e-ISSN: 2582-5208

International Research Journal of Modernization in Engineering Technology and Science

(Peer-Reviewed, Open Access, Fully Refereed International Journal)

Volume:05/Issue:10/October-2023 Impact Factor- 7.868 www.irjmets.com

SOLVING MULTI-DIMENSIONAL IMAGE CLASSIFICATION AND OBJECT DETECTION PROBLEMS USING CLUSTERING AND A ‘STYLIZE-CLASSIFY’ METHOD

Ronav Jaiswal[*1], **Nilesh Jaiswal**[*2]

[*1]Student, The Bishop’s Co-Ed School Kalyani Nagar, Pune, Maharashtra, India.

[*2]Phd Scholar, Lakshmi Narain Institute Of Technology (Lnct), Bhopal, Madhya Pradesh, India.

DOI : https://www.doi.org/10.56726/IRJMETS45537

ABSTRACT

Machine learning based object detection and classification in digital images usually suffers from accuracy issues as the model often gets trained on background pixels and object textures that vary significantly between training and classification stages. In such cases the model identifies objects incorrectly and results in higher than acceptable false-positive and false-negative rates. In this paper the authors propose a novel stylize-cluster-classify approach to improve feature extraction and classification accuracy during object detection in images. The approach has been tested with a custom "Blind Objects" image dataset that consists of everyday objects encountered by people. This research has applications not just for visually impaired persons to identify objects accurately but also in related domains such as medical image prediction and novel body detection in astronomy images. In this approach the authors propose two additional stages in the machine learning workflow: (1) stylization and (2) clustering. Stylization transfers specific rendering styles to the dataset images in order to reduce histogram spread and pixel noise, in addition to improved detection of image features. The clustering stage improves classification accuracy by introducing an additional dimension of cluster identifier to the image feature vector. This additional dimension has a small negative effect on algorithm performance but results in a large improvement in accuracy.

Keywords: Machine Learning, Multi-Dimensional Classification, Stylizing, Clustering, Kmeans, Kmedians.

I. INTRODUCTION

In a world of technologies that are data-driven, we often encounter constraints imposed by the specific requirements of machine learning algorithms. These algorithms demand a certain quantity of data to yield accurate predictions. It's observed that supervised classifiers tend to offer precise predictions when dealing with smaller datasets, as they can accommodate simpler models. While this suffices for a substantial number of datasets, the expanding role of Machine Learning in diverse sectors such as healthcare and travel has brought to light the prevalence of more challenging scenarios. In these cases, data collection is arduous and resource-intensive. For instance, gathering health data related to conditions like Cancer or Down’s Syndrome is an intricate task. Similarly, predicting the behaviours of autonomous vehicles under certain road-conditions faces a lack of historical data, and similarly for acquiring images for object detection and computer vision tasks.

The challenge intensifies in such classification tasks due to the multidimensional nature of the data. In these scenarios, each instance is characterized by multiple class variables, and these variables may have direct or indirect connections with the associated label. Let's denote d as the number of class variables associated with each instance. To illustrate, when dealing with images, every pixel is considered a distinct class variable. For instance, in the case of a 16 × 16 pixel image, d equals 256. Similarly, in the domain of sport analytics, there are numerous parameters like statistics, pitch records, player speed, outfield conditions, injuries, and more, which contribute to a significantly large d. Likewise, tasks related to road performance in autonomous vehicles involve multiple criteria, and healthcare datasets comprise sensor readings, CT and MRI scans, among others, resulting in complex, multi-dimensional datasets. When such datasets necessitate a classification algorithm, they give rise to what is known as a multi-dimensional classification problem, referred to as Image MDC in this paper.

The authors introduce a novel framework with two additional steps in the machine learning workflow – Stylizing and Clustering. The additional steps introduce computational overhead, however the benefits in terms of improvements in accuracy far outweigh the cons. In the context of object detection within images, most state-of-the-art algorithms suffer from poor accuracy due to reliance on object textures rather than object shapes. The 'Stylize-Cluster-Classify' approach introduced in this paper, hypothesizes that images, when transformed into their stylized forms and processed through clustering algorithms before they are classified, are significantly more accurate to identify features from, in terms of object-detection. The authors evaluated multiple stylizing techniques in order to identify the one best suited for this purpose. Stylization reduces the pixel distribution complexities in images, making them more suitable for feature extraction. It also reduces the overall dimensionality of the feature vector, thereby improving performance.

This research paper focuses on tackling Image MDC (Multi-Dimensional Classification) problems that typically involve datasets with a relatively limited number of instances. This approach is applicable when dealing with datasets characterized by d features, and it allows for the condensation of these features into a reduced set of i (cluster ID) values. This reduction occurs when there is a set $c = \{c1, c2, \dots ci\}$ where $i \geq 2$, and c represents the centroids obtained through the clustering process.

The cluster is added as another dimension in the feature vector before the classification stage. This has minimal effect on the computational efficiency, but drastically improves accuracy.

In this study, various combinations of three stylization methods and four classification algorithms were tested, and their performance was assessed by calculating the average accuracy after a 6-fold cross-validation. These outcomes were then contrasted with the results obtained by applying classification directly to the Image MDC datasets, without involving any clustering or stylization, and using the raw data as input to the classifier.

When considering the Blind Objects dataset, it was observed that in the case of three out of four algorithms the approach of stylization combined with clustering yielded superior accuracies and exhibited lower standard deviations. Notably, the most impressive accuracy was achieved with the combination of Photocopy Stylized and the Decision Tree algorithm, boasting an average accuracy of μ = 85.55% and a standard deviation of σ = 2.43.

II. LITERATURE REVIEW

In the existing body of research, various strategies have been proposed to address MDC (Multi-Document Classification) problems. Read et al. have explored the utilization of Bayesian networks as an alternative to traditional classifiers for clustering MDC datasets. Their approach involves employing chained classifiers that capture the relationships between each feature and the data instances, including the inter-feature correlations. [17] This method aims to address overfitting and underfitting issues through algorithmic adjustments, without altering the fundamental nature of the data.

X. Huang et. Al have presented a highly effective method, that enables real-time arbitrary style transfer. It introduces an adaptive instance normalization (AdaIN) layer that aligns the mean and variance of content features with those of style features. The method also allows users to have fine-grained control over various aspects, including content-style balance, style blending, and color and spatial adjustments.[24]

Hae-Gon Jeon et al. have presented a convolutional neural network that is being used in scenarios of 'extremely changing conditions' for visual localization. In this CNN-based approach, they have simulated a disaster environment for the changing conditions. Using various disaster scenes, they have evaluated the effectiveness of their algorithm. They have used stylizing as their technique to boost shape representations in their algorithm. Using 'dominant plane information' they are able to predict CNN-based visual localizations in 6-DoF camera poses. This method is resulting in more reliable camera pose predictions in the changing conditions. [21]

In a related research domain, Haider et al. have introduced the penalty method, which transforms the challenge of deep learning classification from a problem with constraints into one without constraints. They achieve this by incorporating penalties in the cost function for constraint violations, effectively mitigating the issue of underfitting. This method aims to modify algorithms to tackle overfitting and underfitting while leaving the underlying data structure unchanged. [6]

Mitelpunkt et al. have implemented a strategy called 'categorize, cluster, classify' in the context of medical applications. They have demonstrated success in a framework where data is initially categorized and then the centroids of these categories are classified. However, their algorithm is primarily employed to cluster patient subtypes based on variations in disease onset and the involvement of different genetic markers in symptom manifestation [14] .In contrast, our framework takes data that has clear physical interpretations and removes these physical definitions through clustering. This unique approach grants the model the flexibility to establish connections that may not be readily comprehensible to the user, essentially creating a 'black box' that generates labels for data points.

Another interesting paper and study was carried about by Ying et al. They outline a data expansion technique to combat overfitting by enhancing the relationships between features and diminishing the influence of outliers. [8] Additionally, Jia et al. introduce feature manipulation, a direct modification of the dataset that involves identifying correlations between features and incorporating them as additional features. This is done to enhance the differentiation between labels assigned to data instances [1] While these strategies have exhibited success across various datasets, it's important to note that extensive data expansion, especially in the case of smaller datasets, can compromise data quality. Feature augmentation introduces computationally demanding processes that lack built-in safeguards against overfitting and underfitting. Additionally, the introduction of more parameters can easily lead to these errors, resulting in a more complex MDC dataset that is challenging to classify.

III. MATERIALS AND METHODS

3.1. Data Description:

To validate the results obtained through the 'Stylize-Cluster-Classify' approach, the researchers utilized a specially created 'Blind Objects' dataset. This dataset comprises images of common objects that are typically used by blind individuals. It consists of 25 different labels, and each label was trained with roughly 100 images. The images were personally captured by the authors, encompassing a range of lighting and environmental conditions.

3.2. Experiment Design:

The proposed experiment aims to prove the validity of the 'Stylize-Cluster-Classify' approach on image MDC datasets. The average accuracy and standard deviation after 6-fold cross validation for various clustering and classification combinations, along with direct classification without clustering, are compared. Confusion matrix analysis is employed to better make sense of the approach used.

3.3. Experiment Methods:

The following layout is used for conducting the experiments: At the beginning of each experiment, the test subject (blind person) stands in front of the table. The test subject then follows instructions from the experiment to reach out, identify and pick up objects from the table.

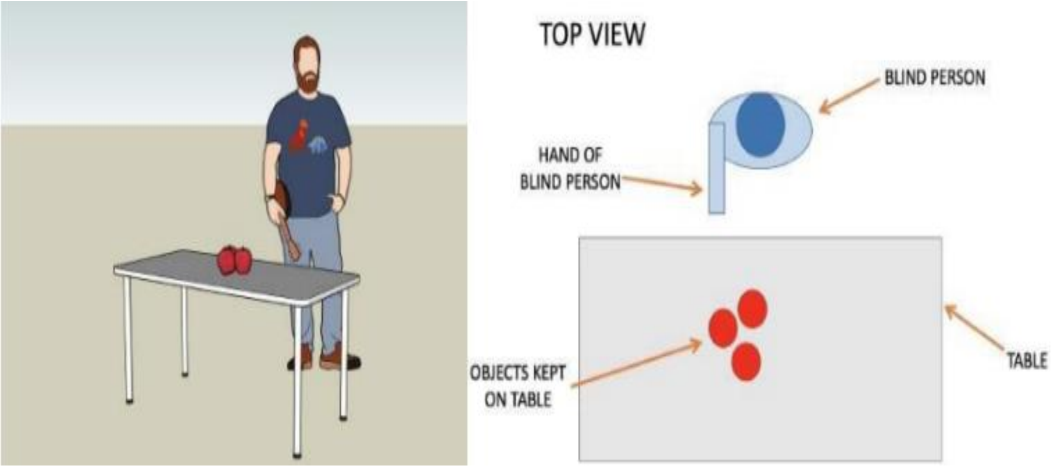

Fig. 1 The experiment methodology carried out

3.4. Experiment Vs. Control

In the experiment, the test subject (blind person) performs the experiment wearing the experimental device Lakshya [22] on their hand. In the control, the test subject performs the experiment without wearing the device.

3.5. Algorithm Design:

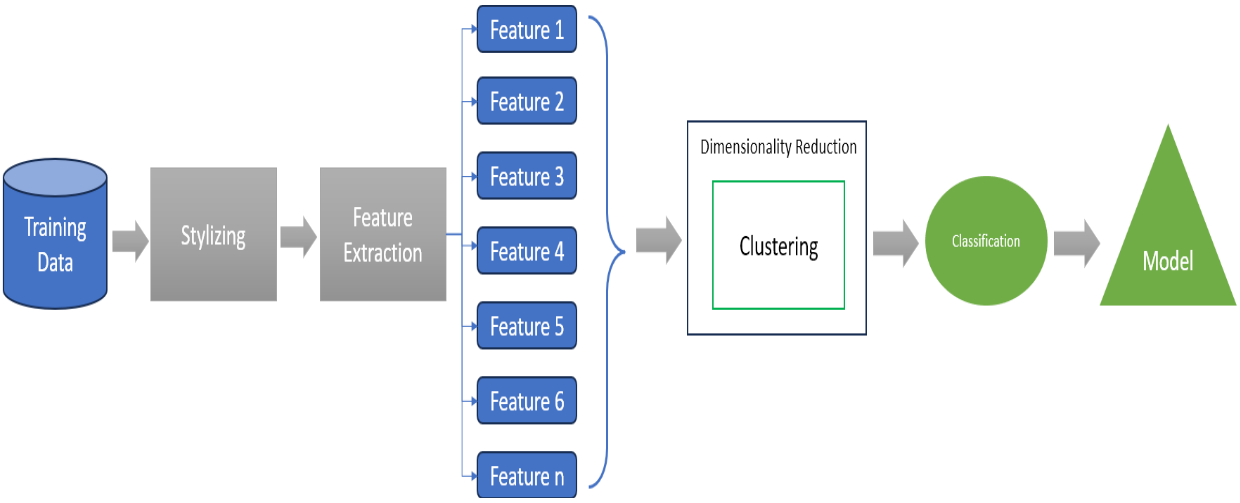

Fig. 2 The architecture of the ‘Stylize-Cluster-Classify’ approach

The Figure 2 concisely describes the ‘Stylize-Cluster-Classify’ approach, in five major stages. Feature selection is included in this diagram and has been performed to increase accuracy. The major stages of the algorithm are Pre-processing (Stylizing and Feature extraction), Dimensionality Reduction (Clustering), and Classifying.

3.6.1. Stylizing Stage:

Stylizing as an image pre-processing step is extremely useful in simplifying images by reducing noise and other distracting features from an image, which don’t contribute as relevant features at the extraction stage. The authors, later, used saliency maps to further ensure that background details don’t contribute to the relevant features necessary for classification.

Using Microsoft Office tools the images have been manually stylized, for the sake of research and experimentation, to test the accuracy of stylization techniques. Below, are some examples of images from the ‘Blind Objects’ dataset, along with their stylized versions:

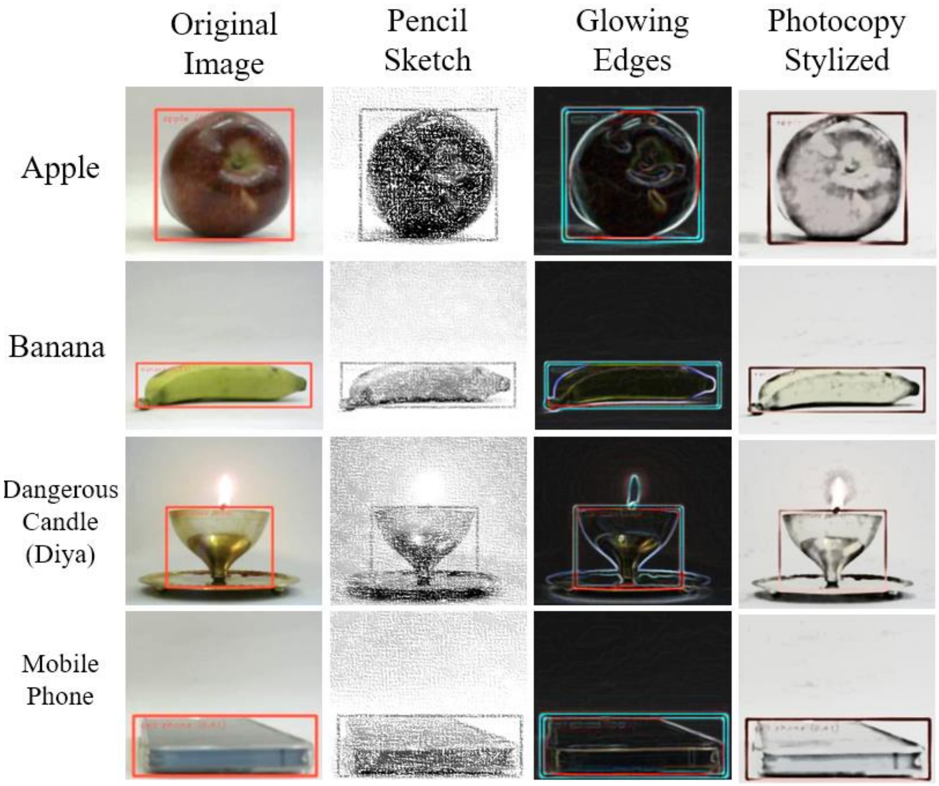

Fig. 3 Various Stylization techniques employed on an array of objects from the ‘Blind Objects’ Dataset

3.6.2. Feature Extraction Stage:

The authors selected features based on the following criteria, in order to ensure that the model provides the most accurate predictions with a wide array of objects:

1. **Independence** (a change in one feature should not change the value of another feature significantly)
2. **Discriminatory** (each feature should have a significantly different value for each different object)
3. **Reliability** (feature should have the same value for all objects in the same class/group)

Features such as Area and Perimeter are rejected because they are dependent on the camera zoom, having low discriminatory value. Some of the significant features that have been selected for use are:

1. Average Colour (Mean pixel value of channels)
2. Rectangularity
3. Circularity
4. Aspect Ratio
5. Edge Features (Edge detection on grayscale image)
6. Histogram of Oriented Gradients (localized HOG)
7. SIFT (Scale-Invariant Feature Transform)

Saliency maps were then used to make sure that background details aren't being used as primary features for classification.

3.6.3. Scaling Stage:

Classifying data doesn't demand the uniformity of units, but for clustering algorithms, the numerical values of features are significant. This necessitates the standardization of the data to ensure that no feature is inadvertently given undue weight due to varying units. Standardization levels the playing field, ensuring that all features are on an equal footing, and it allows the clustering algorithm to allocate weights based on the relative correlations between the features and labels.

To achieve this, non-numeric data is transformed into numeric format, and numeric data is scaled using z-scores, independently for each feature (column-wise). This process can be executed using the provided function.

$$z = \frac{x - \mu}{\sigma} \qquad (1)$$

In this context, $x \in X$, where X represents the data series, and each x denotes a value within that series. μ represents the mean of X, σ is the standard deviation of X, and z stands for the z-score for a particular feature x within the instance X. Let's take it one step further. Suppose there are i instances, then we have a set $z = \{z_1, z_2 \dots z_i\}$. This set z contains the standardized values for the first feature, as standardization is performed column-wise. If there are N features in the dataset, this standardization process is repeated N times.

This standardization procedure scales all values that are one standard deviation away from the mean to a range of [-1, 1], which results in more extreme values being assigned higher or lower values. After standardizing the datasets, only the features from all instances are extracted and assembled into a data frame referred to as $X°$, while the labels of the instances are extracted into a series y. It's important to note that the labels are not subjected to standardization or any additional processing. Consequently, DS standardized is organized as an $N \times i$ matrix and can now be employed for the clustering stage, while y is set aside for later use.

3.6.4. Clustering Stage:

The standardized dataset, DS standardized, is now employed in clustering algorithms with the aim of reducing dimensionality while retaining the correlations between features and labels. This stage allows for the utilization of various clustering algorithms, although this study places emphasis on centroid-based clustering algorithms due to their ability to produce a set of features for subsequent analysis. Specifically, KMeans and KMedians were chosen for comparison and implemented in Python.

KMeans, an unsupervised learning algorithm, partitions an N×1 matrix into k centroids. Its primary objective is to enhance the separation between clusters and the similarity of data points within clusters. This is achieved by minimizing the L2 norm distance and forming k distinct sets, each centred around a centroid. The distance metric used in KMeans is the Euclidean Distance, which computes the squared distance between each data point and the assigned centroid. Initially, the algorithm randomly assigns centroids, and then it reassigns data points to the nearest centroid. For each element $r_i \in r_k$, where r_k represents all points belonging to a cluster, r_i may be reassigned if it minimizes the objective function. [8, 11]

$$J_a = \sum_{n=1}^{k} \sum_{r_i \in r_k} \| r_i - c_n \|_2^2 \quad (2)$$

The calculation for the new centroids is done by:

$$c_k = \frac{1}{|r_k|} \sum_{r_i \in k} r_i \quad (3)$$

KMedians is another unsupervised learning algorithm capable of partitioning an N×1 matrix into k centroids. Its main goal is to minimize the L1 norm distance between each data point and its respective centroid by forming k distinct sets, where the median of each set, denoted as c_k, serves as the centroid. In KMedians, the distance metric used is the Manhattan distance, which is calculated as the sum of the vertical and horizontal distances between two points. Similar to KMeans, the KMedians algorithm initiates with the arbitrary assignment of centroids, followed by the allocation of data points to their nearest centroid. In this process, for each element $r_i \in r_k$, where r_k encompasses all points belonging to a specific cluster, r_i is repositioned if it minimizes the objective function. [12, 13]

$$J_b = \sum_{n=1}^{k} \sum_{r_i \in r_k} \| r_i - c_n \|_1 \quad (4)$$

The calculation for the new centroids is done by:

$$c_k = \underset{y \in r_k}{\arg\min} \sum_{i=1}^{m} \| x_i - y \|_2 \quad (5)$$

Both algorithms converge when no elements are reassigned, which implies that r_k remains constant for all centroids $c = \{c1, .. ck\}$. These algorithms are applied individually to each instance within the dataset. For every row $x \in$ DS standardized, these algorithms operate on the series of N features in x, ultimately reducing them to k centroids. Consequently, DS standardized is transformed into a $k \times i$ dataset, where 'i' represents the number of instances.

The optimal value of k is determined using the Elbow method. This involves plotting the distortion against various values of k, typically ranging from 1 to 50. The point on the graph where a distinct 'Elbow' is observed is considered the ideal cluster value. Selecting the right value for k is crucial to maintain low distortion between the actual features and the centroids. An incorrect value of k can lead to data loss or the inclusion of noisy, unnecessary data, which can adversely affect the model's performance.

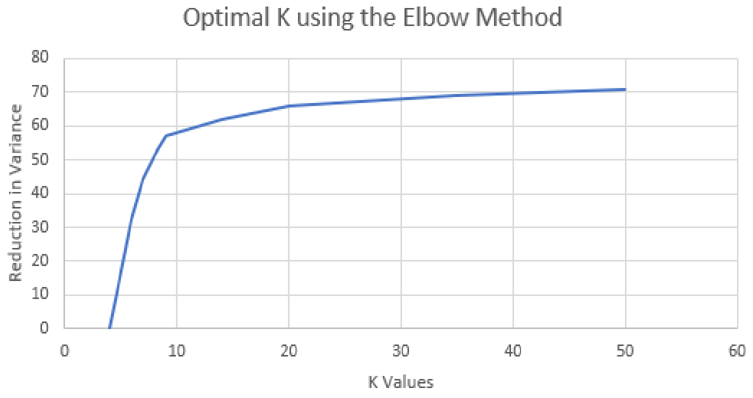

Fig. 4 Elbow curve for the optimal k = 9 of the Blind Objects Dataset

Yaswanth Kumar Alapati et al. have followed the reduction of dataset dimensionality, the next step is to employ clustering algorithms on the reduced dataset. Subsequently, the cluster identifier is incorporated into the dataset. [23]

3.7. Classification Phase:

After standardizing, stylizing, clustering, and organizing the data, we can employ classification algorithms to assign labels relevant to our specific problem. We divide the dataset Xs into two subsets: X_training, which contains 70% of the data for training, and X_testing, which includes 30% of the data for testing. The corresponding labels, denoted as y, are also split into Y_training and Y_testing in the same 70-30 proportion. To develop and evaluate our classification models, we utilize the training data (X_training and Y_training) and implement the selected four classification algorithms using SciKit Learn modules.

Decision Tree: The decision tree algorithm is hierarchical in nature, and it divides a dataset into sub-datasets (nodes) based on rules derived from the characteristics of the data points. This process repeats iteratively until a "leaf" is reached, representing a sub-dataset that can be assigned to a specific label. All data points are then categorized into these labels, which are subsequently provided as predictions. [9, 15]

K Nearest Neighbours: The K Nearest Neighbours algorithm involves comparing test instances with all training instances and selecting the closest data point based on their feature values. The label of this nearest neighbour is then assigned to the test instance, serving as the predicted label. [9]

Naïve Bayes: The Naïve Bayes algorithm is rooted in Bayes' theorem of probability. It assumes strong independence among all variables and calculates the likelihood of a relationship existing between them. Using this information, it categorizes data instances into labels, generating predictions. [9, 12]

Support Vector Machine (SVM): The Support Vector Machine is designed to create a hyperplane in the decision plane, enabling it to separate data instances into their respective labels based on their features. This separation is achieved using a technique called Structural Risk Minimization (SRM), ultimately providing a predicted label. [9, 7]

Each of the classification algorithms underwent a six-fold cross-validation process. This means that each model was executed six times, each time with distinct allocations for training and testing sets. After obtaining predictions for the dataset X_testing from the classifier, these predictions were compared to the actual labels stored in Y_testing. The utilization of 6-fold cross-validation ensures the attainment of statistically significant results, and it's worth noting that p-values were not employed, as Dietterich et al. pointed out that t-tests are not suitable when combined with k-cross validation. [5]

3.8. Criteria of Evaluation:

Every classification gives us a confusion matrix. These generally include false negatives (FN), false positives (FP), true negatives (TN), and true positives (TP).

Table 1. The accuracies and standard deviations of the 20 combinations on the 'Blind Objects' dataset. The cells highlighted in green are the highest scoring models within each category.

Right - Classification Algorithms **Down -** Stylizing Methods	**Metric**	**Decision Tree**	**K Nearest Neighbours**	**Naïve Bayes**	**Support Vector Machines**
Pencil Sketch Stylized	**Accuracy**	81.37%	71.59%	77.71%	72.41%
	Standard Deviation	7.56%	2.82%	3.32%	0.12%
Photocopy Stylized	**Accuracy**	85.55%	63.21%	73.45%	72.54%
	Standard Deviation	2.43%	0.40%	0.52%	0.11%
Non-Stylized	**Accuracy**	61.42%	63.42%	63.33%	67.11%
	Standard Deviation	4.60%	5.43%	5.62%	5.23%
Glowing-Edges	**Accuracy**	72.69%	68.29%	81.27%	77.81%

Stylized	**Standard Deviation**	1.74%	0.91%	0.40%	0.71%
Average of Photocopy and Glowing-Edges	**Accuracy**	79.12%	65.75%	77.36%	75.18%
	Standard Deviation	2.10%	0.65%	0.46%	0.41%

The evaluation metrics for each combination were assessed by executing the algorithms. Initially, after the completion of cross-validation, all 20 combinations were averaged. Then, various algorithms were applied to both datasets, and their accuracy evaluation metrics were computed. The best-performing combinations were identified, and standard deviations were recorded from five runs.

IV. RESULTS AND DISCUSSION

The table illustrates that the Stylize-Cluster-Classify approach consistently exhibits superior accuracy and lower standard deviations across most combinations for both datasets. In the case of the Blind Objects dataset, the highest accuracy was achieved by the combination of Photocopy Stylized and Decision Tree algorithms, with an average accuracy(μ) of 85.55% and a standard deviation(σ) of 2.43%. Notably, Decision Tree emerged as the most successful algorithm with an average accuracy of 75.26% and a standard deviation of 4.16. Moreover, stylizing an image almost always proved to improve accuracies as compared to non-stylized images from the table.

From each dataset, the top three models with the highest scores were selected for further analysis, involving the utilization of the five predefined evaluation metrics. These evaluation metrics were averaged over the six runs of k-fold cross-validation before final calculations. In this context, a higher value for Accuracy, Precision, Sensitivity, Specificity, and a lower False Positive Rate are indicative of a superior model.

Table 2. The results of the three highest scoring combinations on 5 evaluation metrics, for the 'Blind Objects' dataset.

Model Performance	**Naïve Bayes: Glowing-Edges Stylized**	**Decision Tree: Photocopy Stylized**	**Decision Tree: Pencil Sketch Stylized**
	AVG	**AVG**	**AVG**
Accuracy (%)	81.27%	85.55%	81.37%
Precision (%)	87.59%	90.19%	85.43%
Sensitivity (%)	88.13%	90.96%	87.44%
Specificity (%)	63.02%	64.59%	68.31%
FPR (%)	49.10%	43.83%	38.60%

The choice of the best classification algorithm appears to depend on the specific dataset, but overall, the 'Stylize-Cluster-Classify' approach leads to enhanced accuracy. Specifically, there was an impressive 13.79% improvement averaged in accuracy observed on the 'Blind Objects' dataset.

The remarkable efficacy of the 'Stylize-Cluster-Classify' approach, as opposed to direct classification, can be attributed to the existence of inter-feature relationships within the datasets under consideration. The fundamental premise behind multiple features is rooted in the recognition that all features contribute to the label, and assuming that these features are mutually exclusive or do not interact is a misconception.

The significance of inter-feature relationships is exemplified across diverse datasets, and this methodology holds potential for validation across a range of other datasets. Typically, in any multidimensional classification dataset, features exhibit relationships with one another, making this approach particularly valuable. However, it's essential to note that the 'Stylize-Cluster-Classify' approach is not suitable for datasets with an extremely small number of features, as clusters may have very few points associated with them, rendering advanced analysis unfeasible.

Furthermore, the same 'Blind Objects' dataset was subjected to classification using Decision Trees, Naïve Bayes, and SVM, resulting in accuracies of 63.33%, 63.42%, and 67.11%, respectively. In contrast, the 'Stylize-Cluster-Classify' approach achieved accuracies as high as 85.55%. Thus, the 'Stylize-Cluster-Classify' approach enhances accuracy by mitigating overfitting, as demonstrated by the comparison with non-stylized images.

Future research should explore larger datasets, investigate applications in various industries, and incorporate more intricate algorithmic nuances to enhance the efficiency of algorithms and refine the methodology. Exploring a wider array of classification and clustering combinations, potentially adapting these algorithms to leverage the 'Stylize-Cluster-Classify' method, could yield intriguing theories and potentially lead to the development of novel algorithms for addressing machine learning challenges using two-part models. Furthermore, the integration of neural networks within this approach for clustering holds promise.

Another avenue for exploration involves developing tests to identify intra-feature relationships, allowing for a more precise determination of when the 'Stylize-Cluster-Classify' approach proves advantageous. It's worth noting that one major limitation of this study lies in its reliance on a relatively small number of datasets for validation. Consequently, conducting further applications and testing across a broader range of datasets would be invaluable for generalizing the results.

V. CONCLUSION

The ‘Stylize-Cluster-Classify’ method has not been substantially investigated as an approach to improve efficacy of object detection and classification problems before, and thus exists as an independent body of work within this space. In this paper, we substantiated the state-of-the-art workflow of classification with the introduction of additional Stylization and Clustering stages. The successful results obtained by *k*-fold cross-validation demonstrate a lot of promise in this approach for not just object detection but also in related domains such as medical image prediction and novel body detection in astronomy. It has been consistently observed that stylizing images results in an increase in efficacy. The hypothesis of the approach that improved efficacy could be obtained by introducing stylizing and clustering stages is thus verified. The situations where ‘Stylize-Cluster-Classify’ is applicable as well as the reason for the increase in accuracy levels are also discussed. The authors believe that this approach needs further research that could help create a stronger basis for scaling this procedure to additional real-world applications, and demonstrate tangible, positive results.

ACKNOWLEDGEMENTS

The authors would like to thank the Niwant Andh Mukta Vikasalay for helping them carry out research with the visually impaired students and Mr. Ashley Eates, Principal of The Bishop’s Co-Ed School Kalyani Nagar, Pune for his guidance and support while pursuing this research work.

VI. REFERENCES

[1] Jia, Bin-Bin, and Min-Ling Zhang. "Multi-dimensional classification via kNN feature augmentation." Pattern Recognition 106 (2020): 107423.

[2] Dabbura, Imad. "K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks. 2018." (2021).

[3] Bernholdt, David E., Mark R. Cianciosa, Clement Etienam, David L. Green, Kody JH Law, and Jin Myung Park. "Cluster, Classify, Regress: A General Method For Learning Discontinuous Functions." arXiv preprint arXiv:1905.06220 (2019).

[4] Khoury, Nicolas, Ferhat Attal, Yacine Amirat, Latifa Oukhellou, and Samer Mohammed. 2019. "Data-Driven Based Approach to Aid Parkinson’s Disease Diagnosis" Sensors 19, no. 2: 242.

[5] Dietterich, Thomas G. "Approximate statistical tests for comparing supervised classification learning algorithms." Neural computation 10, no. 7 (1998): 1895-1923.

[6] Jabbar, H., and Rafiqul Zaman Khan. "Methods to avoid over-fitting and under-fitting in supervised machine learning (comparative study)." Computer Science, Communication and Instrumentation Devices 70 (2015).

[7] Byun, Hyeran, and Seong-Whan Lee. "Applications of support vector machines for pattern recognition: A survey." In International workshop on support vector machines, pp. 213-236. Springer, Berlin, Heidelberg, 2002.

[8] Ying, Xue. "An overview of overfitting and its solutions." In Journal of physics: Conference series, vol. 1168, no. 2, p. 022022. IOP Publishing, 2019.

[9] D. Gong, "Top 6 Machine Learning Algorithms for Classification | by Destin Gong | Towards Data Science," Medium, Jul. 12, 2022. [Online].

[10] LaValley, Michael P. "Logistic regression." Circulation 117, no. 18 (2008): 2395-2399.

[11] Fedesoriano, "Stroke Prediction Dataset," 2020. [Online]. Available:
https://www.kaggle.com/datasets/fedesoriano/stroke-predictiondataset. [Accessed: Aug. 14, 2022]

[12] Vembandasamy, K., R. Sasipriya, and E. Deepa. "Heart diseases detection using Naive Bayes algorithm." International Journal of Innovative Science, Engineering & Technology 2, no. 9 (2015): 441-444.

[13] M Yasser H, "Wine Quality Dataset," Jan., 2022. [Online]. Available:
https://www.kaggle.com/datasets/yasserh/wine-quality-dataset. [Accessed: Aug. 14, 2022]

[14] Mitelpunkt, A., Galili, T., Kozlovski, T. et al. Novel Alzheimer's disease subtypes identified using a data and knowledge-driven strategy. Sci Rep 10, 1327 (2020).

[15] Quinlan, J. Ross. "Learning decision tree classifiers." ACM Computing Surveys (CSUR) 28, no. 1 (1996): 71-72.

[16] Nwusu, Chidozie Shamrock, Soumyabrata Dev, Peru Bhardwaj, Bharadwaj Veeravalli, and Deepu John. "Predicting stroke from electronic health records." In 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 5704-5707. IEEE.

[17] Read, Jesse, Concha Bielza, and Pedro Larrañaga. "Multi-dimensional classification with super-classes." IEEE Transactions on knowledge and data engineering 26, no. 7 (2013): 1720-1733.

[18] C. Whelan, G. Harrell, and J. Wang, "Understanding the K-Medians Problem - ProQuest," Understanding the K-Medians Problem ProQuest, 2015.

[19] M. Helm, "Use this clustering method if you have many outliers | by Martin Helm | Towards Data Science," Medium, Aug. 06, 2021. [Online]

[20] Biau, Gérard, and Erwan Scornet. "A random forest guided tour." Test 25, no. 2 (2016): 197-227.

[21] Hae-Gon Jeon , Sunghoon Im , Jean Oh and Martial Hebert. "Learning Shape-based Representation for Visual Localization in Extremely Changing Conditions" . ICRA 2020

[22] R. Jaiswal. "Lakshya: Intelligent Device to Help Visually Impaired (Blind) People Visualize and Interact with Objects and People Around Them Using Ai/Ml Models of Google Tensor-Flow" - International Journal for Multidisciplinary Research (IJFMR) 2023

[23] Yaswanth Kumar Alapati et al. "Combining Clustering with Classification: A Technique to Improve Classification Accuracy" IJCSE16-05-06-026

[24] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in Proc. of IEEE International Conference on Computer Vision (ICCV), 2017.