# Position Paper: Neuro-Fuzzy Fusion for Dataset-Agnostic Deepfake Detection

Soumyadeep Chattopadhyay[1], Bhoomi Priya[1], and Pranab K. Muhuri[1]
[1]Department of Computer Science and Engineering, South Asian University, New Delhi, India
soumyadeep.chattopadhyay2025@students.sau.ac.in, bhoomipriya@students.sau.ac.in, pranabmuhuri@cs.sau.ac.in

***Abstract*—The rise of deepfake generation models has made manipulated content increasingly sophisticated and pervasive. While current detectors and domain-specific models perform well on intra-dataset evaluations, they often fail to generalize across datasets generated using unseen generators and suffer from a severe lack of interpretability. This causes a profound explainability gap and limited detection reliability. Henceforth, in this position paper we argue that fuzzy systems, especially neuro-fuzzy systems, can be used to provide an interpretable and generalizable mechanism for decision fusion across heterogeneous detectors. To support this position, we propose a frequency-based ensemble detection network fused through an Adaptive Neuro-Fuzzy Inference System(ANFIS). By coupling multi-domain artifact analysis with neuro-fuzzy reasoning, the proposed network achieves a mean AUC of 0.8705 and a mean F1-Score of 0.9312, matching advanced baselines along with providing interpretable, mathematically derived linguistic rules. This study proves that integrating fuzzy logic into effective deepfake detection networks can be utilized to achieve data-agnostic generalization, enhanced explainability and robust analysis for high stake media forensics.**

***Index Terms*—Frequency based detection, Neuro-Fuzzy, Ensemble Learning**

## I. Introduction

Deepfakes have proliferated across the internet due to advances in technology that allow attackers to forge videos and images nearly indistinguishable from real media. This has created an information crisis in the modern age and poses as a critical challenge for AI systems. In response, numerous methods have been developed to identify and classify media generated using deepfake models. However, these methods often lack cross-dataset generalization abilities. While they perform well on the datasets that they have been trained on, their performance degrades significantly when confronted with media generated by unseen methods. Moreover, even though deep learning networks excel at learning discriminative representations, they offer limited transparency and often struggle to provide reliable confidence scores when encountered with unfamiliar fake content.

Building upon this premise, this paper argues that the primary bottleneck in detection systems is no longer feature extraction, but decision making under uncertain and heterogeneous evidence from multiple sources. We believe that by incorporating fuzzy logic (particularly neuro-fuzzy inference) we can enable soft decision fusion and model prediction uncertainty while providing interpretable reasoning. Fuzzy logic can serve as a bridge between deepfake detection and real-world deployment requirements. This position paper advocates for integrating fuzzy reasoning as a mainstream component in deepfake detection pipelines.

The paper is organized as follows: Section II reviews related work, discussing the current landscape of deepfake detectors and ensemble approaches. Section III outlines the motivation for dataset-agnostic training to improve generalization. Section IV discusses the challenges of domain-specialized ensembles and the need for fuzzy reasoning in the fusion layer. In Section V, we present neuro-fuzzy fusion as an illustrative framework for soft decision fusion and uncertainty modeling. Section VI presents the methodology of the framework. Furthermore, Section VII demonstrates the Experimental Setup and Results of the proposed framework and finally Section VIII concludes the paper.

## II. Related Work

Deepfake detection has seen rapid progress in recent years. Many approaches have been implemented in the deepfake detection ecosystem effectively leveraging the spatial, frequency and noise residuals of the input samples. Existing studies [1]–[5] have consistently shown that deepfake generation techniques have the tendency to introduce character artifacts, especially in the frequency domain which is utilized to detect potential deepfake images [6].

Ensemble approaches that combine multiple detectors have proven to be an effective way to build robust and accurate detection systems [7], [8]. However, most existing ensemble methods rely on simple averaging or majority voting, which limits their ability to explicitly model uncertainty and ambiguous cases.

Fuzzy logic and neuro-fuzzy systems, such as ANFIS [9], have been successfully applied across multiple domains beyond deepfake detection for decision fusion under uncertainty [10], [11]. In particular, fuzzy ensemble approaches have demonstrated strong capability in combining heterogeneous model outputs by handling imprecision and partial confidence in a principled manner [12], [13]. Their adaptability and ability to model vagueness makes them well-suited for complex detection tasks involving unseen or uncertain conditions. Accordingly, this study adopts neuro-fuzzy fusion to bridge the gap between individual detector outputs and reliable, real-world decision-making

## III. Motivation

Current detection techniques present in the literature adopt either a data-specific narrow design, i.e., training on only one dataset or manipulation type, or they utilize a monolithic design i.e., training end-to-end on a mixture of manipulation types. While effective within benchmark settings, such system face several well-documented limitations.

Firstly, these networks overfit to dataset-specific artifacts [14] and lack separate reasoning layers or intermediate models to effectively handle data from multiple sources. This leads to a significant performance drop on cross-dataset evaluation. Secondly, due to the crisp decision boundary, outputs from these network lack meaningful uncertainty quantification. Finally, since the internal decision making is not clear, these networks act as black boxes, where decisions are difficult to explain [15]. Addressing these challenges requires moving beyond accuracy-centric mechanisms and focus the evaluation toward systems that can reason over partial, conflicting, or uncertain evidence.

## IV. Position : Domain-Specialized Ensembles Require Fuzzy Reasoning

We take the position that instead of relying on a single-model for classification tasks, deepfake detection should be transformed as a decision fusion problem for effective generalization. Different manipulation techniques, such as GAN-based synthesis, diffusion models etc. introduce distinct artifacts in the spatial and frequency domain with no single detection model being generalizable across all regimes.

A more robust approach would be to incorporate domain-expert networks and fuse their outputs in a principled manner. Each model exhibits superior performance in their intra-dataset evaluation within a particular generation family. However, naive fusion strategies like average or probability voting doesn't take into account the uncertainty, detector robustness and contextual significance.

By employing fuzzy logic into the fusion step, the network can be provided with a framework for graded reasoning, where detectors contribute soft evidence rather than hard decisions. This aligns with the fuzzy system philosophy of modeling vagueness and partial truth, a core challenge in real-world media forensics.

## V. Neuro-Fuzzy Fusion as an Illustrative Framework

We approach the problem of deepfake detection by augmenting several existing technologies and improving overall performance with a fuzzy network. The aim is to be as dataset-independent as possible, so we propose a modular approach to deepfake detection instead of relying on a monolithic structure. This approach lets us replace sub-components of the whole network as technology progresses, keeping the network highly adaptable to rapidly evolving technologies. As an illustration we propose utilizing the following domain specific detection models:

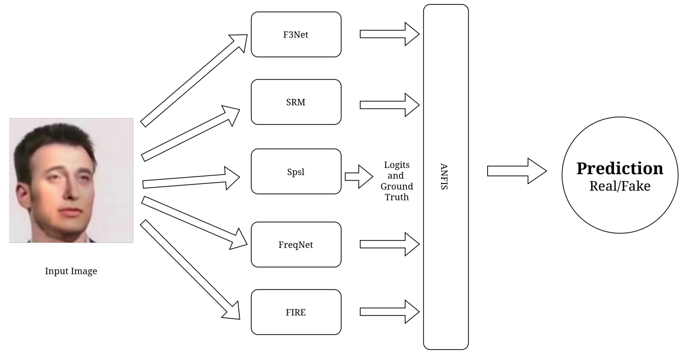



Fig. 1. The proposed Framework

- *F3Net [5]*: Trained on autoencoder, computer graphics, and GAN-based manipulations.
- *SRM [4]*: Focused on residual and noise artifacts.
- *SPSL [1]*: Targeting autoencoder- and graphics-based artifacts.
- *FreqNet [2]*: Specialized for GAN-induced frequency inconsistencies.
- *FIRE [3]*: Designed to detect diffusion-based manipulations.

For the Fuzzy-Fusion component, we propose using ANFIS(Adaptive Neuro Fuzzy Inference System) to learn from the individual domain experts and classify images. The ANFIS module is provided with the logits (raw, non-normalized output score produced by the last layer of a neural network before it is converted into a probability by an activation function)produced by the domain experts as well as the Ground Truth (0 for a real image, 1 for a fake image) As discussed above, instead of enforcing a hard aggregation rule, these outputs are then passed to the Adaptive Neuro-Fuzzy Inference System (ANFIS), which serves as the decision-fusion layer by predicting whether an input image is real or fake as shown in Fig 1.

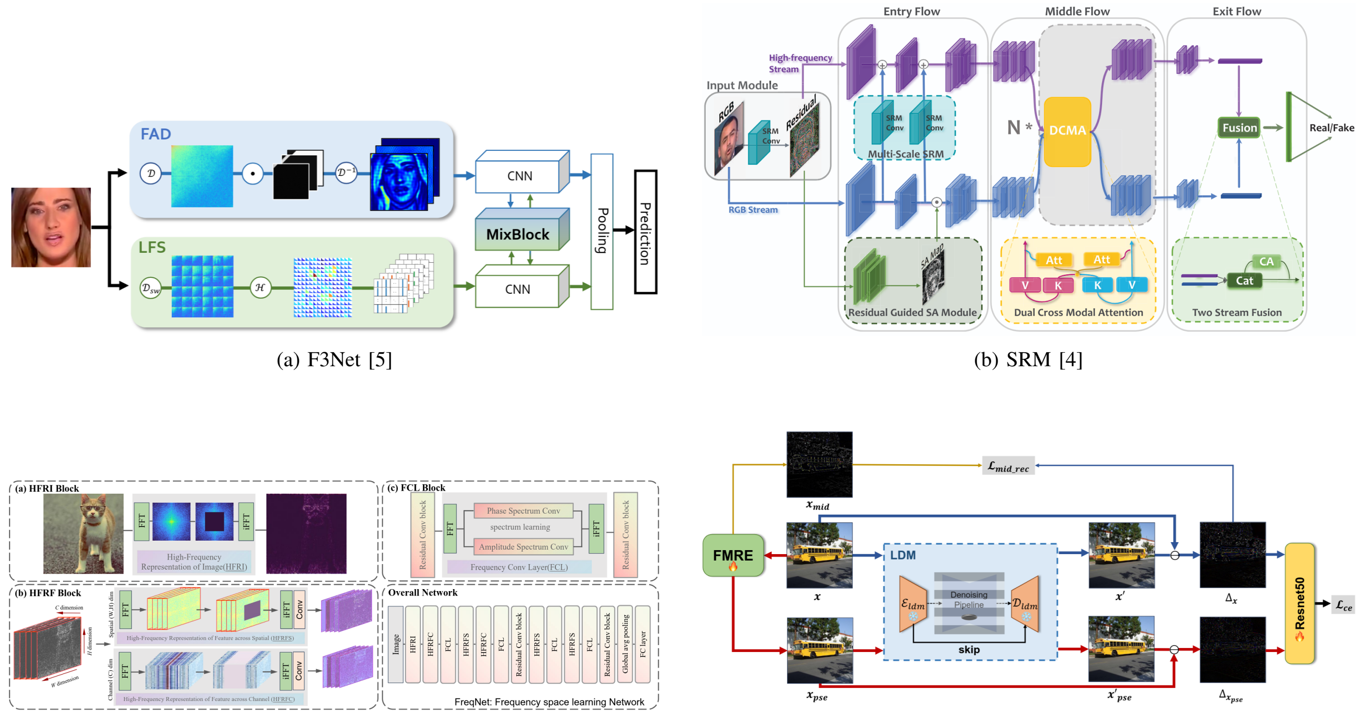
The motivation behind the use of ANFIS was specifically its neuro-fuzzy architecture. Not only does the module learn effective feature representation from data using its neural network component, but it also facilitates interpretability through the inspection of fuzzy rules. Furthermore, the system remains robust under change in input distribution. This framework exemplifies how fuzzy systems can enhance modern AI pipelines without competing directly with deep representation learning. The proposed framework is shown in Fig 1.

## VI. Methodology

### A. Domain Experts

In this section, we discuss the the internal architectures of the domain experts we have proposed to utilize:

*1) F3Net [5]:* As illustrated in Fig 2a, *F3Net* detects generative artifacts using a dual-branch frequency framework. The Frequency-Aware Decomposition (FAD) branch learns forgery patterns across low, middle, and high-frequency spectra, while the Local Frequency Statistics (LFS) branch captures frequency-aware statistical discrepancies between real and




(a) F3Net [5]

(b) SRM [4]

(c) FreqNet [2]

(d) FIRE [3]

Fig. 2. Comparison of pipeline architectures; The architecture diagram of SPSL [1] was not publicly available in the original paper. Image source: [2]–[5]

forged faces. Finally, a cross-attention MixBlock fuses these branches to encourage rich feature interaction.

*2) SRM [4]: Spatial Rich Model (SRM)* Fig 2b filters are used to extract high-frequency noise. The SRM detector uses a two-stream architecture with an aim to process both the RGB image as well as the high-frequency noise data. The interactions between features of the two modalities are modeled and features from the two streams are fused in an attention-based manner for the final classification.

*3) SPSL [1]: Spatial-Phase Shallow Learning* is a dual-stream architecture that leverages the invisible traces in the mathematical phase of an image. Along with a spatial stream, they also add a secondary phase stream where the image is converted into the frequency domain to detect artifacts left by up-sampling, resizing, or blending processes. Instead of using a deep neural network, they employ a shallow learning approach to produce a final fused feature vector which is then passed through a fully connected layer to produce a final classification score.

*4) FreqNet [2]:* As illustrated in Fig 2c, *FreqNet* focuses on both the frequency-domain information derived from images as well as features extracted by a *Convolutional Neural Network (CNN)*. Frequency information is studied using two modules, one to prioritize high-frequency information thus augmenting its sensitivity to significant details and the other to capture broader forgery indicators within the frequency domain by a frequency convolutional layer which diminishes reliance on source-specific characteristics.

*5)FIRE [3]:*As illustrated in Fig 2d, *Frequency-guided Reconstruction Error* works on the drawback that synthetic images possess consistent mid-frequency data that generative algorithms struggle to reconstruct accurately. To leverage the above fact, a Fast Fourier Transform is applied to acquire the full spectral distribution of the input image after which the low-frequency components are centered. Then a module identifies frequency regions that are difficult for a typical image generation model to reconstruct. Furthermore, a frequency mask is applied to obtain an image with its mid-frequencies removed called a pseudo-generated image. Both the original and pseudo-generated images are then input to the network for reconstruction shorten it

To ensure standardized evaluation, pre-trained weights for *F3Net* [5], *SRM* [4], and *SPSL* [1] were sourced from the *DeepfakeBench* [16] framework since the original publications doesn't provide the best model weights. For the remaining domain experts, weights were obtained directly from their respective official GitHub repositories.

### *B. Fuzzy Fusion Using Adaptive Neuro-Fuzzy Inference System*

The Adaptive Neuro-Fuzzy Inference System (ANFIS) [9], is a technique that combines the learning capability of neural networks with the human-like reasoning mechanism of fuzzy logic. It integrates the transparent inference process of the Takagi–Sugeno fuzzy model with the adaptive learning ability of neural networks, enabling the modeling of complex non-linear relationships. As described earlier, the proposed frame-

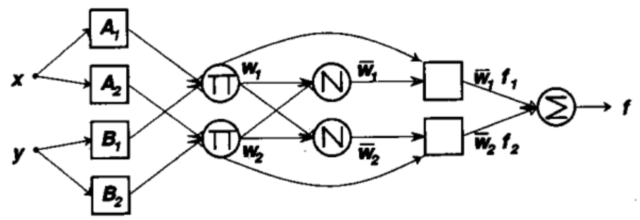



Fig. 3. The five-layer ANFIS architecture used for fusing detector logits, illustrating the flow from fuzzification to final output summation. Image Source: [9]

work employs an ANFIS module to fuse the classification logits obtained from five domain-specialized deepfake detectors.

*1) Fuzzy Rule Base:* For a five-input system, a typical fuzzy IF–THEN rule $R_k$ in the Sugeno model is defined as:

$$\begin{aligned} R_k &: \text{IF } x_1 \text{ is } A_{k1} \text{ AND } \dots \text{ AND } x_5 \text{ is } A_{k5}, \\ &\text{THEN } f_k = \sum_{j=1}^{5} p_{kj} x_j + r_k, \end{aligned} \tag{1}$$

where $k = 1, 2, \dots, n$ denotes the rule index for a total of $n$ rules, $x_j$ denotes the input logits obtained from the five detectors, $A_{kj}$ represents the corresponding linguistic fuzzy sets (e.g., low fake confidence, high real confidence), and $\{p_{k1}, \dots, p_{k5}, r_k\}$ are the consequent parameters learned during training

*2) ANFIS Architecture:* The ANFIS architecture consists of five functional layers, analogous to a neural network structure.

*a) Layer 1: Fuzzification Layer:* Each node in this layer is adaptive and maps a crisp input logit to a fuzzy membership value. In this work, either a Gaussian or a Generalized Bell (GBell) membership function is employed due to its smoothness and continuous differentiability with the choice treated as a tunable design parameter. For the Gaussian Membership Function, The output of each node is given by:

$$O_{1,i} = \mu_{A_i}(x_j) = \exp\left(-\frac{1}{2}\left(\frac{x_j - c_i}{a_i}\right)^2\right), \tag{2}$$

and for the Generalized Bell (GBell) membership function

$$O_{1,i} = \mu_{A_i}(x_j) = \frac{1}{1 + \left|\frac{x_j - c_i}{a_i}\right|^{2b_i}}, \tag{3}$$

where $i$ refers to the $i$-th node index in Layer 2, $\{a_i, c_i\}$ denote the premise parameters corresponding to the width and center of the fuzzy set and $b_i$ controls the slope of the gbell membership function.

*b) Layer 2: Rule Firing Layer:* Nodes in this layer are fixed and compute the firing strength $w_k$ of each fuzzy rule using the product T-norm operator:

$$O_{2,k} = w_k = \prod_{j=1}^{5} \mu_{A_{kj}}(x_j). \tag{4}$$

This layer quantifies the degree to which each rule is activated by a particular combination of detector outputs.

TABLE I
LOMO FOLD ASSIGNMENTS AND GENERATIVE CATEGORIES. CATEGORIES (FR: FACE REENACTMENT, EFS: ENTIRE FACE SYNTHESIS, FS: FACE SWAPPING) ARE ADOPTED FROM THE DF40 DATASET [17].

| Fold | Held-Out Dataset | Generative Type |
|---|---|---|
| Fold 1 | CollabDiff | FE (Diffusion) |
| | DaNet | FR (Image Driven) |
| | InSwap | FS (Used in Roop) |
| | FaceVid2Vid | FR (Landmark Driven) |
| Fold 2 | DDIM | EFS (Latent Diffusion) |
| | FSGAN | FS (Parsing Mask) |
| | PIRender | FR (Image Driven) |
| | DeepFaceLab | FS (High Quality) |
| | FOMM | FR (Image Driven) |
| Fold 3 | PixArt | EFS (Latent Diffusion) |
| | SimSwap | FS (Graphic Based) |
| | UniFace | FS (Disentangle) |
| | FaceSwap | FS (Graphic Swap) |
| | MCNet | FR (Image Driven) |
| Fold 4 | RDDM | EFS (Diffusion) |
| | VQGAN | EFS (GAN based) |
| | MRAA | FR (Image Driven) |
| | SadTalker | FR (Audio Driven) |
| | TPSM | FR (Image Driven) |
| | Wav2Lip | FR (Audio Driven) |

*c) Layer 3: Normalization Layer:* To ensure that the contributions of all fuzzy rules sum to unity, each node in this layer normalizes the firing strength as:

$$O_{3,k} = \bar{w}_k = \frac{w_k}{\sum_{m=1}^{n} w_m}, \tag{5}$$

where $n$ denotes the total number of fuzzy rules.

*d) Layer 4: Defuzzification Layer:* Each node in this layer computes the contribution of the $k$-th rule to the overall output by evaluating:

$$O_{4,k} = \bar{w}_k f_k = \bar{w}_k \left(\sum_{j=1}^{5} p_{kj} x_j + r_k\right), \tag{6}$$

where the parameters $\{p_{kj}, r_k\}$ are optimized during training to emphasize detectors that are more reliable for specific manipulation patterns, such as GAN fingerprints or diffusion-based artifacts.

*e) Layer 5: Summation Layer:* This layer serves as the output layer, where a single node aggregates the contributions of all fuzzy rules to compute the final fusion score:

$$O_{5,1} = \sum_k \bar{w}_k f_k = \frac{\sum_k w_k f_k}{\sum_k w_k}. \tag{7}$$

Here, $O_{5,1}$ represents the final confidence score, which is subsequently passed through a thresholding operation to obtain the binary classification decision (real or fake).

TABLE II
OVERALL PERFORMANCE ACROSS LOMO EVALUATION (MEAN ± STD)

| Method | AUC | EER | Acc | F1 |
|---|---|---|---|---|
| Mean Averaging | 0.8006 ± 0.0841 | 0.2658 ± 0.0435 | 0.8208 ± 0.0689 | 0.8974 ± 0.0427 |
| Weighted Averaging | 0.8093 ± 0.0745 | 0.2611 ± 0.0388 | 0.8623 ± 0.0814 | 0.9224 ± 0.0488 |
| Majority Voting | 0.7529 ± 0.0314 | 0.2348 ± 0.0182 | 0.8108 ± 0.0287 | 0.8925 ± 0.0179 |
| Logistic Regression | 0.8722 ± 0.0351 | 0.2266 ± 0.0356 | 0.8848 ± 0.0620 | 0.9360 ± 0.0369 |
| XGBoost [18] | 0.8797 ± 0.0347 | 0.2156 ± 0.0330 | 0.8415 ± 0.0391 | 0.9105 ± 0.0243 |
| Proposed method | **0.8705 ± 0.0369** | **0.2273 ± 0.0332** | **0.8770 ± 0.0681** | **0.9312 ±0.0408** |

## VII. EXPERIMENTS AND RESULTS

### A. Dataset Construction

To demonstrate real-world effectiveness, our framework was trained and evaluated on a subset of the DF40 [17] dataset, which encompasses 40 diverse generation techniques. For this experiment, we utilized 21 meticulously chosen datasets categorized by their generative architectures out of the 40 datasets available in DF40. Crucially, we excluded any datasets that were originally used to train or evaluate the five base detectors in their respective original papers. The datasets selected for this study and their fold assignments are summarized in Table I. This extensive variety and fold division ensures that the fusion module is able to dynamically weight the expertise of detection modules in the network across diverse artifact distributions. Raw logits were extracted from the five pre-trained detectors across all 21 datasets. To ensure scale parity and numerical stability, these logits were normalized before feeding them into the ANFIS. Additionally, to counter the class imbalance caused by the scarcity of real samples, a stratified sampling strategy was implemented to collect equal number of fake images so that every dataset in the training set has an equal representation preventing the model from developing any kind of bias.

### B. Evaluation Protocol

For evaluation, we implemented a Leave-One-Manipulation-Out (LOMO) strategy which aimed to ensure that in every fold, the test set contains a synthetic manipulation type that the model has never encountered during training. In contrast, the real samples were drawn from two datasets, namely DeepFaceLab and CollabDiff, and were shared across both training and testing splits. This design choice was made due to the limited availability of high-quality real samples within the DF40 dataset. By maintaining a consistent real distribution, the evaluation focuses on the model's ability to generalize across unseen manipulation types rather than variations in real data distributions.

### C. ANFIS Architecture and Design

The architectural design of the proposed fusion module relies on several specific configurations and hyperparameter boundaries to ensure robust, interpretable performance.

*1) Optimization and Configuration:* To implement the ANFIS module, we utilize the GdAnfisClassifier [19] from the Xanfis [19] framework. This modern progression replaces traditional hybrid learning with end-to-end gradient descent using the Adam optimizer. By leveraging autograd capabilities, the model simultaneously optimizes membership functions and fuzzy rules, capturing complex non-linear relationships between detector logits more efficiently than classical implementations.

*2) Constraining the Rule Base (2 to 5 Rules):* To mitigate the common pitfall of 'rule explosion' in neuro-fuzzy systems where models generate hundreds of rules, degrading into a black box, we bounded the search space for the number of fuzzy rules to $R \in [2, 5]$ during hyperparameter optimization. This not only ensures that generated rules map perfectly to human-readable linguistic terms, but also prevents the model from memorizing dataset-specific noise.

*3) Dynamic Membership Function Selection:* Instead of statically assigning membership functions, the shape of the fuzzy sets was treated as a tunable parameter, dynamically selecting between Gaussian and Generalized Bell (GBell) curves, since both functions provide smooth, differentiable curves ideal for gradient-based learning.

*4) Dynamic Threshold Calibration:* Since synthetic artifacts vary across generative families, a static 0.5 boundary is sub-optimal. Henceforth, we implemented dynamic threshold calibration by reserving 15% of each fold's training data as a validation set. The threshold maximizing the F1-score on the validation Precision-Recall curve was then applied to the held-out test set. This adaptive approach optimizes the precision-recall trade-off for each specific training domain

### D. Results

The quantitative analysis of the proposed ANFIS framework is presented alongside five baseline fusion strategies: Mean Averaging, Weighted Averaging, Majority Voting, Logistic Regression, and XGBoost [18]. To mitigate any methodological or architectural bias, identical training protocols were applied in all compared models to ensure that any observed performance gains are strictly attributable to the algorithmic architectures rather than variations in training conditions.

*1) Superiority Over Naive Fusion Strategies:* Table II demonstrates that neural architectures outperform static fusion. Majority voting yielded the lowest performance (AUC: 0.7529, Acc: 0.8108), showing that hard consensus is insufficient for deceptive models. While weighted averaging achieved a competitive F1 (0.9224), its high standard deviation in accuracy reflects susceptibility to variance across evaluation folds.

TABLE III
LEARNED FUZZY RULES FROM ANFIS MODEL FROM THE FOLD THAT GAVE THE HIGHEST AUC OF 0.9109

| Rule | IF–THEN Rule | Weight | Decision |
|---|---|---|---|
| 1 | IF FIRE_Logit is Medium AND FreqNet_Logit is Very High AND F3Net_Logit is Medium AND SPSL_Logit is High AND SRM_Logit is Low THEN weight is 0.5139 | 0.5139 | FAKE |
| 2 | IF FIRE_Logit is Very Low AND FreqNet_Logit is Low AND F3Net_Logit is Very Low AND SPSL_Logit is Very Low AND SRM_Logit is High THEN weight is -0.2971 | -0.2971 | REAL |
| 3 | IF FIRE_Logit is High AND FreqNet_Logit is Very Low AND F3Net_Logit is High AND SPSL_Logit is Low AND SRM_Logit is Medium THEN weight is -0.5796 | -0.5796 | REAL |
| 4 | IF FIRE_Logit is Low AND FreqNet_Logit is High AND F3Net_Logit is Low AND SPSL_Logit is Very High AND SRM_Logit is Very High THEN weight is -0.2829 | -0.2829 | REAL |
| 5 | IF FIRE_Logit is Very High AND FreqNet_Logit is Medium AND F3Net_Logit is Very High AND SPSL_Logit is Medium AND SRM_Logit is Very Low THEN weight is 0.0015 | 0.0015 | FAKE |

*2) ANFIS vs. Traditional Machine Learning Baselines:* The proposed ANFIS architecture proved to be highly effective as a trainable fuzzy fusion mechanism achieving an AUC of 0.8705 and a standout F1-Score of 0.9312. Notably, ANFIS significantly outperformed the advanced tree-based baseline, XGBoost [18], in both accuracy (0.8770 vs. 0.8415) and F1-Score (0.9312 vs. 0.9105). While Logistic Regression resulted in a marginal gain in raw accuracy (+0.0078) and F1 score (+0.0048) as shown in Table II, the performance delta between ANFIS and these machine learning models is statistically negligible given the overlapping standard deviations. More importantly, because these metrics were achieved under a strict Leave-One-Manipulation-Out (LOMO) cross-validation protocol, this highlights the exceptional generalizability of the ANFIS framework.

The primary novelty of the proposed ANFIS architecture lies in providing algorithmic transparency without sacrificing predictive accuracy. Unlike "black-box" ensembles that obscure conflict resolution, ANFIS matches or exceeds baselines while remaining interpretable using mathematically derived linguistic rules (Table III).

## VIII. CONCLUSION

This paper argues that fuzzy logic transforms deepfake detection from pure classification into a problem of reasoning under uncertainty. By combining domain-expert detectors with an ANFIS module, our framework achieves state-of-the-art performance while providing explainability through human-understandable rules and confidence scores. This approach ensures stable, generalizable performance across diverse conditions, motivating the use of fuzzy systems as foundational components in modern detection architectures.

## REFERENCES


[1] H. Liu, X. Li, W. Zhou, Y. Chen, Y. He, H. Xue, W. Zhang, and N. Yu, "Spatial-phase shallow learning: rethinking face forgery detection in frequency domain," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 772–781.

[2] C. Tan, Y. Zhao, S. Wei, G. Gu, P. Liu, and Y. Wei, "Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 5, 2024, pp. 5052–5060.

[3] B. Chu, X. Xu, X. Wang, Y. Zhang, W. You, and L. Zhou, "Fire: Robust detection of diffusion-generated images via frequency-guided reconstruction error," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 12 830–12 839.

[4] Y. Luo, Y. Zhang, J. Yan, and W. Liu, "Generalizing face forgery detection with high-frequency features," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 16 317–16 326.

[5] Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao, "Thinking in frequency: Face forgery detection by mining frequency-aware clues," in *European conference on computer vision*. Springer, 2020, pp. 86–103.

[6] J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, and T. Holz, "Leveraging frequency analysis for deep fake image recognition," in *International conference on machine learning*. PMLR, 2020, pp. 3247–3258.

[7] T. Manju and S. Kalarani, "Deepsight: Enhancing deepfake image detection and classification through ensemble and deep learning techniques," in *2024 5th International Conference on Image Processing and Capsule Networks (ICIPCN)*. IEEE, 2024, pp. 28–35.

[8] S. Das, R. Kumari, U. Tripathi, S. Parashar, A. Dubey, A. Yadav, and R. K. Singh, "Enhanced deepfake detection using cnn and efficientnet-based ensemble models for robust facial manipulation analysis," in *2025 First International Conference on Advances in Computer Science, Electrical, Electronics, and Communication Technologies (CE2CT)*. IEEE, 2025, pp. 677–681.

[9] J.-S. R. Jang, "Anfis: Adaptive-network-based fuzzy inference system," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 23, no. 3, pp. 665–685, 1993.

[10] M. S. Hosseini and M. Zekri, "Review of medical image classification using the adaptive neuro-fuzzy inference system," *Journal of Medical Signals and Sensors*, vol. 2, no. 1, pp. 49–60, Jan. 2012.

[11] K. Uma and R. S. Rajesh, "Anfis based classification of remotely sensed image features," *Natural Sciences Publishing (from 2022 study)*, 2022.

[12] B. Subburaj and R. Ragavendra, "Deepfake detection using spatio-temporal-structural anomaly learning and fuzzy system based decision fusion," *IEEE Access*, 2025.

[13] M. Karaköse, I. Ilhan, H. Yetiş, and S. Ataş, "A new approach for deepfake detection with the choquet fuzzy integral," *Applied Sciences*, vol. 14, no. 16, p. 7216, 2024.

[14] A. V. Nadimpalli and A. Rattani, "On improving cross-dataset generalization of deepfake detectors," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 91–99.

[15] H. Qian, L. Xia, R. Ge, Y. Fan, Q. Wang, and Z. Jing, "From black boxes to glass boxes: Explainable ai for trustworthy deepfake forensics," *Cryptography*, vol. 9, no. 4, p. 61, 2025.

[16] Z. Yan, Y. Zhang, X. Yuan, S. Lyu, and B. Wu, "Deepfakebench: A comprehensive benchmark of deepfake detection," *arXiv preprint arXiv:2307.01426*, 2023.

[17] Z. Yan, T. Yao, S. Chen, Y. Zhao, X. Fu, J. Zhu, D. Luo, C. Wang, S. Ding, Y. Wu, and L. Yuan, "Df40: Toward next-generation deepfake detection," 2024. [Online]. Available: https://arxiv.org/abs/2406.13495

[18] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[19] N. V. Thieu. (2025, Jun.) X-ANFIS: An extensible and cross-learning ANFIS framework for machine learning tasks. [Online]. Available: https://github.com/thieu1995/X-ANFIS