DOI: http://dx.doi.org/10.26483/ijarcs.v9i1.5136

ISSN No. 0976-5697

Volume 9, No. 1, January-February 2018

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

A SURVEY ON DEEP LEARNING TECHNIQUES FOR BIG DATA IN BIOMETRICS

Jaseena K.U.
Research Scholar, Division of Information Technology,
CUSAT, Cochin, Kerala, India &
Department of Computer Applications,
M.E.S College, Marampally, Kerala, India

Binsu C Kovoor
Division of Information Technology
Cochin University of Science & Technology (CUSAT)
Cochin, Kerala, India

Abstract: Big Data and deep learning are two important words in data science now days. The large volumes of data collected by organizations are utilized for various purposes such as for solving problems in marketing, technology, medical science, national intelligence, fraud detection etc. Traditional data processing systems are not adequate to handle, analyze and process as the collected data are unlabelled, uncategorized and very complex. Hence deep learning algorithms which are specialized in analysing such large volumes of unsupervised data can be utilized. The key characteristic that makes deep learning tools the most suitable ones for big data analytics is that they continuously improvise with each set of data they tackle. Deep learning is appropriate for exploiting large volumes of data and for analysing raw data from multiple sources and in different styles. This paper presents an overview of different deep learning techniques for big data in biometrics and discusses some issues and solutions.

Keywords: Big data; Biometrics; Deep learning; Deep Neural Networks; Recurrent Neural Networks

I. INTRODUCTION

Digital data is growing exponentially due to developments in web technologies, social media, and mobile and sensing devices. According to the National Security Agency, the Internet is processing 1,826 Petabytes of data per day [1]. Big data is a term used to identify large size datasets with greater complexity. Advanced analytics and visualization techniques are applied to large data sets in big data analytics to discover hidden patterns, and these hidden patterns and unknown correlations are used for effective decision making [2].

Biometrics systems are big data systems due to the large volume of data and analytics involved in building real-world biometrics. Such systems handle applications which make use of citizen identity cards, voter IDs, and social benefits, as well as secure online and mobile payments [3]. Biometrics is central to identity management and is going to be the hub of the next trend of cognitive systems.

Since data is of huge size and complexity, retrieving valuable knowledge from big data is not an easy task with conventional data processing techniques. Hence the development of advanced technologies is required. Today, machine learning techniques, together with advances in available computational power plays an important role in big data analytics and knowledge discovery [1]. In order to extract useful knowledge and make appropriate decisions from big data, machine learning techniques have been regarded as a powerful solution.

As an extremely active subfield of machine learning, deep learning utilizes supervised or unsupervised strategies to automatically learn hierarchical representations in deep architectures for classification [1]. The popularity of deep learning today can be because of increasing computing power (for example, the advent of graphics processor unit (GPU)) and increasing data size [4]. Deep learning utilizes both developments in computing power and special types of neural networks to learn complicated patterns in big data [5]. Deep learning techniques are currently most popularly used for identifying objects in images and words in sounds. The successful application of deep learning in pattern recognition motivated the researchers to use these techniques in more complex tasks such as automatic language translation, medical diagnosis, and other important problems.

In this paper, we introduce comprehensive survey of all the related work in deep learning for biometric big data. Section II describes big data in biometrics and section III presents literature review. The different issues and solutions to deep learning for big data and the strategies of different deep learning techniques are described in detail in Sections IV and V respectively. Conclusions are followed in Section VI.

II. BIG DATA IN BIOMETRICS

Biometrics is the science of identifying a person or verifying his identity on the basis of his physiological or behavioral characteristics. Fingerprint, iris, face, voice, hand geometry, and dynamic signature are the top six biometrics in the real world [3].

Biometrics systems can operate in two basic modes:

- 1:1 (called authentication or verification)
- 1: N (one-to-many, called matching or identification)

Biometric data in the internal affairs agencies, banks, and databases of criminal law are in a confused state. This causes certain difficulties in identifying the person. Since it ends before the identification process is implemented, biometric data restricts the search process in identifying suspects. By using biometric big data technologies such as authentication, bio cryptography and cloud based architecture, these problems can be solved effectively [6].

Biometric authentication systems are now common to manipulate databases of millions of individuals or for

governmental applications, hundreds of millions or even more. Biometrics systems need to deal with not only a large number of records, but also a variety of multimodal data types such as text, 1D signals, images, and video [3]. In biometrics systems, there are challenges regarding how to manage the ever changing database, rapid access to information, system security and integrity of records.

Both forms of trusted identity systems (1:1 and 1: N) are classic examples of big data systems [6]. Big data systems are associated with four Vs such as volume, velocity, variety and veracity [2, 7]. These dimensions are also associated with biometric data. There is an extremely large quantity of enrolment and verification data in modern biometrics systems and biometrics systems are designed to operate on a wide variety of data types [3].

Various software were developed using big data technologies for purposes such as criminal identification, face recognition, etc. Japanese developers tried to implement the software to recognize the persons who were identified as thieves and Lambda Labs developed face recognition software for Google Glass. The software Superbowl XXXV implemented by the police in Tampa, Florida State was used to recognize criminals using the scanned images of the face. Universities and airports in the United States also implemented software for person identification [6].

III. LITERATURE REVIEW

In this section, we review some papers related to deep learning techniques for biometric big data.

With the huge size of data available today, big data brings vast opportunities for various sectors. In [1], Xue-Wen et al. mainly target on two factors. The primary focus is to determine how deep learning can assist in solving specific problems in big data analytics. The second target is to find out how to improve specific areas of deep learning to reduce certain challenges associated with big data analytics. In this paper, the authors provide a brief overview of deep learning and also highlight current research efforts, the challenges in big data as well as the future trends.

In [2], the issues and challenges related to big data mining are presented. This paper also discusses big data analysis tools like Hadoop Map Reduce and HDFS. Some security and privacy challenges related to big data are also described.

In [3], Ratha et al. describe the "four V" (Volume, Variety, Velocity, and Veracity) challenges of biometric big data and the representative techniques addressing these challenges using different diverse and realistic biometrics applications. Indexing methods are used for dealing with velocity and volume of the biometric database. The authors in their paper describe two indexing methods, one for fingerprints and the other for irises. Risk and accuracy concerns related to biometrics identification for large databases are addressed by using multiple (variety) biometrics. The integrity (veracity) of biometric databases can be ensured by eliminating multiple duplicate records as well as protecting against theft of biometric identifiers.

A systematic overview of emerging research work of deep learning on machine health monitoring is presented in [4]. Four categories of Deep Learning (DL) architectures such as Auto encoder models, Restricted Boltzmann Machines models, Convolutional Neural Networks and Recurrent Neural Networks are proposed. Some research trends and potential future research directions of DL-based machine health monitoring methods are also provided.

Najafabadi et al. [5] have the opinion that the important issues in big data analytics include extracting complex patterns from massive volumes of data, fast information retrieval, semantic indexing, data tagging, and discriminative tasks such as classification and prediction. These above mentioned issues can be solved effectively using deep learning. Deep learning techniques can be used to extract the complex and nonlinear patterns generally observed in big data easily. Deep learning techniques are more beneficial when dealing with learning from massive volumes of input data that is generally unsupervised and uncategorized [5].

According to the author Shui Yu [8], the biggest concern of big data is privacy. In this paper, the author extensively surveys the existing research outputs and achievements of the privacy field in both application and theoretical angles. The author first presents an overview by defining the roles and operations of privacy systems. Then the milestones of the current two major research categories of privacy such as data clustering and privacy frameworks are reviewed. After that, the author discusses the effort of privacy study from the perspectives of different disciplines [8].

David et al. [9] investigated spoofing detection systems for different biometric modalities such as iris, face, and fingerprint based on two deep learning approaches. The first approach makes use of learning suitable convolutional network architectures for each domain and the second approach focuses on learning the weights of the network via back-propagation. The authors consider nine biometric spoofing benchmarks with each one having real and fake samples of a given biometric modality and attack type. The authors then try to learn deep representations for each benchmark by combining and contrasting the two learning approaches. Experiments showed that these approaches gained outstanding classification results for all problems and modalities in eight out of nine benchmarks. The results describe that spoofing detection systems based on convolutional networks are robust to known attacks and can be adapted to image based attacks [9].

Le Cun et al. [10] presents a review of deep learning techniques. Deep learning methods have been successfully implemented in speech recognition, visual object recognition, object detection and many other domains such as drug discovery and genomics. Convolutional neural networks and recurrent neural networks have their own applications. According to the authors convolutional neural networks are more suitable for processing images, video, speech and audio and recurrent neural networks work well with sequential data such as text [10].

A comprehensive study on types of big data in bioinformatics and various deep learning techniques to analyze such big data are presented in [11].The limitations and future research directions in big data in informatics are also provided. The different big data architectures such as Hadoop MapReduce to deal with big data are also discussed.

Various big data applications, opportunities and challenges as well as various big data technologies are demonstrated in [12]. The role of deep learning in big data and the relationship between big data and deep learning are also discussed. In [13], an analysis of the challenges connected with machine learning in big data is defined. An overview of machine learning techniques to deal with the various challenges are also presented.

In [14], a comprehensive review of various deep learning techniques suitable in health Informatics is presented by providing an analysis of the relative merit, and demerits of the technique. Key applications of deep learning in the fields of bioinformatics, medical imaging, medical informatics, and public health are discussed in this paper. How various issues concerned with health informatics can be solved using deep learning are also mentioned.

In [15], machine learning algorithms for effective prediction of chronic disease outbreak in disease frequent communities are proposed. A new convolutional neural network (CNN) based multimodal disease risk prediction algorithm are presented using structured and unstructured data from the hospital. A detailed survey of various deep Learning methods, comparison of frameworks, and algorithms are presented [16]. Then application of deep learning in big data such as speech recognition, computer vision, and Natural language processing, its challenges, open research problems and future trends are presented.

Universal architectures and training method to handle big data using deep neural networks are presented [17]. How big data is related to deep learning is explained in this paper. An efficient and fast visualization technique to represent big data in various dimensions and finally a fast clustering algorithm to find clusters with complex shape are also described.

IV. ISSUES AND CHALLENGES

The main four V's characterized by big data systems are namely volume, variety, velocity, and veracity. These dimensions are also associated with biometric data as explained below. In this section the critical issues related to biometric big data from different perspectives such as volume, variety, velocity and veracity are presented. Also possible solutions to overcome these problems are mentioned briefly.

A. Volume

There is an extremely large quantity of enrolment and verification data in modern biometrics systems. For example, the FBI (Federal Bureau of Investigation) Next Generation Identification has more than 100 million individuals and India's Unique Identity Authority of India has more than 600 million individuals [3]. Large volumes of data are always an issue for both authentication and identification. But they cause a greater challenge for the 1:N matching required by identification than for the 1:1 matching in authentication.

To overcome this issue, distributed frameworks with parallel computing are preferred [18]. Large scale data sets can be dealt with realistic parallel programming methods such as MapReduce. The fault tolerance capability of MapReduce makes it an important tool for tackling the large data sets [2]. The growth of high dimensional data creates a need for dimensionality reduction techniques to transform the high dimensional data into a smaller set. Dimensionality reduction techniques aim at finding and exploiting low dimensional structures in high dimensional data and thus save the computation and storage burden. The most popular and commonly used dimensionality reduction algorithms are Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA).

B. Variety

Biometrics systems are designed to operate on a wide variety of data types, such as 2D grayscale images (fingerprints), colored and nonvisual images (faces and irises), 3D data (faces), video data (faces and gaits), and 1D temporal signals (voices and signatures) [3]. Since the data coming from various sources are of different types, it is always tedious and complex job to learning with such heterogeneous data set.

To handle the problem of heterogeneous data integration, representation learning is preferred. In representation learning, data representation from each data source is learnt first. The learnt features at different levels are then integrated [1, 18]. Data from different heterogeneous sources can also be integrated very effectively using deep learning methods.

C. Velocity

This term refers to the speed at which data is processed. When large systems, such as online commerce and banks, use biometrics for authentication for all of their daily transactions, velocity becomes an important issue. This challenge is further worsened when considering systems with a large number of concurrent users or when applying big data in motion analytics to provide timely alerts based on biometrics [3].

In many real world applications, a task has to be finished within a certain period of time. Otherwise the processing results become less valuable or even worthless. Deep learning can be applied to reduce the time needed to build a useful model.

D. Veracity

The veracity aspect largely worries on how to manage the basic error rates in a biometrics system, particularly the trade-off between false positives and false negatives [3]. The truthfulness can be maintained by using trusted enrolment, trusted verification, and identity credential management. It is important that the biometrics signal comes from a live person is a key requirement in every biometric transaction [18].

The precision and trustworthiness of the source data become an issue nowadays because data come from multiple heterogeneous sources and hence data quality is not all verifiable. Uncertainty and incompleteness are also associated with data quality. With incomplete data, correct interpretations and predictions of data are not possible and hence to tackle these challenges, advanced deep learning methods can be applied.

From this discussion we may conclude that most of the challenges associated with biometric big data can be solved using deep learning techniques.

V. DEEP LEARNING STRATEGIES

The term machine learning refers to the automated detection of meaningful patterns in data. In machine learning, data plays an important role and the learning algorithm is used to discover and learn knowledge or properties from the data. The quality as well as the quantity of the dataset will affect the learning and prediction performance [19].

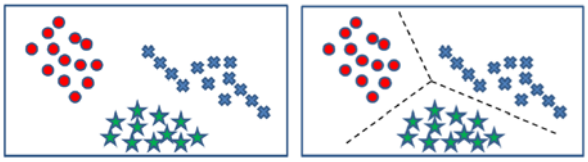
Various types of machine learning are explained below:

a) Supervised learning,
b) Unsupervised learning,
c) Reinforcement learning
d) Semisupervised learning
e) Deep learning

A. *Supervised Learning*

The training set given for supervised learning is the labeled dataset. Supervised learning finds the association between the feature set and the label set and it predicts the class label of test instances using training datasets. The knowledge extracted from supervised learning is often utilized for prediction and recognition [19].

Fig.1 shows supervised learning. Fig (a) presents a three class labelled dataset, where the color shows the corresponding label of each sample. After supervised learning, the class separating boundary could be found as the dotted lines in (b).



(a) (b)

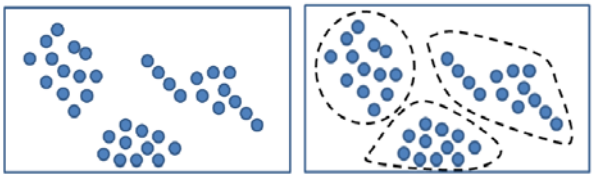
Figure 1. Supervised learning

There are mainly two classes of supervised learning problems and they are regression and classification. Different algorithms are available for solving regression and classification problems. Linear regression and support vector machines are supervised learning algorithms used for solving regression problems and classification problems respectively. But for big data analytics, efficient and advanced supervised methods suitable for parallel and distributed learning are required. Divide and conquer SVM, distributed decision trees and neural networks are supervised learning algorithms suitable for big data in which SVM is the most efficient and commonly used method [11].

B. *Unsupervised Learning*

The training set given for unsupervised leaning is the unlabeled dataset. Unsupervised learning aims at clustering, probability density estimation, finding the association among features, and dimensionality reduction. An unsupervised algorithm may be used to simultaneously learn more than one properties and the results from unsupervised learning can be further used for supervised learning [19].

Fig.2 shows unsupervised learning. Fig (a) presents unlabeled dataset. Unsupervised learning separates the datasets into different clusters as in (b).



(a) (b)

Figure 2. Unsupervised learning

There are two classes of unsupervised learning problems and they are clustering and association problems. K-means and Apriori algorithm are some examples of unsupervised learning algorithms. K-means are commonly used for solving clustering problems and Apriori algorithms are suitable for association rule learning problems.

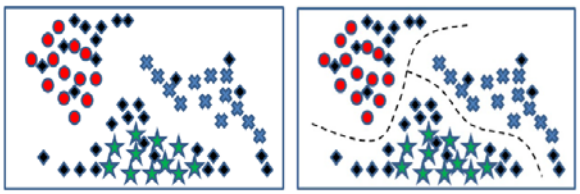
C. *Reinforcement Learning*

Reinforcement learning is a computational method to learning which tries to learn from feedback and the feedback is received by communicating with an external environment. Reinforcement learning is used to solve problems of decision making. Q-learning and Sarsa are some of popular reinforcement learning algorithms [20].

D. *Semisupervised Learning*

There is another machine learning category called semi supervised. It is defined by supervised and unsupervised learning, contains both labeled and unlabeled data, and jointly learns knowledge from them [19].

In Fig.3, labeled dataset is marked with red, green, and blue and unlabeled dataset is marked with black. The distribution of the unlabeled dataset could guide the position of separating boundary.



(a) (b)

Figure 3. Semi-supervised learning

E. *Deep Learning*

Deep Learning is a more sophisticated machine learning artificial neural network approach for getting intelligence from big data. Instead of using shallow structured learning architectures, deep learning mainly makes use of supervised or unsupervised learning strategies. These learning strategies in deep architectures automatically learn hierarchical representations from big data. Deep learning mainly focuses on two characteristics of big data say volume and variety. So deep learning is more suitable for processing high volumes of unstructured and heterogeneous data. Big data can be more accurately predicted using deep learning. Deep neural network models can be trained efficiently using scalable parallel algorithms [11].

The unstructured part of big data contains valuable information and learning patterns. Deep Learning and other technologies are trying to understand the 80 percent of unstructured data contained in EMRs and make that information actionable [21]. The most commonly used deep structures are described below.

Fig.4 shows the layout of a deep learning architecture and the architecture consists of input layers, hidden layers and output layers. Input layer receives input data and in turn they are divided in to multiple samples for data abstractions. The extraction of features from multiple levels and prediction are done by the intermediate layers. The outputs from the intermediate layers are forwarded to the output layer for final prediction [11].

Deep Neural Network (DNN) is a multilayer perceptron with many hidden layers, whose weights are fully connected and are often initialized using either an unsupervised or a supervised pre training technique [19]. Feed forward neural networks and Recurrent Neural Networks are variants of Deep Neural Networks.

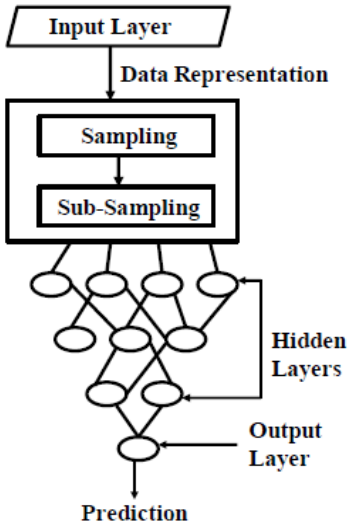


Figure 4. Layout of the deep learning architecture

1) Feed forward neural networks :

In this, the connections between the units do not form a cycle. Convolutional Neural Network (CNN) and Deep Belief Network (DBN) are different types of feed forward neural networks.

Convolutional Neural Network (CNN) is formed by a stack of distinct layers that transform the input volume into an output volume through a differentiable function in order to find the set of locally connected neurons. They are made up of neurons that have learnable weights and biases. Each neuron receives some inputs, performs a dot product and optionally follows it with a non-linearity. Generally three types of layers are used to build CNN architecture and they include Convolutional Layer, Pooling Layer, and Fully Connected Layer. Each Layer accepts an input 3D volume and transforms it to an output 3D volume through a differentiable function. Each Layer may or may not have parameters [21].

The limitations of traditional neural networks, such as its poor performance due to local optima and inefficiency to use unlabelled data, can be rectified by using a deep learning architecture called Deep Belief Network (DBN) [11]. DBN utilizes both labelled and unlabelled data for learning.

DBN comprises of multiple layers of a graphical model having both directed and undirected edges. It consists of multiple layers of hidden units, where each layer are connected with each other. Inputs are given to the bottom layer and the higher layers represent increasingly abstract features of the data. DBNs can be trained in a layer wise fashion.

2) Recurrent neural networks (RNN)

Recurrent Neural Networks can be considered as another class of deep networks for unsupervised as well as supervised learning, where the depth can be as large as the length of the input data sequence. In the unsupervised learning mode, the RNN is used to predict the data sequence in the future using the previous data samples, and no additional class information is used for learning. RNN is called Recurrent since they receive inputs, update the hidden states depend on the previous computations and make predictions for every element of a sequence [19].

RNNs are neural networks with memory as they keep information of what has been processed so far and are powerful dynamic systems for sequential tasks. They can maintain a state vector that implicitly contains information about the history of all the past elements of the sequence. The RNN is very powerful for modeling sequence data such as speech or text.

A neural network having one or more hidden layers with at least one feedback loop is known as a recurrent network. The feedback may be a self feedback, i.e., where the output of the neuron is fed back to its own input.

VI. CONCLUSION

Deep learning techniques can be used to extract the complex and nonlinear patterns generally observed in big data easily. Deep learning combines advances in computing power and special types of neural networks to learn complicated patterns in large amounts of data. In this paper, a comprehensive survey on various deep learning approaches to handling big data in biometrics is presented. The detailed survey presented will help the beginners who are doing research in the field of deep learning to get more useful knowledge.

VII. REFERENCES

[1] Xue-Wen Chen and Xiaotong Lin, "Big Data Deep Learning: Challenges and Perspectives", IEEE, Volume 2, May 2014.

[2] Jaseena K U and Julie M David, " Issues, Challenges, and Solutions : Big Data Mining", NeTCoM, CSIT, 2014, pp. 131–140

[3] N. K. Ratha, J. H. Connell, and S. Pankanti, "Big Data approach to biometric-based identity analytics", IBM J. RES. & DEV. VOL. 59 NO. 2/3 PAPER 4 MARCH/MAY 2015.

[4] Rui Zhao, Ruqiang Yan, Zhenghua Chen, Kezhi Mao, Peng Wang, and Robert X. Gao, "Deep Learning and Its Applications to Machine Health Monitoring: A Survey", Journal of Latex Class Files, Vol. 14, No. 8, August 2015.

[5] Maryam M Najafabadi, Flavio Villanustre, Taghi M Khoshgoftaar, Naeem Seliya, Randall Wald, and Edin Muharemagic, "Deep learning applications and challenges in big data analytics", Journal of Big Data, USA, 2015, vol 2, №1, pp. 1-21.

[6] Shafagat Mahmudova, "Big Data Challenges in Biometric Technology", I.J. Education and Management Engineering, 2016, 5, 15-23 Published Online September 2016 in MECS http://www.mecs-press.net).

[7] Barbara Hammer, Haibo He, and Thomas Martinetz, "Learning and modelling Big data", ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.

[8] Shui Yu, "Big Privacy: Challenges and Opportunities of Privacy Study in the Age of Big Data", IEEE, Volume 4, June 2016.

[9] David Menotti, Giovani Chiachia, Allan Pinto, William Robson Schwartz, Helio Pedrini, Alexandre Xavier Falc˜ao, and Anderson Rocha, "Deep Representations for Iris, Face, and Fingerprint Spoofing Detection", IEEE 2015.

[10] Yann LeCun, Yoshua Bengio and Geoffrey Hinton, "Deep learning", Nature, Vol 521, May 2015.

[11] Hirak Kashyap, Hasin Afzal Ahmed, Nazrul Hoque, Swarup Roy, and Dhruba Kumar Bhattacharyya, "Big Data Analytics in Bioinformatics: A Machine Learning Perspective", Journal of Latex class files, Vol. 13, No.9, 2014

[12] C. L. Philip Chen, "Big Data Challenges, Techniques, Technologies, and Applications and How Deep Learning can be Used", proceedings of IEEE 20th international conference on Computer supported cooperative work in design, 2016.

[13] Alexandra L Heureux, Katarina Grolinger, Hany F Elyamany and Miriam M Capretz, "Machine Learning With Big Data: Challenges and Approaches", Volume 5, IEEE 2017.

[14] Daniele Ravı, Charence Wong, Fani Deligianni, Melissa Berthelot, Javier Andreu-Perez, Benny Lo, and Guang-Zhong Yang. "Deep Learning for Health Informatics", IEEE Journal of Biomedical and Health Informatics, Vol. 21, No. 1, January 2017

[15] Min Chen, Yixue Hao, Kai Hwang, Lu Wang and Lin, "Disease Prediction by Machine Learning Over Big Data From Healthcare Communities", Vol 5, IEEE 2017

[16] Mehdi Gheisari, Guojun Wang, Md Zakirul Alam Bhuiyan, "A Survey on Deep Learning in Big Data", IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC), 2017.

[17] B M Wilamowiski, Bo Wu, and Janusz Korniak, "Big data and Deep Learning", , 20th Jubilee IEEE International Conference on Intelligent Engineering Systems, June 30-July 2, 2016

[18] Junfei Qiu, Qihui Wu, Guoru Ding, Yuhua Xu and Shuo Feng, "A survey of machine learning for big data processing", EURASIP Journal on Advances in Signal Processing (2016) 2016:67.

[19] Wei-Lun Chao, "Machine Learning Tutorial", DISP Lab, Graduate Institute of Communication Engineering, National Taiwan University, 2011.

[20] David Poole and Alan Mackworth, "Artificial Intelligence Foundations of Computational Agents", Cambridge University Press, 2010.

[21] Sri Satish Ambati, "Deep learning: A brief guide for practical problem solvers", New Tech Forum, 2015.