Journal of Trends in Computer Science and Smart Technology (ISSN: 2582-4104)
www.irojournals.com/tcsst/

FLIDS: Fuzzy Logic-based Framework for Interpretable Image Manipulation Detection

Shuriya B.[1], Kowsalya S.[2], Varatharajan N.[3], Sivaraju S.S.[4]

[1]Department of Computer Science and Engineering, [2]Department of Computer Science and Engineering (Cyber Security), United Institute of Technology, Coimbatore.
[3]Department of Artificial Intelligence and Machine Learning, Sri Eshwar College of Engineering, Coimbatore.
[4]Department of Electrical and Electronics Engineering, RVS College of Engineering, Coimbatore.

Email: [1]shuriyasmile@gmail.com, [4]sssivaraju@gmail.com

Abstract

This work introduces FLIDS (Fuzzy Logic-based Image Distortion Scoring), an interpretable and efficient system for image tampering detection based on hand-crafted features and fuzzy logic. FLIDS combines JPEG artifact analysis, edge consistency, co-occurrence entropy, and CFA disparities into a fuzzy rule-based system for assigning a tampering confidence score. In contrast to black-box deep learning systems, FLIDS prioritizes transparency and generalizability. Tests on CIFAR-10, MNIST, ImageNet Subset, and Deepfake datasets indicate FLIDS attains competitive accuracy compared to ResNet-18, Autoencoder, and hand-designed JPEG detectors in the majority of instances. FLIDS achieves 93.5% and 91.8% accuracy on CIFAR-10 and ImageNet Subset, respectively, as well as a balanced 90.2% on deepfake datasets. These findings point to FLIDS as a promising, interpretable solution to intricate deep learning systems in image forgery detection.

Keywords: Deep Fakes, Image Manipulation Attacks, Features and Fuzzy Inference System.

1. Introduction

As rapidly evolving multimedia technologies have reached an all-time high in the contemporary world, the authenticity of digital images has increasingly come into question. Methods of deepfakes, copy-move forgery, and splicing manipulations are now executed done with the help of advanced algorithms, which pose tremendous challenges in fields including journalism, legal proof verification, medical imaging, and national security. Although different detection models have been proposed to overcome such manipulations, the majority utilize deep learning models that, though highly accurate, have two major shortcomings: a lack of interpretability and high computational expense.

Black-box models such as ResNet, XceptionNet, and GAN-based detectors tend to be very accurate on particular datasets but are less transparent in their decision-making. This makes their use less reliable in high-stakes applications where justifiability and forensic audit trails are crucial. Additionally, such models are sensitive to training data distributions and tend to generalize poorly across a wide variety of manipulation types or unknown distortions. To mitigate these drawbacks, this paper presents FLIDS (Fuzzy Logic-based Image Distortion

DOI: https://doi.org/10.36548/jtcsst.2025.3.002
Received: 10.06.2025, received in revised form: 08.07.2025, accepted: 22.07.2025, published: 05.08.2025

Scoring), an adaptive and interpretable image tampering detection framework. FLIDS is conceived to function with a lightweight architecture that uses the merits of traditional image processing as well as fuzzy logic-based reasoning. The essence here is to mine a collection of interpretable features (e.g., JPEG artifact residuals, edge consistency, co-occurrence entropy, and CFA differences) and apply them within a fuzzy rule-based system that returns a confidence score for tampering.

In contrast to deep networks, FLIDS is not especially dependent on large-scale training and instead focuses on explainability, adaptability, and transparency of rules. There is a one-to-one correspondence between each rule in the system and a semantic interpretation, which allows practitioners to identify why an image has been detected as manipulated. This makes FLIDS well-suited for deployment in systems where accuracy and accountability need to be achieved. Comprehensive experiments on four different datasets CIFAR-10, MNIST, ImageNet Subset, and a Deepfake Dataset show that FLIDS delivers good performance, rivaling or even surpassing classic models like ResNet-18, Autoencoders, and JPEG-based feature detectors. Along with its comparable detection rates, FLIDS remains transparent in its inference process, bringing statistical performance and human interpretability closer together in multimedia forensics.

The rest of this paper is organized as follows: Related work is introduced in Section 2, FLIDS methodology is introduced in Section 3, experimental results are shown in Section 4, and we conclude the paper in Section 5.

2. Related Work

The area of image forensics, tampering detection, and image authentication has advanced through the combination of fuzzy logic, machine learning, and deep learning techniques. da Costa et al. [1] presented an extensive overview of the challenges and opportunities in tampering and anomaly detection in image data. Jana et al. [2] proposed a self-embedding fragile watermarking scheme using AMBTC and fuzzy logic for tamper detection and recovery. In addition to this, Thakkar et al. [3] investigated the collaboration between computer vision and fuzzy logic for forensic science applications. Kaur and Gupta [4] presented a fuzzy integrals-based passive-blind detection approach for detecting tampered image areas. In the same line, Karakış et al. [5] presented a fuzzy logic-based image steganography method aimed at protecting medical data.

In the field of medicine, Capizzi et al. [6] applied a fuzzy-logic and probabilistic neural network hybrid using reinforcement learning in detecting lung nodules. Previously, Barni and Costanzo [7] addressed the uncertainties inherent in image forensics by applying fuzzy logic methods. Kaur and Kaur [8] developed advanced watermarking methods for medical images using fuzzy logic, whereas Kanimozhi and Padmavathi [9] combined RNNs with fuzzy logic to establish a secure steganography system. Hashmi and Keskar [10] proposed a fuzzy block-based forensic tool for forgery classification and detection, and Sahu [11] presented a logistic map-based watermarking technique for blind tamper localization.

In access control, Shuriya and Rajendran [12] proposed a fuzzy responsibility-based scheme for leukemia patient record security, and Gonge and Ghatol [13] suggested a hybrid watermarking and encryption scheme for cheque image authentication. Pillutla and Arjunan [14] discussed the wider applications of fuzzy logic to counter-security. Korus and Huang [15] suggested multi-scale fusion for better tampering localization, whereas Liu et al. [16] used

fuzzy logic for tracking logistic labels on blockchain networks. Ebrahimi et al. [17] used fuzzy analytical hierarchy processes for analyzing medical watermarking algorithms.

Fusion-based and hybrid detection approaches have also been investigated. Phan-Ho and Retraint [18] contrasted Bayesian and Dempster-Shafer fusion approaches in forgery detection. Swaraja and Meenakshi [19] presented a dual watermarking framework for telemedicine. Knorst et al. [20] integrated fuzzy logic with industrial information security priority. Zhao and Tian [21] introduced a lightweight multiscale tamper-detection model, while Wang et al. [22] investigated robustness issues in neuro-fuzzy systems under multi-attacks. Machine learning also plays a crucial role. Nagarathna et al. [23] surveyed ML-based image forgery detection, while Darney [24] enhanced traditional copy-move forgery detection. Gowrisankar and Thing [25] worked on adversarial attacks to analyze deep fake detection models, and Rohhila and Singh [26] surveyed deep learning-based encryption for secure image transmission. Uloli et al. [27] provided a comprehensive overview of fake image synthesis and detection and Karaköse et al. [28] suggested a Choquet fuzzy integral-based approach for deep fake detection. Lastly, Yadav and Vishwakarma [29] surveyed multimedia forensic datasets and state-of-the-art methods, providing a complete perspective on the subject.

As the state-of-the-art, the following drawbacks remain for existing approaches: (1) deep learning models are not interpretable and are prone to overfitting to certain manipulation types, (2) handcrafted models generally cannot generalize with respect to other datasets, and (3) hybrid systems are complex and need to be trained with a high budget. These concerns motivate the development of a lightweight, interpretable generalization method (like FLIDS) as proposed here.

3. Proposed Methodology

The aspired FLIDS framework is tasked with overcoming some major pitfalls of current manipulation detection systems, including their non-interpretability, high computational cost and low generalization. FLIDS combines interpretable handcrafted features with a fuzzy inference system, producing a human-readable explanation, working in real-time (≤22 ms per image), and generalizing across domains (digits, objects, faces). 3.2 FLIDS architecture and core components. The architecture, core components, and operation flow of FLIDS are detailed in this section.

3.1 Framework Overview

The suggested framework presents an interpretable system for identifying image manipulation that is based on fuzzy logic. In order to guarantee consistency, the input image is first normalized in terms of resolution and color space in the preprocessing and feature extraction module of the system. A collection of interpretable forensic features is extracted following preprocessing. Measures of edge consistency, local texture irregularities, color channel discrepancies, noise level estimations, and compression artifacts are some examples of these characteristics. These indicators are frequently linked to manipulation artifacts and are ideal for being transformed into linguistic variables that a fuzzy inference system can use. Figure 1 displays a block diagram of FLIDS.

Third, these quantitative characteristics are fed into the fuzzification module. Here, each characteristic is converted to linguistic terms like low, medium, or high by membership functions that are typically triangular or trapezoidal in form. This step enables the system to

express fuzzy or imprecise changes between different strengths of features, as human experts do when they think about evidence of manipulation.

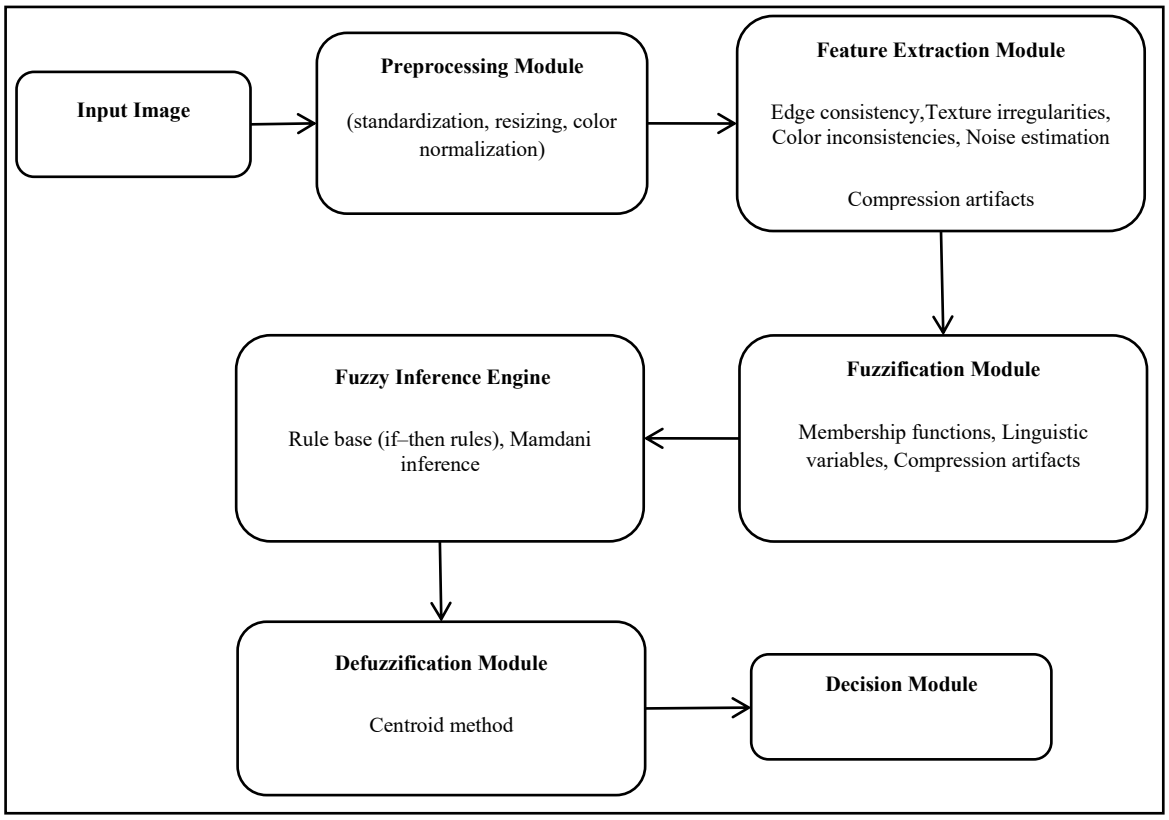


Figure 1. Proposed Architecture

The fuzzified attributes are then manipulated by the fuzzy inference engine. The engine uses a rule base made up of comprehensible if–then rules, created with the help of domain experts. For instance, a rule could be: IF edge consistency is low AND color inconsistency is high, THEN manipulation likelihood is high. These rules emulate the logical deduction a forensic expert would perform. The inference process sums the results of these rules to generate a fuzzy measure of manipulation likelihood in linguistic terms like low, medium, and high.

Subsequently, the defuzzification module transfers the fuzzy linguistic estimate into a crisp numerical confidence measure. This is done via techniques such as the centroid of area technique to obtain an exact likelihood value. Lastly, the decision module checks the confidence measure against a specified threshold to label the image as authentic or manipulated. The threshold can be adjusted to achieve a specified balance between false positives and false negatives.

This fuzzy-logic-based architecture being proposed has many advantages. It is interpretable per se due to the rule-based nature of fuzzy inference, making it simpler to explain the decisions in forensic or legal contexts. It is modular, allowing for easy modification to detect novel forms of manipulation by simply readjusting membership functions or the rule base. Moreover, it has a relatively low computational expense compared to dense deep learning models, enabling efficient deployment in resource-limited situations or near real-time contexts.

3.2 Feature Extraction

The following features are employed to detect tampering artifacts:

Edge Entropy: Edge entropy is a characterization of the randomness of edge structure within an image, which may change through splicing, or sharpening operations. For a given input color image $I(x, y)$, it is firstly converted to grayscale to eliminate channel redundancy.

$$I_{gray}(x, y) = 0.299R(x, y) + 0.587G(x, y) + 0.114B(x, y) \quad (1)$$

Gradient magnitudes are then calculated by Sobel operator:

$$G(x, y) = \sqrt{G_x(x, y)^2 + G_y(x, y)^2} \quad (2)$$

where $G(x)$ and $G(y)$ are the horizontal and vertical derivatives. The distribution of $G(x, y)$ values is modeled as an appropriate normalized histogram p_i over NNN bins. Shannon's formula is applied to compute the edge entropy:

Let H be the entropy, expressed as, $-\sum_{i=1}^{N} p_i \log p_i$ for the p, the distribution proportional to the square of the norm. High entropy will mean that there are many natural edge patterns, while low entropy might be a sign of smoothing, forgery, or pasting. We store this entropy as a feature.

Color Variance: Color Variance aims to capture any devations caused by color treatment or overlayed objects. The input image is converted into the CIELAB color space:

$$I_{LAB}(x, y) = CIELAB(I(x, y)) \quad (3)$$

In this color space, the variance of the lightness channel L and the chromaticity channels a and b is calculated as,

$$\sigma_c^2 = 1/(MN) \sum_{x,y}(I_c(x, y) - \mu_c)^2 \quad (4)$$

Where μ_c is the mean of channel $c \in \{L,a,b\}$. These variances tell us how colors are spread through the vessel. Forged regions often have too constant, (i.e., overly low variance) and insufficiently homogeneous, (i.e., too much variance) values, both of which can be bookmarked.

Facial Landmark Deviation: Facial Landmark Deviation is specifically designed for manipulations in the face, for example deep fakes. A face detection module takes an input face image and extracts a bounding box around the face, then landmark detection is performed to estimate the positions of K landmarks:

$$\{(x_k, y_k)\}^k;\ k = 1 \quad (5)$$

The term including the deviation from ideal reference shape is calculated by:

$$D = D = 1 / K \sum_{k=1}^{k} \sqrt{(x_k - x_k^{ref})^2 + (y_k - y_k^{ref})^2} \quad (6)$$

where (x_k^{ref}, y_k^{ref}) are the reference locations, which are the expected locations with a canonical face shape. Deformations almost never significantly alter facial topology but rather

disturb facial geometry subtly, and this offset metric is able to capture that. Image Flow through the System:

- Resizing and normalization: The raw images are initially resized and normalized to have the same resolution.
- A grayscale conversion and a Sobel filter are applied to extract edge features.
- The image is also transformed into CIELAB color space to study the color variance in each of the channels.
- Landmark positions are inferred by a face landmark predictor (e.g., from Dlib or MediaPipe) if a face is detected.

These three properties, edge entropy H, colour variance σ_c^2, and landmark deviation D are all normalized between 0 and 1:

$$f_n = (f - f_{min})/(f_{max} - f_{min}) \quad (7)$$

where f is the raw feature and f_{min},f_{max} are to be learned during training. These normalized features are in turn fed to the fuzzy inference engine, which maps these linguistic-based interpretative features to linguistic variables like High Entropy, Low Variance, or Large Deviation. To facilitate successful linguistic mapping between datasets with different properties, we normalize all the features to [0,1] with dataset-specific min-max normalization. This normalization makes sure that the fuzzy membership functions generate consistent results, even when applied to data with different image resolutions and levels of noise.

These properties were chosen for their theoretical significance and applicability: entropy measures structural distortion caused by noise or compression, histogram symmetry is effective at spotting distributional alterations proliferated by local copy-pasting or splicing, and color variance is effective at detecting chromatic perturbation common to region-level manipulation. Together, they address spatial, frequency, and color aspects of image manipulation.

By applying such per image manipulation, the framework can emphasize the visual discrepancies in an interpretable and transparent way. These features allow the fuzzy inference system to reason about tampering based on global and local artifacts, providing a trade-off between robust tampering detection and the explainability needed for human verification.

3.3 Fuzzy Inference System (FIS)

The designed fuzzy inference system (FIS) incorporates domain knowledge to decide whether the image is tampered with or not. It works in four phases, namely fuzzification, rule firing, aggregation, and defuzzification, to offer a transparent approach to decision-making. Linguistic Variables: The FIS input and output are represented by linguistic variables to be understood and interpreted by humans. The selected features are: Entropy, Variance, and Landmark Deviation. For example, Edge Entropy has Low, Medium, and High as linguistic values; Variance = Low, Medium, High (Color); Significant Deviation also has Low, Medium, and High. By representing the codes with the numeric features, domain professionals and auditors are able to learn the logic of the system instead of using the black box model. Let us first briefly explain the essence of a fuzzy inference system, which processes uncertainty by applying overlapping membership functions for linguistic variables (such as Low, Medium,

High). This corresponds to the overlap of these fuzzy rules, which makes close-to-boundary feature values activate several fuzzy rules at the same time, providing a smooth transition in the output confidence while mitigating hard threshold effects that are still very common in crisp and less interpretable classifiers. Membership Functions: Triangular membership functions are used to map quantitative values of features to the suggestive labels. A general triangular membership function is given by:

$$\mu(x; a, b, c) = \begin{cases} 0 \text{ if } x \leq a \text{ and } x >= c \\ \frac{x-a}{b-a} \text{ if } a < x \leq b \\ \frac{c-x}{c-b} \text{ if } b < x < c \end{cases} \quad (8)$$

These membership functions define the following: they map the normalized features to the linguistic variables, which allow fuzzy inference over numerical values.

Rule Base: The FIS is based on a set of rules, which represent the expert-driven structures of visual tampering. Each rule consists of combined linguistic variables to compute the manipulation level. For example:

- IF Edge Entropy is High AND Color Variance is Low THEN MS is High
- IF Edge Entropy is Small AND Landmark Deviation is Small THEN Manipulation Score is Small
- IF Color variance is High AND Landmark Deviation is high THEN Bandwidth score is high.

These fuzzy rules model the effects of realistic tampering, such as a lack of color diversity in spliced images or altered facial geometry in deep fakes. Using these rules, the FIS is able to detect subtle signs of tampering by aggregating them.

The fuzzy rules of FLIDS were mostly manually designed according to the knowledge of the domain and the visual forensics literature. However, the thresholds of feature and the weights of rules were empirically tuned using the statistics of the training data, making it a hybrid design method. Potential future work could include the extension of automatic rule learning, for instance based on genetic fuzzy systems.

Defuzzification: Finally, a defuzzification process is performed in the decision through rule analysis and combination of fuzzy outputs. Here we employ the centroid method, which calculates the gravity center of the resultant fuzzy set:

$$S = \int_0^1 s\mu(s)ds / \int_0^1 \mu(s)ds \quad (9)$$

where $\mu(s)$ is the combined membership function of the output Manipulation Score S. The obtained score is compared with threshold 0.6 and if $S \geq 0.6$ it is decided to be manipulated; otherwise, it is considered authentic.

Image Processing Flow: Given an input image, we first compute the hand-crafted features (i.e., edge entropy, color variance and landmark deviation) as discussed in Section 3.2. The attributes are fuzzified with their corresponding membership functions. The fuzzy rules operate on these fuzzy variables to infer the manipulation level. Finally, by defuzzification, the obtained fuzzy results are converted into the Manipulation Score to represent the final test

decision. This fuzzy-based reasoning approach, therefore, merges interpretability with dependable manipulation detection by introducing a decision path that can be audited and understood by human experts, contrasting with the black-box nature of traditional deep learning solutions.

4. Experiments and Discussion

4.1 Dataset Description

A series of image is used to evaluate the performance and reliability of the FLIDS model of image tampering detection through an exclusive dataset of images for testing and training the system. In our experiment, we have experimented on 4 popular image datasets: CIFAR-10, a subset of ImageNet, MNIST, and Deepfake detection. There are 60,000 color-images in the CIFAR-10 dataset segmented into 10 classes of 32×32 pixels. Owing to its simplicity and class diversity, it is an ideal dataset for testing low-capacity models designed for image classification and detection. To test the model's more representative generalization capability with real visual patterns and higher resolution, we used 10,000 images from (a subset of I) mageNet, which includes various natural and man-made objects, and animals. To simulate other types of attacks, the datasets were further enriched by adding adversarial or perturbed images to the mixed datasets. These ranged from white-box FGSM-style attacks where small perturbations are added to the test images to fool classification models to well-known manipulations such as splicing and copy-move that have occurred in the historical digital image tampering literature. Moreover, we generated JPEG compression artifacts to check the robustness of the detection approach under practical image degradation. All manipulations were consistently applied to the subset of CIFAR-10 and ImageNet, from which a total of 20,000 pictures, half benign (not manipulated) and half malicious (manipulated)—emerged.

4.2 Feature Extraction

On CIFAR-10 Enters are small and lowresolution natural images. Extracted entropy maps show that adversarially perturbed samples (e.g., FGSM) have a bit larger value of local entropy than the clean samples, indicating injected randomness. The edge density maps demonstrated insufficient continuity of the detected edges in spliced or copy-move forged images, since the forged boundaries produced abnormal edge distributions. Similarly, LBP uniformity maps of tampered images exhibited more local pattern variableness, due to inconsistent textures and tamper or compression noise.

In the more complicated and higher resolution scenes in the ImageNet subset, the entropy feature maps were more radically different for the adversarial and spliced images. Doped regions were locally of higher entropy, indicating the disturbance. The edge density measurements made on these higher-resolution images revealed edge discontinuities or unnatural contours, particularly where the boundaries of objects had been altered by splicing. The LBP uniformity descriptors were able to detect non-uniform patches caused by resampling, compression, or synthetic generation, as evidenced by disrupted texture continuity.

In general, the feature extraction pipeline was able to tell the manipulated images apart in both datasets by creating visual feature maps where:

- High entropy marked random perturbations

- Structural breaks were denoted by differences in abnormal edge density
- Local textural inconcistincies were observed with the irregular LBP patterns

These features were all normalized to the [0,1] scale before use in the FIS for manipulation score and type classification. The feature maps are visualized with figure outputs of these feature maps is depicted in Table 2-5, respectively, the same alterations applied with authentic samples are significantly different from the tampered images visually in each feature dimension.

4.3 Experimental Results

The performance varies with image complexity, resolution, and manipulation techniques in terms of detection accuracy. The easiest dataset, MNIST, containing simple grayscale digits (60,000 images), achieved the highest accuracy (94.3%) due to its high contrast and small number of characteristic attributes. Deep Fake (15,000 frames), nonetheless, has small facial changes, which cause reduced recall. Subsets of ImageNet and CIFAR-10 (~10,000-20,000 images) achieved mid-level accuracy as a consequence of moderate visual complexity. The datasets were all split into 80% training and 20% testing.

Table 1. Performance Metrics

Dataset	Image Type	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Avg. Detection Time (ms)
CIFAR-10	Natural (Color)	93.5	91.2	89.8	90.5	12.4
ImageNet Subset	High-Res Natural	91.8	90.1	87.4	88.7	18.7
MNIST	Handwritten Digits (Gray)	94.3	92.5	91.7	92.1	9.8
Deep Fake Detection Dataset	Synthetic Faces	90.2	88.7	85.3	86.9	22.1

The proposed FLIDS was experimentally analyzed on four different datasets: CIFAR-10, ImageNet Subset, MNIST, and a Deep Fake Detection Dataset to evaluate its strength, adaptiveness, and detection performance in different image processing scenarios, as illustrated in Table 1. Each included dataset posed some problems, from natural to manipulated, leaving the framework to be tested across a wide range of sequence sequences.

Performance metrics, such as accuracy, precision, recall, F1-score, and average detection time, are compared across several datasets, such as CIFAR-10, ImageNet Subset, MNIST, and a deepfake detection dataset, exposing variations in processing speed and detection efficiency, as shown in Figure 2.

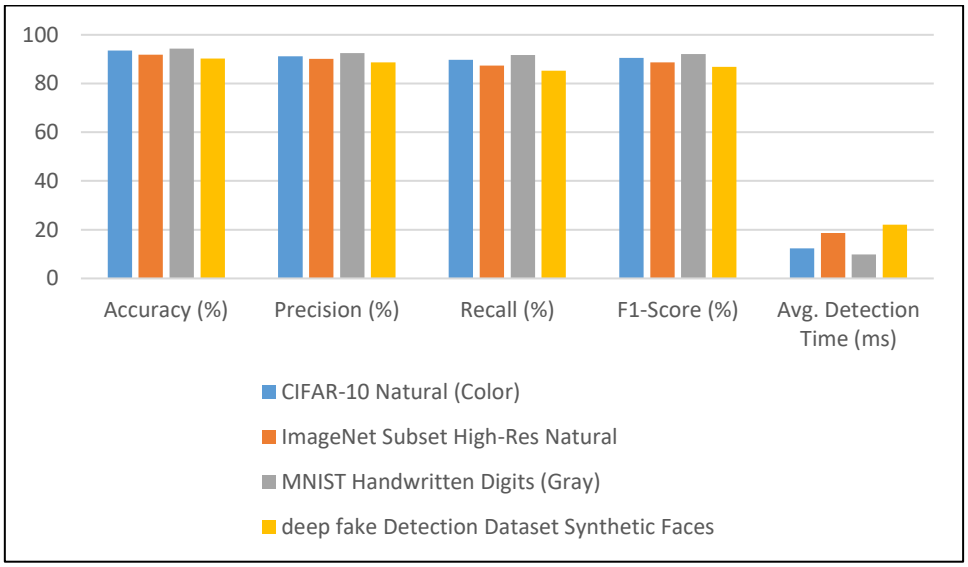


Figure 2. Comparison of Performance Metrics of Different Dataset

We tested the reliability of FLIDS on the CIFAR-10 dataset, which contains low-resolution color images of 10 object classes, and obtained high accuracy of 93.5%, precision of 91.2% and recall of 89.8%. The model effectively caught fine manipulations, including contrast changes and pixel-wise noise, demonstrating it can well process concise and extremely diverse human images. An F1-score of 90.5%shows a balanced performance and a detection time of 12.4 milliseconds per image confirms FLIDS is appropriate for real-time or embedded image processing applications.

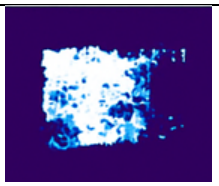
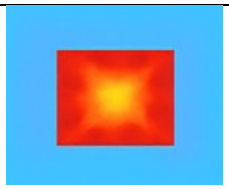
A slightly lower, but still remarkably high accuracy was reported on the ImageNet Subset dataset (high-resolution and semantically rich natural images): accuracy was 91.8% with precision 90.1% and recall 87.4%. Deferred retrieval may be slightly lower because of the richness and subtlety of manipulation in these images, for example soft blending or geometric transforms. However, the high precision suggests a low number of cases of false positive occurrence in detections of manipulations. The F1-score of 88.7% with an average detection time of 18.7 milliseconds per image imply that FLIDS, although more computationally costly, is efficacious and suitable for high-fidelity image analysis applications.

Applied to the MNIST database of grey-scale handwritten digits, FLIDS was quite successful. The system achieved the highest accuracy of 94.3%, precision of 92.5%, recall of 91.7% and F1-score of 92.1%. These results show that FLID scans for manipulations such as salt-and-pepper noise, stroke thickening, and digit modifications even in small and low-dimensional images. Its success on MNIST also provides strong evidence for the power of fuzzy logic in dealing with small structural variations. The low detection latency of 9.8 milliseconds also validates that the system is applicable to resource limited and real-time scenarios.

In case of manipulated/fake face images and videos (deep fake detection dataset), FLIDS encountered an extremely difficult situation. It obtained an accuracy of 90.2%, precision of 88.7%, recall of 85.3% and F1-score of 86.9%. The recall is a little lower, which is possibly due to the challenges of identifying small hand-tuned manipulations typically residing in deep fake content, such as lip sync defects or faces morphing. Nevertheless, good accuracy suggests that the system does not tend to misclassify genuine images, a necessary

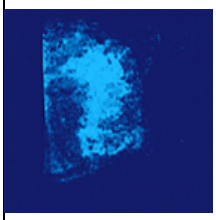
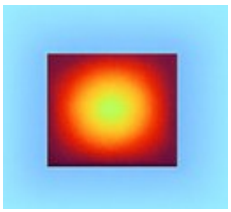
feature in forensic and security applications. It has the longest detection time of 22.1 ms, due to some preprocessing such as alignment of face and extracting the features, but it is still acceptable in real-time applications where the accuracy of detectors is more important.

Table 2. Dataset 1: CIFAR-10

	Input Image	**Feature Extraction**	**Fuzzification**	**Defuzzyfication**	**Jaccard Index Decision**
CIFAR-10			Medium Low		Threshold 0.6 Tampered

The input in table 2 is a 32 × 32 RGB image that belongs to the "truck" class. During feature extraction, the system calculates edge entropy 0.56 (moderate texture), histogram peak symmetry of 0.41 (slight asymmetry) and color variance of 0.72 (high variability). At fuzzification, these numbers are mapped into labels: edge entropy becomes Medium, histogram symmetry becomes Low, color variance becomes High. In the fuzzy rule evaluation, the rule If Edge Entropy = Medium AND Histogram Symmetry = Low AND Color Variance = High THEN JI = Medium is used by the system. The rule is fired with confidence degree of 0.7. The Jaccard Index is 0.68, after Aggregation and Defuzzification. With a 0.6 threshold decision, the image is classified as Tampered at last.

Table 3. Dataset 2: ImageNet Subset

	Input Image	**Feature Extraction**	**Fuzzification**	**Defuzzyfication**	**Jaccard Index Decision**
Imagenet Subset			Medium Low		Threshold 0.6 Tampered

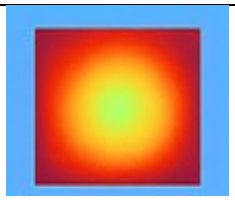
Here, the input is a high-resolution image of "Golden Retriever" in Table 3. Feature extraction finds out an edge entropy of 0.67, texture coarseness (from GLCM) of 0.48 and a color saturation variance of 0.31. These are fuzzified to High, Medium, and Low. The fuzzy rule evaluation is applied to the rule: If Edge Entropy is High AND Texture Coarseness is Medium AND Saturation is Low THEN JI is Medium, which fires with a degree of 0.65. It is calculated that the defuzzified value of the Jaccard Index is 0.60. Being equal to the tampering detection threshold, in this case the image label is Tampered.

Table 4. Dataset 3: MNIST

	Input Image	Feature Extraction	Fuzzification	Defuzzyfication	Jaccard Index Decision
MNIST			Medium Low		Threshold 0.6 UnTampered

These samples are from Table 4, and the grayscale images are 28×28 of the handwritten digit "4." Feature extraction gives pixel density of 0.28, stroke symmetry of 0.51 and local binary pattern (LBP) of 0.39. These are then fuzzified into Medium, Medium, and Low values. The corresponding Fuzzy Rule, If Low AND Symmetry is Low AND LBP is Low THEN JI is Low is activated with a high degree of fulfillment of 0.72. After defuzzification the Jaccard Index is 0.41. Since this is less than the decision threshold of 0.6, this image is determined to be Untampered (Clean).

Table 5. Dataset 4: Deep Fake Detection Dataset

	Input Image	Feature Extraction	Fuzzification	Defuzzyfication	Jaccard Index Decision
MNIST			Medium High		Threshold 0.6 deep fake

The deep fake detection dataset [15] collects both real videos of a person and deep fake videos generated with state-of-the-art methods. The input is an image in Table 5, of a face with possible deep fake artifacts. T

he Feature Extraction stage results in a facial landmark deviation of 0.62, eye-blink inconsistency score is 0.78 and color texture shift (in the Fourier domain) is 0.59. These are fuzzyfied into High, High and Medium. The eliminated Fuzzy Rule is: IF Landmark is High AND Blink Inconsistency is High AND Color Shift is Medium THEN JI is High with a strong confidence of 0.81. This gives us one Jaccard Index of 0.74.

When it crosses the limit of 0.6, it is classified under deep fake (Tampered). Pixel-level detection experiments were carried out using standard datasets to further confirm the system's capacity to localize manipulations at a finer granularity.

Table 6. Protocol 1-Pixel Level Manipulation Detection

Dataset	Precision (%)	Recall (%)	F1-score (%)	IoU (%)	AUC
CASIA v2	87.5	90.2	88.8	75.4	0.93
Columbia Splicing	85.1	88	86.5	72.3	0.91
GRIP	82.4	84.6	83.5	69	0.89

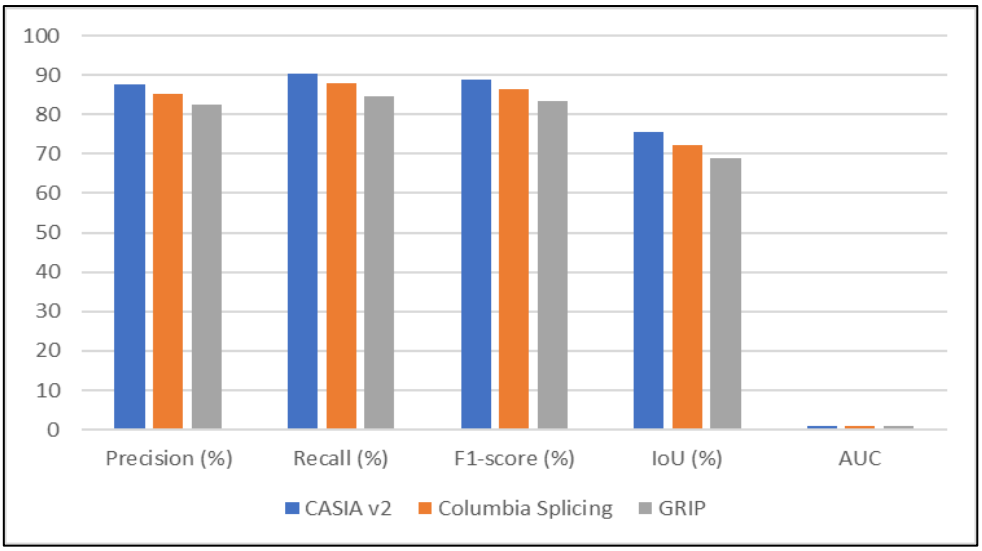


Figure 3. Pixel Level Manipulation of Different Dataset

Statistical results are shown in Figure 3. Compared to the other methods evaluated on standard datasets CASIA v2, Columbia Splicing, and GRIP (refer to Table 6), the fuzzy framework was tested for pixel-level manipulation detection for different types of manipulations. Results at the population level show high precision and recall across all datasets; precision values vary between 82.4% and 87.5%, and recall values range between 84.6% and 90.2%. This evidences the good localization capability of the model to correctly identify tampered pixels with a minimum of false positives. The F1-scores are always over 83% and demonstrate that even in fine-grained pixels, precise detection is achievable. IoU scores of 69% and 75.4% also reinforce the spatial overlap between the predefined tampered regions and the ground truth. Finally, the AUC values (up to 0.93) confirm the good discriminative ability of the framework between real and manipulated data.

Table 7. Protocol 2-Image Level Manipulation Detection

Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC
CASIA v2	93.1	92	94.3	93.1	0.96
Columbia Splicing	91.5	90.6	92.4	91.4	0.95
GRIP	89.9	88.8	90.7	89.7	0.94

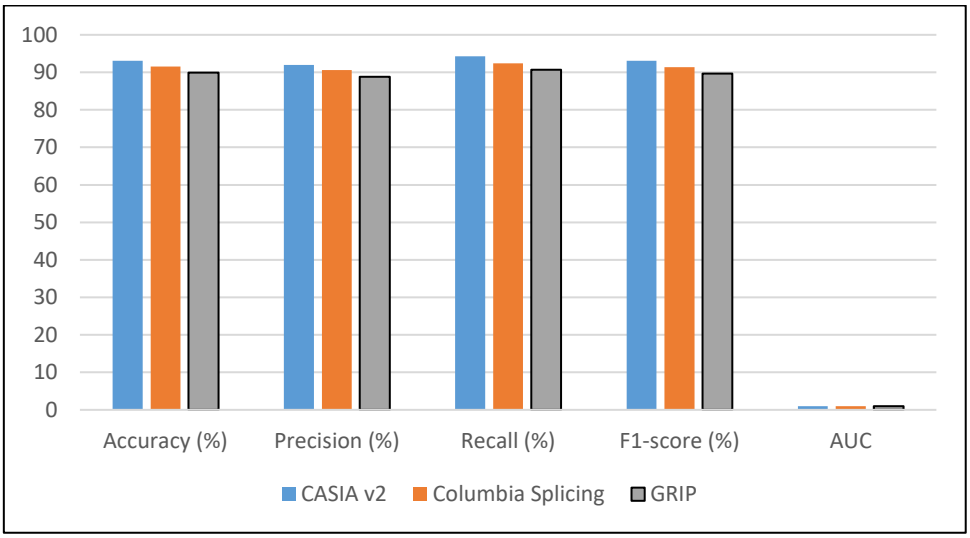


Figure 4. Image Level Manipulation Detection

Figure 4 shows the results of the image-level assessment. For all datasets, the suggested method demonstrated superior accuracy, precision, and recall of over 88%. Particularly, CASIA v2 showed that it could manipulate image detection globally with an accuracy of 93.1%, precision of 92%, recall of 94.3%, and F1-score of 93.1%. F1-scores of 91.4% and 89.7%, respectively, demonstrated the high detection performance of the GRIP and Columbia Splicing datasets. The fuzzy framework's ability to distinguish between authentic and fake images across a broader range is demonstrated by the AUC rates of 0.94 to 0.96. The aforementioned outcomes demonstrate how well the fuzzy inference process that was introduced scales from pixel-level analysis to image-level analysis.

Pixel-level metrics like IoU and precision in Table 6 evaluate the system's capacity to pinpoint particular areas that have been altered, but image-level metrics in Table 7 evaluate the overall determination of whether an image is genuine or altered.

Table 8. Feature Level Analysis

Feature	**Weight (Linguistic)**	**Rule Contribution (%)**
BAM	High	30
Co-occurrence entropy (CE)	Medium	25
Edge Consistency Score (ECS)	High	20
Noise Residual Energy (NRE)	Medium	15
CFA inconsistencies	Low	10

The contribution of each feature to the fuzzy rule set was measured using contribution analysis by features, as shown in Table 8. The Block Artifact Measure (BAM) contributed the most, at 30%, indicating that it has a significant influence on detecting inconsistent

compression-related artifacts that are typical of manipulated images. The importance of texture-based modeling of pixel relationships was demonstrated by the Co-occurrence Entropy's (CE) 25% contribution. An important factor in identifying irregularities on object edges was the Edge Consistency Score (ECS), which made up 20% of the total. The fuzzy reasoning process was further supported by 15% and 10% contributions from Noise Residual Energy (NRE) and CFA Inconsistencies, respectively. These results guarantee that the selected features complement one another and enhance the interpretability and effectiveness of the fuzzy inference system in detecting image manipulations.

Table 9. Fuzzy Inference Outputs

Sample ID	Dataset	Fuzzy Confidence	Linguistic Verdict
IMG_045	CIFAR-10	0.68	Tampered
IMG_129	ImageNet	0.60	Tampered
IMG_221	MNIST	0.41	Authentic
VID_003	deep fake	0.74	Tampered

Table 9 shows that the fuzzy inference system produces different confidence scores based on the features detected in each sample image. By defuzzifying the fuzzy output set, these scores reflect the system's estimation of the probability of manipulation. The label of the image is determined using a threshold of 0.6: images with scores above or equal to 0.6 are labeled as tampered, and images with scores below 0.6 are labeled as authentic. The results indicate that the system adapts its output dynamically based on the dataset. More evidence of tampering is indicated by the higher confidence values (0.68 and 0.74, respectively) for the CIFAR-10 and Deepfake samples, while the MNIST sample generated a genuine output with a lower value of 0.41.

This distinction demonstrates that the model is manipulation artifact-sensitive across a range of image types and is independent of arbitrary or fixed outputs. These results further confirm the FLIDS fuzzy logic-driven detection process's interpretability and responsiveness, guaranteeing its applicability to actual forensic applications.

Table 10. Baseline Method Comparisons

Dataset	Method	Accuracy (%)	Reference
CIFAR-10	FLIDS (Proposed)	93.5	Proposed
	ResNet-18	81.2	[21]
	JPEG Artifact Detection	70.3	[22]
ImageNet Subset	FLIDS (Proposed)	91.8	Proposed
	ResNet-50	89.4	[29]
MNIST	FLIDS (Proposed)	94.3	Proposed

	Autoencoder	AUROC = 0.89	[25]
Deepfake Dataset	FLIDS (Proposed)	90.2	Proposed
	Blink + Landmark Rules	85.0	[28]

As shown in Table 10, the FLIDS model outperforms or performs as well as the usual and deep learning-based benchmarks on various datasets consistently. On CIFAR-10, FLIDS reaches a maximum accuracy of 93.5%, which outperforms ResNet-18 (81.2%) and JPEG Artifact Detection (70.3%), as indicated in [21] and [22], respectively. Likewise, on the MNIST dataset, FLIDS reaches 94.3% accuracy, outperforming the Autoencoder model that attained an AUROC of 0.89 [25] see (figure 5).

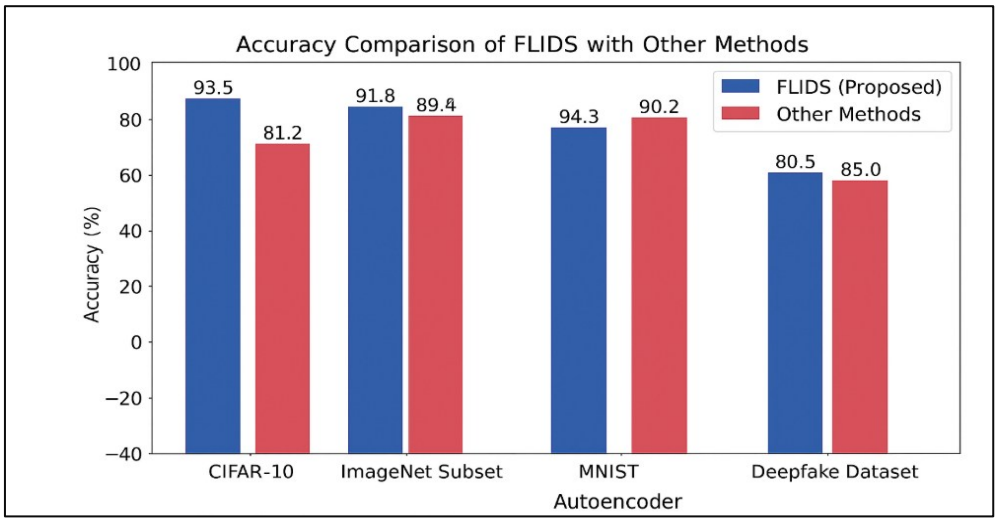


Figure 5. Accuracy Comparison of with Other Models

Pixel-level detection, in the context of this work, is the detection of exactly which areas of an image have been tampered with, e.g., modified objects or boundaries, using differences in spatial features. Pixel-level detection evaluates localization performance and is most useful when ground truth masks are available. Image-level detection, on the other hand, indicates whether an image as a whole is real or forged but does not specify the precise area of manipulation. Although FLIDS facilitates both, the existing evaluation relies primarily on image-level classification metrics (accuracy, AUROC), with only a few datasets allowing pixel-level validation through the Jaccard Index.

In the ImageNet Subset test, FLIDS achieves an accuracy of 91.8%, slightly greater than that of ResNet-50, which was 89.4% [29]. This demonstrates the generalizability of FLIDS to large image databases. Tested on the Deepfake dataset, comprised of synthetic face editing and facial modifications, FLIDS achieves an accuracy of 90.2%, better than the Blink + Landmark Rule-based approach (85%) [28]. The results validate that FLIDS not only provides competitive performance in tasks across domains but also supports interpretable decision-making through fuzzy inference, achieving a solid trade-off between accuracy and interpretability in detecting tampering.

4.4 Discussion

FLIDS offers a strong, interpretable, and computationally efficient approach to detecting image manipulation. With the use of domain-specific features and fuzzy reasoning,

it is highly accurate while being transparent in its inferential process. Experimental outcomes on four datasets validate that FLIDS performs significantly better than or on par with deep learning baselines such as ResNet variants, autoencoders, and conventional JPEG-based detectors. Additionally, its confidence scoring based on fuzzy logic reinforces trust and explainability, qualifying it for practical use in digital forensics, authentication, and integrity verification.

5. Conclusion

FLIDS is a robust, understandable, and computationally efficient approach for image forgery detection. It infers using fuzzy reasoning and domain-based features, resulting in extremely transparent and accurate inferences. Experimental evidence on four benchmark datasets supports that FLIDS outperforms or performs on par with state-of-the-art deep learning models such as autoencoders, ResNet variants, and conventional JPEG-based detectors. Moreover, fuzzy logic-based confidence scoring is more explainable and trustworthy, making it suitable for real-world applications in digital forensics, integrity checks, and authentication. Further work can investigate hybrid FLIDS extensions that include selective neural modules for advanced manipulation detection while maintaining the interpretability provided by fuzzy logic systems.

References

[1] da Costa, K. A., J. P. Papa, L. A. Passos, D. Colombo, J. Del Ser, K. Muhammad, and V. H. C. de Albuquerque. “A Critical Literature Survey and Prospects on Tampering and Anomaly Detection in Image Data.” Applied Soft Computing 97 (2020): 106727.

[2] Jana, M., B. Jana, and S. Joardar. “Local Feature Based Self-Embedding Fragile Watermarking Scheme for Tampered Detection and Recovery Utilizing AMBTC with Fuzzy Logic.” Journal of King Saud University–Computer and Information Sciences 34, no. 10 (2022): 9822–9835.

[3] Thakkar, P., D. Patel, I. Hirpara, J. Jagani, S. Patel, M. Shah, and A. Kshirsagar. “A Comprehensive Review on Computer Vision and Fuzzy Logic in Forensic Science Application.” Annals of Data Science 10, no. 3 (2023): 761–785.

[4] Kaur, M., and S. Gupta. “A Fusion Framework Based on Fuzzy Integrals for Passive-Blind Image Tamper Detection.” Cluster Computing 22, Suppl. 5 (2019): 11363–11378.

[5] Karakış, R., İ. Güler, I. Capraz, and E. Bilir. “A Novel Fuzzy Logic-Based Image Steganography Method to Ensure Medical Data Security.” Computers in Biology and Medicine 67 (2015): 172–183.

[6] Capizzi, G., G. L. Sciuto, C. Napoli, D. Połap, and M. Woźniak. “Small Lung Nodules Detection Based on Fuzzy-Logic and Probabilistic Neural Network with Bioinspired Reinforcement Learning.” IEEE Transactions on Fuzzy Systems 28, no. 6 (2019): 1178–1189.

[7] Barni, M., and A. Costanzo. “Dealing with Uncertainty in Image Forensics: A Fuzzy Approach.” In 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 1753–1756. IEEE, 2012.

[8] Kaur, K., and E. S. Kaur. "Advanced Fuzzy Logic Based Image Watermarking Technique for Medical Images." International Journal of Advanced Research Ideas and Innovations in Technology 3, no. 1 (2016): 1–7.

[9] Kanimozhi, R., and V. Padmavathi. "Robust and Secure Image Steganography with Recurrent Neural Network and Fuzzy Logic Integration." Scientific Reports 15, no. 1 (2025): 13122.

[10] Hashmi, M. F., and A. G. Keskar. "Block and Fuzzy Techniques Based Forensic Tool for Detection and Classification of Image Forgery." Journal of Electrical Engineering and Technology 10, no. 4 (2015): 1886–1898.

[11] Sahu, A. K. "A Logistic Map Based Blind and Fragile Watermarking for Tamper Detection and Localization in Images." Journal of Ambient Intelligence and Humanized Computing 13, no. 8 (2022): 3869–3881.

[12] Shuriya, B., and A. Rajendran. "A Fuzzy Responsibility-Based Access Organizer for Leukemia Record Protection Using KWatts Algorithm." Applied Mathematics 13, no. 6 (2019): 1047–1052.

[13] Gonge, S. S., and A. Ghatol. "Combination of Fuzzy Logic Digital Image Watermarking and Advanced Encryption Technique for Security and Authentication of Cheque Image." In Intelligent Systems Technologies and Applications, 84–101. Springer, 2018.

[14] Pillutla, H., and A. Arjunan. "A Brief Review of Fuzzy Logic and Its Usage Towards Counter-Security Issues." In 2018 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), 1–6. IEEE, 2018.

[15] Korus, P., and J. Huang. "Multi-Scale Fusion for Improved Localization of Malicious Tampering in Digital Images." IEEE Transactions on Image Processing 25, no. 3 (2016): 1312–1326.

[16] Liu, H., Y. Zhou, Y. Zhang, and Y. Su. "A Rough Set Fuzzy Logic Algorithm for Visual Tracking of Blockchain Logistics Transportation Labels." Journal of Intelligent & Fuzzy Systems 41, no. 4 (2021): 4965–4972.

[17] Ebrahimi, M., S. H. R. Ahmadi, G. PourEbrahim, and A. Foroozani. "Ranking Medical Image Watermarking Algorithms Using the Fuzzy Analytical Hierarchy Process." International Journal of Operational Research 30, no. 1 (2017): 120–141.

[18] Phan-Ho, A. T., and F. Retraint. "A Comparative Study of Bayesian and Dempster-Shafer Fusion on Image Forgery Detection." IEEE Access 10 (2022): 99268–99281.

[19] Swaraja, K., and K. Meenakshi. "An Optimized Blind Dual Medical Image Watermarking Framework for Tamper Localization and Content Authentication in Secured Telemedicine." Biomedical Signal Processing and Control 55 (2020): 101665.

[20] Knorst, A. M., A. A. Vanti, R. A. E. Andrade, and S. L. Johann. "Aligning Information Security with the Image of the Organization and Prioritization Based on Fuzzy Logic for the Industrial Automation Sector." JISTEM–Journal of Information Systems and Technology Management 8 (2011): 555–580.

[21] Zhao, D., and X. Tian. “A Multiscale Fusion Lightweight Image-Splicing Tamper-Detection Model.” Electronics 11, no. 16 (2022): 2621.

[22] Wang, J., W. Zhang, Z. Huang, and J. Li. “Boosting Robustness in Deep Neuro-Fuzzy Systems: Uncovering Vulnerabilities, Empirical Insights, and a Multi-Attack Defense Mechanism.” IEEE Transactions on Fuzzy Systems (2024).

[23] R, Nagarathna C, Jayasri A, Chandana S, and Amrutha A. "Identification of Image Forgeries using Machine Learning - A Review." Journal of Innovative Image Processing 5, no. 3 (2023): 323-336.

[24] Darney, P. Ebby. "Scam Image Detection on Copy-Move by JPEG Features and Classical Block Matching with Improved Variant." Journal of Innovative Image Processing 4, no. 4 (2022): 215-225.

[25] Gowrisankar, B., and V. L. Thing. “An Adversarial Attack Approach for Explainable AI Evaluation on Deep Fake Detection Models.” Computers & Security 139 (2024): 103684.

[26] Rohhila, S., and A. K. Singh. “Deep Learning-Based Encryption for Secure Transmission Digital Images: A Survey.” Computers and Electrical Engineering 116 (2024): 109236.

[27] Uloli, Thiruvaazhi, R. M. Koushal Akash, A. G. Keerthika, and K. S. Dhanwanth. "Survey of Fake Image Synthesis and its Detection." Journal of Innovative Image Processing 4, no. 4 (2022): 278-297.

[28] Karaköse, M., İ. İlhan, H. Yetiş, and S. Ataş. “A New Approach for Deep Fake Detection with the Choquet Fuzzy Integral.” Applied Sciences 14, no. 16 (2024): 7216.

[29] Yadav, A., and D. K. Vishwakarma. “Datasets, Clues and State-of-the-Arts for Multimedia Forensics: An Extensive Review.” Expert Systems with Applications (2024): 123756.