Unveiling the Power of Convolutional Neural Networks in Hieroglyph Classification

Ayatollah Elkolally[1], Menna Rasslan[2] and Muhammed Ghareeb[3]

Abstract— **Recent advancements in Artificial Intelligence (AI), particularly in machine learning and deep learning, have opened new opportunities for supporting experts in fields beyond traditional information technology. One such domain is the study of ancient Egyptian hieroglyphics. This project applies convolutional neural networks (CNNs) to classify hieroglyph images from a dataset comprising 4,032 grayscale images of 171 unique hieroglyphs. The dataset was carefully preprocessed to optimize training and improve model performance. We evaluated the effectiveness of CNN architectures, including VGG16, Inception-v3, and Xception, utilizing transfer learning techniques to maximize accuracy. This work demonstrates the potential of deep learning to automate the recognition and classification of ancient scripts, contributing to AI-driven innovations in historical and archaeological research.**

I. INTRODUCTION

Artificial Intelligence (AI) and machine learning have rapidly transformed diverse fields, including archaeology, philology, and the human sciences. These technologies are not only automating tasks once handled manually, but are also enhancing human understanding in ways that were previously unimaginable. As AI applications expand into new domains, there is increasing interest in applying these methods to the study of ancient cultures, where they can provide valuable insights into long-forgotten languages and symbols. Among these, the classification and interpretation of Egyptian hieroglyphs—one of the oldest writing systems—has proven to be an area where AI can significantly contribute to scholarly efforts.

Hieroglyphs represent a vast array of ideograms, which can be classified into 26 categories and used to represent sounds, words, or concepts. These symbols were written in various styles, including monumental and cursive forms, and inscribed on a range of materials such as stone, wood, and papyrus.[1] Despite advancements in AI, the classification of hieroglyphs remains a significant challenge due to the diversity in the visual forms of the symbols and the complexity of their written forms. AI tools, particularly Convolutional Neural Networks (CNNs), offer a promising solution to this problem, enabling the automatic recognition and classification of hieroglyphic symbols with high accuracy.

Several studies have addressed ancient Egyptian language classification and retrieval. For instance, computer vision methods have been applied to identify hieroglyphs in Egyptian cartouches for museum navigation systems[2]. Other works have focused on text retrieval systems for hieroglyphic texts and the transliteration of signs[3]. Despite these efforts, tools for classifying and translating ancient Egyptian texts are still lacking and inaccurate.

In this work, we focus on the classification of individual hieroglyphs using CNNs, which are widely regarded as the most effective tool for visual recognition tasks. We use a dataset containing 4,032 grayscale images, representing 171 unique symbols published[4]. Through this study, we evaluate the performance of several well-established CNN architectures, including VGG16, Inception-v3, and Xception, applying transfer learning techniques to optimize their effectiveness for the task at hand. The experimental results reveal the high potential of deep learning models in classifying hieroglyphs, providing a foundation for developing AI-driven tools to support the study and interpretation of ancient Egyptian texts.

[1]Ayatollah Elkolally, 120210312, Department of Computer Science Engineering

[2]Menna Rasslan, 120210311, Department of Computer Science Engineering

[3]Muhammed Ghareeb, 120210357, Department of Computer Science Engineering

Despite the success of these models, the task of fully automating the recognition and interpretation of hieroglyphs remains complex. The intricacies of the writing system, along with the variations in style and medium, pose significant challenges. Nevertheless, AI offers exciting possibilities for advancing research in Egyptology, enabling researchers to explore historical texts with unprecedented speed and accuracy. This paper discusses the methodology, experimental results, and implications of using deep learning for hieroglyph classification, highlighting the growing role of AI in the humanities and its potential to open new avenues for archaeological and philological research.

II. Methods

The methodology we followed in our work is described below, including how data was preprocessed and the models used for transfer-learning

A. Data

1) About the dataset: The dataset published at , consists of a publicly available collection of grayscale images depicting Egyptian hieroglyphs. This dataset was derived from high-quality photographs taken inside the Pyramid of Unas, an ancient Egyptian monument. The images represent individual hieroglyphic symbols and are labeled in accordance with the widely recognized Gardiner Sign List, a comprehensive catalog of hieroglyphic symbols that assigns a unique alphanumeric code to each hieroglyph.

The dataset comprises 4,032 images, each having a fixed resolution of 75 × 50 pixels as showed in Fig.1 Fig.2, and every image exclusively represents a single hieroglyph. The dataset includes 171 unique hieroglyph types, ensuring coverage of a diverse set of symbols. These hieroglyphs are specific to the Unas Pyramid, making the dataset highly contextualized within a single historical and artistic domain. This homogeneity offers a controlled environment for researchers to explore and test algorithms for hieroglyph recognition without the variability introduced by differing historical periods, artistic styles, or materials.

The images in the dataset are in grayscale format, omitting color information to focus solely on the structural and visual characteristics of the hieroglyphs. The dataset has been preprocessed to ensure uniformity: each hieroglyph is cropped to its rectangular bounding box and resized to maintain a consistent image dimension. This preprocessing standardizes the dataset, facilitating its use in machine learning tasks.

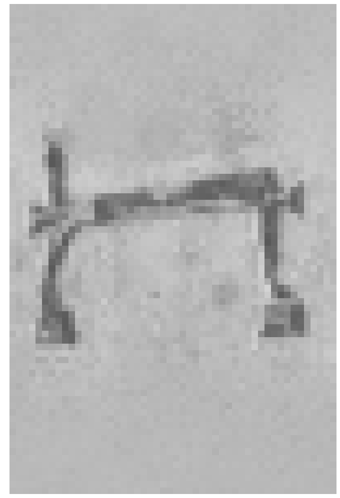

Fig. 1. Sample image from the dataset

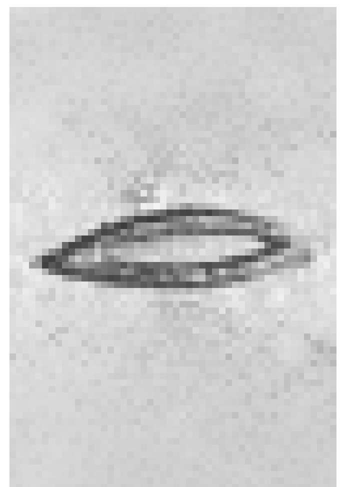

Fig. 2. Sample image from the dataset

Due to its curated nature and controlled environment, the dataset is particularly suited for developing and evaluating algorithms designed for hieroglyph classification, especially in scenarios where contextual consistency is crucial. However, its specificity to the Pyramid of Unas may limit its generalizability to other Egyptian hieroglyphic contexts. Despite this limitation, it provides an essential foundation for hieroglyph recognition studies and serves as a benchmark dataset for researchers in this domain.

2) Data Preprocessing: The data preprocessing pipeline was designed to prepare the dataset for training and evaluation while ensuring consistency, balance, and quality. Initially, the dataset exhibited class imbalance, with some classes containing significantly more images than others. To address this, the number of images per class was standardized to a target of 25 as showed in Fig.3 . For classes with excess images, downsampling was applied, where surplus images were randomly removed. Conversely, for underrepresented classes, upsampling was performed using data augmentation techniques, which included random horizontal flips and rotations of 90°, 180°, or 270°. This approach ensured that all classes were evenly represented in the dataset.

All images were processed to maintain uniformity and facilitate efficient model training. Each image was converted to grayscale, resized to dimensions of 100×100 pixels, and saved in JPEG format. Grayscale conversion reduced color variations, simplifying the feature extraction process, while resizing ensured computational efficiency. Additionally, for images with irregular contours, Gaussian textures resembling the original backgrounds were applied to fill missing areas, preserving the structural integrity of the hieroglyphs while standardizing their appearance.

The preprocessed dataset was stratified and divided into training, validation, and test sets in a 70:15:15 ratio. Stratified sampling ensured that each split maintained an equal proportion of samples across all classes, preserving the representativeness of the dataset. For compatibility with deep learning models, the images were further resized to 224×224 pixels, and pixel intensities were normalized to a range of [0, 1]. This normalization step stabilized the training process and ensured uniform input dimensions for the model.

Class labels, originally represented as folder names, were converted into numerical values using label encoding, allowing for seamless integration with the training pipeline. To further enhance model generalization, additional augmentations, such as random cropping, horizontal flipping, and rotations, were applied during preprocessing. These transformations simulated variations in orientation and framing, making the model more robust to real-world scenarios.

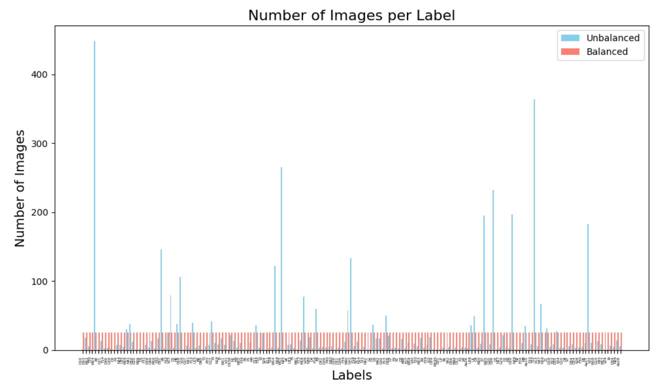

Fig. 3. Data Distribution before and after preprocessing

In summary, the preprocessing pipeline ensured a balanced and standardized dataset, ready for effective training and evaluation. It also addressed class imbalance, enhanced generalization through augmentation, and prepared the images in a format compatible with state-of-the-art deep learning models.

B. Transfer Learning with VGG16

To classify Egyptian hieroglyphs, a transfer learning approach utilizing the VGG16 model was implemented. VGG16, a widely recognized deep convolutional neural network pretrained on the ImageNet dataset, was chosen for its proven ability to extract meaningful features from image data.

The base VGG16 model was loaded without its fully connected layers, retaining only the convolutional layers, which were pretrained to extract low- and high-level image features. These layers were frozen to preserve the pretrained weights, ensuring efficient feature extraction while preventing overfitting on the relatively smaller hieroglyph dataset.

On top of the base model, custom fully connected layers were added to adapt the architecture for the hieroglyph classification task. The custom layers consisted of:

- A flattening layer to convert the convolutional output into a vector.
- Two dense layers, each containing 4096 neurons with ReLU activation, mimicking the architecture of AlexNet to enhance model capacity.
- Dropout layers with a rate of 0.5 applied after each dense layer to reduce overfitting.

- A final dense layer with a softmax activation function, producing class probabilities corresponding to the number of hieroglyph categories.

The model was compiled using the Adam optimizer with a learning rate of 10^{-4}, categorical cross-entropy loss, and accuracy as the evaluation metric. This configuration ensured stable and efficient training.

To improve performance and prevent overfitting, two callbacks were employed:

- **EarlyStopping:** Monitored validation loss, stopping training if it did not improve for 5 consecutive epochs, while restoring the best weights.
- **ReduceLROnPlateau:** Reduced the learning rate by a factor of 0.2 if validation loss plateaued for 3 epochs, with a minimum threshold of 10^{-6}.

The model was trained on the processed dataset using a batch size of 32 for up to 25 epochs, with a 70:15:15 split for training, validation, and testing. This transfer learning approach effectively leveraged pretrained features while adapting to the unique characteristics of the hieroglyph dataset.

C. Transfer Learning with InceptionV3

InceptionV3 is a state-of-the-art convolutional neural network (CNN) architecture that has demonstrated high performance across diverse image classification tasks. The model leverages an innovative design comprising "Inception modules," which utilize multiple kernel sizes and reduce computational complexity. In this study, InceptionV3 was employed to classify hieroglyphs through transfer learning, leveraging its pretrained ImageNet weights.

The InceptionV3 model was loaded without its fully connected layers, retaining the convolutional base for feature extraction. This base was pretrained on ImageNet, enabling it to extract hierarchical features effectively. The layers of the base model were frozen to prevent updates to the pretrained weights during training on the hieroglyph dataset.

On top of the frozen base, custom dense layers were added to adapt the model to the specific classification task:

- A **Global Average Pooling** (GAP) layer to reduce the spatial dimensions of the feature maps.
- A fully connected **Dense layer** with 512 neurons and ReLU activation to increase the model's capacity to learn task-specific features.
- A **Dropout layer** with a rate of 0.5 to mitigate overfitting.
- A final **Dense layer** with softmax activation, outputting probabilities corresponding to the 171 hieroglyph classes.

The model was compiled with the Adam optimizer, a learning rate of 10^{-4}, and categorical cross-entropy loss. To ensure efficient training, the following callbacks were utilized:

- **EarlyStopping:** Training was terminated early if the validation loss did not improve for 5 consecutive epochs, with the best weights restored.
- **ReduceLROnPlateau:** The learning rate was reduced by a factor of 0.2 if validation loss plateaued for 3 epochs, with a minimum learning rate threshold of 10^{-6}.

The training process involved 25 epochs, with a batch size of 16. Data augmentation techniques, such as random rotations and horizontal flips, were applied during training to improve the model's generalization capabilities. The processed dataset was split into 70% training, 15% validation, and 15% testing.

InceptionV3's architectural depth and ability to capture multi-scale features made it particularly effective for hieroglyph classification. Its superior performance underscores the advantages of leveraging transfer learning for domain-specific tasks.

D. Transfer Learning with Xception

To classify Egyptian hieroglyphs, we employed transfer learning using the Xception model, a state-of-the-art deep convolutional neural network known for its efficient use of depthwise separable convolutions. This architecture, pretrained on the ImageNet dataset, is widely regarded for its ability to extract hierarchical features from image data while maintaining computational efficiency.

The pretrained Xception model was utilized by excluding its top classification layers and preserving its convolutional layers, which are optimized for robust feature extraction. These layers were frozen to retain the pretrained weights, preventing overfitting on the hieroglyph dataset.

Custom layers were appended to the base model to adapt it for the hieroglyph classification task. These additional layers included:

- A global average pooling layer to reduce the spatial dimensions of the feature maps.
- A fully connected layer with 1024 neurons and ReLU activation, allowing for high-capacity feature learning.
- A dropout layer with a rate of 0.5, incorporated to mitigate overfitting by introducing regularization.
- A final dense layer with a softmax activation function, generating probabilities corresponding to the 171 hieroglyph classes.

The model was trained using the Adam optimizer with a learning rate of 10^{-4}, employing categorical cross-entropy as the loss function and accuracy as the evaluation metric. To further refine the training process, the following callbacks were implemented:

- **EarlyStopping:** This callback monitored the validation loss and terminated training if no improvement was observed for 5 consecutive epochs, while restoring the model's best weights.
- **ReduceLROnPlateau:** This callback dynamically reduced the learning rate by a factor of 0.2 whenever the validation loss plateaued for 3 epochs, ensuring a more stable optimization process.

The Xception model was trained on the hieroglyph dataset with a batch size of 16 and an epoch limit of 25. Images were resized to 299×299 to align with the input requirements of the architecture. The dataset was split into training, validation, and test sets in a 70:15:15 ratio to ensure fair evaluation.

The overall architecture of the Xception model is depicted in Fig. 4.

By leveraging the pretrained Xception model, the network effectively utilized its depthwise separable convolution layers to extract intricate visual features, enabling robust classification of hieroglyphs. The inclusion of custom top layers ensured that the model adapted specifically to the unique

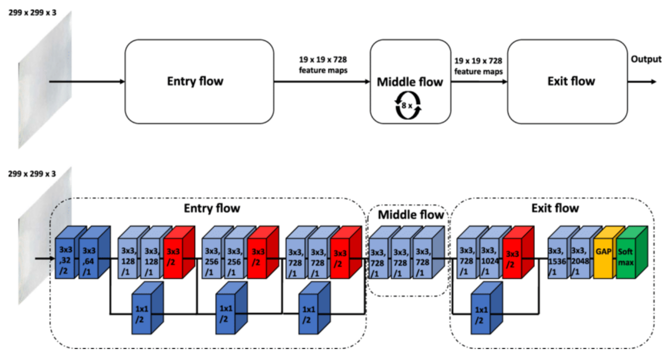

Fig. 4. Xception Architecture: Leveraging depthwise separable convolutions for efficient hierarchical feature extraction and enhanced classification performance.

characteristics of the hieroglyph dataset, achieving high accuracy while minimizing overfitting.

III. Results

The performance of the three Convolutional Neural Network (CNN) architectures—VGG16, InceptionV3, and Xception—was evaluated using the preprocessed hieroglyph dataset. This section provides a detailed comparison of their results, including accuracy, loss metrics, confusion matrices, and other visualizations.

A. Model Performance

TABLE I
TEST ACCURACY FOR EACH ARCHITECTURE

Model	Test Accuracy
VGG16	0.83
InceptionV3	0.84
Xception	0.76

As summarized in Table I, InceptionV3 achieved the highest test accuracy of 0.84, followed closely by VGG16 with 0.83. Xception achieved a slightly lower accuracy of 0.76, likely influenced by the model's sensitivity to dataset characteristics and architectural differences.

B. Training and Validation Performance

The training and validation accuracy and loss curves for all three architectures are shown in Figures 6 and 5. These plots illustrate the convergence trends and overfitting tendencies observed during training.

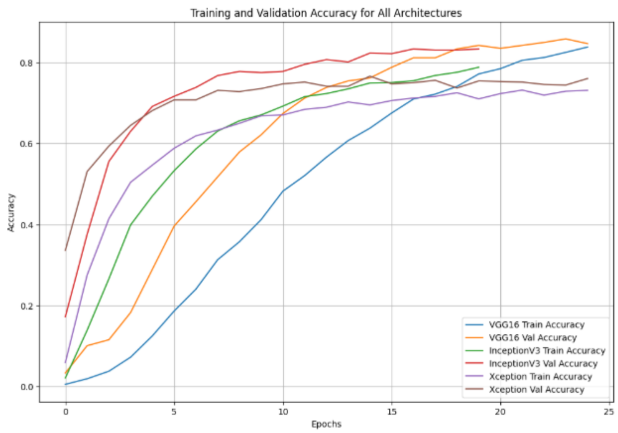

Fig. 5. Training and Validation Accuracy for All Architectures

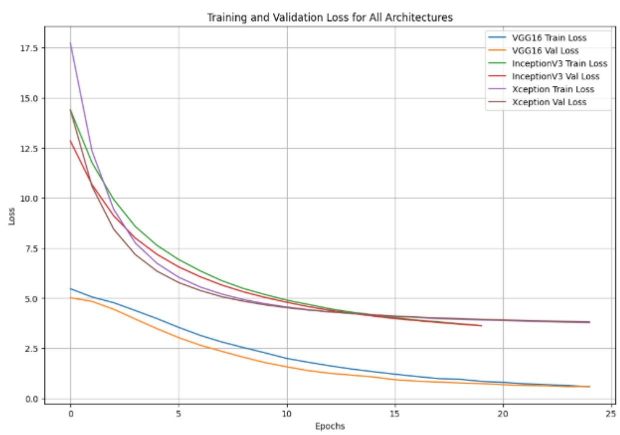

Fig. 6. Training and Validation Loss for All Architectures

Both VGG16 and InceptionV3 demonstrated smooth convergence with minimal overfitting, reflected in their validation accuracy and loss trends. In contrast, Xception showed larger variations in validation accuracy and loss, potentially indicating sensitivity to data distribution.

C. Confusion Matrices

To analyze classification performance in detail, confusion matrices for each model were generated, as shown in Figures 7, 8, and 9. These matrices provide insights into the distribution of misclassifications across the hieroglyph classes.

VGG16 and InceptionV3 showed fewer misclassifications compared to Xception, highlighting their higher precision and recall rates. Xception exhibited more misclassifications, particularly among visually similar hieroglyphs.

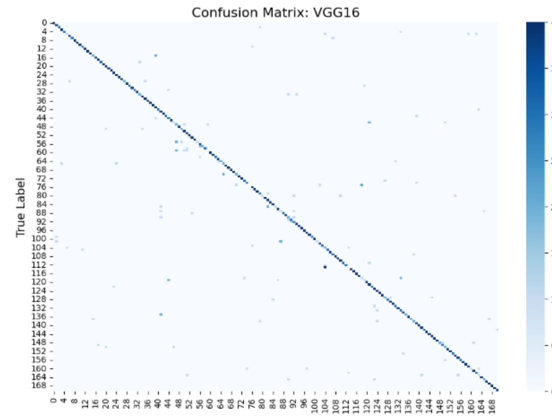

Fig. 7. Confusion Matrix: VGG16

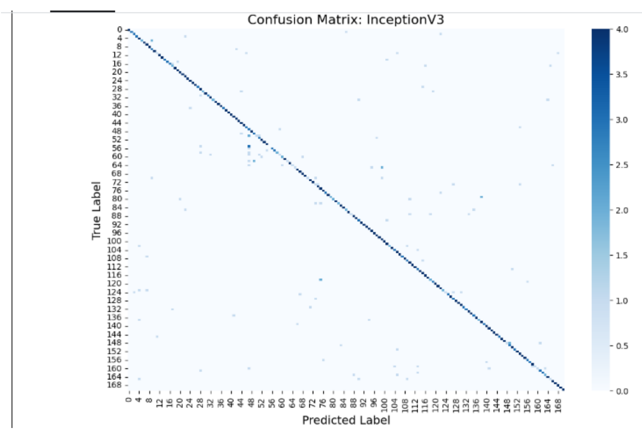

Fig. 8. Confusion Matrix: InceptionV3

D. Overall Metrics Comparison

Figure 10 presents a comparative bar chart of the key performance metrics—accuracy, precision, recall, and F1-score—for all three architectures.

The results confirm that InceptionV3 consistently outperformed the other models across all metrics, followed closely by VGG16. While Xception lagged slightly in comparison, its performance remained competitive.

IV. Discussion

The comparative evaluation of VGG16, InceptionV3, and Xception highlights several key observations regarding the application of convolutional neural networks (CNNs) to the classification of ancient Egyptian hieroglyphs.

A. Model Performance Analysis

The results indicate that **InceptionV3** outperformed the other architectures, achieving the high-

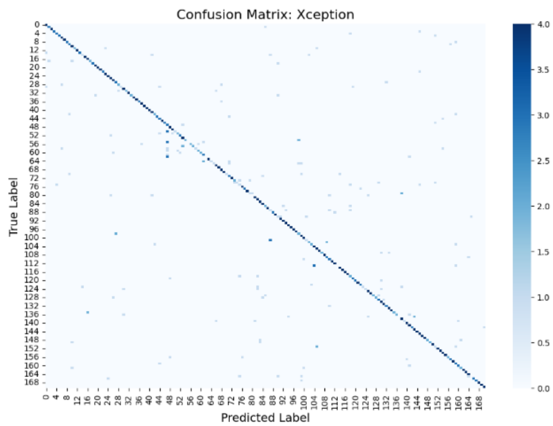

Fig. 9. Confusion Matrix: Xception

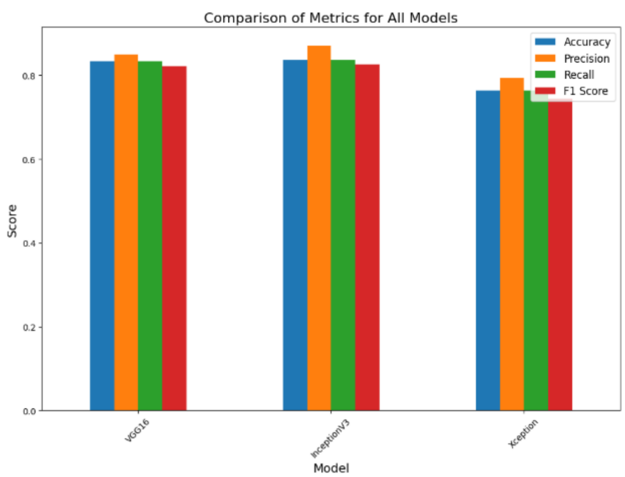

Fig. 10. Comparison of Metrics for All Models

est test accuracy of 0.84. This can be attributed to its architectural innovations, such as Inception modules, which capture multi-scale features efficiently. These features proved to be particularly effective in distinguishing hieroglyph classes with subtle structural variations.

VGG16, with a test accuracy of 0.83, closely followed InceptionV3. Despite its simpler architecture, VGG16 demonstrated robustness and generalizability, aided by its fully connected layers and dropout regularization. Its strong performance underscores its suitability for smaller, well-structured datasets like the hieroglyph dataset.

Xception achieved a test accuracy of 0.76, which, while slightly lower, highlights its potential when combined with additional data augmentation or hyperparameter tuning. The model's reliance on depthwise separable convolutions, which reduce computational complexity, may have limited its ability to capture complex hieroglyphic patterns effectively.

B. Overfitting and Regularization

The training and validation curves indicate that both VGG16 and InceptionV3 converged smoothly, with minimal overfitting. This is evident from their consistent validation accuracy and loss trends. In contrast, Xception exhibited greater variability in validation performance, possibly due to its sensitivity to the dataset's size and characteristics.

The incorporation of **dropout layers** and **early stopping** proved instrumental in mitigating overfitting, particularly for VGG16 and InceptionV3. These techniques ensured that the models generalized well to unseen data, as reflected in their high test accuracies.

C. Impact of Dataset Characteristics

The dataset's structure, comprising grayscale images with uniform dimensions, likely contributed to the strong performance of all three architectures. However, the limited dataset size posed challenges, particularly for deeper models like Xception, which may require larger datasets to fully exploit their capacity. Data augmentation played a critical role in addressing this limitation by introducing diversity into the training data.

D. Limitations and Future Directions

While the results are promising, the study is not without limitations. The dataset's specificity to the Pyramid of Unas restricts its generalizability to other hieroglyphic contexts. Future work could involve expanding the dataset to include hieroglyphs from diverse historical and geographical contexts, enhancing the models' robustness.

Additionally, exploring ensemble methods or fine-tuning pretrained models with domain-specific data could further improve classification accuracy. The application of explainable AI techniques to visualize feature importance would also provide valuable insights into the decision-making process of the models.

V. Conclusion

This study demonstrates the efficacy of CNNs, particularly InceptionV3 and VGG16, in automating hieroglyph classification. By leveraging transfer learning and advanced regularization techniques, the models achieved high accuracy, offering a foundation for future research in applying AI to Egyptology. The findings highlight the transformative potential of deep learning in decoding ancient scripts, contributing to the broader field of AI-driven humanities research.

The comparative evaluation of VGG16, InceptionV3, and Xception highlights the effectiveness of transfer learning for hieroglyph classification. InceptionV3's superior performance can be attributed to its deeper architecture and multi-scale feature extraction capabilities. VGG16 demonstrated robust performance, leveraging its simple yet effective dense layers and regularization. Xception, while achieving a slightly lower accuracy, showed promise with its depthwise separable convolution design and would benefit from further hyperparameter tuning and augmented training data.

In conclusion, the study demonstrates the utility of CNNs in hieroglyph classification, showcasing their potential in advancing automated Egyptology research. These findings lay a foundation for future work in leveraging deep learning for analyzing and interpreting ancient scripts.

References

[1] A. Gardiner, Egyptian Grammar. Oxford, U.K.: Griffith Institute, 1957.

[2] J. Duque-Domingo, P. Herrera, E. Valero, and C. Cerrada, "Deciphering Egyptian hieroglyphs: Towards a new strategy for navigation in museums," Sensors, vol. 17, no. 3, p. 589, 2017.

[3] E. Iglesias-Franjo and J. Vilares, "TIR over Egyptian hieroglyphs," in Proc. 27th Int. Workshop Database Expert Syst. Appl. (DEXA), 2016, pp. 198-203.

[4] M. Franken and J. van Gemert, "Automatic Egyptian hieroglyph recognition by retrieving images as texts," in Proc. 21st ACM Int. Conf. Multimedia. New York, NY, USA: Association for Computing Machinery, 2013, pp. 765-768,

[5] A. Barucci, C. Cucci, M. Franci, M. Loschiavo and F. Argenti, "A Deep Learning Approach to Ancient Egyptian Hieroglyphs Classification," in IEEE Access, vol. 9, pp. 123438-123447, 2021, doi: 10.1109/ACCESS.2021.3110082.