Design of a Deep Learning Model for Bronze Dagger Morphology Classification

Kyungmin Kim [1], Jeongmin Yu[1]

[1] Dept. of Digital Heritage, Korea National University of Heritage, Republic of Korea – gimgyeongmin254@naver.com, jmyu@knuh.ac.kr

Keywords: Bronze Dagger, Deep Learning, Classification, Convolutional Neural Network, Image Analysis

Abstract

In archaeology, buried cultural artifacts serve as important material evidence for distinguishing historical periods, and inferring the morphology and chronology of artifacts excavated from archaeological sites is a crucial task. Buried cultural artifacts can be used as fundamental data for inferring the characteristics of archaeological sites and the scale of past groups that utilized the sites through their morphology and chronological context. In particular, bronze dagger, which are among the buried cultural artifacts excavated from prehistoric sites, exhibit different morphologies according to their periods, making the morphological classification of bronze dagger a key indicator for determining the chronology of archaeological sites. However, the various forms of bronze dagger excavated from the Korean Peninsula to northeastern China show limitations for manual classification by archaeological researchers, and the ambiguous characteristics where forms are not clearly distinguishable create problems in securing consistent objectivity in bronze dagger chronological classification. To overcome these limitations, this paper proposes a framework for automatically classifying bronze dagger morphology using deep learning-based image classification models and quantitative results for bronze dagger classification tasks.

1. Introduction

Buried cultural artifacts serve as important material evidence for chronological classification, and the classification of the morphology and period of use of artifacts excavated from archaeological sites is a core task in archaeology. Generally, archaeology employs typological research methods for the chronological classification of buried cultural artifacts. This research approach estimates the sequential relationships of formal changes based on the assumption that artifacts evolve from simple forms to gradually more complex forms (Jeong et al., 2022). However, such research methods have limitations in that classification criteria are difficult to determine when the overall pattern of artifact changes is not discovered, and securing objectivity in classification criteria is challenging due to differences in evolutionary rates (Lee, 1983; Choi, 1984). In Korean archaeology, bronze dagger present particular difficulties in typological classification due to ambiguous classification criteria. Bronze daggers are bronze weapons or ritual implements used in the Korean Peninsula and northeastern China during the Bronze Age to Early Iron Age period, and are generally classified into ‘*Bipahyeong*’ and ‘*Sehyeong*’ types according to blade morphology. ‘*Bipahyeong*’ type bronze dagger are characterized by broad and short blades resembling the Chinese weapon *Bipa*, while ‘*Sehyeong*’ type bronze dagger feature narrow and long blades. Bronze dagger that cannot be classified into either category due to unclear blade morphology definition are classified as transitional bronze dagger. In Korean archaeology, the morphology of bronze dagger serves as an indicator for distinguishing between Bronze Age cultural spheres and the Early Iron Age. Therefore, the chronological period of archaeological sites is determined based on the morphology of excavated bronze dagger (Heo, 2016).

Bronze dagger excavated from the Korean Peninsula and northeastern China function as key indicator artifacts representing Bronze Age and Iron Age cultural spheres respectively, and the morphology of bronze dagger serves as a major criterion for determining the chronology of archaeological sites. However, regarding transitional bronze dagger that represent the transitional phase from ‘*Bipahyeong*’ type bronze dagger to ‘*Sehyeong*’ type bronze dagger, the typological research approach itself has limitations in that it can vary depending on the method of interpreting detailed characteristics of artifacts, leading to differences of opinion among scholars (Kim, 2007). In particular, bronze dagger extensively excavated from the Korean Peninsula to northeastern China exhibit distinct morphological continuity, making it difficult to apply consistent classification criteria through typological approaches alone, which can create difficulties in securing objectivity in archaeological interpretation. Table 1 presents a classification system for bronze dagger excavated from the Korean Peninsula to northeastern China, showing morphological examples of ‘*Bipahyeong*’, ‘*Sehyeong*’, and transitional bronze dagger respectively.

‘*Bipahyeong*’ type bronze dagger	‘*Sehyeong*’ type bronze dagger	transitional bronze dagger
		

Table 1. Types of bronze dagger excavated from the Korea to northeast China.

Therefore, this paper proposes an automated bronze dagger morphological analysis and typological classification system utilizing artificial intelligence technology to overcome the limitations of existing typological methods and solve the problem of transitional bronze dagger classification. To achieve this goal, the paper proposes a combination of deep learning-based binary classification models using bronze dagger drawing images from archaeological reports, and examines the possibility of establishing objective and practical classification criteria by experimenting with various model architectures and fusion techniques. The purpose is to support highly reliable

morphological analysis in archaeological fieldwork and provide quantitative indicators for bronze dagger classification tasks.
The paper is structured as follows: literature review (Section 2), proposed CNN fusion methodology (Section 3), experimental framework and data preparation (Section 4), classification outcomes (Section 5), and analysis of results with concluding remarks (Section 6).

2. Related Work

2.1 Deep Learning Methods in Digital Cultural Heritage

Recently, researches using classification with AI in cultural heritage studies have steadily increased, demonstrating that AI is rapidly spreading in the cultural heritage field (Gîrbacia, 2024). In the digital heritage field, AI is achieving groundbreaking results in automatic manuscript classification and knowledge extraction, automatic 3D reconstruction, cultural heritage asset identification, choreographic sequence extraction from traditional dance, and improving accessibility to museum and online cultural heritage content.

AI application has proven useful for improving data management, preservation, and sharing performance of cultural heritage compared to traditional methods. AI automates damage detection, object recognition, data analysis, 3D reconstruction, and beyond simple automation, is expanding its application scope to specialized areas that enhance understanding and interaction with cultural artifacts.

In particular, convolutional neural network (CNN) has shown great potential for intangible heritage data preservation and management by effectively capturing both spatial and temporal features simultaneously (LeCun, 2015). CNN-based classification has shown promising results in reconstruction, remote sensing, and collection management in the archaeological field, and is demonstrating excellent effectiveness in cultural heritage image classification and retrieval.

However, CNN-based classification has limitations in that efficiency may decrease when processing complex or morphologically diverse cultural heritage assets. Additionally, practical field application is constrained by the need for vast training datasets and specialized hardware.

2.2 Deep Learning Methods in Archaeological Research

Recently, attempts to introduce deep learning technology into the field of archaeology have been made, leading to active research on image classification and pattern recognition of archaeological artifacts and sites. These studies primarily focus on cultural heritage image classification, morphological recognition, chronological estimation, and artifact type classification, and tend to develop around demonstrating technical feasibility or improving recognition rates for specific artifact types.

As representative examples, Kim (2024) explored the applicability of deep learning to archaeology by attempting chronological classification of ancient roof-end tiles using deep learning-based models. Jeong et al. (2022) performed unsupervised clustering combining CNN and DBSCAN for typological classification and chronological estimation of pedestal dishes. Additionally, Christian et al. (2014) proposed an automated archaeological typology system using machine learning for German Bronze Age pottery. These studies successfully demonstrate the potential for introducing AI technology into the archaeological field by attempting quantification of archaeological data through various technical approaches including high-resolution image collection, transfer learning, data augmentation, clustering, and 3D-based measurement.

However, several areas for improvement exist in terms of practical applicability in field settings. Most studies focus on technical performance validation, resulting in relatively insufficient linkage with the academic establishment of archaeological theory or classification criteria. Additionally, standardization of archaeological artifact classification systems and design as tools that can be immediately applied in actual archaeological practice remain areas requiring further consideration. In particular, while 3D scanning-based analysis, such as in Christian et al. (2014), represents an excellent approach in experimental design, additional improvements are needed for practical tool development in terms of the speed and accessibility required in archaeological fieldwork.

This paper differs from existing research by focusing on practical applicability with emphasis on the following aspects. First, it aims to establish standardized criteria for bronze dagger morphological classification utilizing deep learning, targeting bronze dagger which are key artifacts for distinguishing prehistoric cultural spheres. Second, this paper does not merely stop at developing models. To examine practical applicability, 42 transitional bronze dagger that are difficult to classify morphologically are collected and applied to the model, confirming that the proposed model can practically assist archaeological interpretation by quantifying and presenting the '*Bipahyeong*' and '*Sehyeong'* probabilities for each artifact. This represents a distinction from existing research that focused on technical demonstration or accuracy improvement, as this paper focuses on contributing to solving substantial archaeological fieldwork problems.

3. Proposed Method

This paper presents deep learning binary classification models capable of bronze dagger classification that aligns with the 'typological' research method, which is one of the archaeological artifact classification criteria. However, it is confirmed that limitations exist in using each model individually in archaeological fieldwork. Therefore, after comparing and analyzing the structural characteristics and performance of existing models, this paper proposes a 'proposed fusion model' combination that can maximize the advantages of these models while mutually compensating for their disadvantages. A total of six existing binary classification models are used in the experiments, as follows: ResNet-50, DenseNet-121, EfficientNetV2, RepVGG, ConvNeXT2.

3.1 Pretrained Model for Image Classification

Among the selected deep learning binary classification models, ResNet-50 is an architecture that solves the gradient vanishing problem of deep neural networks through residual connections and enables stable learning even with 50 layers. It provides the advantages of transfer learning through the utilization of ImageNet pre-trained weights and strengths in detailed feature extraction from high-resolution images (Kaiming et al., 2016). In the experiments, the final fully connected layer is modified for binary classification based on ResNet50_Weights.IMAGENET1K_V2, and 50% dropout is applied to prevent overfitting. However, ResNet-50 has limitations in learning subtle morphological differences in bronze dagger drawings due to progressive loss of initial feature information (He et al., 2015). Additionally, it shows tendencies toward degraded generalization performance due to high parameter requirements and low validation accuracy compared to training accuracy. To compensate for these limitations and achieve efficient bronze dagger feature extraction, the DenseNet-121 model is applied in subsequent experiments.

DenseNet-121 is a model that maximizes feature reuse and fundamentally solves the gradient vanishing problem through a dense connection structure where each layer receives direct connections from all previous layers (Huang et al., 2016). It achieves superior performance with fewer parameters than ResNet and provides rich representation while preventing overfitting even with limited datasets, making it suitable for fine morphological classification of bronze dagger drawings. It consists of 4 Dense Blocks and 3 Transition Layers, and transfer learning is performed based on DenseNet121_Weights. DEFAULT. The final classification layer is modified for binary classification and 50% dropout is applied. However, DenseNet-121 has limitations including increased memory usage due to dense connections, slow learning speed, and constraints on mobile environment applications. It may show computational efficiency limitations in real-time classification tasks in archaeological fieldwork, leading to the selection of EfficientNetV2, which mitigates these issues, as the subject of subsequent experiments.

EfficientNetV2 is a model that improves the limitations of existing EfficientNet to provide faster learning speed and higher parameter efficiency (Tan et al., 2021). It improves speed in early learning stages through a combination of Fused-MBConv and MBConv blocks, and induces efficient learning by gradually increasing input image size through a Progressive Learning strategy. EfficientNetV2-S shows excellent performance even with limited computational resources and is suitable for precise morphological classification of bronze dagger drawings. It is implemented with a hierarchical structure consisting of Stem, 6 Stages, and Head, and is directly designed from scratch to optimize for bronze dagger binary classification. The final classification layer is redesigned for 2 classes and 20% dropout is applied. However, EfficientNetV2 still has a relatively complex architecture that may hinder practical field application, and the possibility of overfitting exists due to excessive complexity relative to small-scale datasets. Considering these limitations, RepVGG model is selected as the subject of subsequent experiments to secure appropriate model complexity and generalization performance for small-scale datasets with a simple structure.

RepVGG is a CNN architecture based on structural re-parameterization techniques that secures high representation through a multi-branch structure during the training phase and converts to a single-path structure during the inference phase to maximize execution efficiency as a VGG-style simple network (Ding et al., 2021). The RepVGG-A0 variant is used with width multipliers set to [0.75, 0.75, 0.75, 2.5] to consider the balance between model size and performance. However, RepVGG tends to focus on local features rather than extracting diverse scale features due to its use of single kernels, resulting in limitations in reflecting structural relationships of the overall bronze dagger morphology or subtle morphological differences.

Existing models may suffer from loss of initial information due to hierarchical structures and face constraints in adjusting input image sizes due to fixed architectures. To overcome these limitations, ConvNeXtV2 is introduced. ConvNeXtV2 is an architecture that systematically analyzes the design principles of Vision Transformers and applies them to CNNs (Woo et al., 2023). Modern techniques such as global response normalization (GRN) and fully convolutional masked autoencoder (FCMAE) pre-training are applied, showing stable learning and excellent generalization performance even with small datasets. Considering the balance between efficiency and performance, the ConvNeXt-Tiny variant is used, which consists of a 4-stage hierarchical structure with block depths of [3, 3, 9, 3] and channel numbers of [96, 192, 384, 768], making it suitable for effectively extracting and classifying complex visual features of bronze dagger drawings.

3.2 Proposed Fusion Model Based on ResNet and DenseNet

The ‘bronze dagger typological classification model’ proposed in this paper adopts a feature fusion framework based on a multi-stream architecture method that learns through a systematic workflow from dataset construction through data preprocessing, data augmentation, and automated feature extraction to final prediction and validation. This framework is designed as a structure that combines multiple models learning different feature spaces by extracting image features based on ResNet and DenseNet to overcome the weaknesses of each model. Figure 1 shows the overall workflow of the proposed fusion model.

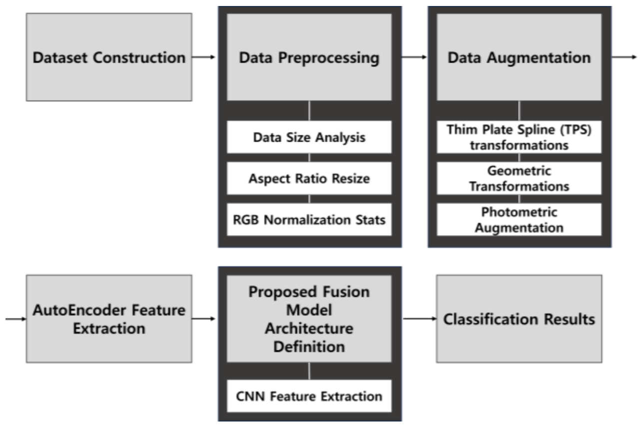


Figure 1. Workflow of the proposed fusion model.

Based on this fusion framework, the data ‘prediction accuracy’ is evaluated through classification performance indicators based on confusion matrices of the six aforementioned models, confirming that the combination with the best compatibility and performance when used in fusion is the model combining the feature points of ResNet-50 and DenseNet-121. Therefore, based on these experimental results, appropriate adapter layers are designed to connect ResNet-50 and DenseNet-121, and the final fusion framework is established by applying staged training, regularization techniques, and additional data augmentation techniques such as data augmentation. Figure 2 shows the workflow of the feature block fusion part of ResNet-50 and DenseNet-121 models within the workflow of the proposed fusion model.

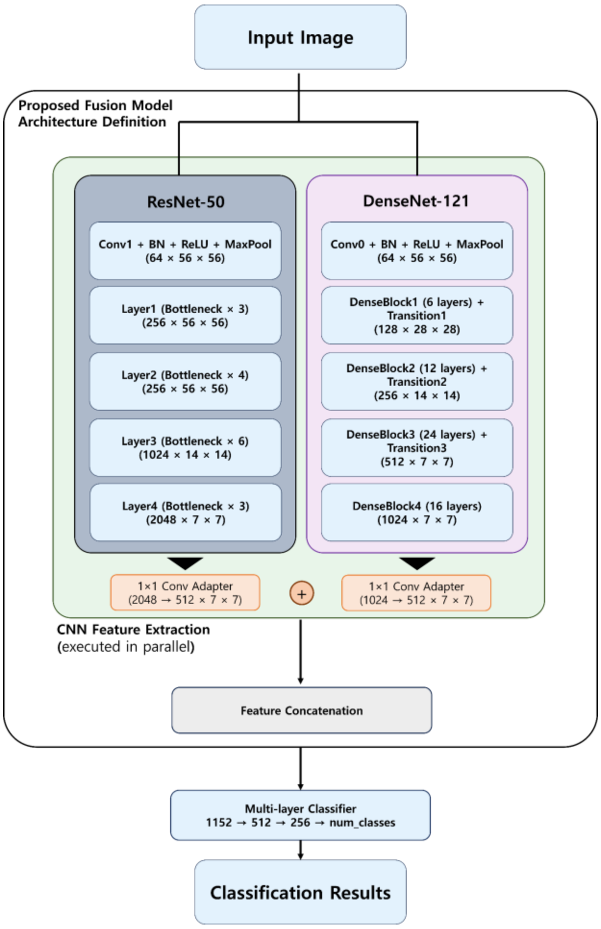


Figure 2. Proposed fusion model architecture definition.

4. Experiment

4.1 Experimental Conditions

The experimental environment is established as follows. The basic environment setup (cloud environment) uses 'Google Colab' environment, the device is Google's 'A100 GPU', and the programming language is 'Python 3'. Hyperparameters are commonly set for all models as 'batch size: 32, epoch: 15, learning rate: 0.001' to suit binary classification. Additionally, all models used in the experiments undergo a common data preprocessing procedure.

4.2 Dataset Preparation and Preprocessing Procedures

The dataset for this paper is constructed using 'drawing' images that archaeologically record the morphology of bronze dagger excavated from the Korean Peninsula to northeastern China. The drawing images are constructed based on reliable sources from bronze dagger-related archaeological papers and excavation reports published from 1968 to 2024.

The dataset consists of a total of 492 items, with 231 '*Bipahyeong*' type bronze dagger and 261 '*Sehyeong*' type bronze dagger. Each dataset is constructed with bronze dagger whose morphological classification has been archaeologically confirmed without disagreement among scholars. Figures 3 and 4 show partial examples of the '*Bipahyeong*' type bronze dagger and '*Sehyeong*' type bronze dagger drawing datasets.

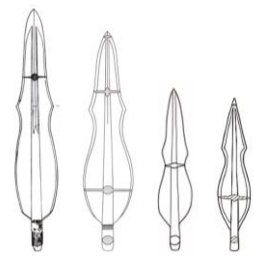


Figure 3. Example of '*Bipahyeong*' type bronze dagger dataset.

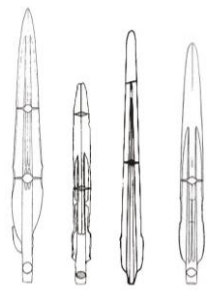


Figure 4. Example of '*Sehyeong*' bronze dagger dataset.

The following data preprocessing procedures are applied to all models used in the experiments. After analyzing the image sizes in the dataset, the mean/median/range of width, height, and aspect ratio are calculated, and a resize strategy (224x224) is applied to achieve optimal aspect ratio while preventing data distortion. Subsequently, the mean and standard deviation for each RGB channel of the data images are calculated to analyze RGB color distribution, and a sharpening filter is applied to solve the problem of blurry outlines. Based on the recalculated statistical values from the improved data through these preprocessing procedures, optimized RGB normalization values are applied to the entire dataset to clearly highlight the morphological features of artifact drawings. Finally, the entire data augmentation process is conducted using Flip, Rotation, and Scaling only on the training dataset.

For model training and evaluation, experiments are conducted by dividing a total of 492 bronze dagger drawing image data, consisting of '*Bipahyeong*' type bronze dagger data and '*Sehyeong*' type bronze dagger data, into training, validation, and test sets. Of the 492 bronze dagger drawing data, 70% is used for training, and 15% each is used for validation and testing. Table 2 shows the training, validation, and test set split ratios of the dataset.

Class	Samples for Class				All Samples			
	total	train	val	test	total	train	val	test
'*Bipahyeong*'	231	164	34	33	492	344	73	75
'*Sehyeong*'	162	180	39	42				

Table 2. Dataset distribution for training, validation, and test set.

4.3 Experimental Results

The evaluation results of the six previously published binary classification models and the proposed fusion model in this paper are shown in the table below. Table 3 and Figure 5 show the performance comparison of binary classification models and ROC curve analysis, respectively.

Binary	Class	Per-class Metrics	Overall

Classification Model		Precision	Recall	F1 Score	Accuracy	Accuracy
ResNet-50	'*Bipahyeong*'	0.87	1.00	0.93	100%	94.67%
	'*Sehyeong*'	1.00	0.88	0.94	90.48%	
DenseNet-121	'*Bipahyeong*'	0.89	1.00	0.94	100%	94.67%
	'*Sehyeong*'	1.00	0.90	0.95	90.48%	
EfficientNetV2	'*Bipahyeong*'	0.86	0.97	0.91	96.97%	92.00%
	'*Sehyeong*'	0.97	0.88	0.93	88.10%	
RepVGG	'*Bipahyeong*'	0.84	0.95	0.89	93.94%	89.33%
	'*Sehyeong*'	0.95	0.86	0.90	85.71%	
CoNeXTV2	'*Bipahyeong*'	0.69	0.94	0.79	93.94%	78.67%
	'*Sehyeong*'	0.93	0.67	0.78	66.67%	
proposed fusion model	'*Bipahyeong*'	0.91	0.97	0.94	96.97%	94.67%
	'*Sehyeong*'	0.97	0.93	0.95	95.86%	

Table 3. Performance comparison of different binary classification models.

The ResNet-50 model shows relatively excellent overall performance, but exhibits mutually complementary characteristics in prediction patterns between the two classes. For the '*Bipahyeong*' class, the recall value is 1.00, showing perfect recall and detecting all actual '*Bipahyeong*' type bronze dagger without omission, but the F1 score remains at 0.93 due to relatively low precision values. Conversely, the '*Sehyeong*' class shows a precision value of 1.00, demonstrating perfect precision where all predictions of '*Sehyeong*' type bronze dagger are accurate, but due to low recall values, it shows a tendency to miss some '*Sehyeong*' type bronze dagger by classifying them as '*Bipahyeong*' type bronze dagger, recording an F1 score of 0.94. The F1 score for the '*Sehyeong*' class is slightly higher than that of the '*Bipahyeong*' class, which appears to be because the perfect performance in precision somewhat offset the lack of recall. Despite these mutually complementary performance patterns, an overall accuracy of 94.67% was achieved, confirming that the model is generally effective in distinguishing between the two classes.

The DenseNet-121 model shows a similar performance pattern to ResNet-50 while demonstrating slight improvements. In the '*Bipahyeong*' class, while maintaining a recall value of 1.00, the precision value improves slightly to 0.89, recording an F1 score of 0.94 and showing reduced misclassification compared to the ResNet-50 model. The '*Sehyeong*' class also maintained a precision value of 1.00 while the recall value improves to 0.90, recording an F1 score of 0.95 and showing that fewer '*Sehyeong*' type bronze dagger samples are missed compared to the ResNet-50 model. In the DenseNet-121 model, higher F1 scores than ResNet-50 are recorded for both classes, with the achievement of an F1 score of 0.95 for the '*Sehyeong*' class being one of the highest F1 score records in this paper. However, despite this balanced performance improvement in both classes, the overall accuracy appears as 94.67%, identical to the ResNet-50 model. This shows that while the two models make errors in different ways, they ultimately achieve similar levels of classification performance.

The EfficientNetV2 model shows different performance characteristics from the previous ResNet-50 and DenseNet-121 models. Unlike the previous models that shows recall values of 1.00 in each class, the EfficientNetV2 model shows balanced performance in both classes. The '*Bipahyeong*' class shows moderate levels with precision and recall values of 0.86 and 0.97 respectively, recording an F1 score of 0.91. The '*Sehyeong*' class also records precision and recall values of 0.97 and 0.88 respectively, achieving an F1 score of 0.92 with balanced performance. The fact that both classes show similar levels with F1 scores in the 0.91-0.93 range means that the model pursues stable and balanced classification performance rather than specializing in extreme aspects. However, with an overall accuracy of 92.00%, which is lower than the accuracy of previous models, it shows disappointing results in comprehensive performance compared to previous models.

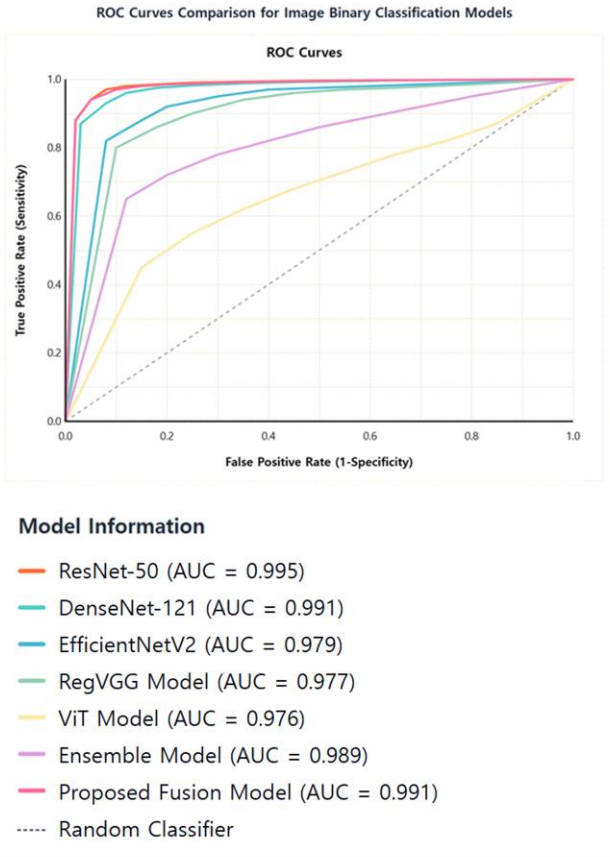


Figure 5. ROC Curves analysis for binary classification models on transitional bronze dagger.

The RepVGG model shows overall lower performance compared to other models in the experiments conducted in this paper. In the '*Bipahyeong*' class, precision and recall values are 0.84 and 0.95 respectively, with precision significantly lower compared to the previously tested ResNet-50, DenseNet-121, and EfficientNetV2 models, resulting in an F1 score of only 0.89. The '*Sehyeong*' class also records precision and recall values of 0.95 and 0.86 respectively, with recall significantly decreases, achieving an F1 score of 0.90. The overall accuracy of the RepVGG model is also 89.33%, showing lower performance compared to previous models. The RepVGG model is introduced to mitigate the problems of the EfficientNetV2 model. Since the EfficientNetV2 model has a structure suitable for learning massive datasets and there are concerns about overfitting for use in bronze dagger typological classification, experiments are conducted with the RepVGG model, which is easier to train even with relatively small datasets. However, this shows that the model does not demonstrate sufficient performance in specific classification tasks.

The ConvNeXtV2 model shows a different performance distribution compared to other models in the experiments conducted in this paper, but records overall low performance. In the ‘*Bipahyeong*’ class, with precision 0.69 and recall 0.93, while recall is high, precision is considerably low, resulting in an F1 score of only 0.79. In the ‘*Sehyeong*’ class, with precision 0.93 and recall 0.67, while precision is high, recall is significantly insufficient, recording an F1 score of 0.78. The overall accuracy is 78.67%, showing the lowest performance. Experiments are conducted with the ConvNeXtV2 model, which applies modern optimization techniques, to overcome the learning instability and overfitting problems of existing binary classification models with small-scale data, but the expected performance improvement is not achieved.

The proposed fusion model is a novel approach proposed in this paper, combining through a ‘Feature Fusion’ method that extracts and combines the characteristic parts of the ResNet-50 model and the characteristic parts of the DenseNet-121 model. It shows improved performance compared to other single models in the experiments conducted in this paper. In the ‘*Bipahyeong*’ class, it shows balanced performance with precision 0.91 and recall 0.97, recording an F1 score of 0.94, and in the ‘*Sehyeong*’ class, it shows stable performance with precision 0.97 and recall 0.93, recording an F1 score of 0.95. The overall accuracy is 94.67%, achieving the same level as the ResNet-50 and DenseNet-121 models, which shows relatively excellent results among existing models, while simultaneously securing high precision and recall above 0.90 in both classes, demonstrating more stable and balanced classification performance. This means that by fusing features from multiple models, the limitations of individual models are compensated and overall classification reliability is improved. Figure 6 shows the confusion matrix of the proposed fusion model, representing classification performance on the evaluation dataset. Among the total 75 evaluation samples, misclassification occurs in 5 samples (6.67%), and these are all cases where ‘*Bipahyeong*’ type bronze dagger is incorrectly classified as ‘*Sehyeong*’ type bronze dagger.

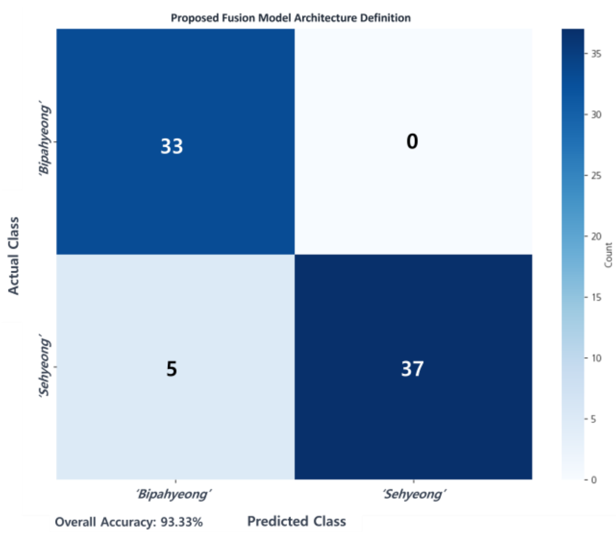


Figure 6. Confusion matrix evaluation results for the proposed fusion model.

Qualitative analysis of the misclassified evaluation data reveals that these samples possess boundary characteristics between ‘*Bipahyeong*’ and ‘*Sehyeong’* classes simultaneously, imposing constraints on the model's classification performance. This inter-class feature overlap is a phenomenon commonly occurring in medical image classification, particularly causing ambiguity in classification boundaries between classes with morphologically similar characteristics. Table 4 shows the original images of the 5 samples identified as misclassified in the confusion matrix, presenting the actual class, predicted class, and prediction confidence for each sample together to provide detailed analysis of misclassification patterns.

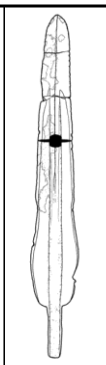
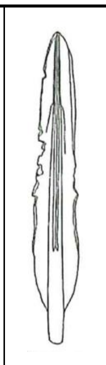
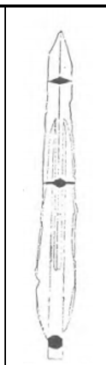
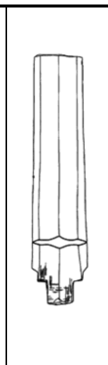
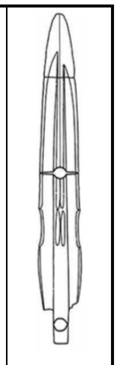
Data Image					
Predicted Class	‘*Bipahyeong*’	‘*Bipahyeong*’	‘*Bipahyeong*’	‘*Bipahyeong*’	‘*Bipahyeong*’
Actual Class	‘*Sehyeong*’	‘*Sehyeong*’	‘*Sehyeong*’	‘*Sehyeong*’	‘*Sehyeong*’

Table 4. Visual examples of misclassified samples from confusion matrix

4.4 Discussion

In this paper, six binary classification models suitable for classifying the morphological features of bronze dagger are selected to choose a binary classification model to be utilized for bronze dagger typological classification. Among the six models, the proposed fusion model is created by extracting the characteristic features of the models most suitable for the dataset of this paper. To determine what type of binary classification model would be suitable for archaeological bronze dagger typological classification models, the performance of the models is compared and evaluated. The results of comprehensive analysis of performance according to the evaluation indicators selected in this paper are as follows.

As the top performance model group, ResNet-50, DenseNet-121, and the proposed fusion model all achieve the same overall accuracy of 94.67%. ResNet-50 shows mutually complementary characteristics, recording perfect recall (1.00) in the ‘*Bipahyeong*’ class and perfect precision (1.00) in the ‘*Sehyeong*’ class. DenseNet-121 shows a similar pattern to ResNet-50 while achieving more balanced performance with improved F1 scores in both classes. The proposed fusion model combines the features of ResNet-50 and DenseNet-121 to show the most stable performance. By simultaneously securing high precision and recall above 0.90 in both classes (‘*Bipahyeong*’: precision 0.91, recall 0.97; ‘*Sehyeong*’: precision 0.97, recall 0.93), it effectively compensates for the limitations of individual models.

The medium performance model group (EfficientNetV2, RepVGG) shows relatively lower performance. EfficientNetV2 shows balanced performance in both classes (F1 score 0.91-0.92) with an accuracy of 92.00%, but overall does not reach the level of the top performance model group. RepVGG shows below-expected performance with an accuracy of 89.33%, despite its suitability for small-scale datasets. The low performance model group, ConvNeXtV2, shows serious performance degradation with an accuracy of 78.67%, failing to achieve the expected performance improvement despite modern optimization techniques.

https://doi.org/10.5194/isprs-archives-XLVIII-M-9-2025-707-2025

In conclusion, the feature fusion model proposed in this paper demonstrates excellent performance in terms of accuracy, balance, and stability, and is judged to be the most suitable model for bronze dagger typological classification. The model proposed in this paper can be utilized as a practical tool that assists archaeological interpretation for bronze dagger that are difficult to classify using existing morphological criteria, beyond binary classification. To assess how the proposed model could be utilized in actual research settings, 42 bronze dagger drawing data for which opinions on morphology are still divided in the archaeological community are separately collected and applied to the model. These bronze daggers are transitional cases with mixed characteristics of ‘*Bipahyeong*’ and ‘*Sehyeong*’, and are artifacts that are difficult to clearly define using existing classification systems. The model presents the probability of each bronze dagger belonging to ‘*Bipahyeong*’ or ‘*Sehyeong*’ in quantified form, confirming that it can provide objective reference indicators for classification. Figure 7 shows the classification probabilities for each test image. The right graph represents ‘*Bipahyeong*’ probabilities, and the left graph represents ‘*Sehyeong*’ probabilities. Each point indicates individual images and their corresponding classification confidence, with all values displayed as colored points with probability values.

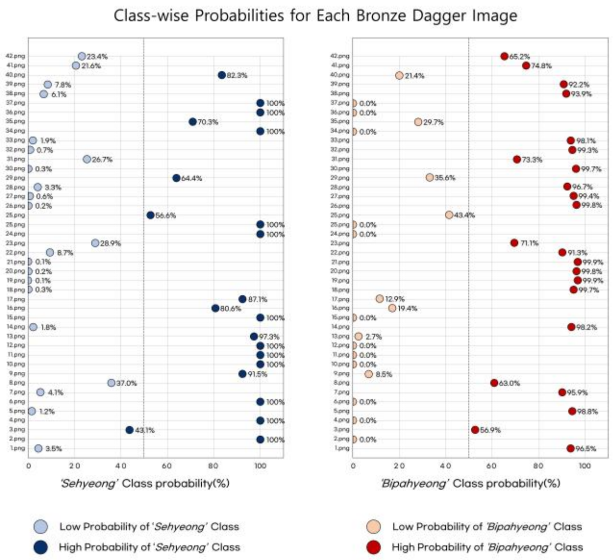


Figure 7. Individual image classification probabilities for transitional bronze dagger.

5. Conclusions and Future Works

This paper evaluated deep learning-based binary classification artificial intelligence models to provide assistance with consistent and objective classification criteria in morphological classification tasks for bronze dagger excavated from the Korean Peninsula to northeastern China, and proposes the most suitable combination for bronze dagger typological classification. Six binary classification models were trained and evaluated using 492 bronze dagger datasets. ResNet-50 and DenseNet-121, which achieved the highest accuracy levels, were combined using a feature fusion approach. Performance was further enhanced by applying data augmentation techniques including TPS transformations and rotation.

The experimental results showed that ResNet-50, DenseNet-121, and the proposed combined model all achieved identical accuracy of 94.67%. However, the proposed combined model demonstrated the most balanced and stable classification performance by securing high precision and recall values above 0.90 for both ‘*Bipahyeong*’ and ‘*Sehyeong*’ classes. The significance of this paper lies in providing quantitative indicators for archaeological researchers' morphological classification work on bronze daggers and presenting a deep learning-based solution to the persistent problem of classification discrepancies among scholars. Particularly, for 42 transitional bronze daggers with ambiguous classifications, separate from the training data, the model can quantify and present classification probabilities for both ‘*Bipahyeong*’ and ‘*Sehyeong*’ types, enabling researchers to obtain data-driven objective reference indicators. However, limitations remain due to the model being trained on a limited dataset of 492 samples.

Future work will focus on constructing large-scale datasets of bronze daggers with diverse morphological characteristics, including transitional types, to enhance the generalization performance of the proposed model.

Acknowledgements

The author gratefully acknowledges the support of Dr. Jeongmin Yu, the corresponding author of this paper.

References

Choi, S. R., 1984. A Study of the Typological Method in Korea. *Journal of the Korean archaeological society Hanguk Koko-Hakbo*, 16, 1-24.

Ding, X., Zhang, X., Ma, N., Han, J., Ding, G., Sun, J., 2021. RepVGG: Making VGG-style ConvNets Great Again. arXiv preprint arXiv:2101.03697. doi.org/10.48550/arXiv.2101.03697.

Gîrbacia, F., 2024. An Analysis of Research Trends for Using Artificial Intelligence in Cultural Heritage. *Electronics*, 13(18), 3738. https://doi.org/10.3390/electronics13183738

He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep Residual Learning for Image Recognition. arXiv preprint arXiv:1512.03385. doi.org/10.48550/arXiv.1512.03385.

Heo, J. Y., 2016. The Establishment Period and Figurative Features of the Korean – Type Bronze Dagger - Focusing on Section Polishing Techniques. *The Korean Archaeological Society*, 99, 4-39.

Heo, J. Y., 2021. The Typology and Study Trend of Korean Bronze Weapons – The papers presented by Youn Moo-Byong in 1966, 1972 and other periods. *Journal of Society for Korean Bronze Culture*, 28, 136–171. doi.org/10.15397/jkbc.2021.28.136.

Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q., 2016. Densely Connected Convolutional Networks. arXiv preprint arXiv:1608.06993. doi.org/10.48550/arXiv.1608.06993.

Hörr, C., Lindinger, E., Brunnett, G., 2014. Machine learning based typology development in archaeology. *Journal on Computing and Cultural Heritage*, 7(1), 1–23. doi.org/10.1145/2533988.

Jeong, H., Chang, Y., 2022. Development of CNN-DBSCAN-based Typological Prediction Model for Estimating the

Chronological Age of Gaya Pottery. *The Korean Society of Culture and Convergence*, 44(3), 251–276. doi.org/10.33645/cnc.2022.03.44.3.251.

Kim, G. G., 2007. A Study of Liaoning Bronze-dagger. *Master's Thesis*, Department of Archaeology and Anthropology, Knu Graduate School, Daegu, Korea.

Kim, Y. H., 2024. Application of Deep Learning for Classification of Ancient Korean Roof-end Tile Images. *Korean Journal of Heritage: History & Science*, 57(3), 24-35.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. *Nature*, 521(7553), 436–444.

Lee, H. J., 1983. Seriation and its Application to Korean Archaeological Data. *Journal of the Korean archaeological society Hanguk Koko-Hakbo*, 14·15, 133-166.

Tan, M., Le, Q.V., 2021. EfficientNetV2: Smaller Models and Faster Training. arXiv preprint arXiv:2104.00298. doi.org/10.48550/arXiv.2104.00298.

Woo, S.H., Dhar, S., Ronghang, H., Chen, X., Liu, Z., Kweon, I.S., Xie, S., 2023. ConvNeXt V2: Co-designing and Scaling ConvNets with Masked Autoencoders. arXiv preprint arXiv:2301.00808. doi.org/10.48550/arXiv.2301.00808.