Detecting AI-generated faces

DOMINIKA WIŚNIEWSKA

Problem Definition

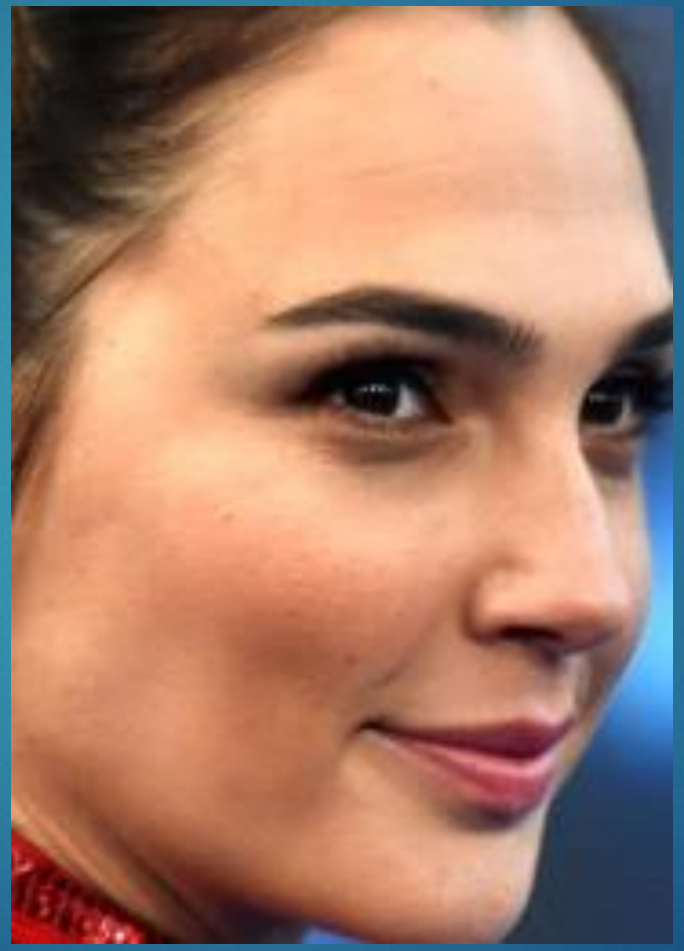
- Objective: Design and implement a machine learning system capable of predicting whether a presented image of a face is AI-generated.
- AI-generated faces have become highly realistic, making it increasingly difficult to distinguish them from real human faces.
- Today everyone can use one of many tools to generate an AI image, but there is no requirement for them to be labelled as such.
- It is especially a problem when such images are used with malicious intent. This poses significant risks, including identity theft, misinformation, and fraud.

Importance for the target group

- This issue is crucial for multiple sectors:
- **Media & Journalism:** Preventing the spread of deepfake-driven misinformation that can distort reality and influence public perception.
- **Cybersecurity:** Protecting users from fraudulent accounts and identity theft in digital spaces.
- **Social Media Platforms:** Enhancing content moderation to identify and flag AI-generated accounts.
- **Law Enforcement:** Assisting in digital fraud investigations by identifying synthetically created identities.

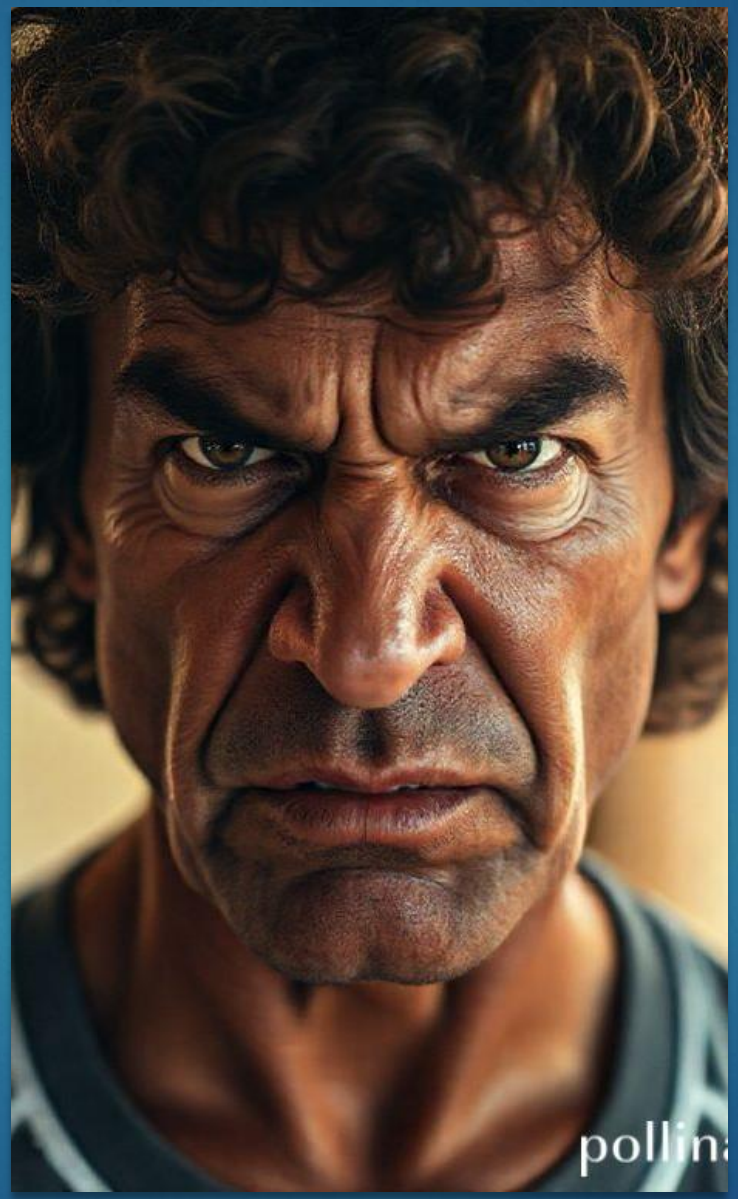
Dataset

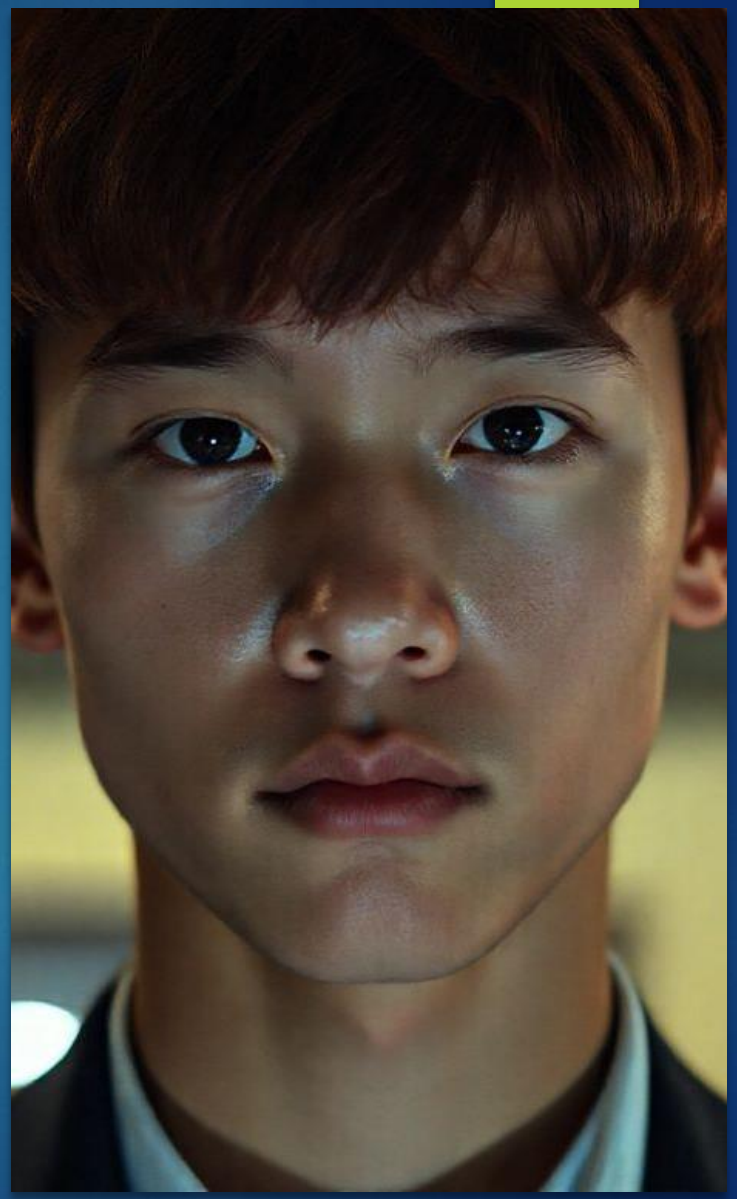
- Example of real faces



Dataset

- Example of AI-generated faces

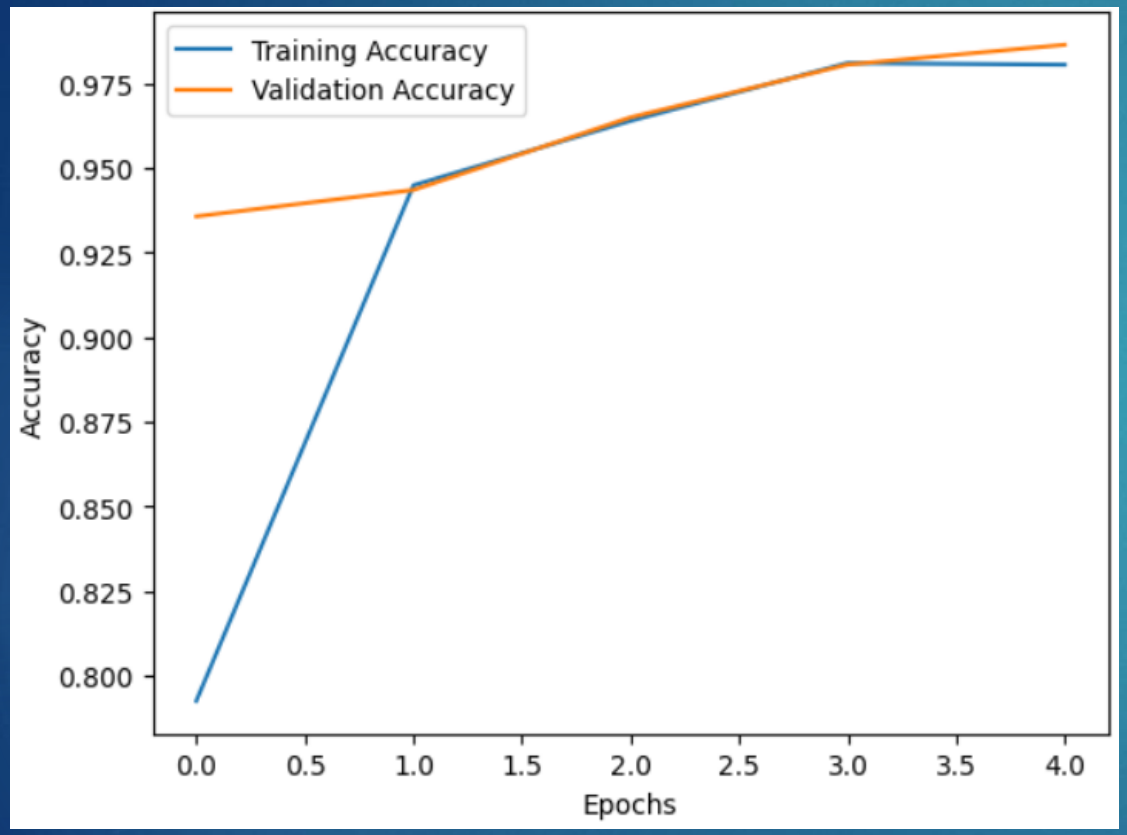


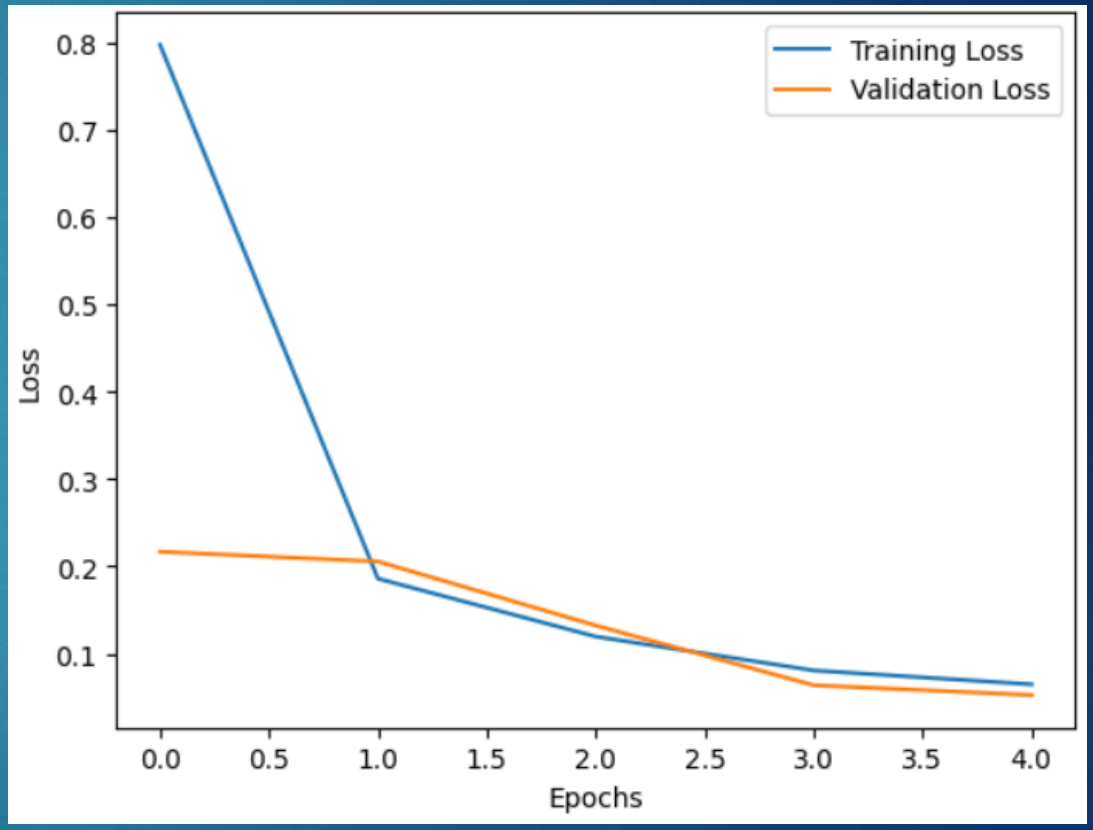


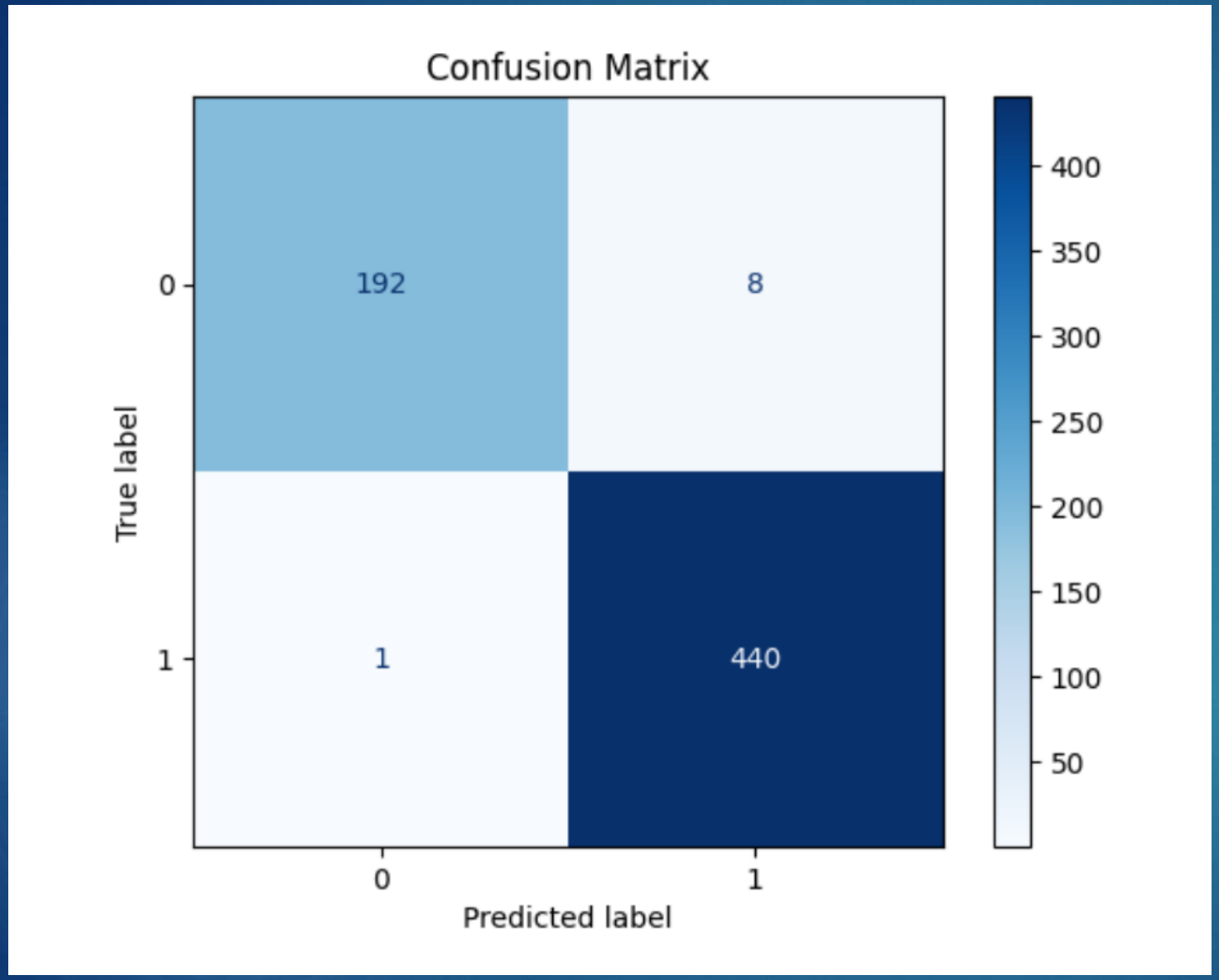
Reasons behind design selections

- **ResNet50 as Base Model:** Chosen for its powerful feature extraction capabilities and pretrained weights on ImageNet, reducing the need for extensive training.
- **Frozen Layers:** Freezing the base model layers ensures that only the custom classification layers are trained, preventing overfitting and speeding up convergence.
- **MaxPooling2D:** Reduces spatial dimensions while retaining key features.
- **Flatten Layer**: Converts the feature maps into a one-dimensional vector, allowing for better feature extraction before classification.
- **Dense Layers (128 neurons):** Helps capture complex relationships before making the final prediction.
- **ReLU Activation**: Introduces non-linearity in the dense layer, improving learning capacity and model performance.
- **Binary Cross-Entropy Loss:** The best choice for binary classification tasks.
- **Adam Optimizer:** Selected for its adaptive learning rate, leading to faster and more stable convergence.

Statistics of the model







Statistics of the model

Originality of the idea/approach

- Right now, the only way to discern an image of an AI face from a real one is to know the common mistakes AI makes when it comes to image generation.
- We have programs that check whether the text is AI-generated or not, but I can't think of a program that would reliably do the same for an image.

Novelty of AI-based solution

- There are countless ways this model could be implemented:
- Automatically detect and flag AI-generated profile pictures, preventing impersonation and fake accounts.
- Assist media organizations in verifying the authenticity of images before publication, reducing the spread of misinformation.
- Enhance identity verification processes for online services, helping banks and businesses prevent AI-generated identity fraud.

Conclusion

- AI-generated faces present a growing challenge for digital authenticity. This deep learning-based detection model offers an innovative and scalable solution. By applying this technology across media, cybersecurity, and law enforcement, we can significantly enhance trust in online interactions. This is a critical step toward mitigating the risks posed by synthetic media.

Thank you!