F1 值表示查准率与查全率之间的一种权衡，前者衡量正确预测为阳性的样本占全部预测为阳性样本数的比例，后者衡量正确预测为阳性的样本占全部实际为阳性的比例。本研究使用的另一个评价指标是 ROC 曲线下的区域(AUC)。ROC 曲线表示负阳性率与真阳性率的一种权衡，前者衡量分类器预测为阳性但实际为阴性的样本占所有阴性样本数的比例，后者衡量预测为阳性且实际为阳性的样本占所有阳性样本总数的比例。分析 ROC 曲线的常用方法是计算曲线下的区域面 AUC。这三种指标的值越大表示越高的预测精度。

使用不同方法建立了 4 个预测模型，表 1 给出了预测性能结果的总结。可以看出，模型的选择对预测性能产生一定的影响。总体来说，使用本研究提出的CNN-LSTM算法可以获得最好的预测性能。相比于LSTM，使用CNN-LSTM框架时，AUC提高了 6.5%，F1 值提高了 12.2%，Accuracy提高了 14.6%。实验结果表明混合深度学习算法更适合时态数据的疾病预测。

Table 1. Forecast results of different models
表 1. 不同模型预测结果

模型	SVM	CNN	LSTM	CNN-LSTM
Accuracy	0.456140	0.578947	0.719298	0.842105
F1	0.491830	0.519999	0.652174	0.742857
AUC	0.578676	0.632353	0.766176	0.819853

根据图 3 整体分析 CNN-LSTM 的 accuracy 和 loss 变化趋势。从图中可以看出，模型的训练集损失

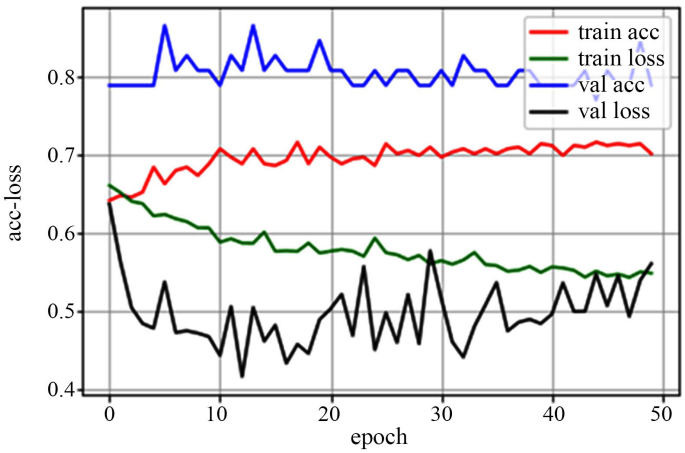

Figure 3. The trend change of accuracy-loss of CNN-LSTM
图 3. CNN-LSTM 的 acc-loss 趋势变化

下降较为平稳，训练集的准确率又能稳定上升，此模型相对其他算法模型具有良好的稳定性。

5. 结论和未来工作

本文提出了一种基于卷积神经网络(CNN)和长时记忆(LSTM)神经网络相结合的混合深度学习神经网络框架，用于单变量和序列相关数据的疾病预测。最近的研究工作已经表明，仅通过LSTM神经网络，就可以实现疾病的高预测精度。本研究进一步证明提出的混合框架优于传统的LSTM神经网络。CNN从原始数据中提取最有用的信息，并将单变量时态数据集转换为多维特征数据，进而促进LSTM的预测性能。

对于本研究的未来工作，我们打算将提出的CNN-LSTM框架应用于更复杂的实际医疗数据集，以验证提出框架的鲁棒性。

基金项目

国家自然科学基金项目(61572442)，福建省高校创新团队发展计划，福建省研究生导师团队，泉州市高层次人才团队项目(2017ZT012)。

参考文献

[1] Wei, W.-Q., Teixeira, P.L., Mo, H., Cronin, R.M., Warner, J.L. and Denny, J.C. (2015) Combining Billing Codes, Clinical Notes, and Medications from Electronic Health Records Provides Superior Phenotyping Performance. *Journal of the American Medical Informatics Association*, **23**, e20-e27. https://doi.org/10.1093/jamia/ocv130

[2] Henriksson, A., Zhao, J., Boström, H. and Dalianis, H. (2015) Modeling Heterogeneous Clinical Sequence Data in Semantic Space for Adverse Drug Event Detection. *IEEE International Conference on Data Science and Advanced Analytics*, Paris, 19-21 October 2015, 1-8. https://doi.org/10.1109/DSAA.2015.7344867

[3] Hsieh, T.J., Hsiao, H.F. and Yeh, W.C. (2011) Forecasting Stock Markets Using Wavelet Transforms and Recurrent Neural Networks: An Integrated System Based on Artificial Bee Colony Algorithm. *Applied Soft Computing*, **11**, 2510-2525. https://doi.org/10.1016/j.asoc.2010.09.007

[4] Socher, R., Lin, C.C., Manning, C. and Ng, A.Y. (2011) Parsing Natural Scenes and Natural Language with Recursive Neural Networks. *Proceedings of the* 28*th International Conference on Machine Learning*, Bellevue, 28 June-2 July 2011, 129-136.

[5] Kong, W., Dong, Z.Y., Hill, D.J., Luo, F. and Xu, Y. (2018) Short-Term Residential Load Forecasting Based on Resident Behaviour Learning. *IEEE Transactions on Power Systems*, **33**, 1087-1088. https://doi.org/10.1109/TPWRS.2017.2688178

[6] Yan, K., Du, Y. and Ren, Z. (2018) MPPT Perturbation Optimization of Photovoltaic Power Systems Based on Solar Irradiance Data Classification. *IEEE Transactions on Sustainable Energy*, **10**, 514-521. https://doi.org/10.1109/TSTE.2018.2834415

[7] Du, Y., Yan, K., Ren, Z. and Xiao, W. (2018) Designing Localized MPPT for PV Systems Using Fuzzy-Weighted Extreme Learning Machine. *Energies*, **11**, 2615. https://doi.org/10.3390/en11102615

[8] Funahashi, K.I. and Nakamura, Y. (1993) Approximation of Dynamical Systems by Continuous Time Recurrent Neural Networks. *Neural Networks*, **6**, 801-806. https://doi.org/10.1016/S0893-6080(05)80125-X

[9] Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2012) Imagenet Classification with Deep Convolutional Neural Networks. *Proceedings of the* 25*th International Conference on Neural Information Processing Systems*, Lake Tahoe, 3-6 December 2012, 1097-1105.

[10] Almalaq, A. and Edwards, G.A. (2017) Review of Deep Learning Methods Applied on Load Forecasting. *Proceedings of the* 16*th IEEE International Conference on Machine Learning and Applications*, Cancun, 18-21 December 2017, 511-516. https://doi.org/10.1109/ICMLA.2017.0-110

[11] Wang, J., Yu, L.C., Lai, K.R. and Zhang, X. (2016) Dimensional Sentiment Analysis Using a Regional CNN-LSTM Model. *Proceedings of the* 54*th Annual Meeting of the Association for Computational Linguistics*, Berlin, Germany, 7-12 August 2016, 225-230. https://doi.org/10.18653/v1/P16-2037

[12] Kumar, U. and Jain, V. (2010) Time Series Models (Grey-Markov, Grey Model with Rolling Mechanism and Singular Spectrum Analysis) to Forecast Energy Consumption in India. *Energy*, **35**, 1709-1716. https://doi.org/10.1016/j.energy.2009.12.021

[13] Wang, Y.-W., Shen, Z.-Z. and Jiang, Y. (2018) Comparison of ARIMA and GM(1,1) Models for Prediction of Hepatitis B in China. *PLoS ONE*, **13**, e0201987. https://doi.org/10.1371/journal.pone.0201987

[14] 马晓梅, 史鲁斌, 其木格. 基于 ARIMA 乘积季节模型和 Holt-Winters 季节模型的梅毒月发病率预测[J]. 郑州大学学报(医学版), 2018, 53(1): 79-84.

[15] Zhang, Y.M., Luo, L. and Yang, J.C. (2019) A Hybrid ARIMA-SVR Approach for Forecasting Emergency Patient Flow. *Journal of Ambient Intelligence and Humanized Computing*, **10**, 3315-3323. https://doi.org/10.1007/s12652-018-1059-x

[16] Kann, B.H., Aneja, S. and Loganadane, G.V. (2018) Pretreatment Identification of Head and Neck Cancer Nodal Metastasis and Extranodal Extension Using Deep Learning Neural Networks. *Scientific Reports*, **8**, Article No. 14036. https://doi.org/10.1038/s41598-018-32441-y

[17] Gu, J.Y., Liang, L.Z. and Song, H.Q. (2019) A Method for Hand-Foot-Mouth Disease Prediction Using Geo Detector and LSTM Model in Guangxi, China. *Scientific Reports*, **9**, Article No. 17928. https://doi.org/10.1038/s41598-019-54495-2

[18] Chae, S., Kwon, S. and Lee, D. (2018) Predicting Infectious Disease Using Deep Learning and Big Data. *International Journal of Environmental Research and Public Health*, **15**, 1596. https://doi.org/10.3390/ijerph15081596

[19] Zeiler, M.D. (2012) ADADELTA: An Adaptive Learning Rate Method.

[20] Jozefowicz, R., Zaremba, W. and Sutskever, I. (2015) An Empirical Exploration of Recurrent Network Architectures. *Proceedings of the* 32*nd International Conference on International Conference on Machine Learning*, Lille, 6-11 July

2015, 2342-2350.

[21] 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

[22] Werbos, P.J. (1990) Backpropagation through Time: What It Does and How to Do It. *Proceedings of the IEEE*, **78**, 1550-1560. https://doi.org/10.1109/5.58337

[23] Ketkar, N. (2017) Convolutional Neural Networks. In: *Deep Learning with Python*, Springer, Berlin, 63-78. https://doi.org/10.1007/978-1-4842-2766-4_5

[24] Goodfellow, I., Bengio, Y., Courville, A. and Bengio, Y. (2016) Deep Learning. MIT Press, Cambridge, 1.