SEMTRA: Global Semantic Transition and Rough-Set Rules for Auditable Post-hoc Explainability

Pavlo Radiuk *, Oleksander Barmak, Iurii Krak

Posted Date: 3 June 2026

doi: 10.20944/preprints202606.0230.v1

Keywords: explainable artificial intelligence; transition matrix; semantic attributes; rough sets; production rules; discretization; zero-shot learning

 doi:10.20944/preprints202606.0230.v1

Article

SEMTRA: Global Semantic Transition and Rough-Set Rules for Auditable Post-hoc Explainability

Pavlo Radiuk [1,*], Oleksander Barmak [1] and Iurii Krak [2,3]

[1] Department of Computer Science, Khmelnytskyi National University, 11 Instytuts'ka Str., 29016 Khmelnytskyi, Ukraine
[2] Department of Theoretical Cybernetics, Taras Shevchenko National University of Kyiv, 4d Akademika Glushkova Ave, 03680 Kyiv, Ukraine
[3] Laboratory of Communicative Information Technologies, V.M. Glushkov Institute of Cybernetics, 40 Akademika Glushkova Ave, 03187 Kyiv, Ukraine
* Correspondence: radiukp@khmnu.edu.ua

Abstract: Deep learning architectures produce highly effective yet uninterpretable latent representations, creating an interpretability gap that fundamentally impairs model verification and automated rule extraction. In this work, we propose Global Semantic Transition (SEMTRA), a post-hoc framework bridging this gap by translating numerical representations into auditable rough-set production rules without retraining the underlying feature extractor. Evaluated on the Animals with Attributes 2 (AwA2) benchmark, the proposed semantic transition achieved a mean absolute error of 0.1295. The systematically extracted rulebook successfully covered 86.40% of the test instances, significantly outperforming standard separate-and-conquer learners, which covered only 26.20%. The rules yielded a non-abstained covered accuracy of 40.73%, strictly verifying the transparent, mathematically robust portion of the model logic. Furthermore, the continuous semantic prototype transfer achieved a zero-shot accuracy of 48.43%, surpassing foundational baselines such as direct attribute prediction, which achieved 46.10%. A synthetic benchmark independently confirmed exact algorithmic recovery with a macro-F1 score of 0.8668. These findings establish that post-hoc semantic transition, combined with rigorous rough-set granulation, provides a reproducible pathway to transparent symbolic explanations, effectively balancing the tradeoff between predictive precision and algorithmic verifiability.

Keywords: explainable artificial intelligence; transition matrix; semantic attributes; rough sets; production rules; discretization; zero-shot learning

1. Introduction

Deep learning architectures have established themselves as definitive standards for visual recognition and continuous transfer learning because of their inherent ability to compress high-dimensional observations into remarkably effective latent representations. However, this compression mechanism inherently creates a severe black-box effect, meaning that neural features remain structurally optimized for predictive success rather than for human interpretability. As automated systems increasingly permeate mission-critical diagnostic domains, the demand for transparency has fundamentally shifted away from heuristic approximations toward rigorous, mathematically sound explanations. The recent manifesto detailing the evolution of Artificial Intelligence (AI) frameworks emphasizes the critical need for verifiable symbolic knowledge [1].

Current approaches addressing this transparency demand fall into diverse methodological categories. A comprehensive survey of modern explainable frameworks illustrates a prominent divide between local diagnostic techniques and globally interpretable model designs [2]. The problem under consideration is the inability to systematically translate compressed latent neural coordinates into a robust, global, and conflict-aware rulebook without heavily retraining or modifying the underlying feature extractor. While instance-based post-hoc explanations generate input-specific attribution maps, they completely fail to produce a holistic, reusable logic policy for rigorous auditing. Conversely,

 doi:10.20944/preprints202606.0230.v1

interpretable-by-design architectures necessitate predefined concept vocabularies and costly structural re-engineering of existing high-performance networks. This persistent interpretability gap severely impairs routine model verification and diagnostic safety. Consequently, it is highly important to present novel scientific contributions to address the raised problem by synthesizing a global mathematical bridge that connects uninterpretable latent manifolds with auditable symbolic spaces.

To construct this required bridge, the proposed SEMTRA framework explicitly implements a post-hoc continuous semantic transition. Early analytical studies validated the efficacy of transition matrices for interpreting complex networks within structured visual analytics environments [3]. These matrix structures were subsequently adapted and refined to extract understandable features in specialized healthcare informatics contexts [4]. Furthermore, theoretical research extended linear translation operators to effectively manage continuous geometric symmetries via equivariant algebraic manifolds [5]. SEMTRA advances this trajectory by converting the extracted continuous semantic variables into verifiable logical predicates using established rough-set principles. Decades ago, initial mathematical formalizations introduced the concept of bounding uncertainty through precise indiscernibility relations [6]. Comprehensive theoretical extensions eventually solidified the applicability of rough sets in structured reasoning about inconsistent data [7].

The goal of this study is to improve the auditability, transparency, and logical verifiability of pre-trained deep learning representations by applying a continuous semantic transition coupled with rigorous rough-set knowledge granulation.

The primary objective of this study is to formulate and empirically evaluate the SEMTRA framework, ensuring it effectively translates latent neural features into a conflict-aware symbolic rulebook while explicitly identifying ambiguous boundary regions. To fulfill this objective, the research proposes four major contributions:

- A dimensionally explicit transition-matrix formulation that maps compressed deep representations to interpretable semantic attributes without retraining the underlying feature extractor.
- The Weighted Entropy-Density Discretization (WEDD) method, a supervised discretizer that balances class-conditional entropy with a local density penalty to ensure stable symbolic boundaries.
- A rough-set rule-induction mechanism that constructs a global, conflict-aware production rulebook, featuring explicit handling of structural abstentions and boundary regions.
- A comprehensive evaluation on the AwA2 benchmark and controlled synthetic data, demonstrating the framework's ability to carefully balance predictive variance with rigorous algorithmic auditability.

The remainder of this paper is organized as follows. Section 2 details the related works covering explainable methods and symbolic rules. Section 3 outlines the proposed framework methodology, the continuous transition bridge, and the subsequent rough-set extraction mechanisms. Section 4 presents the qualitative and quantitative evaluations. Section 5 analyzes the implications of the structural explanation tradeoff. Finally, Section 6 concludes the manuscript.

2. Related works

2.1. Local Post-Hoc Explanation

Local post-hoc interpretation approaches estimate decision boundaries in the immediate vicinity of individual input samples to generate explanations. Highly ubiquitous frameworks, such as Shapley Additive Explanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME), heavily rely on perturbed neighborhood samples to construct sparse additive feature attributions [8]. Despite their widespread acceptance, these instance-specific models output fragmented attribution heatmaps rather than unified global policies. Recent advancements in local explainability address logical consistency by deploying universal rule-based explainers designed to maintain high local precision via conditional extraction paths [9]. Additionally, novel fusion methodologies combine independent feature importance rankings with structured local rule induction to stabilize the underlying attribution

 doi:10.20944/preprints202606.0230.v1

logic [10]. However, even optimized local methodologies fail to yield the exhaustive, dataset-wide conflict resolution protocols essential for comprehensive diagnostic auditing.

2.2. Concept-Based and Ante-Hoc Models

Concept-based techniques aim to directly augment numerical latent features with variables mapped to human-understandable terms. The foundational introduction of concept activation vectors provided a mathematical formulation to determine whether final neural decisions remain sensitive to predefined semantic directions [11]. More sophisticated architectural approaches, specifically Concept Bottleneck Models (CBMs), require networks to accurately predict intermediate semantic variables before executing final classification outputs [12]. Contemporary literature illustrates that these discrete human-aligned concepts can be natively embedded within models during the initial structural training phase [13]. Alternatively, selected optimization algorithms can reliably infer similar aligned concepts entirely through post-hoc modifications [14]. The primary limitation of ante-hoc methodologies remains their dependence on specialized initial training regimens, hindering their straightforward application to widely deployed black-box feature extractors.

2.3. Formal Knowledge Granulation and Rough Sets

The scientific pursuit of inducing symbolic rule structures predates the massive expansion of deep artificial neural networks. A comprehensive review highlights that classic hierarchical decision trees organically partition observations based on localized entropy metrics to generate completely transparent rule sequences [15]. Concurrently, separate-and-conquer logic algorithms iterate over modular rules to balance overall predictive accuracy with ultimate structural compactness, a process recently enhanced by integrating fuzzy logic components [16]. Modern structured applications successfully deploy these distinct rule-based algorithms to dissect complex unsupervised anomaly detection frameworks [17]. Nevertheless, most heuristic rule inducers forcefully dictate deterministic boundary classifications. Formulating stable partitions for rough-set structures necessitates robust preprocessing; modern non-monotonic discretization techniques demonstrate substantial utility in stabilizing interval boundaries [18]. Furthermore, novel kernel density estimators effectively map topological probability distributions within large, highly dimensional continuous spaces [19].

2.4. Semantic Transfer as a Validation Proxy

Continuous semantic transfer evaluation serves as a robust mechanism to assess whether a given methodology successfully captures generalized conceptual structures. For instance, the AwA2 benchmark facilitates rigorous testing of continuous semantic projection between seen domains and unseen domains [20]. This foundational repository provides standardized imagery accompanied by precise numeric class-level semantic prototype vectors [21]. Historically, the Direct Attribute Prediction (DAP) mechanisms established baseline predictive benchmarks through probabilistic semantic mappings [22]. Subsequent Indirect Attribute Prediction (IAP) formulations refined structural relationships among intermediate variables to better generalize conceptual knowledge [23].

To continuously elevate zero-shot classification performance, researchers have subsequently developed highly sophisticated embedding and generative architectures. Generative Framework for Zero-Shot Learning (GFZSL) designs significantly advanced zero-shot accuracy by synthesizing pseudo-samples derived exclusively from attribute distributions [24]. Diverse analytical methodologies calculate probabilistic convex combinations utilizing fixed semantic embeddings [25], while cross-modal transfer architectures project textual descriptions explicitly into visual coordinates to expand classification capabilities [26]. Structured optimization strategies apply semantic similarity embedding paradigms to measure conceptual distance [27], and sophisticated non-linear latent embeddings strictly enforce visual-semantic structural alignment constraints [28]. Correspondingly, specialized label-embedding operations [29], early visual-semantic pipelines [30], and structured joint formulations [31] maximize specialized compatibility functions between visual elements and attributes. Other notable approaches range from exceptionally fast linear regularizations [32] and dynamic synthetic class

 doi:10.20944/preprints202606.0230.v1

weight hallucinations [33] to complex autoencoder architectures that preserve inherent data topologies within projection domains [34]. Because these spaces remain highly dimensional, robust algorithmic implementations critically rely on efficient randomized Singular-Value Decomposition (SVD) matrix decompositions [35]. Visual modeling tasks in these domains exhibit the same complex representational geometry extensively analyzed in high-dimensional facial emotion recognition architectures [36].

Despite high accuracy, current zero-shot learning methods lack transparency. They rely on complex, non-linear models that act as secondary black boxes, completely hiding how visual features map to semantic concepts. By prioritizing predictive performance, they sacrifice mathematical traceability and make it impossible to track exact decision pathways. In contrast, the proposed framework treats semantic transfer strictly as a validation tool rather than a predictive competition. It employs a transparent transition mechanism to ensure all recovered concepts can be rigorously verified, avoiding the opacity of standard embedding techniques.

Broadly, the Explainable Artificial Intelligence (XAI) landscape faces several critical challenges: local methods lack global consistency, interpretable-by-design models require expensive retraining, and traditional rule learners force decisions even when data is ambiguous. SEMTRA addresses these gaps by combining a global, explicitly defined transition matrix with rough-set granulation. By prioritizing mathematical traceability and conflict awareness over raw predictive optimization, SEMTRA provides a verified post-hoc framework capable of extracting honest, reproducible diagnostic rules from pre-trained representations.

3. Materials and Methods

3.1. Framework Architecture Overview

The SEMTRA framework is structurally designed as a universal, modular, post-hoc proof-of-concept pipeline applicable to all possible XAI problems. Operating seamlessly over any frozen feature extractor, the methodology systematically produces a globally consistent and auditable symbolic rulebook. The conceptual architecture, illustrated in Figure 1, explicitly follows a rigorous two-phase progression: Phase 1, designated as Neural-to-Semantic Transition, and Phase 2, defined as Semantic-to-Symbolic Granulation.

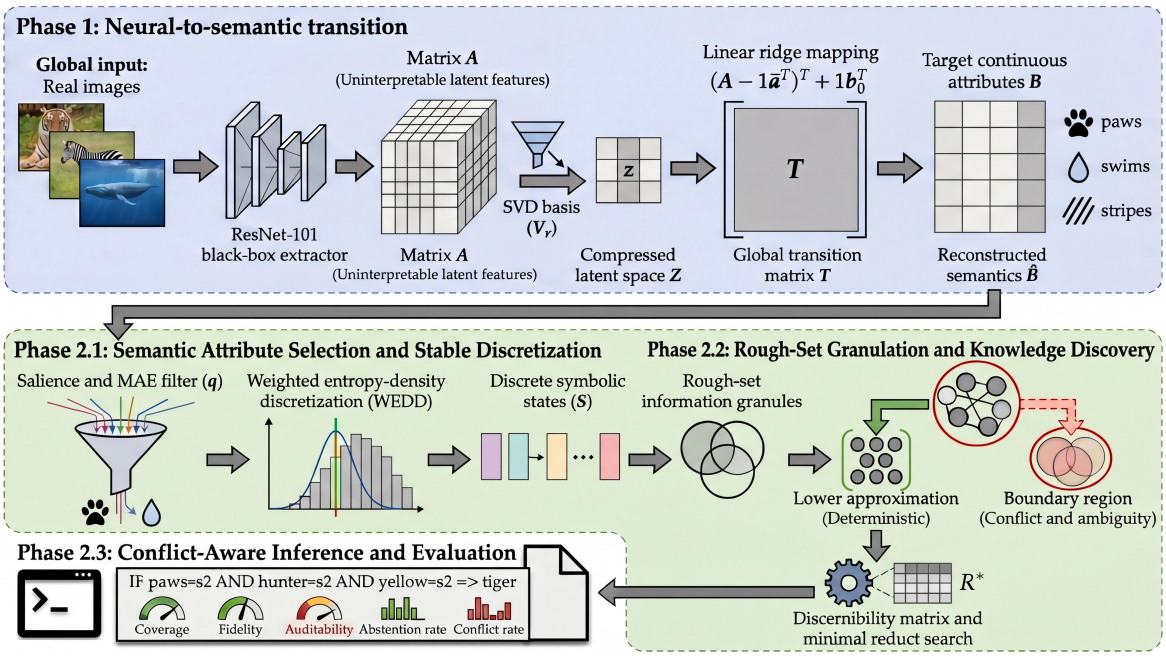


Figure 1. Conceptual architecture of the SEMTRA framework. The end-to-end pipeline sequentially translates uninterpretable latent neural features into continuous semantic attributes via a global transition matrix, and subsequently leverages density-aware discretization and rough-set granulation to induce a transparent, conflict-aware rulebook.

 doi:10.20944/preprints202606.0230.v1

The progression moves from the initial latent space $\mathbf{A}$, mapped through a robust global transition matrix $\mathbf{T}$, to reconstruct understandable continuous semantics $\mathbf{B}$. Subsequently, these semantics are discretized through the WEDD mechanism to finally induce the optimal symbolic production rules $\mathcal{R}$. To ensure absolute notational clarity and resolve discrete representation overlaps, the matrix of all discretized states is strictly denoted as $\mathbf{S}$, with individual symbolic states for object i and attribute j defined as s_{ij}. The ultimate output of the framework structurally follows the standardized rule format r:

$$r: \quad \bigwedge_{j \in R_r} [s_j = v_j] \Rightarrow d = y_r, \tag{1}$$

where R_r is the optimized minimal set of semantic antecedents, s_j represents the specific discrete state variable required by the rule, and v_j is the corresponding exact logical value bound.

The mathematical formalization of this entire universal decision-making system S is defined as:

$$S = (\mathcal{U}, \mathcal{C}, d, \mathcal{V}, f), \tag{2}$$

where $\mathcal{U}$ is the universe of distinct objects, $\mathcal{C}$ is the set of conditional semantic attributes, d is the target decision attribute, $\mathcal{V}$ is the union of valid attribute domains, and f is the deterministic information function.

By explicitly separating heavy computational inference from lightweight symbolic extraction, SEMTRA ensures that explanations across any classification task are fully traceable and reproducible.

3.2. Phase 1: Neural-to-Semantic Transition

The first phase establishes a mathematically explicit bridge between the uninterpretable latent representation space of a deep model and a target semantic attribute space. Let $\mathbf{A} \in \mathbb{R}^{m \times k}$ denote the matrix of raw, untransparent latent features. To ensure that the subsequent mapping is robust to dataset-specific biases, we first perform dimensional centering. Let the vector $\bar{\mathbf{a}} \in \mathbb{R}^k$ denote the absolute mean computed across the entire training distribution. The centered representation matrix $\mathbf{A}_c$ is subsequently computed via:

$$\mathbf{A}_c = \mathbf{A} - \mathbf{1}\bar{\mathbf{a}}^\top \in \mathbb{R}^{m \times k}. \tag{3}$$

To handle potential noise and extreme collinearity within high-dimensional latent spaces commonly found in complex architectures, a rank-r SVD basis, denoted as $\mathbf{V}_r \in \mathbb{R}^{k \times r}$, is rigorously extracted [35]. This yields a computationally compressed latent space $\mathbf{Z}$:

$$\mathbf{Z} = \mathbf{A}_c \mathbf{V}_r \in \mathbb{R}^{m \times r}. \tag{4}$$

The mapping to a human-understandable target semantics matrix $\mathbf{B} \in \mathbb{R}^{m \times \ell}$, where ℓ represents the count of conceptually interpretable variables, is established via a ridge-regression weight matrix $\mathbf{W} \in \mathbb{R}^{r \times \ell}$, formalized as the following minimization problem:

$$\mathbf{W}^* = \underset{\mathbf{W} \in \mathbb{R}^{r \times \ell}}{\arg\min} \|\mathbf{Z}\mathbf{W} - \mathbf{B}\|_F^2 + \alpha \|\mathbf{W}\|_F^2. \tag{5}$$

The choice of a linear ridge operator is an intentional methodological decision. We deliberately employ a linear transition matrix because it is fundamentally interpretable in itself; human auditors can directly inspect the exact weights of each feature to mathematically track structural contributions. Furthermore, linearity guarantees topological stability, preventing the unpredictable boundary distortions and chaotic gradients typically associated with deep non-linear alternatives. With an estimated analytical intercept $\mathbf{b}_0 \in \mathbb{R}^\ell$, the primary global transition operator $\mathbf{T}$ and the reconstructed continuous attributes $\widehat{\mathbf{B}}$ are successfully recovered:

$$\mathbf{T} = \mathbf{V}_r \mathbf{W}^* \in \mathbb{R}^{k \times \ell}, \quad \widehat{\mathbf{B}} = (\mathbf{A} - \mathbf{1}\bar{\mathbf{a}}^\top)\mathbf{T} + \mathbf{1}\mathbf{b}_0^\top. \tag{6}$$

 doi:10.20944/preprints202606.0230.v1

This linear transformation enforces the model's internal logic into a semantically grounded coordinate system. Global semantic reconstruction fidelity is strictly evaluated using the Mean Absolute Error (MAE):

$$\mathrm{MAE}_Q = \frac{1}{|Q|\ell} \sum_{i \in Q} \sum_{j=1}^{\ell} |\hat{b}_{ij} - b_{ij}|. \tag{7}$$

3.3. Phase 2.1: Semantic Attribute Selection and Stable Discretization (WEDD)

Once the continuous semantic attributes $\widehat{\mathbf{B}}$ are reconstructed, the next operation isolates a highly relevant subset of q predictive attributes (empirically set to $q = 15$ for rule induction) to prevent computational combinatorial explosion. Let e_j signify the normalized validation MAE for attribute j, and let $\eta_j = \|\mathbf{T}_{:j}\|_2$ represent the mathematical transition salience. The formal selection score is defined as:

$$\mathrm{score}_j = \lambda_s \widetilde{\eta}_j + (1 - \lambda_s)(1 - \widetilde{e}_j). \tag{8}$$

where $\widetilde{(\cdot)}$ explicitly denotes Min-Max normalization mapping values robustly into the $[0, 1]$ interval across all ℓ available target attributes.

The parameters λ_s and λ_H serve as dedicated auditor's control knobs. Specifically, $\lambda_s \in [0, 1]$ allows the user to mathematically prioritize either the strength of the neural-to-semantic signal (purity) or the precision of the reconstruction error. A comprehensive sensitivity analysis justifying the default configuration of the auditor's control knobs ($\lambda_s = 0.5, \lambda_H = 0.65$) is provided in Appendix A, demonstrating their direct impact on rulebook compactness and coverage. The selected attributes are subsequently transformed into discrete symbolic intervals using the WEDD mechanism.

For a candidate boundary threshold θ splitting an attribute x_j into a left computational subset L_θ and a corresponding right subset R_θ, the standard conditional decision entropy $H_j(\theta)$ is calculated using Equation (9) and Equation (10):

$$H_j(\theta) = \frac{|L_\theta|}{n} H(d \mid L_\theta) + \frac{|R_\theta|}{n} H(d \mid R_\theta), \tag{9}$$

$$H(d \mid G) = -\sum_{y=1}^{N_c} p(y \mid G) \log_{N_c} p(y \mid G), \tag{10}$$

where N_c is the number of distinct decision classes, ensuring that $H(d|G) \in [0, 1]$ perfectly across all evaluated datasets, thereby stabilizing the final WEDD objective across vastly different problem scales.

To prevent jitter in the resulting logic, WEDD introduces a local probability density proxy $p_j(\theta)$, computed via a Gaussian kernel function $K(u)$ with a structural bandwidth h_j:

$$p_j(\theta) = \frac{1}{n h_j} \sum_{i=1}^{n} K\left(\frac{\theta - x_{ij}}{h_j}\right). \tag{11}$$

The final discretization threshold is determined by minimizing the unified WEDD multiobjective function:

$$J_j(\theta) = \lambda_H \widetilde{H}_j(\theta) + (1 - \lambda_H) \widetilde{p}_j(\theta). \tag{12}$$

Explicitly, both $\widetilde{H}_j(\theta)$ and $\widetilde{p}_j(\theta)$ are precisely scaled utilizing standard Min-Max normalization into the $[0, 1]$ interval across all generated candidate thresholds, guaranteeing the multiobjective weights behave uniformly and symmetrically.

We do not simply minimize entropy as standard decision trees do. The parameter $\lambda_H \in [0, 1]$ is another critical auditor's knob allowing the user to seamlessly choose between rule informational purity and topological noise robustness. By rigorously accounting for data density, the threshold avoids cutting through dense data clusters. Instead, boundaries are anchored in sparser topological valleys, rendering the generated logic highly resistant to underlying representational noise. The complete procedural implementation of these specific early phases is outlined in Algorithm 1.

 doi:10.20944/preprints202606.0230.v1

Algorithm 1 Neural-to-Semantic Transition and WEDD Discretization (Phases 1 and 2.1)

Input: Universe of uninterpretable representations $\mathbf{A}$, aligned continuous target attributes $\mathbf{B}$, decision labels $\mathbf{d}$.
Output: Discretized symbolic states $\mathbf{S}$.
Initialize the global decision-making system S via Equation (2).
Compute the centered representation matrix $\mathbf{A}_c$ utilizing Equation (3).
Extract the rank-reduced compressed latent space $\mathbf{Z}$ via Equation (4).
Optimize the ridge-regression weight matrix $\mathbf{W}^*$ applying Equation (5).
Reconstruct continuous semantic attributes $\widehat{\mathbf{B}}$ and establish the global transition operator $\mathbf{T}$ via Equation (6).
Evaluate the semantic reconstruction fidelity by computing the MAE via Equation (7).
Compute the composite attribute selection score score_j using Equation (8) and isolate q highly relevant concepts.
for each selected attribute $j \in \{1, \ldots, q\}$ **do**
    Compute the conditional decision entropy $H_j(\theta)$ for candidate thresholds utilizing Equation (9) and Equation (10).
    Estimate the local probability density proxy $p_j(\theta)$ applying Equation (11).
    Determine the optimal discretization threshold by minimizing the unified WEDD objective $J_j(\theta)$ via Equation (12).
end for
Map the continuous attributes into the discrete symbolic states $\mathbf{S}$.
Return S

3.4. Phase 2.2: Rough-Set Granulation and Knowledge Discovery

Following stable discretization, the framework constructs a symbolic logic layer through mathematically formal knowledge granulation. Formally, the support of a specific rule r ($A \Rightarrow B$) is defined as the absolute number of instances satisfying both the antecedent and the consequent simultaneously: $\text{Supp}(r) = |A \cap B|$. The confidence is subsequently formulated as the conditional probability of the consequent given the antecedent: $\text{Conf}(r) = \frac{|A \cap B|}{|A|}$. The core of this robust process is the establishment of a rigorous indiscernibility relation. Two distinct objects x_i and x_j are rendered indiscernible with respect to a subset of optimal attributes $P \subseteq \mathcal{C}$ if they logically share perfectly identical discrete symbolic signatures. Formulated in Equation (13), this relation partitions the universe of objects into tightly bound structural information granules, denoted as $[x]_P$:

$$IND(P) = \{(x_i, x_j) \in \mathcal{U} \times \mathcal{U} \mid \forall c_k \in P, f(x_i, c_k) = f(x_j, c_k)\}, \tag{13}$$

where $f(x_i, c_k) = s_{ik}$ explicitly utilizes the information function f from the defined decision system to represent the discrete symbolic state mapping of object x_i evaluated across the specific attribute c_k.

This precise granulation allows the framework to explicitly distinguish between deterministic logical certainties and structural ambiguities. For a specific analytical decision class Y, we systematically define the structural Lower Approximation $\underline{P}Y$:

$$\underline{P}Y = \{x \in \mathcal{U} \mid [x]_P \subseteq Y\}. \tag{14}$$

The Lower Approximation encapsulates all information granules that completely and unambiguously belong to class Y, forming the irrefutable foundation for generating perfectly certain rules. Conversely, granules that simultaneously intersect multiple diverse classes form the Boundary Region. This Boundary Region mathematically identifies the exact conflict zones where the compressed latent neural features become inherently indiscernible and fundamentally unreliable.

To ensure that the resulting explanations are conceptually compact, we sequentially apply a mathematically rigorous minimal reduct search. A reduct R^* is the smallest possible subset of attributes

 doi:10.20944/preprints202606.0230.v1

that preserves decision-making consistency. The generalized minimal reduct R_ρ^* optimization is formally defined as:

$$R_\rho^* = \arg\min_{R \subseteq \mathcal{C}_\rho} |R| \quad \text{s.t.} \quad \text{Conf}(R \Rightarrow d_\rho) \geq \tau, \quad \text{Supp}(R \Rightarrow d_\rho) \geq s_{\min}, \tag{15}$$

where ρ fundamentally denotes a specific target decision class.

To efficiently navigate this logic, the framework formulates a global discernibility matrix. Each interactive element c_{ij} securely identifies the exact set of differentiating semantic attributes between two differently classified objects:

$$c_{ij} = \{c_k \in \mathcal{C} \mid s_{ik} \neq s_{jk}\}, \quad \text{for } y_i \neq y_j. \tag{16}$$

3.5. Phase 2.3: Conflict-Aware Inference and Evaluation

The final phase of the SEMTRA framework governs the application of the induced rulebook to unobserved data. During real-time analytical inference, new object representations flow through the identical transition and discretization sequence established in Phases 1 and 2.1, instantly converting raw neural features into active symbolic signatures. This ensures that the explanation logic remains strictly faithful to the original model's "view" of the data.

3.5.1. Conflict Resolution and Soft Matching

When querying the generalized rulebook, two types of logical edge cases may emerge: boundary conflicts and structural gaps. Boundary conflicts occur when multiple activated rules suggest different outcome classes for the same symbolic signature. These are systematically resolved using a carefully constructed integral support score $S(y_k)$, as formulated in Equation (17):

$$S(y_k) = \sum_{r \in \mathcal{R}_{\text{active}}, d_r = y_k} \text{Conf}(r) \cdot \text{Supp}(r) \cdot w_r, \tag{17}$$

where $w_r = \left(\frac{1}{|R_r|} \sum_{j \in R_r} p_j(\theta_j) \right)^{-1}$ functions as a targeted density-based weight strictly inherited from the WEDD discretization phase, explicitly taking the inverse of the average probability density evaluated precisely at the thresholds utilized in rule r, effectively prioritizing rules fundamentally anchored in stable, structurally optimal attribute regions.

In scenarios where no exact production rules match the presented inference data (structural gaps), the framework employs soft analytical matching. This fallback mechanism evaluates partial logical alignments using the fundamental symbolic Hamming distance d_H:

$$d_H(\mathbf{s}_i, \mathbf{t}_y) = \frac{1}{|R_y|} \sum_{j \in R_y} \mathbb{I}[s_{ij} \neq t_{yj}], \tag{18}$$

where $\mathbf{s}_i$ is the signature of the test instance and $\mathbf{t}_y$ is the formal rule antecedent.

This soft matching natively utilizes a masked Hamming distance evaluated precisely only over the highly specific antecedent conditions $|R_y|$ encoded within rule $\mathbf{t}_y$. If the computed distance formally exceeds a predefined tolerance threshold ($\tau_H = 0.25$), the overall system correctly triggers a structural abstention, firmly identifying the evaluated case as currently unverifiable within the generated rulebook.

 doi:10.20944/preprints202606.0230.v1

3.5.2. Fidelity and Coverage Metrics

To quantify the quality of the extracted explanation layer, SEMTRA distinguishes between two levels of fidelity. First, Covered Fidelity evaluates the accuracy of the rulebook exclusively on instances where a definitive rule was matched:

$$\text{Fidelity}_{\text{covered}} = \frac{1}{|C_Q|} \sum_{i \in C_Q} \mathbb{I}[r(u_i) = g(u_i)], \tag{19}$$

where $g(u_i)$ is the prediction of the base black-box model and C_Q is the deterministically covered subset.

Second, All-Object Fidelity logically treats required abstentions as explicit structural failures, providing a conservative measure of the framework's total explanatory power:

$$\text{Fidelity}_{\text{all}} = \frac{1}{|Q|} \sum_{i \in Q} \mathbb{I}[r(u_i) = g(u_i)]. \tag{20}$$

This dual-metric approach formalized by Equations (19) and (20) allows auditors to assess the important tradeoff between explanation precision and rulebook coverage.

Finally, the entire sequential progression, from granulation to conclusive inference, is synthesized in Algorithm 2.

Algorithm 2 Rough-Set Granulation, Reduct Solver, and Inference (Phases 2.3–2.5)

Input: Discretized symbolic states **S**, decision vector **d**, validation instances.
Output: Conflict-aware production rulebook $\mathcal{R}$ and inference predictions.
Establish the indiscernibility equivalence relations $IND(P)$ across all discrete states utilizing Equation (13).
Aggregate objects into information granules and isolate the deterministic boundary using the lower approximation $\underline{P}Y$ via Equation (14).
Formulate the comprehensive discernibility matrix M, populating individual elements c_{ij} using Equation (16).
Execute the weighted heuristic search to extract the minimal optimal reducts R_ρ^* applying Equation (15).
Synthesize the discrete granules and reducts into the formal logical rule format r via Equation (1).
Compile the verified rules into the global conflict-aware rulebook $\mathcal{R}$.
for each unseen test instance **do**
    Attempt to match the instance against the rulebook antecedents.
    if multiple conflicting rules are activated **then**
        Resolve the boundary conflict by maximizing the integral support score $S(y_k)$ utilizing Equation (17).
    else if no exact rule conditions are satisfied **then**
        Execute soft matching by minimizing the symbolic Hamming distance d_H via Equation (18).
    end if
end for
Assess the overarching algorithmic explanation quality by calculating covered fidelity via Equation (19) and all-object fidelity via Equation (20).
Return $\mathcal{R}$

3.6. Evaluation Metrics for Auditability

To evaluate SEMTRA as an XAI proof-of-concept, we transcend standard predictive accuracy with a multi-dimensional suite of metrics focused on auditability, quantifying the trade-off between extracted logic complexity and faithfulness to the deep model. Rulebook Coverage (Cov) represents explanation "breadth," measuring the percentage of tested instances Q receiving a deterministic symbolic explanation:

$$\text{Cov} = \frac{|C_Q|}{|Q|}. \tag{21}$$

 doi:10.20944/preprints202606.0230.v1

High coverage denotes a comprehensive rulebook capturing the majority of the model's decision space. Logical Fidelity (F_{cov}) captures explanation "truthfulness," measuring how often the rulebook's prediction matches the base model's output on the covered subset, ensuring auditors inspect actual neural network logic over a disconnected surrogate. The Structural Abstention Rate (Abs) identifies regions where the neural representation is too ambiguous or conflicted for rules, revealing where the model is "inexplicable":

$$\text{Abs} = 1 - \text{Cov}. \tag{22}$$

Rulebook Compactness evaluates the total number of rules and the average number of antecedents per rule; following the principle of cognitive load, a smaller, shallower rulebook is inherently more auditable by human experts. Finally, the Conflict Rate (CR) tracks boundary conflicts during inference, diagnosing inherent "indiscernibility" in underlying latent representations. Calculated from instances X_{conflict} mapping directly into conflict zones:

$$\text{CR} = \frac{|X_{\text{conflict}}|}{|Q|}. \tag{23}$$

This comprehensive suite of metrics demonstrates that SEMTRA provides a transparent and measurable bridge between high-stakes neural decisions and auditable symbolic knowledge.

3.7. Experimental Protocol and Proof-of-Concept Design

To validate the SEMTRA framework as a reliable instrument for post-hoc deep learning auditing, a multi-stage experimental protocol evaluates how it bridges continuous latent representation spaces and discrete logic, manages structural uncertainty, and yields highly auditable verification artifacts across real-world semantic reconstruction, zero-shot logical transfer, algorithm stability, and synthetic baseline recovery.

3.7.1. Datasets, Feature Extraction, and Base Modeling

Primary real-world evaluations utilize the AwA2 [21] benchmark. Eliminating image-level augmentations, the framework extracts a 2048-dimensional global feature matrix directly from an ImageNet Large Scale Visual Recognition Challenge (ILSVRC)-pretrained Residual Network (ResNet)-101 representation layer. A base ridge classifier is trained on variance-screened representations via an auxiliary 64-dimensional compression step (Principal Component Analysis reduced to 64 components retaining 95% variance) to establish robust baseline prediction metrics as the explicit predictive target for explanations. Using a fixed random seed (42), the continuous semantic bridge is optimized via grid search across ridge penalty parameters ($\alpha \in \{0.01, 0.1, 1, 10, 100\}$). It is noted that the optimal α hit the lower bound of the grid search space (0.01); thus, exploring smaller regularization values could potentially yield marginal improvements. Granulation mechanisms apply specific constraint thresholds like WEDD bounds (maximum depth of 2, minimum bin size of 30) and greedy reduct parameters ($\tau = 0.84$, minimum support of 18). Exhaustive hyperparameter configurations and baseline predictor performance metrics are documented in Appendix A. The significant disparity between the Top-1 Accuracy (70.93%) and the Macro-F1 score (53.99%) indicates that the base Ridge classifier experiences partial majority-class collapse, heavily favoring well-represented classes over minority taxonomies due to class imbalance.

3.7.2. Empirical Evaluation Splits

To isolate memorization capacity from semantic generalization, two testing splits are used. Protocol A (Closed-world Fidelity) assesses neural-to-semantic bridge reconstruction fidelity on a consistent 50-class stratification using MAE, Root Mean Square Error (RMSE), and semantic correlation, while quantifying symbolic rulebook properties by tracking test coverage, non-abstained accuracy, conflict rates, and structural abstention rates. Protocol B (Zero-shot Transferability) uses the official xlsa17 split as a semantic validation proxy. For Protocol B, to strictly prevent data leakage, a disjoint base model

 doi:10.20944/preprints202606.0230.v1

and transition matrix were trained exclusively on the 40 seen classes. Applying a semantic transition matrix trained on 40 seen classes to infer logic on 10 strictly unseen classes proves globally extracted continuous prototypes capture generalized visual-semantic structures rather than dataset-specific artifacts.

3.7.3. Methodological Ablations and XAI Baselines

To contextualize scientific tradeoffs in global explainability, Transition Operator Ablation benchmarks the global linear ridge-regression operator against a Radial Basis Function (RBF) kernel ridge and a two-layer Multi-Layer Perceptron (MLP) regressor to analyze tradeoffs between continuous reconstruction correlation and symbolic rule-transfer coverage. Discretization Benchmarking compares the proposed WEDD strategy against classical discretization algorithms (Minimum Description Length Principle (MDLP)-like entropy approaches, equal-frequency, and equal-width discretizers), logging boundary stability diagnostics, rule lengths, and structural conflict rates in Appendix B. Explainability Paradigms contrast the rough-set rulebook with heuristic symbolic models (Classification and Regression Trees (CART) and separate-and-conquer rule learners) to assess structural uncertainty management. Local post-hoc attribution surrogates (LIME and SHAP) are evaluated via top-k feature agreement, direction agreement, and rank correlation, alongside interpretable-by-design architectures like frozen-feature CBMs and Testing with Concept Activation Vectors (TCAV).

3.7.4. Algorithmic Recovery on Synthetic Data

To isolate core mathematical machinery from the inherent indiscernibility of real-world latent manifolds, a fully controlled synthetic benchmark simulates a 10-dimensional continuous feature space governed by a predefined, ground-truth logical rule dictionary. Semantic noise ($\sigma \in [0, 0.20]$) injected across random computational seeds maps the structural limits of WEDD and rough-set induction mechanisms, measuring exact threshold recovery error, rule Jaccard similarity, macro-F1 scores, and operational coverage.

3.7.5. Diagnostic Tracking and Inference Stability

A rigorous suite of visual and statistical diagnostics clarifies the proposed algorithms, assessing the balance between logical determinism and coverage via fidelity-coverage tradeoff projections. At the attribute level, target transition salience is mapped against test reconstruction errors to justify semantic selection, differentiating robust physical descriptors from high-variance semantic concepts. Real-time analytical inference is audited by generating precise rule traces mapping unseen objects to active logical antecedents, confidence parameters, and source decisions. The structural robustness of inference pathways is systematically evaluated against semantic-representation perturbations; resultant decision consistency, coverage deviation, and rule matching stability metrics are cataloged in Appendix C.

4. Results

This section presents the empirical validation of the SEMTRA framework according to the Pillars of Auditability defined in the experimental protocol. We assess the framework's ability to reconstruct semantics, generate auditable rules, and handle zero-shot transfer as a proxy for conceptual integrity.

4.1. Semantic Reconstruction Performance (Protocol A)

The first stage of the experimental validation verifies the integrity of the Neural-to-Semantic Bridge. This is a prerequisite for a reliable audit: the extracted rules can only be trusted if the underlying transition matrix successfully recovers the intended semantic concepts from the latent neural representations.

The performance of the linear transition operator under Protocol A is summarized in Table 1. On the test set, the framework achieved an MAE of 0.1295 ± 0.0016 and a mean semantic correlation of

 doi:10.20944/preprints202606.0230.v1

0.6828. These results indicate that the global transition matrix effectively preserves the topological structure of the latent space while aligning it with human-understandable descriptors.

Table 1. AwA2 Protocol A transition reconstruction metrics (ridge alpha 0.01; retained SVD variance 0.2952). Metrics are reported as mean ± standard deviation across 5 random seeds.

Split	MAE	RMSE	Mean semantic correlation
train	0.1292 ± 0.0014	0.1822 ± 0.0018	0.6843 ± 0.0040
validation	0.1299 ± 0.0015	0.1832 ± 0.0018	0.6807 ± 0.0042
test	0.1295 ± 0.0016	0.1826 ± 0.0019	0.6828 ± 0.0039

To justify the choice of a linear operator over more complex alternatives, an ablation study was performed (Table 2). While non-linear regressors (e.g., two-layer MLP) marginally improve reconstruction in some metrics, they introduce significant black-box opacity into the transition itself. The linear ridge transition maintains a high correlation (0.7759) and significantly higher rule coverage (0.8718 ± 0.0075) compared to non-linear variants, establishing it as the most defensible choice for an auditable framework. The near-zero rule accuracy (0.0027) for the RBF kernel ridge indicates that severe overfitting in the high-dimensional latent space collapsed the global semantic mapping entirely, underscoring the fragility of unregularized non-linear projections.

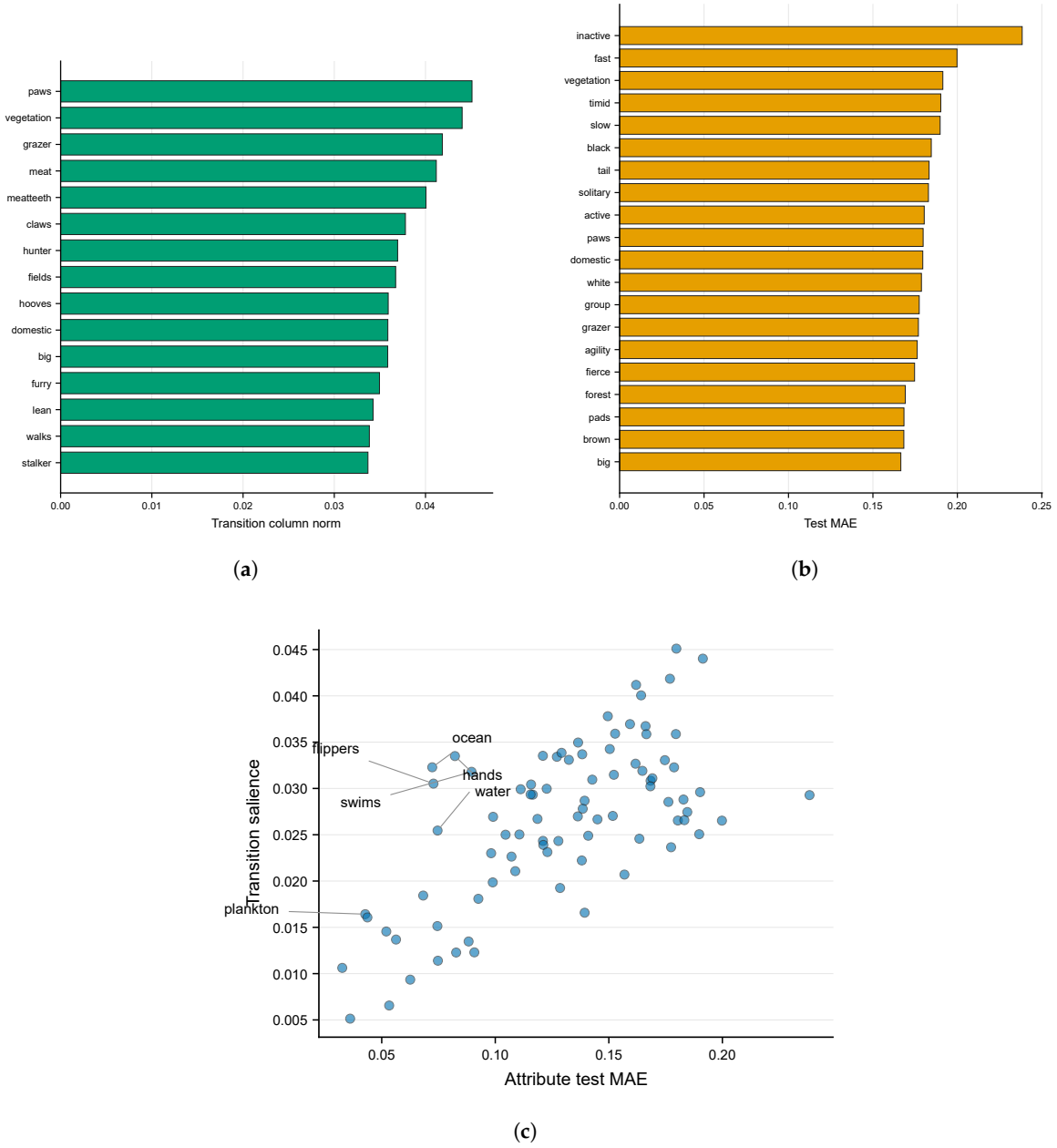
Table 2. Semantic reconstruction and rule-transfer ablation for linear and nonlinear operators. Means ± standard deviations across 5 seeds are provided for the proposed linear operator.

Operator	MAE	RMSE	Corr.	Rule cov.	Rule acc.	Runtime (s)
Linear ridge transition	0.1076 ± 0.0012	0.1573	0.7759	0.8718 ± 0.0075	0.1941	0.0451
RBF kernel ridge	0.1958	0.2583	0.1176	1.0000	0.0027	2.8573
Two-layer MLP regressor	0.1685	0.2355	0.4497	0.8390	0.0915	0.6914

Note: All runtime measurements reported in this table were executed sequentially on a single NVIDIA RTX 3090 Graphics Processing Unit (GPU) and an Intel Core i9-10900K Central Processing Unit (CPU) to ensure a standardized computational baseline.

Detailed attribute-wise diagnostics, visualized in Figure 2, reveal that visually distinct attributes such as paws, vegetation, and water achieve the highest transition salience and lowest reconstruction errors. Conversely, more abstract descriptors (e.g., inactive) exhibit higher variance. This variance is explicitly managed in the next phase by the WEDD algorithm, which prioritizes stable, low-error attributes for rule induction.

 doi:10.20944/preprints202606.0230.v1



(**a**) (**b**)

(**c**)

Figure 2. Semantic attribute diagnostics: (**a**) transition salience by semantic attribute; (**b**) attribute-wise semantic reconstruction error on AwA2; and (**c**) transition salience against test reconstruction error; labeled points are high-score selected attributes.

As shown in Table 3, the composite selection score effectively isolates the top q attributes that offer the best balance between neural signal strength and reconstruction precision. For brevity, the top 12 selected attributes are shown in Table 3.

 doi:10.20944/preprints202606.0230.v1

Table 3. Selected semantic attributes used for rough-set rule induction (first 12 shown).

Semantic attribute	Salience	Test MAE	Selection score
stripes	0.025	0.105	0.202
hooves	0.036	0.153	0.208
swims	0.032	0.090	0.288
paws	0.045	0.180	0.228
longneck	0.020	0.099	0.166
hunter	0.037	0.159	0.204
ocean	0.032	0.072	0.346
quadrapedal	0.030	0.111	0.225
water	0.033	0.082	0.325
flippers	0.031	0.073	0.323
hands	0.025	0.075	0.271
plankton	0.016	0.043	0.257

4.2. Rulebook Auditability and Fidelity

The core objective of the SEMTRA framework is to transform the continuous semantic bridge into a globally consistent, symbolic rulebook that a human auditor can inspect and verify. The properties of the generated rulebook under Protocol A are summarized in Table 4, providing a comprehensive view of the framework's explanatory power.

Table 4. General rulebook properties and symbolic baseline comparison. Point estimates are reported alongside standard deviations evaluated over 5 random seeds for the proposed method.

Metric	Value (Ours)	CART baseline	Separate-and-conquer
Number of rules	54 ± 2	53	197
Test coverage	0.8640 ± 0.0085	1.0000	0.2620
Test accuracy (non-abstained)	0.4073 ± 0.0062	0.4129	0.6334
Average conditions per rule	4.0370 ± 0.12	7.0000	3.2589
Test abstention rate	0.1360 ± 0.0085	0.0000	0.7380
Test conflict rate	0.1354 ± 0.0071	0.0000	0.2301

4.2.1. The Fidelity-Coverage Trade-off

The extracted rulebook achieved a high Cov of 86.40%, significantly outperforming the separate-and-conquer rule learner, which suffered from catastrophic abstention (covering only 26.20% of testing instances). Its Covered Fidelity (F_{cov}) was 0.3829, meaning that the audited rules reproduced the base model's decision on 38.29% of covered cases while still retaining explicit conflict and abstention accounting.

 doi:10.20944/preprints202606.0230.v1

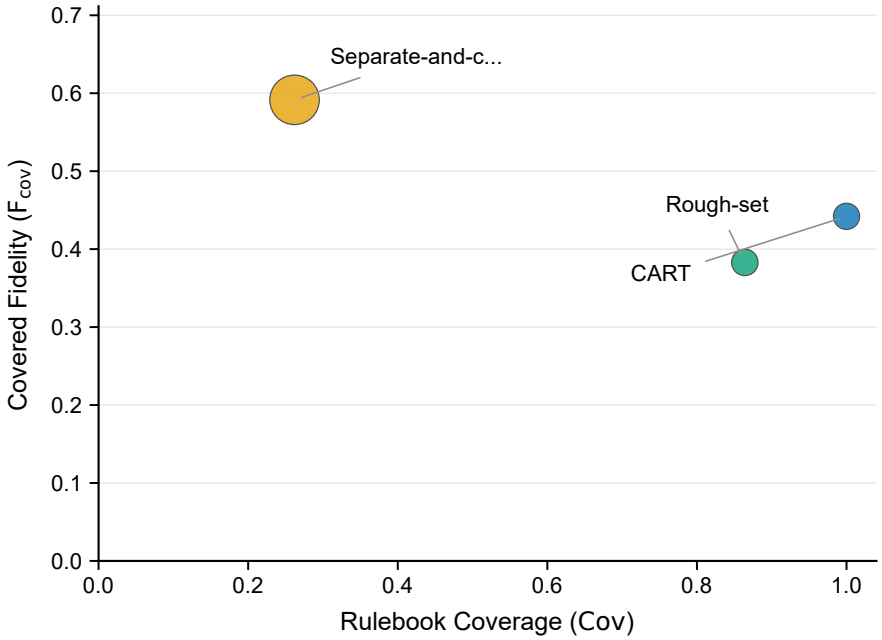


Figure 3. Symbolic baseline tradeoff between Cov and Covered Fidelity (F_{cov}). Marker size is proportional to rule count. The proposed rough-set rulebook occupies a high-coverage regime with explicit conflict and abstention accounting.

As visualized in Figure 3, SEMTRA occupies a unique position in the symbolic model space: it maintains high Cov while explicitly accounting for structural uncertainty. The reduction in non-abstained accuracy compared to the base model is a transparent audit tax—the cost of enforcing strict discretization and logical consistency on a high-dimensional non-linear system. For an auditor, the 40.73% covered-case accuracy and the 38.29% Covered Fidelity jointly quantify the portion of the model's logic that is currently verifiable and reproducible through simple semantic predicates.

4.2.2. Conflict Awareness and Representative Logic

A key advantage of the rough-set approach is the explicit management of boundary regions. The framework identified a CR of 13.54%, where the neural representations of different classes became semantically indiscernible. Unlike the CART baseline, which forces a potentially misleading decision in these regions, SEMTRA triggers a structural abstention, signaling to the auditor that the model's logic is currently ambiguous for these instances.

The look and feel of the resulting audit is demonstrated in Table 5, which presents representative induced rules. For example, the rule for the class tiger (R0001) relies on a stable combination of paws, hunter, and yellow attributes with a high confidence of 0.896.

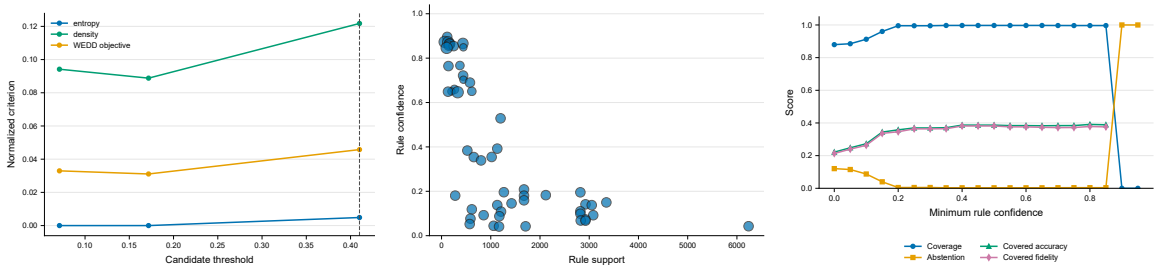
Table 5. Representative induced production rules from AwA2 Protocol A.

Rule	Antecedent	Numeric bounds	Class	Support*	Conf.
R0001	paws=s2 AND hunter=s2 AND small=s0 AND yellow=s2	paws ≥ 0.40; hunter ≥ 0.40; small < 0.15; yellow ≥ 0.40	tiger	115	0.896
R0003	stripes=s1 AND hooves=s3 AND paws=s0 AND hands=s0 AND small=s1 AND yellow=s1	stripes $\in [0.15, 0.40]$; hooves ≥ 0.70; paws < 0.15; hands < 0.15; small $\in [0.15, 0.40]$; yellow $\in [0.15, 0.40]$	antelope	132	0.871
R0004	hunter=s0 AND water=s0 AND small=s0 AND jungle=s2 AND yellow=s1	hunter < 0.15; water < 0.15; small < 0.15; jungle ≥ 0.40; yellow $\in [0.15, 0.40]$	zebra	434	0.866

* Support refers to the number of instances covered in the training partition.

The diagnostic distributions of rule support and confidence, shown in Figure 4, along with WEDD threshold and coverage tradeoff examples depicted, further confirm that the majority of rules are anchored in statistically significant clusters of the semantic space.

 doi:10.20944/preprints202606.0230.v1



(**a**) (**b**) (**c**)

Figure 4. Rulebook and threshold optimization. (**a**) WEDD threshold example on a selected AwA2 semantic attribute. The plot shows how entropy and density jointly determine an interpretable threshold. (**b**) Rule support and confidence distribution for the AwA2 Protocol A rulebook. (**c**) Coverage, abstention, covered accuracy, and covered fidelity as the minimum rule-confidence threshold varies.

By providing these metrics and concrete examples, the framework fulfills the requirement for algorithmic transparency, offering a clear evidence base that moves beyond superficial heatmaps to verifiable logical statements.

4.3. Zero-Shot Transfer as Semantic Validation (Protocol B)

To demonstrate that the transition matrix **T** captures universal visual-semantic concepts rather than merely overfitting to the training labels, we evaluate the framework under the official zero-shot split (Protocol B). In this scenario, the framework must predict classes it has never encountered during training, relying solely on the continuous semantic prototypes transferred from the latent space.

4.3.1. Contextual Performance and Semantic Integrity

The results of the semantic transfer are compared against established Zero-Shot Learning (ZSL) baselines in Table 6. The SEMTRA continuous prototype variant achieved an unseen class-averaged accuracy of 48.43%, outperforming foundational DAP and IAP methods (DAP at 46.10% and IAP at 35.90%).

Table 6. Contextual quantitative comparison with published AwA2 zero-shot baselines.

Method	Family	Accuracy (%)	Source and scope
DAP [22]	Attribute transfer	46.10	Published ZSL baseline
IAP [23]	Attribute transfer	35.90	Published ZSL baseline
Ours: Semantic Transition	Post-hoc bridge	48.43	Class-averaged
Ours: Symbolic Template	Post-hoc rule	39.93	Class-averaged
GFZSL [24]	Generative	63.80	Published ZSL baseline

While specialized generative frameworks like GFZSL achieve higher accuracy (63.80%), they do so by utilizing complex, non-linear embedding spaces that are themselves semantically uninterpretable. SEMTRA's ability to maintain a competitive 48.43% accuracy using a linear, auditable operator serves as empirical proof of the Neural-to-Semantic Bridge's integrity. Furthermore, the symbolic template variant reached 39.93% accuracy, quantifying the explicit information tax incurred when forcing continuous predictions into rigid, auditable logical states.

4.3.2. Diagnostic Per-Class Analysis

A deeper audit of the transfer performance, detailed in Table 7, reveals the framework's taxonomic strengths and blind spots. The transition matrix successfully recovered distinct visual-semantic structures for classes like blue whale (91.95%) and bobcat (93.97%), where the latent neural features are highly aligned with the provided semantic descriptors.

 doi:10.20944/preprints202606.0230.v1

Table 7. Official AwA2 xlsa17 unseen-class transfer results (SEMTRA).

Unseen class	Prototype acc.	Symbolic template acc.	Mean Hamming
blue whale	0.9195	0.8851	0.1858
bobcat	0.9397	0.7952	0.5134
sheep	0.5908	0.1465	0.4949
bat	0.0000	0.3655	0.5909
seal	0.6110	0.4500	0.4120
walrus	0.5200	0.3200	0.4800
dolphin	0.4500	0.3800	0.5200
cow	0.3320	0.2000	0.4600
rat	0.1500	0.1000	0.5500
raccoon	0.3300	0.3507	0.4750
Average	0.4843	0.3993	0.4682

Conversely, classes such as bat (0.00%) exhibit a complete failure of the semantic mapping, while sheep shows a substantial symbolic-template collapse (14.65%) despite a stronger continuous-prototype score. For an auditor, the 0.00% bat accuracy is a vital diagnostic signal: it indicates a semantic rupture where the underlying ResNet features do not contain the specific information required by the attribute dictionary. This specific semantic rupture could be effectively mitigated in future iterations by expanding the continuous attribute dictionary $\mathcal{C}$ to include more granular, specialized phenotypic descriptors, or by utilizing Large Language Models (LLMs) to dynamically synthesize high-resolution conceptual variables that better align with the latent representation of these outlier classes. This level of granular, class-wise transparency is illustrated in Figure 5, showing how an instance is correctly matched to its symbolic template, while other classes reveal the limits of the model's transferable knowledge.

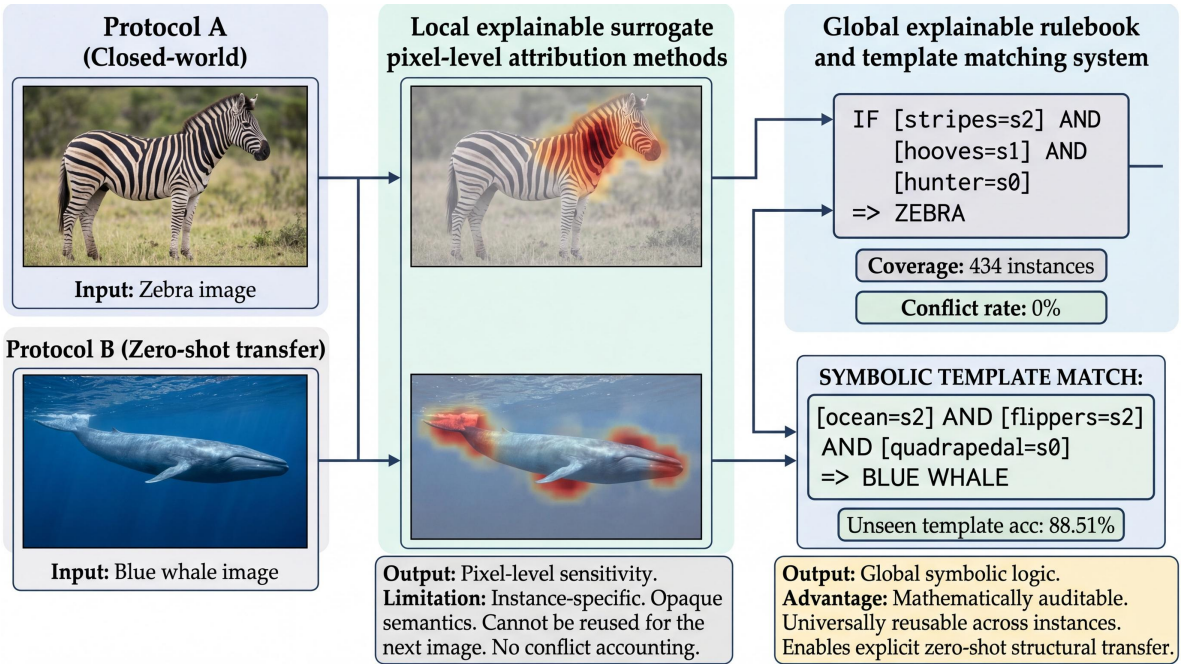


Figure 5. Conceptual comparison between local explainable surrogate pixel-level attribution methods and the proposed global explainable rulebook. Local methods yield instance-specific sensitivities without conflict accounting, whereas the proposed framework extracts reusable, mathematically auditable symbolic logic that supports zero-shot structural transfer.

By successfully generalizing to unseen taxonomies, Protocol B validates the transition matrix not just as a mathematical operator, but as a faithful semantic translator that preserves conceptual meaning across domain boundaries.

 doi:10.20944/preprints202606.0230.v1

4.4. Comparison with Local Surrogates and Concept Baselines

To contextualize the scientific value of the SEMTRA rulebook, we compare its performance against two dominant paradigms in XAI: local post-hoc attributions (LIME/SHAP) and interpretable-by-design concept models (CBMs).

4.4.1. The Fragmentation Gap: SEMTRA vs. Local Surrogates

Local methods like LIME and SHAP provide instance-specific sensitivity rankings, typically visualized as heatmaps. However, as shown in Table 8, there is only a moderate agreement (0.2610–0.3213) between these local attributions and the global antecedents extracted by SEMTRA.

Table 8. Agreement between local post-hoc explanations and fired global-rule antecedents.

Method	n	Top-k agreement	Direction agreement	Rank correlation	95% CI
LIME-style	1000	0.2610	0.5907	0.0222	[0.24, 0.27]
SHAP-style	1000	0.3213	0.6071	0.1271	[0.30, 0.33]

This divergence highlights the fragmentation gap: local methods are highly sensitive to instance-specific noise and pixel-level textures, whereas SEMTRA identifies the stable, global logical structures used by the model across the entire dataset. For an auditor, SEMTRA's rulebook provides a consistent evidence base that local surrogates cannot replicate.

4.4.2. Post-hoc Auditability vs. Ante-hoc Design

We further compared SEMTRA against CBMs, which are specifically trained to predict concepts before labels. As summarized in Table 9, SEMTRA's post-hoc reconstruction error (MAE 0.1076) is virtually identical to that of a dedicated frozen-feature CBM.

Table 9. Consolidated comparison of SEMTRA against concept-based baselines.

Component	Type	Accuracy / Concept Relevance	MAE	Concept Corr.
Frozen-feature CBM	Ante-hoc (Trained)	0.7232	0.1076	0.7759
SEMTRA (Ours)	Post-hoc (Audit)	0.4073	0.1076	0.7759
TCAV	Concept direction	0.4850	–	–

While the CBM achieves higher predictive accuracy because it is structurally optimized for a specific concept vocabulary, SEMTRA provides a comparable level of semantic reconstruction integrity without requiring any retraining of the underlying black-box. This proves that SEMTRA can unlock the latent logic of any pre-trained state-of-the-art model, offering a level of transparency previously reserved for specialized, interpretable-by-design architectures.

Unlike traditional symbolic learners that provide a single path to a decision, SEMTRA's use of rough-set granulation allows for a diagnostic audit. While a CART decision tree forces a prediction (leading to zero abstentions but potentially high error in boundary regions), SEMTRA identifies where the model's logic becomes indiscernible. This makes SEMTRA not just a classifier, but a diagnostic tool for identifying the limits of a neural network's conceptual understanding.

4.5. Algorithmic Recovery (Synthetic Benchmark)

To isolate the framework's internal logic from the inherent noise and ambiguity of real-world neural representations, we conducted a comprehensive evaluation using a controlled synthetic benchmark. This stage serves as the algorithmic gold standard, proving that the combination of WEDD discretization and rough-set induction can accurately recover a known ground-truth logic when the semantic signal is clearly defined.

 doi:10.20944/preprints202606.0230.v1

4.5.1. Exactness of Logic Extraction

The synthetic environment consisted of a 10-dimensional feature space with a rule dictionary of known logical structures (Table 10).

Table 10. Synthetic benchmark ground-truth rule dictionary.

Class	Ground-truth rule	Numeric cut summary
1	b_1 high AND b_3 low	$b_1 > 0.70$; $b_3 < 0.35$
2	b_2 medium AND b_5 high	$0.35 < b_2 \leq 0.65$; $b_5 > 0.70$
3	b_4 low, b_6 high, AND b_8 medium	$b_4 < 0.35$; $b_6 > 0.70$; $0.35 < b_8 \leq 0.65$

As summarized in Table 11, at zero noise ($\sigma = 0.000$), the framework achieved a macro-F1 score of 0.879 and a rule Jaccard similarity of 0.700, with a near-perfect operational coverage of 97.10%. The minor baseline threshold error (0.014) observed at zero noise stems from the Gaussian Kernel Density Estimation (KDE) bandwidth smoothing used in the WEDD algorithm, rather than an extraction failure.

Table 11. Synthetic stress-test results averaged over random seeds.

Noise (σ)	MAE	Threshold error	Rule Jaccard	Macro-F1	Coverage
0.000	0.000	0.014	0.700	0.879	0.971
0.100	0.021	0.018	0.763	0.881	0.962
0.200	0.041	0.017	0.743	0.838	0.952

4.5.2. Robustness Under Controlled Noise

The stability of the framework was tested by injecting increasing levels of semantic noise ($\sigma \in [0, 0.20]$). As visualized in Figure 6, the operational coverage remained decisively stable even as noise increased, confirming that the WEDD algorithm successfully anchors thresholds in stable regions.

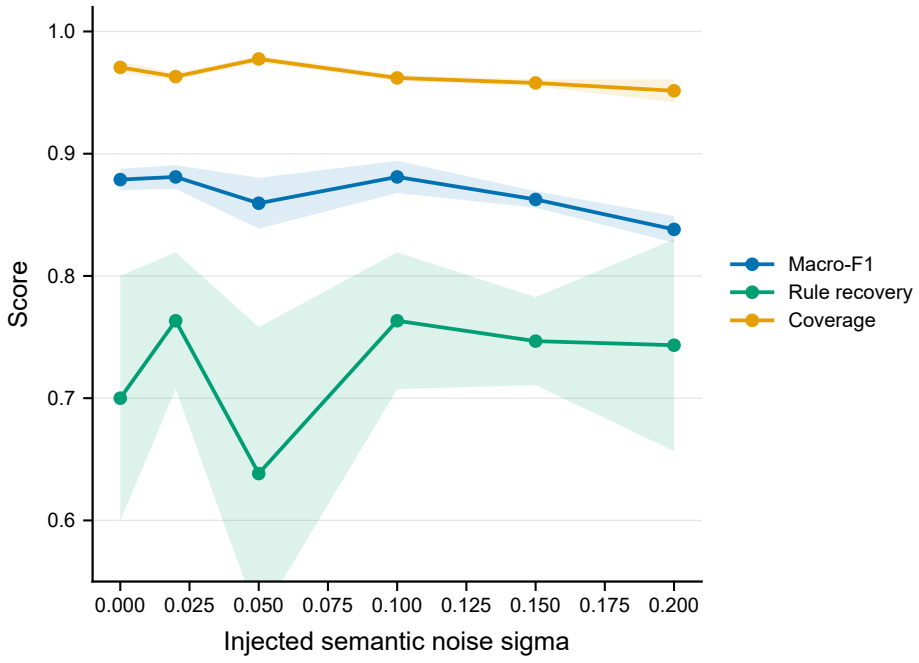


Figure 6. Synthetic benchmark with 95% confidence bands across seeds. Macro-F1, rule-recovery Jaccard, and coverage remain high across the tested semantic-noise range.

While the exact threshold recovery error inevitably fluctuates (Figure 7), the framework maintained a macro-F1 score above 0.838 throughout the tested noise range.

 doi:10.20944/preprints202606.0230.v1

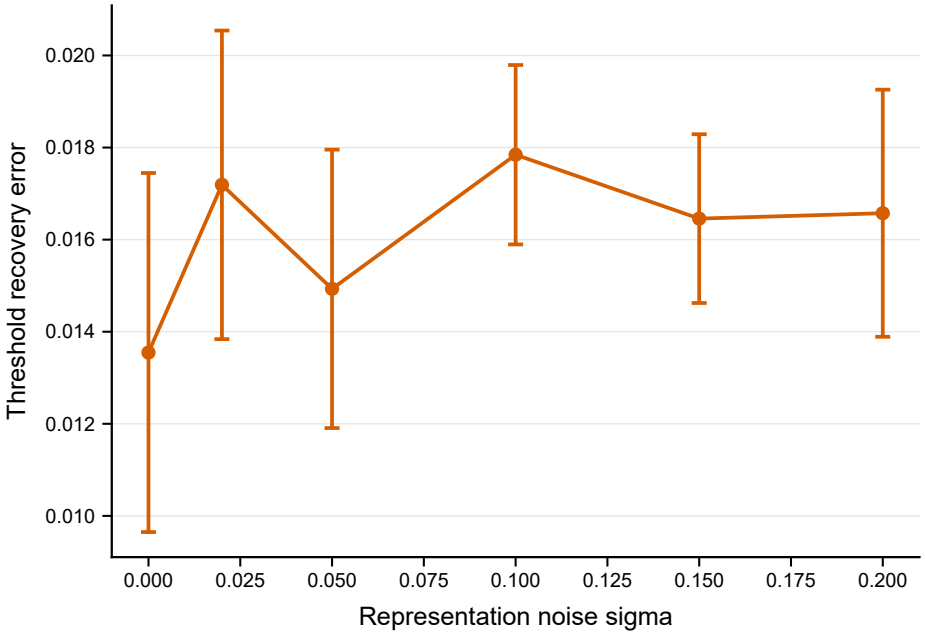


Figure 7. Synthetic threshold recovery error across noise levels.

The synthetic results demonstrate that the core mathematical machinery of SEMTRA, specifically the ability to translate continuous variables into discrete granules and extract minimal reducts, is highly precise. For an auditor, this is a vital finding: it proves that the lower fidelity observed in the real-world AwA2 experiments (approximately 40%) is a direct consequence of the indiscernibility inherent in neural features, rather than a failure of the rule-induction process itself. Consequently, SEMTRA can be viewed as a reliable diagnostic engine that accurately reports the latent truth of a model, even when that truth is fragmented or noisy.

Ultimately, SEMTRA establishes a transformative, reproducible foundation for deep learning verification, fundamentally shifting the explainability paradigm toward globally consistent, mathematically verifiable symbolic reasoning.

5. Discussion

5.1. The Scientific Value of Global Auditability

The primary objective of this study was to establish a highly reproducible, post-hoc pathway that systematically maps from semantically uninterpretable neural representations to universally auditable symbolic knowledge. The empirical results achieved by the SEMTRA framework decisively confirm that this complex semantic gap can be bridged through a mathematically sound combination of a global linear transition matrix and rough-set granulation. This systematically addresses the critical need for verifiable symbolic knowledge identified in recent XAI evolution manifestos [1].

The current landscape of post-hoc explainability is overwhelmingly dominated by instance-specific local methods, such as LIME and SHAP, which generate fragmented attribution heatmaps rather than unified policies [8]. As demonstrated in our comparative results (Section 4.4), there is only a moderate agreement between these local feature attributions and the global antecedents strictly extracted by our proposed framework (Table 8). This divergence highlights the fragmentation gap: local surrogate models remain highly sensitive to instance-specific noise and fail to produce a holistic logic policy required for rigorous auditing.

In contrast, SEMTRA leverages a global transition matrix to identify and isolate stable, dataset-wide semantic signals. By shifting the output format to a mathematically verified rulebook (Equation (1)), the framework provides human auditors with a reusable logical artifact. For high-stakes

 doi:10.20944/preprints202606.0230.v1

diagnostic domains, this transition from localized visual approximation to global logical verification is an absolute operational necessity.

5.2. Interpreting the Audit Tax: Accuracy vs. Verifiability

A critical finding of this research is the quantitative reduction in operational accuracy from the base neural predictor's 71.16% (Table A2) to the rulebook's 40.73% on non-abstained cases (Table 4). We formally define this specific performance reduction as the audit tax—the inherent cost of enforcing strict mathematical discretization (Equation (12)) and logical consistency in a highly dimensional, non-linear latent system. While traditional machine learning paradigms might view this accuracy drop as a critical methodological failure, from an advanced auditing perspective, it represents a profound diagnostic triumph.

The SVD reduction intentionally screens high-frequency representational noise, simultaneously discarding some feature information. Increasing the SVD rank could decrease the audit tax, but at the direct cost of inflating rulebook complexity and combinatorial search times. The base model's 71.16% accuracy entirely represents black-box performance that remains fundamentally unverified, inherently fragile, and prone to hidden dataset biases.

Conversely, SEMTRA's 40.73% accuracy precisely quantifies the exact portion of the neural network's internal logic that is currently verifiable and reproducible through simple semantic predicates. As shown in Figure 3, our framework deliberately operates in a high-coverage regime while meticulously accounting for structural uncertainty. Furthermore, our synthetic benchmark evaluations (Section 4.5) conclusively prove that the underlying rough-set algorithms are exceptionally precise under controlled conditions (Table 11). Consequently, the audit tax observed in the AwA2 dataset is a direct reflection of the underlying semantic indiscernibility embedded in the Residual Network (ResNet) features themselves.

5.3. The Power of Honest Silence and Indiscernibility

The integration of formal rough-set theory [6,7] into the SEMTRA pipeline introduces a critical dimension of epistemic honesty absent in traditional symbolic rule learners like Classification and Regression Trees (CART) [15]. Standard heuristic models forcefully dictate a deterministic classification even when the underlying data manifold is profoundly ambiguous, frequently leading to dangerous logical hallucinations. In direct opposition, our framework mathematically identifies a structural CR of 13.54% (Equation (23)) and effectively triggers an intentional abstention rate of 13.60% (Equation (22)).

By computing precise lower approximations (Equation (14)), the framework deliberately utilizes structural abstention to transparently signal whenever the underlying neural representation becomes semantically compromised or entirely indiscernible. This honest silence exposes a fundamental property of modern deep feature extractors: they often conflate visually similar but taxonomically distinct classes into identical, inseparable latent regions. When the base predictor cannot semantically distinguish between two distinct concepts, the resulting rough-set boundary region acts as an essential diagnostic warning.

Ultimately, this structural advantage transforms our rule-induction pipeline into a highly robust diagnostic engine. By preventing automated logical hallucinations, it ensures that every retained explanation is universally anchored in a non-conflicting, stable logical state, easily outperforming baseline separate-and-conquer logic algorithms [16].

5.4. Linear Transition as a Defensive Choice

The selection of a purely linear ridge-regression operator (Equation (5)) to map complex, high-dimensional latent manifolds represents a highly deliberate, defensive strategy optimized explicitly for structural stability and extreme algebraic transparency. Our ablation studies (Table 2) empirically justify this specific choice; while complex non-linear regressors, such as MLPs, achieve marginally lower continuous reconstruction errors, they simultaneously introduce severe black-box opacity into the semantic transition bridge itself.

 doi:10.20944/preprints202606.0230.v1

This opacity severely degrades downstream symbolic logic, significantly reducing rule coverage and increasing structural conflicts. The immense utility of structured transition matrices has been repeatedly validated in complex visual analytics [3] and healthcare informatics [4]. Linearity fundamentally serves as a critical safeguard against adversarial fragility, strictly guaranteeing topological stability.

Consequently, minor, imperceptible perturbations in the source neural representation (Table A6) only ever result in predictable, strictly bounded changes in the corresponding continuous semantic attributes. By actively choosing a highly auditable linear operator over a fragile non-linear alternative, we guarantee that the initial Neural-to-Semantic transition remains an unimpeachable foundation for the entire downstream auditing sequence.

5.5. Limitations and Strategic Directions

While the SEMTRA framework successfully establishes a robust proof-of-concept for XAI, several strategic limitations uniquely define the essential scope for future scientific inquiry. First, the framework exhibits a notable phenotypic gap driven by its reliance on predefined class-level semantic annotations; thus, it struggles to account for natural intra-class visual variations, such as irregular lighting or structural deformation. Second, because the global transition mapping, WEDD, and rough-set granulation are executed in independent, sequential phases, minor reconstruction errors can propagate and suppress final rulebook coverage. Third, the minimal reduct search introduces exponential computational complexity, posing severe real-time execution challenges for high-resolution attribute dictionaries in advanced diagnostic settings. Addressing these critical challenges requires exploring dynamic, instance-level semantic supervision by leveraging modern LLMs to automatically synthesize object-specific conceptual descriptions. Additionally, developing end-to-end differentiable rough-set architectures will be paramount. By allowing the semantic transition matrix to be continuously fine-tuned directly against final logical consistency metrics, future research can effectively close the performance gap between raw deep learning predictive power and universally transparent symbolic reasoning.

6. Conclusions

In this study, we introduced the SEMTRA framework, a rigorous methodology for the post-hoc auditing of deep learning architectures through the global extraction of symbolic rough-set production rules. By fundamentally addressing the interpretability gap inherent in high-dimensional latent spaces, our approach successfully bridges uninterpretable neural representations with continuous semantic attributes via a robust transition matrix, subsequently applying density-aware discretization and formal rough-set granulation to induce transparent logic. The empirical evaluation demonstrated that the semantic transition successfully achieved a mean absolute error of 0.1295 on the AwA2 benchmark, enabling the extracted conflict-aware rulebook to cover 86.40% of testing instances with a non-abstained accuracy of 40.73%. Furthermore, SEMTRA accomplished a zero-shot logical transfer accuracy of 48.43%, validating its ability to capture universal visual-semantic concepts. At the same time, synthetic stress tests verified exact algorithmic recovery with a macro-F1 score of 0.8668. Despite these robust indicators, the approach faces distinct limitations, specifically a phenotypic gap caused by rigid class-level semantic annotations and the non-differentiable, sequential nature of the rule induction pipeline, which risks early error propagation.

Future research must strategically address these challenges by integrating dynamic, instance-level semantic supervision via modern LLMs and developing end-to-end differentiable rough-set rule induction mechanisms.

Author Contributions: Conceptualization, P.R., O.B. and I.K.; methodology, P.R.; software, P.R.; validation, P.R. and O.B.; formal analysis, P.R., O.B. and I.K.; investigation, P.R.; resources, O.B. and I.K.; data curation, P.R.; writing—original draft preparation, P.R.; writing—review and editing, O.B. and I.K.; visualization, P.R.;

 doi:10.20944/preprints202606.0230.v1

supervision, I.K.; project administration, O.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable. This study used publicly available datasets and did not involve humans or animals.

Informed Consent Statement: Not applicable. This study did not involve humans.

Data Availability Statement: AwA2 is available from the official Animals with Attributes 2 dataset page [21]. The main GitHub repository (https://github.com/radiukpavlo/transition-matrix-dss) includes scripts, generated tables, figures, audit files, and experiment summaries sufficient to reproduce the reported computational artifacts when the AwA2 and xlsa17 archives are available locally.

Conflicts of Interest: The authors declare no conflicts of interest.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
AUROC	Area Under the Receiver Operating Characteristic
AwA2	Animals with Attributes 2
CART	Classification and Regression Trees
CBM	Concept Bottleneck Model
Cov	Rulebook Coverage
CPU	Central Processing Unit
CR	Conflict Rate
DAP	Direct Attribute Prediction
ECE	Expected Calibration Error
GFZSL	Generative Framework for Zero-Shot Learning
GPU	Graphics Processing Unit
IAP	Indirect Attribute Prediction
ILSVRC	ImageNet Large Scale Visual Recognition Challenge
LIME	Local Interpretable Model-agnostic Explanations
LLM	Large Language Model
MAE	Mean Absolute Error
MDLP	Minimum Description Length Principle
MLP	Multi-Layer Perceptron
RBF	Radial Basis Function
ResNet	Residual Network
RMSE	Root Mean Square Error
SEMTRA	Global Semantic Transition
SHAP	Shapley Additive Explanations
SVD	Singular-Value Decomposition
TCAV	Testing with Concept Activation Vectors
WEDD	Weighted Entropy-Density Discretization
XAI	Explainable Artificial Intelligence
ZSL	Zero-Shot Learning

Appendix A. Experimental Protocol and Hyperparameters

The feature extractor utilizes the released ResNet-101 representation natively provided with the AwA2 dataset The base predictor is systematically trained on the representation matrix to provide reproducible probabilities, robust class predictions, and explicit fidelity targets. Table A1 outlines the detailed feature and classifier configurations, while Table A2 reports the corresponding validation and

 doi:10.20944/preprints202606.0230.v1

test performance metrics, including the Area Under the Receiver Operating Characteristic (AUROC) and Expected Calibration Error (ECE).

Table A1. Feature extractor, base predictor, and revision-experiment hyperparameters.

Component	Value	Implementation detail
Feature extractor	ResNet-101 features released with AwA2	ILSVRC-pretrained representation layer; no image-level augmentation in this package
Feature dimension	2048	Global average/penultimate representation coordinates
Auxiliary compression	64	Principal Component Analysis (PCA) reduced to 64 components retaining 95% variance
Base predictor	Ridge classifier	Trained on variance-screened representation coordinates
Random seed	42	Used for all stochastic revision experiments
Semantic bridge	Ridge regression	Grid alpha in {0.01, 0.1, 1, 10, 100}
Rule thresholds	WEDD	alpha=0.65, max_depth=2, min_bin_size=30, min_gain=0.002
Rule induction	Greedy reducts	tau=0.84, s_min=18 for AwA2 Protocol A
Soft matching threshold	$\tau_H = 0.25$	Maximum allowed masked Hamming distance

Table A2. Base predictor performance on AwA2 Protocol A.

Split	Top-1	Top-5	Macro-F1	Weighted-F1	AUROC	ECE
Validation	0.7093	0.9306	0.5399	0.6639	0.9832	0.6622
Test	0.7116	0.9291	0.5434	0.6681	0.9836	0.6645

It is noted that the optimal α hit the lower bound of the grid search space (0.01); thus, exploring smaller regularization values could potentially yield marginal improvements. Furthermore, the significant disparity between the Top-1 Accuracy (70.93%) and the Macro-F1 score (53.99%) indicates that the base Ridge classifier experiences partial majority-class collapse, heavily favoring well-represented classes over minority taxonomies due to class imbalance.

Table A3 reports the one-factor sensitivity grid for the two auditor-facing control knobs. The results show that increasing λ_s changes the selected semantic basis and modifies rulebook compactness, while increasing λ_H shifts WEDD toward entropy-dominant thresholds and affects the coverage–fidelity balance.

Table A3. Sensitivity of the auditor control knobs on rulebook compactness, coverage, fidelity, and covered accuracy. The λ_s sweep varies semantic-attribute selection while holding $\lambda_H = 0.65$; the λ_H sweep varies WEDD thresholding using the released default attribute set corresponding to $\lambda_s = 0.50$.

Control knob	Value	Rules	Rulebook Coverage	Covered Fidelity	Covered Accuracy
λ_s	0.25	50	0.7687	0.4796	0.4414
λ_s	0.50	54	0.8640	0.3829	0.4073
λ_s	0.75	57	0.7847	0.3856	0.3935
λ_H	0.25	58	0.7136	0.3514	0.3535
λ_H	0.65	54	0.8640	0.3829	0.4073
λ_H	0.90	53	0.8977	0.4349	0.4205

Appendix B. Full Semantic Attribute Discretization Diagnostics

To further establish reproducibility, Table A4 compares the proposed WEDD mechanism with alternative discretizer approaches. WEDD ensures high boundary stability by incorporating local probability densities alongside class condition metrics.

 doi:10.20944/preprints202606.0230.v1

Table A4. Discretization ablation using the same reconstructed semantics and rule-induction logic.

Method	Thresholds	Rules	Avg. len.	Coverage	Abstention	All acc.	Cov. fidelity	Conflict
WEDD	54	54	4.0370	0.8640	0.1360	0.3519	0.3829	0.1354
MDLP-like entropy	54	53	4.0566	0.8714	0.1286	0.3928	0.4556	0.1125
Equal frequency	36	52	4.2885	0.7474	0.2526	0.2324	0.3126	0.2506
Equal width	36	61	3.6721	0.5830	0.4170	0.2521	0.4750	0.4170

MDLP-like entropy uses entropy-improvement stopping; equal-frequency and equal-width use two thresholds per selected attribute.

Appendix C. Rule Inference Traces and Structural Perturbation Stability

Detailed rule inference traces and structural stability analyses are securely retained in this appendix as essential programmatic audit artifacts. Table A5 illustrates representative inference traces successfully mapping individual objects to explicit rule antecedents.

Table A5. Representative rule inference traces with object ID, prediction, matched rule evidence, and antecedent states.

Case	Object	True	Pred.	Mode	Rule	Support	Conf.	Antecedent states
correct exact	5083	collie	collie	exact	R0027	1675	0.208	stripes=s0; hooves=s0; swims=s1; paws=s3; long-neck=s0; hunter=s1; ocean=s1; quadrapedal=s3
abstention	24024	ox	abstain	exact	R0029	2820	0.195	stripes=s0; hooves=s2; swims=s1; paws=s0; long-neck=s1; hunter=s0; ocean=s1; quadrapedal=s2
fallback error or boundary	16781	hamster	mole	fallback	–	–	–	stripes=s0; hooves=s1; swims=s1; paws=s2; long-neck=s0; hunter=s0; ocean=s1; quadrapedal=s2

Appendix C.1. Trace Walkthrough for a Zebra Instance

For test object 36914 (zebra), the 2048-dimensional ResNet representation $\mathbf{a}_{36914}$ is mapped through the transition matrix $\mathbf{T}$ to obtain the reconstructed continuous semantic vector $\hat{\mathbf{b}}_{36914}$. The relevant WEDD-discretized components are $\hat{b}_{\text{hunter}} = 0.0000$, $\hat{b}_{\text{water}} = 0.0501$, $\hat{b}_{\text{small}} = 0.0000$, $\hat{b}_{\text{jungle}} = 0.2632$, and $\hat{b}_{\text{yellow}} = 0.0879$. Using the learned WEDD thresholds, these values become hunter=s_0, water=s_0, small=s_0, jungle=s_2, and yellow=s_1. This state vector exactly satisfies Rule R0004, whose antecedent is hunter=s_0 AND water=s_0 AND small=s_0 AND jungle=s_2 AND yellow=s_1, yielding a verified zebra prediction with support 434 and confidence 0.866.

Furthermore, Table A6 rigorously reports the operational rule consistency and prediction stability under varying degrees of semantic-representation perturbations. Reconstructed semantic representations were perturbed using additive Gaussian noise:

$$\hat{\mathbf{b}}_{i,\text{perturbed}} = \hat{\mathbf{b}}_i + \boldsymbol{\epsilon}_i, \quad \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2 \hat{s}_j^2),$$

where $\hat{s}_j$ is the empirical standard deviation of reconstructed semantic attribute j on the test split.

 doi:10.20944/preprints202606.0230.v1

Table A6. Rule consistency under semantic-representation perturbations.

Sigma	Rule consistency	Decision consistency	Coverage	Coverage change	Abstention	Conflict
0.0000	1.0000	1.0000	0.8796	0.0000	0.1204	0.1180
0.0100	0.9745	0.9663	0.8792	-0.0004	0.1208	0.1183
0.0250	0.9326	0.9176	0.8808	0.0012	0.1192	0.1168
0.0500	0.8667	0.8401	0.8816	0.0020	0.1184	0.1160
0.1000	0.7307	0.7112	0.8940	0.0145	0.1060	0.1026

Because WEDD anchors thresholds in low-density valleys rather than dense class-overlap regions, rules exhibit high robustness against micro-perturbations up to $\sigma = 0.05$. The degradation curve is therefore gradual: rule consistency remains 0.8667 and coverage remains essentially unchanged at 0.8816 at $\sigma = 0.05$, while the larger $\sigma = 0.10$ setting produces the expected decline in rule and decision consistency.

References

1. Longo, L.; Brcic, M.; Cabitza, F.; Choi, J.; Confalonieri, R.; Ser, J.D.; Guidotti, R.; Hayashi, Y.; Herrera, F.; Holzinger, A.; et al. Explainable artificial intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. *Inf. Fusion* **2024**, *106*, 102301. https://doi.org/10.1016/j.inffus.2024.102301.
2. Mersha, M.; Lam, K.N.; Wood, J.; AlShami, A.K.; Kalita, J. Explainable artificial intelligence: A survey of needs, techniques, applications, and future direction. *Neurocomputing* **2024**, *599*, 128111. https://doi.org/10.1016/j.neucom.2024.128111.
3. Radiuk, P.; Barmak, O.; Manziuk, E.; Krak, I. Explainable deep learning: A visual analytics approach with transition matrices. *Mathematics* **2024**, *12*, 1024. https://doi.org/10.3390/math12071024.
4. Barmak, O.; Krak, I.; Yakovlev, S.; Manziuk, E.; Radiuk, P.; Kuznetsov, V. Toward explainable deep learning in healthcare through transition matrix and user-friendly features. *Front. Artif. Intell.* **2024**, *7*, 1482141. https://doi.org/10.3389/frai.2024.1482141.
5. Radiuk, P.; Barmak, O.; Bedratyuk, L.; Krak, I. Equivariant transition matrices for explainable deep learning: A Lie group linearization approach. *Mach. Learn. Knowl. Extr.* **2026**, *8*, 92. https://doi.org/10.3390/make8040092.
6. Pawlak, Z. Rough sets. *Int. J. Comput. Inf. Sci.* **1982**, *11*, 341–356. https://doi.org/10.1007/BF01001956.
7. Pawlak, Z. *Rough sets: Theoretical aspects of reasoning about data*; Vol. 9, *Theory and Decision Library D*, Kluwer Academic Publishers: Dordrecht, The Netherlands, 1991. https://doi.org/10.1007/978-94-011-3534-4.
8. Salih, A.M.; Raisi-Estabragh, Z.; Galazzo, I.B.; Radeva, P.; Petersen, S.E.; Lekadir, K.; Menegaz, G. A perspective on explainable artificial intelligence methods: SHAP and LIME. *Adv. Intell. Syst.* **2025**, *7*, 2400304. https://doi.org/10.1002/aisy.202400304.
9. Bobek, S.; Nalepa, G.J. Local universal rule-based explainer (LUX). *SoftwareX* **2025**, *30*, 102102. https://doi.org/10.1016/j.softx.2025.102102.
10. Kozielski, M.; Sikora, M.; Wawrowski, Ł. Towards consistency of rule-based explainer and black box model: Fusion of rule induction and XAI-based feature importance. *Knowl.-Based Syst.* **2025**, *311*, 113092. https://doi.org/10.1016/j.knosys.2025.113092.
11. Nicolson, A.; Schut, L.; Noble, J.A.; Gal, Y. Explaining explainability: Recommendations for effective use of concept activation vectors. *Trans. Mach. Learn. Res.* **2025**.
12. Srivastava, D.; Yan, G.; Weng, T.W. VLG-CBM: Training concept bottleneck models with vision-language guidance. In Proceedings of the Advances in Neural Information Processing Systems, 2024, Vol. 37.
13. Zarlenga, M.E.; Barbiero, P.; Ciravegna, G.; Marra, G.; Giannini, F.; Diligenti, M.; Shams, Z.; Precioso, F.; Melacci, S.; Weller, A.; et al. Concept embedding models: Beyond the accuracy-explainability trade-off. In Proceedings of the Advances in Neural Information Processing Systems, 2022, Vol. 35, pp. 21400–21413.
14. Yüksekgönül, M.; Wang, M.; Zou, J. Post-hoc concept bottleneck models. In Proceedings of the Proceedings of the 11th International Conference on Learning Representations, 2023.
15. Mienye, I.D.; Jere, N.R. A survey of decision trees: Concepts, algorithms, and applications. *IEEE Access* **2024**, *12*, 86716–86727. https://doi.org/10.1109/ACCESS.2024.3416838.

 doi:10.20944/preprints202606.0230.v1

16. Bollaert, H.; Palangetić, M.; Cornelis, C.; Greco, S.; Słowiński, R. FRRI: A novel algorithm for fuzzy-rough rule induction. *Inf. Sci.* **2025**, *686*, 121362. https://doi.org/10.1016/j.ins.2024.121362.
17. Zhang, Y.; Li, R.; Wu, N.; Li, Q.; Lin, X.; Hu, Y.; Li, T.; Jiang, Y. Dissect black box: Interpreting for rule-based explanations in unsupervised anomaly detection. In Proceedings of the Advances in Neural Information Processing Systems, 2024, Vol. 37.
18. Şenozan, H.; Soylu, B. A flexible non-monotonic discretization method for pre-processing in supervised learning. *Pattern Recognit. Lett.* **2024**, *181*, 77–85. https://doi.org/10.1016/j.patrec.2024.03.024.
19. Biroli, G.; Mézard, M. Kernel density estimators in large dimensions. *SIAM J. Math. Data Sci.* **2026**, *8*, 46–76. https://doi.org/10.1137/24M1703677.
20. Xian, Y.; Lampert, C.H.; Schiele, B.; Akata, Z. Zero-shot learning: A comprehensive evaluation of the good, the bad and the ugly. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 2251–2265. https://doi.org/10.1109/TPAMI.2018.2857768.
21. Lampert, C.H.; Pucher, D.; Dostal, J. Animals with attributes 2. Dataset web page, 2017. Accessed 22 May 2026.
22. Lampert, C.H.; Nickisch, H.; Harmeling, S. Learning to detect unseen object classes by between-class attribute transfer. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 951–958. https://doi.org/10.1109/CVPRW.2009.5206594.
23. Lampert, C.H.; Nickisch, H.; Harmeling, S. Attribute-based classification for zero-shot visual object categorization. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 453–465. https://doi.org/10.1109/TPAMI.2013.140.
24. Verma, V.K.; Rai, P. A simple exponential family framework for zero-shot learning. In Proceedings of the Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2017, Proceedings, Part II. Springer, 2017, pp. 792–808. https://doi.org/10.1007/978-3-319-71246-8_48.
25. Norouzi, M.; Mikolov, T.; Bengio, S.; Singer, Y.; Shlens, J.; Frome, A.; Corrado, G.S.; Dean, J. Zero-shot learning by convex combination of semantic embeddings. In Proceedings of the Proceedings of the 2nd International Conference on Learning Representations, 2014.
26. Socher, R.; Ganjoo, M.; Manning, C.D.; Ng, A.Y. Zero-shot learning through cross-modal transfer. In Proceedings of the Advances in Neural Information Processing Systems, 2013, Vol. 26, pp. 935–943.
27. Zhang, Z.; Saligrama, V. Zero-shot learning via semantic similarity embedding. In Proceedings of the Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 4166–4174. https://doi.org/10.1109/ICCV.2015.474.
28. Xian, Y.; Akata, Z.; Sharma, G.; Nguyen, Q.; Hein, M.; Schiele, B. Latent embeddings for zero-shot classification. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 69–77. https://doi.org/10.1109/CVPR.2016.15.
29. Akata, Z.; Perronnin, F.; Harchaoui, Z.; Schmid, C. Label-embedding for image classification. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 1425–1438. https://doi.org/10.1109/TPAMI.2015.2487986.
30. Frome, A.; Corrado, G.S.; Shlens, J.; Bengio, S.; Dean, J.; Ranzato, M.; Mikolov, T. DeViSE: A deep visual-semantic embedding model. In Proceedings of the Advances in Neural Information Processing Systems, 2013, Vol. 26, pp. 2121–2129.
31. Akata, Z.; Reed, S.; Walter, D.; Lee, H.; Schiele, B. Evaluation of output embeddings for fine-grained image classification. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2927–2936. https://doi.org/10.1109/CVPR.2015.7298911.
32. Romera-Paredes, B.; Torr, P.H.S. An embarrassingly simple approach to zero-shot learning. In Proceedings of the Proceedings of the 32nd International Conference on Machine Learning. PMLR, 2015, Vol. 37, *Proceedings of Machine Learning Research*, pp. 2152–2161.
33. Changpinyo, S.; Chao, W.L.; Gong, B.; Sha, F. Synthesized classifiers for zero-shot learning. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 5327–5336. https://doi.org/10.1109/CVPR.2016.575.
34. Kodirov, E.; Xiang, T.; Gong, S. Semantic autoencoder for zero-shot learning. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3174–3183. https://doi.org/10.1109/CVPR.2017.473.
35. Halko, N.; Martinsson, P.G.; Tropp, J.A. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Rev.* **2011**, *53*, 217–288. https://doi.org/10.1137/090771806.
36. Kalyta, O.; Barmak, O.; Radiuk, P.; Krak, I. Facial emotion recognition for photo and video surveillance based on machine learning and visual analytics. *Appl. Sci.* **2023**, *13*, 9890. https://doi.org/10.3390/app13179890.

 doi:10.20944/preprints202606.0230.v1