Trustworthy and Explainable Neuro-Fuzzy Ensembles for Clinical Decision Support

Batyrkhan Sharipbay and Adnan Yazici
Department of Computer Science, *Nazarbayev University*, *Astana, Kazakhstan*
{batyrkhan.sharipbay,adnan.yazici}@nu.edu.kz

***Abstract*—We introduce an interpretable clinical classification framework that couples Adaptive Neuro-Fuzzy Inference Systems (ANFIS) with Fuzzy Random Forests (FRF) to achieve high performance while maintaining transparent decision logic. In the proposed design, ANFIS learns data-driven Gaussian membership functions and optimizes their mean and sigma parameters, which are then used to induce a robust FRF ensemble for decision-making. The end-to-end pipeline follows a leakage-aware methodology: (i) stratified train–test splitting with skewness-aware imputation computed strictly from the training set; (ii) Z-score normalization (zero mean, unit variance) fitted on training data only; (iii) mutual information–based feature selection via SelectKBest to obtain a compact, discriminative subset; (iv) SMOTE class balancing applied exclusively to the training set; (v) ANFIS-based optimization of Gaussian membership parameters; and (vi) FRF construction using ANFIS-optimized fuzzy partitions to produce interpretable predictions. Experiments on two UCI benchmarks (Cleveland Heart Disease and Wisconsin Diagnostic Breast Cancer) show that the ANFIS–FRF framework achieves AUC values of approximately 0.95 and 0.99, respectively, matching or exceeding strong baselines such as Random Forest and XGBoost. Importantly, the model yields a human-readable rule set, whose interpretability is further examined through an LLM-assisted evaluation. By decoupling fuzzy feature learning (ANFIS) from decision logic induction (FRF) and embedding mutual information-driven feature selection within rigorous preprocessing, the proposed approach provides an accurate, robust, and transparently explainable solution for clinical decision support.**

***Index Terms*—ANFIS, Fuzzy Random Forest, Heart Disease, Breast Cancer, Explainable AI, Medical Diagnosis, LLM Evaluation.**

I. Introduction

Recent advances in Artificial Intelligence (AI) have increased the use of machine learning methods in various medical domains [1]. This has also opened many opportunities to enhance the accuracy and efficiency of diagnosis. Yet deploying AI in safety-critical clinical workflows remains challenging due to the requirements for interpretability, auditability, and trustworthiness [2]. Healthcare institutions must not only obtain accurate predictions but also understand the reasoning behind them to support clinician oversight, patient safety, and regulatory compliance [3].

Conventional machine learning, especially Deep Learning (DL), often operates as a "black box" that produces outputs without a transparent rationale [3], [4]. This opacity is problematic in medicine, where clinical acceptance hinges on traceable, human-understandable justifications [3]. Although explainable AI (XAI) has produced both model-agnostic and model-specific techniques to mitigate this gap [4], practical solutions must balance clarity with predictive strength and robustness to real-world data issues (missingness, imbalance, noise).

Fuzzy logic-based, model-specific approaches provide linguistic rules that align with clinical reasoning. Adaptive Neuro-Fuzzy Inference Systems (ANFIS) seem attractive in this regard, combining neural learning with interpretable rule bases [5]. However, ANFIS can suffer from the rule explosion, eroding interpretability and complicating maintenance. At the other end, classical decision-tree ensembles (e.g., Random Forest, XGBoost) offer strong accuracy, fast training, and modest memory usage [6]. Yet they may lack intuitive mapping of inputs to higher-level clinical concepts [6] and can be unstable under slight distribution shifts [7].

Fuzzy Random Forest (FRF) classifiers could be the solution that combines the strengths of ANFIS and classical ensembles. Proposed by Bonissone et al. [8], they have the robustness of diverse tree classifiers and the flexibility of fuzzy logic. Moreover, each fuzzy decision tree bridges the gap between crisp data and interpretable concepts while achieving strong performance rates [6]. FRF has also been shown to exhibit no significant degradation in performance with noisy and imperfect datasets [8], which are often prevalent in the healthcare sector.

We therefore propose a hybrid framework offering the following key contributions:

- ANFIS-guided ensemble learning to define fuzzy clinical features, used for robust and interpretable decision logic;
- FRF rule extraction through a traversal algorithm, providing a global view of the model's decision-making process;
- LIME technique as a model-agnostic explainable AI tool for instance-level transparency;
- LLM-based explanation is incorporated to translate activated fuzzy rules, LIME, and ensemble prediction into concise and clinician-readable rationales.

The remaining part of the paper is structured as follows: Section II reviews related work on neuro-fuzzy modeling, fuzzy random forests and ensembles, and interoperability techniques. Section III provides a detailed description of our proposed pipeline, including dataset descriptions, preprocessing, and LLM prompts. Section IV discusses the results and ablations, with interpretability and robustness analyses, and highlights limitations. Section V concludes the study with future considerations.

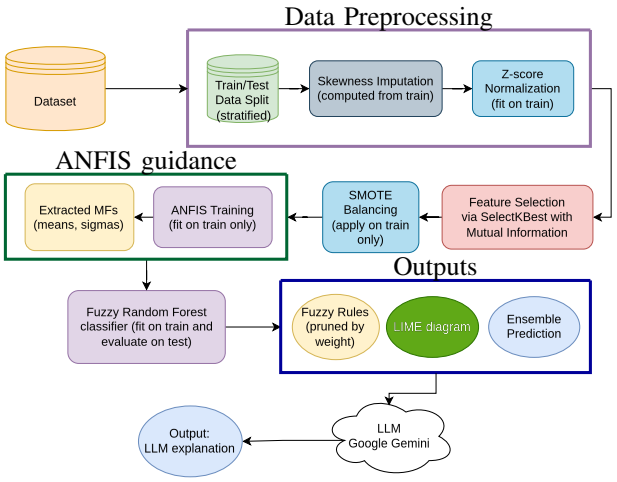

Fig. 1. Proposed ANFIS-FRF framework pipeline: A stratified train-test split is performed first to prevent data leakage. The training data then undergoes skewness-aware imputation, Z-score normalization, and feature selection using SelectKBest with Mutual Information. SMOTE is applied only to the training set to preserve the natural test distribution. ANFIS is trained on the balanced training data to learn MF values, which are then extracted to construct the FRF ensemble classifier. The final explanation is generated by prompting the LLM with the resulting fuzzy rules, LIME diagram, and ensemble prediction.

Note: All the framework code written in Python will be made available upon request. The datasets used in this study are publicly available.

II. Related Work

Neuro-fuzzy models are prominent in clinical decision support because they can learn from data while preserving human-interpretable rule semantics. Recent studies demonstrate that carefully designed fuzzy rule-based systems can achieve strong diagnostic performance on tabular medical benchmarks [9], and that hybrid fuzzy deep learning architectures can improve accuracy while retaining interpretability components [3]. Moreover, ANFIS has been combined with large language models (LLMs) in explainable pipelines, where the LLM performs the main task, and ANFIS provides a reliability or verification layer [5]. These directions motivate using ANFIS not only as an end classifier but also as a learnable module for producing clinically meaningful fuzzy representations that can be transferred to downstream models.

Ensemble methods remain strong baselines for medical tabular classification, and fuzzy extensions aim to improve robustness under uncertainty and noise. Fuzzy Random Forest (FRF) has been applied to heart disease prediction with preprocessing and balancing strategies, illustrating the value of fuzzy partitions inside tree ensembles [7]. Other variants inject fuzzy uncertainty into the split criterion or tree induction, such as fuzzy-entropy–guided forests [10] and intuitionistic fuzzy decision tree ensembles [6]. However, existing fuzzy-ensemble approaches rely on predefined membership functions or heuristic fuzzy partitions, which raises the question of how to learn data-driven fuzzy partitions while keeping the ensemble rule logic interpretable.

In clinical settings, explanations are needed at both the model level (global logic) and the instance level (why this patient was classified this way). Rule-based explanations from fuzzy systems offer transparent decision logic, while model-agnostic approaches (e.g., LIME and SHAP) provide feature-attribution views for arbitrary predictors [4]. Local explanation methods have been emphasized as particularly important in healthcare, where clinicians' trust depends on case-level reasoning [2]. More recently, LLMs have been explored as interfaces for translating model outputs into clinician-readable narratives, but their stochasticity and evaluation remain active concerns [11], [12]. These works implicitly show the potential in combining explicit fuzzy rules with model-agnostic approaches and controlled LLM-generated summaries to deliver explanations that are both structured and accessible.

III. Proposed Methodology

The proposed methodology uses a leakage-aware pipeline in which ANFIS learns data-driven Gaussian membership functions (MFs) and exports their parameters to a Fuzzy Random Forest (FRF) to enable interpretable rule-based classification. The overall architecture is shown in Fig. 1.

A. Datasets

We evaluate on two UCI datasets (Table I). Cleveland Heart Disease has 303 records and 13 input features (plus the target): *age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal* [13]. Following common practice, we convert the original 5-level target to binary classification (disease present vs. not present) [7]. We do not need to apply categorical encodings, as it has already been handled by UCI [13]. Wisconsin Diagnostic Breast Cancer (WDBC) has 569 records and 30 real-valued input features grouped as *mean*, *standard error*, and *worst* measurements for: *radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, fractal dimension* [14].

B. Data split and preprocessing

We use a stratified 80/20 train–test split, and all preprocessing steps are fit on the training split only and then applied to the test split. For Cleveland, missing values are imputed via a skewness-aware strategy [15]: mean if $|\text{skew}| \leq 0.5$, median if $\text{skew} > 0.5$, otherwise mode. Features are standardized using Z-score normalization (fit on train only) [7]. To limit ANFIS rule growth, we apply `SelectKBest` with mutual information and retain the top-k predictors (Cleveland: $k = 7$, WDBC: $k = 8$). Finally, SMOTE (Synthetic Minority Oversampling Technique) is applied *only* on the training set to address class imbalance [7], [10].

C. ANFIS for fuzzy feature learning

In our framework, ANFIS is used as a supervised fuzzy feature learning module: it learns Gaussian MFs per selected feature and exports the MF parameters (μ, σ) for FRF construction. We use three Gaussian MFs per feature:

$$\mu_{f,j}(x_f) = \exp\left(-\frac{(x_f - \mu_{f,j})^2}{2\sigma_{f,j}^2}\right), \tag{1}$$

TABLE I
INFORMATION ON UCI DATASETS.

Dataset name	Number of records	Features number	Number of classes	Missing value present
Cleveland Heart Disease [13]	303	13	5 (2)	Yes
Wisconsin Diagnostic Breast Cancer [14]	569	30	2	No

with $\sigma_{f,j}$ clamped to $\sigma_{f,j} \leftarrow \max(\sigma_{f,j}, \sigma_{\min})$ to avoid degenerate partitions.

Rule firing. Let d be the number of selected features, $R = M^d$ the number of rules, and r index an MF-choice vector $(j_1, \ldots, j_d)$. The (unnormalized) firing strength is computed using a product t-norm:

$$\phi_r(x) = \prod_{f=1}^{d} \mu_{f,j_f}(x_f), \tag{2}$$

and normalized as $\bar{\phi}_r(x) = \phi_r(x) / \left(\sum_{r'=1}^{R} \phi_{r'}(x) + \epsilon\right)$. We use TSK-style consequents with bias-augmented input $x' = [x; 1]$:

$$o(x) = \sum_{r=1}^{R} \bar{\phi}_r(x) \left(x'^{\top} A_r\right), \tag{3}$$

where A_r are learned consequent parameters and the model is trained with cross-entropy loss. We trained ANFIS for 200 epochs with a learning rate of 0.005. After training, we export the learned MF parameters `mf_params`=$(\mu_{f,j}, \sigma_{f,j})$ as fixed fuzzy partitions for the FRF.

D. Fuzzy Random Forest using ANFIS MFs

FRF represents an ensemble of fuzzy decision trees using bootstrap samples and random feature subsets [6], [7]. Our implementation trains an ensemble of 15 fuzzy decision trees with a max depth of 8 for each tree. In contrast to standard FRF, our fuzzy partitions are not hand-crafted: each split uses membership degrees computed from ANFIS-exported `mf_params` (Eq. 1).

Fuzzy entropy and information gain. At a node, each training sample x_i has weight $w_i \geq 0$, and there are C classes. Define weighted class masses $S_c = \sum_{i:y_i=c} w_i$ and $S = \sum_i w_i$, then $p_c = S_c/(S + \epsilon)$. Fuzzy entropy is:

$$H(w, y) = -\sum_{c=0}^{C-1} p_c \log_2(p_c + \epsilon). \tag{4}$$

For the candidate split feature f, let $(M_f)_{i,j} = \mu_{f,j}(x_{i,f})$ be the membership matrix using Eq. 1. For branch j, define child weights $w_i^{(j)} = w_i (M_f)_{i,j}$ and $W_j = \sum_i w_i^{(j)}$. The fuzzy information gain is:

$$G_f = H(w, y) - \sum_{j=1}^{M} \frac{W_j}{W + \epsilon} H(w^{(j)}, y), \tag{5}$$

where $W = \sum_i w_i$. Trees vote by averaging predicted class probabilities across the ensemble.

Algorithm 1 FRF rule extraction (tree traversal) and scoring

Input: forest $\mathcal{M} = \{T_i\}_{i=1}^{N}$, threshold τ, (optional) sample x
Initialize map `RuleScore`$\leftarrow \emptyset$
for all trees T_i in $\mathcal{M}$ **do**
    DFS(T_i.`root`, [], $1/N$)
end for
$\mathcal{R} \leftarrow$ list of unique rules built from `RuleScore`
if x is provided **then**
    Score each $r \in \mathcal{R}$ via $s(x)$; filter scores $< \tau$; sort descending
end if
return $\mathcal{R}$ (ranked if x is provided)

function DFS(node, path, treeWeight)
    if node is leaf **then**
        $A \leftarrow$ Simplify(path), $c \leftarrow$ DominantClass(node), $w \leftarrow treeWeight \times Confidence(node)$
        if $|A| > 0$ **then**
            `RuleScore`$[(A, c)]$ += w
        end if
        return
    end if
    for all children $(j,$ `child`$)$ of node **do**
        DFS(`child`, path $\cup[(node.featureIdx, j)]$, treeWeight)
    end for
end function

E. Rule extraction and scoring

For global interpretability, we traverse each fuzzy tree and extract root-to-leaf decision paths as IF–THEN rules, then merge identical rules across trees (see Algorithm 1). For a path antecedent set A ending in a leaf with class distribution q, define the consequent $c^\star = \arg\max_c q_c$ and confidence $\mathrm{Conf} = q_{c^\star}$. With uniform tree weight $1/N$, the rule weight is:

$$w = \frac{1}{N} \mathrm{Conf}. \tag{6}$$

For a patient x, the rule activation is:

$$\alpha(x) = \prod_{(f,j) \in A} \mu_{f,j}(x_f), \tag{7}$$

and the patient-specific rule score is $s(x) = w\,\alpha(x)$, used to rank rules for instance-level explanations. We prune the weak rules that score less than 0.05.

F. LIME and LLM prompting

Alongside rule-based explanations, we use LIME to provide instance-level feature attributions for individual predictions [4]. To produce clinician-readable narratives, we prompt an LLM (Google's Gemini 2.5 Flash [16]) with (i) the top-ranked extracted rules and (ii) the LIME explanation for the same case, instructing it to focus on clinical meaning rather than ML jargon.

> *You are a clinical AI assistant interpreting ML model predictions for medical professionals. Explain which features most influenced the prediction using the provided patient data: [*dataset row after all preprocessing steps*] (feature, normalized_value).*
> *Reference these explanations: [*filenames for LIME and Fuzzy rules (if ANFIS-FRF)*].*
> *Focus on clinical relevance rather than data science jargon. Explain the impact of high-importance features in relation to their normalized values. Be concise and avoid headings.*
> *Start your response exactly with: 'LLM explanation is as follows:' (without quotes).*

To reduce stochasticity, we use self-consistency by generating multiple candidate responses and selecting the most consistent [17].

IV. Results and Discussion

This section evaluates ANFIS-FRF against Random Forest (RF) and XGBoost in terms of predictive performance and explainability. We report both a stratified 80/20 holdout evaluation and a stratified 10-fold cross-validation (CV). Explainability is analyzed using fuzzy rule extraction, LIME, and clinician-oriented LLM narratives; explanation quality is additionally scored using an LLM-as-a-Judge (LaaJ) protocol.

Due to space constraints, the confusion matrices, LIME diagrams, and LLM outputs are included only for the Cleveland dataset.

A. Heart Disease Cleveland Evaluation

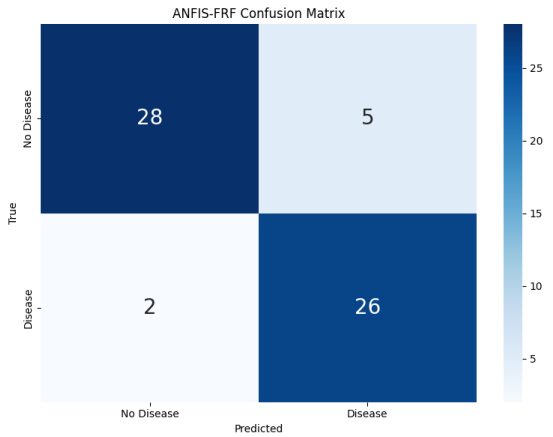

Fig. 2. Cleveland ANFIS-FRF Confusion Matrix.

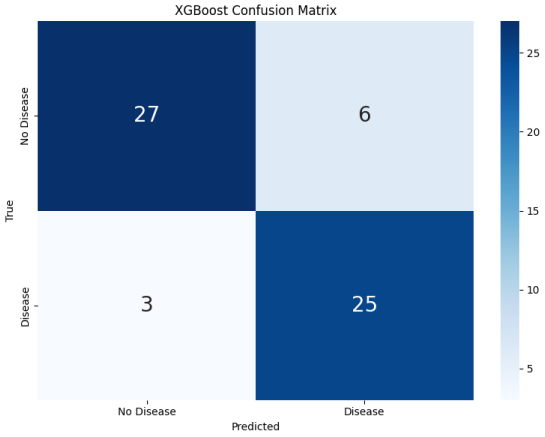

Fig. 3. Cleveland XGBoost Confusion Matrix.

1) Holdout performance (80/20): Table II compares holdout performance under the same preprocessing setup. ANFIS-FRF attains the best overall results on Cleveland, achieving 0.89 Accuracy, 0.85 Specificity, and 0.93 Recall, which yields an F1-score of 0.89 and a Precision of 0.84. It also improves NPV (0.93), MCC (0.77), and AUC (0.95) relative to RF and XGBoost. The corresponding confusion matrices are shown in Fig. 2 and Fig. 3.

TABLE II
Holdout Comparison for Cleveland.

Metrics	RF	XGBoost	*ANFIS-FRF
Accuracy	0.80	0.85	**0.89**
Specificity	0.76	0.81	**0.85**
Recall	0.86	0.89	**0.93**
F-1 score	0.80	0.85	**0.89**
Precision	0.75	0.81	**0.84**
NPV	0.86	0.90	**0.93**
MCC	0.61	0.71	**0.77**
AUC	0.90	0.92	**0.95**

2) 10-fold cross-validation: To assess robustness beyond a single split, we performed a stratified 10-fold CV. In each fold, imputation (if applicable), standardization, feature selection, and SMOTE were fit on the training partition only and applied to the validation partition to avoid leakage. As shown in Table III, RF and XGBoost achieve similar mean accuracy (≈0.79), while ANFIS-FRF attains the highest mean AUC (0.87), indicating stable ranking performance across folds.

TABLE III
10-fold CV for Cleveland (mean ± std).

Model	Accuracy	Precision	Recall	F1-macro	AUC
RF	0.79 ± 0.05	0.79 ± 0.06	0.78 ± 0.05	0.79 ± 0.06	0.86 ± 0.05
XGBoost	**0.79 ± 0.06**	**0.80 ± 0.06**	**0.79 ± 0.06**	**0.79 ± 0.06**	0.86 ± 0.05
*ANFIS-FRF	0.77 ± 0.06	0.78 ± 0.06	0.77 ± 0.07	0.77 ± 0.07	**0.87 ± 0.05**

3) Local explainability (LIME) and LLM narrative: The LIME diagrams are in the format shown in Fig. 4 (example for ANFIS-FRF). We also generate clinician-oriented narratives using the LLM prompt described in Section III-F; for brevity, we report one representative explanation per model.

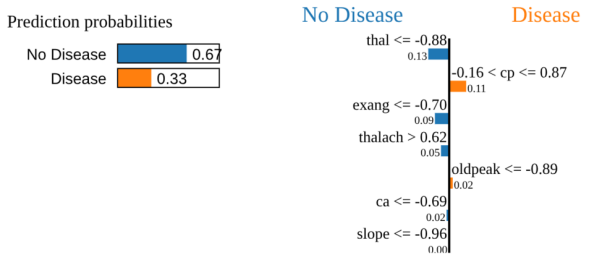

Fig. 4. Cleveland ANFIS-FRF LIME diagram.

The most consistent LLM-generated explanation from LIME and Fuzzy Rule interpretations of ANFIS-FRF is as follows:

ANFIS-FRF LLM explanation is as follows:
The prediction is 'No Disease' (confidence: 67.2%), with strong consensus (80% of decision trees agreeing). The most influential features are:
1. thalach (29.0%): Normalized value 1.415, classified as 'HIGH', strongly supports 'No Disease' and suggests better cardiovascular fitness.
2. exang (25.2%): Normalized value -0.696, classified as 'MEDIUM', supports 'No Disease'; little or no exercise-induced angina argues against obstructive coronary artery disease.
3. thal (21.1%): Normalized value -0.883, classified as 'LOW', supports 'No Disease'; this may indicate normal stress-test blood flow and no defect.
4. slope (17.3%): Normalized value -0.958, classified as 'LOW', supports 'No Disease'; in this context, 'LOW' appears favorable.
5. oldpeak (7.5%): Normalized value -0.8916, classified as 'MEDIUM', supports 'No Disease' and suggests minimal ST depression.

The most consistent XGBoost LLM explanation is as follows:

XGBoost LLM explanation is as follows:
The prediction is 'No Disease' and is influenced by the following features:
1. thalach: Most positive influence (1.42); a higher exercise heart rate is associated with the predicted outcome.
2. cp: Positive influence (0.87); the chest pain type or severity may support the predicted outcome.
3. slope: Most negative influence (-0.96); the exercise ST slope is associated with a decreased likelihood of the predicted outcome.
4. oldpeak: Negative influence (-0.89); exercise-induced ST depression is associated with a decreased likelihood of the predicted outcome.
5. thal: Negative influence (-0.88); a heart blood flow or wall motion characteristic is associated with a decreased likelihood of the predicted outcome.
6. exang: Negative influence (-0.70); exercise-induced angina is associated with a decreased likelihood of the predicted outcome.
7. ca: Negative influence (-0.69); the number of major vessels colored by fluoroscopy is associated with a decreased likelihood of the predicted outcome.

4) LLM-based evaluation (Cleveland): To quantify explanation usefulness, we apply the LaaJ protocol that scores explanations on four criteria (completeness, coherency, faithfulness, and domain plausibility) using a 1–5 scale (higher is better). For this protocol, we used a different instance of Gemini 2.5 Flash to avoid biased judging. Table IV reports the judge scores for the Cleveland case. Note that these scores are illustrative rather than evidentiary: they may not match human clinicians' assessment of explanation quality in clinical practice.

Overall, ANFIS-FRF scores higher on coherency and domain plausibility, reflecting clearer clinical framing and tighter linkage between the model outputs and the final narrative.

TABLE IV
CLEVELAND EXPLANATIONS ASSESSED BY LLM

Criterion	ANFIS-FRF	XGBoost
Completeness	4	4
Coherency	5	4
Faithfulness	4	4
Domain Plausibility	5	4

B. WDBC Evaluation

1) Holdout performance (80/20): Table V reports holdout performance on WDBC. RF achieves the best overall results, while ANFIS-FRF remains competitive: it matches XGBoost in AUC (0.99) and maintains high Recall (0.96) and Precision (0.96), with MCC=0.89.

TABLE V
HOLDOUT COMPARISON FOR WDBC.

Metrics	RF	XGBoost	*ANFIS-FRF
Accuracy	**0.96**	0.95	0.95
Specificity	**0.95**	0.93	0.93
Recall	**0.97**	0.96	0.96
F-1 score	**0.96**	0.95	0.95
Precision	**0.97**	0.96	0.96
NPV	**0.95**	0.93	0.93
MCC	**0.92**	0.89	0.89
AUC	0.98	**0.99**	**0.99**

2) 10-fold cross-validation: We also performed a stratified 10-fold CV on WDBC. Compared with RF and XGBoost, ANFIS-FRF showed a lower mean accuracy and substantially higher fold-to-fold variance, while still maintaining strong AUC, indicating that class ranking remained good even when the final decision boundary was unstable. This variance seems to be primarily from the sensitivity of the ANFIS membership learning stage to fold composition under the current model capacity, rather than from a complete loss of separability. In

particular, the combination of the selected feature dimensionality and three membership functions per feature induces a large fuzzy rule space, which can make the learned partitions fold-dependent on a dataset of this size. The CV results should be interpreted as evidence that the present ANFIS-FRF configuration is not yet robust on WDBC, and that a smaller membership-function count and a stronger per-fold ANFIS training procedure are needed to improve stability.

TABLE VI
10-FOLD CV FOR WDBC (MEAN ± STD).

Model	Accuracy	Precision	Recall	F1-macro	AUC
RF	**0.94 ± 0.03**	**0.94 ± 0.03**	**0.93 ± 0.04**	**0.93 ± 0.04**	0.97 ± 0.03
XGBoost	**0.94 ± 0.03**	0.93 ± 0.03	**0.93 ± 0.04**	0.93 ± 0.03	0.98 ± 0.03
*ANFIS-FRF	0.86 ± 0.11	0.88 ± 0.07	0.88 ± 0.09	0.86 ± 0.11	**0.98 ± 0.02**

3) LLM-based evaluation (WDBC): For a representative malignancy prediction, the LaaJ scores are reported in Table VII. ANFIS-FRF scores higher on completeness and coherency, while XGBoost attains higher faithfulness in this case, consistent with a more direct attribution-style explanation. As mentioned before, these scores are illustrative rather than definitive.

TABLE VII
WDBC EXPLANATIONS ASSESSED BY LLM

Criterion	ANFIS-FRF	XGBoost
Completeness	5	4
Coherency	5	4
Faithfulness	4	5
Domain Plausibility	5	5

C. Studies Comparison

As summarized in Table VIII, ANFIS-FRF achieves competitive performance on both datasets while additionally emphasizing interpretability (fuzzy rules), local explanation (LIME), and clinician-oriented narratives (LLM). Compared to prior work, we consistently report clinically relevant summary metrics such as specificity, MCC, and AUC in addition to accuracy.

In Table IX, our proposition is also the only study that simultaneously supports multiple datasets, data balancing, interpretability, LIME-based local explanations, and LLM-assisted explanations, which collectively strengthen its suitability for decision-support settings.

D. Limitations

While ANFIS-FRF demonstrates promising accuracy–interpretability tradeoffs on two UCI medical benchmarks, several limitations should be considered when interpreting the findings. First, the experimental validation is limited to two datasets (Cleveland Heart Disease and Wisconsin Diagnostic Breast Cancer), which are relatively small and may not capture the full heterogeneity, noise patterns, and dataset shift encountered in real clinical settings. Second, although the pipeline is designed to reduce information leakage by fitting preprocessing transformations on training data only, oversampling-based balancing can still alter the effective decision boundary and does not necessarily reflect real-world prevalence.

Third, the current hyperparameter configuration (e.g., membership function count, ANFIS training settings, and FRF depth/ensemble size) was selected empirically, and systematic sensitivity analyses and ablation studies remain to be conducted. Fourth, the explanation stack combines rule extraction, LIME, and LLM-generated clinician-facing narratives, but both generation and LLM-assisted explanation evaluation introduce potential subjectivity and depend on the chosen LLM (Gemini 2.5 Flash in this study). Finally, scalability remains a practical concern because ANFIS-induced fuzzy rule bases can grow rapidly with the number of selected features and membership functions, potentially limiting straightforward extension to higher-dimensional clinical datasets without additional pruning or structural constraints.

V. CONCLUSION AND FUTURE WORK

This paper proposes ANFIS-FRF, a hybrid clinical classification framework that separates fuzzy feature learning from decision-logic induction by training ANFIS to learn Gaussian membership functions and transferring these learned partitions to a Fuzzy Random Forest classifier. The pipeline enforces leakage-aware preprocessing through stratified splitting, training-only fitting of imputation/standardization/feature selection, and SMOTE applied exclusively to training data to support clinically valid evaluation. Across Cleveland Heart Disease and Wisconsin Diagnostic Breast Cancer, ANFIS-FRF achieved strong discriminative performance (AUC approximately 0.95 and 0.99, respectively), remaining competitive with Random Forest and XGBoost while preserving an explicit, human-readable fuzzy rule set. To support clinician-facing transparency, we complemented rule-based reasoning with LIME explanations and LLM-generated narratives using Gemini 2.5 Flash.

Future work will prioritize external clinical validation on larger, multi-institutional datasets to assess robustness under dataset shift and real deployment constraints. Additional directions include broader ablations and sensitivity studies, improved rule-growth control via pruning or constrained fuzzy partitions, and more rigorous explanation evaluation incorporating clinician studies alongside multiple independent judging protocols.

ACKNOWLEDGMENT

This work was supported by the Ministry of Science and Higher Education of the Republic of Kazakhstan, "Smart-Care: Innovative Multi-Sensor Technology for Elderly and Disabled Health Management" (AP23487613, duration 2024–2026).

REFERENCES

[1] A. Yazici, D. Zhumabekova, A. Nurakhmetova, Z. Yergaliyev, H. Y. Yatbaz, Z. Makisheva, M. Lewis, and E. Ever, "A smart e-health framework for monitoring the health of the elderly and disabled," *Internet of Things*, vol. 24, p. 100971, 2023.

TABLE VIII
PERFORMANCE COMPARISON OF STUDY PROPOSITIONS WITH ANFIS-FRF (HOLDOUT WHERE AVAILABLE).

Dataset	Study	Accuracy	Specificity	Recall	F-1	Precision	NPV	MCC	AUC
Cleveland	**This study***	0.89	0.85	0.93	0.89	0.84	**0.93**	0.77	**0.95**
	Zeinulla et al. [7]	**0.93**	**0.97**	0.90	-	-	-	-	-
	Alharbi [18]	0.89	0.84	**0.94**	**0.90**	**0.87**	-	**0.80**	-
	Saberi et al. [19]	0.90	0.91	0.89	0.90	0.91	-	-	-
Wisconsin	**This study***	0.95	**0.93**	0.96	0.95	0.96	**0.93**	**0.89**	**0.99**
	Savitha and Umadevi [9]	**0.98**	-	0.98	0.97	0.97	-	-	-
	Eftekharian et al. [20]	0.94	0.91	0.97	-	-	-	-	0.95
	Saharan et al. [21]	**0.98**	-	**0.99**	**0.99**	**0.99**	-	-	-

* Direct comparison across studies is limited because preprocessing methods and evaluation settings differ. Therefore, holdout-test results are reported wherever available; for [7], cross-validation scores are shown because no holdout results were reported.

TABLE IX
ANALYTICAL COMPARISON OF STUDY PROPOSITIONS WITH ANFIS-FRF

Study	Data Balancing	Fuzzy Interpretability	LIME/SHAP	Multiple Datasets	LLM Explanation
This study*	**Yes**	**Yes**	**LIME**	**Yes**	**Yes**
Saharan et al. [21]	**Yes**	No	**SHAP**	No	No
Zeinulla et al. [7]	**Yes**	**Yes**	No	No	No
Alharbi [18]	**Yes**	No	No	No	No
Saberi et al. [19]	**Yes**	No	No	**Yes**	No
Savitha and Umadevi [9]	No	**Yes**	No	No	No
Eftekharian et al. [20]	No	No	No	No	No

[2] C. Metta, A. Beretta, R. Pellungrini, S. Rinzivillo, and F. Giannotti, "Towards transparent healthcare: advancing local explanation methods in explainable artificial intelligence," *Bioengineering*, vol. 11, no. 4, p. 369, 2024.

[3] W. Zhou, X. Liu, H. Bai, and L. He, "Intelligent medical diagnosis and treatment for diabetes with deep convolutional fuzzy neural networks," *Information Sciences*, vol. 677, p. 120802, 2024.

[4] J. Cao, T. Zhou, S. Zhi, S. Lam, G. Ren, Y. Zhang, Y. Wang, Y. Dong, and J. Cai, "Fuzzy inference system with interpretable fuzzy rules: Advancing explainable artificial intelligence for disease diagnosis—a comprehensive review," *Information Sciences*, vol. 662, p. 120212, 2024.

[5] M. L. Bangerter, G. Fenza, D. Furno, M. Gallo, V. Loia, C. Stanzione, and I. You, "A hybrid framework integrating llm and anfis for explainable fact-checking," *IEEE Transactions on Fuzzy Systems*, 2024.

[6] Y. Ren, X. Zhu, K. Bai, and R. Zhang, "A new random forest ensemble of intuitionistic fuzzy decision trees," *IEEE Transactions on Fuzzy Systems*, vol. 31, no. 5, pp. 1729–1741, 2022.

[7] E. Zeinulla, K. Bekbayeva, and A. Yazici, "Effective diagnosis of heart disease imposed by incomplete data based on fuzzy random forest," in *2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 1–9, 2020.

[8] P. Bonissone, J. M. Cadenas, M. Carmen Garrido, and R. Andrés Díaz-Valladares, "A fuzzy random forest," *International Journal of Approximate Reasoning*, vol. 51, no. 7, pp. 729–747, 2010.

[9] S. S and U. V, "Fuzzy modelling approach for accurate and explainable breast cancer prediction," *ICTACT Journal on Soft Computing*, vol. 16, pp. 3763–3768, 04 2025.

[10] A. U. Ruby, J. G. C. Chandran, T. S. Jain, B. Chaithanya, and R. Patil, "Rffe–random forest fuzzy entropy for the classification of diabetes mellitus," *AIMS Public Health*, vol. 10, no. 2, p. 422, 2023.

[11] M. Mesinovic, P. Watkinson, and T. Zhu, "Explainability in the age of large language models for healthcare," *Communications Engineering*, vol. 4, no. 1, p. 128, 2025.

[12] F. Gaber, M. Shaik, F. Allega, A. J. Bilecz, F. Busch, K. Goon, V. Franke, and A. Akalin, "Evaluating large language model workflows in clinical decision support for triage and referral and diagnosis," *npj Digital Medicine*, vol. 8, no. 1, p. 263, 2025.

[13] A. Janosi, W. Steinbrunn, M. Pfisterer, and R. Detrano, "Heart Disease." UCI Machine Learning Repository, 1989. DOI: https://doi.org/10.24432/C52P4X.

[14] W. Wolberg, O. Mangasarian, N. Street, and W. Street, "Breast Cancer Wisconsin (Diagnostic)." UCI Machine Learning Repository, 1993. DOI: https://doi.org/10.24432/C5DW2B.

[15] Y. B. Koca and E. Aktepe, "Evaluation of missing data imputation methods and pca techniques for machine learning models in breast cancer diagnosis using wbcd," *Türk Doğa ve Fen Dergisi*, vol. 13, no. 3, p. 109–116, 2024.

[16] G. Comanici, E. Bieber, M. Schaekermann, I. Pasupat, N. Sachdeva, I. Dhillon, M. Blistein, O. Ram, D. Zhang, E. Rosen, *et al.*, "Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities," *arXiv preprint arXiv:2507.06261*, 2025.

[17] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou, "Self-consistency improves chain of thought reasoning in language models," *arXiv preprint arXiv:2203.11171*, 2022.

[18] L. A. Alharbi, "Heart disease prediction of cleveland clinic patients using advanced machine learning algorithms," *Journal of Advanced Research Design*, vol. 126, no. 1, pp. 1–14, 2025.

[19] Z. A. Saberi, H. Sadr, and M. R. Yamaghani, "An intelligent diagnosis system for predicting coronary heart disease," in *2024 10th international conference on artificial intelligence and robotics (QICAR)*, pp. 131–137, IEEE, 2024.

[20] M. Eftekharian, M. Hosseiny, N. Motallebi, and S. Norouzkhani Esterabadi, "Accurate and interpretable breast cancer diagnosis using logistic regression: An evaluation on the wisconsin diagnostic dataset," *InfoScience Trends*, vol. 2, no. 9, pp. 52–60, 2025.

[21] S. Saharan, N. A. Wani, S. Chatterji, N. Kumar, and A. M. Almuhaideb, "A deep learning and explainable artificial intelligence based scheme for breast cancer detection," *Scientific Reports*, vol. 15, no. 1, p. 32125, 2025.