A study on the topic of Unsupervised Machine Learning

Monesa Thoguluva Janardhanan
Computer Science Department
University of North Carolina at Charlotte
Charlotte, United States of America
mthogulu@uncc.edu

***Abstract*—The main aim of this paper is to present the ideas grasped from the study of Unsupervised machine learning. Unsupervised Machine Learning is a branch of machine learning where the data is not labeled or classified or trained like in supervised learning. The objective of unsupervised learning is to discover patterns or clusters from the input data with no help of a human [12]. It has its applications in many industries and with involves huge dataset. 10 research papers which involves unsupervised machine learning algorithms are taken for this study and are presented in this paper. The research mostly deals with feature extraction from the images and some are dealing with medical diagnosis. Some research papers involve both supervised and unsupervised machine learning algorithms as a hybrid architecture. This paper presents a brief idea on some of the unsupervised machine learning Algorithms and gives an overview of the research papers in the Literature Review Section.**

Keywords— Deep Learning, Neural Network, Auto Encoder, K-Means Clustering, K-Prototype clustering, Feature Extraction.

I. Introduction

Unsupervised Learning is a branch is Machine Learning field where there the data is not labeled or trained [12]. It is widely used for classifying or clustering huge number of datasets since the labeling process is expensive, tedious, error-prone and also time-consuming when there are huge datasets. Unsupervised Learning in deep learning algorithms can be adopted in more extensive fields than supervised learning [4]. Unsupervised Learning algorithms are faster compared to supervised learning algorithm in recognition process as there are no training phase and label data for recognition [7]. In the text categorization field as well supervised text categorization requires extra effort to predefine the categories and to assign the category labels to the documents in the training set. This is tedious in a huge and dynamic text database such as world wide web. Also, different human experts may disagree when deciding under which category to a document falls under Due to the rapidly increasing amounts of online documents, the dynamic nature of most text databases makes it difficult to label the categories [8]. Based on the above fact unsupervised learning should be used for nature text classification rather than supervised learning algorithm. Section 2 deals with some of the algorithms which take part in unsupervised machine learning technique such as Deep Learning, Convolutional Neural Networks (CNN), Auto Encoders (AEs), Principal Component Analysis (PCA), Support Vector Machine (SVM), K-Means Clustering, Back propagation, Recurrent Neural Network (RNN), Long Short Term Memory (LSTM), Spatial Information, Temporal Information, K-Prototype Clustering, Bag of Visual Word, Latent Semantic Indexing(LSI), Cluster Based on Retrieval of Images (CLUE), Feature Learning, Convolutional Auto Encoder, Modality Classification based on the used references. Section 3 is about the Literature review or the related work in the unsupervised machine learning and gives an overview of the research, the algorithm used and their accuracies.

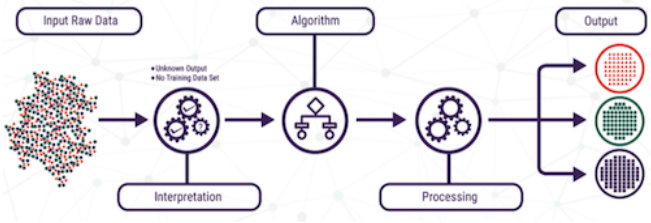

Figure 1: Unsupervised Learning Architecture [13]

II. Machine Learning Algorithms used in this study

This section of the paper deals with some of the algorithms which are used in the study of randomly selected research paper.

A. K-Means Clustering

In K-means classification clusters are formed. These clusters are nothing but collection of data points which are to be classified for splitting of category. So, each clusters classifies the data points and put it into the corresponding category [3]. Normally the features obtained using parameters such as mean, median, standard deviation, minimum and variance and maximum are given as an input to the K-means clustering algorithm. The K-means algorithm is as shown in the Figure 2. K-Means Algorithm is based on the measures in terms of Euclidian distance [3].

- Let D denotes set of data points and P denotes set of cluster centers. Initially randomly select the centers.
- Calculate Distance between data points and centroid using formula

$$P_j = (1/k)\sum_{n=1}^{k} D_j$$

- Data points having minimum distance to that centroid will belong to that cluster.
- Recalculate the new centroid for the clusters.
- Recalculate distance between data points and new centroid.
- Repeat step 3 until each data point fitted correctly.

Figure 2: K- means Algorithm [3]

First, the data points are divided into two clusters. Initially the center points are taken randomly. Then the distance between the points are calculated. Data points which are

having minimum distance towards the centroid belong to that particular clusters and the clusters are formed. This step is repeated until there are no more movement of data points. The drawback of K-means algorithm is that it works only on numeric values, which prohibits it from being used for clustering data set containing categorical data [6].

B. K-Prototype Clustering

The K-Protoype Clustering Algorithm is an extension of K-Means Clustering Algorithm. It integrates K-Means and K-Modes process to cluster data with different attributes [6][11]. It uses K-Modes approach to update the categorical values of the centroids. As per efficiency the K-Prototype clustering is similar to K-Means Clustering. The advantage of K-Prototype Clustering is that it can classify both numerical and categorical attributes [6].

C. Auto Encoders

Auto Encoders are nothing but neural networks that attempt to transform inputs into outputs with least possible distortion [1] which are used for feature extraction. An auto-encoder neural network is an unsupervised machine learning algorithm that consists of two parts, an encoder function that transforms input into low-dimensional representations and a decoder function that produces a reconstruction from the representation [14]. In the hidden layers are considered to have the maximum information. In the deep learning, several auto encoders are learned, and multiple layers are stacked. Another feature extraction method is Principal Component Analysis PCA, but there is a limitation on PCA that it cannot capture non-linear relationship. So AE is preferred over PCA. Figure 3 shows a basic architecture of an auto-encoder [18].

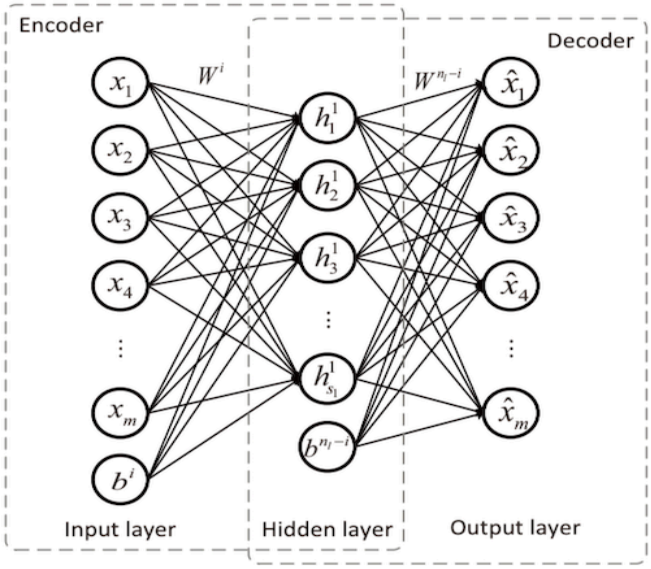

Figure 3 : The architecture of basic auto encoder [1]

D. Support Vector Machine

Support Vector Machine constructs the hyperplane which creates the different classes for given training points. The main aim is to construct a hyperplane having maximum distance between nearest training data points. Support Vector Machine predicts the classes for input data which is feature vector. It classifies training data in terms of two classes 0 or 1 [4].

E. Neural Network

Neural Network consists of processing units called neurons. ANN tries to imitate the human brain. Here the Dendrites are the Input, Synapse via Axon is the output and there is an Activation function which determines whether to activate or deactivate the neurons [16]. So the overall model consists of input, a bias which is by default 1 is added to the input and weight [15]. A sample model of ANN model is displayed as shown in the Figure 2.

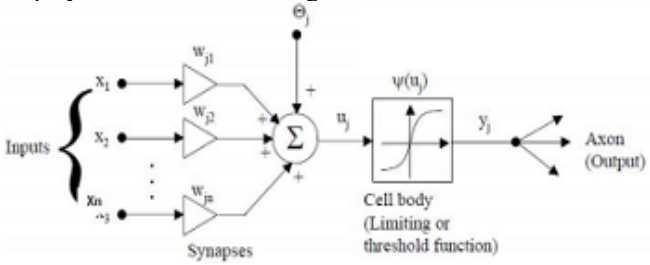

Figure 4: Neural Network [15]

The product of the weight and the input gives the strength of the signal. The weight indicates the connection between the neurons [16]. If the weight is positive, then the neurons are connected and if it is negative the signal strength of the neurons are reduced and if the weight is 0 there are no connection between neurons. This weight changing is based upon the Activation function which will be discussed in this paper in the later sections. This process of adjusting weights of the ANN to get the required output is called learning or training [17]. So, the model can be represented as shown in the below equation.

$$y_j = \Psi \sum_{n=1}^{n} \left(w_{ji}\, x_i + \theta_j \right)$$

Figure 5: ANN representation

Here x is the input, y is the output, w is the weight, θ is the bias. Feedforward Networks, Feedback Networks and Lateral Networks are the architecture of ANN which is discussed below.

F. Backpropagation

Backpropagation training algorithm is based on Gradient Descent. This Algorithm is most commonly used method in training neural network. It comes under supervised machine learning algorithms with feedforward architecture [16].

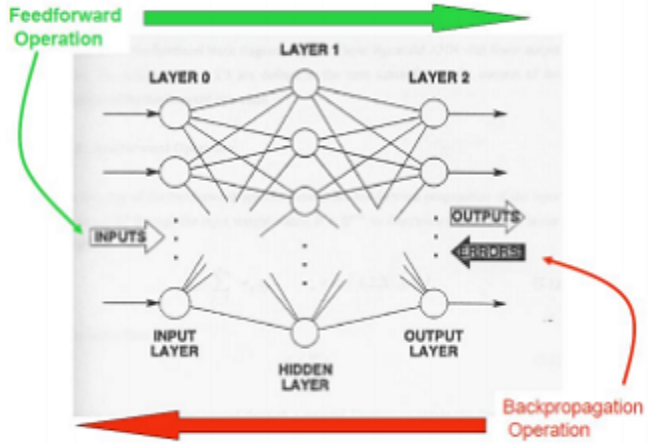

Figure 6 : Neural network with Gradient Descent [16]

It is one of the most frequently utilized neural network techniques for Classification and Prediction. Here once the output is calculated the output is matched with the desired output and the difference between the targeted output and the output obtained on propagated back to the layers [16]. As the flow again starts it changes the weight of the neurons/nodes in the hidden layers. This cycle of going forward from input to output and from output to input is called epoch. This cycle stops until the output obtained is matched with the targeted output with some tolerance [17].

G. Convoluted Neural Network

Convolutional Neural Networks (CNN) is one of the variants of neural networks used heavily in the field of Computer Vision. It derives its name from the type of hidden layers it consists of. The hidden layers of a CNN typically consist of convolutional layers, pooling layers, fully connected layers, and normalization layers. CNN does not use normal activation function. Instead they use CNN and pooling activation function [11]. CNN architecture is displayed as shown in the Figure 7.

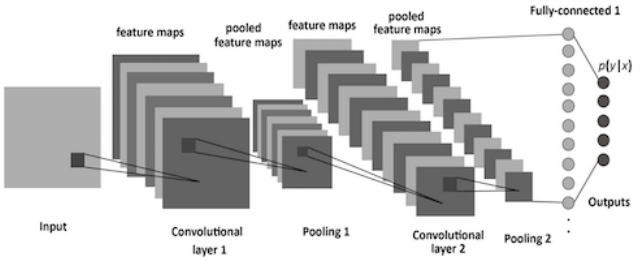

Figure 7 : Convoluted Neural Network Architecture [11]

H. Recurrent Neural Network

Recurrent Neural Networks or RNN in other words, are a very important neural networks which are heavily used in Natural Language Processing. In a general neural network, an input is processed through a number of layers and an output is produced, with an assumption that two successive inputs are independent of each other [11]. RNNs are called recurrent because they perform the same task for every element of a sequence, with the output being depended on the previous computations. Another way to think about RNNs is that they have a "memory" which captures information about what has been calculated so far. In theory, RNNs can make use of information in arbitrarily long sequences, but in practice, they are limited to looking back only a few steps [11]. RNN architecture is as shown in the Figure 8.

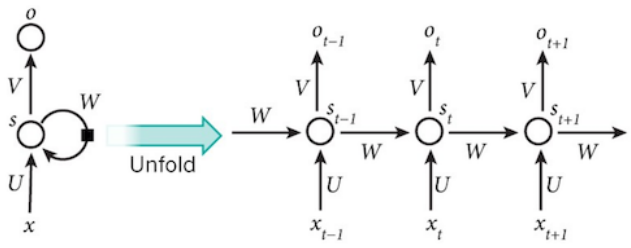

Figure 8 : Recurrent Neural Network Architecture [11]

III. Literature Review

A. Breast Cancer Diagnosis Using an Unsupervised Feature Extraction Algorithm Based on Deep Learning [1]

Reference [1] demonstrates Breast cancer Diagnosis using an unsupervised feature extraction algorithm based on deep learning. The main aim of this paper is to reduce the cost and the time which is taken by supervised learning algorithms as they need to be labeled and trained which is a very tedious process. The data sets used is from Wisconsin Diagnosis breast cancer data set. For feature extraction Auto encoders are being used. This paper proposes a deep learning based unsupervised feature extraction scheme that combines stacked auto encoders with SVM. So, in other words it is SAE-SVM. Figure 9 present a schematic diagram for breast cancer diagnosis model using SAE-SVM.

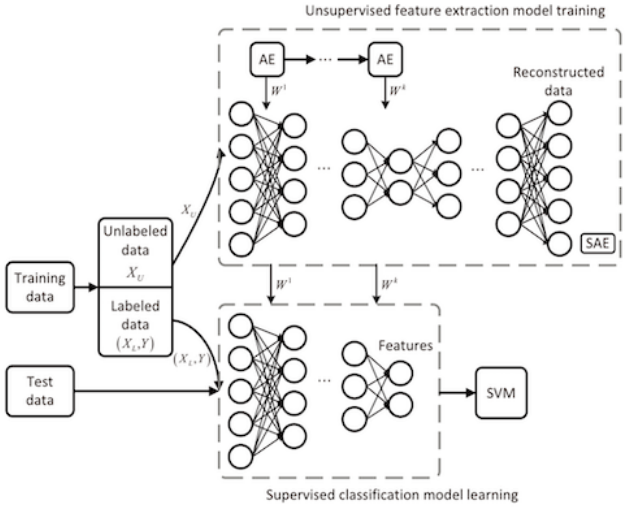

Figure 9 : A schematic diagram of the breast cancer diagnosis model based on SAE-SVM [1]

Here the AE's denotes the weight and this architecture combines both supervised and unsupervised models. The results states that the accuracy acquired is about 98.25%.

B. Unsupervised Deep Learning Features for Lung Cancer Overall Survival Analysis [2]

Reference [2] focuses on Unsupervised Depp Learning Features for Lung Cancer Overall Survival Analysis. This research includes Unsupervised feature learning and Supervised Cox model training. The datasets are split into two one which is labeled with survival time and the other is not labelled with survival time. Residual Convolutional Auto Encoder (RCAE) is used and trained the model without the survival time. Cox proportional model is constructed on the data sets which has survival time. This paper states that Deep learning shows good performance in many medical image analysis tasks however they are not well applied in survival analysis because of the limited data amount.

C. Melanoma skin cancer detection and classification based on supervised and unsupervised learning [3]

Reference [3] is a comparative study of melanoma skin cancer detection and classification of supervised and unsupervised learning. Neural Network, K-means Clustering, Support Vector Machine Algorithms are used for this research and the results indicates that SVM is better than NN and K-means. Fig. 10 shows the flow of data transfer of this research.

Preprocessing is done using median filtering. With the help of median filtering the unwanted part of the skin image is removed, and a smooth image is obtained. Then it goes to histogram equalization. In histogram equalization there are two methods one is global histogram, and another is local histogram. Global histogram considers the overall image while the local histogram considers only a part of the image. After the feature extraction, it is sent to Neural Network, K-Means Clustering and Support Vector Machine as shown in the figure. These features are obtained using parameters such as mean, median, standard deviation, minimum, variance and maximum. Back Propagation algorithm is used in this Neural Network model. The accuracy of the K-means algorithm is 52.63% and for neural network it is 60 to 75% and for SVM it is 80% to 90%.

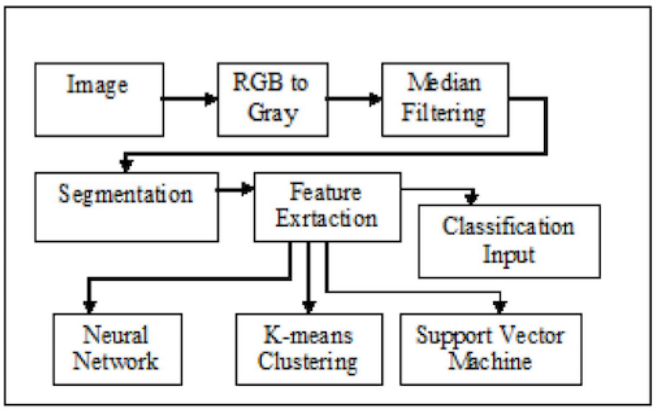

Figure 10: Flow of data transfer for the system [3]

D. Study on Deep Unsupervised Learning Optimization Algorithm Based on Cloud Computing [4]

Reference [4] proposes an unsupervised machine learning algorithm on cloud computing. The model proposed here is CNN-RNN model. This paper states that the optimize algorithm proposed in the paper can better increase the training efficiency of neural network. Gradient Descent Algorithm is used for pre-training. The training mechanism uses neural network with Back-Propagation algorithm. The BP algorithm is used in Map reduce. It states that compared to the conventional algorithms BP algorithms BP algorithm based on Map reduce can increase the training efficiency of neural network.

E. Unsupervised Learning for Forecasting Action Representations [5]

Reference[5] demonstrates a unsupervised learning for forecasting action representations. This paper instead of forecasting the future from the static image they are using unlabeled video clips and extracting the spatial and temporal information from it. So both Spatial and Temporal Information are jointly learned. Spatial information consists of actions and they are all about visual appearance. Temporal information is all about motion trajectory. This paper proposes to modify LSTM in order to forward output of all time steps to a fully constructed layer for a compact output. Thus it avoids getting rid of some historical information at each time step in memory networks.

F. Unsupervised learning for understanding student achievement in a distance learning setting [6]

Reference [6] proposes an unsupervised machine learning technique for understanding Student Achievement in a distance learning setting. It identifies how the student study is affected based on some attributes such as age, gender, demographic locations and their previous education. K-means prototype clustering algorithm is used for this study as the data set comprises of both numeric and categorical attributes. The categorical attributes include gender, region and highest education while the numerical attributes include how many times the user has taken this course, has accessed this course and their credits. The results show that the successful students are more active and are living in the privileged areas and they have higher educational levels compared to the unsuccessful students.

G. An unsupervised approach for traffic sign recognition based on bag-of-visual-words [7]

Reference [7] demonstrates an unsupervised machine learning approach for traffic sign recognition system based on the Bag of visual words. Bag of visual words is a model applied in computer vision. It has its applications in fields such as fingerprint identification, adult image classification. It states that SIFT and SURF methods are widely used for feature extraction. The Key points are grouped into several clusters and the centroid of the cluster is the visual word. Based on this histogram of the visual words are constructed and this histogram becomes the input of the classifier to detect the object. Then K-Means clustering is used to classify the traffic signs. It states that the time consuming of the labeling of data training is the reason to conduct this study. If there are large number of key points the accuracy was more but small number of key points failed to produce high accuracy in this model.

H. A Comparative Study on Supervised and Unsupervised Learning Approaches for Multilingual Text Categorization [8]

Reference [8] demonstrates a comparative study on supervised and unsupervised learning approaches for multilingual text categorization. Support vector machine model is used for supervised learning and Latent Semantic Indexing is used for unsupervised machine learning. There are 6 classifiers covering Astronomy, Physics, Politics, Finance, Medicine and Art. First the data is preprocessed then it is sent to LSI and SVM model for classification. The study states that compared with SVM based supervised technique LSI based unsupervised technique achieved excellent overall performance although its resulting performance was slightly behind the overall performance.

I. An unsupervised learning approach to content-based image retrieval [9]

Reference [9] focuses on unsupervised learning approach to context-based image retrieval. This paper introduces a novel image retrieval scheme, Cluster based retrieval of images by unsupervised learning (CLUE) to handle the semantic problem. It attempts to reduce the semantic gap by providing image clusters instead of set of ordered images. The retrieval process starts with data extraction. The features for the targeted images are computed beforehand and are stored as feature files. The similarity is measures and the results are sent back to cluster the images as shown in the Figure 11.

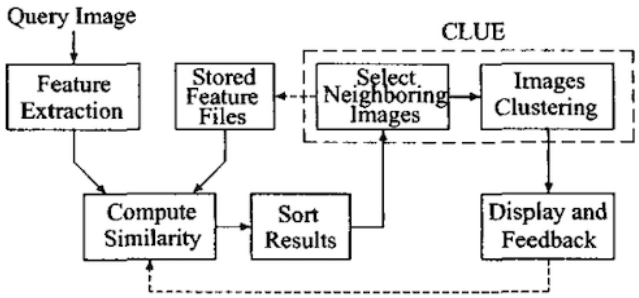

Figure 11 : General CBCIR System [9]

J. *Unsupervised Deep Transfer Feature Learning for Medical Image Classification [10]*

Reference [10] focuses on unsupervised feature learning for medical image classification. This study proposes a new hierarchal unsupervised feature extractor with a convolutional auto encoded placed on the top of pre-trained neural network. The weights of the layer that is to be assigned to CNN is learned by CAE. The activation function used for this convoluted neural network is ReLU. The feature representation is extracted in feed forward manner. Backpropagation algorithm are used in CNN. The reconstruction error is first back propagated and then the weights are updated using stochastic gradient descent. Using this hierarchy this research states that the feature extraction is better compared to other algorithms. An illustration of the CAE-CNN architecture is as shown in the Fig. 12.

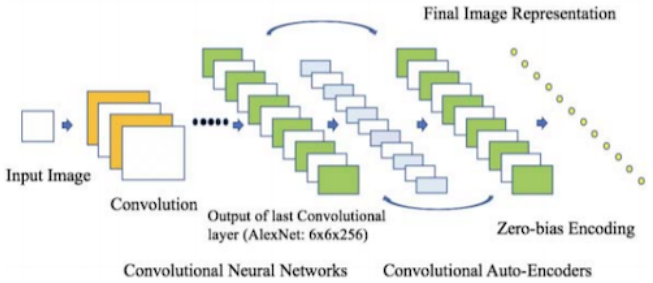

Figure 12 : Architecture of CNN-CAE [10]

IV. Conclusion

This paper demonstrates about the ideas grasped from the study of unsupervised machine learning algorithms. 10 research papers were taken at random. Their study focusses of research and the algorithms used are mentioned in the Literature Review section of this paper. This paper discusses about some algorithms like K-Means Clustering, K-Prototype clustering, Auto Encoders, Support Vector Machine, Neural Network, Back Propagation, Convoluted Neural Network and Recurrent Neural Network. Most of the study indicates that the main aim of using unsupervised machine learning algorithms is because the labeling and training of data is a very complex and time-consuming task in case of supervised machine learning algorithms. Many studies used a hybrid architecture of supervised and unsupervised model for effective results. Some research compared the accuracies between the supervised and unsupervised models, and it states that the supervised model results are better compared to unsupervised models. But the training and labeling of data takes a lot of time.

References

[1] Y. Xiao, J. Wu, Z. Lin and X. Zhao, "Breast Cancer Diagnosis Using an Unsupervised Feature Extraction Algorithm Based on Deep Learning," 2018 37th Chinese Control Conference (CCC), Wuhan, 2018, pp. 9428-9433.

[2] S. Wang et al., "Unsupervised Deep Learning Features for Lung Cancer Overall Survival Analysis," 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Honolulu, HI, 2018, pp. 2583-2586.

[3] H. R. Mhaske and D. A. Phalke, "Melanoma skin cancer detection and classification based on supervised and unsupervised learning," 2013 International conference on Circuits, Controls and Communications (CCUBE), Bengaluru, 2013, pp. 1-5.

[4] H. Yan, P. Yu and D. Long, "Study on Deep Unsupervised Learning Optimization Algorithm Based on Cloud Computing," 2019 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS), Changsha, China, 2019, pp. 679-681.

[5] Y. Zhong and W. Zheng, "Unsupervised Learning for Forecasting Action Representations," 2018 25th IEEE International Conference on Image Processing (ICIP), Athens, 2018, pp. 1073-1077.

[6] S. Liu and M. d'Aquin, "Unsupervised learning for understanding student achievement in a distance learning setting," 2017 IEEE Global Engineering Education Conference (EDUCON), Athens, 2017, pp. 1373-1377.

[7] C. Supriyanto, A. Luthfiarta and J. Zeniarja, "An unsupervised approach for traffic sign recognition based on bag-of-visual-words," 2016 8th International Conference on Information Technology and Electrical Engineering (ICITEE), Yogyakarta, 2016, pp. 1-4.

[8] Chung-Hong Lee, Hsin-Chang Yang, Ting-Chung Chen and Sheng-Min Ma, "A Comparative Study on Supervised and Unsupervised Learning Approaches for Multilingual Text Categorization," First International Conference on Innovative Computing, Information and Control - Volume I (ICICIC'06), Beijing, 2006, pp. 511-514.

[9] Y. Chen, J. Z. Wang and R. Krovetz, "An unsupervised learning approach to content-based image retrieval," Seventh International Symposium on Signal Processing and Its Applications, 2003. Proceedings., Paris, France, 2003, pp. 197-200 vol.1.

[10] E. Ahn, A. Kumar, D. Feng, M. Fulham and J. Kim, "Unsupervised Deep Transfer Feature Learning for Medical Image Classification," 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), Venice, Italy, 2019, pp. 1915-1918

[11] Vibhor Nigam. (2018) Understanding Neural Networks. From neuron to RNN, CNN, and Deep Learning - towardsdatascience.com/

[12] Sanatan Mishra (2017) : TowardsDataScience - https://towardsdatascience.com/unsupervised-learning-and-data-clustering-eeecb78b422

[13] Prena Aditi (2018) Unsuoervised Learning - https://medium.com/@aditi22prerna/unsupervised-learning-a24caf362e79

[14] G. E. Hinton, and R. R. Salakhutdinov, Reducing the dimensionality of data with neural networks, Science, 313(5786):504–507, 2006.

[15] Dishashree Gupta. (2020) Fundamentals of Deep Learning – Activation Functions and When to Use Them? analyticsvidhya.com

[16] Graduate School of Science and Technology, Kumamoto University - Japan cs.kumamoto-u.ac.jp

[17] Shiruru, Kuldeep. (2016). AN INTRODUCTION TO ARTIFICIAL NEURAL NETWORK. International Journal of Advance Research and Innovative Ideas in Education. 1. 27-30

[18] Z. M. Hira, and D. F. Gillies, A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data, Advances in Bioinformatics, 2015: 1–13.