ISSN 2278-3091
Volume 8, No.5, September - October 2019
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse78852019.pdf
https://doi.org/10.30534/ijatcse/2019/78852019

Classification of Mushroom Fungi Using Machine Learning Techniques

Mohammad Ashraf Ottom[1], Noor Aldeen Alawad[2], Khalid M. O. Nahar[2]

[1]Department of Information Systems, Yarmouk University, Jordan
[2]Department of Computer Sciences, Yarmouk University, Jordan

ABSTRACT

Mushroom is one of the fungi types' food that has the most potent nutrients on the plant. Mushrooms have major medical advantages such as killing cancer cells. This study aims to find the most appropriate technique for mushroom classification, and mushroom will be classified into two categories, poisonous and nonpoisonous. The proposed approach will implement a different techniques and algorithms like neural network (NN), Support Vector Machines (SVM), Decision Tree, and k Nearest Neighbors (KNN), on dataset of mushroom images, where the dataset contains images with background and without background. The experimental results shown that the best technique for classifying mushroom images is kNN with accuracy of 94% based on features extracted from images with real dimensions of mushroom types, and 87% based on features extracted from images only.

Keywords: Machine Learning, Mushroom Classification, Supervised Learning.

1. INTRODUCTION

Nowadays, there are different challenges to develop systems that analyze a huge and complex data to make better decisions. This study aims to find new approach working to classify the mushrooms images based on different features using the different techniques of Machine Learning (ML). The purpose of classification process is to predict categorical class labels or the target value [1], for example, feed-forward Artificial Neural Network (ANN), and the purpose of classifier is to map data to predefined classes or groups [2]. In the proposed approach, we used the training dataset that contain the mushroom images to classify it into poisonous and nonpoisonous. Where our approach aims to classifies and predict for the class (groups) of mushrooms when submit the features of the mushrooms to different techniques of machine learning.

A mushroom is one of the fungi types' food that has the most potent nutrients on the plant. Mushrooms have major advantages such as kill cancer cells, viruses and enhancing the human immune system. Currently, the mushroom refers to the process that performed by robot in food industry. This technique used to limit the features such as color. Recently, mushroom system used specific characteristics that improve the selection process of mushrooms. Such system depends on analyzing and investigating the features in order to get better classification based on the well-known features [3].

1.1 Machine Learning

Machine learning (ML) relies under the umbrella of artificial intelligence [4], that allows computer systems to learn based on previous history, experience, examples, and data, it has been making great progress in many directions. Machine learning involving the study of computational learning and pattern recognition theory in the artificial intelligence. In addition, machine learning spots the light on the construction of techniques that can learn and make predictions on available data. In instance, applications such as detection of network intruders or email filtering, optical character recognition (OCR), and computer vision [5].

1.2 Algorithms and Techniques

In this study, we will use different machine learning algorithms and techniques for mushroom classification, some of them are listed below:

- **Neural Network (NN):** is a distributed matrix structure, it used in different applications, such as classifying data and patterns, , predicting new cases or examples, and in pattern recognition applications. NN simulates human biological cells and human capability of thinking and learning [5][6].
- **Decision Tree:** is one of the most popular classification techniques in machine learning, where it used in decision support system. Decision aims to classify objects (instances) to find a track from the great parent node [7][8].
- **kNN:** is an algorithm and classified under machine learning. The kNN featured the low number of training parameters, where the computational complexity is not high, and the performance is satisfactory [7].

2. LITERATURE REVIEW

There are different researches using of different techniques that are used for mushrooms classification. a Mushroom Diagnosis Assistance System (MDAS) was proposed by [3], which involves three components of web application (server), unified database and mobile phone application (client) which is used on mobile phone devices. The Naive Bays and Decision Tree classifiers are used to determine the mushroom types. Firstly, the suggested system chooses the most known mushroom attributes. Secondly, specify the mushroom type. The experiment results show that Decision Tree classifier is better than Naïve Bays classifier in correct and incorrect classified instances, and error measurements.

Kumar and others in [9] compared different classification techniques that are used in data mining for decision systems. A comparison take place among three decision trees algorithms represented by one statistical, one artificial neural network, one support vector machines and one clustering algorithm. The suggested approach uses four datasets from several domains to test the predictive accuracy, error rate, comprehensibility, classification index and training time. The experimental results showed that Genetic Algorithm (GA) and support vector machines algorithms are better compared with the others in the predictive accuracy metric. In decision tree-based algorithms, QUEST algorithm generates trees with smaller breadth and depth. In conclusion, the GA based algorithm is the best algorithm that can be used for their decision support systems.

Babu and others in [10] proposed a new application domain that is used for SVM. The suggested approach uses the Support Vector Machine and Naïve Bayes algorithms for classification of mushrooms. The experiments results showed that SVM is better compared to Naïve Bayer's algorithm in term of accuracy. In conclusion, the SVM is an efficient technique that can be used for application domain. [2] used Multi-Layer Perception for Dataset training to create a model which is used to prediction of classifying. In the experiment, only 8124 of dataset are used for training. The experiment result showed that the best-hidden unit is 2, the best learning rates 0.6, the best activation function is sigmoid, the best moment rate is 0.2 and the best result of epoch is 300.

Onudu in [11] suggested modified K-means technique based on the traditional k-mean algorithm to enhance the clustering categorical dataset and solving the inherent problem in the traditional clustering algorithm. The suggested method is depending on Euclidean distance measure. In the suggested algorithm, the data set converted into numeric values. Then, the algorithm read the input data with normalizes the numeric attributes to avoid the wide range of values. The experiment result showed that the suggested modified K-means techniques faster compared to the existing algorithm.

Al-mejibli and Hamad in [1] developed an application can be applied on a mobile phone and web application named Mushroom Diagnosis Assistance System, the purpose of this application is to realize safety when gathering mushroom. They used decision tree and naïve bays classifiers to group the mushrooms types. They depended on the most famous mushroom attributes to determine the mushroom type. This model has to main phases: training phase and selection phase, to assign most active features in selection process and locate the final decision. The experimental results showed that decision tree was better than naive bays based on error measurements, correctly classified samples and incorrectly classified samples. The authors of [12] analyzed a previous mushroom data set by using different data mining techniques and Weka mining tool. They used nearest neighbor classifier, covering algorithm to collect correct rules, unpruned decision tree and a voted perceptron algorithm. They reached from running the techniques on different groups by stockholders that unpruned tree gives the best accuracy result and then it used on human-machine application based on web to produce interactive mushroom identification.

Chowdhury and S. Ojha in [13] identified a manner to distinguished several mushroom diseases using different data mining classification methods. They used actual dataset gathered from mushroom farm by using data mining like Naïve Bayes, RIDOR and SMO algorithms. They performed comparison based on a statistical way to detect popular symptoms for mushroom to discover mushroom disease. They reached that naïve Bayes gives best result with comparisons to other classification techniques. Beniwal and Das in [14] used data mining classification techniques such as Zero, naïve Bayes and Bayes net to analyze mushroom dataset that contain various kinds of mushrooms, which are poisonous or not poisonous. They evaluated classification techniques by using accuracy, kappa statistic and mean absolute error. They reached that Bayes net gives the lowest mean absolute error and highest accuracy and then naïve Bayes.

3. METHODOLOGY

The aim of this study is to identify Mushroom images and classify it into two categories (poisonous and nonpoisonous) using machine learning techniques.

3.1 Research Phases

In this research our methodology consists of five phases: the first phase is collecting dataset, second phase is preprocessing the dataset, third phase is features extraction, then machine learning model, and finally evaluation phase. Figure 1 shows the research phases for the proposed approach.

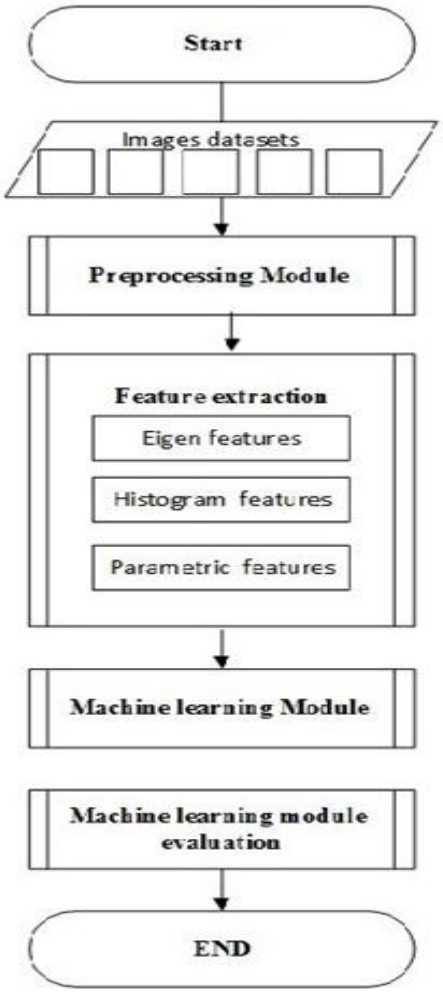

Figure 1: Research phases

3.2 Collecting Dataset

In the first step, we collected our dataset (raw dataset) of mushroom images from [15], where the collected dataset consists of three categories (edible, inedible, and poisonous). There is a different information identify each type of mushroom beside images, such as: family, location, dimensions, and edibility. Figure 2 below shows an example of mushroom images:

Figure 2: Dataset sample

3.3 Features Extracting

In this phase, we used Matlab for extracting all the features from collected images in the raw dataset. Firstly, we extract the Eigen features for each image after resizing it. Secondly, we take the top 100 strongest. We use the dimension's information that available with each type in dataset and include this information with feature matrix for the dataset, i.e. Cap diameter, stem tall and diameter. Finally, we build the feature matrix, which contain each of dimensions with the Eigen features with cap diameter, stem tall and diameter, to build the Machine Learning (ML) Model.

In the proposed approach, the ML model applied by different techniques, such as: Neural Network (NN), Decision Tree (DT), Support Vector Machine (SVM), and KNN. The best results were for KNN with cross validation, where number of folds equal 10, and accuracy 94%. Even though because of difficulty to get the real measurements for mushrooms, we decided to use extracted features from images only.

In order to try to enhance results, we find the width and height for the shape of mushroom inside the pictures using detecting edges in gray scale mode of images as shown in Figure 6. Then add these dimensions to the Eigen features. The experiment results shown that KNN produced accuracy of 86%. We attempted to extract different features from images to enhance the results like histogram features. We applied the same

steps in previous experiment to calculate each of height and width according to detect edges. We built new features matrix for the dataset. The experiment results of histogram features shown the accuracy reached to 87%. To optimize our results, we build an algorithm aims to extract more features we called this group as parametric features, which are:

- Local Contrast Normalization (LCN): used to contrast features within a feature map, as well as across feature maps at the same spatial location, where it inspired by computational neuroscience [16].
- Skewness, Standard deviation, and Kurtosis: are considering as concepts in the statistical meta-features, which are calculated by considering a statistical concept, calculate this for all numeric attributes and taking the mean [17].
- Entropy: is a concept with a complex history and has been the subject of diverse reconstructions and interpretations, where it's defined as the average amount of information produced by a stochastic source of data [18].
- Mean: is useful in assessing expected losses and benefits. For instance, in the proposed approach we used mean to determine in features matrix, especially for calculate the height and width for images.
- Correlation: Correlation is one of the most widely used, where the term "correlation" refers to a mutual relationship or association between quantities [19].
- Homogeneity: is one of the broad categories of distributed data mining, it refers to the process of mining the same set of attributes over all the participating nodes.
- Diameter: the real or virtual diameter of the length of the mushroom stem tall.

3.4 Noise Reduction

Noise reduction is an important factor that influences image quality [20], its working to reduce the errors of image which it has problems. In the proposed approach, we will use noise reduction to remove un-useful sections from original images, such as background of images.

3.5 Features Extraction Images Without Background

In this phase, we used Matlab to extract Eigen features for updated images (i.e. images without background), to build features matrix again. Depending on the edges for the mushroom images, we calculate the height and width for each image, using detecting edges in gray scale mode to build new dataset. We applied NN, DT, SVM, and KNN algorithms on Orange 3 software, where the best results were for KNN with 80% for the accuracy.

3.6 Machine learning model

After feature extraction phase we use Orange3 & Knime to build a machine learning model and apply different algorithms like SVM, Neural Network, Decision Tree and KNN. We use random sampling with training set size 66% but we didn't get a good result due to the small number of the dataset instances which were 380 instances. Therefore, we use cross validation with 10 folds. After building the trained model we evaluate the results in term of accuracy, f-measure, precision and recall. The confusion matrix is used to determine the percentage of wrongly classified instances. Figures 3 and 4 portrayed the training model in orange3 and Knime respectively.

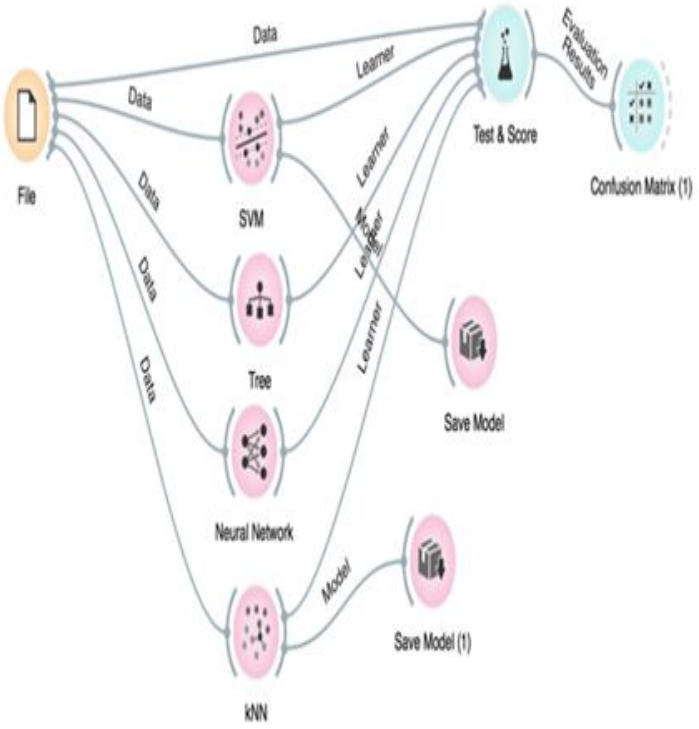

Figure 3: Machine Learning model using Orange3

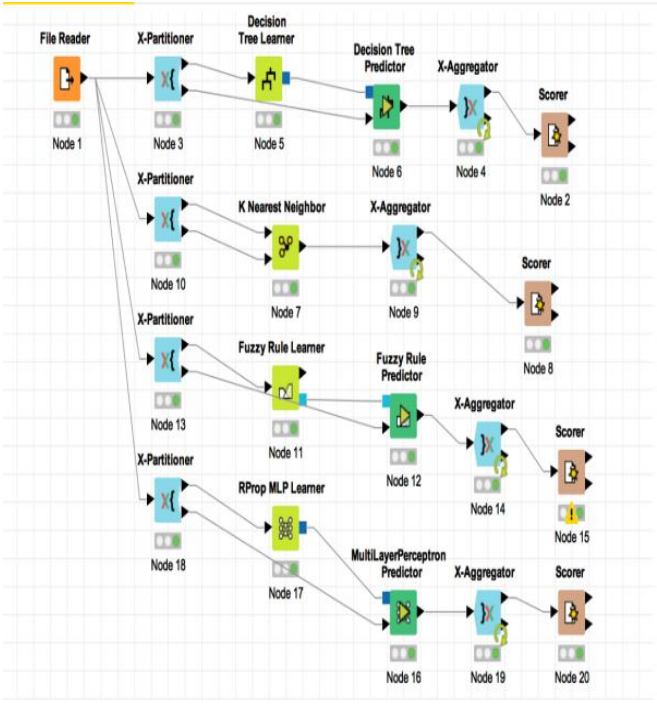

Figure 4: ML model using Knime

After using different tools to build machine learning model we conclude that Knime is much faster than orange3, but orange3 is user friendly and easier to use than Knime.

4. EXPERIMENT RESULTS

After extracting all Eigen features from images, we add it to real dimensions, (cap diameter, stem tall). Figure 5 shows the results of accuracy for Eigen with real dimensions in different machine learning techniques.

Test & Score

Evaluation Results

Method	AUC	CA	F1	Precision	Recall
kNN	0.944	0.896	0.899	0.906	0.896
Tree	0.849	0.878	0.876	0.875	0.878
SVM	0.583	0.775	0.676	0.600	0.775
Neural Network	0.598	0.751	0.703	0.689	0.751

Figure 5: Evaluation results for Eigen features with real dimensions

The experiment results shown KNN technique produced the highest accuracy (0.944) for Eigen with real dimensions. Because the dimensions are not easy to measure when we have only image for mushroom, we worked to extract virtual features from images by calculating heights and widths for mushroom images as it shown in Figure 6 below.

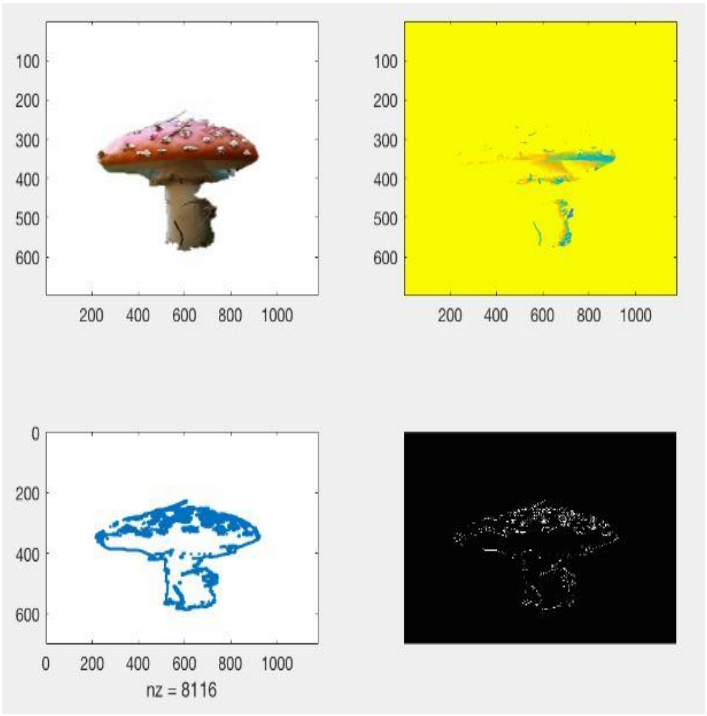

Figure 6: Width & height calculation

The results after calculating virtual dimensions from images and extracting Eigen features from images are shown in Figure 7 below. As we can see the best result obtained by KNN with accuracy of 87%.

Test & Score

Evaluation Results

Method	AUC	CA	F1	Precision	Recall
kNN	0.867	0.854	0.852	0.865	0.854
Tree	0.764	0.801	0.800	0.801	0.801
SVM	0.830	0.674	0.670	0.700	0.674
Neural Network	0.858	0.851	0.848	0.865	0.851

Figure 7: Evaluation results for Eigen features

Next step we tried to extract more features like histogram features, which applied to selected ML and shown the best accuracy gained by KNN technique with accuracy reached 87%. Figure 8 shows result for histogram features.

Evaluation Results

Method	AUC	CA	F1	Precision	Recall
kNN	0.874	0.864	0.862	0.872	0.864
Tree	0.732	0.783	0.783	0.783	0.783
SVM	0.876	0.843	0.841	0.852	0.843
Neural Network	0.845	0.815	0.814	0.816	0.815

Figure 8: Evaluation results for histogram features

We tried to add more features extracted from images which such as contrast, skewness, kurtosis, entropy, mean, standard deviation, energy, correlation and homogeneity. We called this group of features as parametric features. The experimental results for this group of features is shown in Figure 9.

Evaluation Results

Method	AUC	CA	F1	Precision	Recall
kNN	0.866	0.862	0.860	0.871	0.862
Tree	0.792	0.813	0.813	0.814	0.813
SVM	0.861	0.857	0.854	0.873	0.857
Neural Network	0.854	0.851	0.848	0.865	0.851

Figure 9: Evaluation results for parametric features

As we can see there is no enhancement on the results after extracting these features. In order to gain better accuracy, we used another scenario by using Photoshop to remove backgrounds for images, and repeat all previous steps, but this scenario failed to give higher accuracy.

Figure 10 shows the results for all techniques in proposed approach (Neural Network, SVM, Decision Tree, and KNN) with background images, while Figure 11 shows the results for the same techniques without background images.

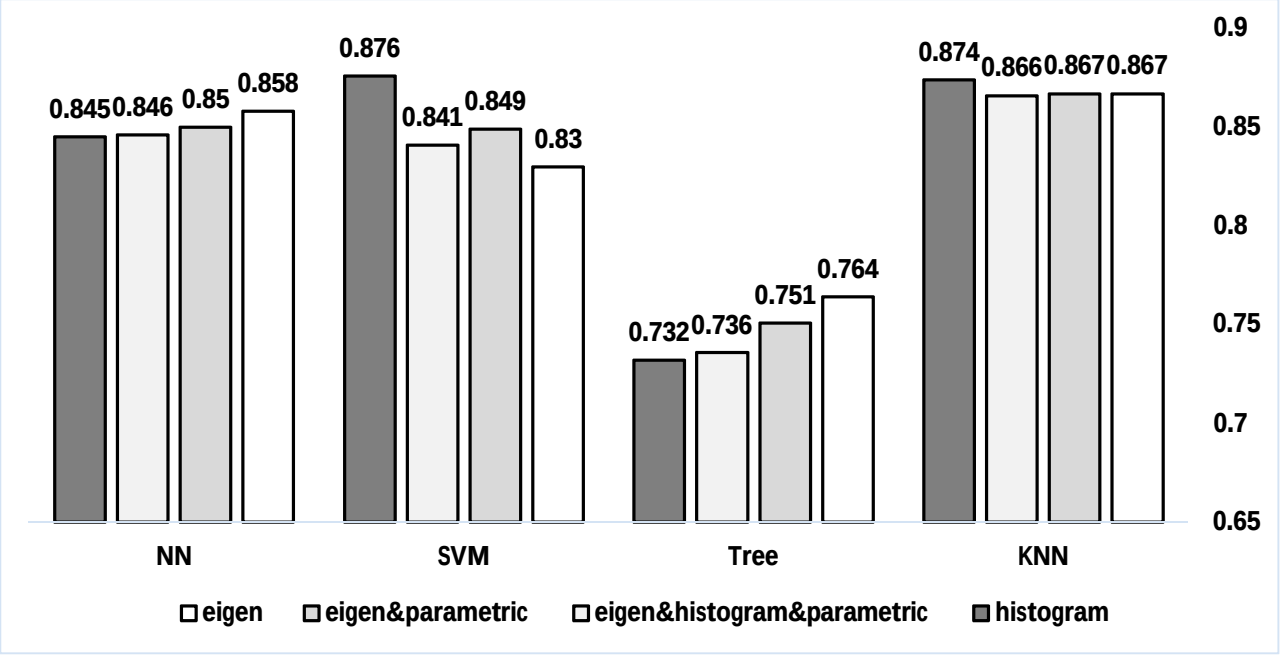

Figure 10: Evaluation results for images with background

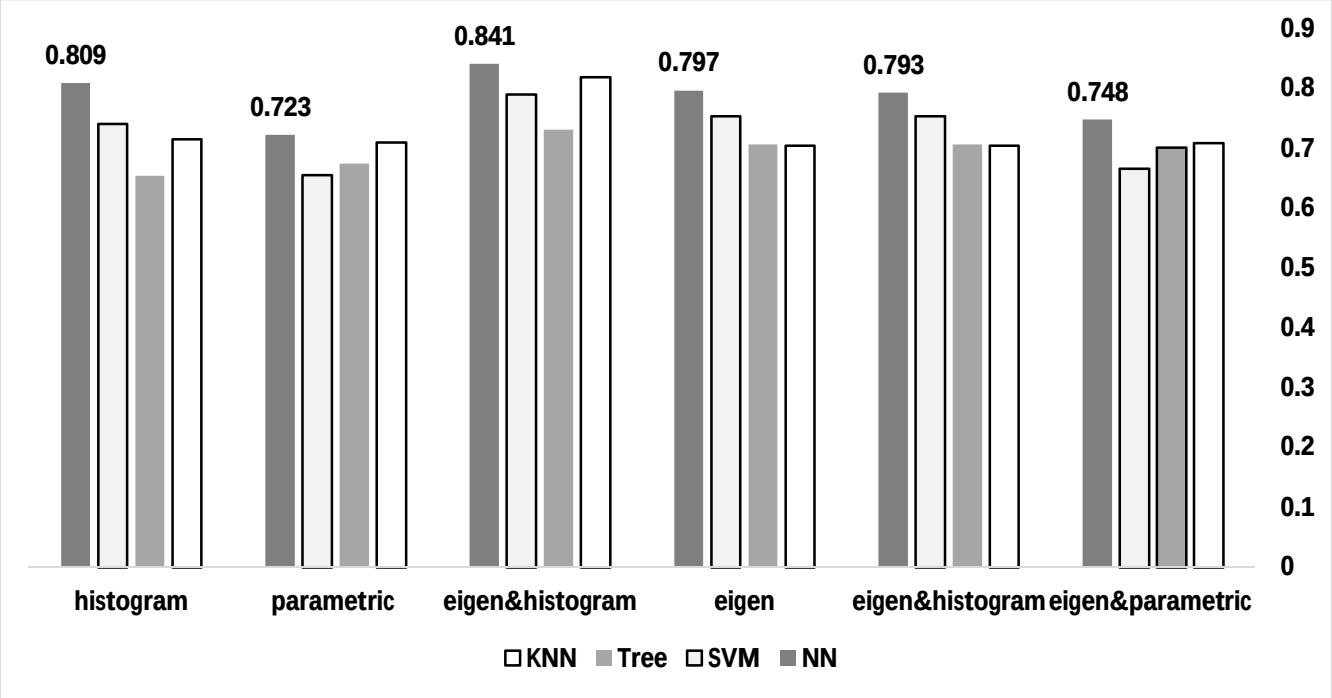

Figure 11: Evaluation results for images without background

5. CONCLUSION

In the proposed approach, we used different algorithms to get best results of mushroom classification, we implement each of neural network (NN), SVM, Decision Tree, and KNN on different scenarios, with background and without background. We extract different features from mushroom images like Eigen features, histogram features and parametric features. In order to improve the results, we remove images background but unfortunately this step failed to improve the result. Finally, the experiment results show advantage for background images, especially when used KNN algorithm, and with Eigen features extraction and real dimensions of mushroom (i.e cup diameter, stem tall and stem diameter) where accuracy reached to 0.944, while the result after replacing real dimensions with virtual dimension (i.e. width and height of mushroom shape inside the images) is 87%. The highest value for KNN after removing images background reached to 0.819 as a maximum value. Our future work we will try to extract some physical dimension from mushroom images like cup diameters, stem tall, color and texture. Also, we will try to expand the dataset and use more images to improve classification process.

REFERENCES

[1] I. Al-Mejibli and D. Hamed Abd, "Mushroom Diagnosis Assistance System Based on Machine Learning by Using Mobile Devices Intisar Shadeed Al-Mejibli University of Information Technology and Communications Dhafar Hamed Abd Al-Maaref University College," vol. 9, no. 2, pp. 103–113, 2017. https://doi.org/10.29304/jqcm.2017.9.2.319

[2] M. Alameady, "Classifying Poisonous and Edible Mushrooms in the Agaricus," *International Journal of Engineering Sciences & Research Technology*, vol. 6, no. 1, pp. 154–164, 2017.

[3] R. LaBarge, "Distinguishing Poisonous from Edible Wild Mushrooms," 2008.

[4] I. Kononenko, "Machine learning for medical diagnosis: history, state of the art and perspective," *Artificial Intelligence in medicine*, vol. 23, no. 1, pp. 89–109, 2001. https://doi.org/10.1016/S0933-3657(01)00077-X

[5] L. Von Ahn, B. Maurer, C. McMillen, D. Abraham, and M. Blum, "recaptcha: Human-based character recognition via web security measures," *Science*, vol. 321, no. 5895, pp. 1465–1468, 2008. https://doi.org/10.1126/science.1160379

[6] M. Tawarish and K. Satyanarayana, "A Review on Pricing Prediction on Stock Market by Different Techniques in the Field of Data Mining and Genetic Algorithm," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 3, no. 23–26, 2019. https://doi.org/10.30534/ijatcse/2019/05812019

[7] N. Bhargava and G. Sharma, "Decision Tree Analysis on J48 Algorithm for Data Mining," *International Journal of Advanced Research in Decision Tree Analysis on J48 Algorithm for Data Mining*, vol. 3, no. 6, pp. 1114–1119, 2013.

[8] A. Deshpande and R. Sharma, "Multilevel Ensemble Classifier using Normalized Feature based Intrusion Detection System," *International Journal of Advanced*

Trends in Computer Science and Engineering, vol. 8, no. 3, pp. 874–878, 2019.

[9] P. Kumar, V. K. Sehgal, D. S. Chauhan, and others, "A benchmark to select data mining based classification algorithms for business intelligence and decision support systems," *arXiv preprint arXiv:1210.3139*, 2012. https://doi.org/10.5121/ijdkp.2012.2503

[10] P. Babu, R. Thommandru, K. Swapna, and E. Nilima, "Development of Mushroom Expert System Based on SVM Classifier and Naive Bayes Classifier," *International Journal of Computer Science and Mobile Computing*, vol. 3, no. 4, pp. 1328 –1335, 2014.

[11] F. E. Onuodu, "K-Modes Clustering Algorithm in Solving Data Mining Problems for Mushroom Dataset," *nternational Journal of Advanced Research in Computer Science and Software Engineering*, vol. 5, no. 9, pp. 596–603, 2015.

[12] C. Eusebi, C. Gliga, D. John, and A. Maisonave, "Data Mining on a Mushroom Database," *Proceedings of Student-Faculty Research Day*, pp. 1–9, 2008.

[13] D. Chowdhury and S. Ojha, "An Empirical Study on Mushroom Disease Diagnosis : A Data Mining Approach," *International Research Journal of Engineering and Technology(IRJET)*, vol. 4, no. 1, pp. 529–534, 2017.

[14] S. Beniwal and B. Das, "Mushroom Classification Using Data Mining Techniques," *International Journal of Pharma and Bio Sciences*, vol. 6, no. 1, pp. 1170–1176, 2015.

[15] "Mushroom Dataset.", Retrevided from http://www.mushroom.world/.

[16] R. Socher, B. Huval, B. Bath, C. D. Manning, and A. Y. Ng, "Convolutional-recursive deep learning for 3d object classification," in *Advances in neural information processing systems*, 2012, pp. 656–664.

[17] G. Wang, Q. Song, H. Sun, X. Zhang, B. Xu, and Y. Zhou, "A feature subset selection algorithm automatic recommendation method," *Journal of Artificial Intelligence Research*, vol. 47, pp. 1–34, 2013. https://doi.org/10.1613/jair.3831

[18] F. Flores Camacho, N. Ulloa Lugo, and H. Covarrubias Martínez, "The concept of entropy, from its origins to teachers," *EDUCATION Revista Mexicana de Física E*, vol. 61, no. December, pp. 69–80, 2015.

[19] R. Socher and B. Huval, "Convolutional-recursive deep learning for 3D object classification," *Advances in Neural …*, no. i, pp. 1–9, 2012.

[20] C. Chang-yanab, Z. Ji-xian, and L. Zheng-jun, "Study on methods of noise reduction in a stripped image," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XXXVII. Pa, no. 1, pp. 2–5, 2008.