



NAZARBAYEV
UNIVERSITY

School of Engineering and Digital Sciences

Bachelor of Science in
Computer Science

**Domain-Agnostic Fuzzy Deep Learning
Framework for Explainable AI in
Multimodal Diagnostics and Decision
Support**

(Final Report for CSCI 409 Senior Project)

by

Batyrkhan Sharipbay
Dias Akhmejanov
Kamilyam Nazarkhanova

Project Supervisor: Adnan Yazici

April 2026

Abstract

Despite achieving strong predictive accuracy in medical diagnosis, the widespread adoption of deep learning-powered clinical decision support systems is hindered by a fundamental conflict between this accuracy and the need for interpretability. This project addresses that tension by designing, implementing, and evaluating the Domain-Agnostic Fuzzy Deep Learning (DFDL) Framework. This multimodal, multidomain clinical decision support system integrates the intrinsic interpretability of fuzzy logic with the representational power of deep neural image encoders.

Existing deep learning models present a core problem: their operation is opaque, functioning as "black boxes." While they generate predictions, they cannot provide human-readable explanations detailing the specific clinical factors that led to a particular decision. This is unacceptable in healthcare, where clinicians require transparent reasoning to validate AI-assisted diagnoses and ensure patient safety. Moreover, no single imaging modality captures the full diagnostic picture, as effective systems must fuse volumetric imaging data with structured electronic health records (EHRs).

The DFDL framework addresses both challenges through a two-stream hybrid architecture. Raw volumetric scans (DICOM, NIfTI) are passed through a frozen MedicalNet-pretrained 3D-ResNet-10 backbone to produce 512-dimensional L2-normalized feature vectors, which are stored as NPZ files for efficient reuse. These image features are processed by a flat MLP encoder (FlatEncoder). At the same time, clinical tabular data is modelled by an Adaptive Neuro-Fuzzy Inference System (ANFIS) using mixed Gaussian and trapezoidal membership functions. A learnable dynamic gate fuses both streams per-sample, automatically falling back to clinical-only inference when imaging is absent. A FastAPI backend exposes the model through structured REST endpoints, and a Retrieval-Augmented Generation (RAG) copilot powered by ChromaDB and NVIDIA NIM (Llama 3.1) translates model outputs into natural-language explanations for clinicians.

The framework was evaluated on the OASIS-1 cross-sectional Alzheimer's dataset (BAcc=0.9052±0.0116, AUC=0.9593±0.0088), OASIS-2 longitudinal Alzheimer's dataset (BAcc=0.8465±0.0260, AUC=0.8836±0.0288), MMIST-ccRCC renal cell

kidney cancer dataset ($\text{BAcc}=0.6506\pm0.0439$, $\text{AUC}=0.6188\pm0.0563$), and UTSW-Glioma molecular brain tumor dataset ($\text{BAcc}=0.8353\pm0.0327$, $\text{AUC}=0.8347\pm0.0545$). All results are produced via unbiased nested cross-validation (5-outer and 3-inner folds). The system's configuration-driven architecture allows for easy generalization to any medical domain, enabling the addition of different imaging modalities or the switching of datasets. Overall, this project demonstrates the design, implementation, and evaluation of a complete computing-based solution, from raw data ingestion to user-facing clinical explanation, addressing a real and significant gap in medical AI deployability.

Table of contents

Abstract	1
Table of contents	3
1 Introduction and Problem Statement	1
2 Identification of constraints	2
3 Background and Related Work	3
3.1 Fuzzy Deep Learning Approaches	3
3.2 Multimodal Data Fusion Approaches	4
3.3 Synthetic data generation and LLM for interpretability	4
4 Methodology	6
4.1 Method selection	6
4.2 Software architecture	7
4.2.1 ML Model Architecture	7
4.2.2 Application Backend architecture	8
4.3 Third-party components and integration	9
4.4 Workflows	10
4.4.1 Prediction workflow	10
4.4.2 RAG Copilot workflow	11
4.5 Features	11
4.6 Algorithms and techniques used	12
4.7 Use cases	12
4.8 Team collaboration and development strategy	13
5 Project Execution	14
5.1 Semester 1: Requirements, Planning, and Initial Prototyping	14
5.2 Semester 2: Full System Implementation and Evaluation	14
6 Evaluation	17
6.1 Model Evaluation	17
6.2 Software Evaluation	20
7 Conclusion & Future Work	22
7.1 Key Findings and Contributions	22
7.2 Future Work	22
8 References	24
Appendix A. Dataset Descriptions	27

1 Introduction and Problem Statement

Deep Learning (DL) has demonstrated remarkable performance in the medical field through training on large datasets and utilizing complex learning methods [1]. However, DL approaches often lack transparency in model predictions, which has sparked criticism regarding their "black-box" nature that could potentially lead to biased and poor predictions without the understanding of clinicians [2]. This opacity prevents Deep Neural Networks (DNNs) from being widely accepted in healthcare, where transparency and interpretability are the highest priority [3]. Recent empirical work confirms that while explainable AI (XAI) methods can increase clinician trust, post-hoc approximations (such as SHAP or LIME) do not reliably reflect the model's actual decision mechanism and may lead to both overtrust and undertrust, reinforcing the need for intrinsic, rule-based interpretability [4][5].

A second, related problem concerns data modality. Clinicians routinely consider multiple heterogeneous data types in their decisions: volumetric image scans (CT, MRI), genomic markers, and structured patient records. As noted by Nie [6], no single-modality image can capture the full information required for a reliable diagnosis, making multimodal fusion a critical requirement for any clinically viable system. Recent comprehensive reviews confirm that combining imaging features with clinical EHR variables significantly enhances diagnostic accuracy and disease subtype differentiation, while identifying data heterogeneity and uncertainty quantification as the two most critical challenges for real-world deployment [7]. Importantly, such modalities as clinical patient records and medical imagery must be considered in a unified representation to enhance diagnostic precision.

Existing solutions either sacrifice interpretability for performance (pure DNN approaches) or sacrifice performance for interpretability (traditional rule-based or pure fuzzy systems). There is currently no well-established framework that simultaneously achieves strong predictive accuracy, genuine intrinsic interpretability, robustness to missing imaging modalities, and cross-domain transferability.

This project addresses all four requirements through the Domain-Agnostic Fuzzy Deep Learning (DFDL) framework, which integrates the intrinsic interpretability of Fuzzy Logic with the adaptive learning capabilities of DNNs. The framework is multidomain by design: it was trained and evaluated on four medical datasets, covering Alzheimer's disease (OASIS-1 [8], OASIS-2 [9]), renal cell carcinoma (MMIST-ccRCC [10]), and glioma molecular classification (UTSW-Glioma [11]), demonstrating that the same architecture generalizes

without modification. A complete software system, including a FastAPI backend and a RAG-based explainability copilot, makes the model accessible to clinicians through a web interface.

This report is organized as follows. Section 2 identifies project constraints. Section 3 reviews related work and justifies the methodology. Section 4 describes the project methodology in detail. Section 5 covers project execution across two semesters. Section 6 presents evaluation results. Section 7 concludes with future directions. Supporting information is provided in the appendices.

2 Identification of constraints

The development of a clinical AI system is shaped by a broad range of constraints that extend well beyond purely technical concerns. The following identifies and assesses each major constraint category.

Ethical and Regulatory. Clinical AI systems that process patient data require institutional ethical approval and, where real patient records are involved, compliance with data protection regulations equivalent to HIPAA or GDPR. Because ethical clearance from clinical partners was not available within the project timeline, all training and evaluation were restricted to publicly available, de-identified datasets (OASIS-1/2, MMIST-ccRCC, UTSW-Glioma). This is appropriate for a proof-of-concept but introduces a risk of reduced generalizability to real clinical data distributions, which may differ from research datasets in terms of scanner protocols, demographic composition, and label noise.

Health and Safety. The framework is intended to support, not replace, clinician judgment. Any erroneous prediction that a clinician acts upon without scrutiny could seriously affect patient safety. This constraint shaped several design decisions: the system reports calibrated probability scores alongside predictions, surfaces the top-fired fuzzy rules for audit, and includes an abstention mechanism in the RAG copilot that refuses to generate explanations when retrieved context does not sufficiently support the query, preventing hallucination in a safety-critical context.

Data Heterogeneity. Medical imaging data exists in multiple incompatible formats (DICOM series, NIfTI `.nii`, compressed `.nii.gz`, Analyze `.hdr/.img`). Normalizing these into a unified pipeline required writing modality-specific loaders and preprocessing routines (HU

windowing for CT, percentile clipping for MRI, log-scale normalization for PET), with a risk of information loss during format conversion and spatial resampling.

Economic. Volumetric 3D image inference is computationally expensive. Running the feature extractor at scale on cloud GPU infrastructure (e.g., Google Colab Pro, NVIDIA NIM API) incurs ongoing costs. The two-stage design (extracting features offline once and caching them as NPZ files) mitigates this by ensuring that training and inference never require GPU access to the backbone.

Social and Stakeholder Impact. The primary beneficiaries are clinicians and, indirectly, patients. Incorrect model outputs presented with high confidence could create automation bias, causing clinicians to over-reliance on AI recommendations. The interpretability layer (ANFIS rules, fusion weights) is designed to counteract this by surfacing the reasoning process, enabling clinicians to detect implausible rule activations.

Sustainability. Caching pre-extracted features eliminates repeated forward passes through the 3D backbone, reducing energy consumption during iterative experimentation. The use of a small, CPU-deployable model (ResNet-10) rather than a large transformer-based encoder was also motivated partly by deployment accessibility in resource-constrained clinical environments.

Dataset and Privacy Constraints. All datasets used are publicly available research datasets with de-identified records. Storage of raw scans and EHR data for future retraining is included as a planned system feature, subject to the proviso that institutions must grant explicit consent before the system stores identifiable data.

3 Background and Related Work

3.1 Fuzzy Deep Learning Approaches

The combination of fuzzy logic and deep learning has been explored from several architectural directions. Zhou et al. [3] proposed the Interpretable Predictor with Deep Convolutional Fuzzy Neural Network (IP-DCFNN) for diabetes detection, which uses a fuzzy antecedent layer to convert numerical inputs into fuzzy rule weights, a deep convolutional network to extract hidden features from those weights, and a fuzzy consequent layer to defuzzify the output. The authors report approximately 7.4% accuracy improvement over standalone ANFIS and DNN baselines on two diabetes datasets. Although limited to single-modal tabular data, this work demonstrates that hybrid models can provide transparent

interpretations while outperforming both pure deep and pure fuzzy systems. More recently, Mahdi et al. [12] proposed a fuzzy evaluation framework combined with explainable ML for diagnosing seven rheumatic and autoimmune diseases across 12,085 patients, demonstrating that fuzzy-XAI hybrids outperform standard ML on complex multiclass clinical datasets and further validating the clinical applicability of the fuzzy-DL paradigm.

Zhang [13] studied a deep neuro-fuzzy framework for speech emotion recognition where fuzzy systems and DNNs cooperate in a closed loop: fuzzy logic prepares features in the fuzzification layer, the DNN refines them, and fuzzy logic interprets the final output in the defuzzification layer. This cooperative-based architecture outperforms baseline deep-only, fuzzy-only, and pre-trained models. Hasan and Khidhir [14] provide a broader taxonomy in their survey of deep fuzzy systems, classifying frameworks into cooperative-based, sequential-based, and parallel-based architectures. The DFDL framework falls into the parallel-based category, where each stream (ANFIS for clinical data, FlatEncoder for image features) processes its modality independently before being combined by the fusion gate.

3.2 Multimodal Data Fusion Approaches

Kumar et al. [15] provide a comprehensive review of multimodal fusion in medical diagnosis, covering four main fusion strategies: early fusion (concatenation of raw features before any model), late fusion (combining predictions from separate modality-specific models), hybrid fusion (combining intermediate representations), and transformer-based fusion. The DFDL framework implements a learnable hybrid fusion, where both streams produce logits that are combined by a dynamic gate. This is more flexible than fixed-weight late fusion and more computationally tractable than transformer-based cross-attention for small medical datasets. Recent reviews confirm this trend: multimodal AI in medicine is advancing rapidly from traditional feature concatenation toward dynamic fusion architectures that adaptively weight each modality based on data availability and stream confidence [7][16].

The MMIST-ccRCC dataset introduced by Mota et al. [10] provides one of the few real-world multimodal benchmarks with five modalities (CT, MRI, histopathology, genomics, clinical data) and pre-defined train/test splits, making it an appropriate challenge dataset for evaluating cross-modal robustness. Their proposed multiple-instance learning approach provides a useful baseline for the DFDL system's survival prediction performance.

3.3 Synthetic data generation and LLM for interpretability

A challenge common to all clinical AI projects is the scarcity of labelled training data. Li et al. [17] introduced EHR-M-GAN, demonstrating that GAN-based synthesis of mixed-type health records from over 141,000 patients can meaningfully augment training sets while passing formal privacy checks. Miura et al. [18] compare synthesizers on small, mixed-type medical datasets, finding that Gaussian Copula models consistently outperform GAN-based methods in univariate consistency and downstream predictive accuracy on small datasets. These findings motivated an early investigation of synthetic data generation for the MMIST-ccRCC clinical features in Semester 1, using SDV's GaussianCopula library. However, experiments showed that the class imbalance in MMIST could be more effectively handled through a combination of 'WeightedRandomSampler' and Smoothed Focal Loss, without introducing distributional artefacts from synthetic data. Synthetic data generation was therefore not included in the final system.

For the explainability component, Van Veen et al. [19] demonstrate in their RadAdapt framework that T5 variants with lightweight LoRA fine-tuning can generate clinically coherent and factually grounded radiology report summaries, achieving high scores in both automatic metrics and clinician evaluations. Their findings justify the use of domain-adapted LLMs in the RAG copilot and support the importance of grounding LLM outputs in retrieved clinical evidence rather than relying on parametric knowledge alone. More recent work has extended this foundation: a 2025 study demonstrated that GPT-4 augmented with RAG achieved 78% accuracy in identifying primary diagnoses compared to 54% for the base model alone, and provided correct differential diagnoses in 98% of cases [20]. A comprehensive review of RAG for LLMs in medicine further shows that while RAG significantly improves factual grounding, standardized evaluation methods and attention to hallucination in safety-critical contexts remain open challenges [21], motivating the DFDL copilot's abstention mechanism. The DFDL copilot uses ANFIS-generated fuzzy rules as a structured context for the LLM, ensuring that the natural-language explanation is anchored to the specific features and thresholds identified by the model for the given patient.

4 Methodology

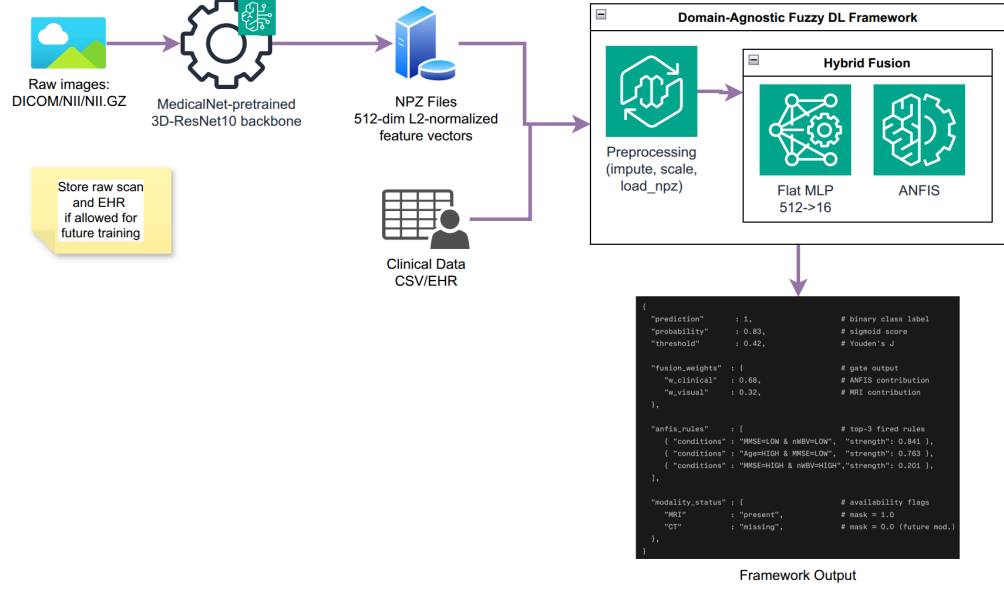


Figure 1. Abstract diagram of the DFDL architecture

4.1 Method selection

The central design challenge was selecting a fusion architecture that simultaneously achieves three objectives that are typically in tension: high predictive performance, intrinsic (not post-hoc) interpretability, and robustness to missing imaging modalities.

Several XAI approaches were considered. Post-hoc methods such as SHAP and LIME generate feature importance scores by approximating the model's behavior around a given prediction, not reflecting the model's actual decision mechanism. ANFIS, by contrast, is intrinsically interpretable: its IF-THEN fuzzy rules are a direct part of the model's computation, not an external approximation. For the clinical tabular stream, this distinction is critical. A clinician reading "IF MMSE is LOW AND nWBV is LOW THEN Dementia" can assess whether the rule is medically plausible. A SHAP attribution score on a black-box model carries no such guarantee.

For image features, applying fuzzy rules directly is not tractable. A 512-dimensional feature vector with 2 fuzzy sets per feature would generate 2^{512} rules, far exceeding any interpretable or computationally feasible rule base. The flat MLP encoder is therefore used for imaging data (it projects the 512-dim NPZ vector into a 16-dimensional latent space),

while the ANFIS handles only the 3–4 selected clinical features where interpretability is achievable. This hybrid design preserves interpretability where it matters most (clinical variables) while retaining the representational power of DNNs for imaging.

Transformer-based fusion was considered and rejected for two reasons: (1) the available datasets are small (OASIS-1: ~416 subjects, OASIS-2: ~373 visits), where attention-based cross-modal fusion tends to overfit without large-scale pretraining; and (2) the cross-attention mechanism does not produce interpretable clinical rules.

4.2 Software architecture

4.2.1 ML Model Architecture

The core ML model architecture (see Figure 2) operates as follows:

- **Offline Feature Extraction (Feature Extractor):** Raw volumetric scans (DICOM, .nii, .nii.gz) are preprocessed (HU windowing for CT, percentile-based normalization for MRI, spatial resampling to $96 \times 128 \times 128$) and passed through a frozen MedicalNet-pretrained 3D-ResNet10 backbone. The output is a 512-dimensional L2-normalized feature vector saved as a .npz file. This step runs once per subject, offline.
- **FlatEncoder (Imaging Stream):** A residual MLP (Linear $\rightarrow$ LayerNorm $\rightarrow$ GELU $\rightarrow$ Dropout) projects the 512-dim NPZ vector to a 16-dim latent space.
- **ANFIS (Clinical Stream):** A TSK ANFIS with mixed Gaussian and 4-parameter trapezoidal membership functions models the nonlinear relationships in 3–4 L1-selected clinical features. Membership function blend (Gaussian vs. trapezoidal), t-norm type (product vs. min), and TSK consequent weights are all learned end-to-end. Auxiliary regularization losses enforce crisp LOW/HIGH membership partitions, sparse rule activation, and diverse rule firing patterns.
- **Fusion Gate:** A 2-layer MLP takes as input the clinical stream logit, the ANFIS firing entropy (uncertainty signal), the visual stream logit, and the per-modality availability masks. It outputs the weights of the clinical (ANFIS) stream and the imaging (FlatEncoder) stream. If no imaging modality is present (all masks = 0), the gate forces the imaging stream to be 0 while clinical to be 1.0, enabling clinical-only inference. Entropy-based confidence refinement further blends gate weights with per-stream prediction certainty.

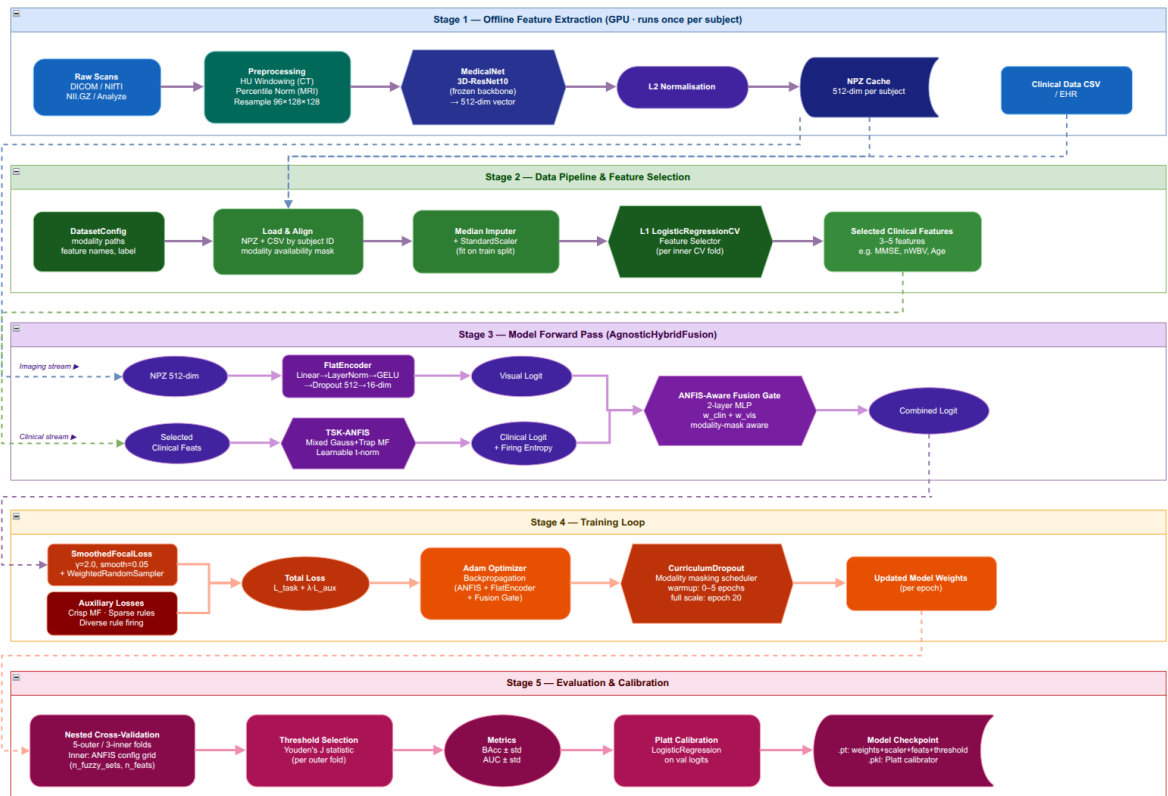


Figure 2. Full Training and Evaluation Pipeline

4.2.2 Application Backend architecture

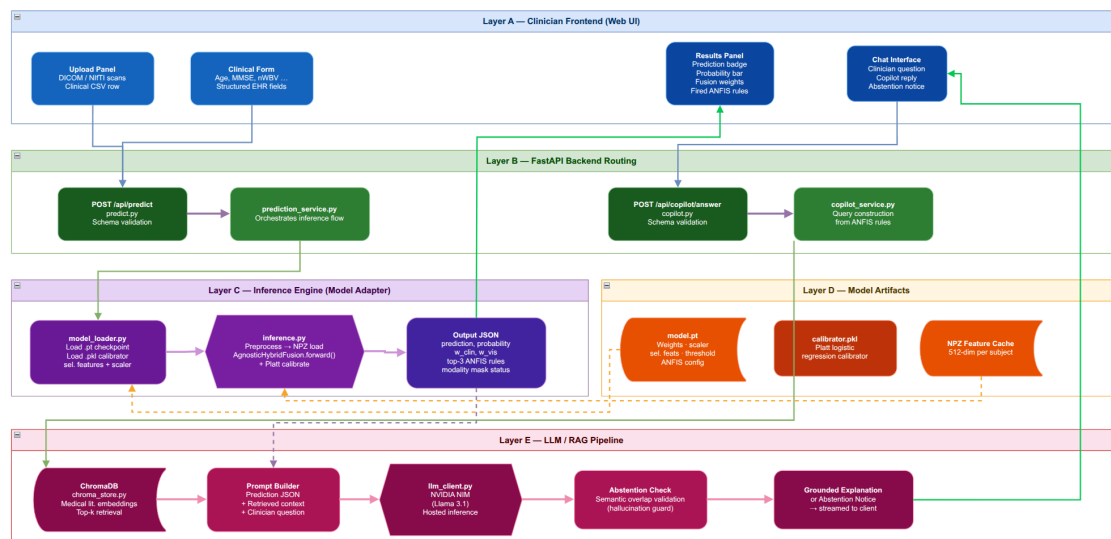


Figure 3. End-to-End System Architecture

Our application is designed as a modular healthcare AI platform that integrates a predictive fuzzy-deep learning model with an explainable AI copilot (see Figure 3). The backend is implemented using FastAPI and structured into clearly separated layers to ensure scalability and maintainability. The architecture includes the following layers:

- **API Layer (Routes)**
 - Handles incoming HTTP requests.
 - Defined in: `app/api/predict.py` and `app/api/copilot.py`
- **Service Layer**
 - Contains core operational logic and orchestrates interactions between components.
 - Key modules:
 - `prediction_service.py`
 - `copilot_service.py`
- **Model Adapter Layer**
 - Encapsulates interaction with the fuzzy deep learning model
 - Files:
 - `model_loader.py` (checkpoint loading)
 - `inference.py` (preprocessing, forward pass, output formatting)
 - Responsible for:
 - Model loading
 - Input preprocessing
 - Output formatting
- **RAG Components:**
 - `chroma_store.py`: Manages the ChromaDB vector database for storing and retrieving medical literature embeddings.
 - `llm_client.py`: manages communication with NVIDIA NIM (Llama 3.1) for natural-language generation.

4.3 Third-party components and integration

The system utilizes several third-party tools to accelerate development and ensure high-quality performance:

Component	Role	Integration
MedicalNet (3D-ResNet10) [22]	Pretrained 3D CNN backbone for volumetric image encoding	Weights loaded via <code>load_medicalnet_weights()</code> in <code>feature_extractor.py</code>

Component	Role	Integration
ChromaDB	Vector database for medical literature embeddings, retrieves top-k relevant documents for a given query	Integrated via <code>chroma_store.py</code> ; retrieves top-k relevant documents for RAG context
NVIDIA NIM (Llama 3.1)	Hosted LLM inference service	Integrated via <code>llm_client.py</code> ; eliminates the need for local GPU infrastructure
PyTorch	Model definition, training, and inference	Core DL framework throughout
nibabel / SimpleITK / pydicom	Medical image format loading	Used in <code>feature_extractor.py</code> for NIfTI, DICOM, and Analyze format support
scikit-learn	L1-penalized feature selection, StandardScaler, metrics	Used in the data pipeline and evaluation
FastAPI	REST API framework	Exposes <code>/predict</code> , <code>/explain</code> , and copilot endpoints

4.4 Workflows

4.4.1 Prediction workflow

1. Client sends patient data (scan files + clinical CSV row) to POST `/api/predict`.
2. “`predict.py`” route receives a request and validates the input schema.
3. “`prediction_service.py`” processes the request.
4. `model_loader` loads the model checkpoint.

5. “inference.py”: preprocesses the input: if raw scans are provided, runs feature extraction; applies median imputation and StandardScaler to clinical features; loads the corresponding NPZ vector.
6. `AgnosticHybridFusion.forward()` runs the inference pass.
7. Response includes: binary prediction, calibrated probability, Youden's J threshold, fusion weights of clinical and imaging streams, top-3 fired ANFIS rules with strengths, and per-modality availability status.

4.4.2 RAG Copilot workflow

1. Client sends a natural-language question and prediction JSON to `POST /api/copilot/answer`.
2. `copilot_service`:
 - Retrieves relevant medical context from vector DB
 - Constructs a prompt using the retrieved context and model outputs
3. NVIDIA NIM generates a response via Llama 3.1
4. The abstention mechanism evaluates whether the retrieved context aligns with the query and the model output.
 - If misaligned, the system abstains instead of hallucinating
5. Final response is either a grounded explanation or an abstention message

4.5 Features

- **Intrinsic Explainability:** Predictions are accompanied by human-readable fuzzy IF-THEN rules (e.g., IF MMSE is LOW AND nWBV is LOW THEN high dementia risk), confidence scores, and per-modality contribution weights.
- **Modality-Agnostic Inference:** The system can process any combination of available imaging modalities and automatically falls back to clinical-only inference when imaging is absent.
- **RAG-Based Clinical Copilot:** Provides context-aware natural-language explanations by retrieving relevant medical literature and synthesizing it with the model's fuzzy rules.
- **Abstention Mechanism:** Prevents hallucination by refusing to generate explanations when the retrieved context does not sufficiently support the query.
- **Domain-Agnostic Configuration:** Switching datasets or adding new imaging modalities (PET, fMRI, WSI) requires only a new `DatasetConfig` instance.

4.6 Algorithms and techniques used

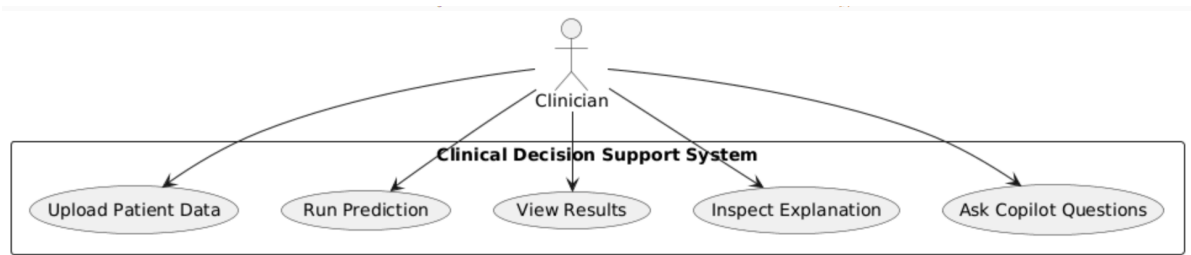
- **TSK-ANFIS with Mixed Membership Functions:** Mixed Gaussian and 4-parameter trapezoidal MFs with a learnable blend parameter and learnable t-norm.
- **CurriculumDropout:** A training scheduler that gradually introduces random modality masking (warmup epochs 0–5, full scale at epoch 20), training the model to be robust under partial data availability.
- **Nested Cross-Validation (5-outer / 3-inner):** Provides unbiased performance estimates while performing inner-loop hyperparameter selection (ANFIS fuzzy sets, number of selected features).
- **Smoothed Focal Loss ($\gamma=2.0$):** Suppresses overconfident predictions on majority-class examples; combined with **WeightedRandomSampler** to handle class imbalance.
- **L1-Penalized LogisticRegressionCV:** Selects the most predictive clinical features before ANFIS training, reducing rule dimensionality and preventing rule explosion.
- **Retrieval-Augmented Generation (RAG):** Embedding-based vector similarity search over clinical literature combined with LLM reasoning for grounded explanations.
- **Rule-Based Abstention Validation:** Checks overlap between retrieved document content, model features, and fired fuzzy rules before committing to a generated response.

4.7 Use cases

The system exposes five primary use cases for the Clinician actor within the Clinical Decision Support System:

- **Upload Patient Data:** The clinician uploads raw scan files (DICOM/NIfTI) and/or a structured clinical data CSV for a patient.
- **Run Prediction:** The system runs the full inference pipeline and returns the structured prediction output.
- **View Results:** The clinician reviews the prediction badge, calibrated probability bar, fusion weight breakdown, and fired ANFIS rules.
- **Inspect Explanation:** The clinician inspects the ANFIS fuzzy rules and their conditions to audit the model's clinical reasoning.

- **Ask Copilot Questions:** The clinician asks follow-up questions in natural language; the RAG copilot retrieves supporting literature and generates a grounded explanation.



4.8 Team collaboration and development strategy

The project team of three members divided responsibilities to leverage individual strengths while maintaining integration awareness across the full system:

- **Batyrkhan Sharipbay** led the development of **AgnosticHybridFusion** architecture, ANFIS module, training pipeline (nested CV, loss functions, curriculum dropout), and **feature_extractor.py**. Also, a project manager.
- **Dias Akhmejanov** was responsible for FastAPI backend, LLM/RAG copilot (ChromaDB, NVIDIA NIM, abstention mechanism), backend–frontend integration.
- **Kamilyam Nazarkhanova** contributed across both workstreams: dataset acquisition and processing for OASIS-1 [8], OASIS-2 [9], MMIST-ccRCC [10], and UTSW-Glioma [11]; model testing and validation; and co-implementation of the feature extractor and fusion components.

Weekly team meetings were held to synchronize progress, review integration issues, and resolve technical blockers. Version control was managed via Git, with pull request reviews before merging into the main branch.

5 Project Execution

5.1 Semester 1: Requirements, Planning, and Initial Prototyping

The first semester focused on defining project requirements, surveying related work, and implementing an initial proof-of-concept prototype. The team implemented an initial version of the IP-DCFNN model (termed IP-FNN in the project), drawing from Zhou et al. [3]. Evaluated on the MMIST-ccRCC dataset, this prototype achieved 66.49% balanced accuracy, which is below expectations but informative. After analyzing, we found two key problems: (1) applying fuzzy rules to high-dimensional image features produces exponential rule proliferation (with 512 image features and even 2 fuzzy sets per feature, the rule base would contain 2^{512} rules), destroying interpretability; and (2) the MMIST-ccRCC dataset has significant class imbalance (alive vs. deceased patients at 12 months), which the initial model did not account for. These findings directly shaped the architectural redesign in Semester 2.

On the data side, the team investigated synthetic data generation as a potential solution to class imbalance and data scarcity, implementing and evaluating GaussianCopula and CTGAN synthesizers on the Cleveland Heart Disease dataset [23]. GaussianCopula produced higher univariate fidelity scores, consistent with findings by Miura et al. [18]. However, both synthesizers introduced distributional artefacts more disruptive than helpful on small datasets, leading the team to abandon this approach.

A mockup of the clinician-facing web application was also produced in Semester 1, establishing the user interface requirements for the Intelligent Chat Companion feature and the prediction results display.

5.2 Semester 2: Full System Implementation and Evaluation

The second semester involved a comprehensive redesign and full implementation of all system components.

The IP-FNN prototype was replaced with the [AgnosticHybridFusion](#) architecture. The key architectural insight of separating the ANFIS for L1-selected clinical features and the flat MLP for the 512-dim image embedding resolved both the rule explosion problem and the performance degradation observed in Semester 1. The ANFIS was extended with mixed membership functions (Gaussian + 4-parameter trapezoidal with learnable blend), a

learnable t-norm, and a four-term auxiliary regularization loss to encourage crisp, sparse, and diverse rule firing.

The team addressed the MMIST class imbalance problem through a combination of `WeightedRandomSampler` (oversampling the minority class during training) and `SmoothedFocalLoss` ($\gamma=2.0$, label smoothing=0.05), which focuses training loss on misclassified minority-class examples. This replaced the earlier synthetic data approach.

A standalone offline feature extraction script (`feature_extractor.py`) was implemented to support DICOM, NIfTI, and Analyze format loading, modality-specific normalization (HU windowing for CT, percentile normalization for MRI), spatial resampling, and batch feature extraction with manifest logging. Integration of MedicalNet pretrained ResNet-10 weights [22] provided strong initialization for the 3D backbone.

We implemented the full FastAPI backend with prediction and copilot endpoints. The RAG copilot was integrated using ChromaDB as the vector store and NVIDIA NIM (Llama 3.1) for inference, avoiding the need to host an LLM locally. The abstention mechanism was implemented to prevent hallucinated medical explanations.

Key Challenges:

- *Hyperparameter sensitivity:* The ANFIS configuration (number of fuzzy sets, number of selected features) had a significant impact on performance, which was addressed via nested cross-validation with an inner hyperparameter search grid.
- *Dataset format heterogeneity:* OASIS-2 stores scans in `.hdr/.img` analyze format rather than NIfTI, requiring separate discovery logic (`find_oasis2_img_files()`). MMIST-ccRCC uses a CSV lookup to map `case_id` to specific exam filenames, which requires a flexible modality lookup system in `scan_modality_files()`.
- *Backend-model integration:* Ensuring that the scaler, feature selection, and training medians saved at training time were correctly loaded and applied at inference time required careful checkpoint design. The final checkpoint stores all pipeline artifacts (model weights, optimizer state, selected features, scaler parameters, training medians, ANFIS configuration, and threshold) in a single `.pt` file.
- *Platt calibrator issue:* The Platt calibrator was trained and saved via pickle in the training script's context, causing an `AttributeError` at server startup. Beyond the serialization problem, calibration analysis revealed a deeper issue: the model's raw sigmoid scores were tightly clustered in the range [0.49, 0.51], providing insufficient

spread for logistic regression (Platt scaling) to fit a meaningful slope. It collapsed to predicting the global class prior (0.216) for all inputs. This is a known failure mode of Platt scaling when the classifier is overconfident or underconfident in a narrow output band. Hence, the calibrator was removed entirely for the current deployment.

6 Evaluation

6.1 Model Evaluation

The DF DL framework was evaluated using an unbiased nested cross-validation strategy (5-outer / 3-inner folds). The outer folds provide performance estimates on held-out test data that was never seen during model training or hyperparameter selection. The inner folds perform a hyperparameter search over a grid of ANFIS configurations (varying number of fuzzy sets and number of selected clinical features). The best configuration across outer folds is selected by majority vote.

Primary metrics are Balanced Accuracy (BAcc) and AUC, preferred over standard accuracy due to class imbalance in all four datasets. The classification threshold is selected via Youden's J statistic on the validation set within each outer fold.

For more information on datasets, see Appendix A.

OASIS-1: Cross-Sectional Alzheimer's Dataset [8]

Metric	Score
Balanced Accuracy (BAcc)	0.9052 ± 0.0116
AUC	0.9593 ± 0.0088
Best config	n_fuzzy_sets=2, n_selected_feats=4

Feature frequency: MMSE (5/5), nWBV (5/5), Age (5/5), Educ (1/5), eTIV (1/5). The perfect selection frequency of MMSE, nWBV, and Age across all folds confirms their well-established clinical significance in Alzheimer's diagnosis. nWBV (normalized whole brain volume) serves as a direct structural biomarker for neurodegeneration. The low variance in both BAcc (± 0.0116) and AUC (± 0.0088) indicates highly robust generalization across folds.

OASIS-2: Longitudinal Alzheimer's Dataset [9]

Metric	Score
Balanced Accuracy (BAcc)	0.8465 $\pm$ 0.0260
AUC	0.8836 $\pm$ 0.0288
Best config	n_fuzzy_sets=2, n_selected_feats=3

Feature frequency across folds: MMSE (5/5), Age (5/5), EDUC (5/5), nWBV (2/5), SES (1/5). The consistent selection of MMSE, Age, and Education across all folds confirms their well-established clinical significance. The slightly higher variance compared to OASIS-1 is consistent with the longitudinal nature of the dataset, introducing inter-visit variability.

MMIST-ccRCC: Renal Cell Carcinoma Multi-Modal Dataset [10]

Metric	Score
Balanced Accuracy (BAcc)	0.6506 $\pm$ 0.0439
AUC	0.6188 $\pm$ 0.0563
Best config	n_fuzzy_sets=2, n_selected_feats=5

Feature frequency across folds: VHL_mutation (5/5), ajcc_clin_metastasis_cm (5/5), ajcc_path_metastasis_pm (4/5), age_diag (4/5), TTN_mutation (3/5). The lower performance (BAcc $\approx$ 0.65) relative to OASIS is expected: MMIST is a survival prediction task inherently harder than binary dementia classification; significant class imbalance is present; and selected features carry substantial noise in 12-month prediction. VHL mutation appearing in 5/5 folds is consistent with its status as the most common somatic driver mutation in ccRCC. The system selecting 5 features (vs. 3–4 for OASIS) indicates ANFIS requires more clinical dimensions to achieve discrimination.

UTSW-Glioma: Glioma Molecular Classification Dataset [11]

Metric	Score
Balanced Accuracy (BAcc)	0.8353 $\pm$ 0.0327
AUC	0.8347 $\pm$ 0.0545
Best config	n_fuzzy_sets=3, n_selected_feats=3

Feature frequency across folds: IDH (5/5), MGMT (3/5), Age_at_Imaging (3/5), Sex_at_birth (2/5), Race (2/5), 1p19Q_CODEL (2/5), IDH_Testing_Method (2/5).

IDH mutation status appearing in all 5/5 folds is clinically well-grounded since IDH mutation is the most prognostically significant molecular marker in glioma under the 2021 WHO CNS tumor classification [24] [25]. MGMT promoter methylation and 1p/19q co-deletion (1p19Q_CODEL) are the other two tier-1 glioma biomarkers, and both appear consistently across folds, further validating the clinical alignment of the L1 feature selector. The presence of IDH_Testing_Method (2/5 folds) is likely a proxy feature: IHC-detected IDH mutations correlate with specific molecular subtypes. The higher variance in AUC (± 0.0545) reflects the smaller cohort size and the categorical nature of molecular marker data.

Cross-Dataset Performance Summary

Dataset	Task	BAcc	AUC	Best Config (#sets, #features)	Top ANFIS Features
OASIS-1 [8]	Dementia (cross-sectional)	0.9052 $\pm$ 0.0116	0.9593 $\pm$ 0.0088	2, 4	MMSE, nWBV, Age
OASIS-2 [9]	Dementia (longitudinal)	0.8465 $\pm$ 0.0260	0.8836 $\pm$ 0.0288	2, 3	MMSE, Age, EDUC

Dataset	Task	BAcc	AUC	Best Config (#sets, #features)	Top ANFIS Features
MMIST-ccR CC [10]	12-month survival	0.6506 $\pm$ 0.0439	0.6188 $\pm$ 0.0563	2, 5	VHL_mutation, ajcc_clin_metast asis_cm
UTSW-Glio ma [11]	Glioma mol. subtype	0.8353 $\pm$ 0.0327	0.8347 $\pm$ 0.0545	3, 3	IDH, MGMT, Age_at_Imaging

The cross-dataset results demonstrate that the same [AgnosticHybridFusion](#) architecture generalizes across four distinct oncological and neurological domains without any code or architectural changes, which is a direct validation of the "domain-agnostic" design claim. The performance gradient (OASIS-1 > OASIS-2 > Glioma > MMIST) correlates with task difficulty. In all four cases, the ANFIS-selected features align with established clinical knowledge.

6.2 Software Evaluation

The software system was evaluated through two complementary approaches.

Unit and Integration Testing. Each major backend component (the prediction service, the model adapter, the ChromaDB retrieval pipeline, and the LLM client) was unit-tested in isolation. Integration tests verified end-to-end correctness of the [/predict](#) and [/api/copilot/answer](#) workflows, including correct handling of missing imaging modalities (mask=0 fallback to clinical-only inference) and the abstention mechanism under low-overlap retrieval conditions.

Latency Benchmarking. Inference latency was measured for two scenarios: (1) clinical-only inference (no imaging, using pre-extracted NPZ features from cache), and (2) full multimodal inference with on-the-fly feature extraction. Clinical-only inference runs on CPU and is suitable for real-time use. Full feature extraction requires a GPU for practical throughput.

Planned User Evaluation. A structured user study with clinicians is planned but has not yet been conducted due to access constraints. The evaluation protocol will involve presenting clinicians with DFDL outputs alongside standard clinical records for a set of de-identified cases, and collecting responses via a structured questionnaire covering (1) prediction plausibility, (2) rule interpretability (can the clinician read and evaluate the fired ANFIS rules?), (3) copilot explanation quality, and (4) overall trust in the system.

7 Conclusion & Future Work

7.1 Key Findings and Contributions

This project designed, implemented, and evaluated the Domain-Agnostic Fuzzy Deep Learning (DFDL) framework, a complete clinical decision support system that achieves strong predictive performance while maintaining genuine intrinsic interpretability through TSK-ANFIS fuzzy rules. The key contributions:

- **A novel hybrid fusion architecture** separating interpretable clinical modelling (ANFIS, 3–5 features) from high-dimensional image feature modelling (flat MLP, 512-dim NPZ), resolving the rule explosion problem that prevented previous end-to-end fuzzy approaches from scaling to volumetric imaging data.
- **Domain-agnostic configurability** demonstrated across four medical datasets (OASIS-1, OASIS-2, MMIST-ccRCC, UTSW-Glioma) without any architectural changes.
- **Strong quantitative performance:** $\text{BAcc} = 0.9052 \pm 0.0116$, $\text{AUC} = 0.9593 \pm 0.0088$ on OASIS-1; $\text{BAcc} = 0.8353 \pm 0.0327$, $\text{AUC} = 0.8347 \pm 0.0545$ on UTSW-Glioma, both under unbiased nested cross-validation.
- **A complete software system:** FastAPI backend, RAG-based explainability copilot with an abstention mechanism, and a clinician-facing web interface.
- **Clinically meaningful feature selection:** MMSE, nWBV, and Age were selected in 5/5 folds on OASIS-1, IDH mutation was selected in 5/5 folds on UTSW-Glioma, confirming alignment between the model's learned reasoning and established clinical knowledge.

7.2 Future Work

Several directions for improvement and extension are identified:

- **Clinical user study.** The most critical next step is to evaluate prediction plausibility, rule interpretability, and copilot explanation quality with real clinicians. This will provide the empirical evidence needed to assess whether the system genuinely assists clinical decision-making.
- **Ablation study.** A systematic ablation comparing: (a) ANFIS-only (clinical features, no imaging), (b) FlatMLP-only (imaging, no clinical), (c) full DFDL hybrid fusion, to quantify each stream's contribution and validate the fusion gate.

- **Additional modalities.** The configuration system supports PET, fMRI, and whole-slide imaging (WSI) as additional modalities. Integrating PET-based amyloid imaging for Alzheimer's staging and WSI for MMIST-ccRCC pathology scoring are natural extensions.
- **Proper probability calibration.** Candidate approaches include isotonic regression (non-parametric, better for non-monotone distortions), temperature scaling (a single learned scalar applied to the logit before sigmoid, robust to small calibration sets), or histogram binning: all of which require a held-out calibration split that is representative of the deployment distribution.
- **LLM domain adaptation.** The current copilot uses a general-purpose Llama 3.1 model. Fine-tuning or parameter-efficient adaptation (LoRA) on domain-specific clinical NLP corpora, as demonstrated by Van Veen et al. [19] for radiology, is expected to improve explanation quality and reduce abstention rates.
- **Prospective clinical validation.** Evaluate on real de-identified patient data from a clinical partner to demonstrate real-world generalizability.
- **CTGAN synthetic augmentation at scale.** Re-evaluate CTGAN for small-cohort augmentation as dataset size grows, per Miura et al. [18].

8 References

- [1] A. Holzinger, C. Biemann, C. S. Pattichis, and D. B. Kell, "What do we need to build explainable AI systems for the medical domain?" arXiv preprint arXiv:1712.09923, Dec. 2017.
- [2] Y. Zheng, Z. Xu, and X. Wang, "The Fusion of Deep Learning and Fuzzy Systems: A State-of-the-Art Survey," *IEEE Transactions on Fuzzy Systems*, vol. 30, no. 8, pp. 2783–2799, Aug. 2022, doi: 10.1109/TFUZZ.2021.3062899.
- [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, May 2024, doi: 10.1016/j.ins.2024.120802.
- [4] R. Rosenbacke, Å. Melhus, M. McKee, and D. Stuckler, "How Explainable Artificial Intelligence Can Increase or Decrease Clinicians' Trust in AI Applications in Health Care: Systematic Review," *JMIR AI*, vol. 3, pp. e53207–e53207, Oct. 2024, doi: <https://doi.org/10.2196/53207>.
- [5] T. Răz, P. De, and M. Reyes, "Explainable AI in medicine: challenges of integrating XAI into the future clinical routine." *PubMed*, vol. 5, pp. 1627169–1627169, Jan. 2025, doi: <https://doi.org/10.3389/fradi.2025.1627169>.
- [6] X. Nie, "Multi-Modal image fusion for medical diagnosis: combining MRI and CT using deep generative models," *Clinical Medicine and Health Research Journal*, vol. 5, no. 03, pp. 1313–1327, Jun. 2025, doi: 10.18535/cmhrj.v5i03.486.
- [7] R. Zhang *et al.*, "Multimodal artificial intelligence in medicine: a task-oriented framework for clinical translation," *Frontiers in Medicine*, vol. 12, Jan. 2026, doi: <https://doi.org/10.3389/fmed.2025.1736272>.
- [8] Open Access Series of Imaging Studies (OASIS): Cross-Sectional MRI Data in Young, Middle Aged, Nondemented, and Demented Older Adults. Marcus, DS, Wang, TH, Parker, J, Csernansky, JG, Morris, JC, Buckner, RL. *Journal of Cognitive Neuroscience*, 19, 1498-1507. doi: 10.1162/jocn.2007.19.9.1498
- [9] Open Access Series of Imaging Studies (OASIS): Longitudinal MRI Data in Nondemented and Demented Older Adults. Marcus, DS, Fotenos, AF, Csernansky, JG,

Morris, JC, Buckner, RL, 2010. Journal of Cognitive Neuroscience, 22, 2677-2684. doi: 10.1162/jocn.2009.21407

[10] T. Mota, M. R. Verdelho, D. J. Araújo, A. Bissoto, C. Santiago, and C. Barata, "MMIST-ccRCC: A Real World Medical Dataset for the Development of Multi-Modal Systems," Institute for Systems and Robotics, LARSyS; Recod.ai Lab, University of Campinas; Lisbon ELLIS Unit.

[11] Reddy, D., Saadat, N., Holcomb, J., Wagner, B., Truong, N., Bowerman, J., Hatanpaa, K., Patel, T., Pinho, M., Yu, F., Zhang, K., Lodhi, S., Madhuranthakam, A., Bangalore Yogananda, C. G., & Maldjian, J. (2026). The University of Texas Southwestern Glioma MRI dataset with molecular marker characterization and segmentations (UTSW-Glioma) (Version 1) [Data set]. The Cancer Imaging Archive. <https://doi.org/10.7937/DFAE-1B86>

[12] M. F. Mahdi, A. Jahani, and D. H. Abd, "Fuzzy evaluation and explainable machine learning for diagnosis of rheumatic and autoimmune diseases," *PeerJ Computer Science*, vol. 11, p. e3096, Aug. 2025, doi: <https://doi.org/10.7717/peerj-cs.3096>.

[13] Q. Zhang, "A deep neuro-fuzzy framework for speech emotion recognition," *Computer Methods in Biomechanics and Biomedical Engineering*, Sep. 2025, doi: 10.1080/10255842.2025.2559060.

[14] D. F. Hasan and A. M. Khidhir, "Toward enhancement of deep learning techniques using fuzzy logic: A survey," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 13, no. 3, pp. 3041–3055, Jun. 2023, doi: 10.11591/ijece.v13i3.pp3041-3055.

[15] S. Kumar, S. Rani, S. Sharma, and H. Min, "Multimodality Fusion Aspects of Medical Diagnosis: A Comprehensive Review," *Bioengineering*, 11(12), 1233, 2024, doi: 10.3390/bioengineering11121233.

[16] A. Tariq, I. Banerjee, H. Trivedi, and J. Gichoya, "Multimodal artificial intelligence models for radiology," *BJR|Artificial Intelligence*, vol. 2, no. 1, Jan. 2025, doi: <https://doi.org/10.1093/bjrai/ubae017>.

[17] J. Li, B. J. Cairns, J. Li, and T. Zhu, "Generating synthetic mixed-type longitudinal electronic health records for artificial intelligent applications," *NPJ Digital Medicine*, 6(1), 98, 2023, doi: 10.1038/s41746-023-00834-7.

- [18] T. Miura, E. Kimura, A. Ichikawa, M. Kii, and J. Yamamoto, "Evaluating Synthetic Data Generation Techniques for Medical Dataset," in *BIOSTEC* (2), pp. 315–322, 2024, doi: 10.5220/0012314500003657.
- [19] D. Van Veen, C. Van Uden, M. Attias, A. Pareek, C. Bluethgen, M. Polacin, et al., "Radadapt: Radiology report summarization via lightweight domain adaptation of large language models," in *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pp. 449–460, 2023, doi: 10.18653/v1/2023.bionlp-1.42.
- [20] O. K. Gargari and G. Habibi, "Enhancing medical AI with retrieval-augmented generation: A mini narrative review," *DIGITAL HEALTH*, vol. 11, Apr. 2025, doi: <https://doi.org/10.1177/20552076251337177>.
- [21] L. M. Amugongo, P. Mascheroni, S. Brooks, S. Doering, and J. Seidel, "Retrieval augmented generation for large language models in healthcare: A systematic review," *PLOS Digital Health*, vol. 4, no. 6, p. e0000877, Jun. 2025, doi: <https://doi.org/10.1371/journal.pdig.0000877>.
- [22] S. Chen, K. Ma, and Y. Zheng, "Med3D: Transfer Learning for 3D Medical Image Analysis," *arXiv:1904.00625 [cs]*, Jul. 2019, Available: <https://arxiv.org/abs/1904.00625>
- [23] A. Janosi, W. Steinbrunn, M. Pfisterer, and R. Detrano. "Heart Disease," UCI Machine Learning Repository, 1989. [Online]. Available: <https://doi.org/10.24432/C52P4X>.
- [24] P. Śledzińska, M. G. Bebyn, J. Furtak, J. Kowalewski, and M. A. Lewandowska, "Prognostic and Predictive Biomarkers in Gliomas," *International Journal of Molecular Sciences*, vol. 22, no. 19, p. 10373, Sep. 2021, doi: <https://doi.org/10.3390/ijms221910373>.
- [25] N. C. D. Truong *et al.*, "Two-Stage Training Framework Using Multicontrast MRI Radiomics for IDH Mutation Status Prediction in Glioma," *Radiology: Artificial Intelligence*, vol. 6, no. 4, Jul. 2024, doi: <https://doi.org/10.1148/ryai.230218>.

Appendix A. Dataset Descriptions

OASIS-1 (Open Access Series of Imaging Studies: Cross-Sectional) [8]

- **Subjects:** 416 subjects aged 18–96; 98 diagnosed with mild-to-moderate Alzheimer's dementia ($CDR > 0$)
- **Task:** Binary dementia classification ($CDR > 0$ = positive class)
- **Modalities:** 3D T1-weighted MPRAGE MRI (Analyze `.hdr/.img`); clinical tabular data (CSV)
- **ANFIS features:** Age, EDUC, SES, MMSE, eTIV, nWBV, ASF

OASIS-2 (Open Access Series of Imaging Studies: Longitudinal) [9]

- **Subjects:** 373 longitudinal MRI sessions from 150 subjects followed over 1.5–5 years; 144 sessions with $CDR > 0$
- **Task:** Binary dementia classification ($CDR > 0$ = positive class)
- **Modalities:** 3D T1-weighted MPRAGE MRI (Analyze `.hdr/.img`; session IDs e.g. OAS2_0001_MR1); clinical tabular data (CSV)
- **ANFIS features:** Age, EDUC, SES, MMSE, eTIV, nWBV, ASF

MMIST-ccRCC (Multimodal Imaging and Survival for clear cell Renal Cell Carcinoma) [10]

- **Subjects:** Multi-institutional cohort; pre-defined 60/20/20 train/validation/test split with significant class imbalance (alive vs. deceased at 12 months)
- **Task:** 12-month survival prediction (binary: alive/deceased)
- **Modalities:** CT, MRI, histopathology whole-slide images (WSI), genomics, and clinical tabular data; DFDL uses CT/MRI for the imaging stream and clinical CSV for the ANFIS stream
- **ANFIS features:** Age at diagnosis, gender, tumor grade, AJCC clinical/pathological staging (cM, pM), mutation status (VHL, PBRM1, TTN), race

UTSW-Glioma (University of Texas Southwestern Glioma MRI Dataset) [11]

- **Subjects:** 1,740 preoperative MRI sessions from 709 subjects (2012–2020); 348 male (54.2%), 251 female (39.1%); glioblastoma is the predominant histological type (377 subjects)

- **Task:** Binary glioma molecular subtype classification (IDH-mutant vs. IDH-wildtype)
- **Modalities:** Multi-parametric MRI: T1w, T2w, FLAIR, T1w contrast-enhanced (NIfTI [.nii.gz](#)); molecular/clinical tabular data
- **ANFIS features:** IDH mutation status, MGMT promoter methylation, 1p19q co-deletion, Age at imaging, Sex at birth, Race, IDH Testing Method, Tumor Type