Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 165 (2019) 104–111

Procedia Computer Science

www.elsevier.com/locate/procedia

INTERNATIONAL CONFERENCE ON RECENT TRENDS IN ADVANCED COMPUTING 2019, ICRTAC 2019

A Review of Dimensionality Reduction Techniques for Efficient Computation

S.Velliangiri[a]*,S.Alagumuthukrishnan[b], S Iwin Thankumar joseph[c]

[a,b]*CMR Institute of Technology, Kandlakoya Village,Hyderabad 501401, India*
[c]*Karunya Institute of Technology, Coimbatore 641114, India*

Abstract

Dimensionality Reduction (DR) is the pre-processing step to remove redundant features, noisy and irrelevant data, in order to improve learning feature accuracy and reduce the training time. Dimensionality reductions techniques have been proposed and implemented by using feature selection and extraction method. Principal Component Analysis (PCA) one of the Dimensions reduction techniques which give reduced computation time for the learning process. In this paper presents most widely used feature extraction techniques such as EMD, PCA, and feature selection techniques such as correlation, LDA, forward selection have been analyzed based on high performance and accuracy. These techniques are highly applied in Deep Neural Network for medical image diagnosis and used to improve the classification accuracy. Further, we discussed how dimension reduction is made in deep learning.

*Keywords:*Dimensionality reduction;Feature Selection; Feature Reduction; Feature Extraction

* Corresponding author. Tel.: +91-9500519166.
*E-mail address:*velliangiris@gmail.com

10.1016/j.procs.2020.01.079

1. Introduction

Deep learning is the most focused area in data science. Deep Learning techniques use High performances Graphical Processing Unit (GPU) to process the massive amount of data and predict classification. Feature Extractions and Feature selection are mixed with Deep learning training model. However, feature selections and feature extractions are an individual part of the machine learning model. The main advantage of deep learning model handle a massive amount of data and takes less testing time [1]. Dimensionality reduction (DR) is the preprocessing step, which can be applied in any high dimensional data analysis, then that data can be visualized and modeled. DR can do in two different methods. In the first method, Feature selection techniques are select essential features from the input data set. The second method is feature extractions, which creates new features from the existing features in the input dataset. Both feature extraction and selection methods are combined or isolated, which used to improve the calculated precision and accuracy of deep learning [2]. In general, features are classified as suitable, unnecessary, or repeated. A subset of features can be generated from all available input features dataset for the efficient method of the training method. The subset dataset with the limited amount of dimensions is mostly contributed while predicting the learning accuracy [3][4]. Feature family and selection plays a vital role, and techniques used such as PCA, Linear Discriminate Analysis (LDA), and Non-Linear Principal Component Analysis (NLPCA). User have to take correct decision based on which algorithm suits and give excellent result for a given problem.

2.Literature Review

Dimensional reduction techniques have been applied in the medical field. Today, a massive volume of data such as patient symptoms, patient medical test reports, and medications are generated in the medical field. The feature is characteristics of a dataset that are used in applications. The feature can be represented in the medical field like patient test report, patient history, side effects, comprise of many factors, sorting the wellbeing status of a patient. The main objective of the section is to review how dimensional reduction techniques are applied in the medical field. Here listed some benefits of dimensionality reduction techniques applied to a dataset.

1. As the number of dimensions comes down, data storage space can be reduced.
2. It takes less computation time only.
3. Redundant, irrelevant, and noisy data can be removed.
4. Data quality can be improved.
5. Some algorithms do not perform well on more number of dimensions taken. So reducing these dimensions helps an algorithm to work efficiently and improves accuracy.
6. It is challenging to visualize data in higher dimensions. So, reducing the dimension may allow us to design and examine patterns more clearly.
7. It simplifies the process of classification and also improves efficiency.

Zhao and Du [5] recommended Dimensions reduction technique to use spectral-spatial feature based classification(SSFC) framework to decrease the spectral dimension. Similarly, complex information set spontaneously extracted using Convolutional Neural network (CNN) framework. Initially, the extracted features are stacked and given as input to Linear Regression classifier in order to perform classification. SSFC is examined with two favorite HIS data sets, and SVM classifier is used for image classification. Yan Xu after al [6] proposed spontaneous removal of piece picture from side to side deep learning. There is a supervised and unsupervised learning method in DNN. PCA is used for dimension reduction and classification purpose. Multiple Instance Learning (MIL) applied. First, learn natural and restrained structures from actual dataset and extract features from an image. The dataset consists of has high-resolution histopathology images of 132 patients. The result shows that automatic feature learning is superior to old-fashioned feature selection.

Min Chen et al. [7] proposed model to help unsupervised picture highlights learning for lung knob unlabeled information using a convolutional autoencoder deep learning system, which needs a little measure of named information for active component learning. Autoencoder separate yield information to recreate input information and contrast it and unique info information. Convolution autoencoder consolidates the neighborhood convolution association with the autoencoder to remake the info information for convolution activity. Dataset comprises of 4500 patients lung CT pictures from 2012 to 2015. Yang et al. [8] reviewed deep learning techniques successfully handle the focal issues in aspiratory knobs diagnosing, containing highlight extraction, knob recognition, false-positive decrease and considerate threatening order for the enormous volume of chest filter father. Deep learning helped choice help for pneumonic knobs diagnosing. The two – dimensional CNN, three-dimensional CNN, and Deep belief network are used for classification. Convolutional autoencoder neural network is used for feature extraction.

Rasool et al. [9]proposed an automatic feature generation method to enhance the detection and diagnosis of various types of cancer. Unsupervised feature learning can be used for detection of cancer and analysis its type from gene expression data. In the feature learning process, they used softmax regression as the learning approach for the classifier. 10-fold cross-validation is performed to appraise performance of classifier, and consequences are presented using the average classification accuracy. Y. Zheng et al. [10] have proposed a diabetic detection technique using artificial neural network. Exploratory outcomes demonstrate that proposed method is a trustworthy method for diabetes location using the reduced amount of computation expense and great exactness. Highlight extraction strategies discovered significantly more reasonable for mechanized recognition of ophthalmologists maladies than highlight determination techniques in light of noisy information.

The big difficult in image analysis in the medical field is shortage of label dataset.. Moreover, the maximum of the datasets that are compared to biomedical contains noisy data somewhat of redundant or irrelevant data. The deep learning is trained, where in huge amount of factors involved, a huge dimension of characterized information remains essential to be a successful model. Medical diagnosis is not easy. It needs endless activities, incredible determinations, and complexity in the assessment of disease shapes. So, professionals are needed to construct a trustworthy dataset for medical image diagnosis applications. In general, supervised learning models, PCA technique is used for feature extraction. Full feature dataset is recommended; randomly selected feature subset is used for classification. Mamta et al. proposed framework includes a system that breaks down the further bits of proof from the picture and interfaces the small forms into extended ones. Proposed system outcomes are prepared contrasted and conventional method. It gives better edge coherence. More typical results can be achieved by using deep learning method[11].

2.1 *Dimensionality Reduction Approaches*

In deep learning, dimensionality refers to the amount of features in the dataset. , Few algorithms fight to train powerful method, if number of entry in the dataset is less than the records of features in the dataset. This technique is the "Curse of Dimensionality," and hence, it needs to be reduced for efficient classifications. There are different dimensionality reduction approaches, here listed in below Table.1.

Table1: Dimensionality Reduction Techniques

S.No.	Dimensionality Reduction techniques
1	High Correlation Filter
2	Low Variance Filter
3	Missing Value Ratio
4	Random Forest
5	Backward Feature Elimination
6	Factor Analysis
7	Projection methods

8	Forward Feature Selection
9	Linear Discriminant Analysis
10	Independent Component Analysis
11	Principal Component Analysis
12	General Discriminant Analysis

There are two essential methods of dimensionality reduction: i)Feature Selection (FS), ii) Feature Extraction (FE)[12]. FS is that no knowledge required to select the features from the original data set. In feature selection, information can be lost since some features should be excluded when the process of feature subset choice by doing this information can be reduced. However, in feature extraction, dimension can be decreased without losing much initial feature dataset. The selection between FS and FE algorithm be influenced by upon the particular dataset used in the projects.

2.1.1 Feature Selection

In feature selection, High dimensional dataset contains an enormous volume of features that be able to be misleading, inappropriate, or duplicate, which also increasing search space dimension and trying to prepare the dataset for the learning process. So, we need to obtain the subdivision of features from original dataset. FS methods choose the suitable features from the original features. Feature selection algorithm uses computing principle to select the features. Feature selection algorithm phase is divided into two phase such as (i) Subset Generation:(ii) Subset Evaluation: In subset Generation we need to generate subset from the input dataset and using Subset Evaluations we have to check whether the generated subset is optimal or not. These operations can be done in three ways, Filtering, wrapper, and Embedded. Filter methods usage feature rank as the standard criteria for feature collection by arrangement. It selects only relevant features, which penetrates more correlation among it from original feature dataset. Figure1 shows the overall method of the feature selection process.

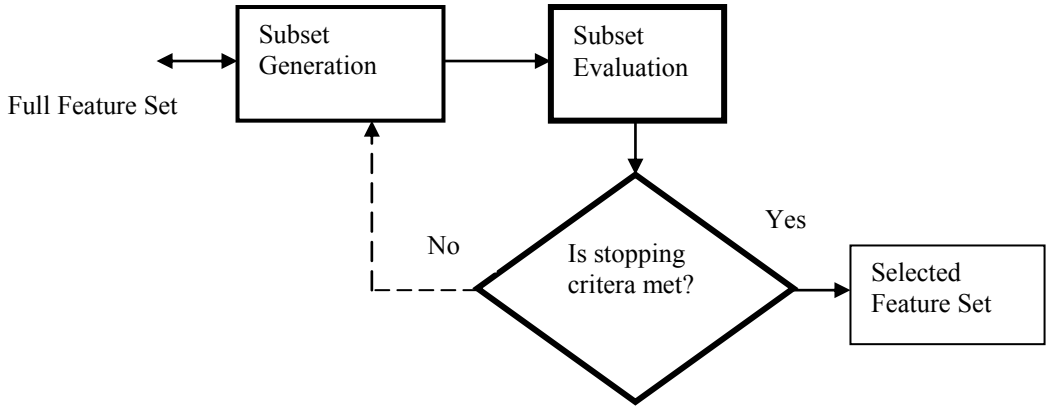

Figure1 Process of Feature Selection

Wrapper approaches feature subset is generated using the classification technique. Wrapper approaches apply the interpreter as a black box. It uses objective functions to maximize evaluations of subset through searching algorithms. Wrapper methods are divided into two types such as i) Sequential Selection Algorithms (SSA) ii) Heuristic Search Algorithms (HAS). The SSA begins with an unfilled set, for example (full dataset) and includes features from dataset until the target output is achieved. In order to seep up the choice process, criterion had been selected, which gradually intensifications the objective function until the maximum is to be reached with the less number of features. Figure2 Depict the Hierarchy of Feature Selection Techniques[13].

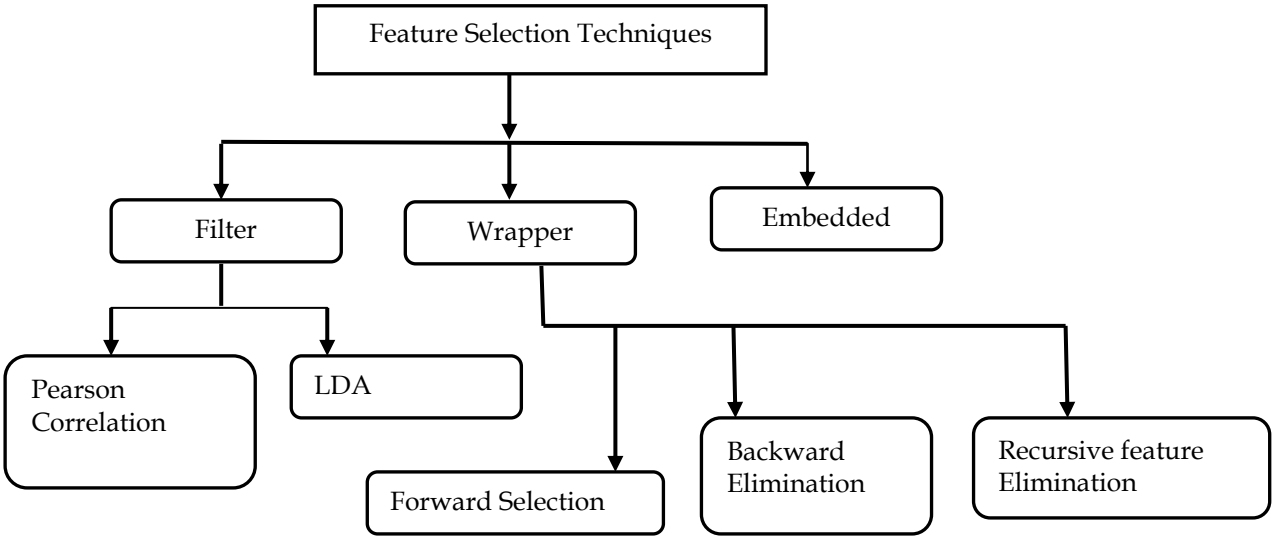

Figure2 Hierarchy of Feature Selection Techniques

2.1.2. *Feature Extraction*

Feature extraction (FE) method extracting new feature from original dataset, and it is very beneficial when we want to decrease the number of resources required for processing without missing relevant feature dataset. Feature extraction can also decrease the number of additional features for an offered study. FE produces remarkable transform of first features to create more significant features. It shows an honest representation of data by reducing the complexity of the data, which describing each variable in feature space. There are different FA method such as PCA, LSA, LDA, ICA, PLS, etc., Karl introduces PCA is most widely feature extraction method.PCA is a straightforward non-factor strategy applied to extricate utmost noteworthy data from the lot of surplus. Principal component analysis is a straight change of information that limits duplication that is estimated through covariance and expands the data that is estimated by the variance. Figure3 Depict the Overall process of Feature Extraction method [14].

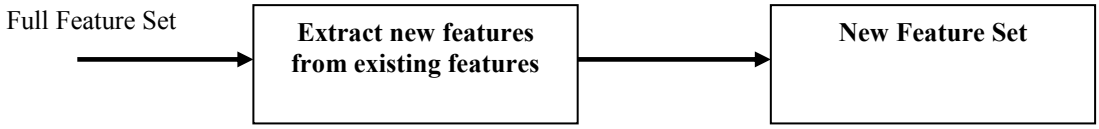

Figure3 Feature Extraction Process

Convolutional Neural Network (CNN) uses deep learning algorithm to extract the features from the input dataset. Automatic feature extraction can be done on CNN. CNN automatically extract features, learn, and organize them.CNN can generally perform superior to anything other surely understood classifiers. In any case, there is no methodical examination which demonstrates that programmed highlight removal in CNN is superior to another necessary element extraction procedure, and there is no investigation which demonstrates that other straightforward neural system models cannot accomplish a similar precision as CNN. CNN, with automated feature extraction, provides a better result than manual feature extractions techniques [15].Similarly, Convolutional deep neural network (CDNN) is utilized to perform relation classification. CDNN mines verbal and verdict level features. These features are combined to get the final mined feature route. Softmax classifier is used to predict the relationship among these features. Figure4 shows the Hierarchy Structure of feature extraction method[16].

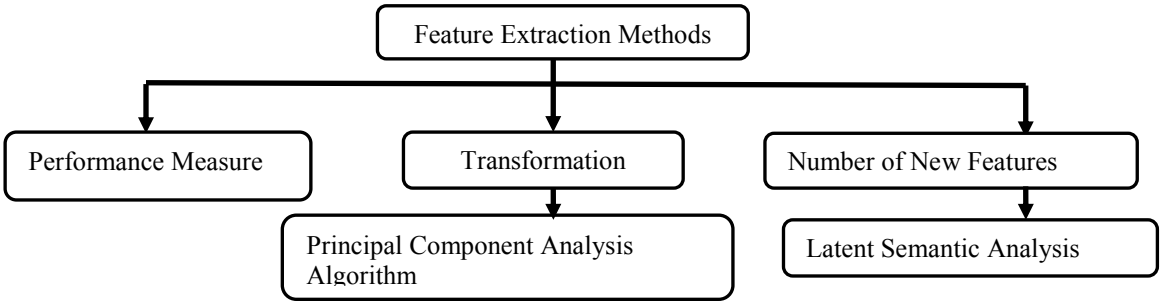

Figure4 Hierarchy Structure of feature extraction

3.Analysis of Feature Selection and Extraction Algorithms

In this section, an overview of FSA with their boundaries discussed. FSA is classified into three methods, such as i) filter methods ii) wrapper methods iii) embedded methods. For each method, separate algorithms to be used. For the filter method, Chi-square test, ANOVA, Correlation coefficient, and LDA are applied to generate a subset of features.

3.1 Feature Selection Algorithms(FSA)

- Pearson's Correlation: It is utilized as a measure for evaluating direct reliance between two consistent factors X and Y. Its value ranges from -1 to +1. Pearson's correlation is given as:
- $\rho X, Y = \frac{COV(X,Y)}{\sigma X \sigma Y}$
- Chi-Square: It is applied to the groups of certain features to evaluate the correlation or association between the features using their frequency distribution.
- LDA: LDA is used to find a linear mixture of features that characterizes two or more group of a categorical variable.
- ANOVA: ANOVA represents the analysis of change. It is like LDA aside from the way that it is worked utilizing at least one absolute autonomous highlights and one consistent ward include. It gives a factual trial of whether the methods for a few gatherings are equivalent or not.

Semwal and et al. [17] proposed Intrinsic Mode Functions for future selection, which has two properties, and one occur among zero exchanges, i.e., it has one mean value of zero. Feature selection is crucial for a specific dataset. Root Square Mean (RMS) and zero crossing rate (ZCR) are essential features, which gives necessary information of a specific data set. IMF values are calculated using the MatLab function. Zou et al. [18]suggested a novel deep-learning-based FS technique. The proposed technique can be applied in remote detecting scene and classification. In scene classification, strong component determination improves the final execution; proposed strategy details the element determination issue as an element reproduction issue. Deep belief network (DBN) include deliberation by limiting the recreation mistake over the entire list of capabilities, and highlights with littler remaking blunders would hold more highlights characteristic for picture portrayal.

Consequently, Deep belief network (DBN) technique chooses very important features that is present in the input dataset. Especially, an incremental method is created to adjust the Deep belief network to deliver asked recreation loads. Amid the iterative element learning process, anomaly includes are dispensed with the dataset. The last remaking weight grid is progressively stable for highlight recreation.

3.2 Characteristics of Feature Subset Selection

Search systems select feature subset from original feature dataset based on the unique feature. The algorithm looks hopeful element subsets by certain assessment measure. To depict target originations of the learning procedure by picking the greatest element subset from feature space. In[19], feature selection method performance is evaluated by diverse datasets from the open area. The quantity of decreased features and their impact on learning execution with some generally utilized strategies have been estimated, at that point assessed. In the process of feature selection, the following aspects must be considered. 1. Starting Point of dataset 2. SearchStrategy3.Evaluation of feature subset and 4.Stopping Criteria. Based on these aspects, Table 2 discusses the comparative study of the feature selection algorithm.

Table 2.Comparative Study of Feature Selection

Method	**Single or Subdivision Feature**	**Initial Fact**	**Exploration Plan**	**Sub-division Creation**	**Sub-division Assessment**	**Ending Rules**	**Used to Reduce**
Statistical	Single	Complete feature set	Random	Weighted	Discrepancy	Relevance	Unrelated Features
Iterative Feature Learning	Single	Full feature set	Random	Weighted	relevant	Relevance	Irrelevant Features
Multiple Instance Learning	Subset	The arbitrarya mount of features	Random	Weighted	relevant	Ranking	Redundan t features
PCA	Subset	Complete feature set	Sequential	Forward selection	Limited Common Information	Ranking	Redundan t features / noisy
Clinical Scores Regressio n	Subset	The arbitrary amount of features	Random	Weighted	relevant	Relevance	Unnecessa ry Features

3.3 Feature Extraction Algorithms(FEA)

Table3: Comparative Analysis of Feature Extraction Algorithms

Algorithm	**Network**	**Application**
Empirical Mode Decomposition	Deep Neural Network	Humaniod Push recovery
Batch normalization	Deep Convolutional Neural Network	Structural Damage Detection
Balanced Local Discriminant Embedding	Convolutional Neural Network	Hyperspectral Image Classification
Principal Component Analysis	Convolutional Deep Belief Network	Audio Classification
Support vector machine - classifier	Deep Neural Network	Medical Image Analysis

Feature extraction process begins from group of calculated data and constructs resulting features in deep learning. It reduces irrelevant, redundant, and noisy data. FEA planned to be instructive, which facilitate to improve the accuracy subsequent feature learning process. High-level dimensional data to be reduced into a low-level dimension

data for better classification. There are various FSA algorithms which suit for classifier on artificial neural networks.

5. Conclusion and Future Scope

The main objective of this study is discussing about basic concept of dimensionality reductions techniques. Dimensionality reduction is the pre-processing method to build a good classification model. Feature extraction and feature selection are the first two techniques used to reduce dimensionality. The feature selections complete dimensions reductions by selecting a subset feature without doing transformations; however, in feature extractions, dimensions reductions are made by creating a new set of feature from the input datasets. Investigation of feature selection and feature extraction algorithm is presented. Feature selection and extraction reduces the Machine learning techniques computation time. Algorithms are discovered, which facilitate new high dimensional data samples can be translated into low dimensional space. So, the substantial storage data set to become less necessary. This study will helpful for researchers to improve their research in the field of dimensionality reduction techniques.

References

[1] P. Kaur, M. Sharma, and M. Mittal, (2018) "Big Data and Machine Learning Based Secure Healthcare Framework," Procedia Comput. Sci., **132**: 1049–1059.

[2] K. U. Leuven and N.- Psychofysiologie, Nikolay Chumerin, Marc M. Van Hulle K.U.Leuven, Laboratorium (2006) "Comparison Of Two Feature Extraction Methods Based On Maximization Of Mutual Information," Mach. Learn: 343–348.

[3] H. M. H. Li, (2002) "Feature selection, extraction and construction," Commun. IICM (Institute Inf. Comput. Mach. Taiwan), **5:** 67–72.

[4] M. Mittal, L. M. Goyal, D. J. Hemanth, and J. K. Sethi,(2018,2019) "Clustering approaches for high-dimensional databases: A review," Wiley Interdiscip. Rev. Data Min. Knowl. Discov., no.: 1–14.

[5] W. Zhao and S. Du, (2016) "Spectral-Spatial Feature Extraction for Hyperspectral Image Classification: A Dimension Reduction and Deep Learning Approach," IEEE Trans. Geosci. Remote Sens. **54(8):** 4544–4554.

[6] Y. Xu, T. Mo, Q. Feng, P. Zhong, M. Lai, and E. I. Chang,(2014) "Deep Learning Of Feature Representation With Multiple Instance Learning For Medical Image Analysis"State Key Laboratory of Software Development Environment , Key Laboratory of Biomechanics and Mechanobiology of Ministry of Education , Beihang University M," Icassp **1:**1645–1649.

[7] M. Chen, X. Shi, Y. Zhang, D. Wu, and M. Guizani(2017), "Deep Features Learning for Medical Image Analysis with Convolutional Autoencoder Neural Network," IEEE Trans. Big Data, **7790**: 1–1.

[8] Y. Yang et al., "Deep learning aided decision support for pulmonary nodules diagnosing: A review," (2018) J. Thorac. Dis., **10(7):** S867–S875.

[9] R. Fakoor, A. Nazi, and M. Huber,(2013) "Using deep learning to enhance cancer diagnosis and classification," Int. Conf. Mach. Learn.,

[10] Y. Zheng et al. (2013), "An automated drusen detection system for classifying age-related macular degeneration with color fundus photographs," Proc. - Int. Symp. Biomed. Imaging,: 1448–1451.

[11] M. Mittal et al.,(2019) "An Efficient Edge Detection Approach to Provide Better Edge Connectivity for Image Analysis," IEEE Access, :1–1

[12] G. Chandrashekar and F. Sahin, (2014) "A survey on feature selection methods," Comput. Electr. Eng., **40(1)**: 16–28.

[13] Veerabhadrappa and R. Lalitha,(2010) "Bi-level dimensionality reduction methods using feature selection and feature extraction," Int. J. Artif. Intell. Appl., **1(4):** 54–68.

[14] O. Access, (2018) "Variable Selection and Feature Extraction Through Artificial Intelligence Techniques," Long-Haul Travel Motiv. by Int. Tour. to Penang, no. tourism : 13

[15] F. Shaheen, B. Verma, and M. Asafuddoula,(2016) "Impact of Automatic Feature Extraction in Deep Learning Architecture," Int. Conf. Digit. Image Comput. Tech. Appl. DICTA.

[16] D. Zeng, K. Liu, S. Lai, G. Zhou, and J. Zhao, (2014) "Relation Classification via Convolutional Deep Neural Network," 25th Int. Conf. Comput. Linguist. COLING, **2011:** 2335--2344.

[17] V. B. Semwal, K. Mondal, and G. C. Nandi,(2017) "Robust and accurate feature selection for humanoid push recovery and classification: deep learning approach," Neural Comput. Appl., **28(3)**: 565–574.

[18] Q. Zou, L. Ni, T. Zhang, and Q. Wang, (2015) "Deep Learning Based Feature Selection for Remote Sensing Scene Classification," IEEE Geosci. Remote Sens. Lett., **12(11)**: 2321–2325.

[19] C. Yun and J. Yang, (2007) "Experimental comparison of feature subset selection methods," Proc. - IEEE Int. Conf. Data Mining, ICDM,: 367–372