Machine Learning Based Approach on Image Classification

Garima, Sunila Godara
Department of CSE, Guru Jambheshwar University of Science & Technology
garimachauhan0530@gmail.com
sunilagodara@gmail.com

***Abstract*--Machine learning has been in the latest trends for several years now.With rapidly changing technology and the constant need to properly store and update the data, various methods have been utilized for proper storage and classification of the former. Different behavioral aspects and a variety of predictions can be extracted from datasets if classified properly using methods that suit best on that particular type of dataset for the particular kind of knowledge extraction from the data. In this paper, we compare the results of different classification algorithms of machine learning using some performance measures. Classification can be applied on various types of datasets such as image datasets,textual,numericalalthough here for applying the classification algorithm, an image dataset has been chosen and worked upon. Some of the algorithms used for classification are Support Vector Machines, Logistic Regression,Naïve Bayes,Decision trees, Multilayer Perceptron. The performance of these algorithms has been evaluated on the basis of their Accuracy,Recall,Precision,F-measure.**

***Keywords*--Support Vector Machines,Logistic Regression,Naïve Bayes,Decision Trees,Multilayer Perceptron.**

I. INTRODUCTION

Image classification is the process of studying the image according to the available visual content of the image. An example of this being classifying whether an image is of an animal or a human, another example is classifying what kind of animal is there in the image a butterfly or an owl, a famous example of it being the classification of the MNIST dataset. Basically in image classification we put input as an image and gets output of a class (image belonging to what particular class) or probability of that the image is of what input class. Image classification is a classical problem of image pre-processing, human & computer vision working in the fields of machine learning.It consists of a systematic arrangement of the image dataset according to the various attributes and categories based on its features extracted from the image preprocessing done by the computer. It basically trains the algorithm being used to categorize the image dataset according to what it sees into the already prescribed categories of different images in the dataset. Here, we explore the classification of images with the help of machine learning approach. The traditional approaches to this study all being part of the Artificial intelligence which is a superset of both machine learning and deep learning. The former one mainly consisting of modules of feature extraction and classification based on the extracted features.

A. Machine Learning

Machine Learning is a subset of Artificial Intelligence that learns from the working of computer and algorithms and improves its efficiency based on the learning.Mainly,it is used to classify data according to its different behavioral aspects and predict various learning that can be further used in various studies.The whole process of working of Machine Learning starts with the different observations from the provided data finding features,various patternsmaking various predictions,better decisions,learning from the previous work done upon the data.The main objective is to learn automatically from the data without any kind of interventions, imitating the learning as humans learn from their vast experiences.Machine Learning methods can be categorized into two parts:Supervised learning and Unsupervised learning.Types of machine learning can be shown in Figure 1.

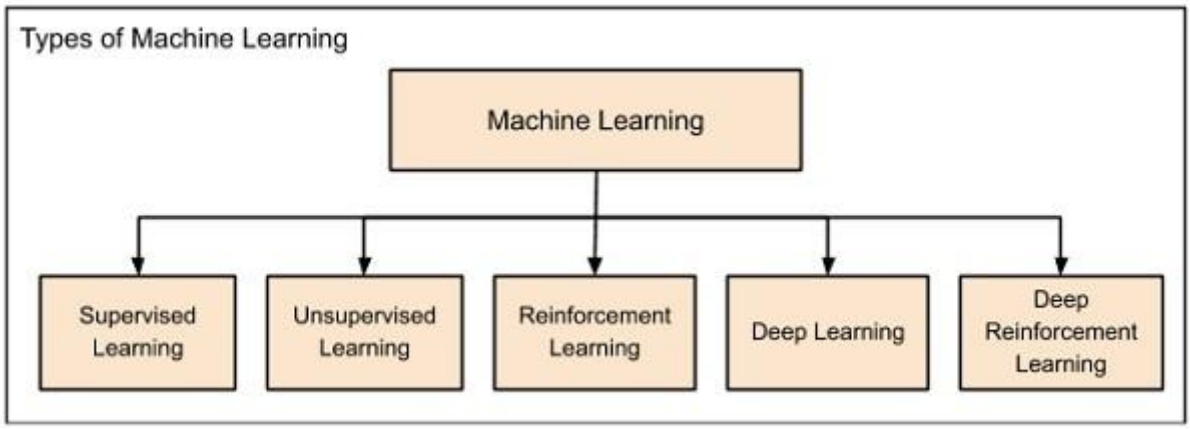

Figure 1. Types of Machine learning.

In Supervised Learning, the labeled classes and data is input to the machine for learning. Some examples of supervised learning algorithms are: Logistic Regression, Linear Regression, Support Vector Machines(SVM),K-Nearest Neighbour(KNN),Random Forest, Decision Trees. Whereas,in case of Unsupervised machine learning algorithms,no labeling of data is done,the data is neither labeled nor classified, no kind of help or supervision is provided to the machine for working upon a set of data,the machine has to train itself to learn from the data by considering the previous learning from the data it had once worked upon.It finds the hidden features,information from the unlabeled set of data on its own,drawing various inferences from the datasets. Clustering is the example of unsupervised learning.

II. RELATED WORK

Sandeep Kumar, Zeeshan Khan, Anurag Jain,"A Review of Content Based Image Classification using Machine Learning Approach",2012[1]. In this paper,various machine learning approaches has been used for image classification such asdecision trees, Markova mode; RBF network and Support vector machines.Features of images have been extracted from shapes,texture,colour of images and then a comparative analysis has been done on the results of all the algorithms used for the classification after extraction of features from images.

Le Hoang Thai, Tran Son Hai, Nguyen Thanh Thuy," Image Classification using Support Vector Machine and Artificial Neural Network" 2012[2],In this paper, two methods i.e. support vector machines and artificial neural networks have been used for image classification,both the approaches have been combined and integrated after the images are preprocessed. In this paper it has been found that the combined approach is easy to design and increases the precision as compared to other techniques as well.

Durgesh K. Srivastava, Lekha Bhambhu," Data Classification using Support Vector Machine",2010[3]. In this paper,support vector machines algorithm has been applied on various datasets having multiple classes each. Different results using different kernel functions have been illustrated and concluded that RBF kernel found to be the best for infinite data and multiple classes.

Q. Cheng, P. K. Varshney and M. K. Arora, "Logistic Regression for Feature Selection and Soft Classification of Remote Sensing Data,"2006,[4].In this paper,logistic regression has been applied for both feature selection and classification of remote images. Using logistic regression in the classification gave some positive results with few restrictive assumptions. The model used was able to reduce the extraction of features to a substantially low level without any notable decrease in the classification accuracy.

Stephan Dreiseitl,Lucila Ohno-Machado,"Logistic Regression and Artificial Neural Network classification: a methodology review",2002[5].In the paper,the dissimilarities and similarities of logistic regression model and artificial neural network based approach has been summarized from a technical point of view concluding that in case of logistic regression,model building process is easier as compared to neural networks. The latter being just the non-linearized generalization of the former one with no significant differences in the classification accuracies.

S. McCann and D. G. Lowe, "Local Naive Bayes Nearest Neighbor for image classification,"2012,[6].In this paper,Local naïve bayes nearest neighbor has been used for image classification.It has been observed that only the local neighborhood of the descriptor makes the significant difference in the probability predictions estimates. The increase in class accuracy has been found while ignoring any kind of adjustment to the distant classes and a significant improvement has been found using the local bayes as compared to the original naïve bayes in the class accuracy.

O. Boiman, E. Shechtman and M. Irani, "In defense of Nearest-Neighbor based image classification," 2008[7]..illustrates Naïve bayes, a on parametric nearest neighbor classifier has been consistently ignored whenever there is a need for classification of images despite it being easy,needing no extra parameters. It has been illustrated in this paper that the quantization of descriptors for using in the parameter based classifiers degrades the accuracy of non parametric classifiers such as the nearest neighbors classifier to significantly low levels. The extracted features that would have been discriminating one features from others are significantly reduced which damages the accuracy of the naïve bayes classifier. Comparisons has been done in this paper between Naive bayes algorithm and other parameter based classifiers and shown how quantization of parameters reduces the accuracy of nearest neighbor classifiers or that of the non parameter based classifiers.

Kun- Che Lu,Don-Lin Yang, "Image Processing and Image Mining using Decision Trees",2009,[8].In this paper,decision trees are used for data preprocessing and data mining of image datasets. Features are extracted from the images in a pixel wise format and changed into a table format upon which algorithms are used. Pixel processing on the images has been done using the decision tree induction method and relationship between the extracted attributes has been studied.Data mining and image processing tools have been merged.

Chun-Chieh Yang, Shiv O. Prasher, Peter Enright, Chandra Madramootoo, Magdalena Burgess, Pradeep K. Goel, Ian Callum,"Application of decision tree technology for image classification using remote sensing data",2003,[9].Classification and Regression tree approach has been used in this paper on the hyperspectral images of agricultural plots to classify the images. It has been illustrated how decision classifiers on the remote sensing can be a very cost effective method in the practices. Decision tress using a top down approach can be effectively used in classification,filtering datasets also relatively interpreting the importance of the relation between various variables in the images.

N. E. Abdullah, A. A. Rahim, H. Hashim and M. M. Kamal, "Classification of Rubber Tree Leaf Diseases Using Multilayer Perceptron Neural Network,"2007[10]. This paper illustrates the classification of diseases in leaves using RGB color models and automation.Images of the diseases are captured,pre processing on the images takes place using image classification filters and techniques through neural networks. In this paper,artificial neural network approach has been used.

Dennis W. Ruck, Steven K. Rogers, Matthew Kabrisky "Feature Selection Using a Multilayer Perceptron",1990[11]shows how to select the best suited features for recognition of the target workload using multilayer perceptrons. In this paper,the approach has been applied on two different image datasets and the results of both are then compared on various training rules.

III. MACHINE LEARNING TECHNIQUES FOR IMAGE CLASSIFICATION

A. Support Vector Machines

Support vector machines are supervised deep learning algorithms that work on classifying datasets,regression and outlier analysis of outlying points in the datasets of various kinds such as linear and nonlinear data. SVM have been found to be working really good on small datasets.In case of linear datasets in which the data when plotted on a graph can be separated by a hyperplane,while in case of nonlinear data,it maps the data to a higher dimension and segregates the data according to the class it belongs.The data points that are found to be the closest to the hyperplane are known as Support vectors.

B. Logistic Regression

Logistic Regression is another example of supervised algorithm that works on the probability of whether a data image belongs to a particular class or not.It is a statistical procedure to find the probability of data belonging to a particular class.It classifies observations by estimating the probability that an observation is in a particular category such as in case of image datasets,if an image consists of an owl or a butterfly.In other words,it models,estimates,predict and classifies the data based on the probability factors.

C.Naive Bayes

Naïve Bayes is a supervised deep learning algorithm that works on the Bayes' Theorem which is in turn is based on the conditional probability i.e. the likelihood of occurrence of an event given that another event has already taken place.

$$P(A/B)=P(B/A)P(A)/P(B)$$

D. Decision Trees

Decision trees are yet another example of supervised deep learning algorithms in which the main dataset is continuously split in smaller subsets based on a condition whether it follows or not making a flow-chart like structure. This algorithm is used both for classification and regression.At each and every node, a question or a condition is put depending on which the dataset will be divided continuously until a definite class is obtained for the dataset.

E.MultiLayer Perceptrons

Multi-layer perceptrons or MLP are feed forward neural network algorithms especially used for binary classificationmostly applied to supervised learning algorithms.In MLP,various layers are present such as input layer,hidden layers and lastly the output layer.We feed the input in the algorithm through the input layer and get the output via the output layer. The number of hidden layers can be customized making the model complex as per the needs.In this algorithm the basic method is multiplying the weights adding bias and then updating the weights whenever any kind of error is found in the classification.

IV. RESULTS AND ANALYSIS

A. Methodology

In Image Classification, the first step is deciding on the dataset.After the dataset has been decided that has to be used for the classification techniques, it is input to the classification tool following which some image pre-processing takes place.In Image Pre-processing,different Image filters are used on the dataset to extract features from the images that would help in the accurate classification. In this paper, SimpleColorHistogramFilter has been used, which gives the most basic distribution of colors and extract them from the features of image such as color, brightness, number of pixels in the image. Then the pre-processed dataset is fed to the classifiers. In this paper, five different classifier algorithms of machine learning which are Support Vector

Machines, Logistic Regression, Naïve Bayes, Decision Trees, Multilayer Perceptrons have been used. The Image dataset pre-processing is done only once before the dataset to be input to different classification.There is no need for the pre-processing process to be done repeatedly. After each classifier has been applied to the dataset,each gives different results,based on which the performance measures are calculated which are: Accuracy, Precision, Recall and F-Measure. The formulae for calculating the above measure has been discussed further.Also,the whole methodology has been depicted in the form of a flow chart in Figure 2.

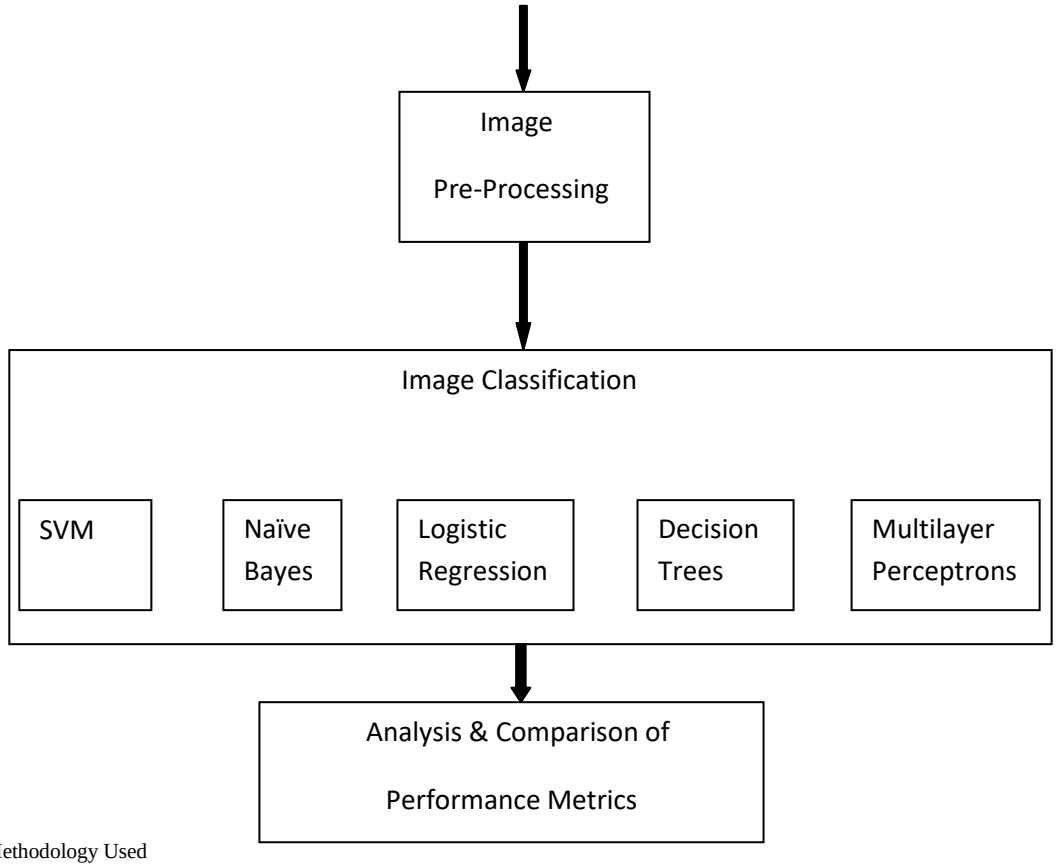

Figure 2. Methodology Used

An Image dataset of vehicle images containing three different classes namely: Car, Plane and Train is used which can be shown in Figure 3.

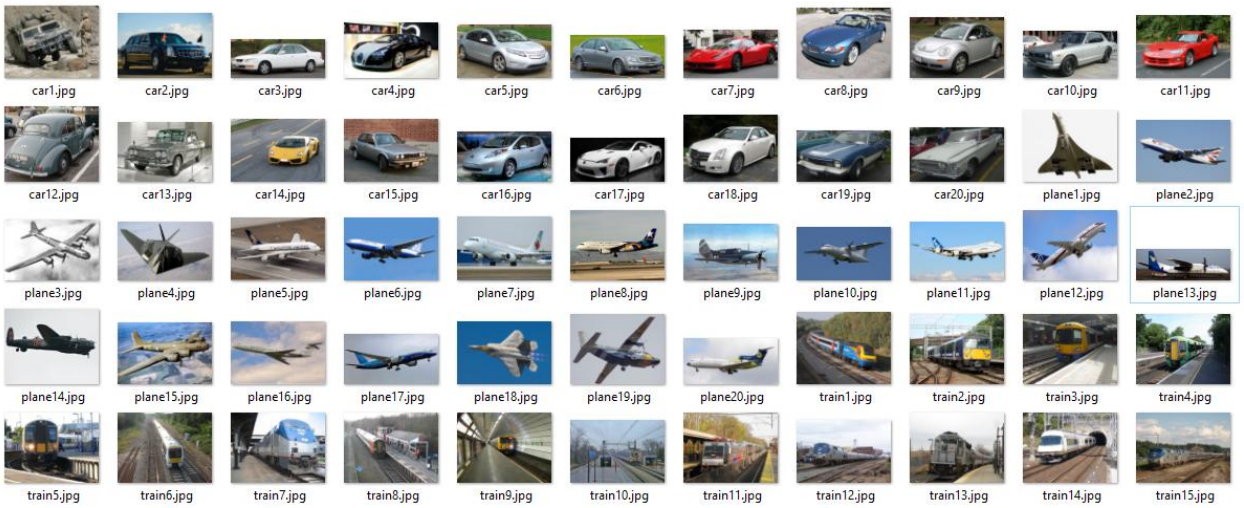

Figure 3.Image dataset used.

The image dataset used in Figure 3 has been taken from the inbuilt library of imageFilters package of WEKA tool.

After some preprocessing on the data using image filters,different classification algorithms were applied on the dataset and the performance measures of each classifier compared with that of the performance of other classifier.The performance measures used were Accuracy which is defined as the number of correctly classified instances of the classes,Recall,Precision and F-measure.

After the classifier has been successfully used on the dataset,a confusion matrix is formed in which we get the values such as TP(True Positive),TN(True Negative),FP(False Positive),FN(False Negative).Confusion matrix can be depicted in the form of table in Table I as below[13].

TABLE I. CONFUSION MATRIX

	Predicted NO	**Predicted YES**
Actual NO	TN	FP
Actual YES	FN	TP

With the help of these we can further calculate the performance measures by using the below formulae[12]:

Overall Accuracy= (TP + TN)/ (TP+TN+FP+FN)

Precision= TP/ (TP+FP)

Recall= TP/ (TP+FN)

F-Measure= 2(Precision * Recall)/ (Precision + Recall)

Following figures from Figure 4 to Figure 8 are screenshots of the results of classification algorithm of Support Vector Machines (SVM), Logistic Regression, Naïve Bayes, Decision Trees, Multilayer Perceptron.

```
=== Summary ===

Correctly Classified Instances          44               73.3333 %
Incorrectly Classified Instances        16               26.6667 %
Kappa statistic                          0.6
Mean absolute error                      0.2926
Root mean squared error                  0.3801
Relative absolute error                 65.8333 %
Root relative squared error             80.6226 %
Total Number of Instances               60

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.900    0.075    0.857      0.900    0.878      0.815    0.932     0.826     PLANE
               0.750    0.200    0.652      0.750    0.698      0.533    0.808     0.596     TRAIN
               0.550    0.125    0.688      0.550    0.611      0.453    0.724     0.538     CAR
Weighted Avg.  0.733    0.133    0.732      0.733    0.729      0.601    0.821     0.654

=== Confusion Matrix ===

  a  b  c   <-- classified as
 18  1  1 |  a = PLANE
  1 15  4 |  b = TRAIN
  2  7 11 |  c = CAR
```

Figure 4. Results of applying SVM on image dataset.

```
=== Summary ===

Correctly Classified Instances          37               61.6667 %
Incorrectly Classified Instances        23               38.3333 %
Kappa statistic                          0.425
Mean absolute error                      0.2624
Root mean squared error                  0.5073
Relative absolute error                 59.0471 %
Root relative squared error            107.6076 %
Total Number of Instances               60

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.650    0.200    0.619      0.650    0.634      0.445    0.766     0.616     PLANE
                 0.800    0.250    0.615      0.800    0.696      0.523    0.784     0.593     TRAIN
                 0.400    0.125    0.615      0.400    0.485      0.315    0.639     0.492     CAR
Weighted Avg.    0.617    0.192    0.617      0.617    0.605      0.428    0.730     0.567

=== Confusion Matrix ===

  a  b  c   <-- classified as
 13  5  2 |  a = PLANE
  1 16  3 |  b = TRAIN
  7  5  8 |  c = CAR
```

Figure 5. Results of applying Logistic Regression on image dataset.

```
=== Summary ===

Correctly Classified Instances          39               65      %
Incorrectly Classified Instances        21               35      %
Kappa statistic                          0.475
Mean absolute error                      0.2622
Root mean squared error                  0.4263
Relative absolute error                 58.9964 %
Root relative squared error             90.4237 %
Total Number of Instances               60

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.850    0.075    0.850      0.850    0.850      0.775    0.948     0.864     PLANE
                 0.600    0.175    0.632      0.600    0.615      0.431    0.814     0.696     TRAIN
                 0.500    0.275    0.476      0.500    0.488      0.222    0.659     0.442     CAR
Weighted Avg.    0.650    0.175    0.653      0.650    0.651      0.476    0.807     0.667

=== Confusion Matrix ===

  a  b  c   <-- classified as
 17  0  3 |  a = PLANE
  0 12  8 |  b = TRAIN
  3  7 10 |  c = CAR
```

Figure 6. Results of applying Naïve Bayes on Image dataset.

```
=== Summary ===

Correctly Classified Instances          36               60      %
Incorrectly Classified Instances        24               40      %
Kappa statistic                          0.4
Mean absolute error                      0.2815
Root mean squared error                  0.4736
Relative absolute error                 63.3412 %
Root relative squared error            100.4698 %
Total Number of Instances               60

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.750    0.075      0.833      0.750    0.789      0.694    0.859     0.760     PLANE
                 0.650    0.325      0.500      0.650    0.565      0.309    0.715     0.474     TRAIN
                 0.400    0.200      0.500      0.400    0.444      0.213    0.610     0.426     CAR
Weighted Avg.    0.600    0.200      0.611      0.600    0.600      0.406    0.728     0.553

=== Confusion Matrix ===

  a  b  c   <-- classified as
 15  3  2 |  a = PLANE
  1 13  6 |  b = TRAIN
  2 10  8 |  c = CAR
```

Figure 7. Results of applying Decision Tree algorithm on Image dataset.

```
=== Summary ===

Correctly Classified Instances          40               66.6667 %
Incorrectly Classified Instances        20               33.3333 %
Kappa statistic                          0.5
Mean absolute error                      0.2797
Root mean squared error                  0.4182
Relative absolute error                 62.9368 %
Root relative squared error             88.7099 %
Total Number of Instances               60

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.750    0.050      0.882      0.750    0.811      0.732    0.900     0.887     PLANE
                 0.750    0.300      0.556      0.750    0.638      0.426    0.785     0.641     TRAIN
                 0.500    0.150      0.625      0.500    0.556      0.373    0.729     0.587     CAR
Weighted Avg.    0.667    0.167      0.688      0.667    0.668      0.511    0.805     0.705

=== Confusion Matrix ===

  a  b  c   <-- classified as
 15  4  1 |  a = PLANE
  0 15  5 |  b = TRAIN
  2  8 10 |  c = CAR
```

Figure 8. Results of applying Multilayer Perceptron on Image dataset.

TABLE II. COMPARISON OF PERFORMANCE MEASURES

Classification method	Accuracy(%)	Precision(%)	Recall(%)	F-Measure(%)
SVM	73.333	73.2	73.3	72.9
Logistic Regression	61.6667	61.7	61.7	60.5
Naïve Bayes	65	65.3	65	65.1
Decision Trees	60	61.1	60	60
Multilayer Perceptron	66.6667	68.8	66.7	66.8

In Table II, we compare the performance measures of the classification techniques which have been used.The screenshots in Figures 4,5,6,7,8 clearly shows that the correctly classified instances in case of SVM is much higher than any other machine learning algorithm with SVM having accuracy of 73%, Logistic Regression with an accuracy of 61.6667%, Naïve Bayes accuracy as 65%, Decision tree as 60% and accuracy of Multilayer Perceptrons as 66.6667% with a significant difference of 13% between the accuracy of SVM and that of Decision Trees, difference of 12% with that of Logistic Regression, difference of 8% with the accuracy results of Naïve Bayes and 7% difference with compared to the accuracy of MLP. The Precision of SVM is 73% which is 5% higher than MLP(Multilayer Perceptrons) whose precision lies at68%, followed by 65% precision of Naïve Bayes,61.7% of Logistic Regression with 11.3% difference in precision as compared to SVM and 61% of Decision Trees making the difference between the precision of SVM with the highest precision and Decision tree with the lowest precision to be 12%.Another Performance measure: Recall with the highest value placed at 73% of SVM classification technique, 66.7% Recall rate of MLP,65% of Naïve Bayes,61.7% of Logistic Regression and 60% Recall rate of Decision tree with the lowest value.The difference in the maximum(that of SVM) and minimum(that of Decision tree) value of Recall lies at 13%.The Last Performance measure is F-Measure,the value of which for SVM is 72.9%,for MLP lies at 66.8%,for Naïve Bayes is 65%, in case of Logistic regression the value of F-Measure lies at 60.5% and that of Decision tree at 60% with the lowest value of F-Measure and a difference of 12.9% with that of SVM with the highest value of F-measure.

Clearly,SVM has the best results as compared to Logistic regression, Naïve Bayes, Decision tress and Multilayer Perceptrons.The performance measures can be shown in the form of a bar graph as well in Figure 9.

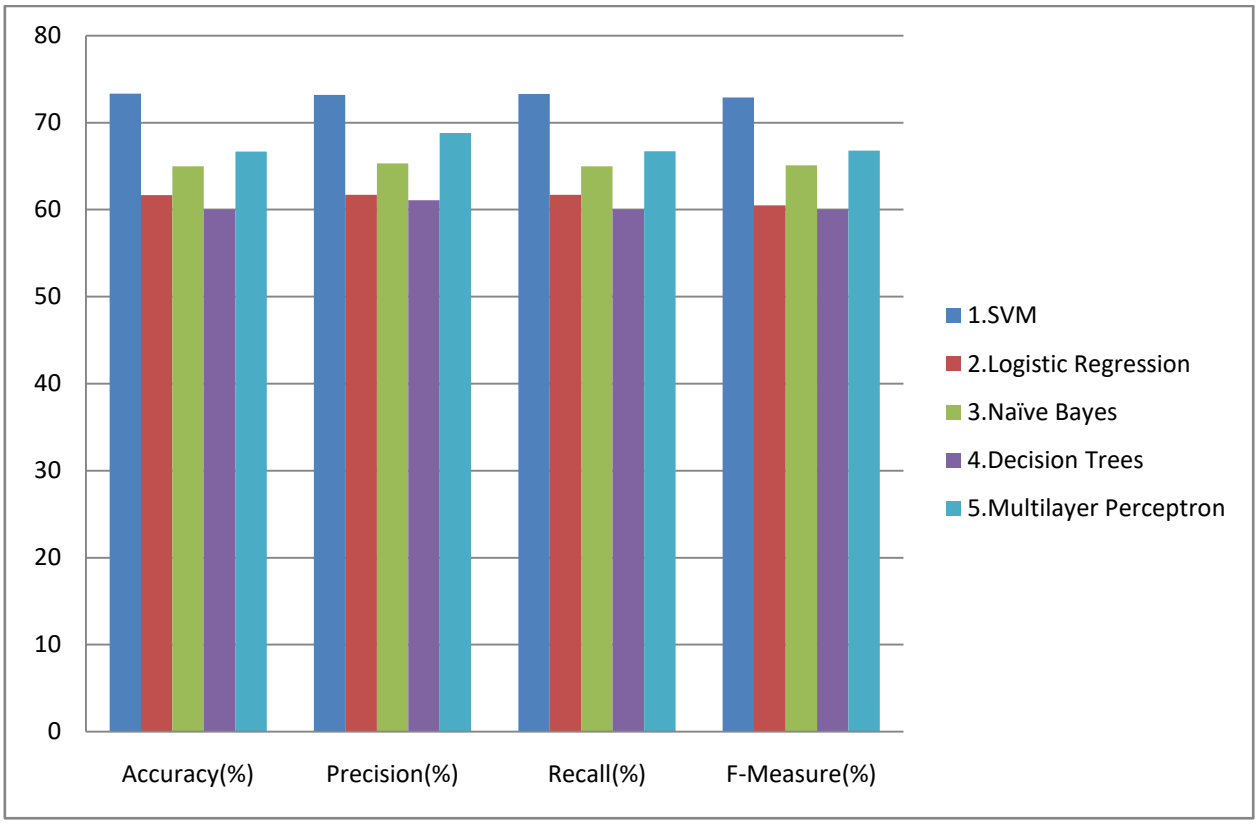

Figure 9. Bar graph Comparison of Performance measures.

V. CONCLUSION

In this paper, Five Machine Learning algorithms: Support Vector Machines, Logistic Regression, Naïve Bayes, Decision Trees, Multilayer Perceptrons have been used on an Image dataset.The results of the above mentioned classification techniques have been compared using performance measures of Accuracy, Precision, Recall, F-Measure. The Comparison of the results illustrates that the best machine learning algorithm to use on image datasets for classification is the Support Vector Machines which shows the highest level of accuracy or the maximum number of correctly identified instances with the highest value of the other measures taken into consideration as compared to the remaining four machine learning algorithms.SVM has the accuracy of 73% with a significant difference in accuracies of 7,8,12 and 13% in comparison to MLP, Naïve Bayes, Logistic regression and Decision tree respectively with the lowest value of all the performance measure in classifying our image dataset to be of Decision tree. Finally, it has been concluded that SVM is the best classification algorithm for image classification on our image dataset.

REFERENCES

[1]. Sandeep Kumar, Zeeshan Khan, Anurag Jain,*"A Review of Content Based Image Classification using Machine Learning Approach*",2012, International Journal of Advanced Computer Research ,vol.2,no.3,pp.2249-7277,2012.

[2]. Le Hoang Thai, Tran Son Hai, Nguyen Thanh Thuy," *Image Classification using Support Vector Machine and Artificial Neural Network*" 2012 I.J. Information Technology and Computer Science, 2012, 5, pp 32-38.

[3]. Durgesh K. Srivastava, Lekha Bhambhu," *Data Classification using Support Vector Machine*",2010 Journal of Theoretical and Applied Information Technology,vol 12,2010.

[4]. Q. Cheng, P. K. Varshney and M. K. Arora, "*Logistic Regression for Feature Selection and Soft Classification of Remote Sensing Data*," in IEEE Geoscience and Remote Sensing Letters, vol. 3, no. 4, pp. 491-494, Oct. 2006, doi: 10.1109/LGRS.2006.877949.

[5]. Stephan Dreiseitl,Lucila Ohno-Machado,"*Logistic Regression and Artificial Neural Network classification: a methodology review*",2002,Journal of Biomedical Informatics,vol 35,pp 352-359.

[6]. S. McCann and D. G. Lowe, "*Local Naive Bayes Nearest Neighbor for image classification*," 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, 2012, pp. 3650-3656, doi: 10.1109/CVPR.2012.6248111.

[7]. O. Boiman, E. Shechtman and M. Irani, "*In defense of Nearest-Neighbor based image classification*," 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, 2008, pp. 1-8, doi: 10.1109/CVPR.2008.4587598.

[8]. Kun- Che Lu,Don-Lin Yang, "*Image Processing and Image Mining using Decision Trees*",,2009,Journal of Information Science and Engineering, 2009, vol. 25, pp-989-1003.

[9]. Chun-Chieh Yang, Shiv O. Prasher, Peter Enright, Chandra Madramootoo, Magdalena Burgess, Pradeep K. Goel, Ian Callum,"*Application of decision tree technology for image classification using remote sensing data*",2003, Agricultural Systems 76,2003,pp 1101-1117.

[10]. N. E. Abdullah, A. A. Rahim, H. Hashim and M. M. Kamal, "*Classification of Rubber Tree Leaf Diseases Using Multilayer Perceptron Neural Network*," 2007 5th Student Conference on Research and Development, Selangor, Malaysia, 2007, pp. 1-6, doi: 10.1109/SCORED.2007.4451369.

[11]. Dennis W. Ruck, Steven K. Rogers, Matthew Kabrisky "*Feature Selection Using a Multilayer Perceptron*" Journal of Neural Network Computing, Volume 2, Number 2, 1990, pp 40-48.

[12].Sunila Godara,Rishipal Singh,"*Evaluation of Predictive Machine Learning Techniques as Expert systems in Medical Diagnosis*",Indian Journal of Science and Technology,vol.9,issue 10,March 2016.

[13].Table II. Confusion matrix,"*https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/*".

[14].Figure 1,"*https://www.tutorialspoint.com/machine_learning/machine_learning_categories.htm",* Typesof Machine Learning by tutorialspoint.