




Improving Isolated Glyph Classification Task for Palm Leaf Manuscripts

Nimol Thuon¹ , Jun Du¹ , and Jianshu Zhang^{1,2} 

¹ University of Science and Technology of China, Hefei, Anhui, China
{tnimol,xysszjs}@mail.ustc.edu.cn, jundu@ustc.edu.cn

² iFlytek Research, Hefei, Anhui, China

Abstract. Digitization of ancient palm leaf manuscripts is gaining momentum due to the limited datasets and complex features of text images of palm leaf manuscripts. Thus far, the previous studies did not deeply analyze the application of the trending techniques on the palm leaf manuscripts, considering how deep learning approaches require large datasets, while some isolated glyphs contain more than one character with complex grammatical components. Therefore, this paper explores the possibilities and practical methods for improving isolated glyph classification. In particular, we focus on both the front-end and the back-end processes involved in the image classification task. For the front-end analysis, we present multi-task preprocessing techniques, including data augmentation techniques, new datasets extraction, and image enhancement techniques to increase the quality and quantity of datasets. For the back-end side, we aim to study the visual backbones of deep learning techniques, especially CNNs (including VGG, ResNet, and EfficientNet) and attention-based models (including ViT, DeiT, and CvT). Furthermore, the analysis and evaluation examined how data augmentation techniques and preprocessing interact with the amount of data used in training. Evidently, we experimented on three palm leaf manuscripts, including Balinese, Sundanese, and Khmer scripts from the ICFHR contest 2018, SluekRith, AMDI LontarSet, and Sunda datasets. Regarding the quality of research, the experiment delivers an effective way of training palm leaf datasets for the document analysis community.

Keywords: Historical document analysis · Vision transformer · Palm leaf manuscript · Neural network

1 Introduction

Historically, thousands of scripts have been recorded on palm leaves to depict significant historical events, Buddhism practices, astrology, or even medical treatment [8]. Southeast Asian Palm Leaf Manuscripts Analysis (SEAPA) offers a known challenge for document analysis tasks. Among these, Balinese [8], Khmer [19], and Sundanese scripts [15] have not only received increasing attention but have also brought fresh challenges for researchers due to their unique writing

formats. SEAPA datasets generally consist of text-line, isolated glyphs, and word/text. While SEAPA research is developing rapidly, most collections of documents are only used during the first step. Deep neural networks have the ability to automatically learn desirable representations of text images rather than designing the custom features manually, which is why this system is best known for its effectiveness when dealing with image classification problems. In addition to Deep Neural Networks, the Convolutional Neural Network (CNN), Very Deep Convolutional Networks (VGGNet) [13], Residual Neural Network (ResNet) [6], and EfficientNet [16] are also some of the well-known deep learning architectures. The recent developments of machine learning also center on ViTs [4, 14, 18], which is one of the trending architectures used to achieve rewarding results in the image classification task. Afterward, hybrid approaches like CNN-ViT [5, 20] were proposed for solving deep learning tasks. However, the applications of those trending approaches on palm leaf datasets remain controversial. The previous studies, ICFHR 2016 [1] and experimental research on Southeast Asian palm leaf benchmarks, focused mainly on handcrafted feature extraction and traditional methods of Balinese scripts. Meanwhile, SEAPA classification tasks have only been analyzed sparingly in recent years. Over the last few years, numerous techniques have been developed and tested, particularly for Latin-based scripts. Regardless, low-resource languages like palm leaf manuscripts still require benchmarking on recent trending approaches for accuracy, data quantity, functional architecture, and training strategies.

Considering these challenges, this paper investigates the performance of front-end and back-end techniques for improving isolated glyph classification. Furthermore, due to the physical degradation of the manuscript datasets, different training ways, such as preprocessing, data augmentation, and visual backbones, are thoroughly investigated to determine their effectiveness and alternation on historical manuscripts.

In summary, the significant contributions of this work are as follows:

1. We present multi-task preprocessing techniques for boosting the accuracy rate. In particular, we extract new datasets from text-line and word datasets. Additionally, image enhancement and data augmentation techniques are presented.
2. We evaluate recent deep learning approaches with data augmentation on different benchmarks toward isolated glyph classification task.
3. We identify the problems with image quality. Based on our analysis, enhancing input images' color from RGB to greyscale to binary could improve their accuracy rates. As a result, binary datasets also showed better performance.
4. We investigate the performances of palm leaf manuscripts using the most effective deep learning approaches and our new datasets with preprocessing techniques.

The rest of this paper, we present relevant information on palm leaf manuscripts in Sect. 2. Section 3 describes the overall framework, including preprocessing and visual backbones of glyph classifications. Finally, Sect. 4 includes the experimental setup and results before we conclude our study in Sect. 5.

2 Palm Leaf Manuscripts from Southeast Asia

This section aims at providing descriptive information on the scales corpus of the palm leaf manuscript for our experimental studies. As shown in Fig. 1, the compilations consist of three palm leaf manuscripts, including Khmer script (Cambodia), Balinese, and Sundanese scripts (Indonesia).

2.1 Corpus and Languages

Balinese [9] is one of Indonesia’s local and traditional languages, which is commonly found in many ancient manuscripts of Bali and is also the native language of Bali people. A total of 100 classes are presented in the Balinese language, including consonants, vowels, and special characters. In the previous contest [10], 133 classes of Balinese were obtained, 11,710 images of which were used for training and 7,673 for testing in the isolated glyph recognition task. Meanwhile, Khmer [19] is the official language of Cambodia, which is also commonly used to record Buddhist literature. The language consists of 111 classes, including consonants, vowels, numerals, special characters, and diacritics. The words can be formed in different ways by combining a few symbols, consonants, or vowels. Buddhist Institute or National Library collected most of the manuscripts, including 113,206 images for training and 90,669 for testing. Lastly, Sundanese [15] originated from Garut, West Java, and Situs Kabuyutan Ciburuy in Indonesia, containing 27 collections, each of which includes 15 to 39 pages. Similarly, Sundanese characters also include numbers, vowels, essential characters, and special characters. However, Sundanese contains only 60 classes, 4,555 images of which were used for training, while 2,816 others were for testing.

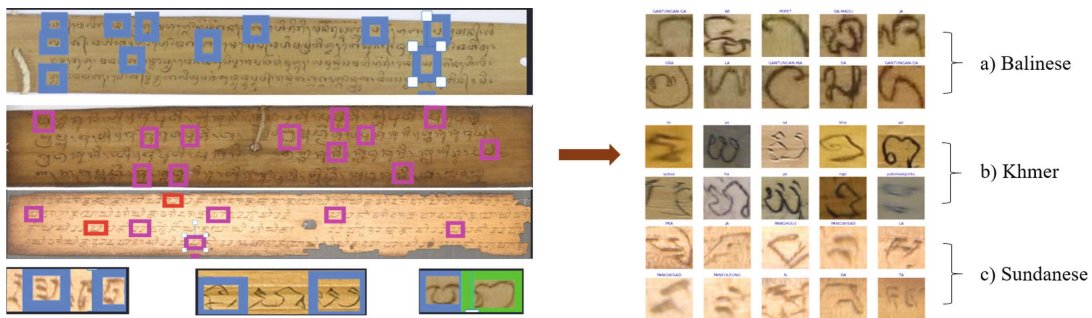


Fig. 1. Sample of data collections from the isolated character palm leaf datasets. (a) Balinese script from Indonesia; (b) Khmer script from Cambodia; (c) Sundanese script from Indonesia.

2.2 Challenges of Isolated Glyph Datasets

Palm leaf manuscripts are generally complicated considering the character classes, alphabets, and numerals, making historical document image analysis more complex to binarize than the scanned documents. In addition, the

characters-based are similar, consisting of more than one way of layering up to form the other letters. Consequently, these ancient manuscripts require a sophisticated and practical system for recognizing and classifying these letters. Moreover, the publicly available datasets for each script of the isolated datasets are small and medium, the smallest of which is Sundanese scripts, in which the data for training and testing consists of approximately ~ 4 K images and ~ 2 K images, respectively. Meanwhile, Balinese and Khmer are considered as medium-sized datasets. From text-line and word datasets, we found that many potential isolated characters could be extracted based on palm leaf datasets [8, 15, 19]. Therefore, extracting more data to support deep learning methods is necessary to balance the training approaches.

3 Overall Frameworks

The overall architecture is shown in Fig. 2, including front-end and back-end. In the front-end, we present a simple and effective method for increasing datasets and improving image quality. Mainly, we extract isolated characters from text-line and word datasets, then perform image enhancement with data augmentation techniques. In the back-end, we present different deep learning techniques, especially on various CNNs and attention-based models.

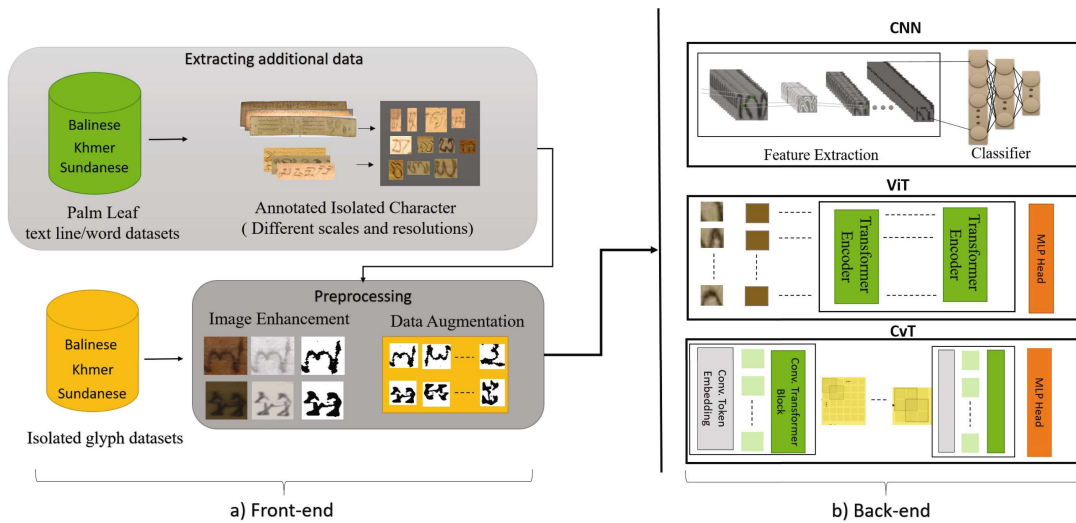


Fig. 2. An overview of our training strategy. As part of the first step of the front-end, we extract new collections from the text-line and word/text palm leaf datasets. Newly additional and the existing datasets are then enhanced using data augmentation and image enhancement techniques. Finally, we inspect the results based on the different data sizes by performing different visual backbones of CNNs and ViTs on the back-end side.

3.1 Data Pattern Generations

Since recent approaches like ViTs and Deep CNNs require a large amount of training data, we present a data pattern generation method to synthesize isolated glyph datasets from palm leaf manuscripts. We enhance and synthesize the isolated character images by taking 2 of the following steps:

1. Obtain additional data collections of text-line and text/word palm leaf datasets.
2. Deploy data augmentation techniques to increase the scale of the existing and the new data collections.

Extracting Additional Datasets. In this step, we aim to extract new collections based on the text-line and word datasets [10]. The publicly available datasets of the isolated glyph images contained limited images; therefore, we extract additional datasets from the previous works containing the three manuscripts. Specifically, 15 local university students from Southeast Asia were selected voluntarily as their comprehensive knowledge and understanding of the Cambodian and Indonesian languages would be more apprehensive in our experiment’s classification and labeling process. As indicated in Fig. 1, we manually crop and label data using friendly interface annotated and label image tools^{1,2}. As shown in some of the samples of the extracting characters, the 15 students are divided into two groups. The first group is the collectors, assigned to extract and identify the characters of the palm leaf images. Based on their general knowledge, they segment and label these palm leaf images upon various qualities of text and resolutions. The second group is responsible for validating the labels collected by the first group based on existing character classes and dictionaries from [10]. As a result, we collected ~ 15 K images from palm leaf datasets. Mainly, we seek to increase 20–50% of the original datasets for training on small datasets such as Sundanese and Balinese manuscripts.

Data Augmentation Techniques. We aim to integrate two straightforward data augmentation techniques into the training of CNNs and ViTs. Firstly, we applied three basic image processing techniques in basic data augmentation using random cropping, random horizontal flipping, and random Gaussian noise. As we randomly resize the patch, crop it according to the scale, then resize it to the original size, we keep the aspect ratio fixed. A random Gaussian noise sample has a mean of 0 and a standard deviation of 0.01. Secondly, we selected regularization and data augmentation (AugReg) into training. In this case, regularization techniques are commonly adopted within the computer vision community [14]. According to [4], we then apply dropout to the intermediate activation of ViT and apply stochastic depth regularization technique [7] that drops layers with a linearly increasing probability. Apart from this, three data augmentation approaches are efficiently utilized to regularize training: RandAug [3], MixUp [22], and CutMix [21].

¹ <https://github.com/donavaly/SleukRith-Set>.

² <https://markuphero.com>.

3.2 Image Enhancement for Palm Leaf Manuscripts (IEPalm)

This section outlines the techniques to enhance the performance of the poor quality palm leaf manuscripts, by presenting a multi-task image processing to clean images called IEPalm in short. We begin by focusing on the contrast balance, and then we perform thresholding binarization to remove some noises from the background and separate the text to provide a better understanding.

Normalization. In the preprocessing step, we primarily focus on balancing the contrast of an image. However, due to the low contrast images, image enhancement is necessary to emphasize certain features or reduce ambiguity between regions of an image, thereby improving thresholding. Inspired by an updated version of contrast limited adaptive histogram equalization (CLAHE), [2] is selected for contrast enhancement in order to correct inconsistencies between text and background. Applying CLAHE can also reduce the noise levels and maintain an image's high spatial frequency content and medial filtering and edge sharpening. In addition to this, adaptive histogram clipping (AHC) can also be applied to the limited contrast technique. By using AHC, clipping levels are automatically adjusted, and over-enhancements are moderated.

With a uniform distribution, the CLAHE technique can be described by:

$$C = Pro(f) * [C_{max} - C_{min}] + C_{min} \quad (1)$$

In exponential distribution, gray level can be described as follows:

$$C = C_{min} - In[1 - Pro(f)] * \left(\frac{1}{\delta}\right) \quad (2)$$

where C_{max} is the maximum pixel value, and C_{min} indicates the minimum pixel value. Therefore, C represents the computed pixel values, $Pro(f)$ is the cumulative probability distribution, and δ is the clip parameter.

Thresholding. WAN [12] was proposed to improve the Sauvola method for more reliability on low-quality images. However, Sauvola's approach struggles to segment text and background. For example, if the contrast between the foreground and the background is down, there may be less noises in the text image. Thus, we present an approach to calculate binarization thresholding for this stage, which is more likely to succeed for these types of degraded documents. The main benefit of WAN is that, it enhances binarization by shifting the threshold for detailed images. Furthermore, the investigation shows that *mean* value m significantly impacts the threshold values. Whenever non-text pixels and text pixels of gray values in an image are close to each other, this can enhance the appearance of the output image. However, if we increase the threshold value, the noise and artifact will remain in the image. Therefore, calculating the maximum threshold value as a replacement for the actual *mean* is necessary. For illustration, the following represents the maximum-mean equation:

$$i_{max} = \frac{\text{mean} + \max(a, b)}{2} \quad (3)$$

where $\max(a, b)$ exemplifies the maximum contrast of the source image, whereas mean , the average contrast of the entire image; both of which allow us to calculate the average of the highest contrast that will consequently recondition the lost features and reduce noises and artifacts in the binarization results. In this case, we estimated the average of the highest contrast and mean of the image. Therefore, WAN is able to restore lost details and reduce noise and artifacts in the binarization results. Specifically, the following algorithm is presented:

$$T = (i_{\max}) \left[1 - k \left(1 - \frac{\sigma}{R} \right) \right] \quad (4)$$

where k and R values use a default value from the Sauvola method. k stands for gray level, m represents mean, σ is the standard deviation, and R stands for color (default value).

3.3 Training CNNs and ViTs

In this section, we investigate the performances of training with different back-ends. CNN and ViT variants have been used successfully in image classification for years. Hence, we explore the back-ends of CNN and ViT variants. The first back-end is CNN-based models. CNNs have also paved ways for convolutional networks, like translation equivalence, object classification, and recognition, which have gained attention in recent years [6, 13, 16]. The second back-end is attention-based models. In recent years, ViTs have performed well in ImageNet image classification tasks [4, 18, 20]. Despite this, implications of both back-ends on low-resource languages like palm leaf manuscripts remain in the studying step. Based on various layers and parameters, the study selects trending architectures as follows:

Very Deep Convolutional Networks (VGG). VGG is an architecture with 16 or 19 layers, introduced by [13] for large-scale image classification. In this case, VGG16 consists of ~ 138 million parameters and VGG19 consists of ~ 144 million parameters. As shown in Fig. 3(a), we fixed the convolution filter size to 3×3 for entire layers to reach deeper implementation while increasing nonlinearity functions for learning complex representations. Furthermore, after the convolution layers, the outputs were maxed before connecting to three fully-connected layers, leading to many different learnable parameters based on VGG16 and VGG19.

Residual Networks (ResNets). ResNet [6] is one of the most popular models in the history of CNN architectures. ResNet is a model that makes of the residual module involving shortcut connections. In this case, when stacking more convolution layers, gradients drop and vanish during back-propagation. Since adding layers without a plan reduces the accuracy and performance, residual learning adds the previous-layer output via “shortcut connections” to the stacked-layer output. These are proven to be less complex than the VGG networks due to the absence of complexity or networks parameters. Therefore, ResNet50 (~ 25 million parameters) and ResNet101 (~ 42 million parameters) are discussed in this study.

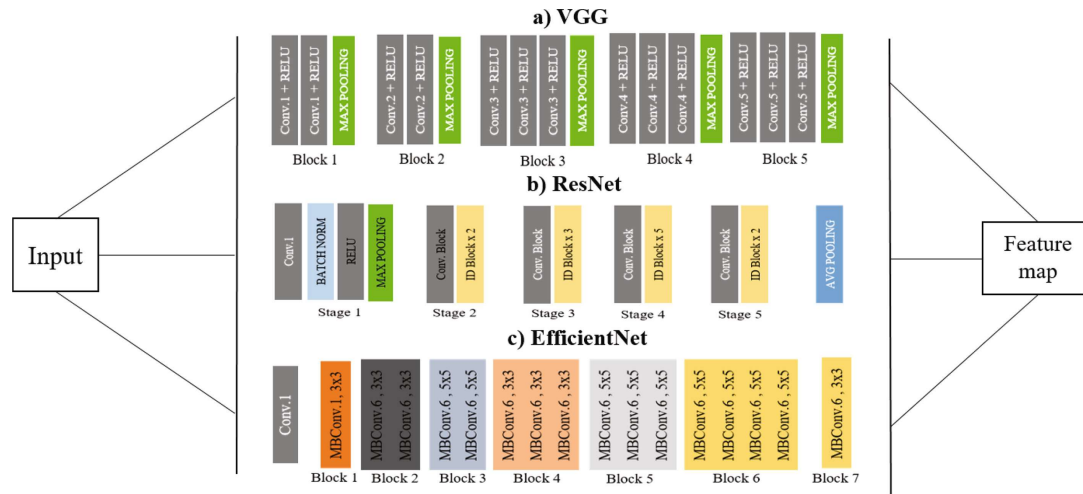


Fig. 3. CNN-based models, such as VGG, ResNet, and EfficientNet, are put into fair comparisons to evaluate their performances at various stages.

EfficientNet. EfficientNet [16] has been one of the trending approaches for improving the accuracy of CNNs. It proves we can achieve excellent results with reasonable parameters by carefully designing our architecture. Hereof, EfficientNet showed excellent results with less parameters. In an effective but simple manner, EfficientNet scales up models using compound coefficients. In contrast to randomly scaling up width, depth, or resolution, compound scaling uniformly scales each dimension with a fixed set of scaling coefficients. Therefore, AutoML and the scaling method are being used. Seven models of various dimensions were developed, and they outperformed the state-of-the-art accuracy and efficiency of most convolutional neural networks. The architecture can be seen in detail in Fig. 3(c). In this study, we evaluate EfficientNetB0 to EfficientNetB3 models ranging from 4–10 million parameters.

Vision Transformers. ViT was proposed with comparable or higher accuracy rates than state-of-the-art approaches for image classification in ImageNet [4]. In most cases, an image is usually split into $K \times K$ overlapping patches. Each input embedding patch yields $K \times K$ input tokens. Figure 4(a) shows a ViT architecture based on transformer multi-attention layers that pair the model over-token intermediate representations. Then, the final grid embedding is used for discrimination. Several approaches use a “class token” to collect contextual information across the entire grid, while others use average global pooling to compact image representation. Lastly, an MLP head outputs a posterior distribution of target classes based on the whole image. Afterward, training on ViTs variants (DeiT) [18] was proposed using supervised pre-training but only with the more regularized and distillate ImageNet-1k datasets to seemingly demonstrate the progress of ViTs. Data-efficient image transformers are far more efficient in training the transformers for image classification tasks. Moreover, it requires far less data and computing resources than the original ViT models. In this case,

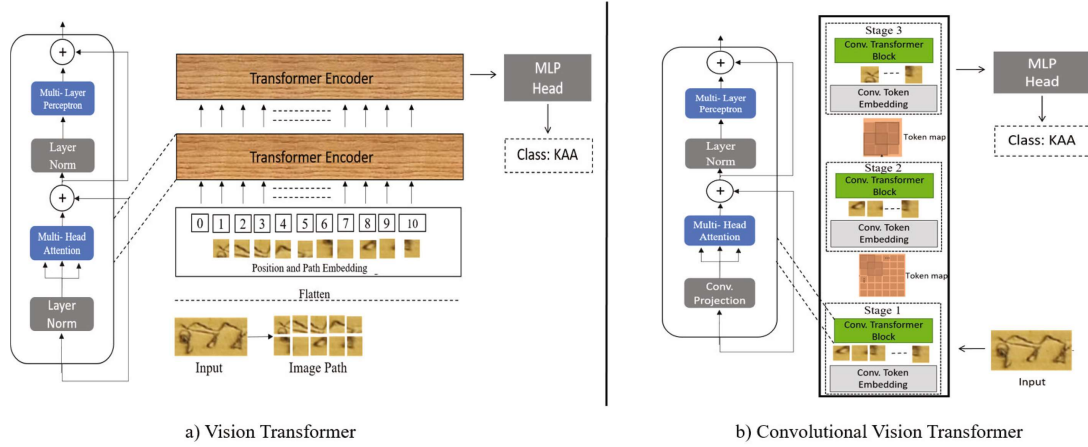


Fig. 4. The overall diagram of ViTs. (a) The standard architecture of ViT (b) The overview of architecture that introduces convolutional to transformer (CvT).

ViT-S and DieT-S with 12 layers and ~ 22 million parameters respectively, while ViT-B and DieT-B with ~ 86 million parameters are selected for evaluations.

Convolutional Vision Transformer (CvT). CvT [20] was proposed to enhance the performance of standard ViT. As shown in Fig. 4(b), this work employs a multi-stage hierarchy design borrowed from CNNs, consisting of three stages in total. In summary, token embedding and convolutional projection are applied to transformer hierarchies and transformer blocks, respectively. As input, convolutional token embedding uses overlapping patches of tokens reshaped to the 2D spatial grid (the stride length can control the degree of the overlapping ones). Each step of CNN architecture lowers tokens (feature resolution) while increasing token width (feature dimension). As transformers stack on each level, the convolutional transformer block uses convolutional projection, whereby the MLP head predicts the output token’s class. CvT13 with ~ 20 million parameters and CvT21 with ~ 32 million parameters are also included in our investigation.

4 Experimental Setups and Results

In this experiment, we aim to determine the front-end and back-end performance for improving isolated glyph classifications. In this case, different experiments are conducted thoroughly to investigate the performances. In addition to this, the experiments are categorized into three questions for quality assurance:

1. Which deep learning technique is the most effective back-end approach for each script in ancient palm leaf datasets?
2. How does the quality of the datasets affect accuracy?
3. Does our new datasets and IEPalm techniques effectiveness improve isolated glyph classification tasks?

4.1 Implementation Settings

Datasets. We evaluate all tasks based on the datasets extracted from the ICFHR 2018 contest [10], SleukRith [19], Sunda dataset [15], and AMADI LontarSet [8], which comprises palm leaf manuscripts for the isolated character/glyph classification task. The Balinese, Khmer, and Sundanese datasets are grouped into classes and divided accordingly for training and testing purposes, as shown in Table 1. The table further compares the number of data collected from the three manuscripts. Based on this table, the study presents the following:

Table 1. The palm leaf datasets contained the original datasets from the ICFHR 2018 contest and mixed them with newly extracted datasets, including classes, training, and testing images.

Script	Dataset	Classes	Training	Testing	Source
Balinese	Track 1	133	11,710	7,673	AMADI LontarSet [8]
	TrackMix 1		16,500		
Khmer	Track 2	111	113,206	90,669	SleukRith Set [19]
	TrackMix 2		120,206		
Sundanese	Track 3	60	4,555	2,816	Sunda Dataset [15]
	TrackMix 3		9,800		

Track 1. Dataset contained 133 classes for the Balinese characters, including 11,710 images of the training set and 7,673 images of the testing set.

TrackMix 1. Dataset contained 16,500 training images of the original Balinese isolated glyph dataset and our additional dataset.

Track 2. Dataset has 111 classes for the Khmer characters, in which 113,206 images were used for training and 90,669 images for testing.

TrackMix 2. Dataset contained 120,206 training images of the original Khmer isolated glyph dataset and our additional dataset.

Track 3. Dataset has 60 classes for the Sundanese characters, including 4,555 and 2,816 images for training and testing, respectively.

TrackMix 3. Dataset contained 9,800 training sets of the original Sundanese isolated glyph datasets and our additional dataset.

Training Settings. In this study, we train the models with NVIDIA 3090 GPUs 24 GB based on TensorFlow and PyTorch. In order to train on palm leaf datasets, we choose CNN architectures such as VGG16, VGG19, ResNet-50, ResNet101, and EfficientNetB0-B3 [6, 13, 16]. Due to limited data training, we perform and analyze data augmentation methods with large-scale pre-trained weights. Particularly, weights for the first CNN levels are frozen, and the remaining parameters are trained. For ViTs, we follow the training recipe of ViT-S, ViT-B, DeiT-S,

DeiT-B, CvT13, and CvT21 [4, 17, 18, 20]. For pre-trained weights, ImageNet-1k (ILSVRC-2012) and ImageNet-21k (ImageNet-21k) are used as the image datasets. Firstly, data should be loaded and transformed into 224×224 in advance to regulate the model on an array of samples. Thus, the size of a loaded image must increase by one for each image with 224×224 pixels and three channels. Moreover, we optimize CNNs, and ViTs using Adam optimizer [11]. The initial learning rate is set to $5e-4$, and the cosine learning rate scheduler is subsequently applied to decrease it. Note that both approaches have been trained with the same configurations of 100 epochs.

Evaluation Metrics. For all experiments, we evaluate the performances of the systems by using accuracy rates of the classification of isolated glyphs palm leaf manuscripts.

4.2 Results

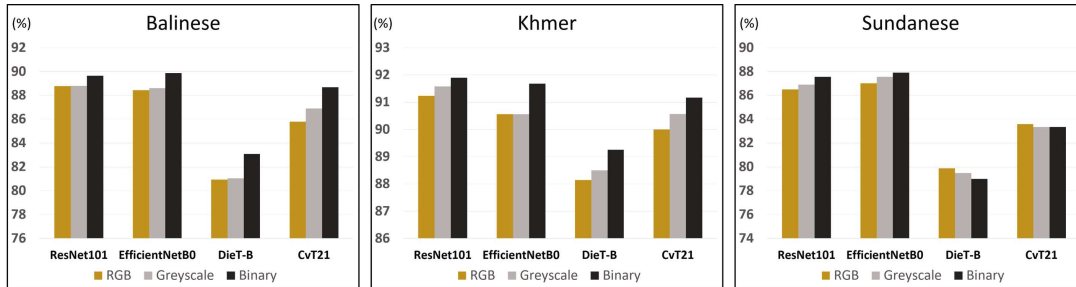
Experiment 1. This study aims to compare results from different deep learning techniques with original datasets [10] with basic data augmentation techniques. As shown in Table 2, this section shows accuracy rates for CNNs and ViTs as well as data augmentation across different manuscripts. According to the results, EfficientNets and ResNets are generally the most reliable methods. Especially, EfficientNets presented impressive performance compared to ViTs methods. However, CvT has shown the best performance among ViT variants. In track 2, CvT also achieved comparable results to CNNs in terms of performances. Furthermore, we found that increasing the parameters of models did not significantly affect the accuracy rates.

Experiment 2. Data transformations were observed in this section. In most of the previous studies, RGB or greyscale images were used in the training step [1, 10]. In this case, we train only the original datasets without any external datasets. Therefore, we compared the results of ResNet101, EfficientNetB0, DeiT-B, and CvT21 using color (RGB), greyscale (GS), and binary (BI) images. In particular, we transformed all training and testing sets into different types of formats. For binary images, we simply applied the WAN method [12] for thresholding. As shown in Fig. 5, isolated glyph classification can be improved by using data transformations. Interestingly, binary datasets outperform greyscale and RGB in Balinese and Khmer scripts, while Sundanese scripts produce equivalent performances. Consequently, preprocessing steps have proven necessary for palm leaf analysis based on the quality of datasets.

Experiment 3. As the central part of this study, we evaluated the effectiveness of our additional datasets and image enhancement methods (IEPalm). Particularly, we applied our strategy to the most effective CNNs and ViTs approaches, such as EfficientNetB0, DeiT-B, and CvT-21. Our entire process is described in Sect. 3. To train these approaches, we used the datasets TrackMix 1, TrackMix 2, and TrackMix 3, whereby each manuscript used the same amount of testing datasets. As shown in Table 3, the results showed that our additional datasets

Table 2. The results of all the architectures trained for Track 1, Track 2, and Track 3 with CNNs, CNNs + data augmentation, ViTs, and ViTs + data augmentation.

Model	Track 1	Track 2	Track 3	Track 1	Track 2	Track 3
	CNNs			CNNs + data augmentation		
VGG16	87.45	88.98	84.50	88.98	89.08	87.20
VGG19	87.90	89.64	84.61	88.91	89.60	87.60
ResNet50	88.47	89.04	85.23	90.04	90.45	88.04
ResNet101	88.78	91.23	86.50	91.65	90.97	88.65
EfficientNetB0	88.45	90.56	87.02	89.49	91.85	90.35
EfficientNetB1	88.57	91.67	86.41	90.45	91.21	88.31
EfficientNetB2	87.44	91.05	86.04	90.77	91.55	89.45
EfficientNetB3	87.95	91.09	87.65	90.65	91.08	89.32
	ViTs			ViTs + data augmentation		
ViT-S_16	79.21	87.12	78.50	82.50	86.25	80.09
ViT-B_16	79.44	88.25	79.50	83.23	87.44	80.47
DeiT-S	80.75	88.04	79.20	85.04	88.45	80.78
DeiT-B	80.94	88.14	79.88	85.50	89.36	81.60
CvT-13	85.50	90.28	83.20	87.04	90.45	84.54
CvT-21	85.80	90.00	83.59	87.50	90.36	85.14

**Fig. 5.** The results of how data transformations can affect the accuracy achieves when performed with color, greyscale, and binary images.

mixed with the original ones had an improved accuracy rate for most cases in this experiment. Similarly, IEPalm techniques also showed their effectiveness in terms of accuracy. Sundanese and Balinese also present an interesting remark, which made significant improvements when used with larger datasets. Therefore, when it comes to boosting the accuracy rates, it is essential to take into account the importance of quality and quantity of the datasets.

Findings and Limitations. For all types of palm leaf manuscripts, the performance of ViTs falls slightly behind CNNs. The evaluation has showcased that both approaches still need significant support from datasets in addition to the

original ones, mainly when trained on Sundanese and Balinese scripts. Specifically, the Sundanese would require more datasets in the training steps to achieve better results. Interestingly, EfficientNet showed impressive results despite using less hyper parameters than ResNet models. For attention based-models, ViTs are also constrained by limited memory and high computing requirements of the expensive quadratic attention in the encoder block. Accordingly, the first observation of CNNs and ViTs also indicates the possibility for improvement by adopting large-scale data such as additional datasets and data augmentation techniques. Lastly, the quality of datasets remains highly significant for boosting the accuracy rates. Thus, cleaning noises or restoring document approaches should be applied before post-processing.

Table 3. The evaluation results of our new collections using IEPalm techniques based on CNNs and ViTs approaches.

	Track 1	TrackMix 1	Track 2	TrackMix 2	Track 3	TrackMix 3
EfficientNetB0	88.45	91.47	90.56	91.47	87.02	91.54
EfficientNetB0 + IEPalm	89.93	92.19	92.91	93.55	91.14	93.08
DeiT-B	80.94	84.46	88.14	91.90	79.88	83.97
DeiT-B + IEPalm	83.64	86.53	89.34	92.44	83.88	86.11
CvT-21	85.8	87.44	90.00	91.57	83.54	88.65
CvT-21 + IEPalm	87.89	89.64	91.55	93.88	85.48	90.23

5 Conclusion

In conclusion, we presented a training way for improving isolated glyph classification tasks using additional datasets and preprocessing techniques. Additionally, we compared the results of CNNs and ViTs using different trending approaches. The results of various experiments indicate that ViTs remain behind CNNs on the isolated glyph classification task. Specifically, EfficientNets and ResNets perform well with or without external datasets and data augmentation techniques. Furthermore, Sundanese and Balinese require more datasets for training those deep learning approaches. In addition, it is essential to consider data augmentation and cleaning datasets to bring out better performance. Considering the results, ResNets and EfficientNets may be used to extract features for the palm leaf recognition system in future works.

Acknowledgements. This research is supported by Chinese Academic Science and World Academy of Science President’s Fellowship (CAS-TWAS).

References

1. Burie, J.C., et al.: ICFHR 2016 competition on the analysis of handwritten text in images of balinese palm leaf manuscripts. In: 2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR), pp. 596–601. IEEE (2016)
2. Chang, Y., Jung, C., Ke, P., Song, H., Hwang, J.: Automatic contrast-limited adaptive histogram equalization with dual gamma correction. *IEEE Access* **6**, 11782–11792 (2018)
3. Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.V.: Randaugment: practical automated data augmentation with a reduced search space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 702–703 (2020)
4. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint. [arXiv:2010.11929](https://arxiv.org/abs/2010.11929) (2020)
5. Graham, B., et al.: Levit: a vision transformer in convnet’s clothing for faster inference. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 12259–12269 (2021)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
7. Huang, G., Sun, Yu., Liu, Z., Sedra, D., Weinberger, K.Q.: Deep networks with stochastic depth. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9908, pp. 646–661. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46493-0_39
8. Kesiman, M.W.A., Burie, J.C., Wibawantara, G.N.M.A., Sunarya, I.M.G., Ogier, J.M.: Amadi_lontarset: the first handwritten balinese palm leaf manuscripts dataset. In: 2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR), pp. 168–173. IEEE (2016)
9. Kesiman, M.W.A., et al.: Benchmarking of document image analysis tasks for palm leaf manuscripts from southeast Asia. *J. Imaging* **4**(2), 43 (2018)
10. Kesiman, M.W.A., et al.: ICFHR 2018 competition on document image analysis tasks for southeast asian palm leaf manuscripts. In: 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR), pp. 483–488. IEEE (2018)
11. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint. [arXiv:1412.6980](https://arxiv.org/abs/1412.6980) (2014)
12. Mustafa, W.A., Yazid, H., Jaafar, M.: An improved sauvola approach on document images binarization. *J. Telecommun. Electron. Comput. Eng. (JTEC)* **10**(2), 43–50 (2018)
13. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint. [arXiv:1409.1556](https://arxiv.org/abs/1409.1556) (2014)
14. Steiner, A., Kolesnikov, A., Zhai, X., Wightman, R., Uszkoreit, J., Beyer, L.: How to train your vit? data, augmentation, and regularization in vision transformers. arXiv preprint. [arXiv:2106.10270](https://arxiv.org/abs/2106.10270) (2021)
15. Suryani, M., Paulus, E., Hadi, S., Darsa, U.A., Burie, J.C.: The handwritten sundanese palm leaf manuscript dataset from 15th century. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), vol. 1, pp. 796–800. IEEE (2017)
16. Tan, M., Le, Q.: Efficientnet: rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning, pp. 6105–6114. PMLR (2019)

17. Tolstikhin, I.O., et al.: Mlp-mixer: an all-mlp architecture for vision. In: Advances in Neural Information Processing Systems, vol. 34 (2021)
18. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning, pp. 10347–10357. PMLR (2021)
19. Valy, D., Verleysen, M., Chhun, S., Burie, J.C.: A new khmer palm leaf manuscript dataset for document analysis and recognition: Sleukrith set. In: Proceedings of the 4th International Workshop on Historical Document Imaging and Processing, pp. 1–6 (2017)
20. Wu, H., et al.: Cvt: introducing convolutions to vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 22–31 (2021)
21. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6023–6032 (2019)
22. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: beyond empirical risk minimization. arXiv preprint. [arXiv:1710.09412](https://arxiv.org/abs/1710.09412) (2017)