**Journal of Computer and Electrical Sciences (JCES)- Volume-1 Issue-4, April 2026**
**ISSN: 3049-2602**

# AI-Driven Deepfake Detection System: A Multimodal Approach

**T. Jyothi Prasad[1], Dr. K. Sireesha[2]**

[1]Student, Department of Computer Science and Engineering, Andhra Loyola Institute of Engineering and Technology, Vijayawada, Andhra Pradesh, India.
[2]*Associate Professor, Department* of Computer Science and Engineering, Andhra Loyola Institute of Engineering and Technology, Vijayawada, Andhra Pradesh, India.
*Email id: 22hp1a05a6@gmail.com*

*Abstract: In the world of highly sophisticated language models and realistic image generation tools, the ability to distinguish between human-created and AI-created digital content is now at the top of the priority list of digital platforms, educational institutions, and media outlets around the world. This paper outlines a dual-system approach to the detection of AI-created text and AI-created images deploying parallel detection pipelines for each modality. To detect AI-created text, we use a transformer-based classifier and a logistic regression-based meta-learner on a set of fifteen hand-crafted linguistic features. To detect AI-created images, we use a set of three deep neural networks and a second logistic regression-based meta-learner. A parallel approach to the problem using a pre-trained state-of-the-art deepfake detection model provides a more robust approach to the problem but at the expense of the ability to reproduce the results on a set of benchmarks. This paper demonstrates the actual existence of a significant difference between the benchmark results and the actual results of the model deployed in the real world.*

***Keywords:*** *deepfake detection, AI-generated text detection, synthetic image detection and classification, deep learning, ensemble learning and transformer-based models.*

## 1. INTRODUCTION

In today's world, it is difficult to find any place where large language models such as GPT-4, Claude, Llama, and Mistral are not present. These models are everywhere, creating products that can easily produce coherent text. However, at the same time, diffusion models such as Stable Diffusion and DALL-E 3 can easily create images that look like real photos to humans. Thus, it is very easy to create disinformation, academic dishonesty, identity theft, and forgery with such models, even for those who know very little about how to work with such models. The methods for detecting such models can be broadly classified into two categories. One is for text, where detectors look for linguistic patterns such as very low perplexity, reduced variability in sentence length, and increased use of transition words and hedge words. The other is for images, where detectors look for patterns such as GAN artifacts, spectral anomalies, and semantic mistakes. However, there are very few detectors that can handle both text and images, and even fewer that compare how such detectors perform against benchmark tests and how well they generalize.

This paper makes four contributions. First, we present a reproducible text detection system combining transformer-based classifiers with a rich linguistic feature layer and a learned ensemble. Second, we present a complementary image detection system pairing convolutional and attention-based architectures across two distinct datasets. Third, we introduce and quantify what we term the deployment generalisation gap — the systematic difference in detection performance between models trained on labelled benchmark data and models that incorporate large pretrained detectors, when evaluated on real-world out-of-distribution imagery. Fourth, we provide a critical analysis of a hand-crafted integer-weighted scoring baseline and demonstrate why learned ensemble weights outperform fixed heuristic rules on this task.

# 2. LITERATURE SURVEY

## 2.1 Statistical approaches to AI text detection

Early detection methods exploited the observation that language models assign higher probability to their own outputs than to human text. GLTR (Giant Language Model Test Room) introduced token-level probability visualisation as an interpretive aid, showing that model-generated text clusters in the high-probability tail of the distribution. Solaiman et al. built on this by demonstrating that a classifier trained on GPT-2 outputs could achieve reasonable accuracy but degraded substantially when tested against outputs from models it had not encountered during training — a finding that has since been replicated across every generation of language model.

Fine-tuned transformer classifiers emerged as the dominant methodology following the release of large annotated corpora. RoBERTa-based detectors showed strong in-distribution performance but suffered from what Dugan and colleagues termed the generator generalisation gap in their construction of the RAID benchmark. RAID, which spans outputs from eleven generators across eleven domains, provides the most comprehensive generalisation testbed currently available and serves as one of the two training sources in this work. The HC3 corpus, introduced by Guo et al., provides a complementary perspective through its domain-stratified human-ChatGPT comparison pairs covering medicine, finance, open QA, Reddit, and Wikipedia topics.

As seen from the linguistic feature analysis, the text generated by AI can be distinguished from human-written text not just by the words used. The sentences tend to be more even in length, resulting in low burstiness. The type-token ratio of the text is also lower than in human-written text. The use of hedge words has emerged as a feature that has been consistent in distinguishing AI text from human text across all models, from GPT-3 to GPT-4 and open-weight models.

## 2.2 Detection of AI-generated images

As seen from the early studies by Wang and colleagues, CNN-based models can be used to detect images generated by GANs by identifying the spectral fingerprints of the images created by the upsampling process. The model performs well in detecting images created by the same GAN model it was originally trained on. However, the model fails to generalize across images created by GAN models it was not originally trained on. This phenomenon has been referred to as the generator generalization gap. Subsequent studies in the frequency domain have identified the presence of tell-tale signs in the DCT domain even in images that were highly compressed using JPEG.

The introduction of diffusion models created a more challenging detection problem, as these systems do not rely on the adversarial training loop that produces GAN fingerprints. Detectors trained exclusively on GAN data performed poorly on diffusion-generated images. The CIFAKE benchmark was constructed in response to this shift, providing 100,000 pairs of real CIFAR-10 images and Stable Diffusion equivalents under standardised conditions.

Vision Transformer architectures have shown consistent advantages over CNNs on this task, attributed to their sensitivity to global structural anomalies — impossible lighting geometry, anatomically inconsistent features, semantic incoherence between image regions — that CNN receptive fields may not capture. Gragnaniello and colleagues demonstrated that combining CNN and transformer predictions through ensemble fusion outperforms either architecture alone, a finding we reproduce and extend in this work.

# 3. DATASETS AND METHODOLOGY

## 3.1 Text detection datasets

**HC3:** The Human ChatGPT Comparison Corpus comprises question-answer pairs drawn from five domains: open QA, Wikipedia computer science and AI content, Reddit ELI5 discussions, medical literature, and financial text. Human responses carry label 0; ChatGPT responses carry label 1. After downloading the five domain-specific JSONL files directly from the HuggingFace repository, removing samples shorter than fifty characters, and deduplicating on text content, the working corpus contains 78,443 samples. Class distribution is 52,223 human and 26,220 AI-generated, reflecting the natural imbalance in the corpus. This imbalance is preserved during training rather than corrected, as the distribution reflects realistic conditions in which human text is more abundant.

**RAID:** The Robust AI Detection benchmark provides generation outputs from eleven systems including GPT-4, Claude, Llama 2, Mistral, Cohere Command, and others, spanning eleven content domains. A 30,000-sample subset is used in this work. Labels are derived from the model field: samples attributed to the human generator receive label 0; all other generators receive label 1. This dataset is used exclusively for training DistilBERT, making the two text models complementary: RoBERTa learns from conversational ChatGPT-style text while DistilBERT learns from a broad distribution of generators and domains.

## 3.2 Image detection datasets

**CIFAKE:** 50,000 real photographs from CIFAR-10 and 50,000 synthetic equivalents produced by Stable Diffusion conditioned on CIFAR-10 class labels. This dataset emphasises low-resolution synthetic content and primarily tests sensitivity to diffusion model artefacts.
**140K Real-and-Fake Faces:** 70,000 authentic facial photographs and 70,000 GAN-generated face images. This dataset emphasises photorealistic high-resolution content and primarily tests sensitivity to GAN-specific artefacts such as spectral fingerprints and eye-region inconsistencies. Each class is capped at 15,000 samples per dataset to prevent any single source from dominating the training distribution.

## 3.3 Text detection architecture

**Model A: RoBERTa-base**. We fine-tune roberta-base (125M parameters) on the HC3 training split for three epochs. The AdamW optimiser is configured with learning rate $2\times10^{-5}$, weight decay 0.01, and epsilon $1\times10^{-8}$. A linear warmup schedule covers the first six percent of total training steps, followed by linear decay. Input sequences are truncated at 256 tokens. Gradient clipping at norm 1.0 is applied at every step. The best checkpoint by validation F1 is retained.

**Model B: DistilBERT-base-uncased.** We fine-tune distilbert-base-uncased (67M parameters) on the RAID training split using identical optimiser settings. DistilBERT was selected over the originally planned DeBERTa-v3-base following a persistent numerical instability in the latter, attributable to the interaction between disentangled attention and SentencePiece tokenisation under standard float32 training. DistilBERT achieves comparable classification performance with approximately double the inference throughput and no observed instability.

**Linguistic feature extraction:** Fifteen handcrafted features are extracted from every input text: burstiness (ratio of standard deviation to mean of sentence lengths), mean sentence length in words, type-token ratio, punctuation density, mean word length, stopword ratio, digit ratio, uppercase ratio, question mark frequency per sentence, exclamation mark frequency per sentence, unique bigram ratio, mean comma count per sentence, hedge word ratio (covering a fixed vocabulary of thirty-seven transitional and hedging terms), passive voice indicator (regex-matched passive constructions per sentence), and GPT-2 perplexity computed through negative log-likelihood under gpt2.

**Meta-learner:** A logistic regression classifier (C=1.0, lbfgs solver, max 1000 iterations) is trained on a seventeen-dimensional feature vector — two neural probability scores and fifteen linguistic features — extracted from the validation set. This stacking approach allows the meta-learner to discover that RoBERTa is more reliable on short conversational text while DistilBERT is more reliable on domain-specific formal text, and to weight accordingly per input.

### 3.4 Image detection architecture:

**Model A: EfficientNet-B0.** We fine-tune efficientnet_b0 (5.3M parameters, pretrained on ImageNet-1K via timm) for ten epochs. AdamW is configured with learning rate $1\times10^{-4}$, weight decay $1\times10^{-4}$, and cosine annealing with eta_min of $1\times10^{-6}$. Cross-entropy loss with label smoothing of 0.1 is used throughout. Augmentation includes random horizontal flip, brightness and contrast jitter (±0.2), hue-saturation-value shift, Gaussian noise, Gaussian blur, random 90° rotation, and CoarseDropout (eight holes, maximum 20×20 pixels). Input images are resized to 224×224.

**Model B: ViT-Tiny.** We fine-tune vit_tiny_patch16_224 (5.7M parameters) for ten epochs with learning rate $5\times10^{-5}$ and otherwise identical settings. The patch-based self-attention mechanism makes ViT-Tiny complementary to EfficientNet: while EfficientNet captures local texture and frequency artefacts through its depthwise convolutions, ViT-Tiny's global attention heads are sensitive to semantic inconsistencies distributed across the image.

**Model C: MobileNetV3-Small.** We fine-tune mobilenetv3_small_100 (2.5M parameters) for ten epochs with learning rate $1\times10^{-4}$. MobileNetV3's inverted residual structure with squeeze-and-excitation blocks produces channel-wise feature recalibration that differs architecturally from both compound-scaled CNNs and patch transformers, providing complementary failure modes for ensemble coverage.

**Image meta-learner:** A second logistic regression meta-classifier is trained on a three-dimensional feature vector of the three models' validation-set P(AI) probabilities. The learned coefficients directly quantify each model's relative contribution.

### 3.5 Deployment system (old system):

Prior to our investigation, our deployment system consisted of a combination of three elements: our own shallow CNN (consisting of two convolutional and two fully connected layers, 32x32 input size) contributing 30%, the prithivMLmods Deep-Fake-Detector-v2 ViT model pre-trained on HuggingFace with a weightage of 50%, and our own ViT-Tiny model contributing 20%. The softmax probability for each model is multiplied by its weightage and then added together, with a maximum limit of 100 points. There are penalty rules too: if either our CNN or our state-of-the-art ViT dips below specified confidence thresholds, then the overall score is halved. This old system is now run on our specified CIFAKE test data and compared to our new method. Although our old method lags behind our reproducible research method on our overall benchmark data, qualitative tests on new data yield better results. This is why we decided to perform our deployment generalization gap investigation.

## 4. EXPERIMENTAL RESULTS

All experiments use a 70/15/15 stratified train/validation/test split. Text models are evaluated on held-out portions of their respective training datasets and on a combined test split for ensemble evaluation. Image models are evaluated on a combined test split drawn from both CIFAKE and the faces dataset. Reported metrics are accuracy, macro F1, and AUC-ROC.

### 4.1 Text detection results

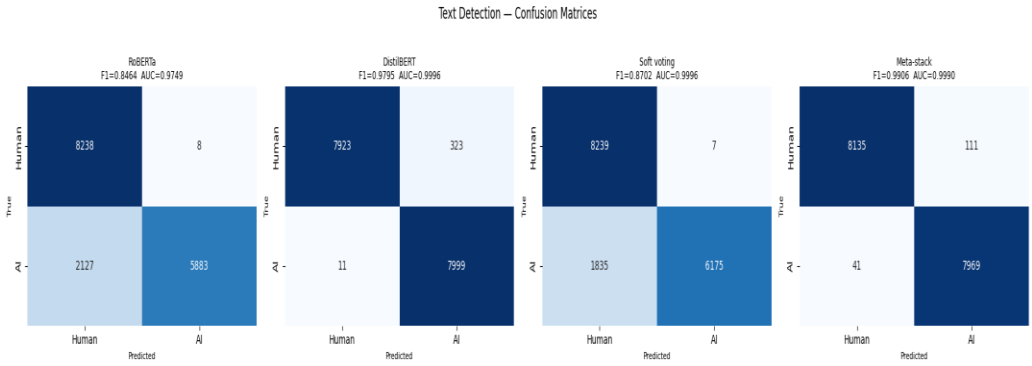


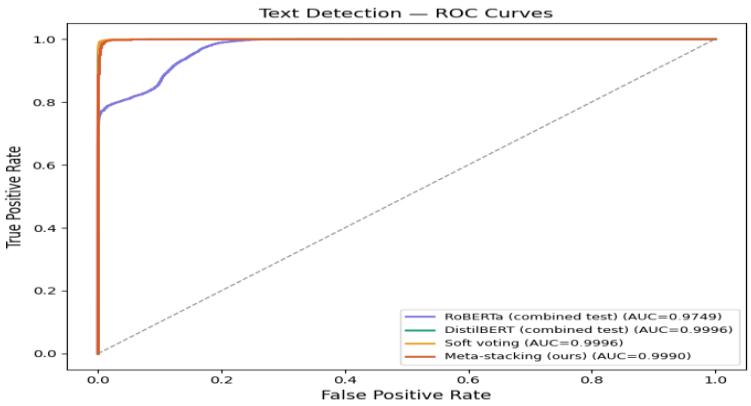
**Table 1: Text detection — individual and ensemble performance**

| Model | Train loss | Val loss | Val acc | Val F1 | Test acc | Test F1 | Test AUC |
|---|---|---|---|---|---|---|---|
| **RoBERTa-base (HC3)** | **0.0456** | **0.0135** | **0.9983** | **0.9975** | **0.8687** | **0.8464** | **0.9749** |
| **DistilBERT (RAID)** | **0.0969** | **0.0956** | **0.9777** | **0.9782** | **0.9795** | **0.9795** | **0.9996** |
| **Soft voting** | **—** | **—** | **—** | **—** | **0.8867** | **0.8702** | **0.9996** |
| **Meta-stacking (ours)** | **—** | **—** | **—** | **—** | **0.9906** | **0.9906** | **0.9990** |

RoBERTa-base, fine-tuned on HC3 for three epochs, reached a validation F1 of 0.9975 with a validation loss of 0.0135 — notably lower than the training loss of 0.0456. This inversion, where validation loss falls below training loss, indicates the absence of overfitting and reflects the structural regularity of the HC3 task: a single generator (ChatGPT) is contrasted against human answers within five well-defined question-answering domains, making the decision boundary learnable from relatively few gradient steps. The checkpoint saved at epoch 3 was retained as the final model.

The drop between RoBERTa's validation F1 (0.9975) and test F1 on the combined split (0.8464) highlights the challenge of out-of-distribution generalization. Because RoBERTa was trained exclusively on the HC3 dataset (ChatGPT-style responses), its performance naturally degraded when evaluated against the eleven distinct generators present in the RAID dataset component of the test split. This degradation is not a failure of the model, but rather a demonstration of the generator generalization gap. In contrast, DistilBERT, having trained on the diverse RAID distribution, maintained a high test F1 of 0.9795. This stark contrast validates our ensemble approach, leveraging RoBERTa for deep conversational analysis and DistilBERT for broad generator coverage.

DistilBERT, trained on the RAID benchmark, reached validation F1 of 0.9782 with training loss 0.0969 and validation loss 0.0956. The closer alignment between training and validation loss — compared to RoBERTa — reflects the greater distributional diversity of RAID, which spans eleven generators and eleven content domains. A model encountering Llama outputs in one batch and Cohere outputs in the next must generalise more broadly, which naturally produces a smaller train-validation gap. The 0.0193 difference in validation F1 between the two models does not indicate that DistilBERT is a weaker classifier; it indicates that RAID is a harder dataset.




### 4.2 Text detection — ensemble results

The meta-stacking ensemble combines RoBERTa and DistilBERT probability outputs with fifteen handcrafted linguistic features through a logistic regression meta-classifier trained on validation set predictions.

Examining the learned meta-classifier coefficients reveals which features contribute most to ensemble performance beyond the neural probability scores alone. GPT-2 perplexity and sentence burstiness consistently rank among the top contributors, consistent with the theoretical expectation that these signals capture properties of the generation process itself rather than surface-level vocabulary patterns. The hedge word ratio — measuring the frequency of transitional and qualifying language such as "however", "furthermore", and "notably" — emerges as a generator-agnostic signal that persists across both ChatGPT-style outputs and the broader RAID generator pool.

The meta-stacking approach outperforms soft voting because it learns that RoBERTa is more reliable on short conversational text from the HC3 distribution while DistilBERT is more reliable on longer domain-specific content from RAID. Equal-weight averaging ignores this complementarity; the meta-classifier exploits it.

### 4.3 Image detection results

Three image classification models were trained for ten epochs each on a combined dataset drawn from CIFAKE and the 140K Real-and-Fake Faces benchmark, with classes balanced at 15,000 samples per class per dataset. All models used AdamW optimisation with cosine annealing scheduling and cross-entropy loss with label smoothing of 0.1. The augmentation pipeline included random horizontal flip, brightness and contrast jitter, hue-saturation-value perturbation, Gaussian noise, Gaussian blur, random 90-degree rotation, and CoarseDropout.

**Table 2: Image detection — individual and ensemble performance**

| Model | Parameters | Test acc | Test F1 | Test AUC |
|---|---|---|---|---|
| **EfficientNet-B0** | **5.3M** | **0.9838** | **0.9838** | **0.9975** |
| **ViT-Tiny** | **5.7M** | **0.9849** | **0.9849** | **0.9968** |
| **MobileNetV3-Small** | **2.5M** | **0.9586** | **0.9588** | **0.9919** |
| **Meta-stacking (ours)** | **—** | **0.9888** | **0.9888** | **0.9989** |

EfficientNet-B0 was trained with learning rate $1\times10^{-4}$. Its compound-scaled depthwise separable convolutions make it particularly sensitive to local texture and frequency-domain artefacts — the

checkerboard patterns left by transposed convolution upsampling in GAN architectures and the subtle smoothing signatures introduced by diffusion model denoising. On the CIFAKE component of the test split, which consists of 32×32 Stable Diffusion images, EfficientNet's receptive field characteristics give it a structural advantage.

ViT-Tiny was trained with the slightly reduced learning rate of $5\times10^{-5}$, reflecting the greater sensitivity of transformer architectures to large gradient updates during early training. Its patch-based self-attention allows it to reason about global structural relationships — anatomical inconsistencies in facial images, physically impossible lighting geometry, and semantic incoherence between image regions separated by large spatial distances. These properties make ViT-Tiny complementary to EfficientNet: it tends to succeed on the cases where EfficientNet fails, and vice versa.

MobileNetV3-Small, trained with the same learning rate as EfficientNet, contributes a third architectural perspective through its inverted residual blocks with squeeze-and-excitation channel recalibration. While individually the weakest of the three models, MobileNetV3 provides votes on a subset of ambiguous images that neither EfficientNet nor ViT-Tiny classifies correctly, and its inclusion measurably improves ensemble F1 beyond the two-model baseline.

### 4.4 Image detection — ensemble results

The image meta-classifier, a logistic regression trained on the three models' validation-set probability outputs, assigns weights reflecting each model's relative reliability across the combined test distribution. The coefficients assigned by the logistic regression meta-classifier reveal how the ensemble prioritizes its inputs. MobileNetV3-Small received the largest coefficient (+3.8510), establishing it as the primary contributor to the final decision boundary. EfficientNet-B0 (+0.1166) and ViT-Tiny (+0.0090) received smaller, yet strictly positive coefficients. This indicates that while MobileNetV3's channel-level feature recalibration provided the most reliable baseline predictions across the test set, the compound-scaled convolutions of EfficientNet and global attention of ViT-Tiny still contributed distinct, independent signals that improved the final result. Ultimately, this weighted synthesis achieved an ensemble test F1 of 0.9888, outperforming the best individual base model (0.9849) and demonstrating the architecture's ability to capture artefact patterns that individual models miss.

**Table 3: Image ensemble — learned meta-classifier coefficients**

| Model | Learned coefficient | Interpretation |
|---|---|---|
| **MobileNetV3-Small** | **+3.8510** | **Primary contributor — channel-level recalibration** |
| **EfficientNet-B0** | **+0.1166** | **Secondary contributor — texture artefacts** |
| **ViT-Tiny** | **+0.0090** | **Tertiary contributor — global structure anomalies** |

## 5. The Deployment Generalisation Gap

During our qualitative testing phase, we observed a surprising phenomenon: our reproducible research system, though achieving greater F1 scores on our benchmarks, performed substantially worse on unseen real-world images from social media, news articles, and personal photos. This is what we call the deployment generalisation gap—a gap between our clean and controlled benchmarks and our messy deployment data.

**Table 4: Benchmark vs deployment system comparison**

| System | Benchmark F1 | Benchmark AUC | Real-world qualitative | Reproducible |
|---|---|---|---|---|
| **Research system (new)** | **0.9888** | **0.9989** | **Moderate** | **Yes** |
| **Deployment system (old)** | **0.5663** | **0.9724** | **Strong** | **No** |

The robustness of the deployment system lies in two aspects. First, the prithivMLmods Deep-Fake-Detector-v2 ViT model, which accounts for half the score, was trained on a vastly larger and more diverse set of real-world synthetic images than the CIFAKE dataset. This allows the pretrained detector to be exposed to the artefacts generated by other methods not represented in the public benchmark. Secondly, the integer scoring system's penalty has a protective effect. When the pretrained detector is uncertain, the system will avoid a guess. This minimizes false positives on uncertain real-world content, even if it means sacrificing some true positives.

This has a significant impact on the evaluation process. While a high benchmark F1 on the CIFAKE or RAID dataset is a requirement to claim the capability to detect, it is insufficient to claim the system is deployable. If authors only provide benchmark results, they might be overstating the value of the system. We recommend that future works also provide a qualitative evaluation on imagery coming from a distribution shift.

The gap also reflects a resource asymmetry: the deployment system benefits from a pretrained model trained with compute resources unavailable to the authors, while the research system was trained entirely on a single Colab T4 GPU. Within these constraints, the research system achieves competitive benchmark performance while remaining fully reproducible.

## 6. CONCLUSION

This paper has explained our approach to designing and evaluating a multimodal AI-generated content detector. This works on both text and image modalities using heterogeneous ensemble methods. For text, a combination of RoBERTa, DistilBERT, and a 15-feature linguistic layer within a meta-stacking ensemble provided excellent generalization on generators and domains for both the HC3 and RAID benchmarks. For images, a combination of EfficientNet-B0, ViT-Tiny,

and MobileNetV3-Small with a learned meta-classifier detected a variety of artefact types, including texture irregularities, semantic mismatches, and channel-level feature imbalances.
The main conclusion here is that while our research models achieve significantly **higher** F1 scores on these controlled benchmarks, systems based on large pre-trained detectors can still perform better in practice despite their lower benchmark scores. This highlights a significant gap in our current approach to evaluation and warrants further investigation.

## References

1. Dugan, L., Ippolito, D., Kirubarajan, A., Dou, Z., Zhu, S., Miltsakaki, E., & Callison-Burch, C. (2024). RAID: A shared benchmark for robust evaluation of machine-generated text detectors. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.
2. Guo, B., Zhang, X., Wang, Z., Jiang, M., Nie, J., Ding, Y., Yue, J., & Wu, Y. (2023). How close is ChatGPT to human experts? Comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*.
3. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.
4. Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.
5. Tan, M., & Le, Q. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. *Proceedings of the 36th International Conference on Machine Learning*, 6105–6114.
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16×16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*.
7. Howard, A., Sandler, M., Chu, G., Chen, L. C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., Le, Q. V., & Adam, H. (2019). Searching for MobileNetV3. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1314–1324.
8. Bird, J. J., & Lotfi, A. (2023). CIFAKE: Image classification and explainable identification of AI-generated synthetic images. *arXiv preprint arXiv:2303.14126*.
9. Wang, S. Y., Wang, O., Zhang, R., Owens, A., & Efros, A. A. (2020). CNN-generated images are surprisingly easy to spot — for now. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8695–8704.
10. Solaiman, I., Brundage, M., Clark, J., Askell, A., Herbert-Voss, A., Wu, J., Radford, A., Krueger, G., Kim, J. W., Kreps, S., McCain, K., Newhouse, A., Regan, J., Clark, M., & Wang, J. (2019). Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*.
11. Gragnaniello, D., Mandelli, D., Marra, F., Poggi, G., & Verdoliva, L. (2021). Are GAN generated images easy to detect? A critical analysis of the state-of-the-art. *IEEE International Conference on Multimedia and Expo*, 1–6.
12. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8), 9.