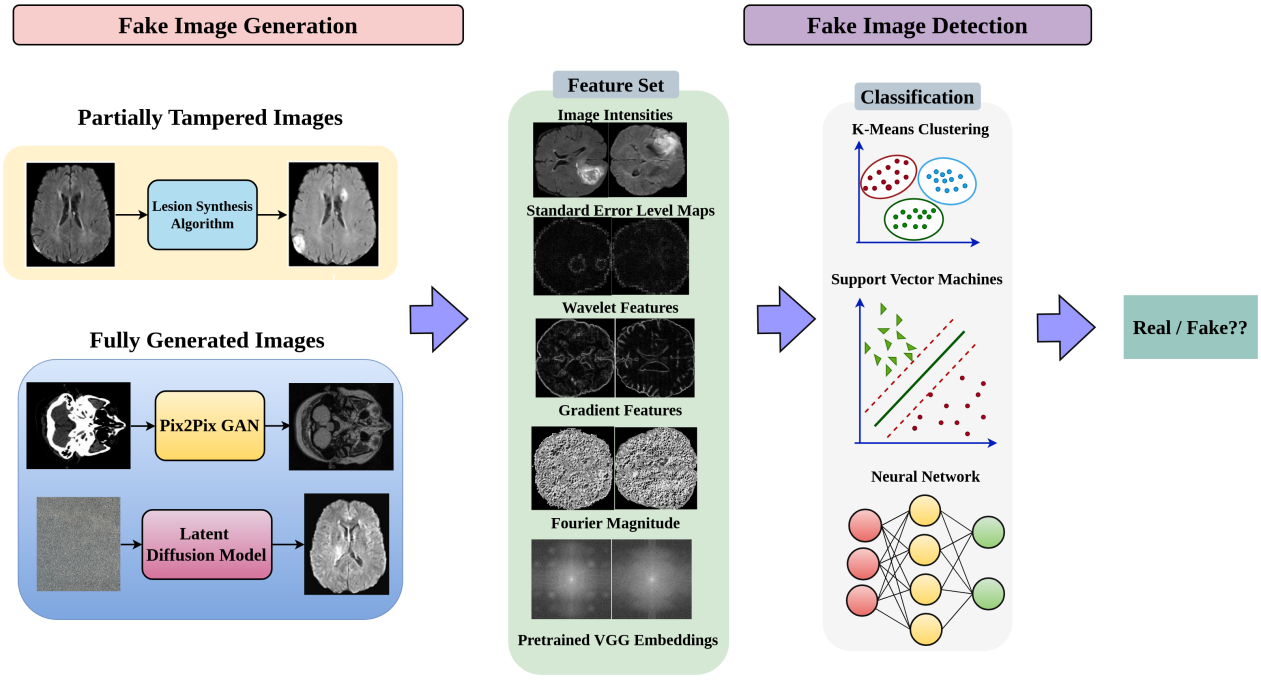
Graphical Abstract

Integrating Fourier Analysis and Deep Learning for Robust Detection of Deep Fake Brain Magnetic Resonance Images

Vaishnavi Ravi, Yogesh K. Sahu, Prabhas R. Onteru, Parag Dutta, Dhanshree Warokar, Padma Murali, Rajesh Katta, Ambedkar Dukkipati, Phaneendra K. Yalavarthy



Highlights

Integrating Fourier Analysis and Deep Learning for Robust Detection of Deep Fake Brain Magnetic Resonance Images

Vaishnavi Ravi, Yogesh K. Sahu, Prabhas R. Onteru, Parag Dutta, Dhanshree Warokar, Padma Murali, Rajesh Katta, Ambedkar Dukkipati, Phaneendra K. Yalavarthy

- Fourier features achieved up to 99.8% accuracy in detecting fake brain MRI scans
- VGG19 embeddings with SVMs showed robust performance even with limited data
- CNNs outperformed SVMs on most features but struggled with very small datasets
- The framework detected various fake types including GAN and diffusion-generated images
- Combining domain-aware features and supervised learning enables reliable detection

Integrating Fourier Analysis and Deep Learning for Robust Detection of Deep Fake Brain Magnetic Resonance Images

Vaishnavi Ravi[a,*], Yogesh K. Sahu[b], Prabhas R. Onteru[b], Parag Dutta[b], Dhanshree Warokar[c], Padma Murali[c], Rajesh Katta[c], Ambedkar Dukkipati[b] and Phaneendra K. Yalavarthy[a]

[a]Department of Computational and Data Sciences, Indian Institute of Science, Bengaluru, 560012, Karnataka, India
[b]Department of Computer Science and Automation, Indian Institute of Science, Bengaluru, 560012, Karnataka, India
[c]Accenture, Innovation Hub, Building 14, Pritech Park, Bellandur, Bengaluru, 560103, Karnataka, India

ARTICLE INFO

Keywords:
Medical Deep Fake Detection
Synthetic data
Fourier Features
Deep Feature Embeddings
Machine Learning
Deep Learning

ABSTRACT

Recent advances in generative models have enabled the creation of highly realistic synthetic medical images, helping to address the challenges of data scarcity. However, these technologies pose a threat by generating realistic fake medical images that can mislead clinical diagnoses. In this work, we proposed a framework for detecting fake brain MRIs using handcrafted features and deep representations. Specifically, we extracted raw intensities, Error Level Analysis (ELA), wavelet, gradient, and Fourier features, along with embeddings from a pre-trained VGG19 network. We evaluated these features, derived from both real and fake MRIs, using machine learning (K-Means clustering, Support Vector Machines (SVM)) and deep learning (MobileNetV2) approaches. Experimental results show that Fourier-based features achieve the highest detection accuracy of 99.5% with SVM and 99.8% with CNNs. Additionally, VGG19 embeddings achieved 98.8% accuracy with SVM even in low-data regimes. These findings highlight the effectiveness of combining domain-aligned features with supervised learning for robust detection of medical deepfakes.

1. Introduction

Deep learning has become a transformative tool in computer vision and has been widely adopted in medical imaging for tasks such as diagnosis, segmentation, and disease classification [16]. However, the effectiveness of these models depends on the availability of large, well-annotated datasets, which remain limited in medical imaging due to privacy constraints, annotation costs, and restricted data access. To address this limitation, generative models such as Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs), and Denoising Diffusion Probabilistic Models (DDPMs) have been increasingly used to synthesize realistic images for data augmentation [5]. While these approaches have improved model training and generalization, their ability to generate visually convincing images has also raised concerns regarding image authenticity, motivating extensive research on deepfake detection. These risks are even more pronounced in medical imaging, as demonstrated in prior work [18], where realistic alterations of 3D medical scans using GAN-based methods were shown to be feasible, potentially leading to misdiagnosis, fraudulent insurance claims, or compromised clinical studies.

Most existing deepfake detection methods are designed for natural images and primarily exploit visual artifacts, semantic inconsistencies, or statistical irregularities. In contrast, medical images are acquired using highly standardized imaging protocols and are characterized by subtle structural and textural variations that are clinically meaningful. As a result, even small perturbations in intensity distributions or frequency components introduced by synthetic generation models can have significant diagnostic implications, while remaining visually imperceptible.

Detecting fake medical images therefore presents distinct challenges, including limited availability of labeled data, low inter-class variability between real and synthetic images, modality-specific noise characteristics, and stringent requirements for interpretability and reliability. Moreover, detection strategies developed for natural images often rely on high-level semantic cues that are largely absent in medical imaging modalities such as MRI, motivating the need for domain-aware detection approaches.

Despite these challenges, most works in medical imaging are focused on synthetic image generation rather than detection. This imbalance is reflected in a PubMed search, where the terms “data augmentation” and “synthetic data generation” yield over 950 articles, compared to fewer than 30 results for “fake medical image detection,” largely related to textual misinformation. This highlights a critical gap in safeguarding the integrity of medical imaging pipelines.

In order to address this, we propose a robust deepfake detection framework that leverages diverse domain-aligned features beyond raw intensities to detect medical deepfakes in brain Magnetic Resonance (MR) images. The key contributions of this work are:

1. A comprehensive framework is presented for medical deepfake detection that systematically integrates handcrafted features such as Error-Level Analysis (ELA),

*Corresponding author
vaishnavi1712@gmail.com (V. Ravi); yogeshsahu@iisc.ac.in (Y.K. Sahu); prabhasreddy@iisc.ac.in (P.R. Onteru); paragdutta@iisc.ac.in (P. Dutta); dhanshree.warokar@accenture.com (D. Warokar); padma.murali@accenture.com (P. Murali); rajesh.katta@accenture.com (R. Katta); ambedkar@iisc.ac.in (A. Dukkipati); yalavarthy@iisc.ac.in (P.K. Yalavarthy)
ORCID(s): 0000-0002-7861-1026 (V. Ravi); 0000-0002-9147-0117 (P. Dutta); 0000-0003-4810-352X (P.K. Yalavarthy)

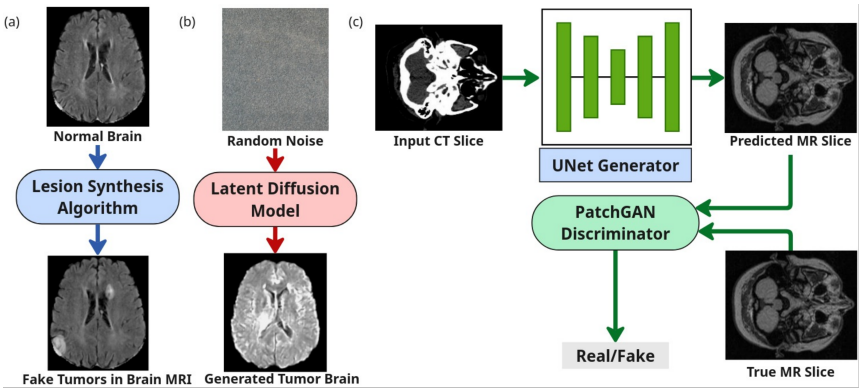


Figure 1: Illustration showing (a) MRI Scans tampered using MedLesSynth [19] Algorithm (b) Fully generated MRI volumes from noise using Latent Diffusion Model, and (c) Fully generated MRI slices from CT slices using conditional Pix2Pix [12] GAN model.

Wavelet, Gradient, Fourier, and deep representations to characterize image manipulations across multi-source clinical datasets.

2. A robust evaluation setup is established by generating both partially tampered images via synthetic lesion insertion and fully synthetic images using GANs and Latent Diffusion Models (LDMs), covering a wide range of manipulation scenarios.
3. Machine learning (K-means, SVM) and deep learning (MobileNetV2) combined with feature selection (Recursive Feature Elimination (RFE)), is employed to evaluate detection performance and relative importance of complementary feature representations.
4. Extensive cross-dataset and cross-modality evaluations are conducted to examine generalizability across diverse datasets & imaging modalities, establishing a strong baseline for future research on medical deepfake detection.

2. Related Work

In literature, there has been a lot of work around detection of various types of deepfakes from videos, images, audio, text & realtime manipulation [24]. Particularly, in image-based deepfake detection, most studies have concentrated on biometrics, such as faces [21] and fingerprints [2]. Identification of fake faces has been investigated using various features like spatio-temporal feature representation [34], RGB-depth integration [14], anisotropic features [11], GAN fingerprinting [23], facial landmarks [8], deep features [33] etc. On the other hand, fingerprint liveness detection is been explored by using handcrafted features such as Local Binary Patterns (LBP) [2]. A fair amount of work has been done on natural images as well [15]. Eventhough researchers well-explored deepfakes for biometrics, investigations specific to medical images remain limited.

Solaiyappan et. al. [31] conducted a comparative study of eight machine learning models, including both classical and deep learning approaches, achieving near-perfect accuracy. In [20], authors introduced an SVM-based method to detect overtly artificial GAN-generated lung CT images, achieving 75.5% accuracy and demonstrating potential as a quality control tool for synthetic datasets. A deep learning approach using multiple YOLO models was proposed in [13] to detect manipulated medical images from knee X-rays and lung CTs. Similarly, Sharafudeen et al. [29] developed Derm-CGAN to generate realistic dermoscopic lesion deepfakes and used Vision Transformers for detection, achieving 97.18% accuracy. Finally, Amiri et al. [3] proposed CFDMI-SEC, a three-stage approach integrating SIFT, Equilibrium Optimization, and Color Histogram Matching to detect copy-move forgeries in medical images, and reported up to 100% accuracy. However, all these existing methods are largely evaluated on single-source datasets, limiting their generalizability across diverse clinical settings. Their dependence on raw image intensities, coupled with the absence of cross-dataset validation, provides limited robustness. Furthermore, approaches like CFDMI-SEC [3], which focus on copy-move forgeries using handcrafted features, are less effective against GAN-generated deepfakes.

To the best of our knowledge, the proposed work is the first one to integrate handcrafted features with deep learning embeddings, establishing a baseline for medical deepfake detection. Moreover, the framework leverages multiple datasets and complementary feature representations to enhance robustness and generalizability.

3. Materials & Methods

This section outlines the datasets used, steps for generating synthetic images, and feature extraction techniques. We then detail the unsupervised, supervised, and deep learning strategies employed to identify fake brain MR images .

3.1. Datasets

Our analysis required two classes of data: real and fake medical images. Unlike previous studies that used single dataset and yielded near-perfect accuracies, we adopted a rigorous approach by integrating multiple datasets from public repositories. We focused on brain MR images as extensive set of high-quality datasets are freely available.

3.1.1. Real Data

Real medical images are obtained from three major publicly accessible sources, encompassing both normal brain MR scans and tumor-affected brain MR images as given in Table.1.

Table 1
Summary of real brain MRI datasets used in this study, including normal and tumor cases from multiple datasets. Five slices per volume were uniformly selected for the experiments.

Brain Type	Dataset	#Patients (#Slices)
Normal	CERMEP-IDB-MRXFDG	19 (95)
Tumor	UPENN	335 (1675)
Tumor	TCGA	51 (255)
Total		**405 (2025)**

1. **CERMEP-IDB-MRXFDG Dataset** [17]: Provided by the CERMEP imaging center (Lyon, France), this multimodal brain imaging dataset comprises data from 37 healthy adult participants aged 23–65 years. It included T1-weighted, FLAIR MRI, low-dose CT, and FDG-PET scans. All images were co-registered and delivered in NIfTI format, adhering to the Brain Imaging Data Structure (BIDS) standard.
2. **UPENN Glioblastoma Dataset** [4]: Released by the University of Pennsylvania, this dataset contains preoperative multi-parametric MRI scans, specifically T1, T1-Gd (gadolinium-enhanced), T2, and FLAIR of 671 glioblastoma patients. The dataset includes expert-annotated tumor segmentations and undergone preprocessing steps, such as co-registration, skull-stripping, and resampling. It is widely utilized for tumor segmentation research and is accessible via The Cancer Imaging Archive (TCIA).
3. **TCGA-GBM Dataset** [28]: Part of the Cancer Genome Atlas initiative, TCGA-GBM dataset provides multiparametric MRI scans (T1, T1-Gd, T2, FLAIR) for 262 glioblastoma patients. Among these, a subset of 135 patients included expert-annotated tumor segmentations and associated radiomic features, facilitating imaging genomics studies.

3.1.2. Fake Dataset Generation Methods

In this study, fake medical images are defined as images that either contain artificially introduced tumors or are fully synthetically generated, resulting in a misrepresentation of true anatomical or pathological features. Accordingly, two categories of fake images are considered:

- **Partially Tampered Images** – Localized manipulations are generated using MedLesSynth [19], which inserts synthetic three-dimensional tumors into MRI scans. Synthetic lesions are embedded into healthy scans (CERMEP) as well as tumor-bearing scans (UPenn and TCGA), producing structurally plausible images with altered pathological content. An example is shown in Figure. 1(a), and a quantitative summary is provided in Table 2.
- **Fully Generated Images** – Two generative models are used to synthesize entirely fake medical images. *(i) Pix2Pix GAN [12]* is trained on paired CT–MRI data from CERMEP to perform modality translation, generating MRI-like images conditioned on CT inputs. *(ii) Latent Diffusion Model [26]*, pretrained on the BRATS dataset, is used to generate synthetic brain tumor MR images from random noise through iterative denoising, producing semantically coherent fake volumes.

Table 2
Summary of synthetic brain MRI datasets used in this study, generated using different methods across various source datasets. Five slices per volume were used.

Generation Method	Dataset	Type	#Patients (#Slices)
MedLesSynth (Local)	UPENN	Tumor	336 (1680)
	TCGA	Tumor	51 (255)
	CERMEP	Tumor	18 (90)
Pix2pix GAN(Full)	CERMEP	Normal	37 (185)
Diffusion (LDM)(Full)	BRATS-like	Tumor	500 (2500)
Total			**942 (4710)**

Table 3
Clustering performance metrics obtained using the penultimate layer of a VGG19 model pretrained on ImageNet.

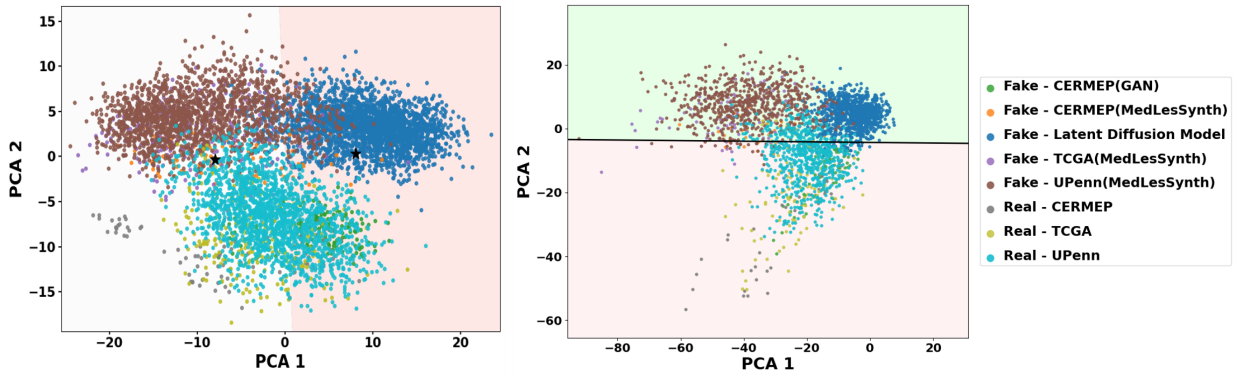
Accuracy(%)	Precision(%)	Recall(%)	F1 Score(%)
40.2	60.1	43.1	50.2

These approaches allow simulating a range of fake medical images, from locally tampered to fabricated ones, supporting comprehensive evaluation of detection methodologies across adversarial scenarios.

3.2. Feature Extraction Techniques

In this study, we employ domain-aligned features beyond raw intensities, including ELA, gradient, wavelet, and Fourier magnitude features to capture MRI-specific signal- and texture-level inconsistencies, along with VGG19 embeddings for higher-level spatial representation.

1. *Raw Pixel Intensities:* Used as a baseline, representing native MRI intensity distributions that reflect tissue-dependent contrast and acquisition characteristics, enabling direct comparison with higher-level domain-aligned feature representations.
2. *Standard ELA* [1]: This method highlights localized reconstruction and compression inconsistencies that are atypical for clinically acquired MR images. Such inconsistencies are particularly relevant for identifying localized or region-specific manipulations, where a small pathological structure may be digitally introduced into an otherwise authentic scan. Given an input image I, it is recompressed at a fixed JPEG quality Q to produce $I_Q = \text{JPEG}(I, Q)$. The pixel-wise absolute difference is computed as $ELA(x, y) = \left|I(x, y) - I_Q(x, y)\right|$, emphasizing regions with abnormal reconstruction behavior arising from synthetic generation or post-processing.
3. *Wavelet-based:* A Discrete Wavelet Transform (DWT) is applied to each image I, decomposing it into one approximation (LL) and three detail sub-bands (LH, HL, HH): $I \xrightarrow{\text{DWT}} \{LL, LH, HL, HH\}$. The LL sub-band captures coarse anatomical structure, while the detail sub-bands encode fine-scale textural patterns. Wavelet features are aligned with the multi-scale textural analysis routinely used in medical image interpretation, making



(a) (b)

Figure 2: (a) t-SNE visualization of the embedding space showing cluster assignments and decision boundaries using K-Means clustering. The lack of clear separation among clusters suggests random assignment, consistent with poor clustering performance. (b) SVM decision boundary visualization of the embedding space of test samples at 50% test ratio, showing clear separation between the classes and effective discrimination of features, reflecting improved classification performance.

them sensitive to subtle texture inconsistencies introduced by synthetic image generation.

4. *Gradient-based:* Gradient maps are extracted from real and fake images using Sobel operators, which highlight intensity changes along horizontal and vertical directions. For an image I, the horizontal and vertical gradients G_x and G_y are computed as $G_x = I * S_x$, $G_y = I * S_y$, where $*$ denotes convolution and S_x and S_y are the Sobel kernels given by

$$S_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}, \quad S_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}.$$

The gradient magnitude is computed as $G = \sqrt{G_x^2 + G_y^2}$. These features emphasize anatomical boundary transitions that are critical in medical imaging, where sharp organ and tissue margins are expected. Blurring or inconsistencies along such boundaries, which may arise during synthetic image generation, are therefore effectively captured by gradient-based representations.

5. *Fourier-based:* To capture global frequency inconsistencies, a 2D Fast Fourier Transform (FFT) is applied to each image I. The 2D FFT is defined as

$$F(u,v) = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} I(x,y)\, e^{-j2\pi\left(\frac{ux}{M}+\frac{vy}{N}\right)},$$

where M and N denote the image dimensions and (u, v) the frequency coordinates. The normalized magnitude spectrum is computed as

$$F_{\text{mag}}(u,v) = \frac{|F(u,v)|}{\max(|F(u,v)|)}.$$

Fourier magnitude features have frequency-domain characteristics of medical image acquisition, particularly MRI, where images are reconstructed from k-space measurements. Biological tissues exhibit smooth and continuous spectral distributions, whereas synthetic images may introduce periodic artifacts or spectral irregularities due to generative upsampling and reconstruction processes. Such deviations are often imperceptible in the spatial domain but are effectively captured in the frequency domain, making Fourier-based features well suited for detecting synthetic medical images.

6. *ImageNet-pretrained VGG19 Embeddings:* High-level feature representations are extracted using the VGG19 network [30] pretrained on ImageNet. Feature embeddings of *4096 dimensions* are obtained from the penultimate fully connected layer. Although trained on natural images, these embeddings capture generic spatial correlations and structural patterns which, when combined with domain-aligned handcrafted features, improve robustness and discrimination in data-limited medical imaging scenarios.

Each representation is used as input to the detection models to assess their effectiveness in detecting fake medical images. This multi-representation approach allows for a comprehensive analysis of the structural and statistical irregularities resulting from image manipulation.

3.3. Classification Methods

This study integrates unsupervised and supervised machine learning & deep learning techniques to detect fake medical images by employing handcrafted and learned features.

3.3.1. Unsupervised Classification

To assess the intrinsic separability of real and fake images without label supervision, *K-Means clustering* is applied to the 4096-dimensional VGG19 embeddings after dimensionality reduction using *Principal Component Analysis (PCA)*. This analysis is included as an exploratory assessment of task difficulty and feature separability, rather than as a detection baseline.

Table 4
Classification accuracies across different test splits using various features and models. The evaluated features included pixel intensity values, Standard ELA, Wavelet, Gradient, Fourier magnitude, VGG19-based embeddings, IFL, and FSC. The rResults are reported for both the SVM and CNN classifiers, with the highest values shown in bold and the second -highest in bold italics. */** indicates $p < 0.05/p < 0.001$ showing the statistical significance of VGG embeddings (SVM) with respect to the corresponding method. Features marked with (✓) denote interpretable, domain-aligned representations, while (×) indicates features with limited inherent interpretability.

Test	Intensities (✓)		Standard ELA (✓)		Wavelet (✓)		Gradient (✓)		Fourier (✓)		VGG Embeds (×)	IFL (✓)	FSC (✓)
Split	SVM	CNN	SVM	CNN	SVM	CNN	SVM	CNN	SVM	CNN	SVM	CNN	SVM
0.3	0.939**	0.996	0.941**	0.925**	0.897**	0.984	0.920**	0.984*	0.993	**0.998**	***0.998***	**1.00**	**0.998**
0.4	0.930**	0.995	0.940**	0.950**	0.896**	0.993	0.915**	0.981	0.995	**0.997***	***0.998***	**1.00**	**0.999**
0.5	0.930**	0.994*	0.936**	0.950**	0.892**	0.984	0.913**	0.985*	0.995	**0.998****	***0.997***	**1.00**	**0.999**
0.7	0.969**	0.993**	0.922**	0.943**	0.894**	0.984	0.905**	0.948**	0.995*	***0.996*****	**0.997**	**1.00**	**0.999**
0.9	0.902**	0.976	0.904**	0.853**	0.874**	0.963	0.882**	0.959**	***0.990****	0.989**	**0.993**	**0.99**	**0.996**
0.95	0.904**	0.869**	0.890**	0.707**	0.864**	0.708**	0.847**	0.926*	***0.988****	0.933	**0.989**	**0.988**	**0.997**

3.3.2. Supervised Classification

- ***Support Vector Machines:*** To assess the discriminative power of various feature representations derived from the images, we trained an SVM [10] classifier using multiple feature modalities, including standard ELA, wavelet, gradient and fourier-magnitude, along with raw pixel intensities, and imageNet pretrained VGGNet19 embeddings. Each of these feature modalities is individually evaluated using SVMs to determine their effectiveness in distinguishing between real and fake medical images.
- ***Convolutional Neural Networks (CNNs):*** To complement traditional machine learning approaches, we employed MobileNetV2 [27] model chosen for its efficiency and strong performance even with limited datasets. This model is evaluated to investigate the ability of the deep learning model to learn deepfake signatures from raw pixels as well as handcrafted features.

3.4. Implementation Details

To perform fake detection analysis, each real brain MRI dataset (UPenn-GBM, TCGA-GBM, and CERMEP-IDB-MRXFDG) is randomly split into two equal subsets, one used as the real class and the other for fake sample generation. Local lesion–based fake images are generated using the MedLesSynth [19] framework, which inserts synthetic tumors into real healthy MR scans to produce realistic manipulated images. For GAN-based synthesis, all available CT scans are used as inputs to a Pix2Pix model to generate corresponding fake MR images, which are treated as the fake class. In addition, a Latent Diffusion Model [26] pretrained on the BraTS dataset is employed to generate 500 fully synthetic tumored brain MR volumes from random noise, providing a complementary generative paradigm based on deep latent priors.

For dataset construction, five axial slices are uniformly sampled from each 3D volume across both real and fake classes to ensure consistency and reduce redundancy. The resulting slice counts for each dataset are summarized in Tables 1 and 2. All experiments are conducted on a Linux workstation equipped with an Intel i9-10980XE CPU, 128,GB RAM, and two NVIDIA RTX A6000 GPUs (48,GB VRAM each).

4. Results

In this section, we present the results of all three methods: (i) Unsupervised K-means clustering, (ii) SVM, and (iii) MobileNetV2 on features extracted from multi-sourced datasets and their integration.

4.1. Experiments with Unsupervised K-Means

To assess unsupervised real–fake separability, K-means clustering is applied to ImageNet-pretrained VGG19 embeddings of brain MR slices, with results reported in Table 3.

As shown in Table 3, unsupervised clustering achieves limited performance, with an accuracy of 40.2% and an F1 score of 50.2%, indicating weak real–fake separability. The t-SNE visualization in Figure. 2, overlaid with cluster centroids and decision boundaries, further confirms this observation, with samples exhibiting substantial overlap and no clear cluster structure. These results highlight the limitations of unsupervised clustering for this task and motivate the use of supervised learning to achieve reliable fake detection.

4.2. Experiments with Supervised approaches

For supervised classification, we employed SVM and MobileNetV2. All handcrafted and image-based features including raw intensities, standard ELA maps, wavelet, gradient-based, and Fourier magnitude representations were evaluated using both models. VGG19 feature embeddings were classified exclusively using SVM. To further extend the evaluation, two additional analyses were conducted: (1) Integrated Feature Learning (IFL) and (2) Feature Selection and Classification (FSC). In IFL, all five handcrafted feature types are concatenated into a single representation and used to train a MobileNetV2 classifier. In FSC, Recursive Feature Elimination (RFE) [9] is applied to PCA-reduced VGG19 embeddings (dimension = 1000) to select the most discriminative features, which are then classified using an SVM.

Performance of each feature-model combination including IFL & FSC are summarized in Table 4. Notably, among individual features, the Fourier ones consistently achieved the highest performance, with SVM and CNN models reaching up to 99.5% and 99.8% accuracies and CNN models

Table 5
Classification results across real and fake datasets using different feature types and classifiers. Values are in count (percentage).

Data Type	Method	Dataset	Intensities	Standard ELA	Wavelet	Gradients	Fourier	VGG Embeds
Real	SVM	Real(Total)	895 (87%)	906 (88%)	848 (83%)	877 (86%)	1006 (98%)	1013 (99%)
		UPENN	729 (86%)	786 (93%)	708 (84%)	739 (88%)	831 (98%)	839 (99%)
		TCGA	125 (96%)	89 (68%)	102 (78%)	107 (82%)	130 (100%)	126 (97%)
		CERMEP	41 (82%)	31 (62%)	38 (76%)	31 (62%)	45 (90%)	48 (96%)
Real	CNN	Real(Total)	1013 (99%)	943 (92%)	1001 (98%)	989 (96%)	1016 (99%)	-
		UPENN	837 (99%)	801 (95%)	822 (98%)	821 (97%)	837 (99%)	-
		TCGA	126 (96%)	99 (76%)	129 (99%)	118 (91%)	129 (99%)	-
		CERMEP	50 (100%)	43 (86%)	50 (100%)	50 (100%)	50 (100%)	-
Fake	SVM	Fake(Total)	2249 (95%)	2258 (95%)	2167 (92%)	2209 (93%)	2356 (99%)	2357 (99%)
		CERMEP	45 (100%)	45 (100%)	41 (91%)	45 (100%)	45 (100%)	45 (100%)
		CERMEP(GAN)	83 (87%)	88 (93%)	81 (85%)	86 (90%)	95 (100%)	92 (96%)
		UPENN(Fake)	760 (90%)	763 (91%)	695 (82%)	730 (87%)	837 (99%)	840 (100%)
		TCGA(Fake)	111 (85%)	113 (87%)	100 (76%)	98 (75%)	129 (99%)	130 (100%)
		DIFFUSION	1250 (100%)	1250 (99%)	1250 (100%)	1250 (100%)	1250 (100%)	1250 (100%)
Fake	CNN	Fake(Total)	2346 (99%)	2269 (96%)	2325 (98%)	2341 (99%)	2360 (100%)	-
		CERMEP	45 (100%)	44 (97%)	45 (100%)	45 (100%)	45 (100%)	-
		CERMEP(GAN)	95 (100%)	87 (91%)	95 (100%)	95 (100%)	95 (100%)	-
		UPENN(Fake)	830 (98%)	774 (92%)	816 (97%)	824 (98%)	840 (100%)	-
		TCGA(Fake)	126 (96%)	114 (87%)	119 (91%)	127 (97%)	130 (100%)	-
		DIFFUSION	126 (96%)	1250 (100%)	1250 (100%)	1250 (100%)	1250 (100%)	-

showing comparable results. This result highlights the discriminative power of frequency-domain representations for detecting synthetic or tampered content in medical images. Although the fake images are visually similar to the real ones in spatial domain, they exhibit distinct patterns in the frequency domain effectively captured by Fourier features, which can reveal subtle inconsistencies introduced during the generation.

In addition VGG19 embeddings showed strong classification capabilities, with SVM accuracies exceeding 99% across test splits. These results demonstrate transfer learning effectiveness, where pretrained networks provide generalizable features when paired with linear classifiers.

While CNNs outperformed SVMs on handcrafted features in moderate data regimes, their performance declined at the 95% test split due to limited training data. SVMs maintain robust performance across splits, particularly with high-dimensional embeddings, suggesting pretrained embeddings enable reliable classification in low-data settings, while CNNs benefit from larger training sets.Further, the IFL approach achieved perfect accuracies (up to 100%), highlighting the benefit of feature integration. FSC yielded near-perfect classification comparable to VGG embeds. Due to space constraints, only accuracies are reported in the main text, while corresponding precision, recall, and F1-scores demonstrating robust performance are provided under Sec.3 of supplementary material.

Figure 3 shows the top 1000 features selected by RFE, with Fourier features dominating among all. This again emphasizes the key role of frequency features in distinguishing real from synthetic images, consistent with Table 4. We conducted a dataset-wise analysis of predictions; results are presented in Table.5. Classification performance across datasets showed CNN-based models outperformed SVMs across all feature types and datasets, showing superior generalization. Fourier features achieved highest accuracy, particularly with CNNs. For real images, UPENN and TCGA datasets achieved high accuracy with CNNs, while SVMs showed more variability with ELA and gradient features. CNN models achieved near-perfect classification on all fake datasets, including GAN and diffusion-generated data, demonstrating robust deep feature representations. VGG embeddings performed well on real and fake datasets, attributed to high-level semantic priors in pretrained networks providing generalizable features even with linear classifiers.

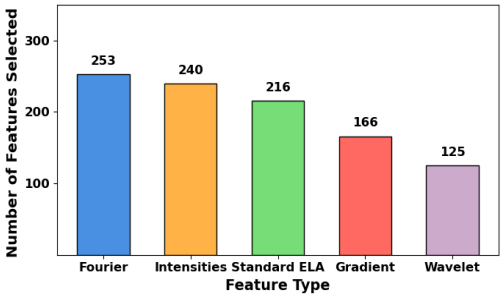


Figure 3: Distribution of the top 1000 features selected by RFE among all integrated features.

To assess computational efficiency, we present training and test times for SVM- and CNN-based models in Table. 6. CNNs require more training time but achieve faster testing speeds than SVMs. VGG embeddings with SVMs are most time-efficient overall, showing the efficiency of pretrained feature spaces. Further, IFL incurs the highest training time due to computational overhead due to processing of all feature maps. Though testing remains moderate, the preprocessing steps to extract all feature maps also contributes additional latency. Finally, FSC with RFE is extremely fast for both training and testing, however the RFE step itself required 79.71 seconds to select the top 1000 features adding a overhead to the processing time while the performance is almost similar to VGG19 embeds alone.

Table 6
Train and test times (in seconds) for various feature representations with SVM and CNN (MobileNetV2).

Mode	Intensities		Standard ELA		Wavelet		Gradient		Fourier		VGG Embeds	IFL	FSC
	SVM	CNN	SVM	CNN	SVM	CNN	SVM	CNN	SVM	CNN	SVM	CNN	SVM
Train	47.73	176.86	89.36	148.71	20.66	126.57	36.50	140.25	28.08	160.74	2.11	295.99	0.18
Test	41.94	7.19	87.05	6.67	20.77	5.14	34.20	6.50	26.54	7.61	1.36	3.81	0.02

5. Discussion

This study presents a comprehensive framework for fake medical image detection, with a focus on brain MRI. By integrating multiple publicly available datasets, the proposed evaluation addresses limitations of prior work that often relied on a single real–fake dataset, thereby enabling a more reliable assessment of generalization.

ELA-based features, traditionally effective for detecting copy–paste forgeries via compression inconsistencies, exhibit limited effectiveness in the presence of modern generative models such as GANs and diffusion models, which produce globally coherent images with minimal compression artifacts. As shown in Tables 4 and 5, both CNN- and SVM-based models relying on ELA features perform poorly, highlighting the need for feature representations that move beyond pixel-level inconsistencies.

To address this limitation, alternative representations were evaluated. Wavelet- and gradient-based features provide limited discrimination, as contemporary generative models increasingly reproduce multi-scale statistics and smooth textures common in medical images. Raw intensity features yield moderate performance with CNNs due to learned spatial hierarchies, but perform poorly with SVMs owing to their lack of spatial inductive bias. In contrast, Fourier-based features consistently achieve strong performance across datasets and classifiers. Unlike localized wavelet representations, Fourier analysis captures global frequency patterns and reveals periodic artifacts and high-frequency suppression introduced by generative models [7]. This is particularly relevant for MRI, where images are acquired in the frequency (k-space) domain, embedding scanner- and protocol-specific signatures that synthetic generation pipelines often fail to replicate. As a result, Fourier-domain representations remain effective forensic cues for fake medical image detection.

Deep features extracted from VGG19 also demonstrate strong performance, particularly when combined with linear SVMs, even in low-data regimes. This observation aligns with prior evidence on the effectiveness of transfer learning in medical imaging [32]. While CNNs perform well when sufficient data are available, their performance degrades under data scarcity. In contrast, SVMs paired with pretrained embeddings remain robust, suggesting that CNNs are better suited for data-rich settings, whereas SVMs are advantageous in low-data scenarios with structured feature representations [6, 22]. Although IFL achieves near-perfect classification by integrating all feature types, it incurs substantial computational overhead (Table 6). FSC offers faster training and inference, but its RFE-based feature selection introduces additional preprocessing latency, limiting its practical advantage. Feature selection analysis (Figure. 3) further indicates that frequency-domain, gradient-based, and wavelet features corresponding to interpretable signal properties such as frequency consistency, anatomical boundary sharpness, and multi-scale texture patterns are most relevant for the detection task. This data-driven preference for domain-aligned features aligns with interpretability perspectives discussed by Reyes et al. [25], where engineered clinical features (e.g., vessel tortuosity) provided more transparent diagnostic cues than raw pixel representations.Overall, the marginal performance gains of IFL and FSC are often outweighed by their computational cost, underscoring a trade-off between accuracy and efficiency. Finally, the proposed framework demonstrates strong generalization across unseen datasets under a leave-one-dataset-out (LODO) evaluation, wherein training is performed on multiple MRI datasets and testing is conducted on a completely held-out dataset. Additionally, consistent performance across imaging modalities is observed through cross-modality evaluation on CT data, indicating robustness to substantial domain and modality shifts. Detailed cross-dataset and cross-modality experimental results are reported in Sections 1 and 2 of the Supplementary Material.

A key limitation of this study is its focus on a single MRI sequence (FLAIR), whereas fake detection can benefit from leveraging multiple complementary sequences which could further improve the detection sensitivity and specificity.

6. Conclusions

This study presents a framework for detecting fake brain MRI scans through a systematic evaluation of handcrafted, domain-aligned features and deep feature embeddings. Experimental results show that Fourier-domain features and ImageNet-pretrained VGG19 embeddings achieve the strongest performance, including in low-data settings. Overall, the proposed framework establishes a reproducible baseline for medical deepfake detection. The current study focuses on single-slice MRI and does not analyze volumetric or temporal consistency, which are important considerations for medical imaging. Future work may extend this framework by incorporating multi-sequence data, volumetric cues, and feature fusion strategies to enhance robustness. Additional directions include developing detection methods resilient to adaptive generative models, integrating attention mechanisms to localize manipulated regions, and exploring zero- and few-shot approaches for detecting previously unseen manipulations, thereby improving clinical applicability.

CRediT authorship contribution statement

Vaishnavi Ravi: Data Curation, Methodology, Formal Analysis, Validation, Writing – original draft. **Yogesh K. Sahu:** Formal analysis, Investigation. **Prabhas R. Onteru:** Formal analysis, Investigation, Software. **Parag Dutta:** Validation, Investigation, Writing – Review & Editing. **Dhanshree Warokar:** Validation. **Padma Murali:** Supervision, Project administration. **Rajesh Katta:** Project administration. **Ambedkar Dukkipati:** Supervision, Resources. **Phaneendra K. Yalavarthy:** Conceptualization, Methodology, Writing – Review & Editing, Resources, Supervision.

References

[1] Abd Warif, N.B., Idris, M.Y.I., Wahab, A.W.A., Salleh, R., 2015. An evaluation of error level analysis in image forensics, in: 2015 5th IEEE international conference on system engineering and technology (ICSET), IEEE. pp. 23–28.

[2] Agarwal, S., Rattani, A., Chowdary, C.R., 2021. A comparative study on handcrafted features v/s deep features for open-set fingerprint liveness detection. Pattern Recognition Letters 147, 34–40.

[3] Amiri, E., Mosallanejad, A., Sheikhahmadi, A., 2024. Cfdmi-sec: An optimal model for copy-move forgery detection of medical image using sift, eom and chm. Plos one 19, e0303332.

[4] Bakas, S., Sako, C., Akbari, H., Bilello, M., Sotiras, A., Shukla, G., Rudie, J., Flores Santamaria, N., Fathi Kazerooni, A., Pati, S., et al., 2021. Multi-parametric magnetic resonance imaging (mpmri) scans for de novo glioblastoma (gbm) patients from the university of pennsylvania health system (upenn-gbm). The Cancer Imaging Archive .

[5] Bond-Taylor, S., Leach, A., Long, Y., Willcocks, C.G., 2021. Deep generative modelling: A comparative review of vaes, gans, normalizing flows, energy-based and autoregressive models. IEEE transactions on pattern analysis and machine intelligence 44, 7327–7347.

[6] Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T., 2014. Decaf: A deep convolutional activation feature for generic visual recognition, in: International conference on machine learning, PMLR. pp. 647–655.

[7] Ferreira, S., Antunes, M., Correia, M.E., 2021. Exposing manipulated photos and videos in digital forensics analysis. Journal of imaging 7, 102.

[8] Gao, Q., Zhang, B., Wu, J., Luo, W., Teng, Z., Fan, J., 2025. Leveraging facial landmarks improves generalization ability for deepfake detection. Pattern Recognition 164, 111528.

[9] Guyon, I., Weston, J., Barnhill, S., Vapnik, V., 2002. Gene selection for cancer classification using support vector machines. Machine learning 46, 389–422.

[10] Hearst, M.A., Dumais, S.T., Osuna, E., Platt, J., Scholkopf, B., 1998. Support vector machines. IEEE Intelligent Systems and their applications 13, 18–28.

[11] Huang, W., Valsecchi, M., Multerer, M., 2025. Anisotropic multiresolution analyses for deepfake detection. Pattern Recognition 164, 111551.

[12] Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A., 2017. Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1125–1134.

[13] Karaköse, M., Yetış, H., Çeçen, M., 2024. A new approach for effective medical deepfake detection in medical images. IEEE Access .

[14] Leporoni, G., Maiano, L., Papa, L., Amerini, I., 2024. A guided-based approach for deepfake detection: Rgb-depth integration via features fusion. Pattern Recognition Letters 181, 99–105.

[15] Lin, X., Wang, S., Deng, J., Fu, Y., Bai, X., Chen, X., Qu, X., Tang, W., 2023. Image manipulation detection by multiple tampering traces and edge artifact enhancement. Pattern Recognition 133, 109026.

[16] Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. Medical image analysis 42, 60–88.

[17] Mérida, I., Jung, J., Bouvard, S., Le Bars, D., Lancelot, S., Lavenne, F., Bouillot, C., Redouté, J., Hammers, A., Costes, N., 2021. Cermep-idb-mrxfdg: a database of 37 normal adult human brain [18 f] fdg pet, t1 and flair mri, and ct images available for research. EJNMMI research 11, 1–10.

[18] Mirsky, Y., Mahler, T., Shelef, I., Elovici, Y., 2019. {CT-GAN}: Malicious tampering of 3d medical imagery using deep learning, in: 28th USENIX Security Symposium (USENIX Security 19), pp. 461–478.

[19] Narayanan, R., Sundaresan, V., 2025. Medlessynth-ld: Lesion synthesis using physics-based noise models for robust lesion segmentation in low-data medical imaging regimes. Pattern Recognition Letters 188, 155–163.

[20] O'Reilly, J.A., Asadi, F., 2022. Identifying obviously artificial medical images produced by a generative adversarial network, in: 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), IEEE. pp. 430–433.

[21] Pei, G., Zhang, J., Hu, M., Zhang, Z., Wang, C., Wu, Y., Zhai, G., Yang, J., Shen, C., Tao, D., 2024. Deepfake generation and detection: A benchmark and survey. arXiv preprint arXiv:2403.17881 .

[22] Raghu, M., Zhang, C., Kleinberg, J., Bengio, S., 2019. Transfusion: Understanding transfer learning for medical imaging. Advances in neural information processing systems 32.

[23] Raj, S., Mathew, J., Mondal, A., 2024. Generalized and robust model for gan-generated image detection. Pattern Recognition Letters 182, 104–110.

[24] Rana, M.S., Nobi, M.N., Murali, B., Sung, A.H., 2022. Deepfake detection: A systematic literature review. IEEE access 10, 25494–25513.

[25] Reyes, M., Meier, R., Pereira, S., Silva, C.A., Dahlweid, F.M., Tengg-Kobligk, H.v., Summers, R.M., Wiest, R., 2020. On the interpretability of artificial intelligence in radiology: challenges and opportunities. Radiology: artificial intelligence 2, e190043.

[26] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B., 2022. High-resolution image synthesis with latent diffusion models, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 10684–10695.

[27] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C., 2018. Mobilenetv2: Inverted residuals and linear bottlenecks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4510–4520.

[28] Scarpace, L., Mikkelsen, T., Cha, S., Rao, S., Tekchandani, S., Gutman, D., Saltz, J.H., Erickson, B.J., Pedano, N., Flanders, A.E., et al., 2016. The cancer genome atlas glioblastoma multiforme collection (tcga-gbm). The Cancer Imaging Archive .

[29] Sharafudeen, M., Chandra SS, V., 2023. Leveraging vision attention transformers for detection of artificially synthesized dermoscopic lesion deepfakes using derm-cgan. Diagnostics 13, 825.

[30] Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 .

[31] Solaiyappan, S., Wen, Y., 2022. Machine learning based medical image deepfake detection: A comparative study. Machine Learning with Applications 8, 100298.

[32] Tajbakhsh, N., Shin, J.Y., Gurudu, S.R., Hurst, R.T., Kendall, C.B., Gotway, M.B., Liang, J., 2016. Convolutional neural networks for medical image analysis: Full training or fine tuning? IEEE transactions on medical imaging 35, 1299–1312.

[33] Wang, S., Zhu, D., Chen, J., Bi, J., Wang, W., 2024. Deepfake face discrimination based on self-attention mechanism. Pattern Recognition Letters 183, 92–97.

[34] Wang, X., Song, W., Hao, C., Liu, F., 2025. Deepfake detection method based on spatio-temporal information fusion. Computers, Materials & Continua 83.