# Algorithm Comparison for Cultural Heritage Image Classification

Radmila Janković

Mathematical Institute of the Serbian Academy of Sciences and Arts
Belgrade, Serbia
rjankovic@mi.sanu.ac.rs

## Abstract

Digitization represents an important part of the development of online systems. As such it includes, among other, the deployment, categorization and preservation of audio, video and textual contents online. Such process is especially interesting from the perspective of cultural heritage, as it allows the long-term preservation and sharing of culture worldwide. This study observes four classification algorithms: (i) the multilayer perceptron, (ii) averaged one dependence estimators, (iii) forest by penalizing attributes, and (iv) the k-nearest neighbor rough sets and analogy based reasoning, before and after attribute classification, and compares these with the results obtained from the convolutional neural network. The obtained results show that the best classification performance was achieved by the multilayer perceptron, followed by the convolutional neural network.

## 1 Introduction

With an increased use of digitization in the domain of cultural heritage, it is possible to preserve and promote the cultural heritage present in every part of the world. Every country worldwide has its own practices, places, values, objects, and arts that are created throughout history, and that represent its cultural heritage. Through cultural heritage knowledge is being shared and passed on from generation to generation. Some examples of cultural heritage include photographs, historical monuments, various types of documents, archaeological sites, and other.

There are three pillars of digital cultural heritage: (i) digitization focusing on conversion of objects into digital form, (ii) access to digital heritage, and (iii) long term preservation of digital objects [IDST12]. Classification is an important part of digitization as it includes building a classification model that groups new inputs into categories, based on the previously available set of data. In terms of cultural heritage, classification is particularly important because it allows the preservation of heritage for future generations. Furthermore, digitization enables promotion of cultural heritage by using innovative technologies to increase accessibility. Through digitization, a country's cultural heritage is being promoted globally, thus contributing to cultural diversity.

Different methods for cultural heritage image classification have been investigated by various authors. Deep learning algorithms were used for image classification in [LlMMZG17]. In particular, AlexNet and Inception V3

In: A. Amelio, G. Borgefors, A. Hast (eds.): Proceedings of the $2^{nd}$ International Workshop on Visual Pattern Extraction and Recognition for Cultural Heritage Understanding, Bari, Italy, 29-Jan-2020, published at http://ceur-ws.org

convolutional neural networks (CNNs) were used, as well as ResNet and Inception-ResNet-v2 residual networks. The dataset included architectural cultural heritage divided into 10 categories. The results showed that deep learning methods perform better than other state-of-the-art methods, particularly when dealing with complex problems [LlMMZG17]. The performance of deep learning methods has also been investigated in [KHP18], where CNNs were used to classify images, audio and video data, while Recurrent Neural Networks (RNNs) were used to classify the text belonging to the cultural heritage of Indonesia [KHP18]. It was observed that the RNN achieved the highest accuracy, while the CNN obtained good accuracy for image and video classification (76%)[KHP18]. Considering other methods, k-nearest neighbor (kNN) classification was used to classify cultural heritage images of 12 monuments and landmarks in Pisa [AFG15], but also to classify and detect alterations on historical buildings with a high accuracy (92%) [MPAL15]. Various decision tree algorithms including J48, random tree, random forest and fast random forest were investigated in [GDPR18]. Classification was performed in WEKA on a set consisting of 3D cultural heritage models, and the results showed that the fast random forest achieves the highest accuracy of 69% [GDPR18]. Different types of image classification techniques were investigated in [AJ19]. In particular, naive Bayes and Support Vector Machine (SVM) algorithms are widely used for tangible and movable cultural heritage, while the intangible cultural heritage is mostly classified using SVM, kNN, CNN, decision trees and Conditional Random Fields - Gaussian Mixture Model (CRF-GMM) algorithms [AJ19]. Tangible and immovable cultural heritage is most commonly classified using CNNs [AJ19].

The aim of this paper is to compare the performance of several classification algorithms for cultural heritage image classification, in particular: (i) the multilayer perceptron (MLP), (ii) averaged one dependence estimators (AODE), (iii) forest by penalizing attributes (Forest PA), and (iv) the k-nearest neighbor rough sets and analogy based reasoning (RSeslibKnn). The performance was observed on a full set of attributes as well as on the reduced set of attributes, and the results were compared. Furthermore, a CNN was also developed for comparison purposes, as deep learning represents a state-of-the-art technique [LlMMZG17, KHP18] and it is interesting to observe and compare its performance with the performance of the other algorithms used in this study.

This paper is organized as follows. Section 2 explains the data and methodology used in this research, while Section 3 presents the results and discussion. Finally, Section 4 contains concluding remarks.

## 2 Data and methodology

### 2.1 Data

The cultural heritage image classification was performed on a public dataset created by [LlMMZG17] and obtained from Datahub (`https://old.datahub.io/dataset/architectural-heritage-elements-image-dataset`). The dataset consists of 10,235 images of size 128 × 128 pixels, but for the purpose of this study, 4,000 images from 5 out of 10 categories were randomly chosen. These images include altars, gargoyles, domes, columns, and vaults (Figure 1).

### 2.2 Methodology

The experiments were performed in WEKA (Waikato Environment for Knowledge Analysis) [WFHP16], a free data mining software based on Java, while the CNN model was developed in Python v.3.7 with the use of the Keras library. The experiments were performed on a Windows machine with a 2.3 GHz processor and 8 GB of RAM.

In Weka, the feature extraction is performed using feature extraction algorithms integrated inside the *imageFilters* package. Three types of filters were applied on the dataset: (i) the edge histogram, (ii) the color layout, and (iii) the JPEG coefficients. Feature extraction for the CNN was not performed manually, as Python automatically scans and extracts features from the dataset.

The edge histogram extracts the MPEG7 edge features from the images. In particular, it detects the directions of edges in images based on the changes in frequency and brightness [WPP02]. There are five types of edges: vertical, horizontal, 45-degree diagonal, 135-degree diagonal, and non-directional [WPP02]. The color layout feature extraction is performed by dividing the image into 64 blocks and calculating the average color for each block, using the *color layout* filter in WEKA. Finally, the JPEG coefficients were extracted by splitting the image based on different frequencies, hence keeping only the most important frequencies [MD13].

After feature extraction, the dataset consisted of 307 attributes. The first results were generated on the full set of attributes, while the second results were generated on the attribute-reduced set of data in order to evaluate the change in performance before and after attribute selection. Feature selection was performed in WEKA using

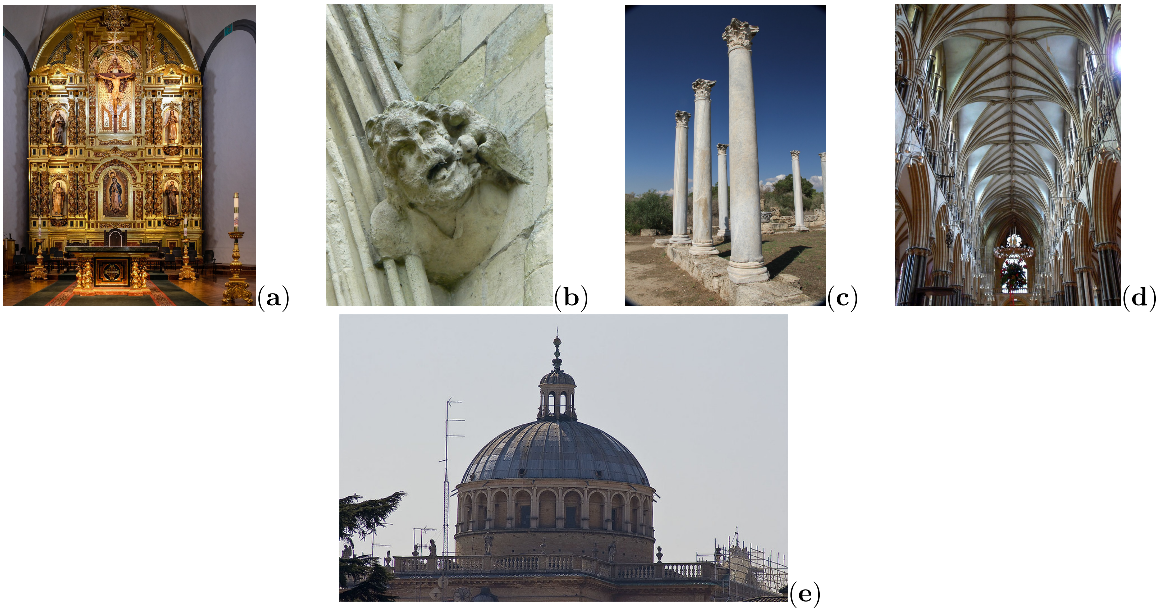


Figure 1: Example of images used in this study for each of the five classes, in particular: (a) altar, (b) gargoyle, (c) column, (d) vault, and (e) dome.

the *AttributeSelection* filter. The attribute search was performed using the best-first method, and attribute evaluation using the CFS subset evaluator. After feature selection, the number of attributes was reduced to 89. The dataset was divided into 70% of images for training and 30% for testing the algorithms. Four algorithms were tested and compared for the purpose of image classification: (i) MLP, (ii) Forest PA, (iii) AODE, and (iv) RSeslibKnn. Additionally, in order to compare the performance of these algorithms with the state-of-the-art techniques, a deep learning CNN model was developed and tested.

The MLP network consists of three layers: an input layer, a hidden layer, and an output layer. The layers consist of neurons, where the neurons in one layer are connected to the neurons in the next layer [PM19]. The MLP uses a nonlinear activation function, making it suitable for different types of problems, without any assumptions regarding data distribution [GD98]. The AODE is a classification technique that calculates the probability of each class and creates a set of one dependence probability distribution estimators [WBW05]. Is holds less rigorous independence assumptions than naive Bayes, hence it is suitable for various range of problems. Forest PA is a relatively new decision forest algorithm developed in 2017 by [AI17] in order to overcome the limitations of the random forest algorithm. The algorithm works in such a way that it uses the full set of attributes for forest creation, but it also assigns weights to those attributes that already participated in the previous decision tree. It generates the bootstrap sample from the training set and creates a decision tree using the attribute weights [AI17]. RSeslibKnn is the k-nearest neighbor classifier that uses a fast neighbor search, thus making it appropriate for using on large datasets [WL19]. A distance measure is calculated based on the weighted sum of distances, which results in creation of the indexing tree [WL19]. The classification is performed by finding the k nearest neighbors in the set. The CNN is a neural network consisting of standard type of layers, as well as of convolution and pooling layers. It is widely used for image processing, as it is able to automatically scan and extract features from images. In particular, each CNN is composed of convolutional layers followed by the pooling layers that reduce the dimensionality of the data. The features extracted through these layers are then transformed into a vector using the flattening layer, and the obtained vector is further distributed to a dense layer, thus forming a fully-connected network.

### *Parameter configuration*

The MLP consisted of one hidden layer with 50 neurons, a learning rate of 0.5, a momentum of 0.6, and a batch size of 32. The sigmoid activation function was utilized in the hidden and output layers. All attributes were standardized. The RseslibKnn included the the city and simple value difference as a distance measure, the inverse square distance as the voting method, and the distance based weighting method. The AODE was configured with the default WEKA parameters. Lastly, the Forest PA included 30 trees in the forest.

The CNN configuration involved four convolution layers with 32 neurons in the first two layers, 64 neurons in the last two convolution layers, one hidden layer with 128 neurons, and an output layer with 5 neurons. The convolutional and hidden layers used the hyperbolic tangent activation function, while the output layer used the softmax activation. The kernel size of the convolutional layers was set to 3 × 3, while the pooling size in the pooling layer was set to 2 × 2. The dropout was set to 0.2. Lastly, in order to avoid overfitting, the early stopping regularization parameter was used with patience set to 3. The number of epochs was set to 50, but the training stopped after 18 epochs because there was no further improvement in the accuracy. The model used 80% of data for training and the rest of the data for validation.

*Evaluation Metrics*

The classification performance can be evaluated using several metrics. In particular, this study observed and analyzed the values of correctly classified instances, precision, recall, F-score, kappa statistics, and ROC area. The percentage of correctly classified instances represents the instances that are correctly classified by the algorithm, while precision shows the fraction of instances that belong to the observed class among the total number of instances that are classified into the observed class by the algorithm. Recall represents the true positive rate of prediction, while the F-measure shows the classification accuracy based on the average value of precision and recall. The F-measure values should ideally be closer to 1 indicating a better classification accuracy. Kappa is the measure of agreement and can have values in an interval of 0–1, with the values in the range 0.81–1 representing an almost perfect agreement, while values close to 0 represent poor agreement [SW05]. Lastly, the ROC area represents the ratio of the true positives and the false positives, and its value should be close to 1, indicating a perfect prediction [FUW06].

## 3 Results

All algorithms used in this study were first tested on the full dataset consisting of 307 attributes. The best performing algorithm in this case is the MLP with 85% of accuracy obtained, followed by the RSeslibKnn, AODE, and Forest PA with 82%, 79% and 78%, respectively (Table 1).

Table 1: Algorithm performance before attribute selection

| Algorithm | MLP | Forest PA | AODE | RSeslibKnn |
|---|---|---|---|---|
| Correctly classified instances | 84.83% | 77.92% | 79.25% | 82.17% |
| Kappa statistics | 0.810 | 0.724 | 0.741 | 0.777 |
| Precision | 0.849 | 0.778 | 0.793 | 0.824 |
| Recall | 0.848 | 0.779 | 0.793 | 0.822 |
| F-measure | 0.848 | 0.778 | 0.791 | 0.820 |
| ROC Area | 0.974 | 0.954 | 0.959 | 0.888 |
| Running time (in seconds) | 1038.09 | 54.02 | 122.72 | 93.14 |

The MLP also obtained the best values in terms of the kappa statistics, precision, recall, and the F-measure, comparing to other three algorithms (Table 1). In terms of the running time, the fastest algorithm is the Forest PA with 54.02 seconds.

After observing the results obtained from the full set of data, the next step involved observing the performance of the algorithms on the reduced set of attributes. The MLP algorithm again performed the best, correctly classifying 98.9% of instances (Table 2). Other algorithms obtained lower classification accuracy, in particular 80.83%, 80.67% and 78.67% for the AODE, RSeslibKnn and Forest PA, respectively. Observing other performance measures, the MLP also obtained the highest value of kappa statistics (0.986), followed by AODE (0.760), RSeslibKnn (0.758), and Forest PA (0.733). As described before in this paper, these values indicate a substantial to almost perfect agreement [SW05]. Moreover, the MLP obtained the highest value of the F-measure (0.986), followed by RSeslibKnn (0.811), AODE (0.807), and Forest PA (0.797). Lastly, observing the value of the ROC area, the results indicate the MLP algorithm performed the best with an obtained value of 0.996, followed by AODE, Forest PA and RSeslibKnn with ROC area values of 0.965, 0.959 and 0.879, respectively. These results show a good classification power of the observed algorithms, with MLP and AODE performing the best (Table

2), but it should be noted that the MLP algorithm requires much longer running time than the other three algorithms.

Table 2: Algorithm performance after attribute selection

| Algorithm | MLP | Forest PA | AODE | RSeslibKnn |
|---|---|---|---|---|
| Correctly classified instances | 98.9% | 78.67% | 80.83% | 80.67% |
| Kappa statistics | 0.986 | 0.733 | 0.760 | 0.758 |
| Precision | 0.989 | 0.787 | 0.808 | 0.811 |
| Recall | 0.989 | 0.787 | 0.808 | 0.807 |
| F-measure | 0.986 | 0.797 | 0.807 | 0.805 |
| ROC Area | 0.996 | 0.959 | 0.965 | 0.879 |
| Running time (in seconds) | 892.02 | 62.01 | 115.17 | 109.27 |

In order to gain more insights about the power of the observed classification algorithms, classification matrices are generated (Table 3). The classification matrix shows the number of correctly (and incorrectly) classified instances by class, where the numbers in the diagonal represent accurate classifications. Before attribute selection, the algorithms mostly miss-classified images of gargoyle, column and vault. In particular, the MLP most accurately classified the dome images, while the highest number of miss-classifications for the MLP algorithm is observed for the vault and column images. The AODE most accurately classified altar images, while Forest PA most correctly classified the dome images. The RSeslibKnn classified altar images most accurately, while the highest number of wrongly classified instances is observed mainly for the column images (Table 3).

Table 3: The confusion matrices for each algorithm, before attribute selection

| Algorithm | Altar | Column | Dome | Gargoyle | Vault | Classified as |
|---|---|---|---|---|---|---|
| MLP | 216 | 6 | 2 | 0 | 19 | altar |
| | 4 | 193 | 9 | 18 | 14 | column |
| | 3 | 16 | 220 | 13 | 1 | dome |
| | 0 | 8 | 15 | 196 | 6 | gargoyle |
| | 20 | 12 | 3 | 13 | 193 | vault |
| AODE | 217 | 9 | 2 | 1 | 14 | altar |
| | 13 | 163 | 27 | 19 | 16 | column |
| | 2 | 13 | 213 | 21 | 4 | dome |
| | 0 | 7 | 25 | 174 | 19 | gargoyle |
| | 32 | 7 | 1 | 17 | 184 | vault |
| Forest PA | 206 | 14 | 1 | 2 | 20 | altar |
| | 7 | 167 | 27 | 21 | 16 | column |
| | 2 | 17 | 215 | 14 | 5 | dome |
| | 1 | 12 | 26 | 170 | 16 | gargoyle |
| | 29 | 13 | 0 | 22 | 177 | vault |
| RSeslibKnn | 231 | 4 | 2 | 0 | 6 | altar |
| | 19 | 169 | 17 | 15 | 18 | column |
| | 8 | 13 | 224 | 8 | 0 | dome |
| | 3 | 12 | 29 | 174 | 7 | gargoyle |
| | 33 | 6 | 2 | 12 | 188 | vault |

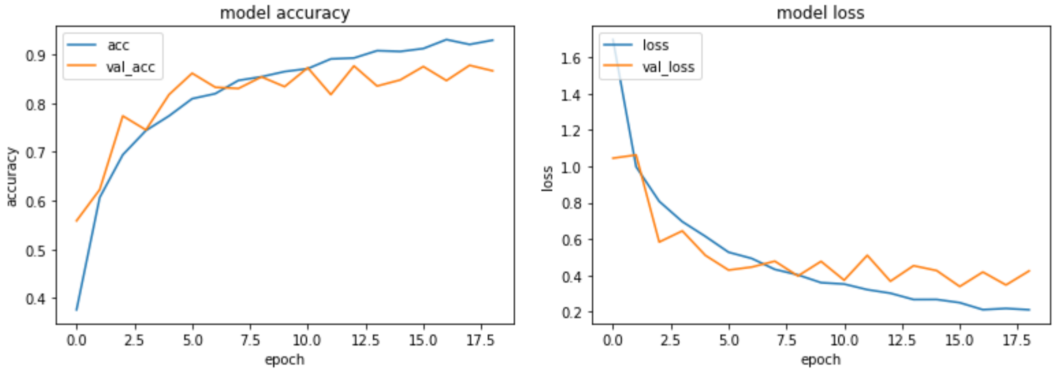
After attribute selection has been applied, the new confusion matrix has been obtained (Table 4). In terms of miss-classifications, the MLP and AODE algorithms mostly miss-classified column images, the RSeslibKnn mostly miss-classified vault images, while Forest PA mostly miss-classified the images of vaults and columns. In terms of accurate classifications, the MLP accurately classified almost all the images, while the AODE, Forest PA and RSeslibKnn most accurately classified the images of altar and dome (Table 4).

The previously described results were compared to the results obtained by using a deep learning algorithm, as it represents the state-of-the-art methodology. For this purpose, the CNN model was developed in Python and the

Table 4: The confusion matrices for each algorithm, after attribute selection

| Algorithm | Altar | Column | Dome | Gargoyle | Vault | Classified as |
|---|---|---|---|---|---|---|
| MLP | 239 | 0 | 0 | 0 | 1 | altar |
| | 1 | 237 | 0 | 0 | 2 | column |
| | 0 | 1 | 238 | 1 | 0 | dome |
| | 0 | 0 | 0 | 238 | 2 | gargoyle |
| | 0 | 2 | 2 | 1 | 235 | vault |
| AODE | 217 | 6 | 3 | 0 | 17 | altar |
| | 10 | 168 | 27 | 18 | 15 | column |
| | 3 | 13 | 217 | 18 | 2 | dome |
| | 0 | 8 | 23 | 179 | 15 | gargoyle |
| | 28 | 11 | 0 | 13 | 189 | vault |
| Forest PA | 201 | 18 | 1 | 3 | 20 | altar |
| | 6 | 177 | 22 | 22 | 11 | column |
| | 1 | 18 | 210 | 22 | 2 | dome |
| | 2 | 13 | 26 | 176 | 8 | gargoyle |
| | 30 | 12 | 0 | 19 | 180 | vault |
| RSeslibKnn | 226 | 1 | 5 | 2 | 9 | altar |
| | 19 | 170 | 22 | 15 | 12 | column |
| | 6 | 13 | 223 | 10 | 1 | dome |
| | 4 | 10 | 29 | 175 | 7 | gargoyle |
| | 40 | 7 | 0 | 20 | 174 | vault |

loss and accuracy results through epochs are plotted and presented in Figure 2. The obtained results demonstrate good accuracy of 93%, with training loss of 0.21 and validation loss of 0.42. Such results are promising as they clearly demonstrate the potential of deep learning techniques. Furthermore, deep learning allows for detailed modifications, thus enhancing the possibility of their application to different types of problems.



(**a**) (**b**)

Figure 2: The (**a**) accuracy and (**b**) loss results for the CNN model.

## 4 Conclusions

The aim of this paper was to compare the performance of several classification algorithms before and after attribute selection. In particular, MLP, AODE, Forest PA, and RSeslibKnn algorithms were applied on the dataset consisting of cultural heritage images, and their performance was compared to the performance obtained by the deep learning algorithm (CNN). Several conclusions can be drawn from this study: (i) the MLP algorithm obtained the best performance both before and after attribute selection, (ii) attribute selection increases the classification accuracy for all algorithms except the RSeslibKnn, and (iii) the deep learning techniques such as the CNN, obtain higher accuracy without needing to reduce the number of attributes. Hence, as deep learning techniques do not require to manually extract features from images, they represent an appropriate method of image classification.

### Acknowledgements

This work was supported by the Serbian Ministry of Education, Science and Technological Development through Mathematical Institute of the Serbian Academy of Sciences and Arts.

## References

[IDST12] Ivanova, K., Dobreva, M., Stanchev, P. and Totkov, G. *Access to digital cultural heritage: Innovative applications of automated metadata generation.* Plovdiv University Publishing House "Paisii Hilendarski", 2012.

[SW05] Sim, J. and Wright, C.C. "The kappa statistic in reliability studies: use, interpretation, and sample size requirements." *Physical therapy* 85, no. 3 (2005): 257-268.

[LlMMZG17] Llamas, J., Lerones, M. P., Medina, R., Zalama, E. and Gómez-García-Bermejo, J. "Classification of architectural heritage images using deep learning techniques." *Applied Sciences* 7, no. 10 (2017): 992.

[KHP18] Kambau, R.A., Hasibuan, Z.A. and Pratama, M.O. "Classification for Multiformat Object of Cultural Heritage using Deep Learning." In *2018 Third International Conference on Informatics and Computing (ICIC)*, pp. 1-7. IEEE, 2018.

[AFG15] Amato, G., Falchi, F. and Gennaro, C. "Fast image classification for monument recognition." *Journal on Computing and Cultural Heritage (JOCCH)* 8, no. 4 (2015): 1-25.

[GDPR18] Grilli, E., Dininno, D., Petrucci, G. and Remondino, F. "From 2D to 3D supervised segmentation and classification for cultural heritage applications." In *ISPRS TC II Mid-term Symposium Towards Photogrammetry 2020*, vol. 42, no. 42, pp. 399-406. 2018.

[MPAL15] Meroño, J.E., Perea, A.J., Aguilera, M.J. and Laguna, A.M. "Recognition of materials and damage on historical buildings using digital image classification." *South African Journal of Science.* 111, no. 1-2 (2015): 01-09.

[AJ19] osovi, M., Amelio, A. and Junuz, E. Classification Methods in Cultural heritage. In *Proceedings of the 1st International Workshop on Visual Pattern Extraction and Recognition for Cultural Heritage Understanding co-located with 15th Italian Research Conference on Digital Libraries (IRCDL*, pp. 13-24, 2019.

[WPP02] Won, C.S., Park, D.K. and Park, S.J. Efficient Use of MPEG-7 Edge Histogram Descriptor. *ETRI journal* 24, no. 1 (2002): 23-30.

[MD13] More, N.K. and Dubey, S. JPEG Picture Compression Using Discrete Cosine Transform. *International Journal of Science and Research (IJSR)* 2, no. 1 (2013): 134-138.

[WFHP16] Witten, I.H., Frank, E., Hall, M.A. and Pal, C.J. *Data Mining: Practical machine learning tools and techniques.* Morgan Kaufmann, 2016.

[PM19] Pal, S.K. and Mitra, S. "Multilayer perceptron, fuzzy sets, and classification." *IEEE Transactions on neural networks* 3 (1992): 683–697.

[GD98] Gardner, M.W. and Dorling, S. "Artificial neural networks (the multilayer perceptron)a review of applications in the atmospheric sciences." *Atmospheric environment* 32, no. 14-15 (1998): 2627-2636.

[WBW05] Webb, G.I., Boughton, J.R. and Wang, Z. "Not so naive Bayes: aggregating one-dependence estimators." *Machine learning* 58, no. 1 (2005): 5-24.

[AI17] Adnan, M.N. and Islam, M.Z. "Forest PA: Constructing a decision forest by penalizing attributes used in previous trees." *Expert Systems with Applications* 89 (2017): 389-403.

[WL19] Wojna, A. and Latkowski, R. Rseslib 3: Library of rough set and machine learning methods with extensible architecture. In *Transactions on Rough Sets XXI*, pp. 301–323. Springer, 2019.

[FUW06] Fan, J., Upadhye, S. and Worster, A. "Understanding receiver operating characteristic (ROC) curves." *Canadian Journal of Emergency Medicine* 8, no. 1 (2006): 19-20.