ASEAN Engineering Journal

Full Paper

ENSEMBLING DEEP CONVOLUTIONAL NEURAL NEWORKS FOR BALINESE HANDWRITTEN CHARACTER RECOGNITION

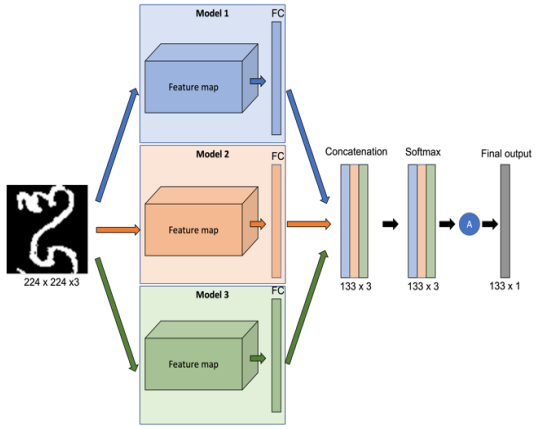
Desak Ayu Sista Dewi[a], Dewa Made Sri Arsa[b*], Gusti Agung Ayu Putri[b], Ni Luh Putu Lilis Sinta Setiawati[a]

[a]Department of Industrial Engineering, Faculty of Engineering, Universitas Udayana, Bali, Indonesia
[b]Department of Information Technology, Faculty of Engineering, Universitas Udayana, Bali, Indonesia

Article history
Received
27 November 2022
Received in revised form
28 April 2023
Accepted
02 May 2023
Published online
31 August 2023

*Corresponding author
dewamsa@unud.ac.id

Graphical abstract



Abstract

While deep learning has proven its performance in various problems and applications, it also opens opportunities in a new way to promote the heterogeneity of cultures and heritages. Balinese script is a cultural heritage in Bali, where it is used to write on palm-leaf manuscripts and contains essential information. Most manuscripts were damaged due to age and lack of maintenance, so a digitalization technique should be developed. In this study, we propose an ensemble of deep convolutional neural networks to recognize the handwritten characters in the Balinese script. We extensively compared various deep convolutional neural network architectures, and the results showed that our ensemble methods achieved the state of the art.

Keywords: Balinese handwritten character, convolutional neural network, ensemble deep learning, recognition, softmax

1.0 INTRODUCTION

Native speakers of languages using the Latin script make up about 70% of the world's population. English, a language written in the Latin script, is also the de facto modern lingua franca, which furthers its dominance. Users of the Chinese script, 10% of Arabic script, and 10% of Devanagari script make up 20%, 10%, and 10% of the global population. For these common writing systems and scripts, standards and tools are mostly developed. Creating extensible standards and universally applicable tools that work with or can be modified for all languages and writing systems has long been a priority for the worldwide research and engineering community. However, low-resource languages, writing systems, and scripts are less likely to benefit from these developments. In order to contribute to a thriving cultural variety and the preservation of cultural heritage, it is crucial to maintain the effort to ensure that all opportunities presented by new technologies are taken advantage of to build the ecosystem.

More than three million people speak the Balinese language, one million use it for daily conversation, and generations of students in Bali's primary schools are learning to speak, read, and write their native tongue. These people use the Balinese script, which has a readership of several million. Bali has a thriving and dynamic history and culture, which has drawn visitors from all over the world and made a considerable economic contribution to the island and Indonesia. Balinese calligraphy may be found and appreciated all around the island since it is often utilised for street signs and signboards. The "lontar", or papers written in Balinese script on palm leaves, have preserved much of Bali's ancient—historical and religious literature. Therefore, creating tools and standards for the Balinese language and script may benefit the Balinese people's ability to preserve their cultural legacy and grow and provide new chances for visitors to engage with the local way of life. Identifying isolated Balinese handwriting from manuscripts on palm leaves is the topic of this study. The goal of this endeavour

is to help find a way to digitise traditional manuscripts so that they may be preserved and distributed.

Even though optical character recognition has been studied for a long time [1], [2], [3], [4], new, possibly more flexible and general techniques have emerged due to machine learning. Convolutional neural networks are one of the most effective techniques used in optical character recognition and have established benchmarks in this and many other fields. Convolutional neural networks extract lower-level to higher-level information at each successive layer [5], thereby validating the potential advantage of deep neural network topologies. For example, Islam et al. [6] proposed a convolutional neural network to recognise characters from the Beowulf Manuscript. Wang et al. [7] developed a method to recognise the chip character from an image to text. Furthermore, Luo et al. [8] and Gan et al. [9] attempted to recognise Chinese characters by proposing new frameworks based on zero-shot learning and pyramid graph transformer, respectively.

For decades, handwritten character recognition has been a focus of attention. The most popular data set is MNist dataset [10]. This data set has ten characters which are numbers from 0 to 9. The current best method for the MNist dataset was proposed by Sanghyeon et al. [11] by utilizing an ensemble convolution neural network with a majority voting technique. The accuracy is 99.9%. In 2021, Kalganova and Dear [12] used a capsule network-based method, and the accuracy was 99.87%. Based on these results, the problem posed by the MNist data set has been overcome since the best error rate is around 0.1%, which corresponds to the error rate of human recognition.

Other scripts, such as the Balinese script, may have received less attention than the Latin script. The Balinese script is a member of the proto-Sinaitic Brahmi family [13]. The Balinese has character similarity to Javanese script, the family of Brahmi script, and Batak and old Sundanese scripts. The Balinese script is used to write traditional Balinese documents on palm leaves. These records often include family trees, details of extraordinary occurrences like plagues, and other historical information. Many Balinese regards these manuscripts as sacred items and are hesitant to alter the palm leaves to read them. Paradoxically, despite the fact that the papers are rarely seen, their quality tends to deteriorate since they need to be appropriately maintained. The digitization of these papers may contribute to the promotion of these historical and cultural artefacts, placing them in a strategic position as objects of interest and value for education, research, culture, and tourism. Kesiman et al. [3] consider the automated recognition of isolated handwritten letters as one of the initial problems in the process, and they explore how to solve this problem.

The physical condition of the script becomes challenging when the script is older for analyzing the Balinese script [4]. The scripts might be turned to yellowing conditions, noise, contrast reduction, and other degradation. Besides the physical condition of Balinese script, the writing style and characters are complex, and there are similarities between characters. The recognition task of Balinese handwritten characters can be identified as an image classification task.

There are two approaches to recognising Balinese characters. The first one uses hand-crafted features. Burie et al. [14] computed the gradient of the image after normalising the image, followed by applying the OTSU method. Kesiman et al. [4] also used gradient-based features and passed the feature to a voting based-classifier. Sudarma et al. [15] calculated several features like the width and length of each character, stopping points, the number of rows and columns number, loop, and horizontal and vertical lines as a semantic feature.

The second approach utilises a convolutional neural network as an end-to-end network. Kesiman et al. [16] compared the convolutional neural network with the hand-crafted feature-based model. In the benchmark by Kesiman et al. [4], the proposed convolutional neural network consists of 3 convolution layers, each followed by batch normalisation and ReLU activation layers. Unlike Kesiman et al. [4], Arsa et al. [17] proposed a residual-based convolutional neural network.

Besides of MNist dataset, the Imagenet dataset is also highly popular for image classification tasks and widely used to build a base model for various tasks, such as object detection, semantic segmentation, or instance segmentation. He et al. [18] made a breakthrough in Imagenet classification by a residual connection idea to prevent the vanishing gradient problem when the convolutional neural network goes deeper. Sandler et al. [19] proposed an inverted residual layer to decrease the number of parameters of the convolutional neural network without reducing the performance. Howard et al. [20] improved MobilenetV2 by searching MobilnetV3 through a network architecture search algorithm. Tan et al. [21] also used a search strategy to build a network called Efficientnet. They proposed eight types of Efficientnet, B0-B7, which have different layers at each stage. Moreover, Tan et al. [22] improved their Efficientnet and proposed the second version by proposing a used inverted linear bottleneck layer. Furthermore, Liu et al. [23] improved the vision transformer called Swin Transformer, which aimed to create a general-purpose backbone.

Moreover, deep convolutional neural neural network have been studied for various studies. Xing et al. [24] used convolutional neural network to extract building area from satelite imagery. Tsai et al. [25] proposed Bisenet which composed of convolutional layers for real time detection purpose. Moudgil et al. utilized capsule networks to recognize devanagari handwritten characters [26]. Cao et al. [27] developed a UNet-like model which purely constructed using transformer modules for medical image segmentation.

Recognising Balinese handwritten characters on palm-leaf manuscripts poses a difficult challenge due to the age of the documents. These manuscripts are considered sacred by the Balinese and have not been well-maintained, resulting in yellowing, decreased contrast, increased noise and complexity, and variations in writing styles among different documents. While traditional image processing and machine learning techniques could be utilised for character recognition, their effectiveness may be limited to certain conditions. Therefore, in this study, we propose a convolutional neural network- based method to recognise Balinese handwritten characters on palm-leaf manuscripts. We developed an ensemble convolutional neural network that has the flexibility to change the base models. Our base models were selected experimentally, and our based models were initialised by pretrained weights, which have been trained on the Imagenet dataset for the classification task. To our knowledge, this is the first study to benchmark the Balinese handwritten character recognition on a convolutional neural network. The base models were first trained separately and combined with their probabilities on inference to produce the final classification.

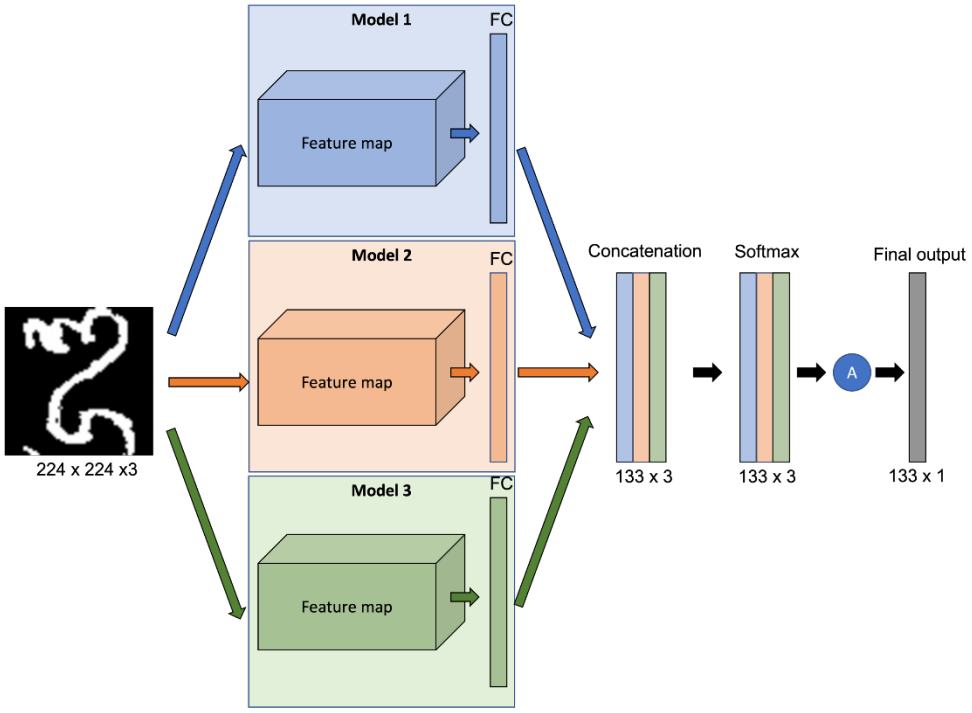


Figure 1 The proposed design of ensemble model

The remaining portions of the paper are organised in the following fashion. The proposed technique is outlined in Section 2. Section 3 describes the experimental setup, which includes the data set. In Section 4, the findings of comparative performance analysis are discussed. We conclude our findings in section 5.

2.0 PROPOSED METHOD

The proposed ensemble framework is shown in Figure 1. Our ensemble framework was consisted of three deep learning models. The deep learning models were selected through experiment. Four based models were chosen, such as Resnet, Efficientnet, Mobilenet, and Swin transformer families. Firstly, we fine-tuned those pre-trained deep convolutional neural networks on the handwritten dataset. Each base model was already pre-trained on the ImageNet dataset, and for each pre-trained model, we re-initialised the fully connected layer (FC) to fit the dataset classes. All models were optimised using stochastic gradient descent, and the loss function was categorical cross-entropy loss. The categorical cross entropy loss $\mathcal{L}_{c\varepsilon}$ can be defined as follows.

$$\mathcal{L}_{c\varepsilon} = -\sum_{i=1}^{c} T_i log\,(Y_i) \tag{1}$$

where T_i and Y_i are the groundtruth and the prediction of ith class respectively.

Secondly, we selected three models which have the best performance by measuring the precision, recall, and f1-score (please see Table I). Thirdly, for each image, we calculated the probability of the image belonging to each class for each model. Suppose FC_i is the output of the model i; then, in the fourth step, we concatenated those FC_i as formalized in equation 2.

$$R = \Phi\,(FC_1, FC_i, \ldots, FC_n) \tag{2}$$

where $\mathrm{R} \in \mathrm{R}^{\mathrm{cxn}}$, c is the number of class and n is the number of models and Φ is the concatenation function. In our study, c is 133 and n is 3. Moreover, in the fifth step, we transformed R into probability which summed up to 1 for each column using equation 3.

$$P = \frac{e^{R^n}}{\sum_{i=1}^{K} e^{R_i^n}} \tag{3}$$

P is the result of softmax activation function which receives vector R as input. After that, we computed the average predicted probability for each class using equation 4.

$$A_j = \frac{1}{n}\sum_{i=1}^{n} P_{ji} \tag{4}$$

where $\mathrm{A_j}$ is the final output probability for class j. Lastly, the final prediction Y_{fin} is finalised by finding the highest probability class on A_j which is defined by $Y_{fin} = argmax(A)$.

3.0 EXPERIMENT SETUP

The dataset used in this study was used in the third challenge at the International Conference on Frontiers in Handwriting

Recognition (ICFHR) in 2016. The experiments were done in RTX 2060, and the codes were developed using PyTorch. ICFHR only provides the training and testing data. So, we split the training data into training and validation data. The size of the validation data is 20% of the original training data. We observed that the data is imbalanced. A weighted mechanism was built in the training phase to avoid this issue. The weight for each class can be calculated as shown in equation 5.

$$w_{C_i} = \frac{1}{n_{C_i}} \tag{5}$$

Where C_i is the i^{th} class, n_{C_i} is the number of data for the i^{th} class, and w_{C_i} is the weight of the i^{th} class. The lower the number of data for the class, the higher the weight will be. It means that the classes are challenging.

Moreover, all images were transformed into MNist's style, black and white and squared shape. The original images have different sizes of images following the shape of Balinese characters and different intensities. So, firstly we converted the images into black and white. Secondly, we put padding with black colour to make the image squared shape. Lastly, we resized the images into 224 to fulfil the required image size to use the pre-trained deep learning on PyTorch.

Four measurements were utilized to measure the performance of the proposed method. The first three are precision, recall, and f1-score and can be calculated using equations 6, 7, and 8, respectively.

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

$$F1score = \frac{2TP}{2TP + FP + FN} \tag{8}$$

where TP, FP, and FN are true positive, false positive, and false negative. The final precision, recall, and f1-score were calculated by averaging precision, recall, and f1-score per class because of the imbalanced problem. The fourth measurement is accuracy. We compared our accuracy with the previous benchmarks, and the accuracy can be computed the same as precision.

Stochastic gradient descent was used to optimize network optimization with 0.001 learning, 0.9 momentum, and 50 epochs. We also adjusted the learning rate every seven epochs by 0.1.

4.0 RESULTS AND DISCUSSIONS

Table 1 shows the results on the accuracy, precision, recall, and f1-score on various pre-trained models. From the Resnet family, Resnet 34 has the highest score on all metrics, while Resnet 101 has the lowest accuracy. In terms of precision, Resnet 101 has a higher score than Resnet18, but a slightly lower recall score means that Resnet 18 produces a lower false negative than Resnet 101.

In the Efficientnet family, Efficientnet V2 L has the best accuracy, recall, and f1-score. Its precision is slightly lower than Efficientnet V2 S by 0.0026 of margin. The precision scores for Efficientnet B0-B4 are below 0.80, which shows that the models have more miss classification—the lowest precision produced by Efficientnet B4. Furthermore, Moblenet V2 is better than Mobilenet V3, small and large, while Swin Transformer B is better than Swin Transformers S and T.

Figures 2-5 presents the training and loss graphs for the best models for each variant. There is no indication of over-fitting because there is not much difference between the training and validation graphs, both on the accuracy and loss graphs. Resnet34 interestingly has ±80% accuracy in the validation phase in the first epoch, while others start from ±60% and then jump aggressively in the next epoch.

Moreover, of the top three models (Resnet34, Efficientnet V2 L, and Swin Transformer B), Efficientnet V2 L has a large number of parameters to achieve good performance. Efficientnet V2 L's parameters are approximately 5 times that of Resnet34 and approximately 1.3 times that of Swin Transformer B.

We generated four ensemble models based on the results in Table I. The first ensemble model was composed of the top three best accuracy models, which are Resnet34, Efficientnet V2 L, and Swin Transformer B. The second ensemble model was constructed by the top three best precision models, which are Resnet34, Resnet50, and Efficientnet V2 S. Then, the third ensemble model was composed of the top three recall models: Efficientnet V2 L, Swin Transformer B, and Swin Transformer T. Moreover, the last ensemble model was composed of the top three f1-score models, which are Resnet34, Resnet50, and Swin Transformer B.

The results can be seen in Table 2. All constructed ensemble models perform better than the single method, as shown in Table I. The improvement is approximately 1% in accuracy, and 2% in precision, recall, and f1-score. The best model based on accuracy is Ensemble 1. However, we preferred Ensemble 3 as the best model with the highest score of f1-score. As mentioned before, the precision and recall were calculated by averaging precision and recall between classes. The f1-score was the harmony between precision and recall. Therefore, Ensemble 3 has better performance as a result. In terms of processing time, Ensemble 3 is quite slow compared to the others because each model has a large number of parameters. However, an accurate model is preferable for recognising Balinese handwritten script because it might change the meaning of the manuscript.

Table 1 Performance on various pretrained models for isolated Balinese character recogntion

Model name	Accuracy	Precision	Recall	F1-score	Number of parameters(Millions)	FLOPS (G)	FPS (CPU)	FPS (GPU)
Resnet 18	0.9148	0.8133	0.8495	0.8189	**11.24**	**1.819**	**103.82**	**521.36**
Resnet 34	**0.9258**	**0.8465**	**0.8711**	**0.8464**	21.52	3.671	68.27	303.1
Resnet 50	0.9192	0.8302	0.8668	0.8377	23.78	4.11	46.11	202.43
Resnet 101	0.9109	0.8183	0.8452	0.8199	42.77	7.832	26.49	103.81
Efficientnet B0	0.8901	0.7739	0.8393	0.7915	**4.18**	**0.401**	**59.47**	**152.66**
Efficientnet B1	0.8973	0.7789	0.8409	0.7936	6.68	0.591	43.59	105.78
Efficientnet B2	0.9034	0.7927	0.8382	0.8017	7.89	0.681	42.12	104.42
Efficientnet B3	0.8943	0.7731	0.8382	0.792	10.90	0.992	34.81	91.75
Efficientnet B4	0.8463	0.6959	0.7785	0.7139	17.79	1.543	27.53	75.21
Efficientnet B5	0.9116	0.8083	0.8589	0.8227	28.61	2.411	19.63	59.61
Efficientnet B6	0.9139	0.8049	0.848	0.8155	41.04	3.43	15.83	51.23
Efficientnet B7	0.9161	0.8218	0.8622	0.8309	64.13	5.265	13.07	42.84
Efficientnet V2 S	0.9163	**0.8281**	0.8688	0.8374	20.35	2.876	25.59	67.52
Efficientnet V2 M	0.9155	0.8243	0.8673	0.8338	53.03	5.406	16.38	45.68
Efficientnet V2 L	**0.9206**	0.8255	**0.8763**	**0.8374**	117.40	12.309	9.75	33.37
Mobilenet V2	**0.9073**	**0.8136**	**0.8409**	**0.8133**	2.39	0.313	103.05	267.77
Mobilenet V3 small	0.8555	0.7125	0.7815	0.7249	**1.65**	**0.058**	**138.83**	**247.36**
Mobilenet V3 large	0.8721	0.7434	0.7979	0.7519	4.37	0.225	101.77	203.71
Swin Transformer S	0.9196	0.8213	0.8706	0.8349	48.94	8.768	15.32	61.33
Swin Transformer B	**0.9202**	**0.827**	**0.8759**	0.8396	86.68	15.467	12.23	60.99
Swin Transformer T	0.9131	0.8189	0.8718	**0.8333**	**27.62**	**4.509**	**30.59**	**121.1**

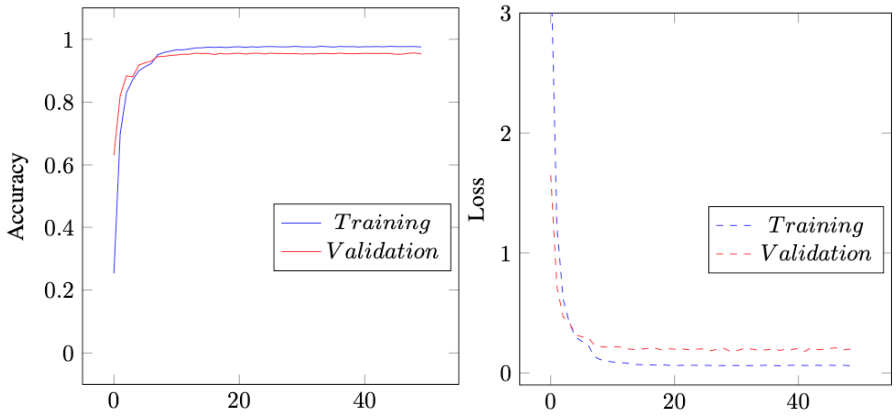


Figure 2 The graph for accuracy and loss on training and validation phase for Swin Transformer B

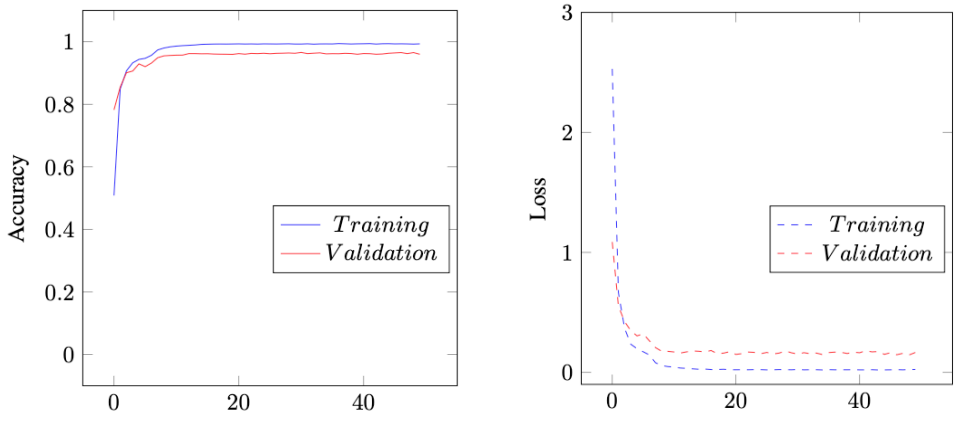


Figure 3 The graph for accuracy and loss on training and validation phase for ResNet34

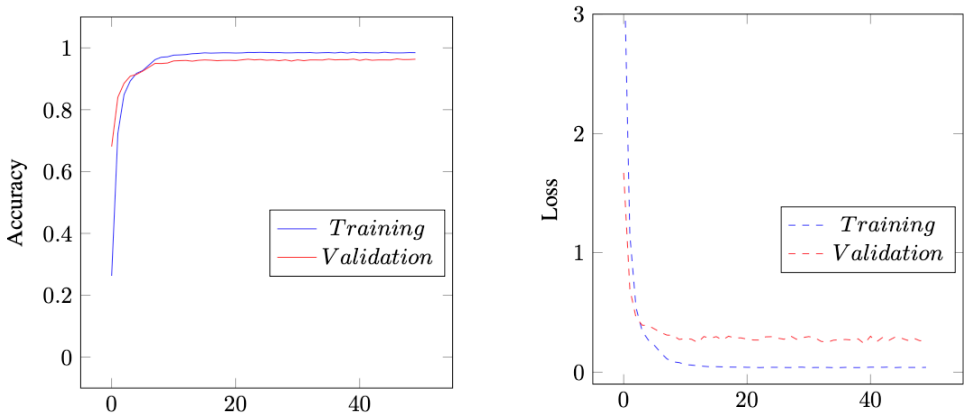


Figure 4 The graph for accuracy and loss on training and validation phase for EfficientNet V2L

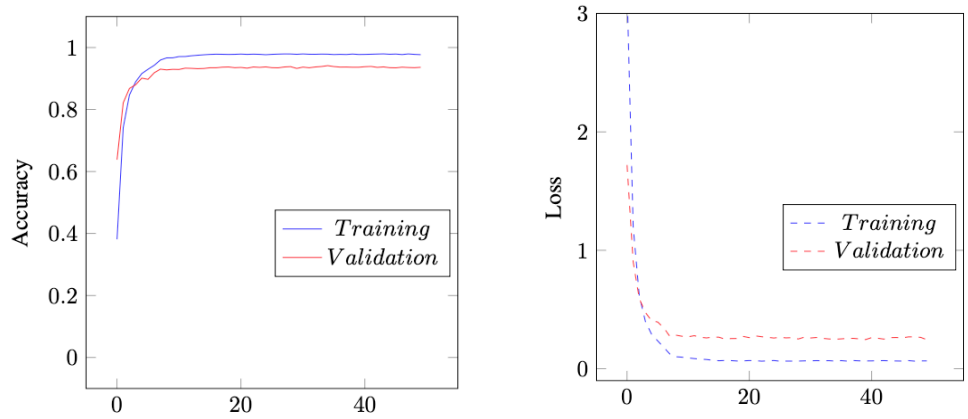


Figure 5 The graph for accuracy and loss on training and validation phase for MobileNet V2

The comparison with previous studies can be found in Table 2 These results were reported from the papers and used the original image size. Most previous research used handcrafted features and only measured the accuracy. The highest accuracy in the previous study was reported by Kesiman et al. [18] with 88.39% of accuracy. Our ensemble methods have better performance compared to it. Arsa et al. [19] have better precision than our ensemble but lower recall and f1-score, which means their method has more false negative predictions.

Table 2 Ensemble models performance and previous studies results

Model name	Accuracy	Precision	Recall	F1-score
Ensemble 1	**0.9360**	0.8532	0.889	0.8598
Ensemble 2	0.9329	0.8469	0.8889	0.8577
Ensemble 3	0.9343	**0.8644**	**0.894**	**0.8699**
Ensemble 4	0.9341	0.8583	0.8915	0.8635
Handcrafted Feature with k-NN [18]	0.8516	-	-	-
Handcrafted Feature with NN [20]	0.8551	-	-	-
Handcrafted Feature with UFL + NN [20]	0.8563	-	-	-
CNN 1 [18]	0.8431	-	-	-
CNN 2 [4]	0.8539	-	-	-
ICFHR G1: VCMF [21]	0.8744	-	-	-
ICFHR G1: VMQDF [21]	0.8839	-	-	-

5.0 CONCLUSION

In this study, we proposed an ensemble of deep convolutional neural networks. Our ensembles consist of the three best pre-trained deep learning from extensive experiments. The experiment results present that ResNet 34 outperformed other pre-trained deep learning on single deep learning-based classification. Our ensemble method outperformed all single deep learning-based models and surpassed previous studies. Based on the designed ensembles, the third ensemble achieved the best performance in precision, recall, and f1-score by 0.8644, 0.8940, and 0.8699, respectively.

Even though this study presents an outstanding result, some rooms exist to make improvements. The first one is to put some preprocessing to minimize the noise. If the script is clear, it will be easier to recognize through training. Secondly, ensembling is one way to get higher performance, but it also increases the complexity of the models. So, deploying the model in a real-time process is not reasonably possible. Therefore, we can transfer the knowledge from the ensemble to the lighter model through knowledge distillation technique.

Acknowledgement

This study was funded by Universitas Udayana under Penelitian Unggulan Program Studi no: B/116/Un14.2.5.11/PT.01.03/2021.

References

[1] T. Wang, Z. Xie, Z. Li, L. Jin, and X. Chen, 2019. "Radical aggregation network for few-shot offline handwritten chinese character recognition," *Pattern Recognition Letters.* 125:821–827.

[2] J. I. Olszewska, 2015. "Active contour based optical character recognition for automated scene understanding," *Neurocomputing.* 161:65–71.

[3] M. W. A. Kesiman, J.-C. Burie, G. N. M. A. Wibawantara, I. M. G. Sunarya, and J.-M. Ogier, 2016. "Amadi lontarset: The first handwritten balinese palm leaf manuscripts dataset," in *15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, IEEE. 168–173.

[4] M. W. A. Kesiman, D. Valy, J.-C. Burie, E. Paulus, M. Suryani, S. Hadi, M. Verleysen, S. Chhun, and J.-M. Ogier, 2018. "Benchmarking of document image analysis tasks for palm leaf manuscripts from southeast asia," *Journal of Imaging.* 4(2): 43.

[5] D. Matthew and R. Fergus, 2014. "Visualizing and understanding convolutional neural networks," in *Proceedings of the 13th European Conference Computer Vision and Pattern Recognition, Zurich, Switzerland.* 6–12.

[6] M. A. Islam and I. E. Iacob, 2023. "Manuscripts character recognition using machine learning and deep learning," Modelling. 4(2):168– 188.

[7] X. Wang, Y. Li, J. Liu, J. Zhang, X. Du, L. Liu, and Y. Liu, 2022. "In- telligent micron optical character recognition of dfb chip using deep convolutional neural network," *IEEE Transactions on Instrumentation and Measurement*. 71: 1–9.

[8] G.-F. Luo, D.-H. Wang, X. Du, H.-Y. Yin, X.-Y. Zhang, and S. Zhu, 2023. "Self-information of radicals: A new clue for zero-shot chinese character recognition," *Pattern Recognition.* 140: 109598.

[9] J. Gan, Y. Chen, B. Hu, J. Leng, W. Wang, and X. Gao, 2023. "Characters as graphs: Interpretable handwritten chinese character recognition via pyramid graph transformer," *Pattern Recognition.* 137: 109317.

[10] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, 1998. "Gradient-based learning applied to document recognition," *Proceedings of the IEEE.* 86(11): 2278–2324.

[11] S. An, M. Lee, S. Park, H. Yang, and J. So, 2020. "An ensemble of simple convolutional neural network models for mnist digit recognition," arXiv preprint arXiv:2008.10400.

[12] A. Byerly, T. Kalganova, and I. Dear, 2021. "No routing needed between capsules," *Neurocomputing.* 463: 545–553.

[13] D. Ghosh, T. Dube, and A. Shivaprasad, 2010. "Script recognition—a review," *IEEE Transactions On Pattern Analysis And Machine Intelligence*. 32(12): 2142–2161.

[14] J. Burie, M. Coustaty, S. Hadi, M. W. A. Kesiman, J. Ogier, E. Paulus, K. Sok, I. M. G. Sunarya, and D. Valy, 2016. "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).* 596–601.

[15] M. Sudarma and I. W. A. Surya, 2014. "The identification of balinese scripts' characters based on semantic feature and k nearest neighbor," *International Journal of Computer Applications*. 91(1).

[16] M. W. A. Kesiman, S. Prum, J.-C. Burie, and J.-M. Ogier, 2016. "Study on feature extraction methods for character recognition of balinese script on palm leaf manuscript images," in *2016 23rd International Conference on Pattern Recognition (ICPR).* IEEE. 4017–4022.

[17] D. M. S. Arsa, G. A. A. Putri, R. Zen, and S. Bressan, 2020. "Isolated handwritten balinese character recognition from palm leaf manuscripts with residual convolutional neural networks," in *2020 12th International Conference on Knowledge and Systems Engineering (KSE).* IEEE. 224–229.

[18] K. He, X. Zhang, S. Ren, and J. Sun, 2016. "Deep residual learning for image recognition," in *Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition*. 770–778.

[19] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, 2018. "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition*. 4510–4520.

[20] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan et al., 2019. "Searching for mobilenetv3," in *Proceedings Of The IEEE/CVF International Conference On Computer Vision.* 1314–1324.

[21] M. Tan and Q. Le, 2019. "Efficientnet: Rethinking model scaling for con- volutional neural networks," in *International conference on machine learning. PMLR.* 6105–6114.

[22] M. Tan and Q. Le, 2021. "Efficientnetv2: Smaller models and faster training," in I*nternational Conference on Machine Learning, PMLR.* 10096–10106.

[23] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, 2021. "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 10012–10022.

[24] J. Xing, Z. Ruixi, R. Zen, D. M. S. Arsa, I. Khalil, and S. Bressan, 2019. "Building extraction from google earth images," in *Proceedings of the 21st International Conference on Information Integration and Web-based Applications & Services*. 502–511.

[25] T.-H. Tsai and Y.-W. Tseng, 2023. "Bisenet v3: Bilateral segmentation network with coordinate attention for real-time semantic segmentation," *Neurocomputing.* 532: 33–42.

[26] A. Moudgil, S. Singh, V. Gautam, S. Rani, and S. H. Shah, 2023. "Handwritten devanagari manuscript characters recognition using capsnet," *International Journal of Cognitive Computing in Engineering*. 4: 47–54.

[27] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, 2023. "Swin-unet: Unet-like pure transformer for medical image segmenta- tion," in *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel,* October 23–27, 2022, Proceedings, Part III. Springer. 205– 218.