A novel explainable AI framework for medical image classification integrating statistical, visual, and rule-based methods

Team Members
Muhammad Waqas | 25I - 7649
Muhammad Hamza | 25I - 7639

Base Paper Methodology Summary

The paper proposes **SVIS-RULEX**, an **explainable AI framework** designed to make deep learning–based medical image classifiers transparent.
It combines **statistical**, **visual**, and **rule-based** explanation methods hence the name.

Step-by-Step Workflow

1. Pre-processing

- **Operations:** ROI cropping → data augmentation → resizing (224×224).
- **Goal:** Improve input quality and ensure consistent size for the CNN.
- **Augmentations:** rotation, flipping, brightness/contrast change, shifting, RGB shifts, fog effect.

2. Deep Feature Extraction (MobileNetV2)

- A **custom MobileNetV2** model is fine-tuned to extract high-level image features.
- Layers: convolutional backbone + custom dense layers (ReLU, BatchNorm, Dropout, Softmax).
- Features are taken from the **second dense layer** → compact deep feature vector.
- Output: $\mathbb{R}^{}$ for each image.

3. Statistical Feature Engineering

- Converts deep features into **26 interpretable statistical metrics**:
    - Mean, variance, skewness, entropy, kurtosis, range, SNR, etc.
- These **statistical summaries** provide interpretable, quantifiable representations of deep features.
- Output: $\mathbb{R}^{\times 26}$

4. Feature Selection (ZFMIS)

A **two-stage feature selection** process called **Zero-based Filtering with Mutual Information Selection (ZFMIS)**:

1. **Zero-based filtering:** removes sparse or uninformative features (≥50% zeros).
2. **Mutual Information (MI):** ranks remaining features by correlation with target labels.
3. **Hierarchical subset selection:** top-k (3, 6, 9, …18) features tested to optimize performance.

5. Rule-Based Modeling (Decision Tree & RuleFit)

- Selected features are used to train:
  - **Decision Tree (DT):** produces hierarchical if–then rules.
  - **RuleFit:** ensemble of rules + linear regression for enhanced accuracy and interpretability.
- Rules are **human-readable** and explain how features lead to specific classifications.

6. Visual Explanation (SFMOV)

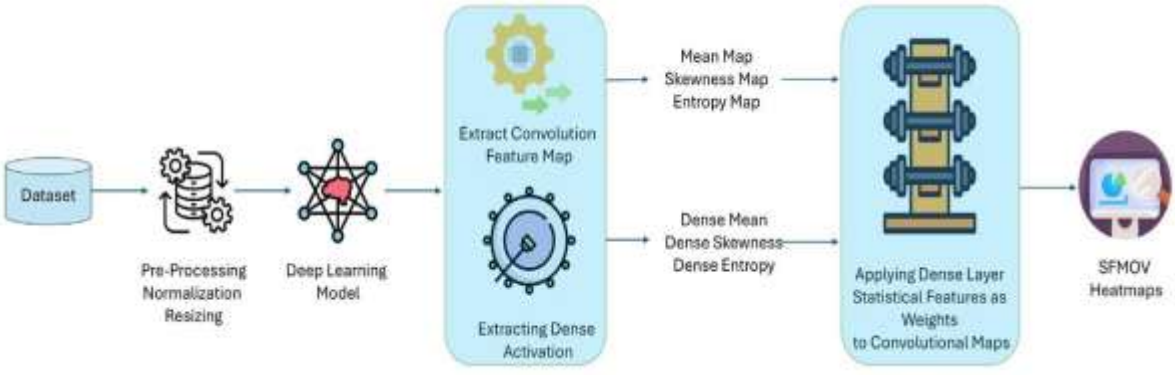
- Proposes **Statistical Feature Map Overlay Visualization (SFMOV)**:
  - Inspired by Grad-CAM, but based on **statistical measures** instead of gradients.
  - Creates **heatmaps** from mean, skewness, and entropy maps.
  - Combines them using weighted sum:

$$\square\square(\square,\square) = \square_1\square\square_{\square}(\square,\square) + \square_2\square\square_{\square}(\square,\square) + \square_3\square_{\square}(\square,\square)$$

  - Overlays heatmap on the input image → localized interpretability.

7. Evaluation

- Applied to **five medical datasets** (COVID-19 X-ray, Breast ultrasound, Brain MRI, Histopathology, Fundus).
- Metrics: accuracy, precision, recall, F1-score, and qualitative interpretability confirmed by radiologists.



Gap of the base paper

The existing framework provides interpretable rule-based and statistical explanations but lacks attention-driven, context-aware, and quantitatively validated interpretability.

Integrating a Vision Transformer (ViT) can bridge these gaps by providing global attention-based explanations directly derived from the model's internal reasoning.

Proposed Pipeline: Vision Transformer Integration for Attention-Based Interpretability (A-SVIS-RULEX)

1. Pre-processing

- **Input:** Raw medical images (e.g., X-ray, MRI, Ultrasound, Fundus, Histopathology).
- **Operations:**
    - ROI extraction (to isolate diagnostic region).
    - Data augmentation (rotation, flipping, contrast adjustment, etc.).
    - Resizing images to ViT input resolution (e.g., 224 × 224).
- **Goal:** Normalize and diversify data for better generalization.

2. Feature Extraction using Vision Transformer (ViT)

- Replace the **MobileNetV2** backbone from the base paper with a **Vision Transformer (ViT-B/16 or Swin Transformer)**.
- **Process:**
    - The image is divided into fixed-size patches (e.g., 16×16).
    - Each patch is linearly embedded and passed through **self-attention layers**.
    - ViT outputs contextualized **patch embeddings** and a **[CLS] token** representing the global image representation.
- **Fine-tuning:** Pre-trained ViT is fine-tuned on medical datasets used in the base study.
- **Output:** Global attention-aware embeddings that capture spatial and contextual dependencies across the image.

3. Attention-Based Interpretability Layer

- Extract **attention weights** from the final transformer layers.
- Apply **attention rollout** or **mean attention head aggregation** to produce visual **attention maps**.
- **Overlay** these maps on original medical images to highlight diagnostically relevant regions.
- **Output:** Model-intrinsic, global attention heatmaps representing *where* and *how strongly* the model focused during prediction.
  (This replaces or complements the SFMOV visualization of the base paper.)

4. Feature Transformation and Rule-Based Modeling

- Use ViT's global embedding (e.g., [CLS] token or pooled representation) as the **input vector** for rule-based interpretability.
- Perform:
  - **Statistical summarization (optional):** Compute mean, variance, entropy, etc. on ViT embeddings to retain statistical interpretability from SVIS-RULEX.
  - **Feature selection:** Employ the same **ZFMIS** (Zero-based Filtering + Mutual Information Selection) technique to retain informative features.
- Train:
  - **Decision Tree (DT)** — for explicit *if–then–else* rules.
  - **RuleFit model** — for sparse rule ensembles capturing linear + non-linear relations.
- **Output:** Human-readable rules explaining decision logic (e.g., "If entropy > 0.45 and attention_weight > 0.70 ⇒ Malignant").

5. Combined Explainability Output

- **Attention Map:** Shows *where* the model looked (global interpretability).
- **Rule-Based Explanation:** Describes *why* the model decided (logical interpretability).
- The two outputs together form a **hybrid explainability framework** offering both visual and logical transparency.

6. Evaluation

- **Quantitative metrics:** Accuracy, Precision, Recall, F1-score, AUC.
- **Explainability metrics:**
  - Attention–ground-truth overlap (IOU).
  - Rule fidelity (agreement between rule model and ViT).
- **Qualitative:** Visual comparison of ViT attention vs. SFMOV and expert radiologist validation.