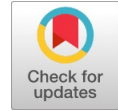# Classifying Flowers Images by using Different Classifiers in Orange

Vijaylakshmi Sajwan, Rakesh Ranjan

Abstract: This paper presents the first step towards looking for an advanced solution of image classification using distinct Classifiers in the Orange Data Mining Tool. The objective of the paper is to decide the ability of distinct classifiers for flowers image classification involving a small sample; Deep learning models are used to calculate a feature vector for every image of the Iris flower database. The used classifiers involved logistic regression, Neural Network, AdaBoost, Support Vector Machine, Random Forest and K-NN. The result indicates that the Logistic Regression, Neural Network, AdaBoost classifiers perform best in classifying a small sample of Iris flower images, and SVM and Random Forest classifiers perform less classification accuracy then above classifiers while K-NN performs worst with the lowest classification accuracy.

Index Terms: Logistic Regression, Neural Network, AdaBoost, SVM, Random Forest, K-NN, Supervised Machine Learning

I. INTRODUCTION

Machine learning is used to make a forecast and better understand the system. Ma- chine learning has a wide range of application such as Manufacturing, Telecommuni- cation, Retail, Finance, Medicine, and Bioinformatics [1]. Machine learning can be categorized into two way, supervised learning, unsupervised learning. Here we are considering the supervised learning methods. Supervised classification techniques are algorithms that 'learn' patterns in data to forecast an associated discrete class. They are flexible statistical prediction techniques collectively referred to as pattern in data to predict an associated discrete class [1]. Machine learning is defined as "program- ming computers to optimize an achievement criterion using example data or past oc- currence" [2]. Supervised machine learning techniques that have been trialed include Logistic Regression [7], Neural Network [4], AdaBoost [6], Random Forest [5], Sup- port Vector Machine (SVM) [3], and K- Nearest Neighbour (K-NN) [2]. The aim of this research is to evaluate how well a classification algorithm works when fed with image descriptor data from a small sample in the flower image context. I particular, deep learning is used in Orange software for feature extraction. Then, based on their value, Logistic Regression, Neural Network, AdaBoost, SVM, Random Forest, and K-NN algorithm are applied in order to predict the image classes.

Ms. Vijaylakshmi Sajwan, Assistant Professor in Uttaranchal Institute of Management,Uttaranchal University.
Prof.(Dr.) Rakesh Ranjan, Pro-Vice Chancellor in Himgiri Zee University.

II. RELATED WORK

In [8], the authors used deep learning technique to classify images of architectural heritage. Two convolutional neural networks were applied for images classification tasks: (1) AlexNet, and (2) Inception V3. Accuracy results obtained from this study were very good, with a mean value of over 0.93 [8]. The authors concluded that deep learning methods are very suitable for classifying heritage images. Moreover, the authors suggested using fine-tuning methods, if there are limited computational re- sources, or if the dataset is smaller [8]. Research by [9] investigated the efficiency of three data mining technique: (i) decision tree, (ii) Support Vector Machine, and (iii) maximum likelihood algorithm to asses land cover changes from remotely sensed data. In this study, it was noted that the decision tree algorithm performed better than the other two classification methods. In [10], a building façade classification based on an architectural style of the building was performed, considering only the windows. The approach was based on clustering and learning of local features [10]. A high classification rate was obtained, proving the efficiency of the proposed method. The authors in [11] applied an alternative dimension of CBIR in order to organize heritage images in two classes: one class involving human activities, and the other involving non-human activities such as images of antique objects. The authors also applied a Naïve Bayes algorithm in order to classify images based on the features from the edge histogram [11]. Good accuracy was obtained by this application of image classification, and the approach is considered suitable for the automatic classification of heritage images. The proposed approach uses Orange, an open source data mining and machine learning software, to classify a small sample of Iris flower dataset. This approach is based on the fact that image classification needs to be more simple and human independent, i.e. it needs less manual input and smaller sample size. This can be accomplished by using deep learning model of Orange for feature extraction and then applying classification algorithms on extracted features. In [15] the author used a unique system to build to reveal the information of the image based on its category and relevant labels associated with it. To accomplish this task different data mining techniques are used for image classifications using Bag of Visual Words (BoVW) feature extraction algorithm. This research investigated the efficiency of different data mining technique, that includes neural networks (NN), decision trees (DT), Bayesian networks (BN), discriminant analysis (DA) and K-nearest neighbour (KNN). In this study, it was noted that the BN and DA algorithm performed better than the other three classification method

The paper is organized as follows. In section 3 the database and methodologies used in this research are described. Then, in Section 4, the analysis is presented followed by a description of the result. Lastly, Section5 draws a conclusion from the study and proposed recommendations.

III. DATABASE AND METHODOLOGY

A. Database

We have used Iris flower data set in which 10 classes of flower images and each class have a different number of images, a total of 210 images in that database. A few images from the sample are shown in Figure
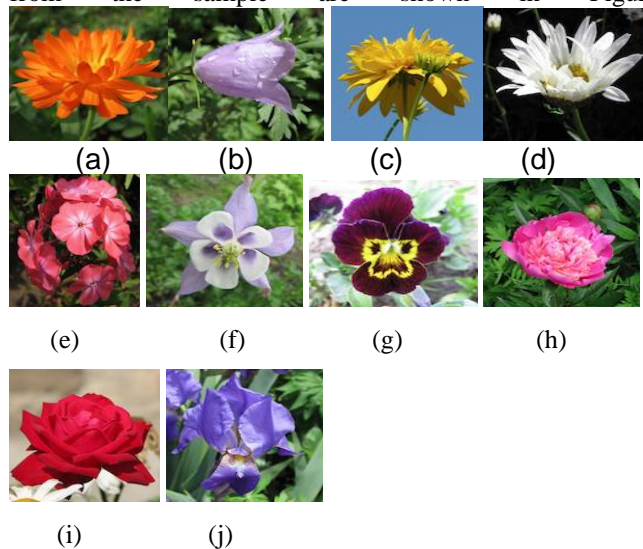

Figure 1: (a) Coltsfoot (b)crocus (c) Daindelion (d)Daisy (e) Linium (f) Ophyts (g) Pancy (h) Papaver (i) Rose (j) Satosa

B. Methodology

The main task of this paper is to classify flower images using Orange Data Mining tool, based on extracted features. The extraction of the features was made in Orange, using Deep learning model [14]. After the feature extraction, six classifiers were applied to the dataset in order to classify images based on their type. The applied classifiers were Logistic Regression, Neural Network, AdaBoost, Support Vector Machine (SVM), Random Forest and K-NN. 10-fold cross-validation was used for testing the model [14]. The results were then compared in order to evaluate which algorithm performs best for image classification problems with the small sample size.

C. Classification Methods

We have many supervised classification methods. Some of them, we following is a brief description of the classification method.

i. Support vector machine

Support Vector Machine (SVM) is a supervised machine learning technique which can be used for both distribution and regression challenges [3]. However, it is primarily used in distribution problems. In this technique, we plot each data item as a point in n-dimensional space (where n is the number of attributes) with the value of each attribute being a particular coordinate. Then, we perform classification by finding the hyperplane that discriminates the two classes very well [3].

ii. Neural Network

Neural networks are the most vital algorithm for artificial intelligence and machine learning. Multilayer perceptron (

MLP) has at least three layers: input, middle and output layer which is the type of Neural Network. Each node is a neuron that uses a nonlinear activation function, rather than an input node [4]. MLP are useful in re- search for their proficiency to short out the problems stochastically. It allows impre- cise solutions for tremendously complex problems like fitness estimation. In the 1980s, MLPs were very popular to finding applications in diverse fields such as im- age recognition, speech recognition, and machine translation software, but thereafter overlooked biggest challenges from much simpler SVM. Being the success of deep learning, interest in back-propagation network returned [4].

iii. Random forest

Random forest is the most pliable and easy to use supervised learning technique. A forest is composed of a tree. Forest is more vigorous if it has more tree [5]. A random forest constitutes decision trees on arbitrarily selected data samples, gets a forecast from each tree and selects the best result by means of voting. It also provides a perfect indicator of feature importance. The random forest has a range of applications, like recommendation engines, image classification, and attributes selection [5].

iv. K- Nearest Neighbor

Universal K-NN classification technique decides about the class of an image by searching for the k images of the training set most similar to the image to be classified and by performing a class weighted frequency analysis [2]. The k closest images are recognized relying upon a similarity scope among the images. The classification is based on a majority vote of the k-nearest neighbors, based on Euclidean distance in feature space, whereby k specifies the number of neighbors to be used. It does not require a training step to be performed but can be turned to determine the optimum value of k on which to base the classification. The training data were measured so that they have a standard deviation of 1 and a mean of 0. This was necessary so that each attribute was considered uniformly significant in the classification [5].

v. AdaBoost

AdaBoost or Adaptive Boosting is one of the boosting classifier presented by Yoav Freund and Robert Schapire in 1996 [6]. It combines numerous classifiers to hike the efficiency of classifiers. The basic concept behind AdaBoost is to set the weights of classifiers and training the data sample in each repetition such that it guarantees the exact forecasts of unusual observations [6].

vi. Logistic regression

In the field of social and natural sciences, Logistic regression is one of the most important analytic tools. It can be used to classify an observation into one of two classes (like 'positive sentiment' and 'negative sentiment'), or into one of many classes. Logistic regression classifies the data by considering outcome variables on extreme ends and tries makes a logarithmic line that distinguishes between them [7].

K-NN	0.987	0.829	0.828	0.847	0.829

IV. ANALYSIS AND RESULT

Performance & Measures

In order to calculate the classification specification, the following performance measures were adopted: (1) AUC, (2) CA (3) F1 score, (4) Precision, (5) Recall.

The **AUC (Area under the Curve)** is the area enclosed by the ROC curve. Completely random classifier has AUC = 0.5 and an ideal classifier has AUC = 1. The Scope of possible AUC value is [0, 1]. However, if our AUC is below 0.5, it means we can reverse all the outputs of our classifier and get a better score, so we did something wrong.

Classification Accuracy (CA) is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations [13]. Symmetric datasets where values of false positive and false negatives are almost the equal, accuracy is a great measure.

$$CA = TP + TN / TP + FP + FN + TN \qquad (1)$$

F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than classification accuracy, especially if we have an uneven class distribution. If the cost of false positives and false negatives are very different, it's better to look at both Precision and Recall.

$$F1 \, Score = 2 \times Recall \times Precision/Recall + Precision \qquad (2)$$

Precision is the ratio of correctly predicted positive observations of the total predict- ed positive observation. High precision relates to the low false positive rate [13].

$$Precision = TP / TP + FP$$

(3)

A recall is the ratio of correctly predicted positive observations to all observations in an actual class [13].

$$Recall = TP / TP + FN \qquad (4)$$

B. Obtained results

The results of the analysis are shown in Table 1. It is observed that the Logistic regression, Neural Network, AdaBoost classified accurately (CA) 0.990% of instances, followed by the SVM and Random Forest with 0.982% of Classified accurately of instances. K-NN achieved the lowest classification rate of 0.829%.

Table I. Comparison of different classifiers for image classification

Method	AUC	CA	F1	Precision	Recall
Logistic Regression	1.00	0.990	0.990	0.991	0.990
Neural Network	1.00	0.990	0.990	0.991	0.990
AdaBoost	0.999	0.990	0.990	0.991	0.990
SVM	1.00	0.981	0.981	0.982	0.981
Random Forest	0.999	0.981	0.981	0.982	0.981

Moreover, the results indicate substantial Neural Network, Logistic Regression to almost best on the values of AUC (Area under the curve) that is 1.00 and the AdaBoost value is 0.999. This analysis also shows that the highest Precision and Recall are achieved for the Logistic Regression algorithm, Neural Network, AdaBoost with values of 0.991 for precision and 0.990 for Recall. K-NN, on the other hand, per- formed the worst in this case with precision value is 0.847 and recall value is 0.829. Moreover, the analysis shows that the Logistic Regression, Neural Network, AdaBoost achieved the highest F1-measure value of 0.990 while the K-NN achieved the lowest F1-measure value of 0.828.

Figure1, Figure 2, Figure 3, Figure 4, Figure 5, Figure 6 presents the confusion matrices generated from Logistic Regression, Neural Network, AdaBoost, SVM, Random Forest, K-NN algorithms. In below confusion matrices, each row corresponds to a correct class, while columns represent the predicted classes. The rightmost column gives the number of instances from each class and bottom row gives the number of classified into each class. The diagonal elements represent the right prediction of the each class and other elements represent the wrong prediction of the each class.
It can be noted that Logistic Regression, Neural Network, AdaBoost algorithms most correctly classified the images of different classes, while the SVM and Random Forest correctly classified the 7 classes of dataset and SVM Correctly classified the 6 classes of dataset. And worst case classifier is K-NN which correctly classified only 3 classes of images.
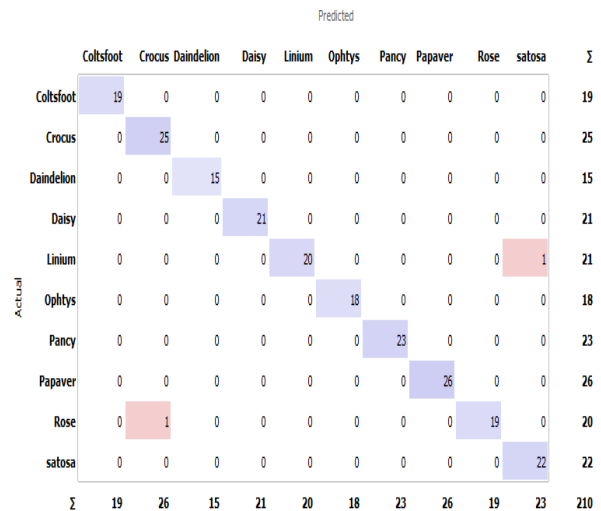

Figure 1: Logistic Regression Confusion Matrix

1059

Predicted

Actual	Coltsfoot	Crocus	Daindelion	Daisy	Linium	Ophtys	Pancy	Papaver	Rose	satosa	Σ
Coltsfoot	19	0	0	0	0	0	0	0	0	0	19
Crocus	0	25	0	0	0	0	0	0	0	0	25
Daindelion	0	0	15	0	0	0	0	0	0	0	15
Daisy	0	0	0	21	0	0	0	0	0	0	21
Linium	0	0	0	0	20	0	0	0	0	1	21
Ophtys	0	0	0	0	0	18	0	0	0	0	18
Pancy	0	0	0	0	0	0	23	0	0	0	23
Papaver	0	0	0	0	0	0	0	26	0	0	26
Rose	0	1	0	0	0	0	0	0	19	0	20
satosa	0	0	0	0	0	0	0	0	0	22	22
Σ	19	26	15	21	20	18	23	26	19	23	210
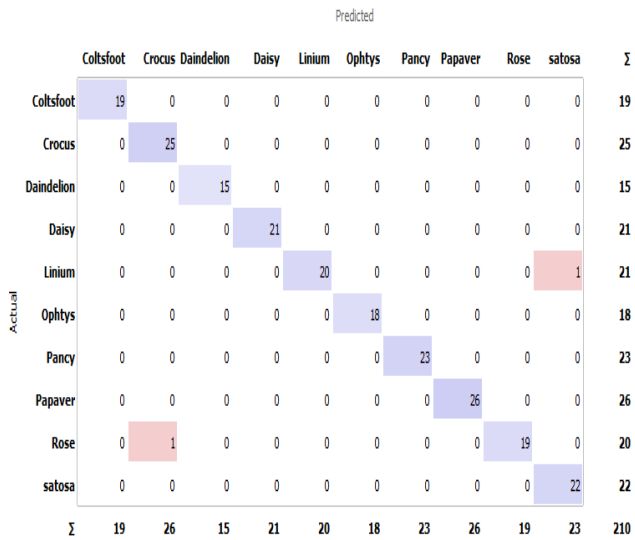
Figure 2: Neural Network Confusion Matrix

Predicted

Actual	Coltsfoot	Crocus	Daindelion	Daisy	Linium	Ophtys	Pancy	Papaver	Rose	satosa	Σ
Coltsfoot	18	0	0	0	0	0	1	0	0	0	19
Crocus	0	25	0	0	0	0	0	0	0	0	25
Daindelion	0	0	15	0	0	0	0	0	0	0	15
Daisy	0	0	0	21	0	0	0	0	0	0	21
Linium	0	0	0	0	21	0	0	0	0	0	21
Ophtys	0	0	0	0	0	18	0	0	0	0	18
Pancy	0	0	0	0	0	0	23	0	0	0	23
Papaver	0	0	0	0	0	0	0	26	0	0	26
Rose	0	1	0	0	0	0	0	0	19	0	20
satosa	0	0	0	0	1	0	1	0	0	20	22
Σ	18	26	15	21	22	18	25	26	19	20	210
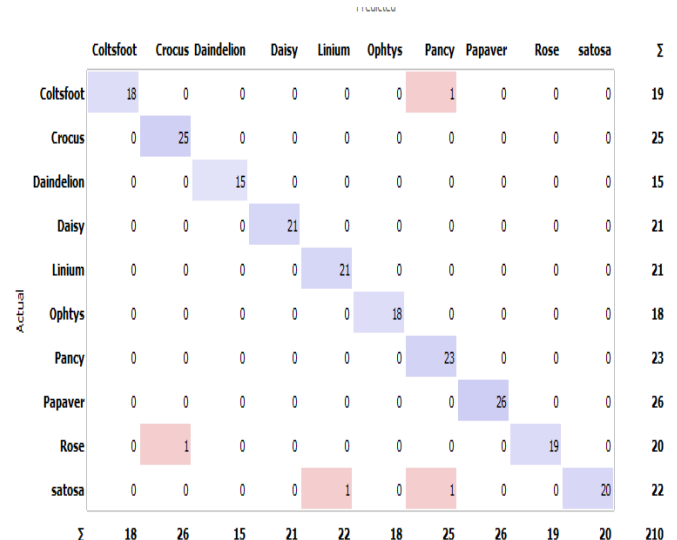
Figure 5: Random Forest Confusion Matrix

Predicted

Actual	Coltsfoot	Crocus	Daindelion	Daisy	Linium	Ophtys	Pancy	Papaver	Rose	satosa	Σ
Coltsfoot	19	0	0	0	0	0	0	0	0	0	19
Crocus	0	25	0	0	0	0	0	0	0	0	25
Daindelion	0	0	15	0	0	0	0	0	0	0	15
Daisy	0	0	0	21	0	0	0	0	0	0	21
Linium	0	0	0	0	21	0	0	0	0	0	21
Ophtys	0	0	0	0	0	18	0	0	0	0	18
Pancy	0	0	0	0	0	0	23	0	0	0	23
Papaver	0	0	0	0	0	0	0	26	0	0	26
Rose	0	1	0	0	0	0	0	0	19	0	20
satosa	0	0	0	0	1	0	0	0	0	21	22
Σ	19	26	15	21	22	18	23	26	19	21	210
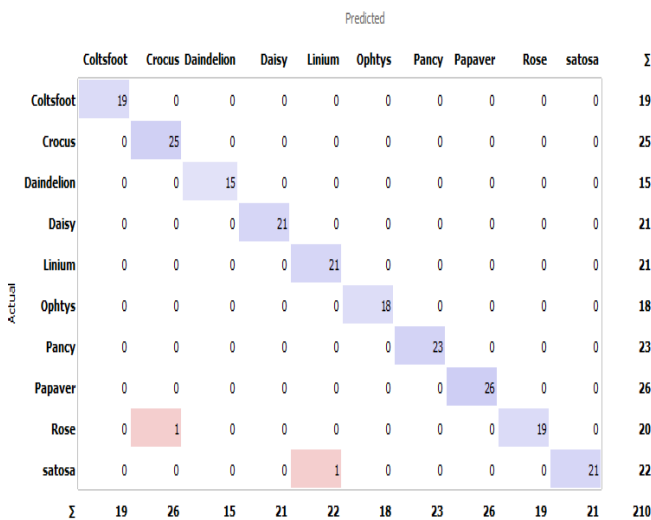
Figure 3: Adaboost confusion Matrix

Predicted

Actual	Coltsfoot	Crocus	Daindelion	Daisy	Linium	Ophtys	Pancy	Papaver	Rose	satosa	Σ
Coltsfoot	19	0	0	0	0	0	0	0	0	0	19
Crocus	0	20	0	1	1	0	1	1	0	1	25
Daindelion	4	0	10	1	0	0	0	0	0	0	15
Daisy	0	0	0	21	0	0	0	0	0	0	21
Linium	0	2	0	0	16	0	3	0	0	0	21
Ophtys	0	1	0	0	1	14	2	0	0	0	18
Pancy	0	0	0	0	0	0	23	0	0	0	23
Papaver	0	0	0	0	0	2	2	20	1	1	26
Rose	0	1	0	0	0	1	2	2	14	0	20
satosa	0	3	0	0	0	0	2	0	0	17	22
Σ	23	27	10	23	18	17	35	23	15	19	210

Figure 6: K-NN Confusion Matrix

Predicted

Actual	Coltsfoot	Crocus	Daindelion	Daisy	Linium	Ophtys	Pancy	Papaver	Rose	satosa	Σ
Coltsfoot	19	0	0	0	0	0	0	0	0	0	19
Crocus	0	25	0	0	0	0	0	0	0	0	25
Daindelion	1	0	14	0	0	0	0	0	0	0	15
Daisy	0	0	0	21	0	0	0	0	0	0	21
Linium	0	0	0	0	20	0	0	0	0	1	21
Ophtys	0	0	0	0	1	17	0	0	0	0	18
Pancy	0	0	0	0	0	0	23	0	0	0	23
Papaver	0	0	0	0	0	0	0	26	0	0	26
Rose	0	1	0	0	0	0	0	0	19	0	20
satosa	0	0	0	0	0	0	0	0	0	22	22
Σ	20	26	14	21	21	17	23	26	19	23	210

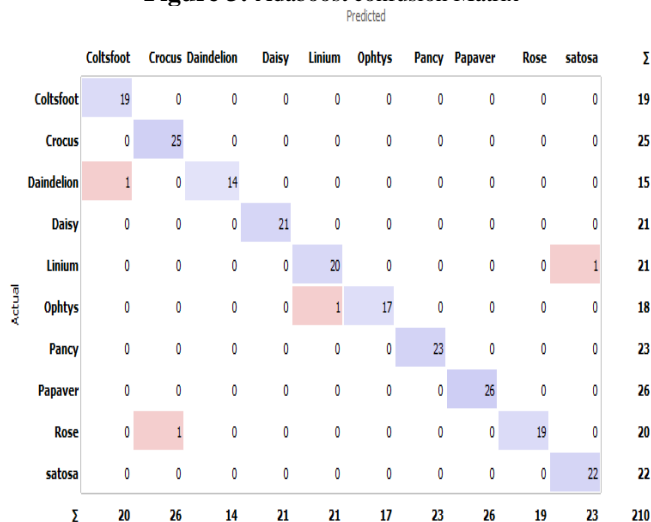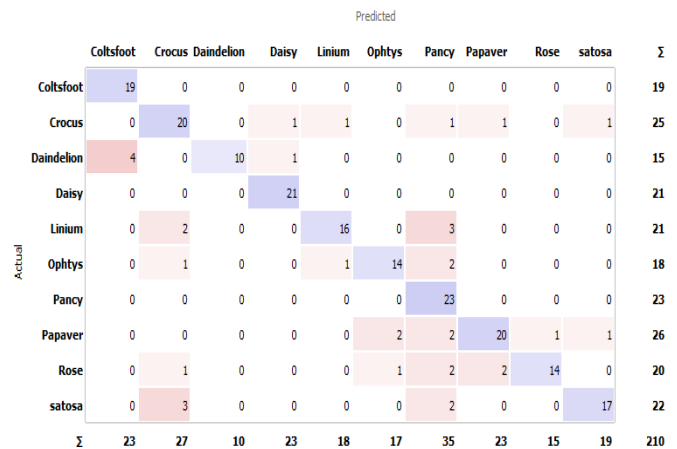**Figure 4:** SVM Confusion Matrix

V. CONCLUSION

This paper analyzed the application of methods in orange data mining tools for image classification using the different supervised classifier. The dataset included a small sample of flower images divided into ten classes, which included a total a 210 images. Six supervised classifier used : (1) Logistic Regression, (2) Neural Network, (3) AdaBoost, (4) SVM, (5) Random Forest and (6) K-NN. The obtained results indicate that, in terms of classification accuracy, the best one to use is the Logistic Regression, Neural Network, and AdaBoost. K-NN classifier achieved the lowest performance with 82% classification accuracy. This research is an ongoing work only involving a small sample of images, in which a total of 210 images within 10 classes. Future work will increase the sample size and create a much larger training database.

REFERENCES

1. R. J, "Supervised Classification Techniques," in *Remote sensing digital image analysis: An introduction*, Verlag Berlin Heidelberg: Springer, 2013, pp. 1–494.
2. D. Stephens and M. Diesing, "A comparison of supervised classification methods for the prediction of substrate type using multibeam acoustic and legacy grain-size data," *PLoS One*, vol. 9, no. 4, 2014.

3. S. R. Alty, S. C. Millasseau, P. J. Chowienczyk, and A. Jakobsson, "Cardiovascular disease prediction using support vector machines," pp. 376–379, 2006.
4. R. Bala and D. Kumar, "Classification Using ANN: A Review," *Int. J. Comput. Intell. Res. ISSN*, vol. 13, no. 7, pp. 973–1873, 2017.
5. A. Cutler, D. R. Cutler, and J. R. Stevens, "Random forests," *Ensemble Mach. Learn. Methods Appl.*, pp. 157–175, 2012.
6. H. Fleyeh and E. Davami, "Multiclass Adaboost Based on an Ensemble of Binary AdaBoost's," *Am J. Intell. Syst.*, vol. 3, no. 2, pp. 57–70, 2013.
7. A. B. Musa, "Logistic Regression Classification for Uncertain Data," *Res. J. Math. Stat. Sci. Res. J. Math. Stat. Sci*, vol. 2, no. 2, pp. 2320–6047, 2014.
8. J. Llamas, P. M. Lerones, R. Medina, E. Zalama, and J. Gómez-García-Bermejo, "Classification of Architectural Heritage Images Using Deep Learning Techniques," *Appl. Sci.*, vol. 7, no. 10, p. 992, 2017.
9. J. R. Otukei and T. Blaschke, "Land cover change assessment using decision trees, support vector machines and maximum likelihood classification algorithms," *Int. J. Appl. Earth Obs. Geoinf.*, vol. 12, no. SUPPL. 1, 2010.
10. G. Shalunts, Y. Haxhimusa, and R. Sablatnig, "Architectural Style Classification of Building FacadeWindows," in *International Symposium on VisualComputing*, 2011, pp. 280–289
11. J. Polpinij and C. Sibunruang, "Thai heritage images classification by naïve bayes image classifier," in *Proceeding - 6th International Conference on Digital Content, Multimedia Technology and Its Applications, IDC2010*, 2010, pp. 221–224.
12. J. Kurian and V. Karunakaran, "A Survey on Image Classification Methods," vol. 1, no. 4, pp. 69–72, 2012.
13. R. Jankovi, "Classifying Cultural Heritage Images by Using Decision Tree Classifiers in WEKA," in *1st International Workshop on Visual Pattern Extraction and Recognition for Cultural Heritage Understanding*, 2019, pp. 119–127.
14. Iris dataset at http://archive.ics.uci.edu/ml/datasets/iris
15. A. S. Rao, A. V. Ramana, and S. Ramakrishna, "Implementing the Data Mining Approaches to Classify the Images with Visual Words," *Int. J. Recent Technol. Eng.*, vol. 7, no. 6, pp. 901–909, 2019.

AUTHORS PROFILE

Ms. Vijaylakshmi Sajwan currently working as Assistant Professor in Uttaranchal Institute of Management,Uttaranchal University. She has done BCA from C.C.S university, MCA from IGNOU, New Delhi and M-Tech in IT from Graphic Era University, Dehradun. Currently, pursuing Ph.D from Himgiri Zee University, Dehradun. Field of specialization-Image Processing. Has attended Seminars and Conferences at various levels, three paper published in international journals (including SPRINGER).

Prof.(Dr.) Rakesh Ranjan currently working as Pro-Vice Chancellor in Himgiri Zee University. He has done B.Sc(Engg) in Electrical Engineering from MIT Muzzaffarpur, Bihar University, M.E from BIT Mesra and Ph.D from BITS Pilani. Has guided 9 Ph.D student, 25 M.Tech thesis. Has published 79 International Reaserach paper , 06 text book .He got Siksha Rattan-2013. Educationist of the year at 7th India Excellence Award, and Academy Award for Academic Excellence in Engineering Education.