King Saud University
College of Computer and Information Sciences
Information Technology Department

جامعة الملك سعود
King Saud University

كلية علوم الحاسب والمعلومات
قسم تقنية المعلومات

IT 362 Data Science Project

Prepared by

Student name	Student ID
Nouf Alaskar	443200456
Nora Albyahi	443200479
Lana Albogami	444201031
Layan Aldbays	444200653
Noor algumlas	444200811

Supervised by:
I. Mashael Aldayel

Second Term 1445 H

Table of Contents

1. Introduction:

Artificial intelligence has made significant advancements in recent years, particularly in the field of image generation. Cutting-edge AI models are now capable of producing photorealistic images that closely mimic real-world photographs. This progress raises intriguing questions about whether artificial intelligence is currently advanced enough to generate images that are indistinguishable from real photographs. This project aims to investigate this very question, exploring the capabilities of current AI models. We will use a classification model to examine the accuracy and generalization of machine learning in this context. Through this research, we seek to contribute to the understanding of AI's current capabilities and limitations in generating and recognizing photorealistic images.

2. Data Source:

Primary Data Source: CIFAKE Dataset https://www.kaggle.com/datasets/birdy654/cifake-real-and-ai-generated-synthetic-images/data .

The primary data source is the CIFAKE dataset, which consists of:

1. Real Images: 60,000 images sourced from the CIFAR-10 dataset, covering ten categories such as airplanes, cats, and cars. These images are provided in JPG format with corresponding class labels.
2. AI-Generated Images: 60,000 synthetic images created using the Stable Diffusion v1-4 model, mimicking the same ten categories as the real images. These images are also in JPG format.

Data Details:

- Number of Observations: 120,000 images (60,000 real + 60,000 synthetic).

- Features: Images stored in JPG format, each associated with a categorical label indicating its class.
- Data Types: Numerical pixel values (image data) and categorical labels (classification target).

Potential Biases:

- Source Bias: CIFAR-10 images come from specific datasets, potentially underrepresenting certain object variations or backgrounds found in real-world scenarios.
- Synthetic Image Generation: The Stable Diffusion model generates images based on learned patterns, which may result in artifacts or stylistic differences compared to real-world images, making it easier for models to distinguish real from synthetic images.

3. Objectives:

The objectives of this study are to explore the capabilities and limitations of machine learning models in distinguishing real and AI-generated images. This includes measuring the model's accuracy, identifying key visual features used for classification, and understanding common mistakes in the classification process. Additionally, the study aims to evaluate the impact of preprocessing techniques on the model's performance and evaluate it by comparing it with human performance in identifying real and synthetic images. Insights from these objectives will provide a comprehensive understanding of AI's current ability to handle this task.

1. How accurately can a classification model distinguish between real and AI-generated images in this dataset without predefined categories?
2. What elements in the image help AI determine whether it is real or AI-generated? (For example: colors, edges, shadows?
3. How do image characteristics preprocessing steps (resizing or adjusting lighting) affect AI’s ability?
4. What are the common mistakes AI makes when classifying images?

5. How effectively do humans distinguish between real and AI-generated images compared to the classification model?

4. Data collection method:

For this project, we collected a dataset containing real images (REAL) and AI-generated images (FAKE), The dataset was organized into training and testing sets, with each category stored separately. We obtained images from public image datasets such as Kaggle. For real images, we made sure to obtain a diverse set of real images to help the model generalize well and not rely on a specific pattern to classify images. We also ensured the dataset included a variety of lighting conditions, backgrounds, colors, and categories to avoid biases in classification. For AI images we made sure to include various styles and qualities of AI-generated images to prevent the model from learning only one type of synthetic pattern since AI-generated images may contain subtle differences from real images (such as texture, edges, or lighting inconsistencies), which the model may learn to detect.

What will we do with this data?

1. Evaluating Classification Model Accuracy

Approach:

- Develop a classification model (such as CNN or Vision Transformer) to distinguish between real and AI-generated images.
- Split the dataset into training (80%) and testing (20%) sets.
- Evaluate the model's performance using accuracy, F1-score, precision, and recall.

2. Identifying Key Elements in AI Classification

Approach:

- Use Saliency Maps or Grad-CAM to visualize which parts of the image influence the model's decision.
- Analyze the impact of image characteristics such as colors, edges, shadows, textures, and noise.
- Compare feature importance using different models.

3. Effect of Preprocessing on AI's Ability

Approach:

Apply various image preprocessing techniques such as:

- Resizing
- Brightness adjustment
- Denoising

Compare model performance before and after preprocessing using the same evaluation metrics.

4. Common AI Mistakes in Classification

Approach:

- Analyze misclassified samples (False Positives and False Negatives).
- Identify recurring patterns in errors (such as does the model focus on incorrect details?).
- Compare different models (ResNet, EfficientNet, or Transformer-based models) to observe if the same mistakes persist.

5. Comparing Human vs. AI Classification

Approach:

- Conduct a survey where participants are asked to classify images as either real or AI-generated.
- Record human classification accuracy and compare it with the AI model.
- Perform a statistical analysis (such as T-test or Chi-square) to determine if there is a significant difference between human and AI performance.

5. Challenges:

Challenges Faced During Data Collection:

1. Nature of Unstructured Data:

The data used in this project, including jpg images, is considered unstructured data. Unlike data organized in tables, images require additional steps to process them and extract useful information from them. Such as: color distribution, edges, and lighting patterns, which increases the complexity of the analysis process.

2. Large Dataset Size:

The data used in this project is unstructured data, JPG images. Unlike data organized in tables, images require additional steps to process them and extract useful information from them. Such as: color distribution, edges, and lighting patterns, which increases the complexity of the analysis process.

3. Representation Bias:

The images contain only specific groups such as (airplanes, cars, cats) which limits the model's ability to identify images in the real world. This lack of diversity may make the model unable to generalize or provide accurate performance when dealing with new categories that are not present in the data.

4. Variability in Image Quality:

The quality of real images is clearly different from images manufactured by artificial intelligence. Real images usually have more detailed features and higher resolution than those generated by artificial intelligence. This difference in quality may cause the model to focus more on quality differences rather than the basic characteristics that distinguish actual images from fake ones.

5. Limited Generation Tools:

The Stable Diffusion model was used to create the image generated by artificial intelligence. This may cause the model to be biased towards identifying patterns for this tool due to its reliance on a single tool. Therefore, when the model is used on images produced by other AI tools, such as DALL-E or GANs, its performance can be affected.

Recommendations to Address Challenges:

1. Nature of Unstructured Data:

- Use image processing packages, such as OpenCV or PIL, to extract useful information from images.

- Resize or compress images while retaining important information needed for analysis to ensure efficient processing.

2. Large Dataset Size:

- Start with smaller, randomly selected parts of the data set for initial testing and model evaluation.

- Use cloud computing services like AWS or Google Colab to access more processing capacity, which can reduce processing time.

3. Repre Train the model on more images from different and new categories.

- To make the dataset more diverse for real-world situations,

- To avoid biasing the model toward certain groups, make sure the data set is balanced, with an equal number of images in each category.

4. Variability in Image Quality:

- Use different processing methods such as lighting normalization and noise reduction to reduce the quality discrepancy between generated images and real images.

- To improve the generalization of the model across many datasets, train it using images of varying quality levels.

5. Limited Generation Tools:

- To expand the diversity of the dataset, combine AI-generated images from additional tools, such as DALL-E and GANs.

- To evaluate the model's resilience and generalizability, test it frequently using pictures from various sources.

Challenges Faced During Data EDA & Preprocessing:

During the Exploratory Data Analysis (EDA) and Data Processing phases, we encountered several challenges that required careful consideration and appropriate solutions. Below are the key challenges we faced and the strategies we used to address them:

1. Identifying and Handling Outliers:

Challenge: Some images had significantly low or high values in features like edge density and contrast, indicating potential outliers. However, it was unclear whether these were naturally occurring variations or issues in the data.

Solution: We used the Z-score method to standardize contrast and edge density values and identify images that deviated significantly from the mean. After visually inspecting these images, we decided to retain them since they represented natural variations rather than errors.

2. Dealing with Color Distribution Variability:

Challenge: The color distribution of real and AI-generated images varied widely, but no significant differences were observed between the two categories. This made it unclear whether color features could effectively contribute to classification.

Solution: We plotted pixel intensity histograms for RGB channels separately for real and fake images. The analysis revealed that while both categories exhibit diverse color distributions, there were no strong distinguishing patterns. Given this, color features may not be a primary classification factor. However, if further analysis shows subtle differences, we may consider color normalization or contrast adjustments to refine the dataset.

3. Dataset Size and Computational Limitations:

Challenge: The dataset contained a large number of images, making some computations slow and resource-intensive.
Solution: Instead of analyzing the entire dataset at once, we sampled subsets (e.g., 30 images for visualizations) to speed up analysis while maintaining accuracy. Additionally, we performed histogram-based analysis separately for REAL and FAKE images to identify patterns more efficiently.

4. Deciding Whether to Remove Outliers:

Challenge: After detecting outliers, a major decision was whether to remove them or retain them for model training.
Solution: We visually inspected the outliers and found that most represented low-contrast natural scenes (e.g., sky, water, grass) rather than data errors. As a result, we decided to keep these images in the dataset and only consider removing them if they negatively impact model accuracy later.

Conclusion: By addressing these challenges methodically, we ensured a more robust dataset for training our classification model. The insights gained from EDA and data processing will help optimize the model's learning process and improve its ability to distinguish between real and AI-generated images.

6. Observations and insights gained for our research questions:

1. How accurately can a classification model distinguish between real and Al-generated images in this dataset without predefined categories?

Our analysis shows that the classification model was able to distinguish between real and AI-generated images without needing to know the specific object category (e.g., whether the image was of an animal, car, etc.). The model achieved an accuracy of **98.41%** using the CIFAKE dataset, demonstrating its ability to differentiate between REAL and FAKE images based solely on visual characteristics rather than object type.

This suggests that the model effectively learns to recognize differences in textures, edges, color distributions**,** and artifacts that separate AI-generated images from real ones, regardless of the image content. The high accuracy highlights the model's robustness in identifying AI-generated images even in a diverse dataset containing multiple object categories.

2. What elements in the image help AI determine whether it is real or AI-generated? (For example: colors, edges, shadows?

Through Exploratory Data Analysis (EDA)**,** we identified key visual characteristics that the AI model relies on to differentiate between real and AI-generated images. These features significantly influence classification accuracy:

Feature	Real Images	AI Images	Effect on Classification
Edges	Natural, noisy	Smooth/sharp	High impact
Texture	Varied	Flat/repetitive	Important
Artifacts	Rare	Often distorted	Distinctive
Color	Varied, not key	Also varied	Not decisive alone

EDA and model performance analysis revealed that classification is primarily based on **edge sharpness, contrast, fine details, and artifacts**. Our analysis of color distribution showed no clear differences between REAL and FAKE images, suggesting that **color alone is not a strong distinguishing factor**.

However, when images were converted to **grayscale**, model accuracy dropped significantly. This decline does not indicate that color was the primary factor but rather that **grayscale conversion affected not only color but also contrast and fine details**, which are crucial for classification.

Conclusion Color is **not the main feature the model relies on**, but it may serve as a **supporting factor** when combined with other features like **contrast and edges**. The decrease in accuracy after grayscale conversion suggests that the transformation altered how the model perceives contrast, making classification more challenging, but it was not the sole reason for performance degradation. Understanding these key elements helps refine AI training methods to improve classification accuracy and robustness.

3. How do image characteristics preprocessing steps (blurring or adjusting brightness) affect AI's ability?

Our analysis demonstrates that image preprocessing techniques significantly impact the model's ability to accurately classify real and AI-generated images. While the model initially achieved **98.41% accuracy** on untransformed images, various transformations caused a decline in performance, particularly for **REAL** images.

Transformation	Effect on Real	Effect on Fake	Notes
Flipping	High misclassification	Minimal effect	Disrupts spatial patterns
Brightness	More REAL → FAKE errors	Easier to detect FAKE	Alters cues
Rotation	Improves some, introduces confusion	Reduces confidence	Orientation shift
Blurring	Most harmful	Heavy misclassification	Destroys fine details
Grayscale	Moderate impact	Less than blur	Affects color + contrast

Overall Impact on AI's Classification Ability:

- Most transformations caused the model to overclassify images as FAKE, especially REAL images.
- Feature shifts due to preprocessing led to increased misclassification, highlighting the need for careful selection of transformations.
- Blur and grayscale transformations had the most detrimental effects**,** reducing the model's ability to accurately distinguish between real and AI-generated images.

Image preprocessing techniques can significantly alter a model's perception of image features, sometimes leading to unexpected biases and misclassifications. Our findings emphasize the importance of carefully selecting preprocessing methods to ensure that essential image characteristics are preserved, preventing performance degradation in AI classification models. Future research should focus on identifying optimal preprocessing strategies that improve robustness while maintaining classification accuracy across diverse transformations.

4. What are the common mistakes AI makes when classifying images?

Based on our analysis, the AI model demonstrates several recurring misclassification patterns when distinguishing between real and AI-generated images. These errors are influenced by image transformations, feature extraction limitations, and biases in training. The most notable mistakes include:

1. **Overclassification as Fake**
   - The model tends to classify real images as **FAKE** more frequently, especially after applying certain transformations such as blurring and grayscale conversion. This suggests that essential features distinguishing real images are lost or altered.
2. **Misclassification Due to Rotation**

- Rotation introduces ambiguity, leading the model to incorrectly label real images as fake. This implies that the model struggles with recognizing objects from different orientations, possibly due to insufficient rotational invariance in training data.

3. **Blurring and Detail Loss**
   - The **blur transformation** primarily affected real images, causing them to be **misclassified as FAKE** due to the loss of fine details. However, for **AI-generated images**, the effect was less severe, as they were already being classified as FAKE. This indicates that fine details play a crucial role in distinguishing real images, and removing them leads to confusion in classification.
4. **Grayscale Conversion Effects**
   - Converting images to grayscale increases the likelihood of real images being misclassified as fake. This suggests that **color information** is a key factor in determining authenticity, and removing it disrupts the model's ability to differentiate real from AI-generated images.
5. **Bias Toward Fake Images**
   - Most transformations shift the model's decision boundary, making it lean toward classifying images as **FAKE** rather than **REAL**. This highlights the need for improved preprocessing strategies to ensure the model maintains balanced classification performance.

These common misclassifications indicate that the model is highly sensitive to transformations that alter image features, particularly **edges, color details, textures, and lighting patterns**. Our findings emphasize the importance of designing robust AI models that can generalize well across real-world variations without being misled by preprocessing changes. Future improvements should focus on enhancing the model's ability to retain key features under different transformations while minimizing bias toward one class.

5. How effectively do humans distinguish between real and AI-generated images compared to the classification model?

Two surveys were conducted to evaluate human performance in distinguishing real from AI-generated images. The **first survey**, completed by 57 participants, used the **same CIFAKE dataset** that was utilized for model training and evaluation. The average human accuracy in this survey was approximately **56.14%**, indicating the difficulty humans face in differentiating real from synthetic images. However, due to the relatively small size and limited clarity of the CIFAKE images, it was scientifically necessary to conduct a **second survey** using a new set of **manually selected, clearer images** to provide a fairer evaluation environment. In the **second survey**, involving **36 participants**, the average human accuracy was approximately **47%**. In both evaluations, the AI model (achieving **98.41% accuracy**) significantly outperformed human participants, demonstrating its superior ability to detect subtle visual differences.

7. Summary of New Insights and Hypotheses:

1. Summary of Insights Gained

Based on the **Exploratory Data Analysis (EDA)** conducted in the project, several key insights have emerged regarding the differences between real and AI-generated images.

A. Edge Sharpness and Texture Details

- **Observation**: AI-generated images often exhibit either overly sharp or unnaturally smooth edges, while real images have more natural, irregular contours.
- **Hypothesis**: The model may rely heavily on edge clarity and texture inconsistencies to differentiate between real and AI-generated images.

B. Color Distribution Analysis

- **Observation:** No significant differences in color distribution were found between real and AI-generated images.

- **Hypothesis:** Color alone is not a strong classification feature, but it may serve as a supporting factor when combined with other features like contrast and edges.

C. Impact of Preprocessing on Model Performance

- **Observation:** Blurring significantly reduced classification accuracy, suggesting that the model depends on fine details and sharp edges for decision-making.
- **Hypothesis:** Any preprocessing step that reduces texture clarity or edge sharpness negatively impacts classification performance.

2. New Hypotheses Generated

Based on these insights, the following hypotheses were formulated for further validation:

1. AI-generated images may contain artificially smooth textures or unnatural patterns compared to real images.
2. Color distribution is not a primary distinguishing factor, but it may play a supporting role when combined with other features.
3. Neural networks rely on unnatural edge smoothness and texture repetition to classify AI-generated images.
4. Preprocessing techniques significantly impact classification: Blur reduces accuracy, indicating reliance on edge clarity.
5. AI models may struggle more with certain textures, particularly those with overly smooth surfaces.

8. Modelling Task:

In this project, three models were tested to classify images as "REAL" or "FAKE."

- **Model 1** was a pre-trained Vision Transformer (ViT) used without extra training. It achieved **98% accuracy**, but it did not provide full evaluation results like precision, recall, or training curves. It was fast for testing but not ideal for deeper analysis.

- **Model 2** was a small CNN built and trained from scratch on 32×32 images. It reached **92.69% validation accuracy** and provided balanced precision, recall, and F1-score. It was lightweight and worked well for simple or limited-resource environments.
- **Model 3** was a deeper CNN trained on 128×128 images, with dropout layers to avoid overfitting. It achieved **93.52% validation accuracy** and **95.47% training accuracy**, and showed strong, stable performance during training.

Conclusion:

Model	Type	Accuracy	Strength	Weakness
Model 1	ViT (Pre-trained)	98%	Fast, accurate	No training curves or detailed metrics
Model 2	Custom CNN	92.69%	Lightweight, explainable	Lower accuracy
Model 3	Deeper CNN	93.52% (Val), 95.47% (Train)	Best overall performance	Heavier model

9. References:

- Krizhevsky, A., & Hinton, G. (2009). Learning multiple layers of features from tiny images.
- Bird, J.J. and Lotfi, A., 2024. CIFAKE: Image Classification and Explainable Identification of AI-Generated Synthetic Images. IEEE Access.
- Real images are from Krizhevsky & Hinton (2009), fake images are from Bird & Lotfi (2024).