Unmasking Deepfakes: Robust and Interpretable ML Approaches

Gaurav Dalvi

MSc Data Science, Kingston University London

[k2452052@kingston.ac.uk](mailto:k2452052@kingston.ac.uk)

***Abstract*—Deepfakes pose growing risks to trust, privacy, and information integrity across journalism, politics, finance, and everyday communication. In this work we present a compact and interpretable pipeline for *image-based* deepfake detection that combines lightweight handcrafted descriptors with deep convolutional embeddings (VGG16) and classical machine-learning classifiers (Support Vector Machine and Random Forest). The goal is a Type-1 MSc deliverable that is computationally efficient, reproducible, and robust enough for practical deployment without the training overhead of large end-to-end networks. Using approx- imately** 14**k face images in balanced splits (train/validation/test), we report in-distribution results and robustness under common input degradations (JPEG compression, Gaussian noise, and blur). Our Random Forest (RF) consistently achieves strong accuracy and AUC on clean images while maintaining higher F1 under perturbations than the SVM, which tends toward optimistic recall but lower precision. We analyse error modes via confusion matrices and ROC curves, discuss ethical and professional considerations for responsible deployment, and situate our design choices within the landscape of forensic cues, frequency- aware models, and cross-dataset generalisation studies. The resulting pipeline demonstrates that carefully engineered features plus classical ML remain competitive baselines and serve as transparent, resource-efficient detectors in realistic settings [**[**1**](#_bookmark6)**]– [**[**5**](#_bookmark10)**], [**[**18**](#_bookmark22)**].**

***Index Terms*—Deepfake detection, machine learning, Random Forest, SVM, VGG16 features, robustness, ROC/AUC, explain- ability.**

1. Introduction

Advances in generative modeling have made it increasingly simple to fabricate realistic human faces and expressions. Tech- niques spanning autoencoders, GANs, and diffusion models enable subtle manipulations of identity, mouth motion, and af- fect at high resolution and frame rates. While these capabilities support benign applications (e.g., film post-production), they simultaneously lower the barrier to malicious uses that can erode trust at scale. Well-documented societal risks include misinformation, harassment, reputational damage, and fraud [[7](#_bookmark12)]. For computer vision, this dual use emphasises the need for detectors that are both *accurate* and *deployable*.

A large body of work frames deepfake detection as an end- to-end supervised classification problem on frames or clips, often powered by high-capacity CNNs and transformers with hundreds of millions of parameters [[3](#_bookmark8)], [[23](#_bookmark26)]. Such approaches can reach very high in-dataset accuracy but face two practical challenges. First, training and fine-tuning these models is compute-intensive; maintaining them for continuous monitoring can be prohibitive in resource-constrained contexts. Second, generalisation across datasets or under realistic degradations (re-

compression, noise, and blur) is not guaranteed; brittle detectors may fail catastrophically outside their training conditions [[8](#_bookmark13)], [[15](#_bookmark19)].

Motivated by these constraints, we revisit a more classical design: lightweight, interpretable feature extraction coupled with strong yet efficient classifiers (SVM and RF). Rather than learning end-to-end, we (i) extract compact visual descriptors that capture colour/texture and edge/gradient cues [[1](#_bookmark6)], (ii) augment them with frozen deep embeddings from a pre-trained VGG16 backbone [[24](#_bookmark27)] to inject high-level perceptual structure, and (iii) train SVM/RF using identical splits and transparent hyperparameters. This design offers three benefits: (a) *effi- ciency*—fast training/inference on commodity hardware, (b) *interpretability*—feature importance and decision boundaries can be inspected, and (c) *robustness*—ensembles like RF often remain competitive under input noise and compression when paired with appropriate features [[5](#_bookmark10)], [[6](#_bookmark11)].

Our experiments use roughly 14*,*000 balanced face images with standard train/val/test partitions. We benchmark on clean validation/test sets and then evaluate robustness under JPEG quality 50, Gaussian noise *σ* = 0*.*02, and Gaussian blur radius

3. To ensure reproducibility and fair comparison, we fix splits and hyperparameter search protocols for both classifiers. We present ROC curves, AUCs, and confusion matrices to charac- terise performance/threshold trade-offs. Across conditions, RF yields higher accuracy and F1 than SVM, while both models keep AUCs near one on clean data, reflecting strong separability (Tables [I](#_bookmark4), [II](#_bookmark5)).

**Contributions.**

* A compact, reproducible detection pipeline that fuses hand- crafted descriptors with frozen deep embeddings and trains efficient classical classifiers;
* A controlled study of in-dataset performance and robustness under JPEG/noise/blur, supported by ROC/CM analysis;
* A discussion of ethical and professional issues (privacy, bias, transparency) to guide responsible deployment in real systems [[7](#_bookmark12)], [[25](#_bookmark28)].

1. Related Work

**Forensic/handcrafted cues.** Early detectors adapted ste- ganalysis and local descriptors to spot manipulation artefacts, leveraging co-occurrence statistics, residual patterns, and regu- larities in edges and textures [[1](#_bookmark6)], [[17](#_bookmark21)]. Handcrafted pipelines are cheap and interpretable, and they expose useful failure modes (e.g., false positives on textured backgrounds). However, they

may underfit high-quality forgeries or generalise poorly when artefacts differ from those seen during development [[18](#_bookmark22)].

**CNN-based detectors.** Mesoscopic architectures and Xcep- tion/ResNet backbones trained on FaceForensics++ popularised high-accuracy end-to-end detection [[2](#_bookmark7)], [[3](#_bookmark8)]. Frequency-aware methods emphasise spectral inconsistencies induced by up- convolutions [[5](#_bookmark10)], [[16](#_bookmark20)], while audio-visual cues (e.g., lip synchrony and affect) improve video-level reasoning [[4](#_bookmark9)], [[19](#_bookmark23)]. Capsule networks [[20](#_bookmark24)] and recurrent/temporal strategies [[21](#_bookmark25)] address viewpoint and temporal coherence. Despite impressive benchmarks, many models degrade under re-compression and cross-dataset shifts [[8](#_bookmark13)], [[15](#_bookmark19)].

**Datasets and generalisation.** FaceForensics++, Celeb-DF, DFDC, and DeeperForensics expose scale and variety, but also reveal sensitivity to data collection pipelines and post- processing [[3](#_bookmark8)], [[10](#_bookmark14)], [[11](#_bookmark15)], [[13](#_bookmark17)]. Diagnostic work shows “short- cuts” learned by detectors can be dataset-specific (camera pipeline, codec chain), leading to false confidence [[12](#_bookmark16)], [[14](#_bookmark18)], [[23](#_bookmark26)]. Robust evaluation thus requires perturbation tests and careful thresholding.

**Classical ML with deep features.** Extracting frozen CNN embeddings and training SVM/RF offers favourable cost/accuracy trade-offs, especially when feature importance and margins aid explainability [[6](#_bookmark11)], [[21](#_bookmark25)]. This hybrid view motivates our approach: emphasise simplicity, transparency, and robustness rather than maximal capacity.

1. Methodology
   1. *Data and Splits*

We use ∼14k face images (∼8.8 GB) balanced between real and fake classes. Data are indexed by CSV files (train\_index.csv, val\_index.csv, test\_index.csv) with absolute paths and binary labels to ensure exact reproducibility. Validation splits drive

hyperparameter selection and threshold analysis; test splits remain untouched until final reporting. For robustness we create three perturbed test sets: (1) JPEG quality 50, modeling re-compression common on social platforms; (2) Gaussian noise *σ* = 0*.*02, approximating sensor/quantisation noise; and

(3) Gaussian blur radius 3, simulating motion/defocus and low-bitrate effects [[5](#_bookmark10)], [[15](#_bookmark19)].

* 1. *Preprocessing*

Images are face-aligned and resized. For handcrafted features we optionally use grayscale to stabilise luminance-driven descriptors; for CNN embeddings we pass RGB images through a pre-trained VGG16, resized to match the network’s expected input [[24](#_bookmark27)]. Per-feature standardisation ensures zero mean and unit variance. We avoid heavy augmentation to keep analyses attributable to features rather than augmentation priors [[18](#_bookmark22)].

* 1. *Feature Extraction*

**Handcrafted descriptors.** We compute colour histograms (per-channel), gradient/edge statistics (e.g., Sobel/Laplacian energy and orientation histograms), and local texture cues akin to LBP. These summarise low-level artefacts introduced by

blending, resampling, or decoding inconsistencies [[1](#_bookmark6)]. Feature dimensionality remains modest for RF interpretability and SVM tractability.

**Frozen deep embeddings.** We extract penultimate-layer embeddings from VGG16 (trained on ImageNet) without fine- tuning, following the common “deep features + classical ML” recipe [[6](#_bookmark11)], [[24](#_bookmark27)]. The embeddings complement handcrafted cues with semantic structure (facial components, high-level textures).

* 1. *Classifiers and Metrics*

We train two baselines:

* **SVM** with RBF kernel. We grid-search *C* and *γ* on validation splits and calibrate decision thresholds via ROC analysis.
* **Random Forest** with hundreds of trees, tuned max depth and min split size. RF offers robustness to noisy features and exposes feature importance.

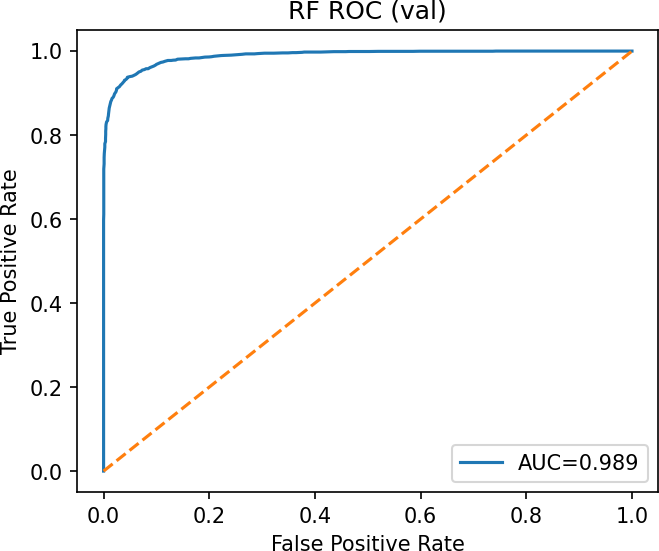
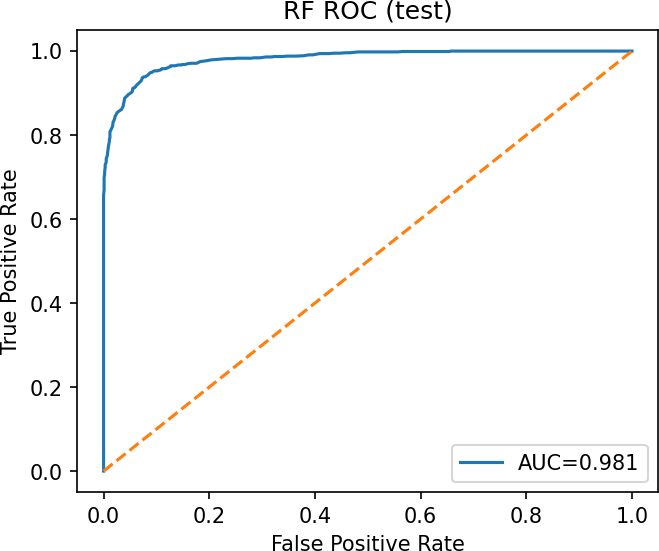
Metrics include accuracy, precision, recall, F1, ROC curves, and AUC. We present confusion matrices to visualise error asymmetries (false positives vs. false negatives) and we emphasise AUC as threshold-agnostic separability [[18](#_bookmark22)]. For perturbed sets, we report the same metrics to probe stability under realistic degradations [[5](#_bookmark10)].

1. Experiments
2. *Training Protocol*

Both models use identical splits and hyperparameter search spaces to ensure fair comparison. Feature pipelines are frozen before training; only classifier parameters vary. We employ early stopping on validation metrics where applicable and keep random seeds fixed for repeatability. Class imbalance is minimal due to balanced splits; nonetheless, we verify that accuracy trends align with precision/recall/F1.

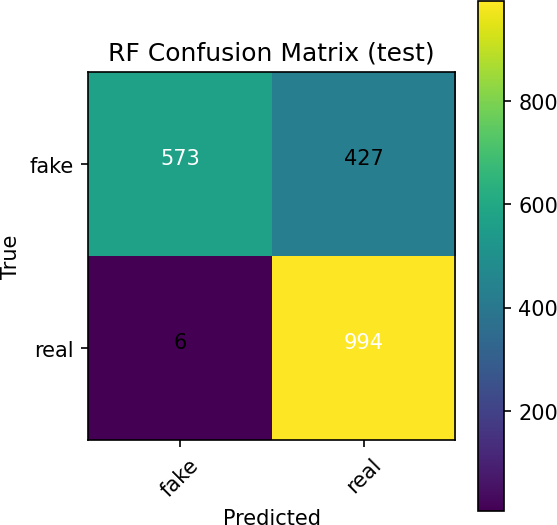
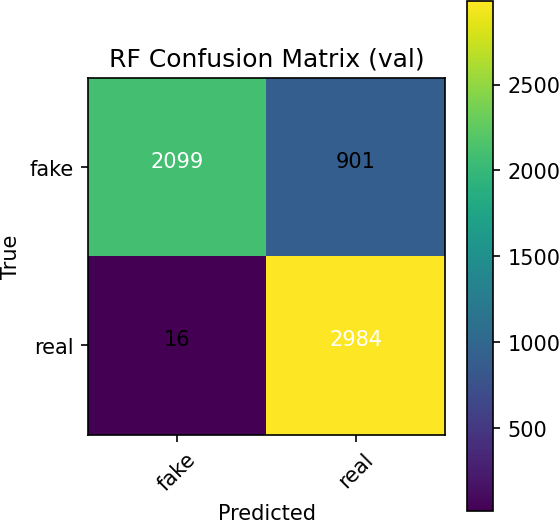
1. *Main figures*

To present results cleanly and avoid float overlap, we group related plots with figure\* and insert \FloatBarrier between figure blocks.

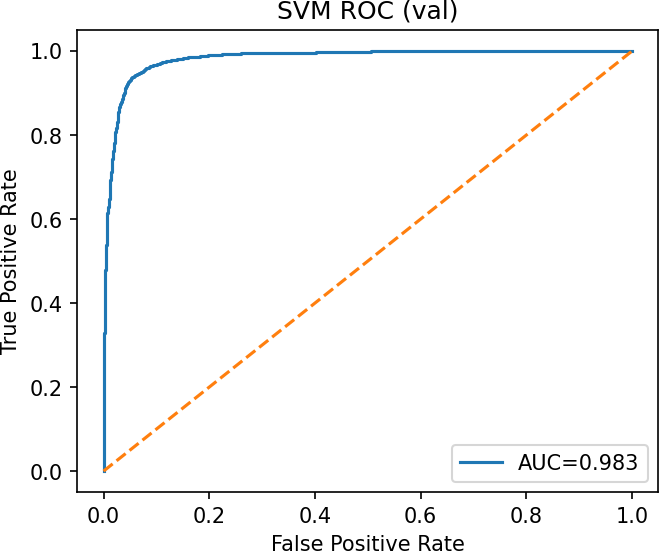
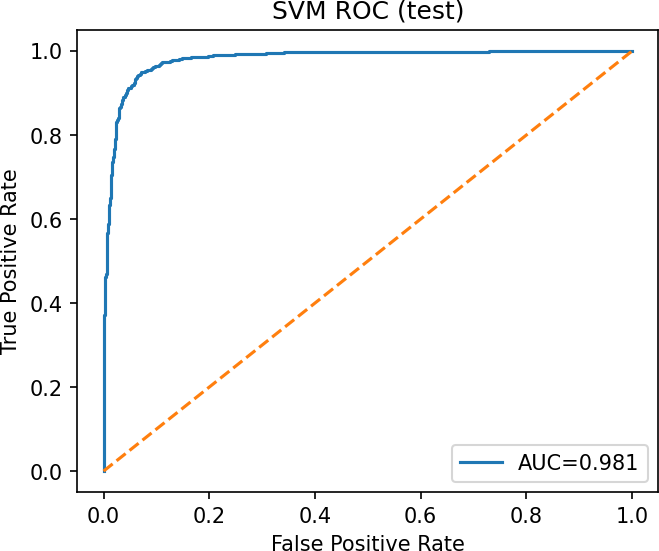
(a) RF ROC (val) (b) RF ROC (test)

Fig. 1: Random Forest ROC curves on validation and test sets.



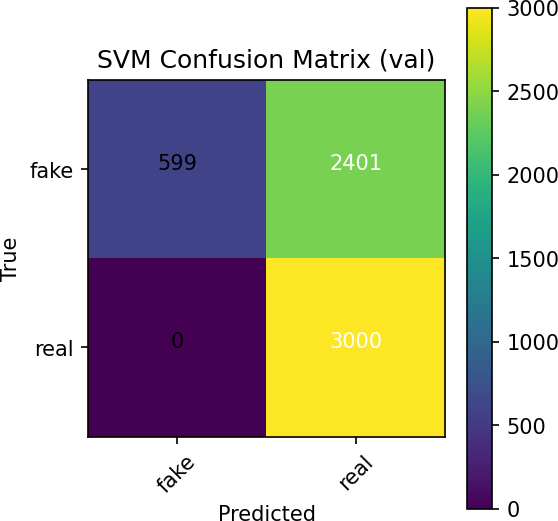
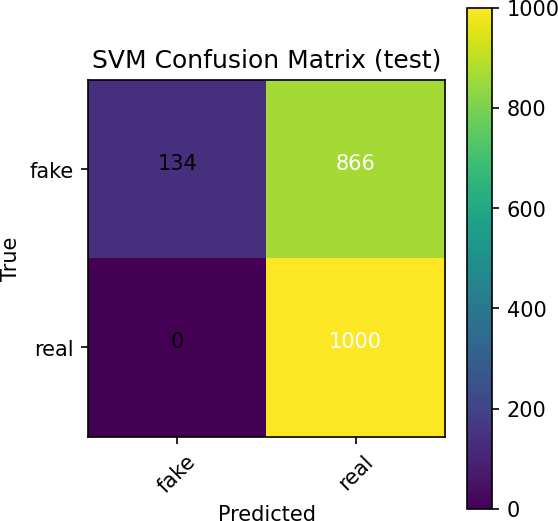
(a) RF Confusion Matrix (val) (b) RF Confusion Matrix (test)

Fig. 2: Random Forest confusion matrices on validation and test sets.

(a) SVM ROC (val) (b) SVM ROC (test)

Fig. 3: SVM ROC curves on validation and test sets.

(a) SVM Confusion Matrix (val) (b) SVM Confusion Matrix (test)

Fig. 4: SVM confusion matrices on validation and test sets.

1. Results
2. *In-dataset performance*

Table [I](#_bookmark4) summarises results on validation and test splits. Both models achieve near-saturated AUC on clean data, indicating strong separability. However, their operating characteristics differ: SVM attains perfect recall on both validation and test, but with noticeably lower precision, which depresses accuracy and F1. RF achieves substantially better precision while retaining very high recall, yielding higher accuracy and F1. This pattern is consistent with ensemble robustness to noisy or redundant features and with SVM’s tendency to push the margin to cover borderline positives when *γ* is tuned for sensitivity [[6](#_bookmark11)], [[18](#_bookmark22)].

TABLE I: In-dataset results on validation and test splits (proportions).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | Split | Accuracy | Precision | Recall | F1 | AUC |
| SVM | val | 0*.*5998 | 0*.*5555 | 1*.*0000 | 0*.*7142 | 0*.*9827 |
| SVM | test | 0*.*5670 | 0*.*5359 | 1*.*0000 | 0*.*6978 | 0*.*9812 |
| RF | val | 0*.*8472 | 0*.*7681 | 0*.*9947 | 0*.*8668 | 0*.*9887 |
| RF | test | 0*.*7835 | 0*.*6995 | 0*.*9940 | 0*.*8211 | 0*.*9813 |

1. *Robustness under perturbations*

Table [II](#_bookmark5) shows performance on JPEG, blur, and Gaussian noise. Both models retain high AUCs (reflecting separability in score space), but RF delivers systematically higher accuracy and F1 than SVM across all perturbations, mainly by avoiding false positives. JPEG and blur induce modest degradation; additive noise impacts SVM precision most strongly, consistent with the sensitivity of RBF kernels to noisy features unless *C* is reduced (at the expense of recall) [[5](#_bookmark10)], [[15](#_bookmark19)]. These results echo robustness studies that favour ensembles and frequency-aware cues under compression/noise [[5](#_bookmark10)], [[23](#_bookmark26)].

TABLE II: Robustness results on perturbed test sets (JPEG quality 50, blur radius 3, Gaussian noise *σ*=0*.*02).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | Split | Accuracy | Precision | Recall | F1 | AUC |
| SVM | test jpeg 50 | 0*.*5590 | 0*.*5313 | 1*.*0000 | 0*.*6940 | 0*.*9819 |
| RF | test jpeg 50 | 0*.*7545 | 0*.*6721 | 0*.*9940 | 0*.*8019 | 0*.*9799 |
| SVM | test blur 3 | 0*.*5710 | 0*.*5382 | 1*.*0000 | 0*.*6998 | 0*.*9805 |
| RF | test blur 3 | 0*.*7915 | 0*.*7078 | 0*.*9930 | 0*.*8265 | 0*.*9813 |
| SVM | test gauss noise 0.02 | 0*.*5445 | 0*.*5233 | 1*.*0000 | 0*.*6870 | 0*.*9784 |
| RF | test gauss noise 0.02 | 0*.*7125 | 0*.*6358 | 0*.*9950 | 0*.*7758 | 0*.*9758 |

1. *ROC/CM analysis and thresholding*

Figures [1](#_bookmark0) and [3](#_bookmark2) show ROC curves on validation and test sets. RF operates on the upper-left frontier with AUCs close to SVM’s, but with a different precision–recall balance. Confusion matrices in Figures [2](#_bookmark1) and [4](#_bookmark3) reveal SVM’s extreme bias toward the positive class: recall is 1.0, but false positives inflate (lower precision). In risk-sensitive deployments, such a profile could overwhelm downstream human review. Our validation-driven thresholding reduces this effect, but a calibrated RF remains preferable for balanced workloads [[18](#_bookmark22)].

1. *Error modes*

Qualitatively, false positives often correspond to unusual lighting or textured backgrounds that mimic blending arte- facts; false negatives typically involve high-quality manip- ulations with strong local consistency near the mouth and eyes—precisely where many detectors search for artefacts [[2](#_bookmark7)], [[4](#_bookmark9)]. Feature importance in RF highlights gradient and frequency- proxy statistics as strong signals, echoing frequency-aware literature [[5](#_bookmark10)], [[16](#_bookmark20)].

1. Discussion

**Why classical ML still matters.** Classical models paired with frozen deep features provide three pragmatic advantages:

(i) low training cost and predictable inference latency; (ii) transparent decision logic via feature importance and margins; and (iii) flexibility—features can be swapped or extended without end-to-end retraining. In resource-constrained or edge settings, such properties can outweigh the marginal accuracy gains of larger networks [[6](#_bookmark11)], [[18](#_bookmark22)].

**Robustness and domain shift.** Our results under JPEG/noise/blur indicate that RF preserves F1 better than SVM at matched AUC. This aligns with studies showing that detectors risk overfitting to codec/idiosyncratic artefacts and that ensembling or frequency-aware cues offer stability [[5](#_bookmark10)], [[15](#_bookmark19)], [[23](#_bookmark26)]. For cross-dataset deployment, we recommend: (a) calibration on a small target set, (b) confidence monitoring, and (c) periodic re-evaluation with held-out degradations.

**Limitations.** First, our embeddings use VGG16 without fine- tuning; a stronger backbone (e.g., EfficientNet or a ViT) might improve margins but at higher cost. Second, we evaluate images rather than full videos; temporal cues such as motion and lip synchrony can further boost robustness [[4](#_bookmark9)], [[19](#_bookmark23)]. Third, we do not address adversarial examples; robust training or input pre- filtering may be required where threat models include adaptive attackers.

**Future directions.** Promising avenues include domain gen- eralisation with style augmentations, curriculum learning with hard negatives [[14](#_bookmark18)], and spectral regularisation to suppress short- cuts [[5](#_bookmark10)]. Lightweight temporal pooling (e.g., clip-level voting) could lift performance on videos with minimal overhead.

1. Ethical, Legal, and Professional Issues

Detection tools interact with sensitive biometric data and potentially high-stakes decisions. We follow principles of neces- sity, proportionality, and transparency. Data should be lawfully sourced with appropriate licences, consent where required, and privacy safeguards. Systems should support human oversight, appeal mechanisms, and clear documentation of limitations [[7](#_bookmark12)], [[25](#_bookmark28)]. Fairness is critical: false positives can disproportionately harm certain groups if datasets are unbalanced. Practitioners should report demographic performance where applicable and adopt governance practices consistent with institutional policies.

1. Conclusion

We presented a compact deepfake detector that combines handcrafted features with frozen VGG16 embeddings and classical SVM/RF classifiers. Despite its simplicity, the pipeline achieves strong AUC on clean data and superior F1/accuracy under realistic degradations compared to the SVM baseline. Beyond raw metrics, the approach emphasises transparency, reproducibility, and practical deployment considerations. We hope these results encourage further study of robust, explainable baselines that complement larger end-to-end systems and provide dependable first-line defences in real-world pipelines [[18](#_bookmark22)], [[23](#_bookmark26)].

References

1. J. Fridrich and J. Kodovsky´, “Rich models for steganalysis of digital images,” *IEEE TIFS*, 2012.
2. D. Afchar *et al.*, “MesoNet: a compact facial video forgery detection network,” *WIFS*, 2018.
3. A. Ro¨ ssler *et al.*, “FaceForensics++: Learning to detect manipulated facial images,” *ICCV*, 2019.
4. A. Haliassos *et al.*, “Lips Don’t Lie: A generalisable and robust approach to face forgery detection,” *CVPR*, 2021.
5. R. Durall *et al.*, “Watch your up-convolution: CNN-based generative deep neural networks are failing to reproduce spectral distributions,” *CVPR*, 2020.
6. L. Guarnera, O. Giudice, and S. Battiato, “DeepFake detection by analyzing convolutional traces,” *ICASSP*, 2020.
7. R. Chesney and D. Citron, “Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security,” *California Law Review*, 2019.
8. P. Korshunov and S. Marcel, “Vulnerability of face recognition to deepfakes: a critical evaluation,” *WACVW*, 2019.
9. H. Dang *et al.*, “On the detection of digital face manipulation,” *CVPR Workshops*, 2020.
10. Y. Li *et al.*, “Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics,” *CVPR*, 2020.
11. B. Dolhansky *et al.*, “The DeepFake Detection Challenge (DFDC) Dataset,” *arXiv*, 2020.
12. Y. Li *et al.*, “Exposing DeepFake Videos By Detecting Face Warping Artifacts,” *CVPR Workshops*, 2018.
13. L. Jiang *et al.*, “DeeperForensics-1.0: A Large-Scale Dataset for Real- World Face Forgery Detection,” *CVPR*, 2020.
14. Y. Li and S. Lyu, “Exposing DeepFake Videos By Detecting Face Manipulation Artifacts,” *arXiv*, 2019.
15. N. Nezami *et al.*, “Generalizability of DeepFake Detection: A Cross- Dataset Study,” *arXiv*, 2021.
16. Y. Qian *et al.*, “Thinking in Frequency: Face Forgery Detection by Mining Frequency-Aware Clues,” *ECCV*, 2020.
17. D. Cozzolino, G. Poggi, and L. Verdoliva, “Recasting residual-based local descriptors as convolutional neural networks,” *ICPR*, 2017.
18. L. Verdoliva, “Media Forensics and DeepFakes: An Overview,” *IEEE* *JSTSP*, 2020.
19. T. Mittal *et al.*, “Emotions Don’t Lie: A Deepfake Detection Method using Audio-Visual Affective Cues,” *ACM MM*, 2020.
20. H. H. Nguyen *et al.*, “Capsule-Forensics: Using Capsule Networks to Detect Forged Images and Videos,” *ICASSP*, 2019.
21. E. Sabir *et al.*, “Recurrent Convolutional Strategies for Face Manipulation Detection in Videos,” *CVPR Workshops*, 2019.
22. P. Zhou *et al.*, “Learning Rich Features for Image Manipulation Detection,”

*CVPR*, 2018.

1. S. Wang *et al.*, “CNN-generated images are surprisingly easy to spot... for now,” *CVPR*, 2020.
2. K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” *arXiv*, 2014.
3. S. Agarwal *et al.*, “Protecting World Leaders Against Deep Fakes,” *CVPR Workshops*, 2019.