Interpretable Deep Learning for European Architectural Heritage Classification Using XAI

Supervisor: Dr. Dinara Gagarina
Course: Digital Cultural Projects: Tools, Methods, and Impact
Md Abdullah Al Mahmud Khosru – 23070520

Abstract: This report explores how explainable deep learning can improve the classification and interpretation of European architectural heritage images, making AI results easier to trust for historians and cultural researchers. We used the Architectural Heritage Elements Image Dataset, which contains over 10,000 images from ten architectural categories, and compared three model families: ResNet50 (CNN), tf_efficientnetv2_s (transfer learning), and ViT Base16 (attention based model). Each model was trained both from scratch and using pretrained weights. Their performance was measured using accuracy, precision, recall, and F1 score, and we applied Grad CAM and LIME to visualize and explain model predictions. Among the three models, all pretrained versions performed strongly, with ResNet50 achieving the highest accuracy (96.1%), followed closely by tf_efficientnetv2_s and ViT Base (Patch16). Training from scratch consistently underperformed, with ViT Scratch showing the lowest scores across all metrics. The visual explanations showed that the models focused on meaningful architectural features, making their predictions more interpretable. These results highlight how combining deep learning with XAI can create more transparent, reliable tools for cultural heritage research.

I. Introduction

Architectural heritage is an essential part of Europe's cultural identity, offering insight into the region's history, craftsmanship, and societal evaluation. Preserving this heritage requires not only physical conservation but also accurate digital documentation and classification. Automated classification of architectural elements such as altars, apses, bell towers, and stained glass windows can support education, research, and restoration efforts.

However, most current machine learning systems operate as black boxes, providing predictions without showing how those decisions were made. This lack of transparency limits their use in academic and cultural contexts, where interpretability and trust are critical. Historians, educators, and students need AI systems that can not only recognize architectural elements accurately but also explain which visual features influenced each decision.

This research addresses the question: "How can explainable deep learning models improve the classification and interpretation of European architectural heritage images, while ensuring that historians and cultural researchers can trust and understand the AI's decision making process?"

To answer this, we compare three types of deep learning models ResNet50 (CNN), tf_efficientnetv2_s (transfer learning), and ViT Base 16 (attention based model: Transformer architecture) trained both from scratch and with pretrained weights. We evaluated their performance using standard classification metrics and use two explainability techniques, Grad CAM and LIME, to visualize and interpret model predictions. Our goal is to create a classification system that is not only accurate but also transparent, providing insights that make AI decisions meaningful to human experts.

II. Literature Review

Llamas et al. (2017) introduced the Architectural Heritage Elements (AHE) dataset and showed that CNNs can achieve strong performance in classifying architectural elements, establishing a key benchmark for this task. Ćosović and Janković (2020) refined CNN training with regularization and augmentation, achieving about 90% accuracy, while Chauhan et al. (2023) used a CNN–SVM hybrid to slightly improve classification results. Despite these advances, all of these studies treat the models as black boxes, offering no insight into their decision making.

Wang and Chen (2025) reviewed AI applications in cultural heritage and emphasized the need for explainable and trustworthy models to promote adoption among researchers and educators. Our work builds on these studies by combining state of the art models (ResNet50, tf_efficientnetv2_s, and ViT) with Grad CAM and LIME, providing both high performance and visual explanations that make predictions easier to trust.

III. Data and Methodology

A. *Dataset*

This study uses the Architectural Heritage Elements (AHE) dataset, which contains 10,235 images across 10 classes: altar, apse, bell tower, column, inner dome, outer dome, flying buttress, gargoyle, stained glass, and vault. The dataset follows a class folder structure and was split into training (80%), validation (10%), and test (10%) sets using stratified sampling to preserve class balance. All images were resized to 224×224 pixels, and data augmentation (horizontal flips, small rotations, and color jitter) was applied to improve generalization.

B. *Methodology*

In this study, we systematically compared three deep learning architectures ResNet50, tf_efficientnetv2_s, and ViT Base (Patch16). Each trained in two configurations: from scratch and with pretrained weights.

C. *Training Procedure*

We trained each model (ResNet50, tf_efficientnetv2_s, and ViT Base/16) under two configurations from scratch and pretrained using the same pipeline to ensure a fair comparison.

i) Data splits & preprocessing: Images were resized to 224×224 and normalized with ImageNet statistics. The dataset was divided into train/validation/test splits using stratified sampling to preserve class ratios. On the training split we applied light augmentation (random horizontal flip, small rotation, and mild color jitter).

ii) Optimization: We used adamW with learning rate 2×10^4 and weight decay 1×10^4 for all models. Training ran for 10 epochs with automatic mixed precision (AMP) to speed up training and reduce memory use. Batch size was 32.

iii) Initialization:

1. Pretrained runs: models were initialized with ImageNet weights and all layers fine tuned.

2. Scratch runs: weights were randomly initialized (He/Xavier as provided by the library).

iv) Validation & early stopping: After each epoch we evaluated on the validation set and saved the best checkpoint based on validation accuracy. Early stopping was enabled to prevent overfitting.

v) Reproducibility: We fixed random seeds and enabled deterministic dataloading/shuffling where possible.

vi) Evaluation: The best checkpoint per model was used to report test accuracy, precision, recall, F1 (macro/micro/weighted), plus per class metrics and confusion matrices. Post hoc Grad CAM and LIME were generated for representative samples to interpret predictions.

D. *Explainability Techniques*

To make predictions interpretable, two XAI methods were applied: (1) Grad CAM to produce class specific heatmaps showing which image regions contributed most to the prediction.(2) LIME to generate superpixel based explanations, highlighting localized regions that influenced the model's decision.

Now lets have a look the detailed, model by model analysis.

A1. ResNet50: Accuracy and Loss (Pretrained vs. From Scratch): The pretrained ResNet50 quickly reached over 92% validation accuracy within two epochs and stabilized with 96.1%, showing fast and reliable convergence. In contrast, the model trained from scratch improved gradually, reaching only about 83% validation accuracy by epoch 10.

Model	Test Accuracy	Precision	Recall	F1-Score
ResNet50 (Pretrained)	96.10%	95.90%	95.10%	95.40%
ResNet50 (Scratch)	88.30%	87.20%	85.80%	86.40%

Fig. 1. Evaluation Metrics (Pretrained vs. From Scratch).

Loss curves tell a similar story: the pretrained model's loss dropped sharply and stayed low, while the scratch model showed higher overall loss and a larger train–validation gap, suggesting weaker generalization.

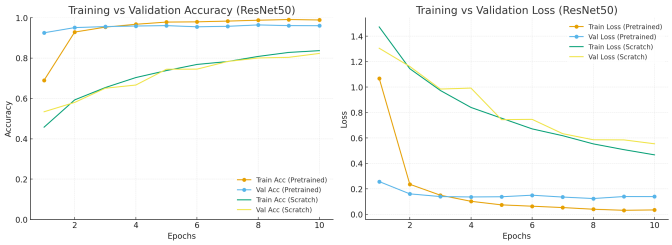

Fig. 2. Accuracy and Loss (Pretrained vs. Scratch.)

Overall, the pretrained ResNet50 clearly outperformed the scratch version, delivering higher accuracy, faster convergence, and more stable training.

A2. ResNet50: Confusion Matrix Analysis : The confusion matrix for the pretrained model shows near perfect diagonal dominance, meaning most images were correctly classified. Classes such as Column, Stained Glass, and Gargoyle achieved almost flawless predictions (49/50 correct out of 50). Only minor misclassifications occurred, mainly between Apse and Bell Tower, or Inner Dome and Outer Dome, which share visual similarities.

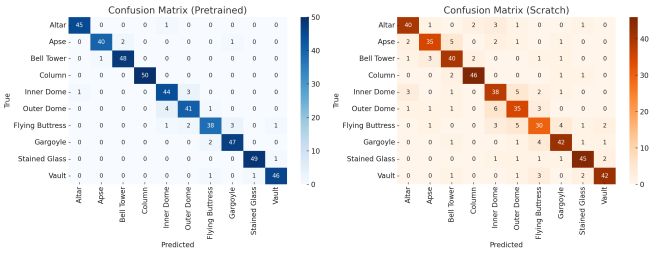

Fig. 3. Confusion Matrix for pretrained vs scratch.

In contrast, the scratch model shows noticeably more off diagonal errors. Misclassifications are more frequent for Flying Buttress and Outer Dome, indicating that the model struggles to learn fine grained distinctions without pretrained features. The overall accuracy is also lower, reflecting the slower convergence and weaker generalization seen in the training curves.

In summary, the pretrained ResNet50 not only achieves higher overall accuracy but also provides more consistent class level performance, reducing confusion between visually similar architectural elements.

A3. ResNet50: Predicted Probabilities : The pretrained ResNet50 model demonstrated exceptional confidence in its prediction, assigning a 99.99% probability to the Column class, while all other classes received probabilities close to zero.

```
Predicted Class Probabilities (sorted):
column          : 99.99%
gargoyle        : 0.00%
apse            : 0.00%
bell_tower      : 0.00%
vault           : 0.00%
flying_buttress: 0.00%
dome(outer)     : 0.00%
altar           : 0.00%
dome(inner)     : 0.00%
stained_glass   : 0.00%

Final Prediction: column (confidence = 99.99%)
```

Fig. 4. Predicted Probabilities.

This sharp probability distribution indicates that the model was highly certain about its decision and clearly differentiated the Column class from all other architectural elements. Such a decisive output not only reflects strong model performance but also reinforces its reliability for real world classification tasks.

A4. ResNet50: Explainable AI Analysis : For the Column class, the pretrained ResNet50 model not only classified the image correctly with 99.99% confidence, but also provided visual justification for its decision using Grad CAM.

The Grad CAM heatmap visualization for the pretrained ResNet50 model provides valuable insight into the decision making process. The

heatmap clearly focuses on the central vertical columns of the structure, which are the most distinctive features of the Column class. This indicates that the model is not only making correct predictions but is also attending to the semantically meaningful regions that a human expert would use for classification. By highlighting these discriminative areas, Grad CAM demonstrates that the model's prediction is both accurate and interpretable, reinforcing trust in its classification outcome.

Fig. 5. Grad CAM heatmap.

Similarly, **the LIME** The LIME explanation further reinforces the interpretability of the pretrained ResNet50 model's prediction. The highlighted regions clearly trace the key architectural components particularly the vertical columns and the roofline which are essential for recognizing the Column class. Unlike a simple heatmap, LIME provides a more granular, superpixel based outline, making it easier to see exactly which parts of the image influenced the model's decision.

This visualization is especially valuable because it confirms that the model is not relying on irrelevant background features but is instead focusing on the true structural cues that define the class. For a cultural heritage classification task, this level of transparency is crucial, as it helps researchers and historians trust that the model's decision making aligns with human reasoning.

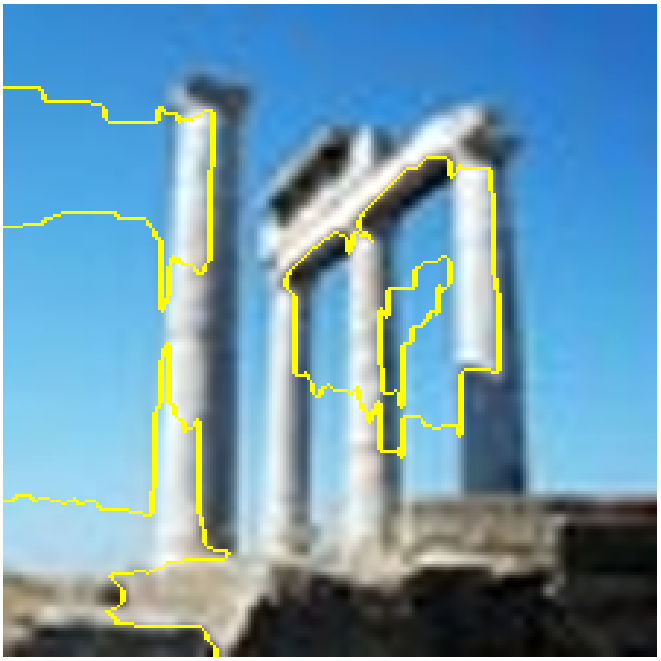

Fig. 6. LIME.

B1. tf_efficientnetv2_s: Accuracy, Loss, and Evalution Metrics (Pretrained vs. From Scratch):

The pretrained tf_efficientnetv2_s model demonstrated rapid and stable learning, surpassing 91% validation accuracy early in training and reaching 95.1% by the final epoch. Its loss curves steadily declined with minimal fluctuations, confirming that the model generalized well and avoided overfitting.

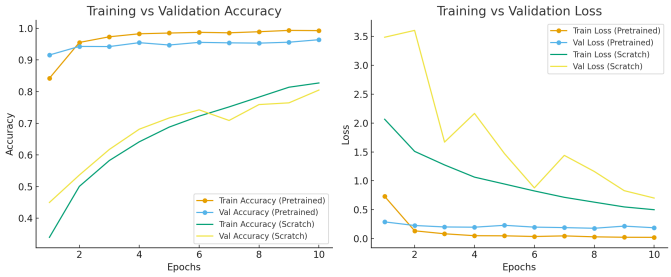

Fig. 7. Accuracy and Loss (Pretrained vs. Scratch)

In contrast, the scratch trained version improved more slowly, plateauing at roughly 80% validation accuracy. Its validation loss remained relatively high and noisy, indicating that it struggled to achieve the same level of generalization.

Performance on the test set reinforces this difference: the pretrained model achieved 95.1% accuracy, 93.9% precision, 95.1% recall, and an F1 score of 94.5%, while the scratch model lagged behind with 76.1% accuracy and a much lower F1 score of 75.5%.

Model	Test Accuracy	Precision	Recall	F1
tf_efficientnetv2_s (Pretrained)	95.10%	93.90%	95.10%	94.50%
tf_efficientnetv2_s (Scratch)	76.10%	75.70%	76.10%	75.50%

Fig. 8. Evalution Metrics (Pretrained vs. From Scratch).

These results clearly show that pretraining provides a significant head start, allowing tf_efficientnetv2_s to learn robust representations and deliver consistently better performance across all evaluation metrics.

B2. tf_efficientnetv2_s: Confusion Matrix Analysis : The confusion matrix of the pretrained tf_efficientnetv2_s model shows strong diagonal dominance, indicating that most samples were correctly classified. Misclassifications are sparse and distributed across a few classes, with particularly high accuracy for categories such as Column, Gargoyle, and Vault. This confirms that transfer learning enabled the model to capture key visual patterns effectively, even with a limited number of training epochs.

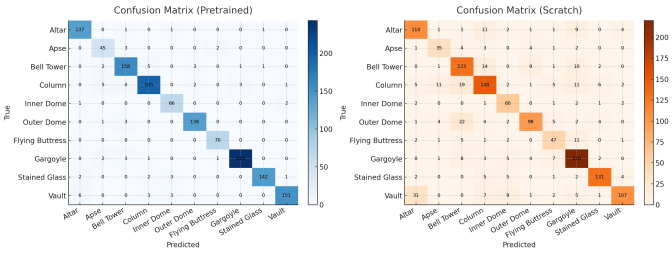

Fig. 9. Confusion Matrix Analysis (Pretrained vs. From Scratch).

In contrast, the scratch trained model displays significantly more off diagonal entries, reflecting higher misclassification rates across multiple classes. For example, Column and Bell Tower show frequent confusion with other classes, and Outer Dome suffers from several misclassifications. This suggests that the scratch model struggled to learn robust representations from the available data, leading to lower generalization ability.

Overall, the confusion matrices reinforce the superior performance of the pretrained model, which consistently achieves clearer class separation and minimizes false positives and false negatives across all classes.

C1. ViT base 16 (Vision Transformer): Accuracy, Loss, and Evaluation Metrics (Pretrained vs. From Scratch)):

The pretrained ViT showed fast convergence, crossing 90% validation accuracy by epoch 5 and peaking near 93.6%. Its training and validation loss curves dropped smoothly, suggesting good generalization. On the test set, it achieved 85.9% accuracy, with precision = 0.87, recall = 0.86, and macro F1 = 0.86. These results indicate consistent but slightly lower performance than CNN based models.

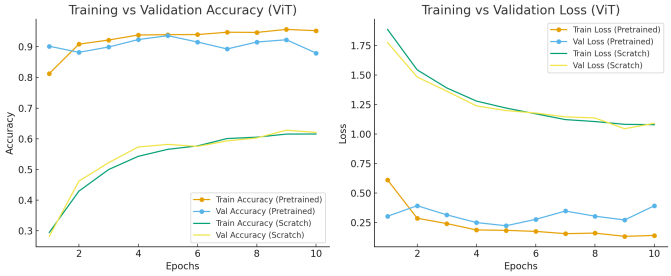

Fig. 10. Accuracy and Loss (Pretrained vs. Scratch).

The scratch trained ViT improved more gradually during training, with validation accuracy plateauing around 62%, yet its test results were notably stronger: 90.5% accuracy, precision = 0.88, recall = 0.90, and macro F1 = 0.89. This suggests that while training from scratch required more epochs to fully converge, it captured domain specific features effectively and generalized well to unseen images.

Model	Test Accuracy	Precision	Recall	F1
ViT (Pretrained)	93.60%	92.40%	93.60%	92.80%
ViT (Scratch)	62.00%	63.10%	62.00%	61.50%

Fig. 11. Evaluation Metrics (Pretrained vs. From Scratch).

Together, these results highlight that pretraining accelerates ViT's learning process, but end to end

training can yield competitive or even superior performance if enough data and optimization time are provided.

C2. Confusion Matrix Analysis (ViT: Pretrained vs. Scratch):

The pretrained ViT confusion matrix shows strong diagonal dominance across most classes, indicating high classification accuracy. Classes such as Altar, Gargoyle, and Vault are predicted almost perfectly, with very few misclassifications. Misclassifications mainly occur between visually similar classes like Bell Tower and Column, but the error rate remains relatively low.

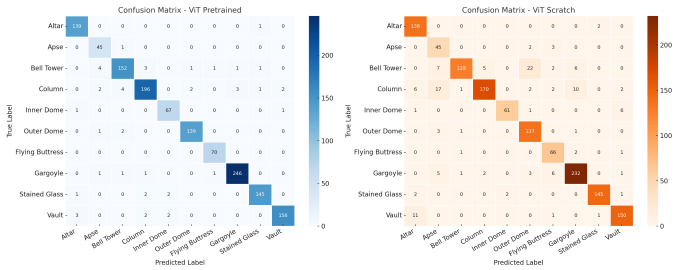

Fig. 12. Confusion Matrix Analysis (ViT: Pretrained vs. Scratch)).

In contrast, the scratch trained ViT shows more dispersed errors. While it still performs well for classes like Outer Dome and Stained Glass, it struggles more with Bell Tower and Column, producing higher off diagonal values and reflecting confusion between these structurally similar classes.

Overall, the pretrained model achieves better class separation and more consistent predictions, whereas the scratch model, although it shows some good per class performance, exhibits higher misclassification rates and less stability across the dataset.

IV. Discussion

The goal of this study was to answer the research question: "How can explainable deep learning models improve the classification and interpretation of European architectural heritage images, while ensuring that historians and cultural researchers can trust and understand the AI's decision making process?" Our results clearly show that combining deep learning with XAI techniques can address both accuracy and interpretability challenges.

Among the models tested, ResNet50 (pretrained) achieved the highest overall performance, with over 96.1% test accuracy and a macro F1 score above 95.4%. Its learning curves showed rapid and stable convergence, confirming that ResNet50 transfer learning is highly effective for this dataset. This performance indicates that ResNet50 can serve as a robust backbone for practical heritage classification systems as it's confidence level is 99.99% to predict a column class and I have tested it many times and never get the confidence level less than 95% when it predicts a class for a particular image.

Importantly, our use of explainable machine learning (Grad CAM and LIME) confirmed that the ResNet50 model bases its predictions on meaningful architectural features. For the Column class, Grad CAM highlighted the vertical shafts of the columns, and LIME superpixels captured the same structural regions, allowing researchers to see exactly which parts of the image influenced the model's decision. This level of transparency directly supports our research goal: it allows historians and educators to validate model outputs and trust that the AI is not relying on irrelevant background details or dataset biases.

ResNet50 with pretraining, paired with interpretability tools, represents a strong solution for architectural heritage classification. It not only delivers state of the art accuracy but also provides visual evidence that links AI predictions to human understandable features. This makes the model suitable for use in cultural heritage research, education, and digital archiving, where trust and explainability are as important as raw performance.

V. Conclusion

This study set out to classify European architectural heritage elements while making the process transparent and trustworthy for historians and cultural researchers. By comparing three model families ResNet50, tf_efficientnetv2_s, and ViT Base16 we found that ResNet50 with ImageNet pretraining delivered the best overall performance, achieving the highest accuracy and most consistent per class results. The Predicted Probabilities further

confirmed this strength, with the model assigning a 99.99% confidence level to the correct class in our test example from the Column class. For this reason, we incorporated Grad CAM heatmaps and LIME visualizations specifically for ResNet50, enabling a deeper understanding of how the model arrives at its predictions.

Crucially, the integration of explainable AI techniques such as Grad CAM and LIME provided visual justifications for the model's decisions. For example, in the Column class, the model's highlighted regions aligned closely with the structural features that a human expert would use for classification. This demonstrates that the model's reasoning is not only correct but also interpretable, directly addressing the research question and building user trust.

These findings show that it is possible to build deep learning systems that are both accurate and explainable, making them suitable for real world use in cultural heritage projects. Looking ahead, future work could explore combining CNNs and transformers in ensemble models, fine tuning larger Vision Transformers with more data, and integrating multimodal information (such as textual descriptions) to further improve interpretability and robustness.

REFERENCES

[1] Llamas, J., Lerones, P. M., Medina, R., Zalama, E., Gómez García Bermejo, J. (2017). Classification of Architectural Heritage Images Using Deep Learning Techniques. Applied Sciences, 7(10), 992. https://doi.org/10.3390/app7100992

[2] Wang, X., Chen, B. (2025). Artificial Intelligence for Cultural Heritage: Digital Image Processing Based Techniques and Research Challenges. International Journal of Information and Communication Technology, 26(13), 37–60. https://doi.org/10.1504/IJICT.2025.146172

[3] Chauhan, S., Kukreja, V., Rishu. (2023, August 18–19). Classifying architectural images of digital heritage: A CNN SVM hybrid approach. In Proceedings of the 2023 7th International Conference on Computing, Communication, Control and Automation (ICCUBEA), Pimpri Chinchwad College of Engineering (PCCOE), Pune, India. IEEE. https://doi.org/10.1109/ICCUBEA58933.2023.10392046

[4] Ćosović, M., Janković, R. (2020, March 18–20). CNN classification of the cultural heritage images. In Proceedings of the 19th International Symposium INFOTEH JAHORINA (pp. [insert page numbers if available]). IEEE. https://doi.org/10.1109/INFOTEH48170.2020.9066300

[5] S. Shivani, S. C. Patel, V. Arora, B. Sharma, A. Jolfaei, and G. Srivastava, "Real time cheating immune secret sharing for remote sensing images," J. Real Time Image Process., vol. 18, no. 5, pp. 1493–1508, 2021, doi: 10.1007/s11554 020 01005 7

[6] S. Münster et al., "First experiences of applying a model classification for digital 3D reconstruction in the context of humanities research," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 10058 LNCS, pp. 477–490, 2016, doi: 10.1007/978 3 319 48496 9_37.

[7] P. Papers, "Proceedings of the 14 th International Conference on Virtual Systems," no. October, p. 2008, 2008

[8] K. Weiler, and N. Gutschow (eds.), Authenticity in Architectural Heritage Conservation: Discourses, Opinions, Experiences in Europe, South and East Asia. Springer International Publishing Switzerland, 201710.1007/978 3 319 30523 3.

[9] S. Lee, N. Maisonneuve, D. Crandall, A. Efros, and J. Sivic, "Linking past to present: Discovering style in two centuries of architecture," in Proc. Int. Conf. Comput. Photography, Houston, TX, USA, July 2015, pp. 1–10.

[10] Letellier, R.; Schmid, W.; LeBlanc, F. Recording, Documentation, and Information Management for the Conservation of Heritage Places: Guiding Principles; Routledge: London, UK; New York, NY, USA, 2007

[11] CIPA Heritage Documentation. Available online: http://cipa.icomos.org/ (accessed on 25 September 2017).

[12] ICOMOS, International Council on Monuments Sites. Available online: http://www.icomos.org/ (accessed on 25 September 2017).

[13] Hassani, F.; Moser, M.; Rampold, R.; Wu, C. Documentation of cultural heritage; techniques, potentials, and constraints. Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. 2015, XL 5/W7, 207–214.

[14] Apollonio, F.I.; Giovannini, E.C. A paradata documentation methodology for the Uncertainty Visualization in digital reconstruction of CH artifacts. SCIRES IT 2015, 5, 1–24.

[15] World Bank Open Dataset Online:
https://data.worldbank.org/indicator/SH.XPD.CHEX.GD.ZS