Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 12, Issue 10, October 2021: 1424-1434

A REVIEW ON-MACHINE LEARNING BASED MODEL FOR SECURE DATA ANALYSIS

Vivek Kothuru
vivek.kothuru@gmail.com
Department of Computer science Engineering
NIIT UNIVERSITY, NH-8, Rajasthan, India – 301705
Prashanth Kumar Manji
prashanth.manji@gmail.com
Masters in Computer Science / Department of Computer Science
The George Washington University, 2121 I St NW, Washington, DC 20052
Greeshmanjali Bandlamudi
greeshmanjali.bandlamudi@gmail.com
Masters in Data Science / Department of Data Science
The George Washington University, 2121 I St NW, Washington, DC 20052
Likitha Chimirela
likitha123.chimirela@gmail.com
Department of Computer science Engineering
ICFAI University, Donthanapally, Hyderabad, Telangana, India - 501203

Abstract--In data science, there is a lot of interest in **Data Analytics** and **Machine Learning**. Large volumes of domain-specific data are being collected by government agencies and private organisations alike, and this data may provide useful information on national security, cyber security, fraud detection, and marketing strategies. Google and Microsoft, for example, examine vast amounts of data for business assessments and choices that have an effect on current and future technological developments, microsoft also does this. Through a hierarchical learning process, machine learning systems extract complicated high-level abstractions as data representations. At each step of the hierarchy, complicated abstractions are learnt through building on the simpler abstraction defined in the previous step. Machine Learning is used in **Data Analytics** because it is capable of analyzing and learning from vast volumes of unlabeled and uncategorized raw data. As a result, it's a useful tool for data analytics.

Our study examines how Machine Learning may be used to tackle some of the most difficult Data Analytics problems, such as extracting complex patterns from large volumes of data, Ontology Indexes, Data Tagging, and Fast Information Retrieval. Streaming data, high-dimensional data scalability, and distributed computing are all topics of Machine Learning research that need to be explored further in order to address particular Data Analytics challenges. We also delve into these subjects in depth. Finally, we create data sampling criteria and domain adaption modelling, as well as requirements for the generation of appropriate data abstractions, to provide new insights into critical future endeavors.

Index Terms--Machine Learning, Semantic Indexing, Data Tagging, Fast Information Retrieval, and Simplifying Discriminative Tasks.

I. INTRODUCTION

One of the primary goals of ***Machine Learning*** algorithms is to generalize learned patterns so that they may be used on previously unknown data. For a basic machine learner, good data representation may lead to excellent performance; for a more complex machine learner, bad data representation is likely to lead to worse performance. A key aspect of machine learning is feature engineering, which concentrates on the creation of features and data models based on unstructured data. For a machine learning operation, feature engineering takes a large amount of time, it is domain-specific, and requires a lot of human input. These approaches were developed specifically for computer vision, and are known as HOG [2] and SIFT [3.] Popular feature engineering techniques in computer vision include these: Conducting hand - crafted features in a more automatic and thorough approach would be a major breakthrough in machine learning since practitioners would no longer need to rely on direct human input."

Machine learning approaches might be used to automate the extraction of high-level Abstraction features, according to researchers. Higher-level features are represented by these algorithms in terms of reduced (less abstract) qualities, culminating in a layered, unified framework for learning and interpreting data. Hierarchical learning architecture has been driven by artificial intelligence that mimics Machine Learning Algorithms' basic sensory regions in the human brain (the neo-cortex) and pulls characteristics and abstractions from that input automatically (see [4]-[6]).

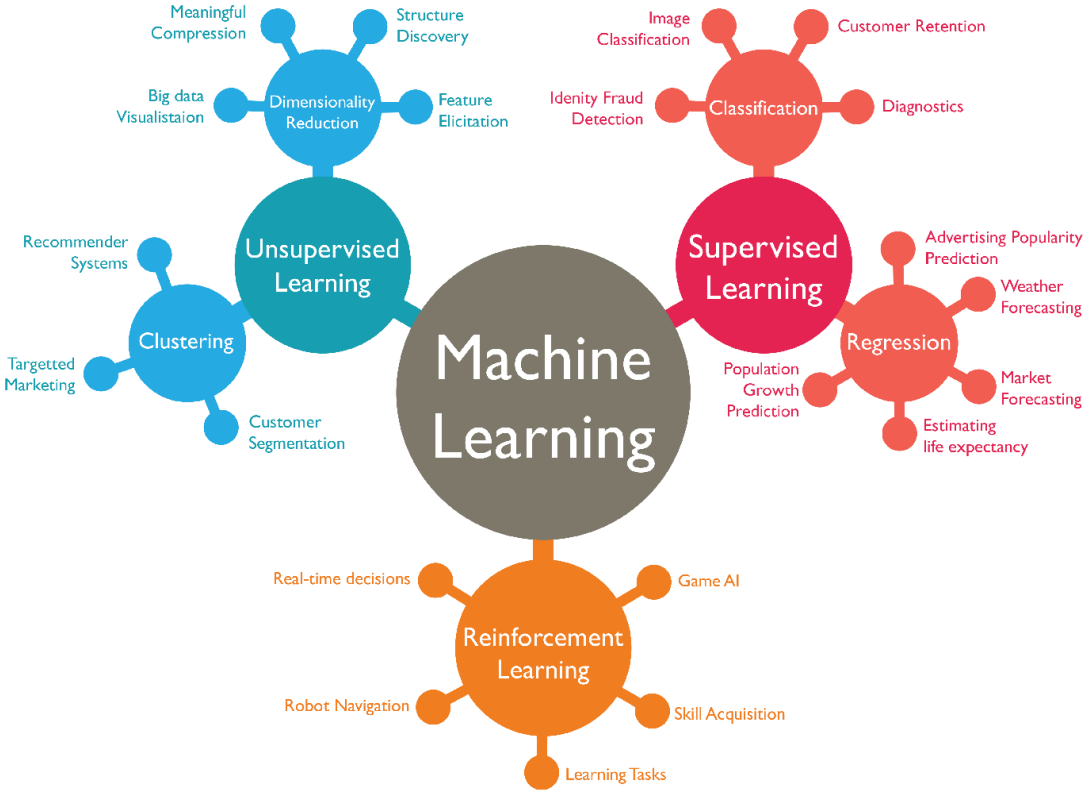

Figure 1: Machine Learning Algorithms.

When dealing with huge volumes of unsupervised data, a greedy layer-wise technique for training data representations is beneficial [7],[8]. For instance, generative probabilistic models that stack non-linear feature extractors produce better samples and have stronger data representation invariance (as in Machine Learning). Speech recognition [12-16], computer vision [7,8,17], and ***Natural Language Processing*** [18-20] are just a few of the numerous applications for machine learning approaches.

Using Machine Learning, data analysis activities like categorization and prediction may be performed using linear models that are more easier and faster to build. Researchers in both fields will benefit from this work since it is the first to examine how Machine Learning algorithms may be utilized to address crucial Data Analytics problems.

The study is divided into two parts:

(1) How Machine Learning can assist with certain Data Analytics difficulties, and

(2) How certain aspects of Machine Learning might be enhanced to reflect specific Data Analytics concerns.

In this paper, we look at how Machine Learning may be used for specialized Data Analytics, such as learning from large amounts of data, semantic indexing, discriminative tasks, and data tagging. Second, we are looking at the unique obstacles that Machine Learning confronts in light of current Data Analytics concerns, such as learning from streaming data, dealing with high-dimensional data, and scaling models across distributed and parallel contexts.. Last but not least, we identify key future areas in Machine Learning for Data Analytics that require innovation, such as data sampling for the generation of useful high-level abstractions, domain (data distribution) adaptation, defining criteria for the extraction of good data representations for discriminative and indexing tasks and semi-supervised learning and active learning.

The following is a description of the paper's layout. Information about applying AI for data mining and data analysis may be found in this paper. In Section "***Data Analytics***," which covers essential aspects of data and identifies specific data analysis challenges faced in Data Analytics, a quick review of Data Analytics is presented. The section "***Applications of Machine Learning in Data Analytics***" investigates how Machine Learning may be used to Data Analytics problems and gives a thorough appraisal of works investigating Machine Learning-based data analysis solutions. "***Machine learning problems in data analytics***" highlights the challenges Machine Learning experts face as a result of Data's particular data problems that require. It expresses our thoughts on what further needs to be done to widen the application of Machine Learning in Data Analytics and presents important problems for domain specialists to consider. Section "Future Machine Learning Work" The "***Conclusion***" section summarizes the work that has been done and reiterates the paper's main conclusion.

II. REVIEW OF LITERATURE

“As previously stated, Algorithms that use machine learning to extract meaningful abstract representations from raw data employ hierarchical multi-level learning, in which more abstract and complex representations are learned at a higher level of the learning hierarchy from less abstract

concepts and representations at a lower level. Despite the fact that tagged data may be used by machine learning [4],[5] if there is a sufficient quantity of it, it is better suited to learning from large volumes of unlabeled/unsupervised data, making it ideal for extracting meaningful patterns and representations from data. [4],[5]].

After Machine Learning has learned hierarchical data abstractions from unsupervised data, more traditional discriminative models may be trained with less labeled/supervised/labeled data." Finding global and non-local patterns in data requires the use of machine learning algorithms outperform shallow learning frameworks. [4].

Data Analytics provides a significant potential for the development of unique algorithms and models to handle specific data concerns. For data analytics specialists and practitioners, Machine Learning ideas give one such solution venue.

2.1 Semantic Indexing

It makes data easier to access and comprehend by helping search engines to operate more quickly and effectively, for example. The use of machine learning to create high-level abstract data representations for semantic indexing may be preferable to using raw input for data tagging.

Hinton et al. [1] to learn document binary codes, develop a Machine Learning generative model Machine Learning represents the document's word-count vector as high-dimensional data at the lowest layer, while binary code is represented at the highest level of the network. The authors show that the binary codes of semantically linked texts may be located in the Hamming space using 128-bit codes. Information in documents may be obtained via binary coding that is included in them. The Hamming distance between any two query documents is calculated. Once the top D comparable have been retrieved, the results are shown. To calculate the Hamming distance between two binary codes, methods like fast-bit counting are used, which reduces the amount of storage needed for binary codes. As a result, searches now complete significantly more quickly. Binary code retrieval is more accurate and quicker than semantic analysis, according to the authors.

Shorter binary codes can be generated using Machine Learning generative models by restricting the quantity of variables used at the top of the learning hierarchy. After that, the smaller binary codes may be utilised as memory addresses in larger programmes and systems. As a result of using semantic hashing, a tiny Hamming-ball surrounding the memory location will include documents with semantically comparable contents. Information may be gathered from an enormous number of documents in a short time using this method, no matter how big the collection is. If the query document's memory address differs by only a few bits, then finding all memory addresses that do the same will provide documents that are similar. Because it detects all memory locations that vary by a few bits from a document's address, semantic hashing is a common approach for obtaining data.

Ranzato et al. [2] in which machine learning parameters are learnt from supervised as well as unstructured data, share your findings. With this strategy, a huge collection of data does not have to be labelled entirely (since some unlabeled data is anticipated) and the model already has some previous knowledge (from the supervised data) to capture important class/label information within the data. To put it another way, the model must discover data representations that produce correct input reconstructions and document class label predictions that are both accurate.

Machine Learning models outperform shallow learning methods for learning compact representations, the authors show. Compact representations are efficient because they perform fewer calculations and need less store space when indexing.

2.2 Natural Language Processing (NLP)

Using Google's "word2vec" tool, semantic representations may also be automatically extracted from large datasets. A big text corpus is used to construct the word vectors. Once the vocabulary is constructed from the training text input, word vector files may be utilized as an element in many NLP and machine learning applications.

It is possible to generate high-quality word vectors from datasets comprising hundreds of millions of diverse vocabulary items, according to ***Miklov et al. [3].*** (Including roughly 1.6 billion). Artificial neural networks, which they learn to employ, are used to represent words (ANNs). To train the network on such a huge quantity of data, the large-scale distributed framework DistBelief is employed. On the basis of enormous volumes of data, researchers discovered strong semantic links between words in word vectors. This involves ties between cities and the countries to which they belong. Semantic word vectors offer the potential to improve a wide range of natural language processing (NLP) applications now in use.

Using word2vec for natural language translation is demonstrated in a related study by ***Miklov et al. [4],*** for example. Machine learning techniques that generate nonlinear representations of word occurrences may capture texts' high-level semantic components (which could not normally be learned with linear models). Large volumes of data must be collected for the input corpus in order to capture these complex representations, and categorizing this enormous quantity of data is a tough operation. Using unlabeled documents (unsupervised data), Machine Learning may get access to additional input data while using less supervised data to enhance data representations and make them more suitable for particular learning and inference tasks. Data representations derived from document retrieval have proven to be particularly valuable for search engines, making them tremendously profitable. As with textual data, machine learning may be used to derive semantic representations from the input corpus, allowing for semantic indexing. Because of the recent emergence of Machine Learning, additional research is required to harness its hierarchical learning process for semantic indexing of data. It is a contentious problem how to define similar when trying to develop indexing data representations (recall, data points that are semantically similar will have similar data representations in a specific distance space).

One benefit is that using Machine Learning to extract features gives the data analysis a nonlinear twist, which ties discriminative tasks to Artificial Intelligence. Another advantage is that Data Analytics is more computationally efficient because of the use of basic linear analytical models on the retrieved characteristics.

2.3 Audio Video Indexing System

Audio and video files may be searched using speech using the MAVIS system from Microsoft Research. MAVIS is a machine learning-based voice recognition system based on artificial neural networks. MAVIS automatically creates closed captions and keywords from digital audio and video signals, making audio and video files containing speech content more accessible and discoverable.

To better recognise picture objects, ***Hinton et al. [6]*** employed Machine Learning and Convolutional Neural Networks, and their technique performed better than earlier ones. Hinton's team used the ImageNet dataset, one of the largest in image object recognition, to illustrate the benefits of Machine Learning in boosting picture searches.

In order to train an artificial neural network using ImageNet, ***Dean et al. [7]*** employed a similar Machine Learning modelling strategy in combination with a large-scale software infrastructure and achieved even greater results. By employing their technique to analyse natural language words, ***Socher et al. [8]*** demonstrate that it is a natural tool for predicting tree architectures. This highlights Machine Learning's use in deriving data representations from a wide range of data sources.

The volume of digital picture collections has exploded in recent years due to the fast growth of the Internet and the number of online users. Social networks, global positioning satellites, picture-sharing systems, medical imaging systems, military monitoring, and security systems are just a few of the places where this information comes from today.” Google has built picture search algorithms based only on image file names and document contents, with little regard for the image's actual content (such as Google Images search engine). Practitioners should go beyond the basic relationship between an image and its text to acquire artificial intelligence when looking for photos because these representations are not always accessible in large image archives To better understand, search, and retrieve these large picture data sets, experts should organise and gather them better. When you work with large amounts of picture data, you may automate the tagging and semantic information extraction processes. It is now feasible to generate sophisticated picture and video data representations at relatively high abstraction levels using Machine Learning, which may then be used for image annotation and tagging for indexing and retrieval. Machine Learning may prove effective in Data Analytics for semantic tagging and data discrimination.

Recurrent neural networks may be used to build a relevant search area for Machine Learning-based specifically created search, according to ***Kumar et al. [9].***

Using machine learning and an independent variation evaluation that identifies invariant spatio-temporal properties from video data, action sequences may be recognized and video data annotated. When employing Machine Learning techniques like stacking as well as convolution to create hierarchical representations, their solution outperforms previous methods." SIFT and HOG were previously available, which have been video adaptations of hand-drawn picture characteristics.

A recent study by ***Le et al. [10]*** illustrates that taking features from video data and applying them to other domains is an interesting research avenue. ***Zhou et al. [11]*** the incremental feature learning technique quickly converges in a large-scale online scenario to the optimum amount of features. Incremental feature extraction is beneficial when the distribution of data in online data streams changes over time. Adapting to new internet-based large-scale data flows may require the usage of RBM and other Machine Learning methods, such as incremental feature learning and extraction. For big datasets, it also avoids the requirement for time-consuming cross-validation analysis for choosing the number of features to use.

Calandra et al. [12] the use of adaptive machine belief networks demonstrates the potential of Machine Learning in the online context to learn from both stagnant and flowing data. To replicate samples from the original data, they leverage the generative feature of Machine belief networks. These samples, together with newly observed samples, are then used to form a new Machine belief network that is tailored for usage with the new data. However, an adaptive Machine belief network has the issue of requiring constant memory utilisation.

Table 1: Comparative analysis of algorithms.

Authors	Algorithms Applied	Findings
Hinton et al. [1]	Machine Learning generative model	Using 128-bit codes, the authors demonstrate that the binary codes of semantically linked documents are relatively near in Hamming space. The binary coding of the papers allows for retrieval of information.
Ranzato et al. [2]	Supervised ML and unstructured ML for data.	Concise representation are more efficient since they use less storage space when indexing and conduct fewer calculations.
Miklov et al. [3]	Natural Language Processing (NLP)	The researchers discovered that word vectors trained on enormous amounts of data show deep semantic links between words, such as the relationship between a city and the country to which it belongs - for instance, Paris refers to France and Hamburg to Germany.
Li et al. [5]	Artificial Neural Networks	In the context of Data Analytics, Machine Learning might be useful for discriminative semantic labelling of data.
Calandra et al. [12]	Machine Learning	

In the above table 1 the comparative analysis is given.

III. PROBLEM IDENTIFICATION

Data storage, indexing, and labelling are also key concerns in Big Data Analytics, as is the need for rapid access to retrieved information. Modern data analysis and data management technologies are required when working with data. The large dimensionality of data in the data analytics sector was addressed in a recent work in which we analysed feature selection algorithms.

IV. PROPOSED WORK OBJECTIVES

To build a secure technique to provide secures storage for that data.

Build authentication mechanism to provide secure access for user's data.

Use secure hash function to make technique to build a mechanism to provide secure

auditing for the user's data.

In that way secure framework for secure data analysis and access user's data over Machine Learning is provided.

V. PROPOSED METHODOLOGY

In proposed method an improved token information checking calculation over space procedure depends on examination of beforehand work over data analysis in machine learning web dataset, In the current strategy like weighted token , hyper incited look, security calculation are ready to perform productive answer for the observing. In this manner Enhanced calculation method is proposed. In the current strategy calculation utilizes essential structure to decide the significance of web data. Furthermore, allocate token for the page in light of information checking is done.

A stream outline for the propose strategy is appeared. In that system an arranged request of the web token and security is given which give a superior approach to look information. In this part a stream chart for the proposed strategy is introduced. To actualize this strategy a token dataset is utilized which contains information about various session which can be performed by the interloper, programmer, or other unapproved clients. In propose system first dataset are stacked in to the database, at that point proposed method is connected to shape diverse bunches from the dataset, the groups are framed on the premise of convention. At that point datasets are isolated into various areas. At that point secure session is connected to discover distinctive connections.

Three Descriptions of Proposed Methodology

The methodology has major three components;

1. Structural damage and oil outflow calculation.

2. Consequence assessment and,

3. Design Comparison.

The result for the first two steps feed into the design comparison. The division of each step into tasks is shown in fig 2 .Each task involves both theoretical and methodologies challenges.

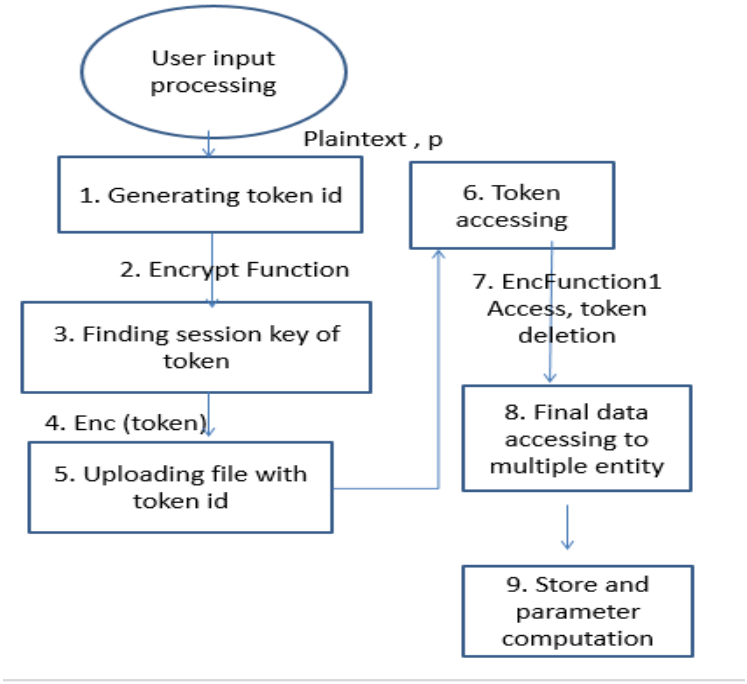

Figure 2: Flow diagram of proposed method

In the figure 2, a complete proposed flow diagram is presented, which is proposed approach and token access.

PROPOSED ALGORITHM

SHKTA (Secure hash key token approach)

Input: Input login credentials, token id tid, session creation framework;

Output: Token creation, security management, secure data management;

Steps:

Begin [

While (session>0)

{

Initialize token tidparameter();

Initialize all user Uid();

Initialize input session for token();

Foreach transaction (1-n)

{

Loading user inputs;

File processing();

Token embed();

};

Token usage()

{

If(session is available)

{

Token access();

Encryption token data();

Observed detail action();

Return ct,cc, bw parameters()

VI. CONCLUSION

Computer science may be able to solve some of the issues that occur when dealing with large amounts of data. It helps automate the approach for extracting complex data representations from large volumes of unsupervised data. In order to benefit from Big Data Analytics, large volumes of unstructured data must be analyzed. Machine learning's hierarchical learning and extraction of numerous layers of intricate data abstractions make semantics indexing, data tagging, retrieval of information, and racist and discriminatory activities such as classification methods simpler for machine learning.

REFERENCES

[1]. Hinton G, Salakhutdinov R: **Discovering binary codes for documents by learning deep generative models.** *Topics Cogn Sci* 2011,**3**(1):74–91. 10.1111/j.1756-8765.2010.01109.x

[2]. Ranzato M, Szummer M (2008) Semi-supervised learning of compact document representations with deep=networks. In: Proceedings of the 25th International Conference on Machine Learning. ACM. pp 792–799.

[3]. Mikolov T, Chen K, Dean J (2013) Efficient estimation of word representations in vector space. CoRR: Computing Research Repository: 1–12. abs/1301.3781

[4]. Mikolov T, Le QV, Sutskever I (2013) Exploiting similarities among languages for machine translation. CoRR: Comput Res Repository: 1–10. abs/1309.4168

[5]. Li G, Zhu H, Cheng G, Thambiratnam K, Chitsaz B, Yu D, Seide F (2012) Context-dependent deep neural networks for audio indexing of real-life data. In: Spoken Language Technology Workshop (SLT), 2012 IEEE. IEEE. pp 143–148

[6]. Hinton GE, Osindero S, Teh Y-W: **A fast learning algorithm for deep belief nets.** *Neural Comput* 2006,**18**(7):1527–1554. 10.1162/neco.2006.18.7.1527

[7]. Dean J, Corrado G, Monga R, Chen K, Devin M, Le Q, Mao M, Ranzato M, Senior A, Tucker P, Yang K, Ng A (2012) Large scale distributed deep networks. In: Bartlett P, Pereira FCN, Burges CJC, Bottou L, Weinberger KQ (eds)Advances in Neural Information Processing Systems, 1232–1240.

[8]. Socher R, Lin CC, Ng A, Manning C (2011) Parsing natural scenes and natural language with recursive neural networks. In: Proceedings of the 28th International Conference on Machine Learning. Omnipress. pp 129–136

[9]. Kumar R, Talton JO, Ahmad S, Klemmer SR (2012) Data-driven web design. In: Proceedings of the 29th International Conference on Machine Learning. icml.cc/Omnipress

[10]. Le QV, Zou WY, Yeung SY, Ng AY (2011) Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In: Computer Vision and Pattern Recognition (CVPR) 2011 IEEE Conference On. IEEE. pp 3361–3368

[11]. Zhou G, Sohn K, Lee H (2012) Online incremental feature learning with denoising autoencoders. In: International Conference on Artificial Intelligence and Statistics. JMLR.org. pp 1453–1461.

[12]. Calandra R, Raiko T, Deisenroth MP, Pouzols FM: **Learning deep belief networks from non-stationary streams.** In *Artificial Neural Networks and Machine Learning–ICANN 2012*. Springer, Berlin Heidelberg; 2012:379–386. 10.1007/978-3-642-33266-1_47

[13]. Chen M, Xu ZE, Weinberger KQ, Sha F (2012) Marginalized denoising autoencoders for domain adaptation. In: Proceeding of the 29th International Conference in Machine Learning, Edingburgh, Scotland

[14]. Coates A, Ng A (2011) The importance of encoding versus training with sparse coding and vector quantization. In: Proceedings of the 28th International Conference on Machine Learning. Omnipress. pp 921–928

[15]. Goodfellow IJ, Warde-Farley D, Mirza M, Courville A, Bengio Y (2013) Maxout networks. In: Proceeding of the 30th International Conference in Machine Learning, Atlanta, GA

[16]. Coates A, Huval B, Wang T, Wu D, Catanzaro B, Andrew N (2013) Deep learning with cots hpc systems. In: Proceedings of the 30th International Conference on Machine Learning. pp 1337–1345

[17]. Glorot X, Bordes A, Bengio Y (2011) Domain adaptation for large-scale sentiment classification: A deep learning approach. In: Proceedings of the 28th International Conference on Machine Learning (ICML-11). pp 513–520

[18]. Chopra S, Balakrishnan S, Gopalan R (2013) Dlid: Deep learning for domain adaptation by interpolating betweendomains. In: Workshop on Challenges in Representation Learning, Proceedings of the 30th International Conference on Machine Learning, Atlanta, GA

[19]. Suthaharan S: **Big data classification: Problems and challenges in network intrusion prediction with machine learning.** In *ACM Sigmetrics: Big Data Analytics Workshop*. ACM, Pittsburgh, PA; 2013.

[20]. Wang W, Lu D, Zhou X, Zhang B, Mu J: **Statistical wavelet-based anomaly detection in big data with compressive sensing.** *EURASIP J Wireless Commun Netw* 2013, **2013:** 269.