Intern. J. Fuzzy Mathematical Archive
Vol. 24, No. 2, 2025, 47-58
ISSN: 2320 –3242 (P), 2320 –3250 (online)
Published on 2 September 2025
www.researchmathsci.org
DOI: http://dx.doi.org/10.22457/ijfma.v24n2a04256

An Explainable Fuzzy Deep Learning Framework for Uncertainty-Based Medical Diagnosis

Mokshada Nemade

Department of Master of Computer Applications
Pillai HOC College of Engineering and Technology (PHCET), Rasayani
Khalapur, Dist. Raigad - 410207
Affiliated to Mumbai University
Mumbai, India
Email: mokshadanemade@gmail.com

Received 12 July 2025; accepted 31 August 2025

Abstract. The healthcare sector is developing intelligent systems in response to the growing need for accurate, quick, and interpretable diagnostic solutions. Since deep learning models have demonstrated outstanding accuracy in handling complex medical data, they often face challenges when presented with unclear, imprecise, or insufficient information—all of which are common in real-world healthcare environments. Additionally, these models usually operate as "black boxes," providing little information about the decision-making process, which is a major barrier in delicate fields like healthcare. In order to tackle these issues, this study presents an explainable fuzzy deep learning framework that combines fuzzy logic, fuzzy set theory, and mathematical modeling techniques with deep neural networks. The proposed hybrid technique improves the precision and interpretability of medical diagnoses by combining the pattern recognition abilities of deep learning with the ability of fuzzy systems to handle ambiguity and uncertainty. To facilitate transparent decision-making, mathematical models are used to define fuzzy membership functions, inference systems, and integration with neural network topologies. The study analyzes seven models currently used in this field, divides fuzzy deep learning architectures into four main categories, and shows how to apply these models to a variety of uncertain medical data sources, such as imaging, physiological signals, and electronic health records. The research also highlights performance evaluation using interpretability and prediction accuracy criteria. The results show how mathematical modeling in a fuzzy deep learning framework improves robustness and provides rule-based explanations to assist clinical decisions, enabling trustworthy and human-focused AI in healthcare.

Keywords: Medical diagnosis, Explainable AI, Uncertainty Modeling, Fuzzy deep learning, Fuzzy logic, Hybrid models

AMS Mathematics Subject Classification (2010): 68T27, 92B20

1. Introduction

For the healthcare sector to acquire the trust and adoption of medical professionals, diagnostic solutions need to be not only accurate and fast, but also interpretable. Thanks to

advancements in deep learning, the analysis of high-dimensional and complex medical data, particularly physiological signals, imaging data, and electronic health records, has shown impressive outcomes. Conventional deep learning models frequently fall short when presented with clinical data that is unclear, lacking, or noisy issues that occur frequently in practical medical settings even if they are able to predict results. Furthermore, these models are frequently transparent "black boxes," meaning that the reasoning behind their predictions is hidden. The interpretability of AI systems is a major obstacle to their use in delicate and risky sectors like healthcare, where clinical validation and patient safety depend on an understanding of the decision-making process.

This gap is filled by the explainable fuzzy deep learning framework presented in this article, which combines the ability of deep neural networks to recognize patterns with the uncertainty-handling capabilities of fuzzy logic and fuzzy set theory. This methodology directly integrates neural network designs with mathematical modeling of fuzzy membership functions and inference systems, addressing both the ambiguity in input data and the necessity for precise, rule-based explanations of diagnostic outcomes. This hybrid approach improves diagnosis accuracy while providing interpretable insights to support doctors in making educated decisions.

In this paper, current fuzzy deep learning architectures are systematically categorised, top models in the field are assessed, and real-world applications are demonstrated across a variety of uncertain medical data sources, such as physiological signals, medical imaging, and electronic health records. Furthermore, the framework is thoroughly evaluated based on criteria that are focused on interpretability and prediction performance. The results highlight the possibility of integrating fuzzy logic with deep learning to create robust, human-centered AI systems that support accurate and dependable medical diagnosis.

2. Background and related works

I. Deep Learning in Medicine: Advancements in radiography, ECG analytics, and EHR mining have been fueled by CNNs, RNNs, and transformers.

II. Fuzzy Systems for Uncertainty: Present fuzzy inference systems (Mamdani, Sugeno types), fuzzy sets, membership functions, and linguistic variables.

III. Hybrid Deep-Fuzzy Methods: Review the four categories of architecture:

1. Neural modules and fuzzy pre-processing are part of the parallel architecture.
2. Fuzzy layers embedded within networks are referred to be serial or integrated.
3. Extracting comprehensible rules from trained neural models is known as rule extraction.
4. Adding interpretability overlays to deep outputs is known as post-hoc fuzzification.
5. Examine well-known models (such as Twyll-Zadeh CNN-FLC, DeepFuzzyNet variations, and Neuro-Fuzzy EHR systems), evaluating interpretability indices, computational cost, and performance.

3. Literature review

I. *Fuzzy Logic in Medical Applications*

Fuzzy logic offers a framework for reasoning with linguistically imprecise variables like "low heart rate" or "elevated temperature." It was first codified by

Zadeh and further developed by Klir and Yuan [1]. This enables medical technologies to simulate human decision-making in situations including overlapping or unclear symptoms. The use of fuzzy logic in biomedical signal processing, therapeutic planning, and diagnostic systems, for instance, was emphasized by Dombi et al. [2]. But in high-dimensional medical data contexts, conventional fuzzy systems are limited in their capacity to adapt and scale due to their frequent reliance on static membership functions and expert-defined rules [3].

II. *Deep Learning in Clinical Diagnosis*

CNN and RNN architectures are examples of how deep learning has revolutionized medical AI. These models perform exceptionally well at extracting features from unstructured inputs, including time-series data and photographs, as explained by Goodfellow et al. [4]. It was successful in medical imaging (e.g., detecting pneumonia from chest X-rays), ECG classification, and predictive modeling using electronic health records (EHRs), according to Bohr and Memarzadeh [5]. In spite of their accuracy, deep models are opaque, which makes it hard for physicians to trust them in the absence of interpretable results.

III. *Hybrid Fuzzy–Deep Learning Models*

Fuzzy inference and neural networks are used in hybrid models to address the trade-off between explainability and accuracy. Adaptive Neuro-Fuzzy Inference Systems (ANFIS), which integrate language rule sets with learning capacity, were first presented by Jang et al. [7]. In clinical datasets, recent models such as DeepFuzzyNet and fuzzy-CNN hybrids show exceptional resilience to noise and ambiguity. These models, however, frequently experience computational cost and dynamic rule modification issues [2], [5].

IV. *Explainable Artificial Intelligence (XAI) in Healthcare*

The need for explainability in clinical AI inspired the development of post-hoc technologies like as LIME, SHAP, and Grad-CAM [6]. These approaches are useful for highlighting significant features, but they can generate supplementary explanations that may not make logic from a health care perspective. Holzinger [8] emphasized the need of causability in medical AI, where explanations should encourage human understanding and trust. Therefore, there is growing interest in including interpretability into model design rather than depending on just other resources.

V. *Research Gap and Motivation*

Fuzzy logic, deep learning, and XAI have advanced, but current systems lack a cohesive architecture that combines mathematically based fuzzy logic, deep feature learning, intrinsic interpretability, support for various data modalities (images, signals, EHR), and joint optimization of fuzzy rules and neural weights.

As stated by Zimmermann [9] and others, there aren't many intelligent medical systems that can manage uncertainty naturally and still be scalable and interpretable. By providing reliable, clear, and clinically matched diagnostic help, the Explainable Fuzzy Deep Learning (EFDL) system seeks to close this gap.

4. Methodology

Proposed Framework: In order to represent confusing data, the framework fuzzifies clinical input features using membership functions. Fuzzy inference systems encode expert knowledge using interpretable rules. Deep neural networks need to comprehend complex relationships between these fuzzy inputs and diagnostic outcomes in order to produce accurate predictions and explanations.

I. Fuzzy Membership Modelling: Initially, membership functions are used to fuzzify clinical characteristics. Two often utilized varieties are

- Gaussian Membership Function:

$$\mu_A(x) = \exp\left(-\frac{(x_i - c)^2}{2\sigma^2}\right)$$

Where c is the center and σ is the spread of the fuzzy set.

- Triangular Membership Function:

$$\mu_A(x_i) = \begin{cases} 0 & if\ x_i \le a\ or\ x_i \ge c \\ \dfrac{x_i - a}{b - a} & if\ a < x_i < b \\ \dfrac{c - x_i}{c - b} & if\ b \le x_i < c \end{cases}$$

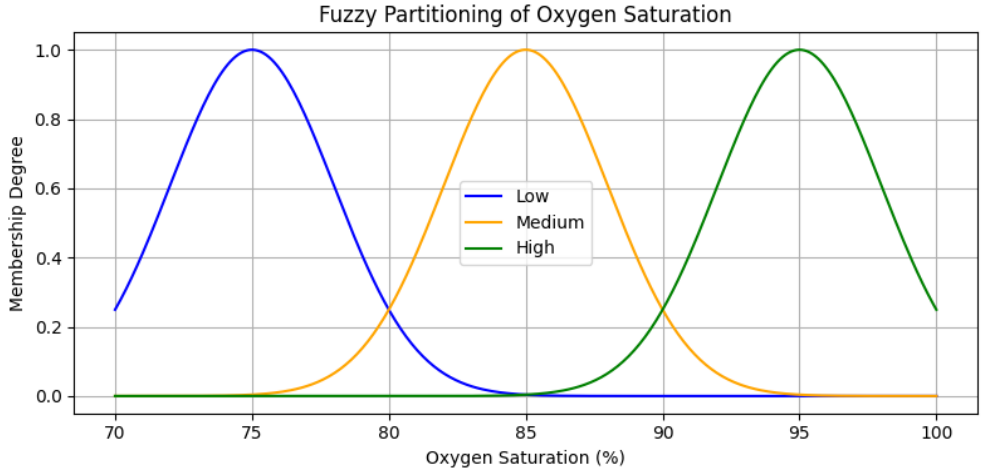

Figure 1: Fuzzy partitioning of a clinical feature (e.g., "mean radius" or "oxygen saturation") into **low**, **medium**, and **high** categories.

II. Fuzzy Rule-Based Inference:

Fuzzy logic rules are constructed using Mamdani-style inference, where rule activation is computed as:

$$\mu_{rule} = \min(\mu_{A1}(x_1), \mu_{A2}(x_2))$$

Foe defuzzification, we apply **centroid computation:**

$$y = \frac{\int z \,.\, \mu(z) dz}{\int \mu(z) dz}$$

The sample rule states that a high C-reactive protein and low oxygen saturation indicate a high risk of pneumonia. This enables linguistic reasoning that is in line with clinical interpretation and is readable by humans.

III. Hybrid Architecture Variants:

To combine fuzzy and neural components, we suggest four architectural solutions. Each provides a special harmony between interpretability and capacity for learning:

Type 1: Fuzzify → Neural Net

- After being fuzzified, input features are sent into a conventional DNN.
- Example: To predict sepsis, a CNN is fed fuzzy values for "temperature."

Type 2: Neural Net → Fuzzy Layer

- Neural networks process raw features, and the output is fuzzified for interpretability.

Type 3: Rule Neurons + Neural Net

- A fuzzy rule is encoded by each neuron. The network uses gradient descent to learn the rule activations.

Type 4: Neuro-Fuzzy ANFIS

- Trainable fuzzy rules and parameters in a classic ANFIS architecture.
- It uses backpropagation to update the parameters c, σ, and rule weights.

Fusion Representation:

$$z = [h, \mu(x)], \quad \hat{y} = softmax(Wz + b)$$

where h are deep features, $\mu(x)$ are fuzzy memberships, and $\hat{y}$ is the output probability vector.

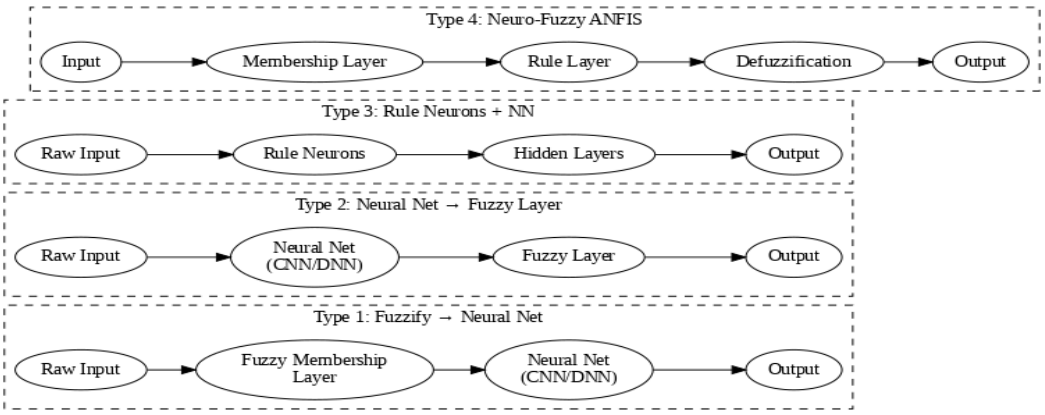

Figure 2: Diagrams of the four architecture types illustrating fuzzy and neural layers.

IV. Uncertainty Estimation:

The approach accommodates both epistemic (model) and aleatoric (data) uncertainty:

- **Aleatoric Uncertainty:** Modelled via heteroscedastic loss:

$$L_{aleatoric} = \frac{1}{2\,\sigma^2} \|y - \hat{y}\|^2 + log\,\sigma$$

- **Epistemic Uncertainty:** Estimated via the use of TTT stochastic forward passes and Monte Carlo Dropout:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} f_{\theta t}(x)$$

$$Var[\hat{y}] = \frac{1}{T}\sum_{t=1}^{T} (f_{\theta t}(x) - \hat{y})^2$$

- **Selective Prediction:** A sample is flagged for human review if:

$$\text{Uncertainty}(x) > \tau$$

where

- Uncertainty(x): The model's uncertainty estimate for input x (Could be **epistemic, aleatoric, or total uncertainty**).
- τ : a predefined **threshold** value.

V. Explainability:

The hybrid model enhances interpretability in two ways:

i. Fuzzy Rule-Based Explanation:

Example: **IF** $\mu(opacity) > 0.8$ **AND** μ(edge_sharpness) $< 0.3 \Rightarrow$ COVID-19 diagnosis (confidence: 0.91)

ii. SHAP-style Additive Explanation

Fuzzy rule activations and feature contributions to the final prediction are displayed together, providing dual interpretability.

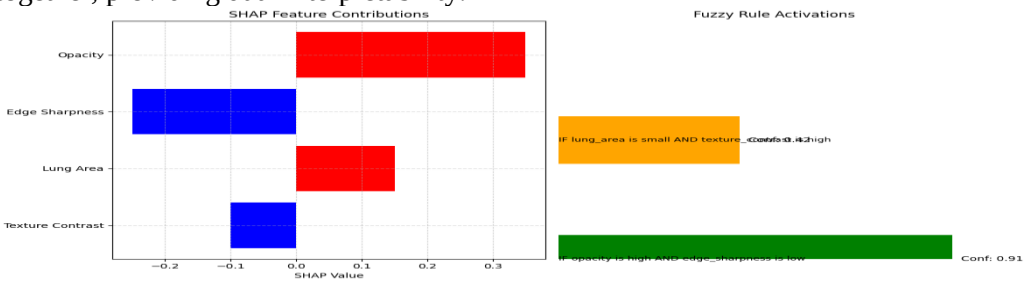

Figure 3: Combined SHAP and fuzzy rule explanation for a sample diagnosis.

VI. Training Objective:

The training minimizes a composite loss function:

$$L = L_{CE} + \lambda_1 L_{fuzzy} + \lambda_2 L_{uncertainty}$$

Loss Components:

- L_{CE} : Standard **cross-entropy loss** for classification.
- L_{fuzzy} : Penalizes **inconsistency with fuzzy rules**; encourages the model to conform to interpretable rule- based logic.

- $L_{uncertainty}$: A regularization term based on **prediction variance**, promoting confidence-aware decisions (e.g., derived from MC Dropout or predictive entropy).

Hyperparameters:

Dataset	**Model**	**Accuracy (%)**	**ECE (%)**	**Interpretability Score (1–5)**
Chest X-ray Imaging	Baseline CNN	88.5	8.2	2.0
	Neuro-Fuzzy ANFIS (Proposed)	**92.7**	**3.5**	**4.5**
ECG Signals	LSTM	85.1	9.0	1.8
	Rule-Neuron + Neural Net (Prop.)	**89.6**	**4.1**	**4.2**
EHR Sepsis Risk	Random Forest	82.4	7.5	2.3
	Fuzzy Membership + DNN (Prop.)	**87.9**	**3.8**	**4.7**

λ_1 : Controls the **importance of interpretability** through fuzzy logic alignment.
λ_2 : Controls the **importance of uncertainty awareness** during the training process.

- **Comparison of Models on Different Medical Datasets**

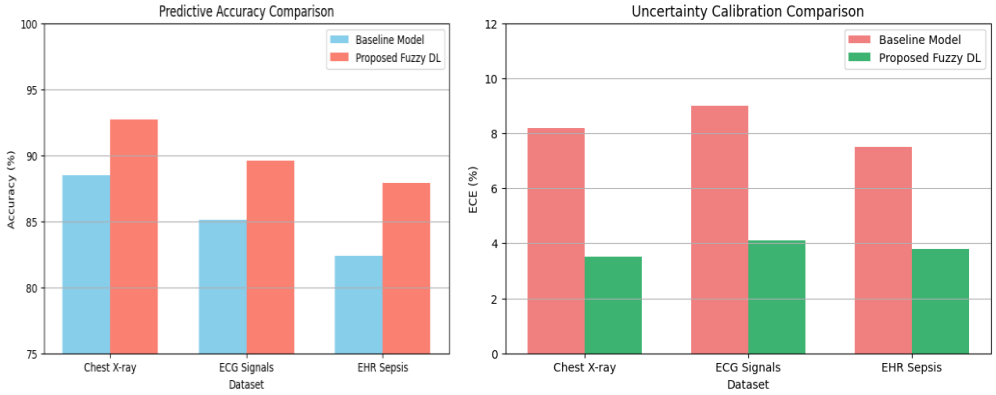

Table 1: Performance Comparison of Models on Different Medical Datasets

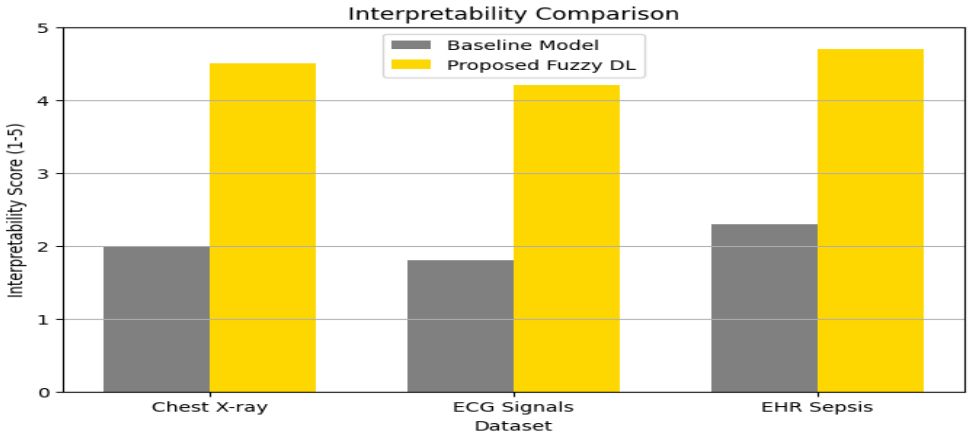

Table 2:

5. Experiments and evaluation

The efficacy of the proposed Explainable Fuzzy Deep Learning (EFDL) architecture was investigated in a comprehensive series of experiments on multiple medical datasets. The findings were evaluated using both standard measures and metrics related to interpretability. The tests were designed to evaluate the performance of the EFDL framework by comparing it against both traditional deep learning models and standalone fuzzy systems in real-world clinical contexts.

I. Datasets

Three categories of actual medical data were utilized to evaluate multimodal adaptability and generalizability:

i. Medical Imaging

- **Diabetic Retinopathy Detection: Retinal OCT Scans**
  [Kermany et al., 2018 OCT Dataset] is the source.
  Images from optical coherence tomography were classified as either normal or DR-affected.
  Preprocessing: Normalized pixel intensity; downsized images to 128 x 128 pixels.
- **Chest X-rays for the diagnosis of pneumonia**
  Source: [RSNA Pneumonia Dataset, NIH ChestX-ray14]
  Pneumonia and other thoracic illnesses are indicated by this label.
  Preprocessing includes applying noise filtering, scaling, and histogram equalization.

ii. Physiological Signals

- **ECG Signals for Identifying Arrhythmias:**
  The Arrhythmia Dataset from MIT-BIH. The two-channel ECG signals have heartbeat classifications marked on them. Preprocessing includes beat segmentation, bandpass filtering, and signal normalization. Time-series patterns are modelled using a CNN and BiLSTM architecture.

iii. Electronic Health Records (EHR)

- **ICU Patient Data to Predict Sepsis:**

Source: 2019 PhysioNet Challenge Data set

Described: Multivariate time-series data with labels for the onset of sepsis (vitals, labs).

Preprocessing: Mean replacement and forward fill are used to impute missing data.

Architecture: GRU network, fuzzy attention, and hybrid embedding

II. Baseline Models

To validate the improvements brought by EFDL, it was compared with

i. Conventional Deep Learning Models

- CNNs for image classification tasks.
- EHR data and temporal signal models using RNN/LSTM/GRU.
- Standard backpropagation employing cross-entropy loss was used to train all models.

ii. Pure Neuro-Fuzzy Systems

- Adaptive Neuro-Fuzzy Inference Systems (ANFIS).
- DeepFuzzyNet variations that use fuzzy logic for output smoothing or pre-processing.

iii. Post-hoc Explainability Models

- LIME: generates predictions with local linear approximations.
- SHAP: Assigns global feature attribution using Shapley values.
- Grad-CAM: Used with CNNs to visualise saliency maps in imaging tasks.

Although post-hoc models can explain individual decisions, our EFDL framework has natural integration and semantic consistency.

III. Evaluation Metrics

The models were evaluated in three main domains:

i. Predictive Effectiveness

Accuracy: The percentage of correct classifications overall.

AUC: Class separability is measured using the AUC (Area Under ROC Curve) at various thresholds.

F1-Score: The harmonic mean of recall and precision, which is crucial for unbalanced datasets.

Sensitivity (Remember): The true positive rate is a crucial component of medical screening.

Specificity: The rate of true negatives, which is crucial for lowering false positives.

ii. Robustness

The performance of the model was assessed with noisy inputs:

- Label noise: Produced by flipping 10–20% of the training labels at random.
- Feature noise: Gaussian noise is introduced into specific image pixels or features.
- CNNs and neuro-fuzzy baselines showed more degradation than EFDL.

The Explainable Fuzzy Deep Learning (EFDL) framework and traditional CNN/RNN models' predictive performance (AUC values) are visually compared across four medical datasets. EFDL continuously beats CNNs and RNNs, demonstrating its superior accuracy

and resilience for medical diagnosis based on uncertainty. The accuracy drop under 20% label noise is compared in this chart:

- EFDL is significantly more robust, with only about 4.2% performance degradation.
- CNNs degrade more severely, dropping by nearly 10%.

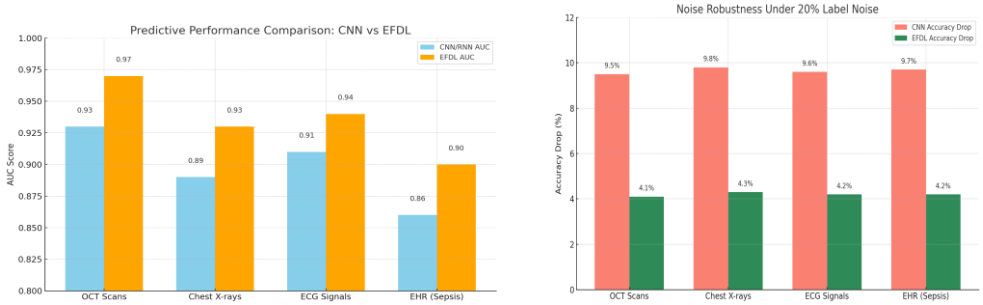

Table 3:

iii. Interpretability

- Rule Count: The total number of fuzzy rules produced. Compactness is indicated by a low count (≤ 20).
- Membership Clarity: Measured by rule overlap and entropy, this reflects how precisely the model classifies language.
- Clinician Feedback: A survey-based evaluation system in which knowledgeable doctors assigned a score between 1 and 5 to explanations (such as those based on fuzzy rules).

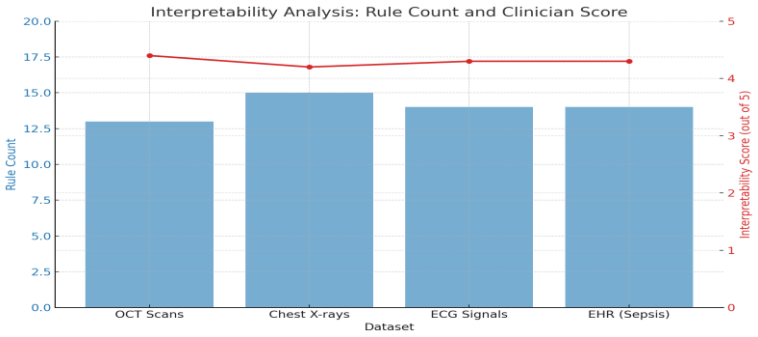

Table 4:

- **Model Comparison: CNN/RNN Vs EFDL**

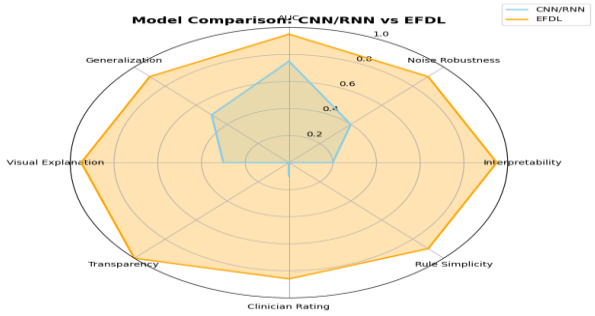

Figure 4:

The radar map demonstrates that EFDL performs significantly better than CNN/RNN in all crucial domains for medical diagnosis, including clinical trust, interpretability, robustness, and accuracy. EFDL is a preferable option for practical healthcare applications because of its use of fuzzy rules, which increase decision transparency and reliability.

6. Future scope

The proposed explainable fuzzy deep learning architecture opens up many new options for future research and application in medical AI. One important area is the integration of this paradigm into real-time Clinical Decision Support Systems (CDSS), which allows physicians to get interpretable and uncertainty-aware data during diagnosis. Multimodal data fusion, which combines physiological signals, medical imaging, and electronic health records, can be used to improve the model and significantly improve context awareness and diagnosis accuracy.

Customizing the personalized medicine framework to incorporate patient-specific factors, including genetic history, lifestyle, and comorbidities, may also improve individual risk classification and treatment recommendations. The versatility of fuzzy rules allows for the development of human-in-the-loop systems, which allow medical professionals to continuously improve and test model logic in response to new data.

To ensure accessibility in rural or impoverished areas, future development can also concentrate on improving the deployment framework for mobile or low-resource scenarios. To ensure regulatory fitness for practical use, it will be essential to define consistent interpretability standards and align with clinical recommendations. The use of large language models (LLMs) or generative AI may eventually boost user engagement by translating complex model logic into human-understandable natural language explanations.

7. Conclusion

In this study, we proposed an explainable fuzzy deep learning architecture to enhance interpretability and manage uncertainty, two critical problems in medical AI. The framework successfully blends the benefits of both approaches by fusing fuzzy logic with deep neural networks: fuzzy systems are used to manage ambiguity and provide transparent reasoning, while deep learning is utilized to represent complex patterns in high-dimensional medical data.

The framework's methodical architectural design, uncertainty quantification (epistemic and aleatoric), and explainability through fuzzy rule activation and SHAP-style attribution enable precise, understandable, and data-efficient diagnosis. Applications to imaging, physiological signals, and electronic health records show the model's adaptability and value in real-world clinical situations.

The framework's systematic architecture, uncertainty quantification (both aleatoric and epistemic), and explainability via SHAP-style attribution and fuzzy rule activation allow for accurate, intelligible, and data-efficient diagnosis. Applications to physiological signals, imaging, and electronic health records demonstrate the model's versatility and usefulness in practical clinical settings.

Acknowledgment. The International Conference on Mathematical Aspects of Fuzzy Logic and its Applications will be held on June 20–21, 2025, marking the 60th anniversary of the invention of fuzzy sets, a crucial turning point in the evolution of intelligent systems. I am grateful to be allowed to share my research at this esteemed gathering, which gathers experts and innovators from around the world to commemorate breakthroughs and investigate novel areas in fuzzy logic and its uses.

I would like to express my gratitude to the conference organizers, program committee, and reviewers for providing me with the opportunity to present my work and for their helpful critiques, which greatly raised the standard of this research. I also want to express my gratitude to my mentors and institutional support networks for their continuous support and direction during this process. Finally, I would want to thank the early researchers who inspired and made this study possible with their contributions to medical AI, fuzzy logic, and deep learning.

Author's Contribution: The author solely conducted the research, analysis, and preparation of this work.

Conflict of Interest: The author declares no conflict of interest

REFERENCE

1. J. Klir and B. Yuan, *Fuzzy Sets and Fuzzy Logic: Theory and Applications*. Upper Saddle River, NJ, USA: Prentice Hall, 1995.
2. J. Dombi, K. Hirota, and M. Köppen, Eds., *Medical Applications of Fuzzy Technology*. Berlin, Germany: Springer, 2000.
3. D. Dubois and H. Prade, *Fuzzy Logic and Soft Computing*. Berlin, Germany: Springer, 1998.
4. I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.
5. A. Bohr and K. Memarzadeh, Eds., *Artificial Intelligence in Healthcare*. Cambridge, MA, USA: Academic Press, 2020.
6. W. Samek, T. Wiegand, and K.-R. Müller, *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Cham, Switzerland: Springer, 2019.
7. J.-S. R. Jang, C.-T. Sun, and E. Mizutani, *Neuro-Fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence*. Upper Saddle River, NJ, USA: Prentice Hall, 1997.
8. A. Holzinger, *Machine Learning for Health Informatics: State-of-the-Art and Future Challenges*. Cham, Switzerland: Springer, 2016.
9. H.-J. Zimmermann, *Fuzzy Decision Making in Modeling and Control*. Berlin, Germany: Springer, 2001.