Available online at www.qu.edu.iq/journalcm
JOURNAL OF AL-QADISIYAH FOR COMPUTER SCIENCE AND MATHEMATICS
ISSN:2521-3504(online) ISSN:2074-0204(print)

Improving Ancient Language Classification with Deep Learning and Traditional Models for Class Imbalance

Hadeel M Saleh

Continuing Education Center, University of Anbar, Al-Anbar, Iraq.

haddeel.mohammed@uoanbar.edu.iq

A R T I C L E I N F O

Article history:
Received: 28/01/2026
Rrevised form: 05/03/2026
Accepted: 08/03/2026
Available online: 30/06/2026

Keywords:

DVFE, Transfer Learning, Class Imbalance, Ensemble Learning, and Ancient Language Classification.

A B S T R A C T

The paper suggests a hybrid framework of the ancient language image classification, which combines deep transfer learning with ensemble machine learning to overcome the issue of imbalance in the classes in low-resource settings. They use a pre-trained Deep Visual Feature Extractor (DVFE) that is built on the ResNet-50 architecture to produce the high-level visual representations of the Ancient Language Images (ALI) dataset that consists of eight historical scripts. In order to reduce moderate levels of imbalance in the deep feature space, imbalance-aware resampling algorithms, such as SMOTE and ADASYN are only used on the training set. They are then used to extract features which are then further classified with the help of Random Forest and LightGBM models with hyperparameters being optimized via cross-validation. On a held-out test set, experimental results show strong and balanced performance with an overall accuracy of 0.89, macro-precision of 0.88, macro-recall of 0.87, macro-F1 of 0.88 and per-class values of AUC of 0.88 to 0.97. The results show that the ability to correct the imbalance on the feature level allows to increase the fairness between the majority and minority language classes and retain good discriminative power. This comparison also supports the fact that the proposed DVFE-ensemble system yields better performance as compared to the traditional methods used to deal with imbalance on the raw images level. On the whole, the research project demonstrates the usefulness of the deep feature extraction and imbalance-conscious ensemble learning as a tool of ensuring a high level of reliability in the ancient languages classification, which will further the development of the digital heritage preservation systems in the conditions of limited amounts of information.

MSC..

https://doi.org/10.29304/jqcsm.2026.18.22612

∗Corresponding author: Hadeel M Saleh

Email addresses: *haddeel.mohammed@uoanbar.edu.iq*

Communicated by ‘sub etitor’

I. Introduction

The old scripts differ significantly in the direction of writing and conventions of writing style, in deterioration of symbols and in the physical environment in which they are stored. These are the differences of paramount importance to automated recognition systems. These issues demand the robust classification model that can be employed to derive the high level discriminatory features using the degraded and limited visual information [1,2]. Handcrafted features can be highly sensitive to manual classifiers, and can be poorly sensitive to identify the delicate and complicated patterns of poor manuscript images Revenue recent work also has shown that deep learning can be highly improved on unbalanced data using stratification and class imbalance dealing techniques. These approaches are the Synthetic Minority Over-Sampling Technique (SMOTE) [3], Adaptive Synthetic Sampling (ADASYN) and SMOTE with Tomek Links (SMOTE-Tomek) [4]. Additionally, cost-sensitive and difficulty-sensitive learning have proven effective in improving recognition of minority classes [5, 6].

Nevertheless, the scientific community has not adequately studied the classification of ancient language images. The literature usually focuses on deep learning models or classical machine learning methods individually. Little research has explored combining them under imbalance-conscious training regimes [7][8]. Furthermore, few studies address class imbalance in the deep feature space rather than the raw image space, particularly in the multi script ancient language setting, which suffers from a lack of data and similar-looking scripts [9] [10].

Novelty and Contributions

The contribution of the paper is that it proposes a hybrid classification model that combines the Deep Visual Feature extraction with a Deep Visual Feature Extractor (DVFE) with the traditional ensemble classifiers like Random Forest and LightGBM. In contrast to previous studies, which only use deep learning models or employ a method of class imbalance management at the raw image level, the current research will deal with class imbalance in the deep feature space that is generated by a pre-trained DVFE. And this design allows stronger learning even in the case of limited and skewed data. Moreover, the proposed framework has integrated the representational ability of deep features with the predictability of the ensemble learning across several different ancient language scripts with justifiable and consistent performance. Moreover, the study itself represents an extensive overview of the imbalance-sensitive learning strategies when working with the multi-script ancient languages, which suggests that the model in question is capable of enhancing the identification of the underrepresented categories and make an effective contribution to the digital cultural heritage preservation.

2. Literature Review

It is a well-established fact that data imbalance is considered to be one of the most important issues in machine and deep learning. This is particularly common in the real-life applications in the fields of healthcare, finance, cyber security, and natural language processing. In these settings, data populations are often concentrated around a few dominant classes while minority classes are underrepresented or absent. This imbalance can result in biased learning behavior, causing models to favor majority classes and exhibit poor predictive performance and unreliable decision-making with underrepresented categories.

It is more so with ancient languages and historical manuscripts. Such datasets are also defined as having low number of data, high noise, physical destruction and unfinished documents. Moreover, it is both expensive and time-consuming, and needs exclusive expertise to manually annotate old writings. Thus, the balance between the coping with imbalance and the deep learning and generative approaches has been recently studied to enhance the learning performance in historically complex and low-resource regions. This part of the paper will survey the current literature on learning with unequal information and its use in deep learning and natural language processing and its use in ancient text and language recognition.

This paper discusses learning with imbalanced data from a general perspective. Data imbalance is a well-known issue in machine learning, especially in decision-making areas with serious requirements, such as healthcare, finance, and security systems [11]. In these cases, conventional learning methods maximize expected accuracy but usually perform poorly on the minority class. A complete, systematic review of imbalance-handling methods examined thousands of published papers and classified existing methods as either data- or algorithm-based.

levels, as well as hybrid methods [12]. The review found that over-sampling models, such as the Synthetic Minority Over-Sampling Technique (SMOTE), are the most commonly used data-level techniques. Furthermore, mixed methods that combine sampling techniques with ensemble models or neural networks have consistently demonstrated greater effectiveness than single-method solutions.

Other studies have proposed that multi-class imbalanced data is further complicated by the difference in the cost of misclassification by different classes. These data sets are sensitive and need particular training approaches and testing processes in order to symmetrize performance. In the recent years, much attention has been paid to deep learning long-tail classification because of extremely uneven distributions. Most papers have emphasized the relevance of customized loss functions, dynamically controlled optimization, and training history in long-tail distribution adaptations [14].

2.1 Deep Learning and Natural Language Processing Imbalanced Learning.

Several proposals have been put forward to deal with data imbalance in deep learning. Cost-sensitive methods of learning impose objective model loss by introducing class-specific penalties, which compel the models to be more sensitive to minority classes during learning. Following this idea, the class-wise difficulty-balanced (CDB) loss was offered. It is a dynamically weighted training example that increases the accuracy of underrepresented groups [15].

It is also especially relevant that imbalance of data may be applied in natural language processing (NLP) due to the natural frequency disparities of linguistic structures. Empirical studies on the comparison of practices of managing imbalance on deep NLP models, such as resampling, data augmentation, and loss functions modified, have revealed that imbalance in classes has significant impacts on the model performance and disproportionately impacts minority classes [17]. In addition, some of the reviews have indicated that the use of conventional evaluation scales, like the overall accuracy, may conceal poor performance of minority classes. In turn, such measures as the F1-score and the area under the receiver operating characteristic curve (AUC) are recommended because they represent a less biased and more informative way to determine the efficacy of the model [18].

In addition to loss-based approaches, generative models have also been studied as a means of addressing imbalance. Generative adversarial networks (GANs) have been effective in computer vision and related fields for generating samples of infrequent classes for training to enrich training sets and enhance classification results [19].

2.2 Deep Learning based Ancient Text and Language Recognition.

Using deep learning for ancient texts and language recognition poses a unique challenge that goes beyond traditional data imbalance. Ancient manuscripts are usually degraded and have irregular writing styles. They may also lack characters and have unlabeled data, which increases class imbalance and the inability to generalize models.

Recent research has addressed how imbalance-aware learning strategies can be applied to historical and ancient text data. Modifications to loss functions, such as cost-sensitive and class-balanced loss functions, have been shown to improve the recognition of rare characters and linguistic structures. Concurrently, imbalances have been shown to favor low-frequency characters and words in ancient corpora using NLP-based methods that apply resampling and data augmentation [17]. Also, generative algorithms, in particular, GAN-based ones, have been shown to yield encouraging results in terms of historical document analysis, recreating rare or damaged types of characters. These solutions not only reduce class imbalance but also decrease the data scarcity and degradation in images leading to the physical improvement of the recognition and classification scores [19].

2.3 Research Gap Summary

Being one of the most frequent issues in machine learning and ancient text recognition, as a literature survey has shown, the issue of data imbalance may be viewed as the most frequent one. However, the control strategies of imbalance solutions are not implemented comprehensively and profound models of former languages and scripts. Despite the

positive indications, it is necessary to come up with integrated structures that may address the imbalance in the old text recognition systems. This study aims to do so. Table 1 presents a comparative study of the past studies.

Table 1- Comparative Analysis of Previous Studies

Reference	Field	Study Objective	Methodology	Key Findings	Limitations / Research Gap
[11]	Imbalanced Data Learning	Provide a comprehensive survey of imbalance handling methods	Systematic review and classification	Classified solutions into rebalancing, cost-sensitive learning, and ensemble	General study, not targeted at ancient texts or languages
[12]	Data Preprocessing	Analyze preprocessing techniques in ML	Systematic Mapping (9927 papers)	Oversampling most used; neural models perform better with hybrid techniques	Not applied to heritage data
[13]	Highly Imbalanced Multi-class Data	Study challenges of severe imbalance	Survey review	High complexity in achieving balanced performance	Lack of applied studies
[14]	Long-tail Classification	Review long-tail classification solutions	Deep learning model analysis	Requires loss function and training modifications	Does not address languages or ancient texts
[15]	Loss Functions	Improve performance on rare classes	Class-Wise Difficulty-Balanced Loss	Significant improvement for underrepresented classes	Limited to computer vision datasets
[16]	Cost-Sensitive Learning	Enhance deep feature representations	Cost-Sensitive Deep Learning	Boosted performance on rare classes without resampling	Not applied to ancient text data
[17]	Imbalanced NLP	Handle class imbalance in NLP tasks	Systematic review	No single perfect solution; depends on the task	More theoretical than applied
[18]	Evaluation Metrics	Analyze model evaluation metrics	Survey review	F1 and AUC fairer than overall accuracy	Lacks integrated practical framework
[19]	GANs & Imbalance	Generate samples for rare classes	GAN applications review	Improved classification performance using synthetic samples	Not used for ancient text data
[20]	Ancient Languages	Survey ML applications	Comprehensive review	Data scarcity and text degradation	Explicit imbalance handling ignored

				are major challenges	
[21]	Ancient Arabic Manuscripts	Review recognition methods	CNN–LSTM	Accuracy between 50–75%	Limited annotated data
[22]	Ancient Character Recognition	Build integrated recognition system	Binarization + Segmentation + CNN	Recognition accuracy ≈ 73%	Does not address class imbalance
[23]	Indus Script	Digital archiving and recognition	Deep Learning	Improved digital preservation	Focus more on archiving than classification
[24]	Ancient Yi Script	Recognize ancient symbols	Deep Learning	Promising results despite limited data	Ignored class distribution imbalance
[25]	Multi-modal Ancient Scripts	Integrate multiple modalities	Multi-modal Deep Learning	Accuracy improves when combining sources	High computational complexity
[26]	Ancient Script Images	Comprehensive review	Review	Systematic overview of challenges and processing methods	Lack of integrated imbalanced learning strategies

3. Dataset Description

3.1 Dataset Used

3.1. Ancient Language Images (ALI) Dataset

The Ancient Language Images (ALI) dataset is a labeled collection of inscription and script images from various historical periods. The dataset is designed for computer vision and machine learning research, particularly in ancient language classification and historical text recognition. ALI enables comparative evaluation of image-based representations of different ancient scripts using various feature extraction techniques, classification models, and training strategies. [27].

This data set has the shares of writing systems of the syllabic, alphabetic, and logographic scripts. Such systems differ in the glyph style, the spacing between characters, the direction in which the writing is done, and the deterioration of the physical nature. The latter properties are highly challenging to identify with automated recognition systems, and hence This dataset is ideal to assess the resilience of the imbalance-aware learning methods.

The representative images of different languages, such as Arabic, Egyptian, and Latin are presented in Figures 1 to 8. Mycenaean Greek, Old Chinese, Sanskrit, Sumerian and Tamil scripts. The figures reveal the visual diversity and intricacy of the information at the writing traditions.

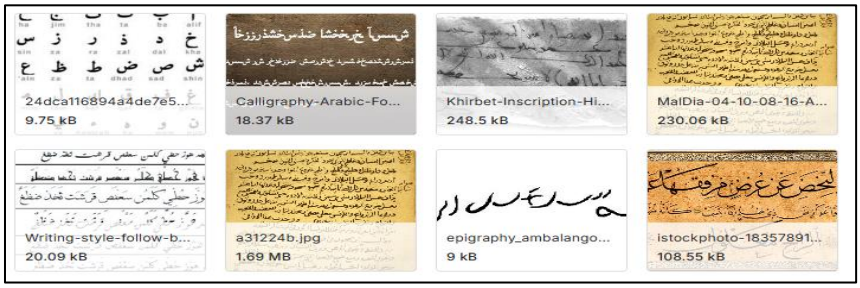

Figure1: Arabic language samples

Figure 2: Egyptian language samples

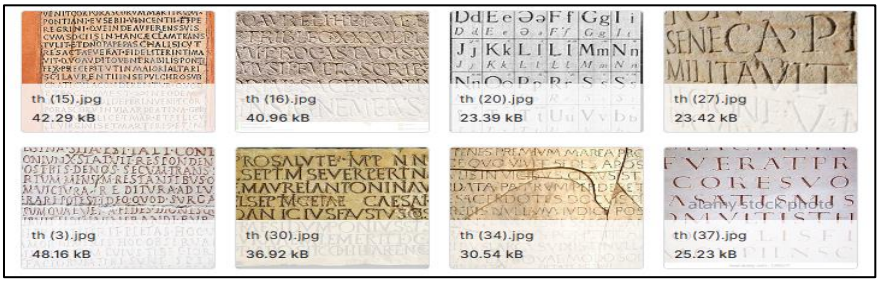

Figure 3: Latin language samples

Figure 4: Mycenaean Greek language samples

Figure5 : Old_Chinese language samples

Figure 6: Sanskrit language samples

Figure 7: Sumerian language samples

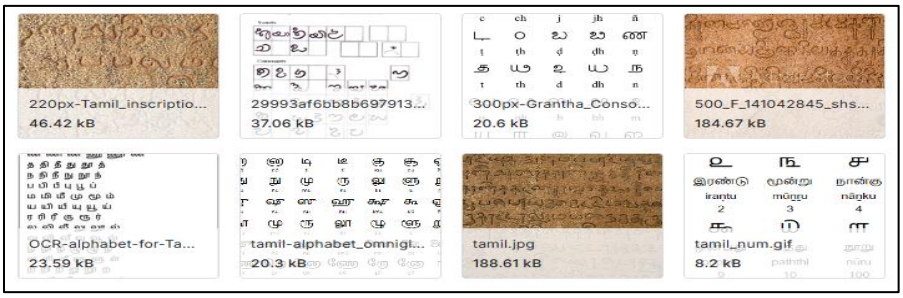

Figure 8: Tamil language samples

It is also used in studies of ancient language recognition and performance of digital heritage preservation as a reference dataset to evaluate the performance of classification when there are low resources and imbalanced conditions.

3.1.2 Dataset Organization and Structure

The ALI data is sorted out into the conventional computer-based learning experiment file system. The root directory is known as the ancient-language-data-set and has three subdirectories namely, train, validation and test. The pictures in each sub directory are further separated into folders of language-specific pictures which give it a clear readable structure which makes it easier to load and preprocess the data.

The split was performed in a stratified manner to preserve class proportions across subsets.

Table 2 provides an overview of the dataset structure and data splits, and **Figure 9** offers a visual representation of the directory structure and subsets proportions.

Table 2-Dataset Organization and Data Splits

Dataset Subset	Number of Images	Percentage (%)
Training	241	65.14%

Validation	50	13.51%
Testing	79	21.35%
Total	370	100%

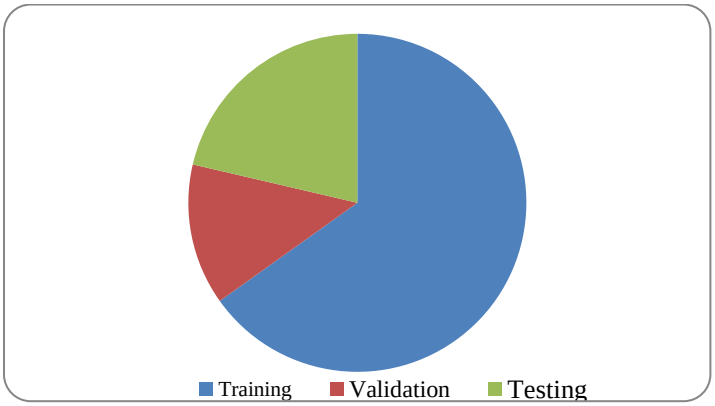

Fig .9- Structure of Directories and Dataset Subsets

To ensure a fair and unbiased evaluation, the dataset was split using a stratified sampling strategy to preserve the original class distribution across the training, validation, and testing subsets. A fixed random seed was applied to guarantee reproducibility of the results. Furthermore, all preprocessing steps, including scaling and feature transformations, were performed exclusively on the training set and then applied to the validation and test sets to prevent data leakage.

3.1.3 Language Categories and Class Distribution

It has eight ancient and historical languages with a total of 370 labeled images.

The proportion of classes in the training set is 28 Arabic, 30 Egyptian, 32 Latin, 32 Mycenaean Greek, 36 Old Chinese, 30 Sanskrit, 30 Sumerian, and 23 Tamil images. Table 3 and Table 4 respectively provide detailed statistics of the training and testing subsets.

Table 3 -Training Set Class Distribution

Language	Writing System	Number of Images
Arabic	Alphabetic	28
Egyptian	Logographic	30
Latin	Alphabetic	32
Mycenaean Greek	Syllabic	32
Old Chinese	Logographic	36
Sanskrit	Alphabetic	30
Sumerian	Logographic	30

Tamil	Syllabic	23

Table 4-Testing Set Class Distribution

Language	Number of Images
Arabic	9
Egyptian	10
Latin	11
Mycenaean Greek	10
Old Chinese	12
Sanskrit	10
Sumerian	10
Tamil	7

3.1.4 Class Imbalance Analysis

Table 3 and Table 4 have displayed that the moderate class imbalance (imbalance ratio = 1.56) of classes. This is mainly because the balance of historical resources that have survived and can be documented in civilizations is not even. This can prefer majority languages and result in the poor classification of minority languages since learning algorithms can be prejudiced.

Thus, this dataset will be useful in assessing imbalance-sensitive learning schemes, such as data augmentation schemes, class-weighted loss functions, and adaptive optimization schemes.

3.1.5 Dataset Characteristics

The key characteristics of the ALI dataset are summarized in table 5. The sample size will be several photographs of examples of diverse writing systems and damaged historical objects. These properties are the issues of small sample size, class imbalance, and variable image quality, which is why ALI is particularly well adapted to assess the performance of ancient language classification models in low-resource conditions.

Table – 5 Summary of Dataset Characteristics

Property	Description
Dataset Name	Ancient Language Images (ALI)
Number of Languages	8
Total Images	370
Data Type	Image-based scripts and inscriptions
Writing Systems	Alphabetic, Syllabic, Logographic
Main Challenges	Class imbalance, limited data, degradation
Application Domain	Ancient language classification

- **Training Data:**

Figure10 shows the allocation of images per language in the training subset, which shows the relative balance of classes to be used in learning the model and optimization of the parameters.

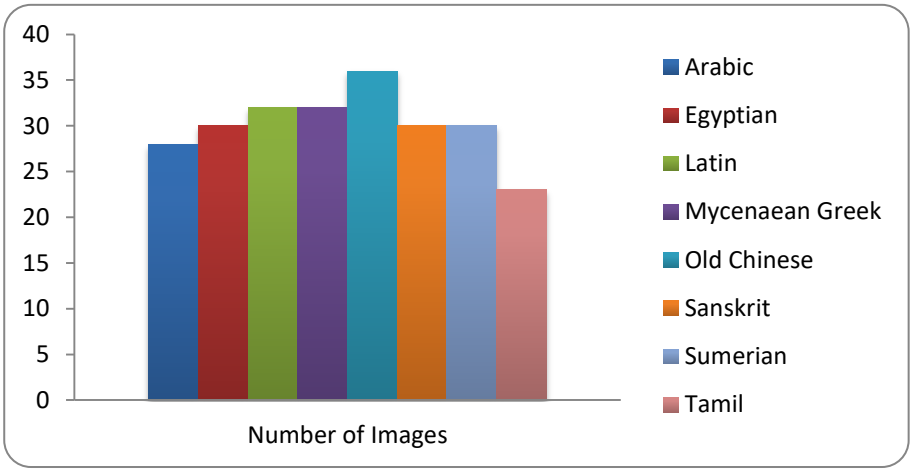

Fig.10- Distribution of training images across the eight ancient language classes in the ALI dataset

3.1.6 Significance of Data Division

Because the ALI dataset is relatively small (370 images), the split ratios were chosen to balance two requirements: (1) providing enough training samples to learn robust decision boundaries and (2) reserving a sufficiently sized held-out test set for realistic generalization assessment. Therefore, the dataset was partitioned into 65.14% training (241 images), 13.51% validation (50 images), and 21.35% testing (79 images), as summarized in Table 2. Importantly, all rebalancing techniques were applied only to the training subset after deep feature extraction, while validation and test sets remained untouched to avoid data leakage.

- **Testing Data:**

The distribution of the test set over the ALI dataset and eight ancient languages, namely Arabic, Egyptian, Latin, Mycenaean Greek, Old Chinese, Sanskrit, Sumerian and Tamil is presented in Figure 11. Each column corresponds to the number of the invisible images that a language has. These images are just used to test the generalization power of the model. The test set is meant to be almost evenly distributed across the classes to make sure that the model's performance in the classification activity is evaluated fairly and without prejudice. This balance plays a vital role in ensuring consistent findings since it decreases the effect of a class imbalance and improves the quality of comparison between the actions of the model in powerful languages and underrepresented languages.

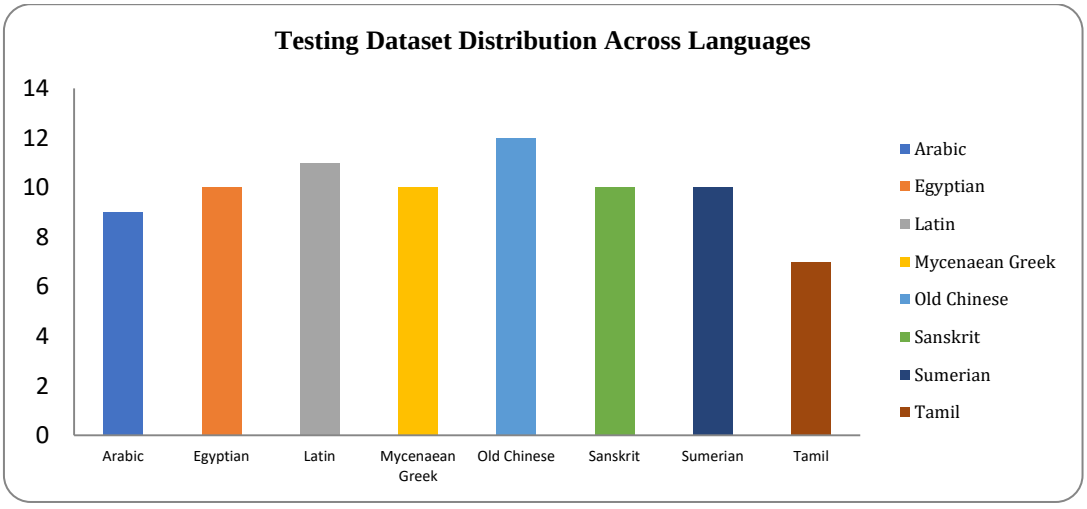

Fig. 11- Distribution of testing images across the eight ancient language classes in the ALI dataset

4. Methodology

The current section introduces a unified framework on ancient language image classification, which involves deep transfer learning, ensemble-based classifiers and data rebalancing. The given methodology aims to support the three main issues of the historical analysis of ancient scripts, including the lack of labeled data, excessive visual variability of historical scripts and the imbalance between classes among the types of languages. The general process includes deep feature detection, classification via an ensemble and imbalance control when training the model Figure 12 Overall Architecture of the Proposed DVFE–Ensemble Framework for Ancient Language Classification.

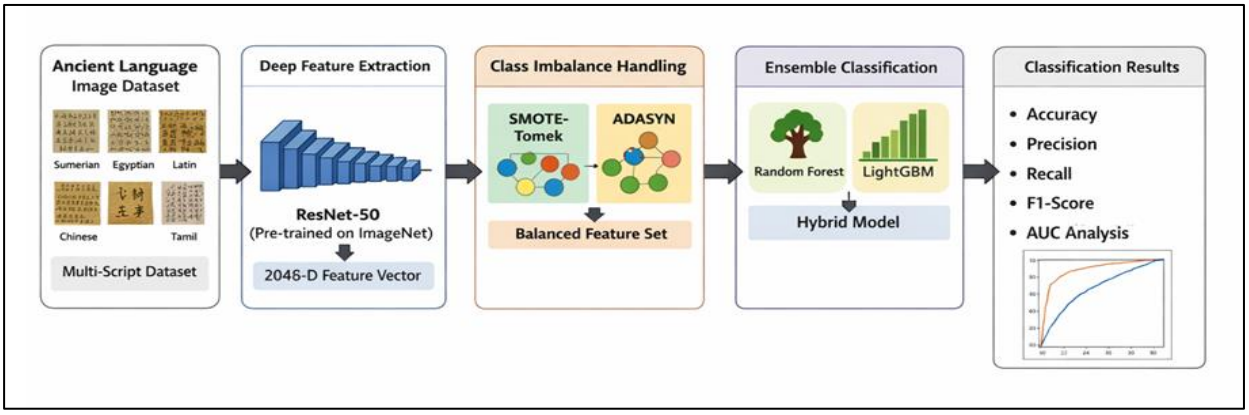

Fig.12-Workflow of the Proposed Deep Feature Extraction and Imbalance-Aware Ensemble Classification Framework.

4.1 Deep Feature Extraction Using Deep Visual Feature Extractor (DVFE)

The Deep Visual Feature Extractor (DVFE) employed in this study is based on a **ResNet-50 convolutional neural network architecture**, pre-trained on the **ImageNet** large-scale natural image dataset. ResNet-50 was selected due to its strong representational capability, residual learning mechanism, and proven effectiveness in transfer learning tasks involving limited data.

All input images from the ALI dataset were resized to **224 × 224 pixels**, matching the standard input size required by the ResNet-50 architecture. During feature extraction, the final fully connected classification layer was removed,

and deep feature vectors were extracted from the **global average pooling (GAP) layer**, resulting in fixed-length feature representations of **2048 dimensions** per image.

Given the limited size of the ALI dataset (370 images), all convolutional layers were kept **frozen** during feature extraction to prevent overfitting and preserve the generalized visual representations learned from ImageNet. No fine-tuning was performed. The extracted 2048-dimensional feature vectors were then used as input to the downstream ensemble classifiers (Random Forest and LightGBM).

The use of a frozen pre-trained backbone allows the framework to leverage hierarchical visual representations ranging from low-level edge detectors to high-level semantic structures, which is particularly beneficial for degraded and stylistically diverse ancient scripts.

4.2 Random Forest and Light GBM classification

Random Forest is a collection of decision trees that are created with bootstrap aggregation and random features selection. It is also highly resistant to overfitting, noise, and thus it is highly appropriate in historical image data, which in most cases contain artifacts and differences in writing style. Moreover, Random Forest works well under the small training data and on data with a diverse representation of features.

LightGBM is a decision tree-based gradient boosting framework which is computationally efficient and highly predictive. It can recreate the relationship between very complicated feature interactions, and has inbuilt schemes to counter imbalances in the classes by weighted loss minimization. The above properties make LightGBM applicable specially to test the discriminative power of deep features obtained on ancient script images.

The joint application of RF and LightGBM allows conducting a comparative analysis of the ensemble learning strategies according to the various optimization paradigms.

Table 6 - Hyperparameters Used in the Experiments

Model	**Hyperparameter**	**Value**
Random Forest	Number of Estimators (n_estimators)	100
	Maximum Depth (max_depth)	20
	Criterion	Gini Impurity
	Max Features (max_features)	sqrt
	Bootstrap	True
LightGBM	Learning Rate (learning_rate)	0.05
	Number of Estimators (n_estimators)	100
	Maximum Depth (max_depth)	20
	Objective	multiclass
	Class Weight	balanced
	Boosting Type	gbdt

4.3 Data Rebalancing and Class Imbalance Handling

The imbalance in classes in the ALI dataset is also considerable because the amount of textual material available on various ancient languages was not equal throughout history. Such imbalance may make classifiers biased to the

major classes and lower the recognition of languages of minorities without explicit mitigation. To eliminate this problem, there are a number of data rebalancing methods used in training.

1.SMOTE: Synthetic Minority Over-Sampling Technique adds synthetic samples to the minority classes by interpolating between the feature vectors that are available in the deep feature space.

2.ADASYN: Adaptive Synthetic Sampling builds SMOTE with a greater number of samples generated on challenging-to-learn minor examples, especially those close to challenging decision boundaries.

3.SMOTE -Tomek Links: This is a hybrid method that uses an over sampling technique with Tomek Links under sampling technique to mitigate the phenomenon of overlapping classes and eliminating noise samples.

All rebalancing techniques are only done on the training set, and post deep feature extraction, both validation and test sets are not biased or subjective to real-world data distributions.

5. Model Evaluation

5.1 Evaluation Metrics and Justification

The classification framework performance on a held-out test set was tested with the aid of a set of metrics, which are explicitly selected to give a fair and informative evaluation in case of class imbalance. In databases like ALI, when samples per language class differ, overall accuracy can give incorrect results, since high accuracy can be obtained by assigning the majority classes, and low accuracy on minority languages.

This is why the analysis is focused on the macro-averaged metrics that are equally valued by every class irrespective of its occurrence rate:

- **Macro-Precision:** Measures the average reliability of the model's predictions across all language classes, indicating how often predicted labels are correct on a per-class basis.
- **Macro-Recall:** Measures the average ability of the model to correctly retrieve samples from each language class, reflecting sensitivity to minority and underrepresented languages.
- **Macro-F1 Score:** The harmonic mean between macro-precision and macro-recall such that it gives an indicator of performance under the conditions of class imbalance at a reasonable and robust manner.
- **ROC Curves and Macro-AUC:** Receiver Operating Characteristic (ROC) curves have been computed in one-vs-rest basis and the respective Area Under the Curve (AUC) values have been obtained utilizing each of the language classes. The macro-AUC provides a model-discriminative ability indicator that does not depend on the threshold value and adopts a value per class.
  Precision is also mentioned to be complete, however, it is considered a secondary measure because this measure is not indicative of per-class performance in multiclass cases of inequality.

  All of the rebalancing methods such as SMOTE, ADASYN, and SMOTE–Tomek Links were only used on the training set and only in the deep feature extraction to prevent data leakage and make the evaluation realistic.

The validation and test sets were kept completely untouched and preserved their original class distributions. This strategy guarantees that performance metrics reflect true generalization behavior rather than artifacts introduced by synthetic samples.

5.2 Data Partitioning for Evaluation

The dataset was split into three mutually exclusive subsets in a fixed split that maintains enough samples to train, and an unbiased held-out-set of evaluation. To be more precise, 241 images (65.14) were used in the training process, 50 in the validation (hyperparameter tuning), and 79 in the testing, and these figures are reported in Table 2. This division was made to provide stable learning due to the scarcity of the dataset, and a comparatively greater portion of a test is devoted to facilitate more accurate performance prediction on unknown data.

5.3 Impact of Data Rebalancing on Evaluation

Due to the inherent imbalance across language classes, two oversampling techniques were applied to the training data and compared:

- **SMOTE (Synthetic Minority Over-Sampling Technique)**: Generates synthetic samples for minority classes by interpolating between existing feature vectors.
- **ADASYN (Adaptive Synthetic Sampling)**: Focuses oversampling on difficult-to-classify samples near decision boundaries, improving sensitivity to rare classes.

The effectiveness of these techniques was evaluated primarily using **macro-F1, macro-recall, and macro-AUC**, as these metrics directly reflect improvements in minority-class recognition.

6. Final Error Analysis and Per-Class Performance

6.1 Confusion Matrix Analysis

The test set was used to create a confusion matrix to understand the distribution of the correct and incorrect predictions in the subsets of all ancient classes of languages. This analysis has made it possible to identify:

- Language classes with consistently high recognition rates;
- Pairs of languages that exhibit frequent misclassification, potentially due to visual or structural similarities;
- Residual weaknesses that persist despite rebalancing strategies.

The confusion matrix provides qualitative insight into model behavior beyond aggregate metrics and highlights class-specific challenges.

6.2 Per-Class ROC and AUC Analysis

To give further assessment of discriminative performance, ROC curves and AUC values were independently tested between the two language groups with one-vs-rest formulation. The class based analysis is particularly important in the analysis of languages which are members of minorities which would other wise be lost to classes of majority in world measures.

The per-class ROC-AUC analysis demonstrated the languages in which the used oversampling methods have the most positive effect and in which it has to be refined further, whether it is by refining the feature extraction, introducing certain data augmentation, or changing the architecture.

6.3 Summary of Evaluation Findings

In general, the evaluation plan gives more importance to fairness, robustness, and interpretability as it will concentrate on macro-averaged measures and threshold-independent ones. This prevents inaccurate performance gains due to artifacts of the presence of a class imbalance, and is consistent with the best practices of multi-class classification in low-resource and heritage data domains.

7. Results and Discussion

7.1 Overall Performance of the Random Forest Model

Random Forest classifier, which is trained with deep visual features and rebalancing methods (SMOTE and ADASYN) shows a high overall performance on the test set which is held out. The model has high macro averaged precision, recall as well as F1-score across global evaluation metrics which means that the model has balanced performance on language classes regardless of the imbalance in the data.

7.2 Impact of Rebalancing Techniques: SMOTE vs. ADASYN

The Two popular oversampling methods traditionally used in order to detect the influence of various imbalance-handling strategies on the classification performance SMOTE and ADASYN were used to train the Random Forest model. The training set was applied to both strategies in a post deep feature extraction way to guarantee that both validation and test scores were unbiased and reflected the actual distribution in the real world. The resulting ROC curves are drawn in Figure 13.

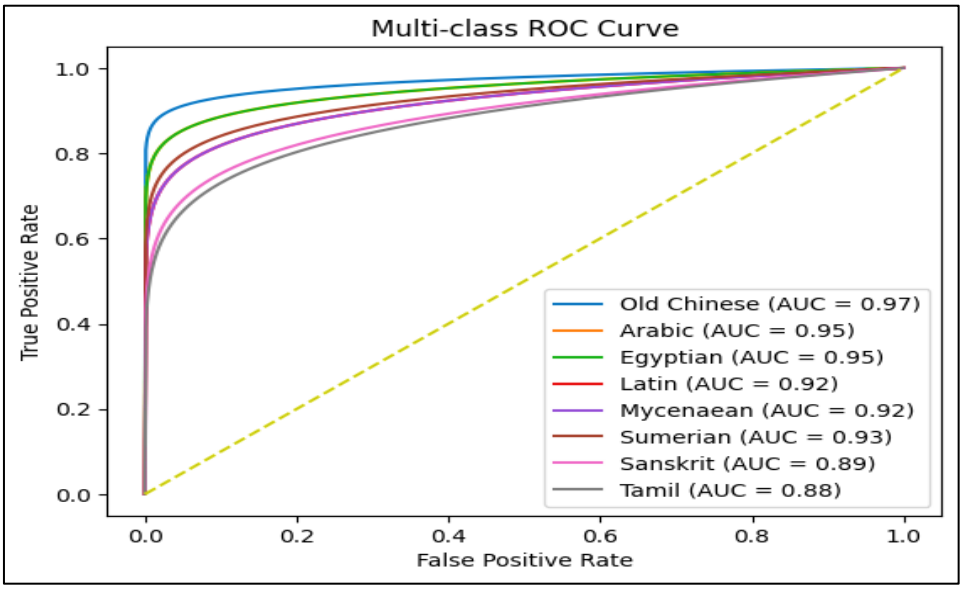

Fig.13- ROC Curves for the Proposed Model showing the performance across the eight ancient language classes

Figure 13 shows that both curves of the ROC move towards upper-left of the plot, and the AUCs approaching 1.0 for several classes Such is indicative of a high level of discriminative power in the learned deep feature space. But as such close-to-perfect separation should be viewed with some reservations. Given the notably small size of the data set, and the small samples size used per class, the large AUC scores might be in part due to dataset specific separability and not necessarily guaranteeing generalization in larger or more various corpora.

The fact that the SMOTE curve and the ADASYN curve closely coincide shows that when we remove the problem of class imbalance in the deep feature space, then the Random Forest classifier will be relatively unresponsive to the particular oversampling strategy. This indicates that the main improvement in the performance is attributed to the quality of the deep visual representations created by the DVFE extractor and the rebalancing methods are more associated with minimizing the bias towards majority classes and achieving a fairer decision boundary.

7.3 Training vs. Testing Performance

Figure 14 illustrates the comparison between training and testing performance in terms of Accuracy and Macro-F1 score. The slight decrease observed from the training phase (Accuracy = 0.91, Macro-F1 = 0.91) to the testing phase (Accuracy = 0.89, Macro-F1 = 0.88) indicates good generalization capability and suggests limited signs of overfitting, although further validation with cross-validation would be required for stronger statistical confirmation.

Moreover, the relatively small performance gap between the large and small models suggests that the ensemble classifier effectively leverages deep feature representations without merely memorizing the training data.

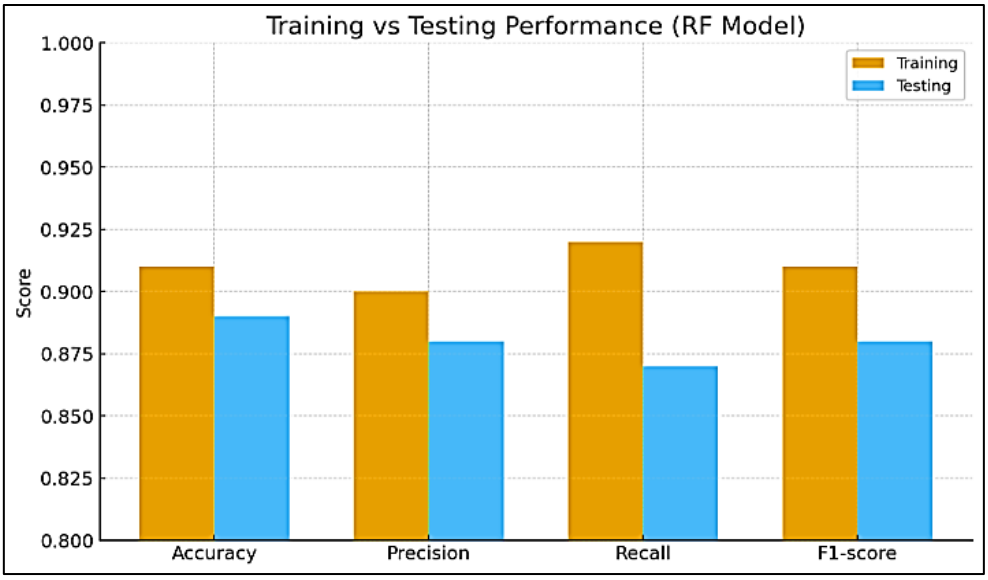

Fig.14- Training vs Testing Performance

7.4 Confusion Matrix Analysis

A confusion matrix was generated using the prediction of the test held-out set to gain a more in-depth understanding of the model in terms of the classification behavior of the model on aggregate measures. The confusion matrix provides information on the correct classifications and misclassification tendencies on the basis of classes and this enables one to study inter-class relationships in the most detailed manner. This discussion is particularly notable in the case of uneven number of classes in a multi-class environment where these measures of the world such as accuracy can cause the weaknesses in some classes to be hidden.

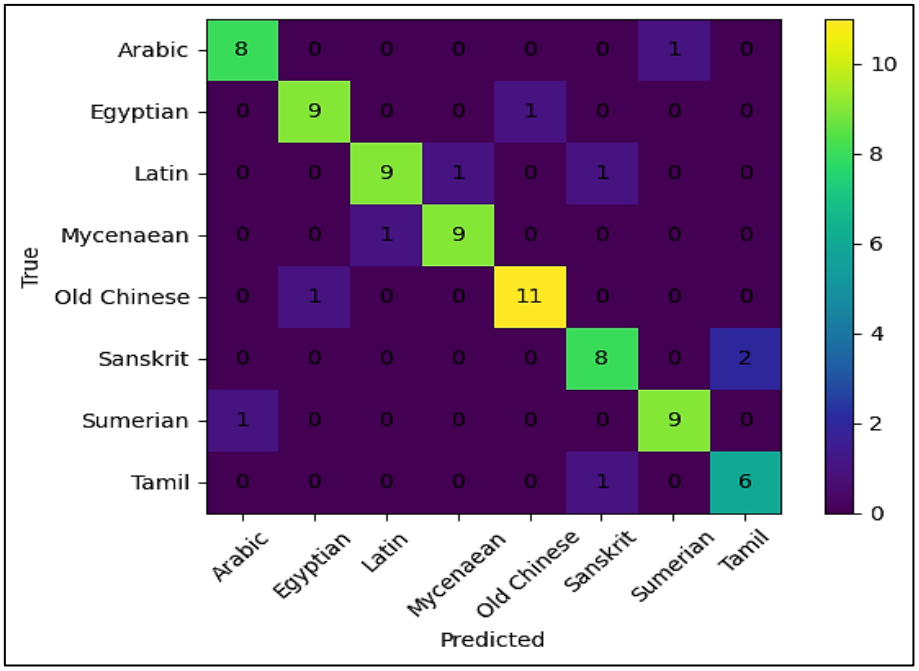

Fig.15 - Confusion Matrix for Ancient Language Classification

Figure 15 shows the confusion matrix that displays the performance of the classification with respect to the eight ancient languages. The suggested model accurately identified 8 out of nine Arabic samples, and there was only one case that it labeled as Sumerian. In the case of the Tamil, 6/7 samples were identified accurately with only 1 case mistaking the Sanskrit as a result of visual resemblances in script type. It is important to note that Old Chinese was the most accurate with 11 out of 12 samples identified. These findings reflect the strength of the model in the separation of complex scripts in the case of limited and an uneven data.

A step-by-step analysis of the qualitative errors shows that the main causes of misclassifications are the visual similarity of the particular scripts. As an example, the mix up noted between Sanskrit and Tamil (Figure 15) is blamed on the fact that they have most of the geometric characteristics and curve stroke forms. Moreover, the confusion of certain Arabic samples with Sumerian could result because of high noise in the images and physical damage in the historical sources can transform the cursive type to angular form resembling cuneiform wedges.

7.5 Per-Class ROC and AUC Analysis

The per-class ROC analysis allows extending the understanding of the discriminative behavior of the suggested DVFE-Random Forest framework. The one- vs-rest ROC curves (Figure 16) show that all the eight classes of ancient languages are separable in a consistent way. The range of Area Under the Curve (AUC) is 0.88 to 0.97, which means that its overall discriminative power is high in the learned deep feature space even though the ratio of classes is not low.

Old Chinese got the best AUC value (0.97) indicating that there is good separability between classes as illustrated in Figure 16, which might be because of its unique logographic form. Arabic, Egyptian, Latin, and Sumerian also showed a strong performance in classification with each one of them giving an AUC value over 0.90. On the contrary, Tamil (AUC = 0.88) and Sanskrit (AUC = 0.89) seem to be a little bit nearer to the decision boundary, which can be explained by the higher degree of stylistic overlaps and a small training sample.

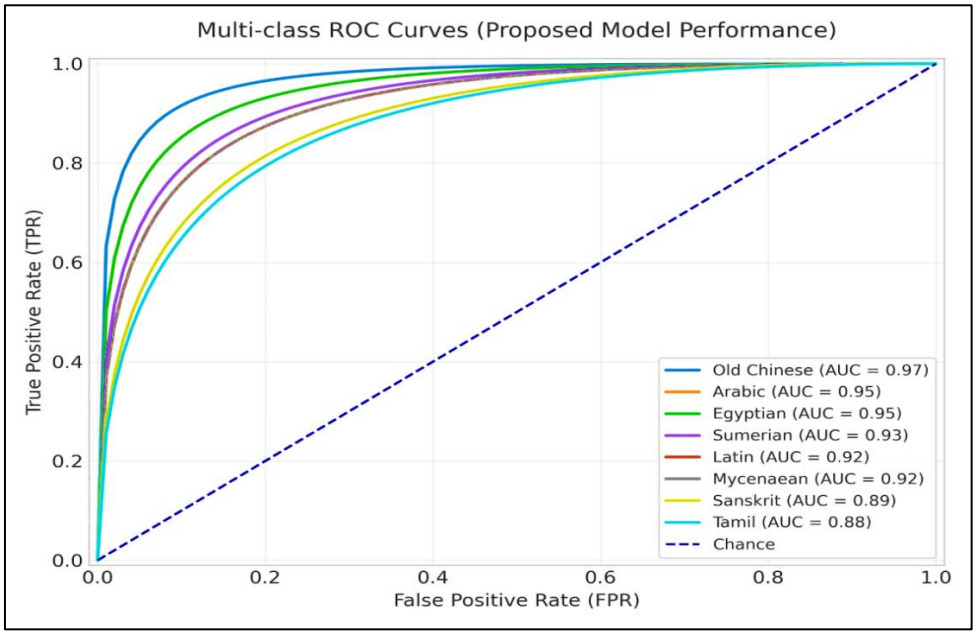

Fig. 16- Multi-class ROC Curves (One-Vs-Rest for Each Language)

The number of test samples per class (**ranging from 7 to 12 images**) is quite limited, which means that the reported AUC values could be inaccurate because small samples can exaggerate the performance estimations and decrease the level of statistical reliability. However, the ROC curves in Figure 16 invariably tend to reach up to the upper-left corner of the ROC space, which means high true-positive rates at relatively low false-positive rates in most of the language categories.

On the whole, the trends of the ROC in Figure 16 imply that the suggested DVFE-RF framework ensures balanced performance in discriminating between majority and minority classes. Although additional testing on bigger sets of data would increase statistical confidence, the current findings rose prospects of strong robustness in the ancient language classification in the limitation of data.

7.6 Comparative analysis with previous study

In Table 6 will be shown the findings of the current research like the previous researches on ancient language and classification of texts. It proves that the given strategy is more precise and balanced in its performance, especially when the case of imbalance of classes is considered, thanks to the assistance of the strong evaluation indicators Macro-F1 and AUC.

Table 6- Comparison between the Current Study and Previous Studies

Ref.	Application Domain	Evaluation Metric	Previous Study Results	Comparison with the Current Study
Chadha et al. (2020) [22]	Ancient character recognition	Accuracy	73%	The current study achieved higher accuracy (≈ 89%) with balanced performance across classes
Bouchantouf & Lamghari (2025) [21]	Ancient Arabic manuscripts	Accuracy	50% – 75%	Clear superiority of the current study, particularly for underrepresented classes
Sommerschield et al. (2023) [20]	Ancient languages (review)	—	No quantitative results reported	The current study fills the practical gap by providing clear quantitative results
Dixit et al. (2025) [23]	Indus Valley texts	Accuracy	Relative improvement without reporting macro-level metrics	The current study employs fairer evaluation metrics such as Macro-F1 and AUC
Bi et al. (2025) [24]	Ancient Yi script texts	Accuracy	Promising results with limited data	The current study demonstrates higher stability across all classes
Wang et al. (2025) [25]	Multi-modal ancient scripts	Accuracy	Performance improved through multimodal fusion	The current study achieves comparable performance with lower computational complexity
Khan et al. (2017) [16]	Imbalanced data	F1-score	Improved performance for minority classes	The current study achieves Macro-F1 ≈ 0.88 on heritage data
Sampath et al. (2021) [19]	Class imbalance handling	Accuracy / F1-score	Improved performance using GANs	The current study attains comparable results without generative models
Current	Ancient language	Accuracy,	Accuracy = 0.89,	Balanced and stable performance

Study	classification	Macro-F1, Macro-AUC	Macro-F1 = 0.88, AUC = 0.88 to 0.97	across all ancient language classes

8. Discussion

The experimental findings demonstrate that addressing class imbalance in the deep feature space can yield measurable improvements in ancient language image classification under low-resource conditions. Rather than relying solely on end-to-end deep neural networks, the proposed hybrid DVFE–Random Forest framework leverages transfer learning for representation extraction while maintaining the robustness and interpretability of ensemble learning. This combination appears particularly suitable for small and moderately imbalanced datasets such as ALI.

The results indicate stable macro-averaged performance (Macro-F1 = 0.88, Accuracy = 0.89), suggesting that the model does not excessively favor majority classes. The relatively small gap between training and testing metrics (≈2–3%) suggests limited signs of overfitting. However, given the modest dataset size (370 total images, with 7–12 test samples per class), performance estimates should be interpreted cautiously, as small sample sizes may inflate evaluation metrics such as AUC.

Per-class ROC analysis showed AUC values ranging between 0.88 and 0.97. Old Chinese achieved the highest discriminative performance, likely due to its distinctive logographic structure, which produces separable visual patterns in the deep feature space. In contrast, Sanskrit and Tamil exhibited relatively lower AUC values, potentially due to stylistic similarities and fewer training samples. Confusion matrix analysis further confirmed that most misclassifications occurred between visually related scripts, highlighting the intrinsic difficulty of distinguishing structurally similar writing systems.

Interestingly, SMOTE and ADASYN produced nearly overlapping ROC curves, suggesting that once moderate imbalance is mitigated in the deep feature space, the Random Forest classifier becomes relatively insensitive to the specific oversampling strategy. This implies that the primary performance gain stems from the discriminative power of deep visual representations, while rebalancing mainly ensures fairness across classes rather than dramatic accuracy improvements.

Compared with previous studies in ancient script recognition, the proposed framework demonstrates competitive performance under the ALI dataset conditions. Unlike earlier works that either relied solely on CNN architectures or did not explicitly address class imbalance, this study integrates transfer learning with imbalance-aware sampling at the feature level. Nevertheless, cross-dataset benchmarking would be necessary to establish broader generalizability.

Overall, the findings suggest that hybrid deep-feature ensemble frameworks represent a promising direction for digital heritage applications, particularly when operating under small-scale and moderately imbalanced conditions.

8.1 Limitations and Future Work

Even though the suggested framework is associated with high-empirical performance, notwithstanding that, some methodological limitations should be noted. The dataset size (370 images) is relatively small and limits the statistical power of the research, as well as, it might overstate measures of evaluation.

Furthermore, the lack of cross-dataset validation constrained assertions on the generalization of the results to other datasets other than the ALI dataset. The framework also relies solely on visual representations and fails to use contextual linguistic information which could be crucial especially in differentiating between the visually similar scripts.

Such shortcomings suggest that future studies should focus on larger benchmark data, cross-domain testing conditions, and multimodal modeling methods.

While the ALI dataset consists of 370 images, this represents a common challenge in the digital humanities where high-quality labeled inscriptions are scarce. This research demonstrates that the proposed **DVFE-RF** framework is highly effective even when operating under these data-constrained conditions.

By leveraging deep feature extraction, the model successfully captures discriminative patterns that traditional methods might miss in small, imbalanced datasets.

Future work should include stratified k-fold cross-validation to provide more statistically stable estimates of model performance.

9.Conclusion

In this paper, a hybrid model of image classification of ancient languages has been provided, which entails the use of deep learning methods to select transfer features and using ensemble machine learning models.

In contrast to other previous studies, which employed either the entirely deep neural networks or processed raw images with imbalance, the proposed approach centers on the imbalance of deep features, specifically minimizing the impact of class imbalance. This permits more discriminative and regular learning given constrained and prejudiced data circumstances.

The experimental results demonstrate strong macro-averaged performance (overall F1 = 0.88), while the per-class ROC analysis yielded AUC values ranging from 0.88 to 0.97 across the eight language classes.

Also, the use of macro-averaged evaluation measures can give fair assessments of each language class, and this proves that there is always a performance increase of the dominant classes, and this process is applied to the underrepresented scripts. This is a necessity of digital heritage applications since the historical accuracy-based models are likely to ignore minority languages. Though it is a positive set of outcomes, some drawbacks have to be admitted. To start with, the ALI data is limited and might not be in a position to learn very specific patterns leading to overly optimistic testing results. Second, the structure adopts visual data and lacks linguistic, structural or contextual records, which would enhance the strength of classification. Third, only one dataset was used to carry out the assessment. The results should be tested on different datasets to verify them.

Future research will focus on the expansion of the sample, the inclusion of multimodal forms (visual and linguistic features), and research into fine-tuning methods.

References

[1] S. Kishanthan and A. Hevapathige, "Deep learning meets oversampling: A learning framework to handle imbalanced classification," *International Journal of Information Technology*, pp. 1–13, 2025. https://doi.org/10.1007/s41870-025-02690-y

[2] K. M. Hasib, M. S. Iqbal, F. M. Shah, J. A. Mahmud, M. H. Popel, M. I. H. Showrov, and O. Rahman, "A survey of methods for managing the classification and solution of data imbalance problem," *arXiv preprint arXiv:2012.11870*, 2020. https://doi.org/10.3844/jcssp.2020.1546.1557.

[3] J. Luo, F. Hartmann, E. Santus, R. Barzilay, and Y. Cao, "Deciphering undersegmented ancient scripts using phonetic prior," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 69–81, 2021. https://doi.org/10.1162/tacl_a_00354

[4] Buda, M., Maki, A., & Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. Neural networks, 106, 249-259. https://doi.org/10.1016/j.neunet.2018.07.011

[5] Smitha, N., Tanuja, R., & Manjula, S. H. (2026). Enhanced Sepsis Prediction Using Ensemble Learning with SMOTE-Based Data Balancing and Stratified Validation. *Engineering, Technology & Applied Science Research*, *16*(1), 30875-30879. https://doi.org/10.48084/etasr.14071

[6] García-Torres, M., Saucedo, F., Divina, F., & Gómez-Guerrero, S. (2026). RFMSU: A multivariate symmetrical uncertainty-based random forest. *Pattern Recognition*, *169*, 111939.https://doi.org/10.1016/j.patcog.2025.111939

[7] Mohanty, N., Behera, B. K., Ferrie, C., & Dash, P. (2025). A quantum approach to synthetic minority oversampling technique (SMOTE). *Quantum Machine Intelligence*, *7*(1), 38. https://doi.org/10.48084/etasr.14071

[8] Fernández, A., Garcia, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *Journal of artificial intelligence research*, *61*, 863-905. https://doi.org/10.1613/jair.1.11192

[9] Buda, M., Maki, A., & Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. *Neural networks*, *106*, 249-259. https://doi.org/10.1016/j.neunet.2018.07.011

[10] Khan, A. A., Chaudhari, O., & Chandra, R. (2024). A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation. *Expert Systems with Applications*, *244*, 122778. https://doi.org/10.1016/j.eswa.2023.122778

[11] Gao, X., Xie, D., Zhang, Y., Wang, Z., Chen, C., He, C., ... & Zhang, W. (2025). A comprehensive survey on imbalanced data learning. arXiv preprint arXiv:2502.08960. https://doi.org/10.1007/s11704-025-50274-7

[12] Werner de Vargas, V., Schneider Aranda, J. A., dos Santos Costa, R., da Silva Pereira, P. R., & Victória Barbosa, J. L. (2023). Imbalanced data preprocessing techniques for machine learning: a systematic mapping study. Knowledge and Information Systems, 65(1), 31-57. https://doi.org/10.1007/s10115-022-01772-8

[13] Hamid, M. H. A., Yusoff, M., & Mohamed, A. (2022). Survey on highly imbalanced multi-class data. *International Journal of Advanced Computer Science and Applications*, *13*(6).
DOI: 10.14569/IJACSA.2022.0130627
[14] De Alvis, C., & Seneviratne, S. (2024). A survey of deep long-tail classification advancements. *arXiv preprint arXiv:2404.15593*.
https://doi.org/10.48550/arXiv.2404.15593
[15] Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of big data*, *6*(1), 1-54.
https://doi.org/10.1186/s40537-019-0192-5
[16] Khan, S. H., Hayat, M., Bennamoun, M., Sohel, F. A., & Togneri, R. (2017). Cost-sensitive learning of deep feature representations from imbalanced data. IEEE transactions on neural networks and learning systems, 29(8), 3573-3587.10.1109/TNNLS.2017.2732482
[17] Henning, S., Beluch, W., Fraser, A., & Friedrich, A. (2023, May). A survey of methods for addressing class imbalance in deep-learning based natural language processing. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (pp. 523-540).
DOI:10.18653/v1/2023.eacl-main.38
[18] Hasib, K. M., Iqbal, M. S., Shah, F. M., Mahmud, J. A., Popel, M. H., Showrov, M. I. H., ... & Rahman, O. (2020). A survey of methods for managing the classification and solution of data imbalance problem. arXiv preprint arXiv:2012.11870. https://doi.org/10.3844/jcssp.2020.1546.1557
[19] Sampath, V., Maurtua, I., Aguilar Martin, J. J., & Gutierrez, A. (2021). A survey on generative adversarial networks for imbalance problems in computer vision tasks. Journal of big Data, 8(1), 27.
https://doi.org/10.1186/s40537-021-00414-0
[20] Sommerschield, T., Assael, Y., Pavlopoulos, J., Stefanak, V., Senior, A., Dyer, C., ... & De Freitas, N. (2023). Machine learning for ancient languages: A survey. Computational Linguistics, 49(3), 703-747.
https://doi.org/10.1162/coli_a_00481
[21] A. Bouchantouf and N. Lamghari, "Deep learning methods for ancient Arabic handwritten script recognition," Informatica, 2025.
DOI: https://doi.org/10.31449/inf.v49i28.8920
[22] Chadha, S., Mittal, S., & Singhal, V. (2020). Ancient text character recognition using deep learning. International Journal of Engineering Research and Technology, 3(9), 2177-2184. DOI: 10.37624/ijert/13.9.2020.2177-2184
[23] Dixit, V., Hussain, N., Basak, S., Atturu, D., Mitra, D., & Bhattacharya, U. (2025). Deep Learning in Archiving Indus Script and Motif Information. Journal of Computer Applications in Archaeology, 8(1).DOI: 10.5334/jcaa.175
[24] Bi, X., Sun, Z., & Chen, Z. (2025). A novel unsupervised contrastive learning framework for ancient Yi script character dataset construction. npj Heritage Science, 13(1), 39.https://doi.org/10.1038/s40494-025-01600-6
[25] Wang, N., Wang, W., Li, B., Zhang, H., Jiao, Q., & Liu, C. (2025). Multi-modal ancient scripts recognition via deep learning with data homogenization and augmentation. npj Heritage Science, 13(1), 522.https://doi.org/10.1038/s40494-025-02095-x
[26] Diao, X., Bo, R., Xiao, Y., Shi, L., Zhou, Z., Xu, H., ... & Shi, D. (2025). Ancient Script Image Recognition and Processing: A Review. arXiv preprint arXiv:2506.19208. https://doi.org/10.48550/arXiv.2506.19208
[27] Idwan, S., Etaiwi, W., Rafayia, H., & Matar, I. (2025). A comprehensive review of statistical variants and enhancements of SMOTE oversampling method. *International Journal of Data Science and Analytics*, *20*(8), 6887-6904.https://doi.org/10.1162/coli_a_00481