Big Data Deep Learning Framework using Keras: A Case Study of Pneumonia Prediction

Conference Paper · December 2018

DOI: 10.1109/CCAA.2018.8777571

2 authors:

Karan Jakhar
Chandigarh University
2 PUBLICATIONS **2** CITATIONS
SEE PROFILE

Nishtha Hooda
Thapar University
18 PUBLICATIONS **34** CITATIONS
SEE PROFILE

2018
4th International Conference on Computing Communication and Automation (ICCCA)

Big Data Deep Learning Framework using Keras: A Case Study of Pneumonia Prediction

Karan Jakhar
Computer Science Engineering Department,
Chandigarh University
Mohali, Punjab, India
karanjakhar49@gmail.com

Nishtha Hooda
Computer Science and Engineering Department
Chandigarh University
Mohali, Punjab, India
27nishtha@gmail.com

***Abstract*—Big Data predictive analytics using machine learning techniques is currently a much active area of research in medical science. With increasing size and complexity of medical data like X-rays, deep learning gained huge success in prediction of many fatal diseases like pneumonia. In this research work, DCNN (deep convolutional neural networks) an efficient predicting model for big data, having deep layers is a proposed, which can classify whether a person is having a pneumonia or not. The experiments are carried after extracting the features of high quality X-ray images data and achieved an prediction accuracy of 84% and AUC of**

Promising results are found, when the results of the DCNN framework is compared with the regular classifiers like SVM, random forest, etc. using different evaluation metrics like accuracy, sensitivity, etc. With the appearance of increasing cases of pneumonia, tactful implementation of deep learning can play a big part in improving the performance of prediction of many fatal diseases in the future.

Keywords—big data; machine learning; prediction; deep learning; pneumonia

I. INTRODUCTION

Around one million adults are diagnosis with pneumonia and every year about fifty thousand die from this deadly in the US alone [1]. Pneumonia is effecting a lot children who are under age of five and also common cause of death of them worldwide [2]. Predicting Pneumonia is important in the medical field. Various tests can take some time but predicting by X-ray of chest will help the doctor to get an idea of the disease and steps can be taken accordingly. Detecting pneumonia by observing X-ray of chest is a complex task and is an active area of research.

Deep learning as well as Big Data are two popular fields in the rapidly growing digital world [3]. While Big Data has numerous definitions, this research work refer it to the veracity i.e. unstructured data as presented in the Figure 1, defining important Vs of big data [4]. The medical data is vast, complex, and difficult to analyze using conventional data analysis techniques. Hence, deep learning offers a great solution in harvesting valuable knowledge from such complex medical data.

In this research work, the proposed prediction model is implemented by convolutional deep neural networks (CNN) using Python programming language. CNNS, also known as ConvNets in deep learning. After pre-processing of data, different machine learning algorithms are trained to measure the performance of CNN with popular and modern classifiers. Promising results are achieved, when the results of the suggested framework is compared with the regular classifiers like SVM, random forest, adaboost, etc. using different estimating metrics like accuracy, specificity, area under the curve, and sensitivity, etc.

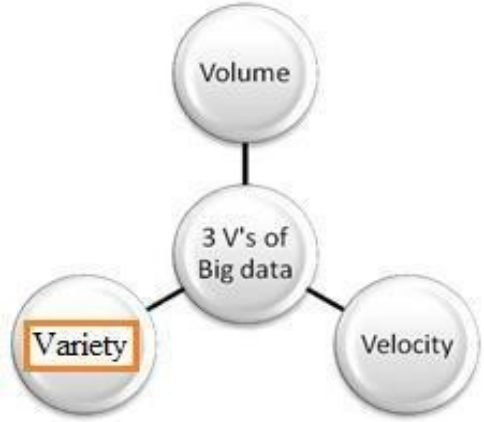

Figure 1 3 V's of Big Data [3]

Rest part of the paper is organized as follows: Section 2 discusses about the related work. Section 3 presents brief discussion of classification methods that are used in the suggested framework. Section 4 gives detail about the data, its features and experimental setup. Section 5 explains summary of the experiment outcomes, graphs, and performance measures. Section 6 presents conclusion and about future scope.

978-1-5386-6947-1/18/$31.00 ©2018 IEEE

II. RELATED WORK

Researchers are utilizing the results of machine learning predictions for solving problems of medical science [12, 13, 14, 17, 18]. Medical images have large volume of information which can be extracted and used for future prevention of dangerous diseases [19]. Many researchers have implemented machine learning algorithms using Python and R language for extracting information from the medical images [17].

Use of ensemble methods for optimizing the results of prediction accuracy is much in trend today. Ensemble classifiers focuses on hybridization for improving the results of machine learning prediction model [20].

Recently, deep learning is much active area of research in medical science. Greenspan et al. has reviewed the present and future perspectives of deep learning in medical science [21]. Prediction model using Convolutional Neural Networks (CNN) helps in providing much better experimental results for high dimensional image data [21].

High dimensional data consists of the medical images which has large number of feature descriptors. Feature extraction techniques are applied on good quality X ray images to extract the numerous feature descriptors. Deep learning neural networks are trained with the extracted data to build the prediction model [21]. Research is also carried for prediction of pneumonia using machine learning classifiers [22].

III. MATERIAL AND METHODS

The main purpose of exploring the field of machine learning is get a trained model for the classification and prediction of pneumonia patients, considering available X-ray data. The outcome of DCNN proposed framework helps to predict whether a person has pneumonia or not.

A. Proposed framework

The outcome of proposed DCNN framework helps to predict whether a person has pneumonia or not based on the X-ray image of chest. Normally, in real scenario problems, there is less control over the quality of images. Some regular pre-processing like removal of corrupted images, cleaning, etc. are always required [6]. Machine learning aim is to adopt efficient techniques to process large and complex data also considering cost. The abstract and detailed view of DCNN framework is displayed in the Figure 2 and Figure 3 respectively. Image is first preprocessed in required format for feeding to neural network and also checked for any corrupted image and removed it. Then, the converted data goes through the DCNN where various features are extracted at each level. At the end of the DCNN there is fully connected layer and then the last layer which is output layer which expect '1' or '0' , '1' for pneumonia and '0' for normal. With the help of back-propagation the network learns the right weights. After the model is trained it can be used for predict output on data which it has not seen earlier.

The data on which we want to predict should be in same format as is training data. We tested our model on basis of various metrics (Accuracy, AUC , error rate, sensitivity, etc.)
.

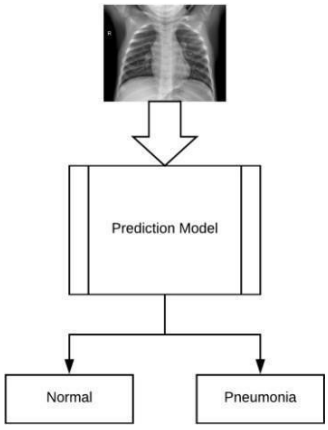

Figure 2 Proposed Prediction Framework

Consider all these performance metrics, models can be compared well because it is good to know what is TP rate as it shows in how many cases the is going to do the right classification.

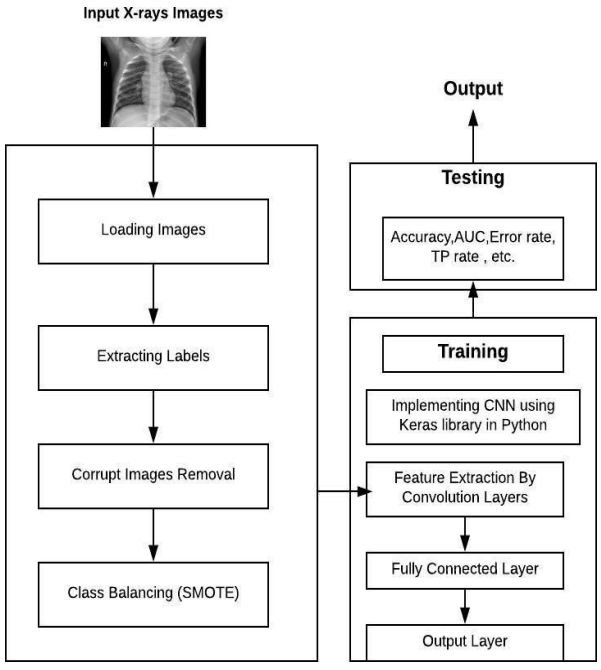

Figure 3 DCNN Framework

The person having pneumonia should be classified as positive but if a person not having pneumonia is classified as positive is not a big issue as it can be further rectified.

B. Machine Learning Classifiers

i. Neural Network: The idea is based on human brain working, like neuron communicates in human brain the same concept is applied here. There are different layer of neurons and they activate other neurons, like this is learns right weight for prediction [8].

ii. Random Forest : It ensembles results of different decision trees and take their average, by doing so improves accuracy and also avoid over-fitting [7].

iii. Support Vector Machine: Using vector, this method finds a hyperplane between the datasets. The hyperplane acts like a wall between the different classes. Checked the category of new unseen data (in which group it falls as all are separated by hyper-plane) accordingly and results are also shown here. The dimension depends on the number of features [9].

iv. Adaboost : It is an ensemble based method in which the output of one become input of next tree after some changes. Doing so improves the accuracy and over-fitting [10].

v. Logistic Regression : This is a classification method which learn some link in the dependant variable (label) and independent variables (features) by considering the probability [11,15].

vi. Decision Tree: It is a graph based machine learning classifier [16]

IV. EXPERIMENTAL INVESTIGATION

This section discusses about the dataset and experimental setup.

A. Dataset

Chest X-ray Images (pneumonia) for classification from the medical database [4]. The dataset consists of 5,863 X-Ray images with two labels i.e. Pneumonia or Normal.

B. Experimental Setting

Python's sklearn library is used to perform the various tasks like pre-processing of images and model building techniques. For implementation of convolutional neural network Keras library is used. The aim is to measure the classification accuracy of the classifiers after training them then test on new samples which where were not shown to the model before and checking the classification strength. To measure the performance of the suggested framework, seven parameters namely accuracy, MCC, F measure, error rate, TP Rate, FP rate, and also area under the curve (AUC) are used.

V. RESULTS AND DISCUSSION

This section discusses parameter evaluation metrics to measure the performance of various machine learning algorithms. The results are discussed much in detail and are also presented graphically.

A. Performance Evaluation

The results and performance of the suggested framework is evaluated with different parameters shown in confusion matrix Table 1. The various evaluation metrics calculated from the Table 1 are presented in Table 2. Based on various metrics DCNN performed better than other models as shown in Figure 4. As it was imbalanced data, we cannot totally depend on accuracy so comparing other metrics results DCNN gives good results. Neural Network and Random Forest also quite good and are strong competitors. Comparing TP rate and FP rate DCNN maintaining its stand. Overall DCNN is giving efficient result on unseen data.

Table 1 Confusion Matrix

	True Reference	
Predicted Condition	Condition Positive	Condition Negative
Pneumonia Positive	T P (A)	F P (C)
Pneumonia Negative	F N (D)	T N (B)

Table 2 Performance Metric Formula

Sensitivity	A/(A + B)
Specificity	B/(D + B)
Accuracy	(A + B)/(A + C + D + B)
F Score	(2 *A)/((2 *A) + (D + C))
MCC	(A * B)–(D * C)/SQRT((A + D)*(A + C) * (B + D) * (B + C))

B. Experimental Results

After experimentation and testing of different models, the results of various metrics are represented in the Table 3. It can easily be observed that accuracy and other parameters of the DCNN are the best among all other models. The results are also graphically depicted in Figure 4.

Random forest and Neural network are also showing good results but the performance of DCNN is better than the state-of-the art methods. False positive rate is low for DCNN and True positive rate is high which is good as they show that model is working good on unseen data. The patient having pneumonia is more likely be detected.

The chance of cases of patients having pneumonia but classified as normal is low which is shown by False positive rate. Many models have accuracy close to one another but when we consider other metrics then we can compare the models easily. TP rate is also an important metric to consider when comparing the models. To check the robustness of proposed DCNN model, K fold cross validation method is used.

With using 10 iterations, the stability of DCNN is presented graphically in the Figure 5 and Figure 6.

Table 3 Comparison of DCNN performance with different state-of-the art methods using machine learning performance metrics

Classifier	Accuracy (%)	Error rate (%)	TP Rate	FP Rate	F Score	MCC
Decision tree	77	23	0.83	0.24	0.62	0.50
Adaboost	78	22	0.92	0.25	0.60	0.53
Random forest	82	18	0.88	0.20	0.70	0.65
SVM	76	23	0.91	0.26	0.56	0.50
Logistic	77	23	0.90	0.26	0.57	0.50
DCNN	**84**	**16**	**0.92**	**0.11**	**0.77**	**0.66**
Neural Netwok	81	18	0.72	0.15	0.76	0.62
Naive Bayes	72	27	0.63	0.22	0.62	0.40

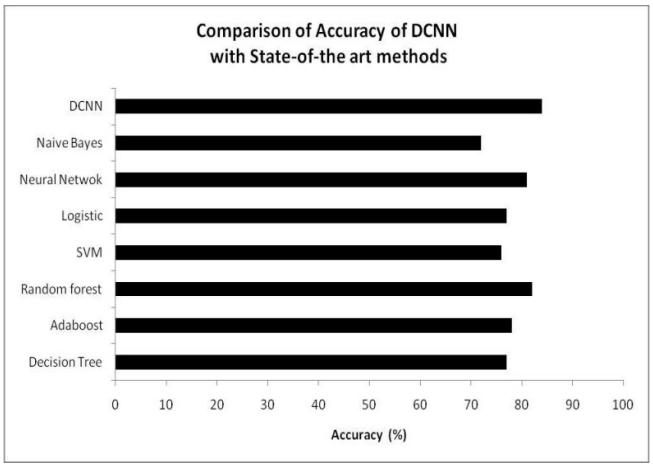

Figure 4 Comparison of Accuracy of DCNN with state-of-the-art methods

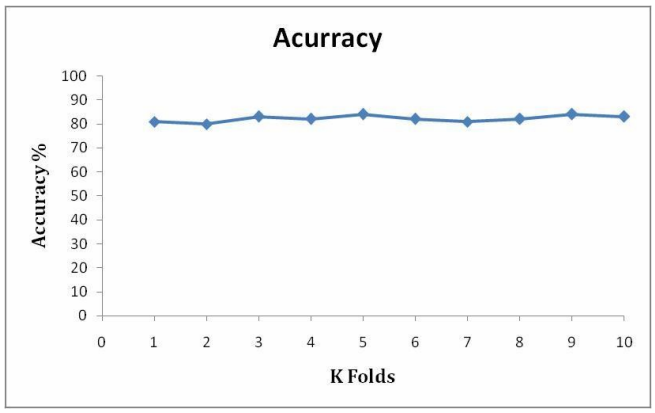

Figure 5 K fold cross validation of accuracy

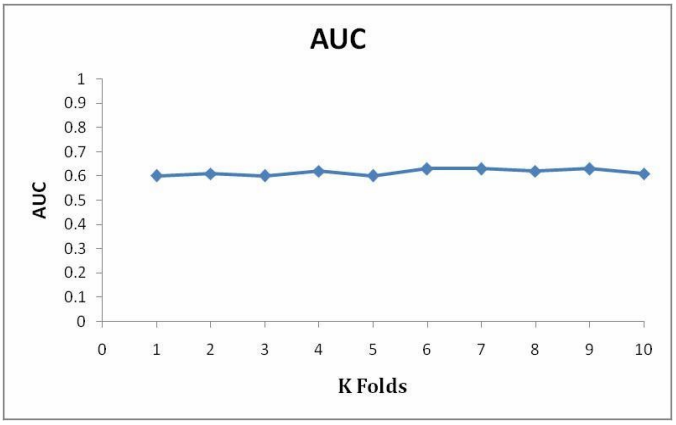

Figure 6 K fold cross validation of AUC

At last, K fold cross validation (with K=10) is performed to test the robustness of DCNN framework. The result for K fold validation for accuracy and AUC are depicted graphically in Figure 5 and Figure 6 respectively.

As it can be observed from the graphs, the values of accuracy and AUC are quite stable in all ten folds of cross validation. Hence, promising results are achieved for the prediction of pneumonia by the proposed framework.

VI. CONCLUSION AND FUTURE SCOPE

Pneumonia is life-threatening if it is not diagnosed properly in patients. Around two third of the global population lacks access to radiology diagnostics in India, according to an estimate by the World Health Organization. In this research, chest X ray image reports are utilized to train an efficient deep machine learning based prediction model for predicting pneumonia in patients. Deep learning makes this task more effective as deep learning is efficient in case of image data processing.

An efficient model is built using deep learning algorithms in Python language which will help doctors to detect this deadly disease. The proposed framework is compared with state-of-the art methods of machine learning and found to be more efficient in prediction with an average accuracy of 84%, which is found to be better than all other classifiers.

For future work, optimization of results will be done for improving the performance of prediction. Further, more volume of image data will be collected and data processing is done on the top of Hadoop framwork.

REFERENCES

[1] P. Rajpurkar et al. "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning."*arXiv preprint arXiv:1711.0522,2017.*

[2] WHO. *Pneumonia*, 2016 [Online] Available: http://www.who.int/news- room/fact-sheets/detail/pneumonia [Accessed: May 24, 2018]

[3] X. Chen. "Big data deep learning: challenges and perspectives."*IEEE access*", pp. 514-525, 2014.

[4] A. Gandomi et al. "Beyond the hype: Big data concepts, methods, and analytics."*International Journal of Information Management* vol 35(2), pp.137-144, 2014.

[5] Chest XRay data, 2018, [ONLINE] Available: http://dx.doi.org/10.17632/rscbjbr9sj.2#file-41d542e7-7f91-47f6-9ff2-dd8e5a5a7861 [Accessed: May 24, 2018]

[6] H. Nishtha et al. "B2FSE framework for high dimensional imbalanced data: A case study for drug toxicity prediction."*Neurocomputing* vol. 276, pp.31-41, 2018.

[7] Liaw, Andy, and Matthew Wiener. "Classification and regression by randomForest."*R news* vol. 2(3), pp. 18-22, 2002.

[8] A. Rowley et al. . "Neural network-based face detection."*IEEE Transactions on pattern analysis and machine intelligence* vol. 20(1), pp. 23- 38, 1998.

[9] M. Hearst, et al. "Support vector machines."*IEEE Intelligent Systems and their applications* vol. 13(4), pp. 18-28, 1998.

[10] R.. Takashi Onoda, and K-R. Müller. "Soft margins for AdaBoost."*Machine learning, vol.* 42(3), pp. 287-320, 2001.

[11] Hosmer Jr, David W., Stanley Lemeshow, and Rodney X. Sturdivant. *Applied logistic regression*. Vol. 398. John Wiley & Sons, 2013.

[12] P. Pedro et al. . Community-acquired pneumonia: identification and evaluation of non responders. Therapeutic advances in infectious disease, 1(1), pp. 5-17, 2013.

[13] M. Aydogdu et al. Mortality prediction in community-acquired pneumonia requiring mechanical ventilation; values of pneumonia and intensive care unit severity scores. Tuberk Toraks, vol. 58(1), pp. 25–34, 2010.

[14] D. Mollura et al. White paper report of the rad-aid conference on international radiology for developing countries: identifying challenges, opportunities, and strategies for imaging services in the developing world. *Journal of the American College of Radiology*, vol. 7(7), pp. 495–500, 2010.

[15] Press, S. James, and Sandra Wilson. "Choosing between logistic regression and discriminant analysis." *Journal of the American Statistical Association* 73.364 (1978): 699-705

[16] Safavian, S. Rasoul, and David Landgrebe. "A survey of decision tree classifier methodology." *IEEE transactions on systems, man, and cybernetics* 21.3 (1991): 660-674.

[17] Kononenko I. Machine learning for medical diagnosis: history, state of the art and perspective. Artificial Intelligence in medicine. 2001 Aug 1;23(1):89-109.

[18] Magoulas GD, Prentza A. Machine learning in medical applications. InAdvanced Course on Artificial Intelligence 1999 Jul 5 (pp. 300-307). Springer, Berlin, Heidelberg.pneumonia pattern using RNA-Seq and machine learning: challenges and solutions. BMC genomics. 2018 May;19(2):101.

[19] Wernick MN, Yang Y, Brankov JG, Yourganov G, Strother SC. Machine learning in medical imaging. IEEE signal processing magazine. 2010 Jul;27(4):25-38.

[20] Dietterich, Thomas G. "Ensemble methods in machine learning." *International workshop on multiple classifier systems*. Springer, Berlin, Heidelberg, 2000.

[21] Greenspan H, Van Ginneken B, Summers RM. Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique. IEEE Transactions on Medical Imaging. 2016 May;35(5):1153-9.

[22] Choi Y, Liu TT, Pankratz DG, Colby TV, Barth NM, Lynch DA, Walsh PS, Raghu G, Kennedy GC, Huang J. Identification of usual interstitial pneumonia pattern using RNA-Seq and machine learning: challenges and solutions. BMC genomics. 2018 May;19(2):101.