Transition from Observation To Knowledge To Intelligence (TOKI)

Editors
Dr. Victor ODUMUYIWA, Dr. Olufade ONIFADE,
Prof. Amos DAVID & Prof. Charles UWADIA

Victor ODUMUYIWA
Department of Computer Sciences,
University of Lagos
Nigeria

ISBN: 978-978-976-000-8

Transition from Observation to Knowledge to Intelligence

3rd Biennial International Conference on Transition from Observation to Knowledge to Intelligence (TOKI)
15-16 August 2019
University of Lagos, Nigeria

Editors

Dr. Victor ODUMUYIWA
Dr. Olufade ONIFADE
Prof. Amos DAVID
Prof. Charles UWADIA

Development of Deep Learning Model on Mushroom Dataset towards Classifying Poisonous Mushroom with Feature Selection

AYORINDE Ibiyinka Temilola
Department of Computer Science
University of Ibadan, Ibadan. Nigeria

BADMOS Zainab Oyeronke
Department of Computer Science
University of Ibadan, Ibadan. Nigeria

Abstract. Deep learning is a machine learning technique that immitates human action in the area of learning and forming patterns. It can automatically learn features from a large number of image dataset. A multilayer perceptron (MLP) is a deep, artificial neural network that is composed of more than one perceptron. Since the percentage of edible mushrooms is very low compared with the poisonous ones, there is need to know the features that characterize these two types of mushrooms. Hence, this study develops a deep learning model that helps to accurately classify edible and poisonous mushrooms using multi-layer perceptron neural network. The dataset was collected from the UCI Machine Learning repository. Odour and spore-print colour are the two major features that help in the classification. Two different models namely initial model (without principal component analysis (PCA)) and model with PCA were developed with MLP and implemented using TensorFlow. The result shows that the model with the PCA performed better with an accuracy of 98.34% and a running time of 1.88s compared with the initial model which gives an accuracy of 97.48% and running time of 2.3s. Hence, the ability to better differentiate between edible and poisonous mushrooms using the PCA model will save more lives.

.

Keywords: Deep learning, Multilayer perceptron (MLP), Classification and Mushroom

1. Introduction

Deep learning is a branch of machine learning technique that is based on a set of algorithms that attempt to model high-level abstractions in data by using model architectures, with complex structures or otherwise, composed of multiple non-linear transformations. (Deng & Yu, 2014 ; Yoshua, 2009 ; Bengio, Courville & Vincent, 2013 ; Schmidhuber, 2014). It can automatically learn features from a large number of image data set (Zhu et al., 2018). Deep learning is based on learning representations of data. An image, for instance, can be represented in many ways such as a vector of intensity values per pixel, or in a more abstract way as a set of edges and regions of particular shape among others. Deep learning has helped replaced handcrafted features with efficient algorithms for unsupervised or semi-supervised feature learning. (Song & Lee, 2013).

Various deep learning architectures have been applied to fields like automatic speech recognition and natural language processing where they have produced state-of-the-art results on different tasks. (The Wikipedia Guide, 2019). Layers that have been used in deep learning include hidden layers of an artificial neural network and sets of complicated propositional formulas (Yoshua, 2009).

Multi-Layer Perceptron (MLP) is a widely used Feed Forward neural network (FFNN) that is composed of an input layer that receives the signal, an output layer that makes a decision and in between those two, an arbitrary number of hidden layers that are the true computational engine of the MLP (Azami & Escudero, 2015 ; Hemalatha & Ran, 2017). Multilayer perceptron are often applied to supervised learning problems. They train on a set of input-output pairs and learn to model the correlation between those inputs and outputs. Training involves adjusting the parameters or the weights of the model in order to minimize error. Back propagation is used to make those weights adjustments relative to the error. The error itself can be measured by root mean squared error (RMSE).

Feature selection, also known as variable selection and attribute selection, can be used in data pre-processing. (Srinidhi, 2018). This

helps in finding accurate data models (Jovi, Brki & Bogunovi, 2015). It is effective and efficient in preparing high-dimensional data for data mining and machine learning problems. Feature selection helps to build simpler and more comprehensible models and improve data mining performance. (Li et al., 2017). It speeds up the learning task and also eliminate redundant and irrelevant features. In practice, feature selection is a frequently used technique, it is especially suitable in domains where there are many features and comparatively few samples (Wang, 2015).

Mushroom can be classified into several categories based on edibility. Mushroom is one type of fungus type plant containing no chlorophyll. There is around 45000 type fungus available in the world. Among them, around 2000 fungus are edible vegetable food. Mushroom can be edible and non-edible. Cultivating mushroom in scientific ways reduces the probability to occur poison in mushroom yield (Chowdhury & Ojha, 2017). Hence, for easy recognition of edible mushrooms, this study develops a deep learning model that helps to accurately classify edible and poisonous mushrooms using multi-layer perceptron neural network.

2. Related Works

Development of a Mushroom Growth Measurement System Applying Deep Learning for Image Recognition: This work proposed a smart mushroom measurement system using image processing technology in deep learning. The system automatically measures and records the size of the mushroom cap and the growth rate during the fruiting body formation. The harvest time can be estimated by the obtained data. The study limitation is such that growth of mushrooms is difficult to quantify clearly. Also the relationship between the growth status of mushrooms and environmental conditions, such as temperature, humidity, and carbon dioxide is not known (Lu et al., 2019).

Mushroom Classification Using ANN and Adaptive Network-based Fuzzy Inference System (ANFIS) Algorithm : This work described classification of Mushroom dataset, which was split for training and testing purposes. Different sizes of training data were used to check the performance of classifier. The performance of different classification algorithm such as ANN, ANFIS, and Bayes Net classifier were compared on the basis of Mean absolute error, Accuracy, Kappa statistic for mushroom datasets. Prescription of the research is that size of training set as well as selection of classification technique depends on the data to be analyzed, therefore performance of all the techniques is low when dataset size is small (Verma & Dutta 2018).

Development of Mushroom Expert System Based on Support Vector Machine (SVM) Classifier and Naive Bayes Classifier: In this work, SVM and Naïve Bayes algorithms were used for classification of mushrooms. Mushroom expert system was developed for classification of mushrooms and to predict the class of mushrooms on submission of characteristics of the mushrooms. This system is a web based application for online users. The study shows that SVM Classifier gives better accuracy when compared to Naïve Bayes classifier on mushroom data (Babu et al., 2014).

Deep Shrooms - Classifying Mushroom Images: This study used a machine learning classifier based on a convolutional neural network to create a smartphone application which classifies user inputted pictures of wild mushrooms to edible or not edible. The application only provides the user with the probability of having an edible mushroom and not clear prediction of the edibility, leaving the user to do more research on the edibility of the mushroom (Koivisto, Nieminen & Harjunpää, 2017).

3. Methodology

This section discusses the method used in this study. Figure 1 shows the overall methodology. The mushroom dataset used in this study was from the UCI Machine Learning repository. It consists of descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family. A three-layer

(input layer, hidden layer and output layer) neural network is adopted to solve this classification problem. The number of neurons in the input layer equals the number of attributes each instance of mushroom has, which is 22 in the instance mushroom dataset.

3.1. Data cleaning

This removes non-related data, missing value or treatments to any other abnormalities.

3.2. Data encoding

This transforms categorical data into numerical data for further training and classification. Originally, the dataset contains 21 valuable features that are all categorical data, which is not suitable to directly fit some machine learning models. One-Hot-Encoder was adopted in this study since it has the advantage of giving binary result rather than ordinal. After One-Hot-Encoder, the clean dataset was transformed into a new one with 116 columns of dummy features. Principal Component Analysis (PCA) was used for further analysis. Among all the 116 expanded features, only 50 are possible defining features.

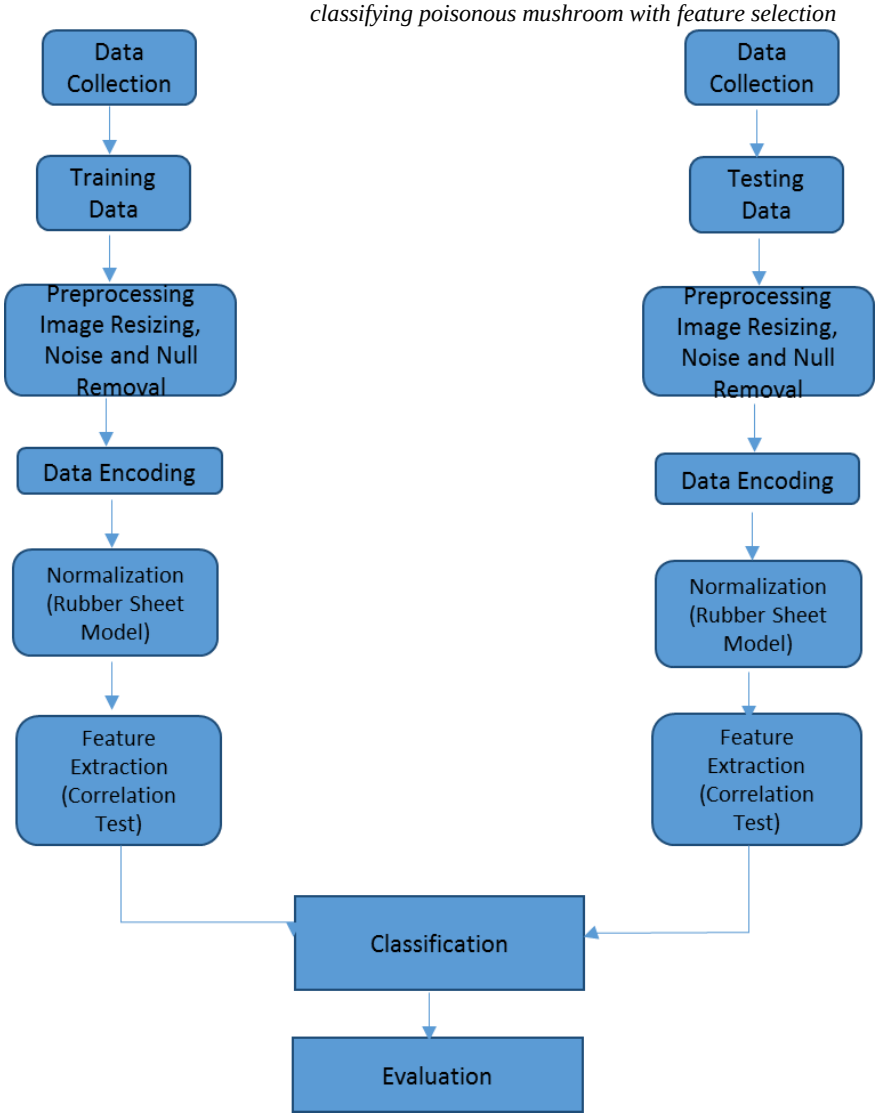

Figure 1: The Overall Methodology of Mushroom Classification

Figure 2 shows the Multi-layer perception used in identifying the mushrooms.

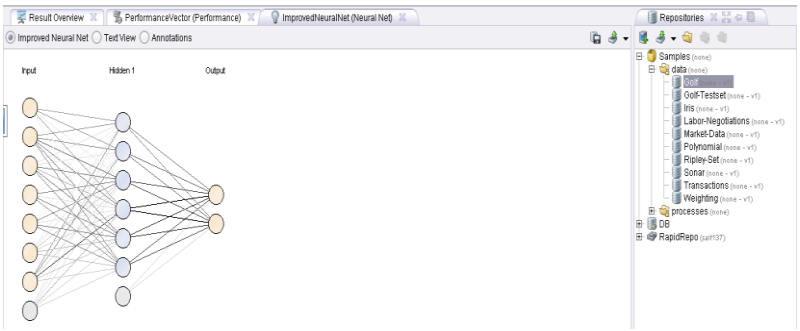

Figure 2: Multi-layer perceptron used in identifying mushroom

3.3. The Metrics Used

The three evaluation metrics used for regression problems are:

(i). Mean Absolute Error (MAE): This is the mean of the absolute value of the errors. It is a measure of difference between two continuous variables. This is shown in equation (1).

$$\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad \ldots\ldots \quad (1)$$

(ii). Mean Squared Error (MSE): This is the mean of the squared errors. It it is the average squared difference between the estimated values and what is estimated. This is shown in equation (2).

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad \ldots\ldots \quad (2)$$

(iii).Root Mean Squared Error (RMSE): This is the square root of the mean of the squared errors. It is a measure of the differences between values predicted by a model or an estimator and the values observed. This is shown in equation (3).

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad \ldots\ldots \quad (3)$$

All of these are loss functions because the goal is to minimize them.

4. Implementation and Discussion of Results

The results of the three different models developed with MLP are discussed in this section. TensorFlow was used for the implementation. Figure 3 shows a balance check between the edible and the poisonous classes.

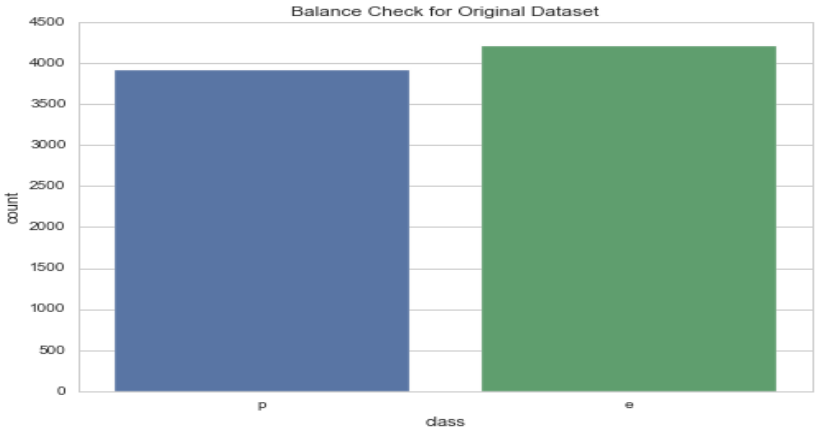

Figure 3: Balance check between edible (e) and poisonous (p) classes

4.1. Distribution of Selected Features

Figure 4 shows the distribution for some features. Some features draw special attention for unbalanced distribution of certain attributes, indicating differences between edible and poisonous mushrooms. For the "Odor", edible mushrooms mostly have no special smell, while poisonous ones have foul smell. For the "spore-print-color", edible ones are significant in "black", but the "poisonous" ones are with high quantity of "buff" and "chocolate".

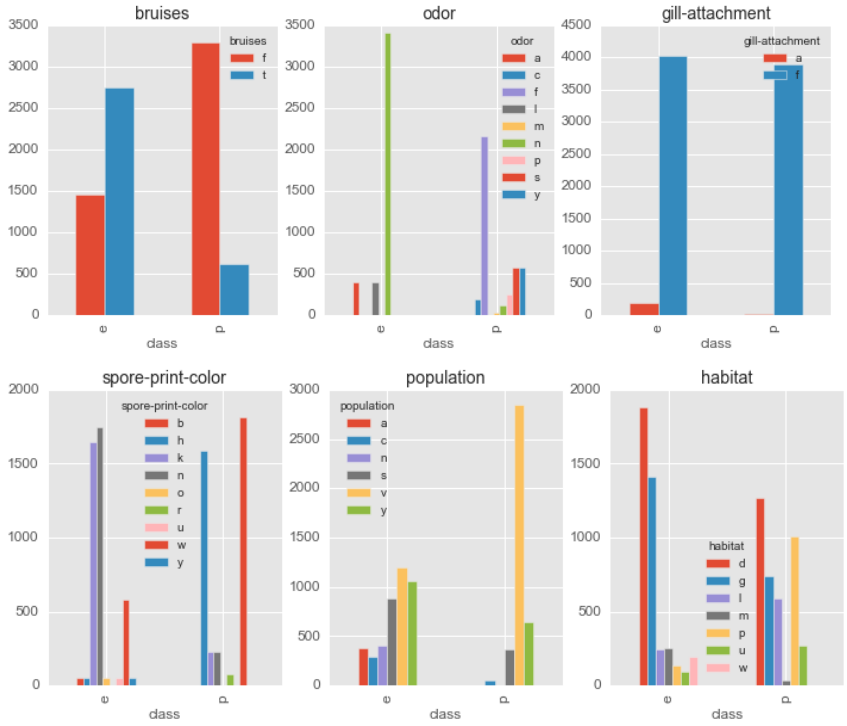

Figure 4: Distribution of selected features

4.2. Feature Importance

The importance of the features, based on PCA is about how many features are important in the expanded dataset. Xgboost was utilized to draw the feature importance chart for original dataset with 21 features. This is shown in Figure 5. The chart indicates that “odor” and “spore-print-color” color are the top two important features when determining if a mushroom is edible or poisonous.

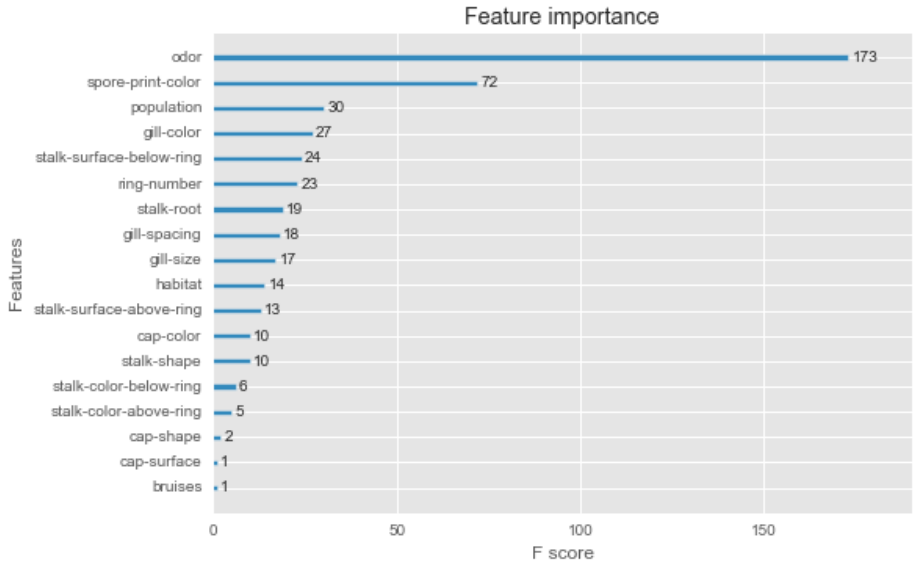

Figure 5: Feature Importance of the Mushroom Dataset

4.3. Accuracy Result

Figure 6 shows the accuracy of both the initial model (without PCA) and that of PCA with multiple components. The initial model has an accuracy of 97.48% while PCA model has 98.34%.

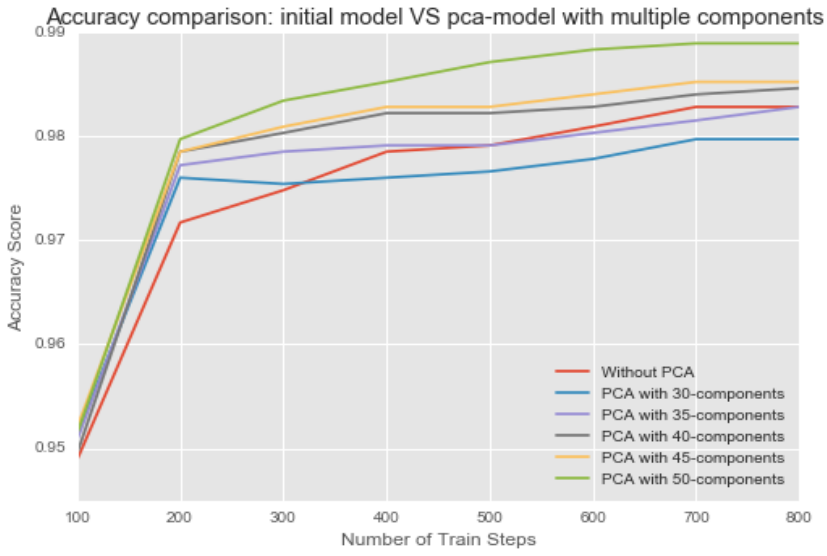

Figure 6: Accuracy Result of the Models

4.4. Running Time Result

Figure 7 shows the running time of both the initial model (without PCA) and that of PCA with multiple components. The initial model has a running time of 2.3s while PCA model has 1.88s.

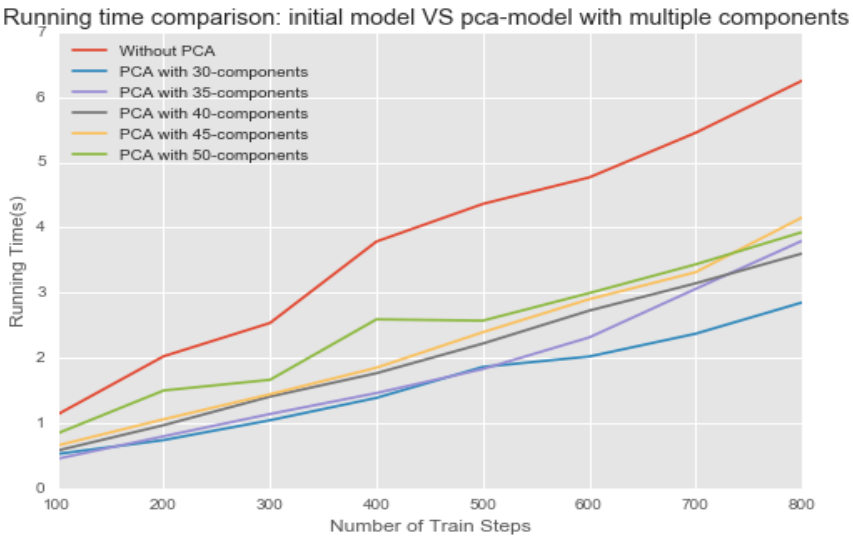

Figure 7: Running Time Result of the Models

Table 1 shows the summary of the overall results of the two models.

Table 1: Overall Result of the Models

	Initial Model (Without PCA)	Model With PCA
Accuracy	97.48%	98.34%
Running Time	2.3s	1.88s

5. Conclusion

The result shows that the model with the PCA performed better with an accuracy of 98.34% and a running time of 1.88s compared with the initial model which gives an accuracy of

97.48% and running time of 2.3s. Hence, PCA model is better in classifying edible and poisonous mushrooms.

List of References

Azami, H. & Escudero, J. (2015) A Comparative Study of Breast Cancer Diagnosis based on Neural Network Ensemble via Improved Training Algorithms, 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Milan, Italy, 26/08/15 - 29/08/15, pp. 2836-2839, 2015

Babu, P., Thommandru,R., Swapna, K. & Nilima, E. (2014). Development of Mushroom Expert System Based on SVM Classifier and Naive Bayes Classifier. International Journal of Computer Science and Mobile Computing - IJCSMC (Vol.3, No. 4)

Bengio, Y, Courville, A. and Vincent.,P.(2013). "Representation Learning: A Review and New Perspectives," IEEE Trans. PAMI, special issue Learning Deep Architectures, 2013

Chowdhury & Ojha, (2017). An Empirical Study on Mushroom Disease Diagnosis: A Data Mining Approach. International Research Journal of Engineering and Technology (IRJET). Vol 4. Issue 1. January, 2017.

Deng, L. & Yu D. (2014) "Deep Learning: Methods and Applications" http://research.microsoft.com/pubs/209355/DeepLearning-NowPublishing-Vol7- SIG-039.pdf

Hemalatha K. and Ran K. U. (2017).Advancements in Multi-Layer Perceptron Training to Improve Classification Accuracy. International Journal on Recent and Innovation Trends in Computing and Communication. ISSN: 2321 8169. Volume: 5 Issue: 6

Jovi, A., Brki, K., & Bogunovi, N. (2015). A review of feature selection methods with applications. Pp 25 -29 (2015)

Koivisto, T., Nieminen, T., & Harjunpää, J. (2017). Deep Shrooms : Classifying Mushroom Images.

Li, J., Cheng, K., Wang, S., Morstatter F., Trevino, R. P., Tang, J. and Liu, H. (2017). Machine Learning. ACM Computing Surveys (CSUR), 50(6): 94:1-94:45, 2017.

Lu, C., Liaw, J., Wu, T. & Hung, T. (2019). Development of a Mushroom Growth Measurement System Applying Deep Learning for Image Recognition. Agronomy 2019, 9, 32; doi:10.3390/agronomy9010032. www.mdpi.com/journal/agronomy

Schmidhuber, J. (2014). "Deep Learning in Neural Networks: An Overview" http://arxiv.org/abs/1404.7828, 2014

Song, Ah H., and Lee, S.(2013). "Hierarchical Representation Using NMF." Neural Information Processing. Springer Berlin Heidelberg, 2013.

Srinidhi, S. (2018). What is Feature Selection and why do we need it in Machine Learning? A Blog post from https://blog.contactsunny.com/data-science/what-is-feature-selection-and-why-do-we-need- it-in-machine-learning

The, Wikipedia Guide (2019). Introduction to Machine Learning. https://en.wikipedia.org/wiki/Book:Machine_Learning_–_The_Complete_Guide

Verma, S. K., & Dutta, M. (2018). Mushroom Classification Using ANN and ANFIS Algorithm, *8*(1), 94–100.

Wang, S. (2015). Feature Fusion , Feature Selection and Local N-ary Patterns for Object Recognition and Image Classification. Ph.D. Thesis. University of Technology, Sydney. January, 2015.

Yoshua, B.(2009). "Learning Deep Architectures for AI" (PDF). Foundations and Trends in Machine Learning **2** (1).

Zhu, L., Li, Z., Li, C., Wu, J., & Yue, J. (2018). High performance vegetable classification from images based on AlexNet deep learning model, 11(4), 217–223. https://doi.org/10.25165/j.ijabe.20181104.2690