DEVELOPMENT OF A DEEP LEARNING MODEL FOR SYNTHETIC VS. REAL IMAGE CLASSIFICATION

Bernardo Gandara – Ignacio Alarcon Varela, CCT College Dublin May 2025

ABSTRACT

This project develops a deep learning model to classify images as either AI-generated or real, addressing the growing challenge of synthetic media detection. Using the DeepGuardDB dataset and guided by the CRISP-DM methodology, we implemented and compared three Convolutional Neural Networks (CNNs) architectures via transfer learning. The best-performing model was further optimised using hyperparameter tuning and fine-tuning techniques The resulting model achieved strong accuracy and generalisation, making it a promising candidate for real-time deployment and practical use across diverse industries.

RESEARCH QUESTIONS

The four questions below represent the core focus of the entire study:

Q.1. What preprocessing steps are required to prepare the dataset for compatibility with different CNN architectures?

Q.2 Which CNN architecture delivers the best performance for our binary classification task?

Q.3 How can the selected model be further optimised to enhance its accuracy and generalisation capabilities?

Q.4 Does the final model meet the defined business success criteria, demonstrating potential for integration into real-world applications?

CONCEPTUAL FRAMEWORK

This project followed the CRISP-DM methodology (Cross-Industry Standard Process for Data Mining), a widely adopted framework for structuring data science workflows. CRISP-DM divides the project into six logical stages forming a cyclical process that supports iterative improvement. *Cirillo, (2017)*

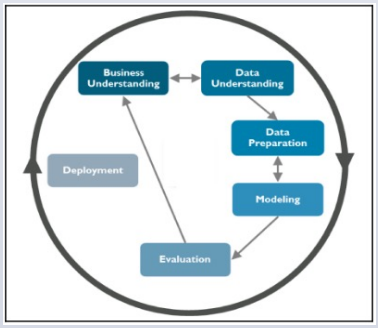

Figure 1 – CRISP-DM stages. Source: Cirillo, (2017)

BUSINESS UNDERSTANDING

The rapid growth of generative AI has made it increasingly difficult to distinguish real images from synthetic ones. This poses serious risks in the spread of fake news and misleading narratives, where AI-generated visuals can be used to manipulate user perception. As authenticity becomes harder to verify, developing accurate detection tools is essential for safeguarding information integrity.

DATA UNDERSTANDING

We used the DeepGuardDB dataset, containing 13,000 images evenly split between real photos (from MS-COCO and Flickr30k) and AI-generated images produced by Stable Diffusion 3, Imagen, and DALL-E 3. This dataset is designed to support research in distinguishing real from synthetic visual content.

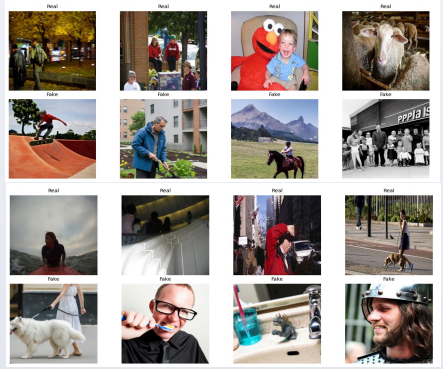

Figure 2 - Random samples taken from original dataset

DATA PREPARATION

The dataset was split into training, validation, and test sets to ensure independent model development, tuning and evaluation.

We applied Data Augmentation to the training set using random transformations—such as flips, rotations, zooms, and brightness/contrast adjustments–to enhance generalisation and reduce overfitting.

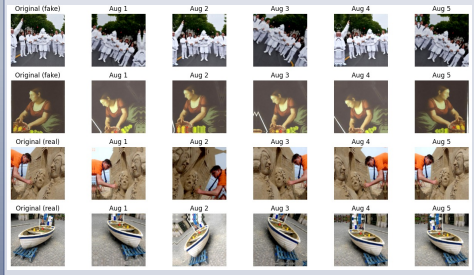

Figure 3 – Data Augmentation applied in a few samples from training set.

MODELLING

Training a CNN from scratch is intensive due to the complexity of tuning kernels, filter sizes, and weight parameters through backpropagation. (Mirza Rahim Baig et al., 2020)

We applied Transfer Learning using three increasingly sophisticated architectures: VGG-16 (plain network), ResNet-50 (residual learning) and EfficientNet (compound scaling).

To ensure a fair comparison of feature extraction performance, we applied the same custom classification head to all three models and evaluated them on the same test set.

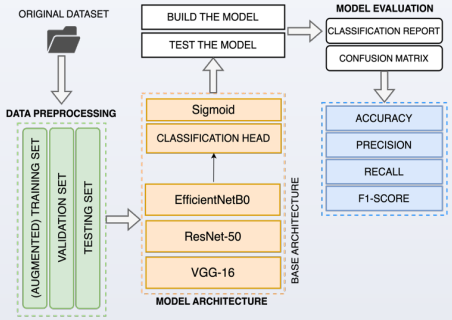

Figure 4 – Modelling process from data preparation to model evaluation.

Custom Classification Head

All models used a shared custom classification head consisting of a fully connected layer with 128 neurons and ReLU and the final layer was a fully connected output neuron with a Sigmoid activation.

EVALUATION

We evaluated all models using classification report metrics and confusion matrices to assess classification performance. This analysis guided the selection of the best-performing architecture for further optimisation.

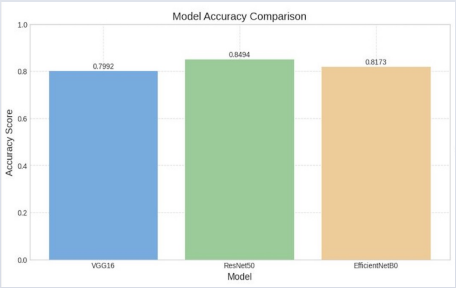

Figure 5 – Accuracy Comparison across multiple architectures.

ResNet-50 achieved the highest accuracy (84.9%), outperforming both EfficientNetB0 and VGG-16. Based on its superior performance, it was selected for further optimisation through hyperparameter tuning and fine-tuning.

HPO & Fine-Tuning

To improve model performance, we first applied hyperparameter optimisation using Keras Tuner's RandomSearch to identify the best configuration for the classification layers (Team, n.d.) . Then, we performed fine-tuning by unfreezing and retraining the top layers of the ResNet-50 base model, allowing it to learn higher-level features tailored to our specific dataset. (TensorFlow, 2025)

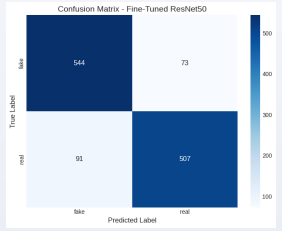

Figure 6 – Final Confusion matrix displaying the TP, TN, FP and FN of the final model.

The confusion matrix shows strong classification capability with relatively low misclassification rates. These results confirm the model's reliability in distinguishing AI-generated content from real photographs.

DEPLOYMENT

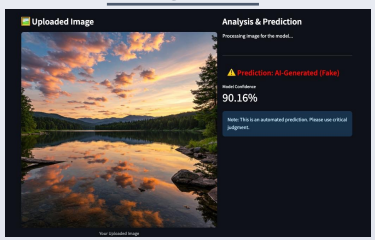

Figure 7 – Model deployed in Streamlit web application performing classification.

We deployed the fine-tuned ResNet-50 model through a Streamlit web application. The app offers an interactive interface for real-time image classification. With an accuracy of 86.5% and balanced precision and recall, the model is well-suited for practical use in tasks like media verification and content moderation.

CONCLUSIONS

A summarised answer to the research questions are presented below:

Q.1. All images were resized to 224x224 pixels and converted to .jpg format for consistency. Data augmentation was applied to the training set, and each architecture followed its specific preprocessing pipeline.

Q.2. ResNet-50 outperformed VGG-1 and EfficientNetB0, delivering the highest accuracy and most balanced results across key classification metrics.

Q.3. We used Keras Tuner to identify the optimal hyperparameters and applied fine-tuning to retrain the top layers of the ResNet-50 base, leading to improved performance and better generalisation.

Q.4. Yes, the model achieves high accuracy, maintains balanced performance across metrics, generalises across different image sources, and is fully deployable through a Streamlit app, showing strong potential for real-world integration.

REFERENCES

Cirillo, A. (2017). *R data mining : implement data mining techniques through practical use cases and real-world datasets*. Birmingham, UK: Packt Publishing.

Ikram Reghioua, Mouna Yasmine Namani, Gueltoum Bendiab, Mohamed Aymen Labiod, Stavros Shiaeles, 2024. DeepGuardDB: Real and Text-to-Image Synthetic Images Dataset. Available at: https://dx.doi.org/10.21227/10ap-pk52.

Mirza Rahim Baig, Joseph, T.V., Nipun Sadvilkar, Mohan Kumar Silaparasetty and So, A. (2020). *The Deep Learning Workshop*. Packt Publishing Ltd.

Team, K. (n.d.). *Keras documentation: KerasTuner*. [online] keras.io. Available at: https://keras.io/keras_tuner/.

TensorFlow. (2025). *Transfer learning with a pretrained ConvNet | TensorFlow Core*. [online] Available at: https://www.tensorflow.org/tutorials/images/transfer_learning.