Investigating Degradation Levels in Palm Leaf Document Images through Clustering techniques: Unveiling Inherent Patterns

Potula Radha Nishant
Department of Computer Science and Engineering
Amrita School of Computing, Bengaluru
Amrita Vishwa Vidyapeetham, India
bl.en.u4cse21161@bl.students.amrita.edu

Rejeti Kartik
Department of Computer Science and Engineering
Amrita School of Computing, Bengaluru
Amrita Vishwa Vidyapeetham, India
bl.en.u4cse21170@bl.students.amrita.edu

Sanjay Baitha
Department of Computer Science and Engineering
Amrita School of Computing, Bengaluru
Amrita Vishwa Vidyapeetham, India
bl.en.u4cse21185@bl.students.amrita.edu

***Abstract*—Palm leaf manuscripts are important for historical text and hold significant importance for the connection to the past and hence need to preserved. The script in the degraded palm leaf manuscript undergoes extraction of the characters, conversion into readable words and sentences and prediction for which type of degradation status of each image. This will ultimately be used to make a model that can accurately predict the characters in the document. Classification models are trained with the best hyper-parameters using a grid search method. The features were selected based on the co-relation factor and the similar features were dropped. Then, different clustering methods K-Means, Agglomerative, and others were trained on the dataset along with training of different classification models and the results were compared to yield the better results. Finally, an image analysis is done on basis of the type of degradation in the manuscript.**

***Index Terms*—Degradation, Classification, Grid Search, Clustering, K-Means, Agglomerative**

I. Introduction

Documents that are of historical importance are meant to be preserved for coming generations. The older the document, the less likely they remain in good condition. There can be many factors that could lead to these situations, such as the natural environment which slowly degrades the quality of the document, and manmade causes due to invasion, war, arson or purposefully destroying them. Our goal is to make sure that these documents can be preserved in our digital age and make them available to everyone. Palm leaf manuscripts are very old documents, where some of the oldest documents were recorded before the invention of the printing press. These artifacts have some of the oldest writings in our culture and their scripts are very old compared to our modern counterparts. These naturally degrade over time and sometimes are difficult to read after centuries of wear and tear.

Image processing is a process in which an image is taken, and analyze it using a model (pixel by pixel) and use that to identify the object, character, or background of the image. Palm leaf characters or glyphs are variably difficult to analyze depending on how preserved it is. An image processor is used to analyze the glyphs, find patterns in the way the characters are written, and classify them accordingly. This is where the uncertainty comes in, when the manuscript is partly destroyed or illegible. Here, the best that can only be made are partial guesses and solve what character might be.

II. Literature Review

Windu Antara Kesiman et.al. [1] propose a spatial detection and glyph recognizer K -means Clustering and Neighborhood Pixels Weight Neural Networks (NN) and Unsupervised Feature Learning (UFL) on Balinese characters or glyphs on the AMADI_LontarSet dataset and these were the results that are achieved. The ASC class recognizer consists of 7 glyphs, consisting of 860 training data and 921 testing data, has a 92.73% accuracy using the NN model, and a 93.16% accuracy using the UFL NN model. The DESC class recognizer consists of 1,860 glyphs, consists of 1,860 training data and 593 testing data, has an 85.84% accuracy with the NN model, and an 88.03% accuracy with the UFL NN model. The BASE class recognizer consists of 49 glyphs, consisting of 5,070 training data and 4,392 testing data, has an 87.46% accuracy with the NN model, and an 87.43% accuracy with the UFL NN model. The ASC-BASE class recognizer consists of 16 glyphs, consisting of 1,170 training data and 208 testing data, has a 75.48% accuracy with the NN model, and a 75.96% accuracy with the UFL NN model. The BASE-DESC class recognizer consists of 40 glyphs, consists of 2,550 training data and 1,309 testing data, has an 86.40% accuracy with the NN model, and an 86.63% accuracy with the UFL NN model.

Gede Aditra Pradnyana et. al. [2] propose the Otsu binarization, sliding patch algorithm, Zoning methods and Latent

Semantic Index (LSI) using top-N Patch Learning on Balinese characters or glyphs on the AMADI_LontarSet dataset and these were the results that are achieved. It is able to rank given patches, hierarchically from same page, but they still showed poor performance in ranking the patches which are in different pages. Their Mean Average Precision (maP) values averaged at 0.44 and recalled a mean value of 30.95 giving the accuracy not very satisfactory.

Rapeeporn Chamchong et. al. [3] compare between multiple models including, the Otsu binarization (OT) technique, Sauvola and Pietikainen's technique (SAU), Adaptive Logical Level technique (ALL), Improvement of Integrated Function Algorithm (IIF), Background Estimation (BE), Kohonen Self Organizing Map (KSOM) and Local Maximum and Minimum techniques (LMM) on the DIBCO 2011 dataset for identifying Thai characters or glyphs and these were the results that are achieved. The F-measure scores of all the methods were obtained. The OT method obtained 77.4661. The SAU method obtains 73.8697. The ALL method receives 80.2972. The IIF method receives 80.4826. The BE method obtains 81.6674. The LMM method acquires 85.4494. The KSOM obtains 80.1970. The BE, LMM hybrid method acquires 85.4794. The BE, ALL, LMM hybrid method receives 87.2591. Here the BE, ALL and LMM combination model fives them the best result.

Kavitha Subramani et. al. [4] propose the Binarization of images using the Otsu method, mean-shift algorithm, and Gaussian noise removal on the GOML dataset on Tamil glyphs or characters and these were the results that they are achieved. The Peak signal-to-noise ratio (PSNR) value measures 46.2 which is a good result and shows quite accurately.

R. S. Sabeenian et. al. [5] presents the training of data in CNN and compare it with Support Vector Machines (SVM), K-Nearest Neighbor Classifier (KNN), and Fast Artificial Neural Network (FFNN), from the total parameters, 2,238,441 parameters were learned and from this, 2,235,945 parameters were trainable and 2,496 parameters were non-trainable. The prediction rate of the SVM model is 85.46%, the prediction rate of the KNN model is 77.21%, the prediction rate of the FFNN model is 89.21%, and the prediction rate of the CNN is 96.21%. The proposed CNN model has given the best result for this dataset.

Shijian Lu et. al. [6] propose document image binarization which is used for digital image processing. They have used a model that consists of polynomial smoothing, document background estimation, text stroke edge detection and local threshold estimation on the DIBCO 2009 dataset and compared with Otsu's binarization, Niblack's method, Sauvola's method, Gato's method and Su's method and these are the results that are achieved. The F-measure of the Otsu's method is 78.72, and the PSNR is 15.34. The F-measure of the Niblack's method is 55.82, and the PSNR is 9.89. The F-measure of the Sauvola's method is 85.41, and the PSNR is 16.39. The F-measure of the Gato's method is 85.25, and the PSNR is 16.50. The F-measure of the Su's method is 91.06, and the PSNR is 18.50. The F-measure of their proposed method is 91.24, and the PSNR is 18.66. Their proposed method is the best for dealing with the DIBCO dataset.

Vaisakh et. al. [7] propose the Artificial CNN model which includes an image binarization with convolutional layers, maxpooling layers and flattening layers sequentially, on the Amrita_MalCharDb dataset consisting of 17236 samples for training and 6030 images for testing and these were the results that are achieved. The total accuracy attained from proposed method in this paper is 91%, the F1-score comes at 92%, the precision comes at 93%, and recall comes at 89%.

Peeta Basa Pati et. al. [8] propose a custom multilayer CNN has been utilized to recognize segmented characters from palm leaves. It has been considered only forty-eight characters of Malayalam to test upon, consisting of line and character segmentation, having six convolution layers and three dense layers. The training and testing accuracy differences are in the order of 8% between the Custom CNN model and Resnet50 model. 4 glyphs are recognized with accuracies of 98%, 94%, 93%, and 91%. 23 glyphs are predicted to have an accuracy in the range of 80-90%. 8 glyphs are predicted to have an accuracy in the range of 78-80%.

Remya Sivan et. al. [9] propose a CNN model, Which involves character segmentation, and training of Malayalam handwritten character set using Deep learning Models. Using a Kaggle dataset along with VGG16, LeNet, Mobile Net, Resnet152v2, and VGG19 architectures and these were the results that they have received. 10 glyphs were not classified and recognized with accuracy in the range of 92-96%. 6 glyphs were recognized and attained an accuracy in the range of 83 to 90%. The Custom CNN model in training data gives an accuracy of 98% and that of testing data 96%. With the augmentation, and the training data attained an accuracy of 92% and a testing data of 82% without the augmentation.

Bolan et. al. [10] propose document image binarization which is used for digital image processing. They have used a model that consists of Contrast Image Construction, High Contrast Pixel Detection, and Historical Document Thresholding on the DIBCO 2009 dataset and compared with Otsu's binarization, Niblack's method, Sauvola's method, and Lu and Tan method and these are the results that are achieved. The F-measure of the proposed methodology is 89.93%, while the Lu and Tan method achieved a score of 88.53%, the Otsu method achieves a score of 66.13%, the Niblack method achieves a score of 77.34%, the Sauvola method achieved a score of 80.44%. The proposed method gives the best digital image to be used.

Aditi et. al. [11] propose the study about recognition of Devanagari script using CapsNet-based model . Although CNN can be used for Recognition, due to it's limitation like feature extraction and improper recognition on overlapping characters, CapsNet-based model would yield the better result. The model yields the results, implemented in Keras python Library, which gives an accuracy of 94.6%. hen the proposed model is compared with other models like Multi-layer Perceptron, K-Nearest Neighbor, Convolutional neural network Which yields the accuracies of 92.8%, 93.8% and 93.73% which gives us

using CapsNet based model as efficient algorithm. Although, it yields better than rest of the algorithms, the accuracy is still unsatisfactory and hence it can be concluded that the results are not satisfactory.

Srikanta M. K. et. al. [12] state a unique approach to remove noisy data in the degraded historic documents. Two methods, Curvelet transform and Mathematical Morphology used in this study, where curvelet transform is used to remove the noisy data in the image and Mathematical Morphology is used to remove the background of the document image. This process yields the results as follows. The Peak Signal to Noise Ratio (PSNR) is taken as the primary metric to find the efficiency, where 10 image samples are taken at random and compared the results. Two PSNR'S (PSNR1 and PSNR2) Where PSNR1 is for removing noise in image data and PSNR2 for removing the background. For the 6th Image, it was observed the highest PSNR1 and PSNR2 with a total of 27.7096 and 57.5540.

T. J. Alexander et. al. [13] states the binarization approach to restore the palm leaf manuscript. In this study, the palm leaf document image is binarized using various global and local techniques. After binarization, Whale Optimization algorithm is applied which yields a precision of 95%. It is then compared with the binarization techneques like Global threshold, Local Threshold and others. On comparison, it is observed that the Whale Optimization algorithm attains highest precision and recall.

III. Dataset

The dataset used is patches_gabor_15816_1. This dataset has filename, class and 24 features as its column entries and consisting of 4311 observations. The images that are attributed by their name, have undergone feature extraction. The number of Gabor kernels that have been employed is 25, with 5 values of θ and 5 values of λ each in a loop. The parameters used in the kernel are λ, which represents the wavelength of sinusoidal component, θ which represents the orientation of the normal to the parallel stripes, ψ represents the phase offset, σ represents the standard deviation and γ represents the spatial aspect ratio. The features, are hence also called as filters for the image.

IV. Methodology

A. Feature Selection

The 24 features that are present in the dataset, may not all contribute to the application as strong as some features. Some features may be similar to one specific feature. Some features may not strongly improving the accuracy in predicting the label value. This is the reason for applying feature reduction to obtain the best features that will be helpful to remove any noise or redundancy in the data. The feature selection can be applied using the metric correlation, which tells how related one feature is to another. The level of correlation applied can differ and it all has to be done by trial and error. The features are compared to each other and the highly correlated features are removed. It is then compared to the target label and the least correlated features are removed to obtain the finalized dataset.

B. Clustering

Clustering is a type of unsupervised learning that takes only the input data and assigns it into a group of clusters having a cluster centroid. These clusters change every iteration and the outlier values are calculated differently. Usually, a distance measure like Euclidean distance or Manhattan distance to calculate how far the observation is from the centroids. This is a learning that can be decided by the programmer, to control the number of clusters involved based on the problem statement. There are multiple clustering that have been employed in the study.

k-Means clustering is a method which takes k random points as cluster centers and measures the distance of all points to the cluster centers. The smallest distance is identified, and placed with that corresponding cluster. The mean value of cluster points is obtained, becoming the new cluster centers. The cycle is repeated until there is no major improvement in the cluster positions.

Agglomerative clustering is a hierarchical clustering method that groups together multiple data points each other based on distance measure. This method provides clearer positions of the clusters as the grouping happens first and hence the separability of each cluster label is much easier. This bottom approach is very efficient in finding separated groups in the data.

Density-based Spatial clustering of Applications with Noise (DBSCAN) clustering is a density based clustering which takes the radius around the data point in the cluster and checks for the minimum number of points from the cluster present. It recursively checks for the points and if there are not enough, try to find another cluster that fits. If no other cluster values are present, then declare them as noise.

Gaussian Mixture Model (GMM) clustering employs the Naive Bayes distribution to obtain the maximum likelihood of the image being part of the specific cluster. An estimation to the closest value is done and then maximizing the labels values is done so to give the most optimal solution.

C. Classification

Classification is a process where a machine takes the observation vectors and undergoes training to predict specific label values. This in turn is compared with the original labels that have been assigned with according metrics. There are many classification model that have been employed in this study.

The k-NN or k – nearest neighbors' classification is a classification model which deals the class values and the testing data takes values from k neighbors, and predicts the closest among them and it is assigned that class value. This gives an idea of how close the data is predicted to the original value and can be used for modelling our palm leaf data images.

Decision tree is a flowchart tree structure, where each node represents attributes to be tested, and the leaf node represents the class label. The training data is divided into subsets using decision trees, based on the attributes until a stopping condition comes, such as maximum depth of the tree.

Based on ensemble learning, Random Forest is a machine learning algorithm that consists of decision trees and takes the average to improve the accuracy of that dataset. The higher the number of trees in the forest, the higher the accuracy. A function that is imported which is RandomForest Classifier from sklearn. Then use the parameters "n_estimators" to find the number of decisions take need to be made.

A perceptron is a binary classifier that comes under a neural network, where the algorithm acts as neuron where the inputs to the model acts as a sensory activation and it goes through hidden layers and reaches the end acting like out output. There are two parts to the perceptron, a weighted summation and an activation function.

A logistical regression model actually acts as classification solution and it acts as a regression. It takes the linear regression input and a sigmoid function is applied to it to closely, get the approximate predicted class value and compare with the target class value.

The support vector machine model tries to separate the class values in a bi-class classification problem. It does this by creating a hyperplane that acts as a separator. This tells how far can the classes be spread apart and choosing the best hyperplane gives a fairly accurate positioning of the classes.

Naive Bayes classification is a probability classifier that deals with the probability of the class values. These probabilities are calculated for the training data and will be used to predict the class values of the testing data and is compared with the target class values.

D. Image Analysis

For each cluster label, an analysis of images can be done to obtain any insights into the clusters assigned. The metrics that have been taken to check the images are:

- Border present in the leaf
- Punch-hole present in the leaf
- Visibility of the image
- Scratches present in the leaf
- Stained text
- Discoloration of the text
- Fungus in the palm leaf
- Condition of the palm leaf

In total; for k-Means model, 30 images have been analyzed; for DBSCAN model, 25 images have been analyzed; and for Agglomerative and Gaussian Mixture models, 10 images have been analyzed. It strikes clearly that the Agglomerative clustering and the GMM clustering have nearly the same images and hence the analysis for both clusters is the same.

E. Workflow of the dataset

First, importing the dataset and check for data types, and then the encoding of attributes. In the encoding, some of the attributes are nominal while others are ordinal. For nominal attributes, one hot encoding is applied since they consist of only two types of data, it is divided and encoded. After encoding a check for missing values is done and the ways of handling them is identified. It is preferred to use the mean function to deal with missing values. Outliers are the attributes that are too far spread from the expected value. The outliers are then damped with z-score normalization onto a suitable scale. The data is split into 70% consisting of training data, and 30% consisting of testing data. In the actual data, it is noticed that the target labels are imbalanced with label 1 having 621 observations, label 2 having 223 observations and label 0 having 449 observations in the testing data. Hence the use of Synthetic Minority Oversampling technique (SMOTE) is applied on the training data and the testing of the models remain as usual. During every phase, the models are subject to hyper-parameter training using grid search method to get the most optimal results. The classification models are employed with data before it is normalized. After normalizing the data, a retraining of the models is done. After feature selection is employed, a final retrain on the classification models is done. Clustering models is then employed on the reduced dataset and image analysis on the clusters is presented. The figure Fig. 1 gives us the flowchart of the data moving through the system before being trained with a model.

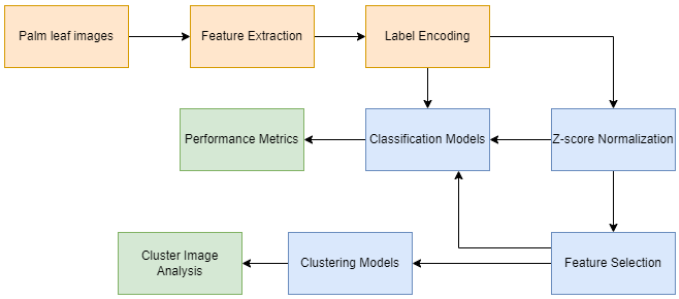

Fig. 1. Flowchart to depict the data flow

V. Result

A. Metrics Used

The metrics that are useful that be employed for the classifiers are visualized using a confusion matrix. If the values which are labelled as the class value are predicted correctly, then they are true positive (TP). If the values, which are labelled as the class value are predicted incorrectly, then they are false negative (FN). If the values, which are not labelled as the class value are predicted correctly, then they are true negative (TN). If the values which are labelled as the class value are predicted incorrectly, then they are false positive (FP). The metrics are hence defined as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1 - score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{4}$$

The metrics that will be used for the clustering methods are Silhouette score, Davies-Bouldin (DB) index, and Adjusted Rank Index (ARI). The Silhouette score tells how the data points are close to the cluster centers compared to its outliers. The Davies-Bouldin index shows the similarity of the cluster to its nearest cluster, where lower the DB index, the more the separable all clusters are present. The ARI score measures how much the cluster labels are separated and ideally should be close to 1.

B. Classification Models Performance

For classification models, the predicting of label values and comparing it with the original labels is done in three phases. Before normalization, Random Forest model gives the best results with an accuracy score of 76.02%, a precision of 75.54%, a recall of 76.02%, and an F1-score of 75.61% taking 69.2890s to compute. Support Vector machine could not run and it showed no results in training or testing. The other models range between an accuracy of 60% and 75%.

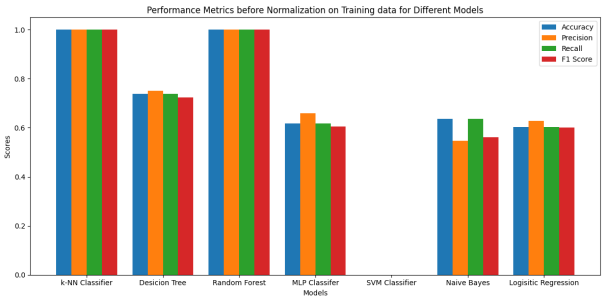

Fig. 2. Results for Training data before normalization

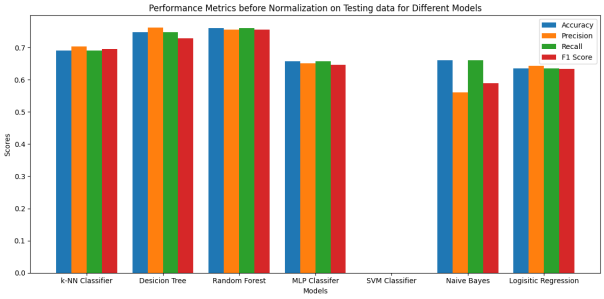

Fig. 3. Results for Testing data before normalization

After normalization, Support Vector Machine gives the best results with an accuracy of 78.11%, a precision of 77.77%, a recall of 75.56%, and an F1-score of 75.70% taking 75.3815s to compute. Naive Bayes performs the worst with an accuracy of 66.04%, a precision of 56.09%, a recall of 66.04%, and an F1-score of 58.85% taking 0.1259s to compute. The other models range between an accuracy of 67% and 76%.

After feature selection, Support Vector Machine gives the best results with an accuracy of 75.56%, a precision of 75.46%, a recall of 75.56%, and an F1-score of 74.22% taking 1.3209s to compute. Multi-layer Perceptron performs the worst

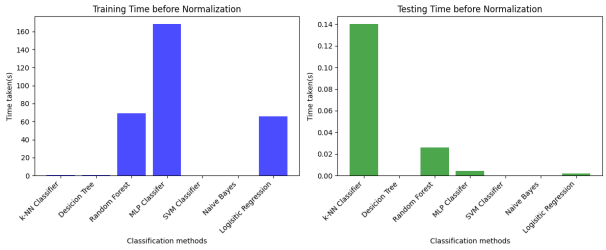

Fig. 4. Computational times before normalization for both training and testing data

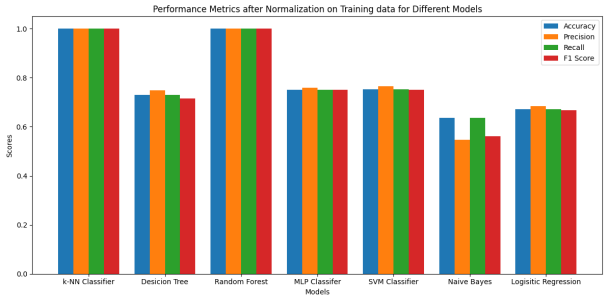

Fig. 5. Results for Training data after normalization

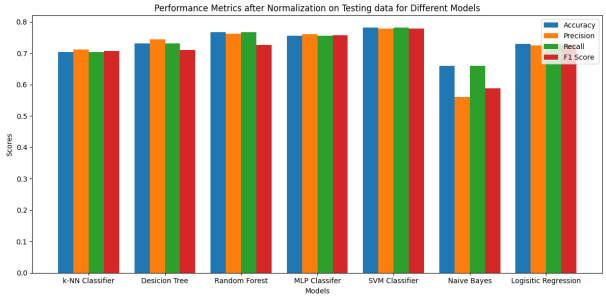

Fig. 6. Results for Testing data after normalization

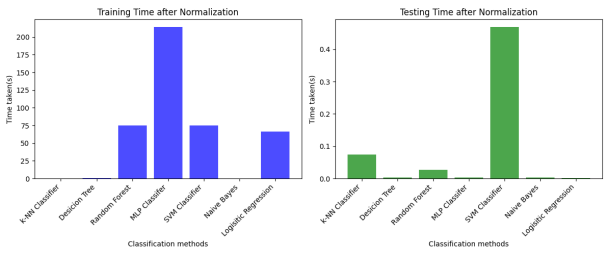

Fig. 7. Computational times after normalization for both training and testing data

among all the models with an accuracy of 40.83%, a precision of 66.90%, a recall of 40.83%, and an F1-score of 42.99% taking 119.2250s to compute. The other models range between an accuracy of 59% to 74%.

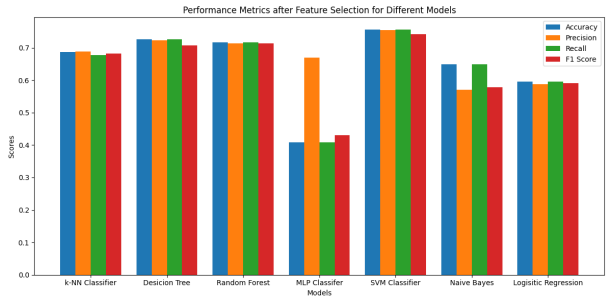

Fig. 8. Results for Testing data after feature selection

Fig. 9. Computational times after feature selection for testing data

This gives an insight that despite applying normalization and feature selection, the performance of classification models did not improve substantially and even became worse after feature selection. Hence with re-examination of the data, clustering models are applied on the input data.

C. Clustering models performance

The k-Means clustering did give a total of 6 clusters with a Silhouette score of -0.1887, a DB index of 29.1169, and an ARI score of 0.1280 taking 1.2791s to compute. The Agglomerative clustering did give a total of 2 clusters with a Silhouette score of 0.2919, a DB index of 2.6032, and an ARI score of 0.2058 taking 20.0449s to compute. The GMM model also did the give the exact same scores as the Agglomerative clustering for 2 clusters taking 120.8077s to compute. The DBSCAN clustering did give a total of 5 clusters with a Silhouette score of -0.3678, a DB index of 9.5601, and an ARI score of 0.2186 taking 31.0702s to compute.

D. Image Analysis

Upon closer inspection, each model is generating clusters that represented features present in the images. For k-Means

TABLE I
HYPER PARAMETERS FOR CLASSIFICATION MODELS(BN-BEFORE NORMALIZATION AND AN-AFTER NORMALIZATION)

Classifier Model	Hyper parameters before feature selection	Hyper parameters after feature selection
k-NN	'n_neighbors': 1	'n_neighbors': 1
Decision Tree	'max_depth': 10, 'max_features': 'log2', 'max_leaf_nodes': 15 (BN) 'max_depth': 10, 'max_features': 'sqrt', 'max_leaf_nodes': 15 (AN)	'max_depth': 10, 'max_features': 'log2', 'max_leaf_nodes': 15
Random Forest	'n_estimators': 100	'n_estimators': 50
MLP	'activation': 'tanh', 'alpha': 0.8, 'hidden_layer_sizes': (100, 3), 'learning_rate': 'constant', 'solver': 'lbfgs'	'activation': 'relu', 'alpha': 0.8, 'hidden_layer_sizes': (100, 3), 'learning_rate': 'constant', 'solver': 'lbfgs'
SVM	Not known (BN) 'C': 30, 'kernel': 'rbf' (AN)	'C': 30, 'kernel': 'rbf'
Naïve Bayes	'var_smoothing': 10.0	'var_smoothing': 10.0
Logistic Regression	'solver': 'liblinear'	'solver': 'liblinear'

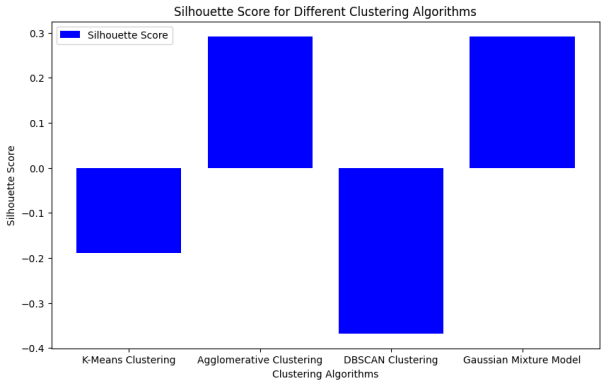

Fig. 10. Silhouette scores for each clustering model

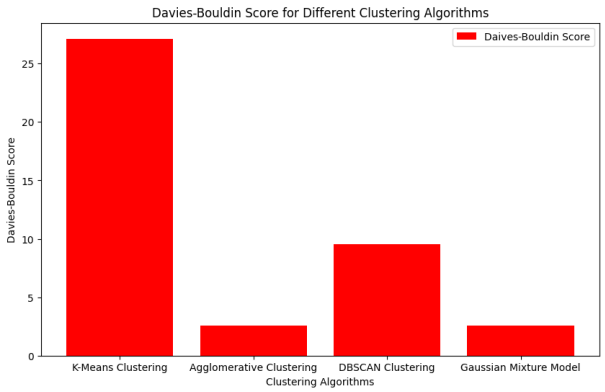

Fig. 11. DB index for each clustering model

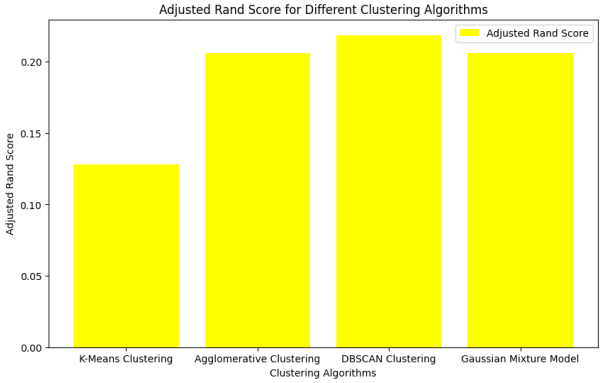

Fig. 12. ARI score for each clustering model

TABLE II
HYPER PARAMETERS FOR CLUSTERING MODELS

Clustering Model	Hyper parameters
k-Means clustering	'n_clusters': 6
Agglomerative clustering	'n_clusters': 2, 'linkage': 'single', 'affinity': 'cosine'
DBSCAN clustering	'min_samples': 7, 'metric': 'euclidean', 'algorithm': 'auto'
GMM clustering	'max_iter': 100, 'covariance_type': 'tied', 'n_components': 2, 'init_params': 'kmeans'

clustering, cluster label 0, represents images which are clearly visible with no major scratch marks, slight discoloration and having fungus in the leaf. Cluster label 1, represents images

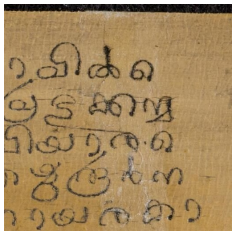

Fig. 13. Image from k-means cluster label 0

that are clearly visible with no major scratch marks but with sometimes faded text, or maybe too much ink and having fungus in the leaf. Cluster label 2, represents images that not

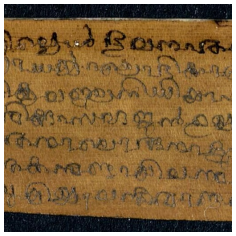

Fig. 14. Image from k-means cluster label 1

clearly visible and subject to insect bites slowly deteriorating the palm leaf. Cluster label 3, represents images that are

Fig. 15. Image from k-means cluster label 2

clearly visible and have no physical degradation, stains or discolorations. Cluster label 4, represents images that are

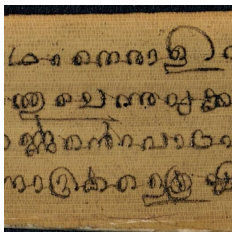

Fig. 16. Image from k-means cluster label 3

clearly visible or mostly visible but are torn in some places in the palm leaf making it sometimes difficult to read. Cluster

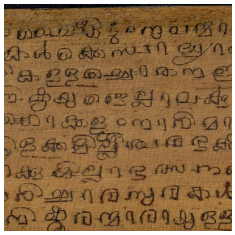

Fig. 17. Image from k-means cluster label 4

label 5, represents images where the palm leaf has either no data at all or the document is brown and the characters are practically difficult to spot with a particular low contrast with visible peeling of the leaf.

Fig. 18. Image from k-means cluster label 5

For Agglomerative clustering and GMM clustering, cluster label 0, represents images that are visible to a varying degree, may contain discoloration and or stains but is not of the worst quality. Cluster label 1, represents images that are dull, having a pigment of green sometimes and is not visible to the naked eye and having discoloration of the characters.

For DBSCAN clustering, cluster label -1, represents images that are not fully visible, having tears and peeling in

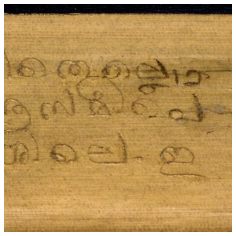

Fig. 19. Image from Agglomerative/GMM cluster label 0

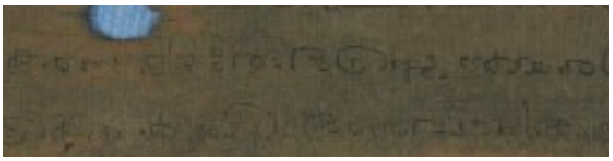

Fig. 20. Image from Agglomerative/GMM cluster label 1

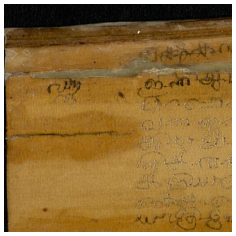

Fig. 21. Image from DBSCAN cluster label -1

label 0, represents images that are mostly visible with no major discoloration, but a couple tears and scratch marks and sometimes fungus on the palm leaf. Cluster label 1, represents

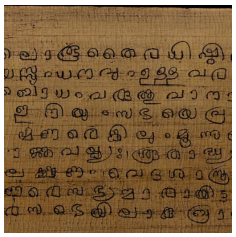

Fig. 22. Image from DBSCAN cluster label 0

images that are very clearly visible with no major stains or discolorations but with very little scratch marks and no major defects. Cluster label 2, represents images that are not fully visible, however are affected by insect bites and major discoloration in characters. Cluster label 3, represents images that are not fully visible, but have major discoloration without any major stains or physical deterioration.

VI. Conclusion

The use of clustering to classify images to identify patterns in the palm leaves is a now an important role in identifying the right algorithm for optimizing visible data with no flaws or

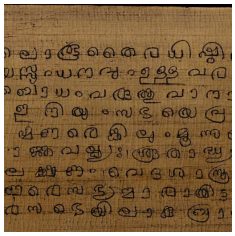

Fig. 23. Image from DBSCAN cluster label 1

Fig. 24. Image from DBSCAN cluster label 2

even data that is not visible with stains, insect bites fungus or even tearing of the leaf. There is still more that can be explored in the field of palm leaf document classification, which can optimize the search even better.

References

[1] Kesiman, M.W.A., Burie, J.C. and Ogier, J.M., 2017, November. A complete scheme of spatially categorized glyph recognition for the transliteration of balinese palm leaf manuscripts. In 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR) (Vol. 1, pp. 125-130). IEEE.

[2] Kesiman, M.W.A. and Pradnyana, G.A., 2019, October. A Complete Scheme of Word Spotting System for the Balinese Palm Leaf Manuscripts. In 2019 11th International Conference on Information Technology and Electrical Engineering (ICITEE) (pp. 1-5). IEEE.

[3] Chamchong, R., Jareanpon, C. and Fung, C.C., 2013, July. Generation of optimal binarisation output from ancient Thai manuscripts on palm leaves. In 2013 International Conference on Machine Learning and Cybernetics (Vol. 4, pp. 1643-1648). IEEE.

[4] Subramani, K. and Murugavalli, S., 2017, January. A novel binarization method for degraded tamil palm leaf images. In 2016 Eighth International Conference on Advanced Computing (ICoAC) (pp. 176-181). IEEE.

[5] Sabeenian, R.S., Paramasivam, M.E., Anand, R. and Dinesh, P.M., 2019. Palm-leaf manuscript character recognition and classification using convolutional neural networks. In Computing and Network Sustainability: Proceedings of IRSCNS 2018 (pp. 397-404). Springer Singapore.

[6] Lu, S., Su, B. and Tan, C.L., 2010. Document image binarization using background estimation and stroke edges. International Journal on Document Analysis and Recognition (IJDAR), 13, pp.303-314.

[7] Vaisakh, V.K. and Das, L.B., 2020, February. Handwritten Malayalam Character Recognition System using Artificial Neural Networks. In 2020 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS) (pp. 1-4). IEEE.

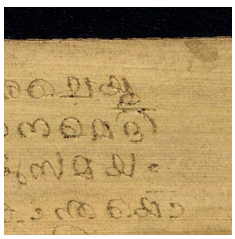

Fig. 25. Image from DBSCAN cluster label 3

[8] Sivan, R., Singh, T. and Pati, P.B., 2022, December. Malayalam Character Recognition from Palm Leaves Using Deep-Learning. In 2022 OITS International Conference on Information Technology (OCIT) (pp. 134-139). IEEE.

[9] Sivan, R., Singh, T. and Pati, P.B., 2022, December. Malayalam Character Recognition from Palm Leaves Using Deep-Learning. In 2022 OITS International Conference on Information Technology (OCIT) (pp. 134-139). IEEE.

[10] Su, B., Lu, S. and Tan, C.L., 2010, June. Binarization of historical document images using the local maximum and minimum. In Proceedings of the 9th IAPR International Workshop on Document Analysis Systems (pp. 159-166).

[11] Aditi Moudgil, Saravjeet Singh, Vinay Gautam, Shalli Rani, Syed Hassan Shah, 2023. Handwritten devanagari manuscript characters recognition using capsnet. International Journal of Cognitive Computing in Engineering, 4, pp.47-54.

[12] Gangamma, B., Murthy, K.S., Chandra, G.P., Kaushik, S. and Kumar, S., 2010, December. A combined approach for degraded historical documents denoising using Curvelet and mathematical morphology. In 2010 IEEE International Conference on Computational Intelligence and Computing Research (pp. 1-6). IEEE.

[13] Alexander, T.J. and Kumar, S.S., 2020. A novel binarization technique based on Whale Optimization Algorithm for better restoration of palm leaf manuscript. Journal of Ambient Intelligence and Humanized Computing, pp.1-8.