

Neuro-symbolic Rule Induction over Multimodal LLM-Generated Features for Explainable Image and Video Classification

Matěj Koval¹, Lukáš Sýkora¹, Michal Ahmet Ay¹, Vojtěch Balek¹, Yuliia Yershova¹ and Tomáš Kliegr¹

¹Department of Information and Knowledge Engineering, Faculty of Informatics and Statistics, Prague University of Economics and Business, W. Churchill Sq. 4, 130 67 Prague 3, Czech Republic

Abstract

While black-box classifiers for video and image analysis deliver high predictive accuracy, their opaque decision logic is at odds with strict transparency requirements. We propose a novel neuro-symbolic methodology that bridges unstructured visual media and classical symbolic rule induction. Rather than relying on post-hoc visual approximations or continuous embeddings, our approach utilizes a local multimodal large language model (MLLM), such as Qwen2.5-VL, strictly as a perceptual translator to map raw pixels into a discrete semantic feature space. This intermediate representation enables the direct application of rule learners, such as RuleKit, to induce human-readable IF . . . THEN rules governing the final classification. Our primary methodological contribution is the systematic design and empirical characterisation of this pipeline — coupling stochastic MLLM perception with deterministic rule induction — to yield predictive models where the downstream decision logic is inherently interpretable. We instantiate this framework in *Media Feature Lab*, an open-source system supporting both standalone rule classification and hybrid soft-voting ensembles. Evaluated on four-class video memorability prediction (MediaEval), binary pneumonia detection, and binary mammography classification on the public VinDr-Mammo benchmark, the rule-based methodology achieves predictive performance competitive with simple baselines while rendering the classification path more transparent than standard black-box approaches. We provide a critical analysis of the variance introduced by the LLM-based translation step and its impact on rule stability. Code and a live demo are publicly available.

Keywords

large language models, multimodal LLM, rule induction, neuro-symbolic AI, RuleKit, explainable AI, feature generation, image and video classification

1. Introduction

A classifier that predicts whether a short clip will be memorable, go viral, or be flagged as a strong advertisement is of limited use to a media analyst if it cannot explain *why*. The dominant approaches to media classification — hand-crafted descriptors or deep embeddings from convolutional neural networks (CNNs) and Vision Transformers (ViTs) — are accurate but opaque: their predictions can be explained post hoc with tools such as SHAP (SHapley Additive exPlanations) [1], Local Interpretable Model-agnostic Explanations (LIME) [2], or Gradient-weighted Class Activation Mapping (Grad-CAM) [3]. These methods are applicable to opaque classifiers, but explain their predictions through feature attributions, local surrogates, or saliency maps. Our objective is instead to make the downstream decision logic itself directly inspectable as human-readable rules. Transparency and explainability are also increasingly reflected in regulation; the EU Artificial Intelligence Act introduces transparency requirements for specified AI systems [4].

Multimodal large language models (MLLMs) make a different design possible. Models such as Qwen2.5-VL [5] can analyse video content and temporal dynamics and generate natural-language responses. Instead of training a specialised network end-to-end, one can use the MLLM as a *translator*

RuleML+RR'26: Companion Proceedings of the 10th International Joint Conference on Rules and Reasoning, August 24–26, 2026, Vilnius, Lithuania

✉ kovm23@vse.cz (M. Koval); lukas.sykora@vse.cz (L. Sýkora); ayimi00@vse.cz (M. A. Ay); balv05@vse.cz (V. Balek); poty01@vse.cz (Y. Yershova); tomas.kliegr@vse.cz (T. Kliegr)

🆔 0000-0002-6700-5610 (L. Sýkora); 0009-0008-6236-0489 (Y. Yershova); 0000-0002-7261-0380 (T. Kliegr)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

from raw pixels into a small set of meaningful features, and then learn another — preferably intrinsically interpretable — machine learning model on top of them.

This paper presents *Media Feature Lab*, a tool that realises this idea end-to-end by coupling automatic MLLM feature engineering with interpretable rule induction. Conceptually, the resulting pipeline is neuro-symbolic: a neural MLLM constructs the interpretable feature representation, while symbolic rule induction operates over that representation.

Positioning in Neuro-symbolic AI. This work addresses an open question frequently posed in modern explainable AI: *How can neuro-symbolic AI systems combine the perceptual power of multimodal LLMs with the transparent, logical reasoning of classical rule induction?* The idea of using a neural model (LLM) to produce a small set of interpretable features for downstream classification was demonstrated in our previous work for text by Balek et al. [6], who analysed the predictive contribution of the generated features with SHAP. A related approach for biomedical X-ray classification based on LLM-generated textual representations was studied by Ay [7]. Our system extends this general direction to *video* and, importantly, establishes a loosely coupled neuro-symbolic pipeline by using RuleKit [8] as an explicit mechanism for inducing human-readable rules over the LLM-extracted attributes. Based on our survey, this specific neuro-symbolic combination — MLLM feature extraction feeding rule induction — had not been applied to video classification before [9].

Contributions and research questions. We make three contributions, framed as the questions the tool is built to answer:

- **(RQ1, neuro-symbolic automation)** A web tool that *designs and extracts* interpretable features from media using a neural perception layer (a locally run MLLM), seamlessly feeding a symbolic rule learner with no manual feature set and no code (Section 3).
- **(RQ2, accuracy)** An evaluation of the resulting predictive models against baselines and a classical-machine-learning ensemble on three domains, including negative and inconclusive results (Section 4).
- **(RQ3, reproducibility)** An analysis locating important sources of variability in stochastic feature discovery and extraction, with a built-in repeatability test (Sections 3 and 4).

Paper organization. Section 2 introduces MLLM-as-feature-extractor and the RuleKit rule learner. Section 3 describes the tool: architecture, the five-phase pipeline, the discovery/extraction prompts, the models, and the explanation mechanism. Section 4 presents the three case studies. Section 5 discusses related work and Section 6 concludes with open-science resources.

2. Background

2.1. MLLMs as feature extractors

Since Contrastive Language–Image Pre-training (CLIP) aligned image and text in a shared space [10], multimodal models have advanced quickly, from Flamingo [11] and Video-LLaMA [12] to later video-capable models such as VTimeLLM [13], LongVA [14], and Qwen2.5-VL [5]. Video-MLLM pipelines commonly represent video with sampled frames, sometimes combined with subtitles, audio, or audio-derived text [15, 16]. These models remain imperfect — hallucination and poor temporal perception are well documented [17, 18]. However, in our work we hypothesise that they are sufficiently capable to serve as a *description layer*: a way to convert raw media into structured text from which a small number of measurable attributes can be read.

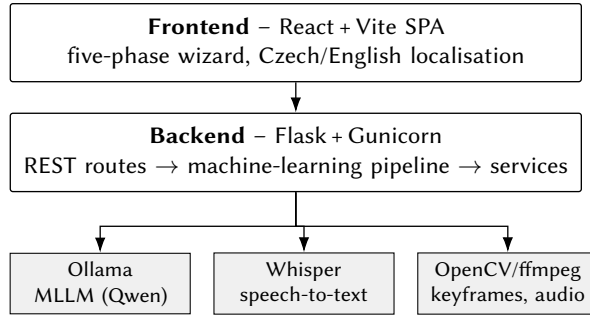


Figure 1: Three-layer architecture. The backend orchestrates a service layer that runs the multimodal model, speech-to-text, and deterministic video processing locally.

2.2. Rule induction with RuleKit

The rule-learning tradition — AQ [19], CN2 [20], RIPPER [21] — produces models that state their decisions as explicit conditions. We use RuleKit [8], an open-source suite for rule-based learning (classification, regression, survival) that induces compact IF . . . THEN rules with associated quality measures. RuleKit is the source of interpretability in our system: unlike post-hoc explainers such as SHAP [1] and Local Interpretable Model-agnostic Explanations (LIME) [2], which provide feature-attribution or local-surrogate explanations for opaque models, the induced rules constitute the predictive model in the RuleKit-only configuration and can be read without further tooling. We chose RuleKit for its versatility, as it can handle both regression and classification tasks. However, other rule learning systems could also be used.

3. The Media Feature Lab System

Media Feature Lab is a web application that takes a small labelled sample of media and, without any coding, walks the user from raw clips to interpretable rules and predictions. It is built around two main ideas: the feature set is *discovered automatically* by an MLLM, and rule induction provides *human-readable decision logic* over the extracted features.

3.1. Architecture

The system is a three-layer application (Figure 1): a React/Vite single-page application (SPA) frontend; a Flask/Gunicorn backend exposing the pipeline as representational state transfer (REST) endpoints; and a service layer wrapping the media tooling — a locally run Ollama server hosting the multimodal model, Whisper for speech-to-text, and OpenCV/ffmpeg for keyframe and audio extraction. For video inputs, the current implementation also appends deterministic OpenCV descriptors such as motion magnitude, optical flow, saturation, brightness, scene-cut count, and face-presence ratio; these descriptors are not proposed by the MLLM but are treated by the downstream models as ordinary numeric features. The first two case studies reported in Section 4 used the MLLM-extracted feature sets documented in [9]; the effect of combining them with the deterministic descriptors is left for future ablation. Running the MLLM locally keeps data on-premises and avoids per-call API cost, at the cost of latency and the 7B-scale quality ceiling discussed in Section 4. Each browser session is isolated by an X-Session-ID header and its state is persisted to disk, so a refresh or a server transfer does not lose work.

3.2. The five-phase pipeline

The user advances through five phases (Figure 2, Table 1). The processing phases run asynchronously: the corresponding endpoints return a `job_id` and the frontend polls for progress. The tool uses ZIP archives for media batches, comma-separated values (CSV) files for labels, and JavaScript Object Notation (JSON) for feature specifications.

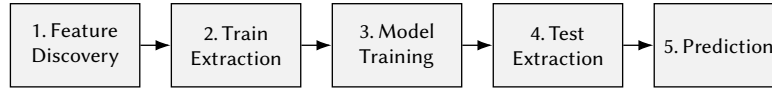


Figure 2: The five-phase pipeline. The MLLM is used in phases 1, 2 and 4; RuleKit induces the rules in phase 3.

Table 1

The five phases of the pipeline.

Phase	User provides	System produces
1. Discovery	sample clips + target name	feature specification (JSON)
2. Train extract.	training ZIP + labels CSV	training feature matrix
3. Training	target column	rules, metrics, importances
4. Test extract.	test ZIP	test feature matrix
5. Prediction	(optional ground truth)	predictions + metrics

3.3. Feature discovery and extraction

We describe the more complex video-processing pipeline below. The image pipeline is analogous, with video-specific steps such as keyframe extraction omitted.

Discovery is a two-step LLM procedure. First, up to `DISCOVERY_MAX_SAMPLES` (default 10) supplied sample clips are each *observed*: the model is asked for an objective bullet-point description (temperature 0.1). For video, keyframes are selected by scene-change detection (Hue–Saturation–Value, HSV, histograms compared by Bhattacharyya distance, threshold 0.35, falling back to uniform sampling) and an audio transcript is attached. Second, the observations are *synthesised* into a specification of 5–8 measurable MLLM-extracted features (temperature 0.3), each either a numeric range `[min, max]` or a categorical domain `["a", "b", ...]`. Listing 1 shows the core of the synthesis prompt; the output is a JSON object. When training labels are supplied, the model receives aggregate target context: the class distribution and example class values for classification, or summary statistics and example target values for regression. Test labels and explicit media-to-label mappings are not provided during discovery.

Listing 1: Synthesis prompt for Feature Discovery (excerpt).

```

You are a machine learning feature engineer.
Your goal is to predict: '{target_variable}'.
Below are observations from N media sample(s): ...
Define EXACTLY 5 to 8 measurable features that:
- Can be extracted from ANY media clip of this type
- Are likely to correlate with '{target_variable}'
- Cover DIVERSE perceptual dimensions, are independent
Output STRICTLY a JSON object. Each value MUST be a numeric
range [min, max] or a categorical domain ["a", "b", ...].

```

Extraction (phases 2 and 4) uses a describe-then-extract structured prompt: the model first describes the clip in 2–3 sentences and then produces a JSON object with exactly the specified keys. The intermediate description is intended to provide explicit perceptual context for the subsequent structured extraction. To dampen the stochasticity of a single call, extraction can be run k times per clip (`EXTRACTION_PASSES`) and aggregated — median for numeric features, majority vote for categorical ones. Values are validated and clamped to the declared ranges; within an extraction batch, missing numeric values are filled with the column median and missing categorical values with the modal category.

3.4. Models and the explanation

For *regression* the system trains a RuleKit `RuleRegressor`. The predictions are produced by the induced rule model. For *classification*, the current implementation trains a RuleKit `RuleClassifier`

together with Random Forest (RF) and Gradient Boosting Trees (GBT). When all three models are available, the ensemble averages the class-probability distributions of RuleKit, RF (200 trees, `class_weight="balanced"`), and GBT (100 trees). For a sample x and class c , the implemented prediction is

$$\hat{p}_c(x) = \frac{1}{3}p_c^{\text{RK}}(x) + \frac{1}{3}p_c^{\text{RF}}(x) + \frac{1}{3}p_c^{\text{GBT}}(x), \quad \hat{y}(x) = \arg \max_c \hat{p}_c(x). \quad (1)$$

If RF/GBT training fails, classification falls back to RuleKit only.

In the RuleKit-only configuration, predictions are generated by RuleKit. For classification predictions, the application reports the RuleKit component’s predicted class and up to three label-consistent covering rules when available, using the first as the primary displayed rule. If no such rule is available, the interface reports this explicitly. In the hybrid configuration, these rules explain the RuleKit component rather than necessarily the blended decision; the interface indicates when the ensemble prediction differs from RuleKit. Figure 3 shows the training screen with the induced rules and metrics.

3.5. Reproducibility support

Because the feature values come from a stochastic model (LLM), the tool includes a built-in *repeatability test*: a single clip is processed N times and, for each numeric feature with a non-zero mean μ , the standard deviation σ and the coefficient of variation (CV) are calculated.

$$\text{CV} = 100 \cdot \frac{\sigma}{|\mu|} \% \quad (2)$$

For categorical features, the modal category and its frequency are reported instead.

4. Case Studies

We demonstrate the tool on three tasks spanning video and biomedical imagery, ranging from a few dozen to a thousand media items. The aim is to demonstrate the workflow and characterise its behaviour rather than to establish general superiority in predictive accuracy.

The first two case studies (Movie Memorability and Pneumonia) were processed using the Qwen2.5-VL-7B model and the third one (Mammography) using the Qwen3.6-35B model.

4.1. Movie memorability (MediaEval)

The target is the long-term memorability of short movie excerpts, a challenging prediction task [22]. The continuous memorability score was binned into four classes (0–3) by training-set quartiles; the test set has 106 clips. Phase 1 autonomously proposed eight features (Table 2) — a mix of visual, audio, and temporal dimensions — with no manual input.

Table 2

Features auto-discovered for the memorability task (Phase 1).

Feature	Domain
audio_intensity	numeric [0, 10]
audio_repetition	{none, low, high}
emotional_tone	{calm, tense, neutral}
scene_activity	{high, low}
scene_type	{indoor, outdoor, mixed}
temporal_change	numeric [0, 1]
visual_complexity	numeric [0, 10]
visual_dynamics	{static, dynamic}

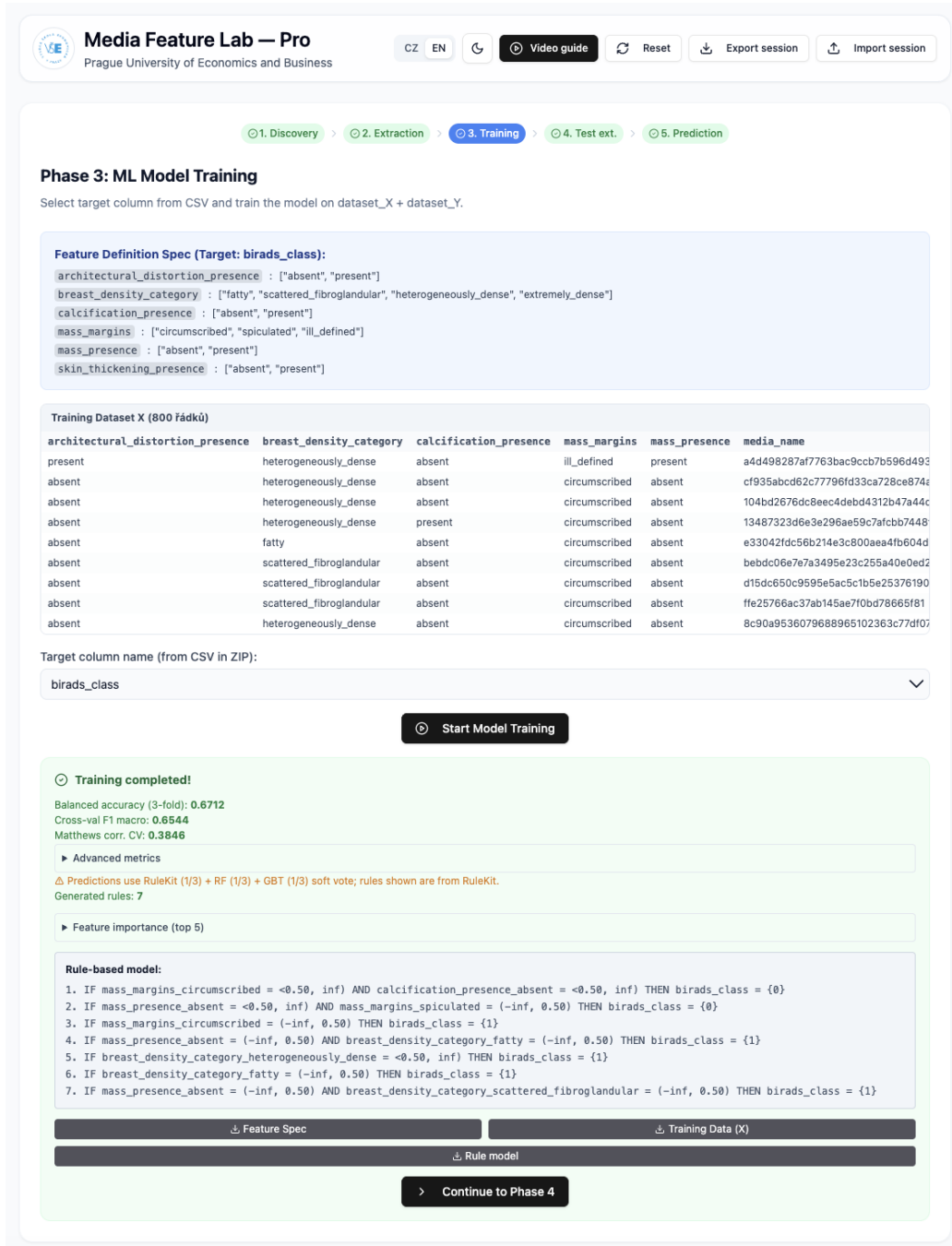


Figure 3: Phase 3 (training) for the mammography task in the public demo instance: the auto-discovered feature specification, the extracted feature matrix, the cross-validated metrics, and the seven induced RuleKit rules.

The induced RuleKit model contains 19 rules with on average 1.8 conditions (median 2) — compact enough to read in full. Representative rules reported in [9]:

IF visual_complexity < 6 AND temporal_change < 0.5 THEN memorability_class = 0;
IF audio_intensity ∈ [4, 7.5) AND visual_complexity < 4.5 THEN memorability_class = 1.

Accuracy (RQ2). Table 3 summarises the test-set results; balanced accuracy is the unweighted mean of recall computed separately for each class, and MCC denotes the Matthews correlation coefficient. RuleKit achieved accuracy 0.330 (95 % Wilson interval [24.8, 42.4] %), compared with 0.302 for the majority-class baseline, 0.349 for logistic regression, and 0.311 for the hybrid RuleKit/RF/GBT model. These absolute differences are small, and no formal paired significance test was performed. The

results therefore do not establish an accuracy advantage for the rule-based approach; its distinguishing property is the availability of human-readable rules. Because MediaEval evaluates memorability as regression using Spearman correlation, these classification metrics are not directly comparable to official leaderboard results.

Table 3

Movie memorability, four-class, test set of 106 clips.

Model	Acc.	Bal. acc.	F1 _{macro}	MCC
Majority-class baseline	0.302	0.250	0.116	0.000
Logistic regression	0.349	0.323	0.312	0.112
LLM + RuleKit (iter. 1)	0.330	0.304	0.289	0.083
LLM + RuleKit/RF/GBT (iter. 2)	0.311	0.287	0.270	0.057

Reproducibility (RQ3). Running extraction five times on one clip exposes where variance lives (Table 4): categorical features were perfectly stable, `audio_intensity` had CV 0 %, but `visual_complexity` varied with CV 51.6 %. This illustrates limited extraction stability for some features of the local 7B model. Re-running *Phase 1* on identical sample clips also produced a different feature set with markedly different predictive power, suggesting that stochastic feature discovery is an important source of performance variability.

Table 4

Per-feature variation across five extractions of one clip (excerpt). CV = coefficient of variation.

Feature	Mean $\pm$ Std	CV
<code>audio_intensity</code>	3.00 ± 0.00	0.0 %
<code>visual_complexity</code>	4.20 ± 2.17	51.6 %
<code>emotional_tone</code>	neutral (100 %)	–
<code>scene_type</code>	indoor (100 %)	–

4.2. Pneumonia detection (cross-domain)

To probe generality, we applied the unmodified application to binary pneumonia detection on an internal chest imagery dataset (80 training, 34 test; majority-class baseline 73.5 %). Phase 1 autonomously proposed seven radiology-oriented features (e.g., `lung_field_opacity` $\in [0, 10]$, `abnormal_lung_patterns`).

Three exploratory runs (Table 5) examined the effects of feature specification and classifier choice. Run 1 used RuleKit with independently discovered spec A and reached 41.2 %. RuleKit with spec B (Run 3) reached 76.5 %, a difference of 35.3 percentage points between the two independently discovered specifications. On spec B, the RF/GBT ensemble (Run 2) reached 82.4 %, 5.9 percentage points above Run 3. Thus, in these runs, the change in feature specification was associated with a substantially larger accuracy difference than the change in classifier on spec B. Because the test set contains only 34 samples and no formal significance test was performed, these numbers should be interpreted as an exploratory portability check rather than evidence of a general performance improvement. For context, specialised deep convolutional networks such as CheXNet reach radiologist-level pneumonia detection on large chest X-ray benchmarks [23]; the present workflow does not aim to compete with such models on accuracy and instead trades peak performance for a fully transparent decision path.

Representative rules from the RuleKit model in Run 3, reported in [9], included IF `lung_field_opacity` < 6.5 THEN `class` = 0 and IF `heart_size` $\neq$ large THEN `class` = 1. These examples illustrate that the induced decision conditions are directly inspectable, while their domain validity still depends on the faithfulness of the MLLM-extracted features.

Table 5

Pneumonia detection, binary, test set of 34 samples. Runs 1 and 2 use independently discovered specs; Run 3 reuses Run 2’s spec with the RuleKit classifier.

Model	Acc.	Bal. acc.	F1 _{macro}	MCC
Majority-class baseline	0.735	0.500	0.424	0.000
LLM + RuleKit (run 1, spec A)	0.412	0.564	0.410	0.141
LLM + RF/GBT (run 2, spec B)	0.824	0.738	0.755	0.519
LLM + RuleKit (run 3, spec B)	0.765	0.591	0.595	0.283

4.3. Binary BI-RADS classification (VinDr-Mammo)

The target is the BI-RADS assessment of a screening mammogram, taken from VinDr-Mammo [24], a public benchmark of 5,000 four-view exams double-read by radiologists. Because BI-RADS is assigned per breast, each image was labelled from its own breast: BI-RADS 1–2 form class 0, BI-RADS 4–5 form class 1, and BI-RADS 3 was excluded. Images were converted from DICOM to 8-bit PNG following the dataset authors’ reference procedure [25]. We drew a class-balanced sample from the official VinDr-Mammo splits, which are patient-disjoint, so no woman appears in both training and test – 800 training and 200 test images, that is 1,000 images of 500 breasts from 375 women. Within the sample, the two classes were additionally drawn from disjoint patients, so no woman contributes a breast to both classes.

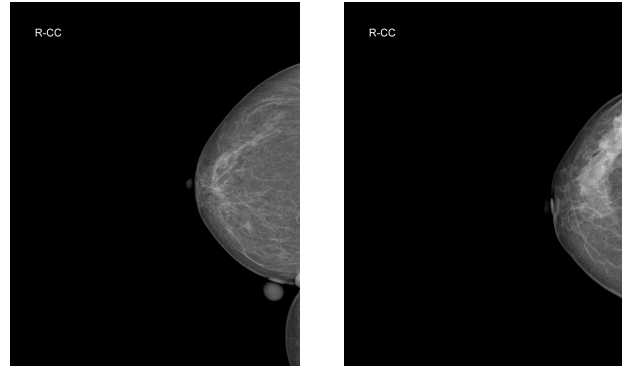


Figure 4: Example input images from VinDr-Mammo [24], both right craniocaudal views of ACR density B breasts. Left: BI-RADS 1 (class 0), no annotated finding. Right: BI-RADS 5 (class 1), with a mass and clustered suspicious calcifications in the dataset annotations.

Table 6

Features auto-discovered for the mammography task (Phase 1).

Feature	Domain
architectural_distortion_presence	{absent, present}
breast_density_category	{fatty, scattered fibroglandular, heterogeneously dense, extremely dense}
calcification_presence	{absent, present}
mass_margins	{circumscribed, spiculated, ill-defined}
mass_presence	{absent, present}
skin_thickening_presence	{absent, present}

The induced RuleKit model contains seven rules with on average 1.6 conditions (median 2):

```

IF mass_margins = circumscribed AND calcification_presence = absent THEN birads_class = 0;
IF mass_presence = absent AND mass_margins ≠ spiculated THEN birads_class = 0;
IF mass_margins ≠ circumscribed THEN birads_class = 1;
IF mass_presence = present AND breast_density_category ≠ fatty THEN birads_class = 1;

```

IF *breast_density_category* = *heterogeneously dense* THEN *birads_class* = 1;
 IF *breast_density_category* $\neq$ *fatty* THEN *birads_class* = 1;
 IF *mass_presence* = *present* AND *breast_density_category* $\neq$ *scattered fibroglandular* THEN
birads_class = 1.

The first three reproduce the standard reading that margin morphology, rather than the presence of a mass alone, separates benign from suspicious findings. The remaining four, however, key the suspicious class to breast density, which is a risk factor rather than a finding; the rule IF *breast_density_category* $\neq$ *fatty* THEN *birads_class* = 1 is close to a constant prediction: only 8 of the 1,000 images (0.8 %) are annotated ACR density A. Rule 5 is subsumed by rule 6. Density nonetheless carries the largest share of rule weight (0.36, against 0.27 for *mass_presence* and 0.27 for *mass_margins*). These rules do not dominate the decisions: five of seven rules predict class 1, yet the model predicts class 0 for 147 of 200 test images. Reading the rule set thus exposes a latent dependence on a risk factor rather than a finding — a defect the current weighting happens to suppress, and that a feature-attribution explanation would not surface. Figure 3 shows the corresponding training screen.

Accuracy (RQ2). Table 7 reports performance on the held-out test set. All predictions and evaluation metrics were computed at the individual image level ($n = 200$), without aggregating the craniocaudal and mediolateral oblique views into a breast-level diagnosis. The RuleKit configuration achieved an accuracy of 0.675 and the hybrid ensemble 0.670, both exceeding the 0.500 majority-class baseline. To account for sample non-independence (200 images representing 75 distinct patients), we report a conservative patient-clustered ($n = 75$) 95% Wilson confidence interval for RuleKit accuracy of [56.3, 77.0] %.

The RuleKit confusion matrix (91 true negatives, 9 false positives, 56 false negatives, 44 true positives) yields a class 1 precision of 0.830 and a recall of 0.440. Although the induced density rule quoted above applies to over 99% of the test set, the model output only 53 class 1 predictions in total. This discrepancy demonstrates that the final decision logic — whether driven by internal rule weighting or the ensemble’s soft-voting mechanism — does not strictly follow individual rule coverage. Clinically, while the model isolates suspicious features with relatively high precision, missing 56 out of 100 class 1 images (0.440 recall) fails to meet the high sensitivity thresholds required for mammography screening [26].

Table 7

Binary BI-RADS classification, held-out test set of 200 images.

Model	Acc.	Bal. acc.	F1 _{macro}	MCC
Majority-class baseline	0.500	0.500	0.333	0.000
LLM + RuleKit	0.675	0.675	0.656	0.397
LLM + RuleKit/RF/GBT	0.670	0.670	0.656	0.371

Reproducibility (RQ3). Table 8 reports five repeated extractions of the BI-RADS 5 mammogram of Figure 4; as every discovered feature is categorical, the table gives the modal category together with how often it was returned. Agreement is high: five of the six features returned the same value on all five passes, and *architectural_distortion_presence* on four. Stability, however, is not grounding. Of the five features for which VinDr-Mammo records ground truth, the model is right about the mass, the calcifications and the absence of skin thickening, but reports an architectural distortion that no radiologist annotated, and returns *heterogeneously dense* on every pass for a breast graded ACR B. Both errors fall on features that were perfectly or near-perfectly stable, and the density error is the one that propagates: four of the seven rules above test breast density. Margin morphology is not recorded in the dataset, so *mass_margins* cannot be scored.

Table 8

Five repeated extractions of one mammogram (the BI-RADS 5 image of Figure 4), against the dataset’s annotations.

Feature	Modal category (frequency)	Annotated ground truth
architectural_distortion_presence	present (80 %)	absent
breast_density_category	heterogeneously dense (100 %)	scattered fibroglandular
calcification_presence	present (100 %)	present
mass_margins	spiculated (100 %)	not recorded
mass_presence	present (100 %)	present
skin_thickening_presence	absent (100 %)	absent

4.4. Discussion and limitations

Across all three domains, RuleKit provides directly inspectable decision conditions over the extracted features. Rule interpretability, however, should not be conflated with feature faithfulness. An MLLM-extracted feature may itself be inaccurate, hallucinated, or insufficiently grounded in the input media. Consequently, a RuleKit explanation is faithful to the rule-based component conditional on the quality of the extracted representation, but is not automatically a faithful explanation of the original image or video. Carandang et al. [27] operationalise consistency across repeated iterations using the same prompt as intra-prompt stability, while evaluating correctness separately against a reference. Accordingly, the repeatability test measures extraction stability rather than grounding accuracy.

In the mammography case study, a further limitation concerns the unit of evaluation. Because BI-RADS is a breast-level assessment, a clinically meaningful evaluation would aggregate the two views into one per-breast decision. The pipeline has no such step, so the reported sensitivity is a per-image quantity.

Future work will address feature faithfulness and discovery stability, compare MLLM-extracted and deterministic descriptors by ablation, and evaluate the practical utility of the generated rules with domain experts.

5. Related Work

Our system sits at the intersection of LLM-based feature generation and rule learning, contributing to the growing field of neuro-symbolic AI. Balek et al. [6] generate interpretable features from text with an LLM and study them with SHAP; Ay [7] studies explainable biomedical image classification using LLM-generated textual descriptions and embedding representations, while Jeršova [28] studies explainable image classification based on word embeddings. We extend this line to video by using RuleKit [8], in the tradition of AQ/CN2/RIPPER [19, 20, 21], as an intrinsically interpretable predictive configuration and as the rule-based explanation component of the hybrid classifier. This contrasts with post-hoc explainers such as SHAP and LIME [1, 2], which explain predictions of already-trained models. On the perception side, our frame-based description layer follows common MLLM video practice [15, 16, 29]. Known MLLM limitations include hallucination [17] and poor temporal perception [18]; our RQ3 separately evaluates the repeatability of extracted features.

6. Conclusion and Open Science

Media Feature Lab demonstrates a practical neuro-symbolic architecture: a local multimodal LLM serves as an automatic neural feature engineer whose output feeds a symbolic rule learner, yielding a pipeline from raw media to human-readable IF . . . THEN rules without manual feature definition or programming. The three case studies do not establish a consistent predictive advantage over conventional baselines: the first two do not improve on the strongest reported baseline, whereas on the mammography assessment task the held-out accuracy of 0.675 exceeds the 0.500 baseline with an interval

that excludes it, though at a per-image sensitivity too low for clinical use. They demonstrate an end-to-end, reproducibility-oriented workflow and show that performance is sensitive to the automatically discovered feature representation. In the RuleKit-only configuration, the feature-to-decision mapping is directly inspectable. In the hybrid configuration, RuleKit provides an interpretable component of the decision rather than a complete explanation of the ensemble. In both cases, interpretation remains conditional on the faithfulness of the MLLM-extracted features.

Resources. The complete source code of Media Feature Lab is available at <https://github.com/kovm23/BP-featurelab>, including a link to the live demo. The tool ingests media as ZIP archives and labels as CSV, and exports feature matrices, rules, and predictions as CSV/JSON for downstream use.

Acknowledgments

The authors would like to acknowledge support from the grant IGA 37/2026 of the Faculty of Informatics and Statistics, VSE Prague. We thank the Department of Information and Knowledge Engineering for support.

Declaration on Generative AI

During the preparation of this work, the authors used Grammarly, Gemini, ChatGPT, and Claude in order to: draft content, check grammar and improve writing style, paraphrase and reword, and assist with programming. After using these tools/services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

- [1] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: *Advances in Neural Information Processing Systems*, volume 30, 2017, pp. 4765–4774.
- [2] M. T. Ribeiro, S. Singh, C. Guestrin, “why should I trust you?”: Explaining the predictions of any classifier, in: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144. doi:10.1145/2939672.2939778.
- [3] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: Visual explanations from deep networks via gradient-based localization, in: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626. doi:10.1109/ICCV.2017.74.
- [4] European Parliament and Council, Regulation (eu) 2024/1689 laying down harmonised rules on artificial intelligence (artificial intelligence act), *Official Journal of the European Union*, 2024. <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>.
- [5] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, et al., Qwen2.5-VL Technical Report, Technical Report, Alibaba Group, 2025. ArXiv:2502.13923.
- [6] V. Balek, L. Sýkora, V. Sklenák, T. Kliegr, LLM-based feature generation from text for interpretable machine learning, *Machine Learning* 114 (2025) 241. doi:10.1007/s10994-025-06867-1.
- [7] M. A. Ay, Explainable Biomedical Image Classification Using Large Language Models, Bachelor's thesis, Prague University of Economics and Business, 2025.
- [8] A. Gudyś, M. Sikora, L. Wróbel, RuleKit: A comprehensive suite for rule-based learning, *Knowledge-Based Systems* 194 (2020) 105480. doi:10.1016/j.knosys.2020.105480.
- [9] M. Koval, Video classification based on descriptions generated by large language models, Bachelor's thesis, Prague University of Economics and Business, 2026.
- [10] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in:

- Proceedings of the 38th International Conference on Machine Learning (ICML), 2021, pp. 8748–8763.
- [11] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, et al., Flamingo: a visual language model for few-shot learning, in: *Advances in Neural Information Processing Systems*, volume 35, 2022, pp. 23716–23736.
 - [12] H. Zhang, X. Li, L. Bing, Video-LLaMA: An instruction-tuned audio-visual language model for video understanding, in: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2023, pp. 543–553. doi:10.18653/v1/2023.emnlp-demo.49.
 - [13] B. Huang, X. Wang, H. Chen, Z. Song, W. Zhu, VTimeLLM: Empower LLM to grasp video moments, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 14271–14280.
 - [14] P. Zhang, K. Zhang, B. Li, G. Zeng, J. Yang, Y. Zhang, Z. Wang, H. Tan, C. Li, Z. Liu, Long context transfer from language to vision, *arXiv preprint arXiv:2406.16852* (2024).
 - [15] C. Fu, Y. Dai, Y. Luo, L. Li, S. Ren, R. Zhang, Z. Wang, C. Zhou, Y. Shen, M. Zhang, et al., Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis, *arXiv preprint arXiv:2405.21075* (2024).
 - [16] H. Wu, D. Li, B. Chen, J. Li, LongVideoBench: A benchmark for long-context interleaved video-language understanding, in: *Advances in Neural Information Processing Systems 37 (NeurIPS), Datasets and Benchmarks Track*, 2024.
 - [17] Z. Bai, P. Wang, T. Xiao, T. He, Z. Han, Z. Zhang, M. Z. Shou, Hallucination of multimodal large language models: A survey, *arXiv preprint arXiv:2404.18930* (2024).
 - [18] Y. Liu, S. Li, Y. Liu, Y. Wang, S. Ren, L. Li, S. Chen, X. Sun, L. Hou, TempCompass: Do video LLMs really understand videos?, in: *Findings of the Association for Computational Linguistics: ACL 2024*, 2024, pp. 8731–8772.
 - [19] R. S. Michalski, I. Mozetič, J. Hong, N. Lavrač, The multi-purpose incremental learning system AQ15 and its testing application to three medical domains, in: *Proceedings of the Fifth AAAI National Conference on Artificial Intelligence (AAAI-86)*, 1986, pp. 1041–1045. URL: <https://cdn.aaai.org/AAAI/1986/AAAI86-171.pdf>.
 - [20] P. Clark, T. Niblett, The CN2 induction algorithm, *Machine Learning* 3 (1989) 261–283. doi:10.1023/A:1022641700528.
 - [21] W. W. Cohen, Fast effective rule induction, in: *Proceedings of the 12th International Conference on Machine Learning (ICML)*, 1995, pp. 115–123.
 - [22] R. Cohendet, K. Yadati, N. Q. K. Duong, C.-H. Demarty, Annotating, understanding, and predicting long-term video memorability, in: *Proceedings of the ACM International Conference on Multimedia Retrieval (ICMR)*, 2018, pp. 178–186.
 - [23] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpan-skaya, M. P. Lungren, A. Y. Ng, CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning, *arXiv preprint arXiv:1711.05225* (2017). URL: <https://arxiv.org/abs/1711.05225>.
 - [24] H. H. Pham, H. Nguyen Trung, H. Q. Nguyen, VinDr-Mammo: A large-scale benchmark dataset for computer-aided detection and diagnosis in full-field digital mammography, *PhysioNet* (2022). doi:10.13026/br2v-7517, version 1.0.0.
 - [25] VinBigData Medical, VinDr-Mammo: dataset reference code, <https://github.com/vinbigdata-medical/vindr-mammo>, 2023. Accessed 2026-08-11.
 - [26] S. M. McKinney, M. Sieniek, V. Godbole, J. Godwin, N. Antropova, H. Ashrafi, T. Back, M. Chesus, G. S. Corrado, et al., International evaluation of an AI system for breast cancer screening, *Nature* 577 (2020) 89–94. doi:10.1038/s41586-019-1799-6.
 - [27] K. A. M. Carandang, J. M. Arana, E. R. Casin, C. Monterola, D. S. Tan, J. F. B. Valenzuela, C. Alis, Are LLMs reliable? an exploration of the reliability of large language models in clinical note generation, in: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, Association for Computational Linguistics, 2025, pp. 1413–1422. URL: <https://aclanthology.org/2025.acl-industry.99/>. doi:10.18653/v1/2025.acl-industry.99.

- [28] J. Jeršova, Explainable image classification based on word embeddings, Master's thesis, Prague University of Economics and Business, 2025.
- [29] K. Hu, F. Gao, X. Nie, P. Zhou, S. Tran, T. Neiman, L. Wang, M. Shah, R. Hamid, B. Yin, T. Chilimbi, M-LLM based video frame selection for efficient video understanding, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025, pp. 13702–13712.