

GRAPH-BASED CONCEPT FORMATION FOR EXPLAINABLE ARTIFICIAL INTELLIGENCE: STRUCTURAL REDUCTION APPROACH

Lapin M., Bokhan K., Parzhyn Y., Perevoznyk K.

Summary

Subject matter: This article presents a novel approach to explainable artificial intelligence based on graph-based representation of contour images, where structure itself serves as the carrier of explanations. The approach addresses the fundamental limitation of modern AI systems that operate as opaque pipelines, making it difficult to understand decision-making processes and validate reasoning paths.

Goal: The goal is to develop and validate a graph-based representation method that enables transparent, interpretable AI through explicit structural encoding, where visual patterns are represented as attributed graphs and multiple instances are reduced to stable concept attractors suitable for classification.

Tasks: The research addresses the following tasks: (1) define a minimal vocabulary of node and edge types with semantic attributes for graph-based contour representation; (2) develop normalization procedures ensuring scale and translation invariance; (3) formulate structural reduction rules that iteratively merge training samples into generalized concepts; (4) implement parametric generalization strategies that preserve shared topology while abstracting instance-specific variation; (5) validate the approach on handwritten digit recognition with few-shot learning capability.

Methods: The methodology employs bipartite graph representation where Point nodes (endpoints, corners, junctions) and Line nodes (segments) alternate with semantic attributes and geometric properties. Concept formation proceeds through iterative structural reduction using custom operations: endpoint removal, intersection point merging, and path pruning. Parametric generalization transforms numeric attributes into range representations, preserves consistent categorical properties, and retains only universal list elements. Classification uses Graph Edit Distance with custom cost functions incorporating semantic and geometric constraints.

Results: Validation on MNIST handwritten digits demonstrates that concepts formed from 5-6 training samples achieve 82.35% classification accuracy through graph matching. The approach requires no gradient-based optimization or deep feature extraction, relying solely on explicit structural comparison. Concept graphs remain compact (3-12 nodes) with interpretable topological signatures: simple linear structures yield minimal graphs (mean degree 1.33), while closed-loop digits exhibit higher connectivity (mean degree 2.00) with intersection points. Confusion patterns trace directly to structural similarity, confirming that decision boundaries are queryable and verifiable.

Conclusions: The graph-based structural reduction approach advances explainable AI by making semantics explicit in graph topology rather than opaque weight matrices. Few-shot learning capability from minimal training data (2-6 samples per concept) demonstrates the viability of concept formation through direct structural composition without episodic meta-learning. By encoding decision paths as navigable graph elements, the approach provides inherent interpretability suitable for domains requiring traceable reasoning and trustworthy AI systems.

Keywords: explainable artificial intelligence; graph-based representation; structural reduction; few-shot learning; concept formation; graph edit distance; neural networks; pattern recognition

1 Introduction

Modern artificial intelligence systems, particularly deep neural networks, achieve impressive accuracy across diverse perception tasks yet fundamentally operate as opaque computational pipelines. This opacity creates significant challenges for practitioners who need to understand why a particular decision was reached, where evidence resides in the input, or how to diagnose failures when they occur. Post-hoc explanation techniques such as LIME and SHAP attempt to address these limitations by reverse-engineering model behavior after training, but recent research has exposed fundamental vulnerabilities: both methods can be adversarially manipulated to hide discriminatory behavior, exhibit high sensitivity to hyperparameters, and produce inconsistent attributions that are difficult to validate against actual model reasoning.

The gap between model performance and interpretability undermines trust in AI systems, slows diagnosis and debugging, complicates compliance in regulated domains, and hinders knowledge transfer across tasks. This problem is particularly acute in critical applications such as medical diagnosis, autonomous systems, and legal decision support,

where stakeholders require not only accurate predictions but also comprehensible justifications that can be validated, audited, and contested when necessary.

The fundamental limitation of post-hoc explanations stems from their disconnection from the learning process itself: explanations are generated after the model has been trained, making it inherently difficult to corroborate them against the model’s actual reasoning process. An alternative paradigm would embed interpretability directly into the representation from the outset, making structure the carrier of explanations rather than treating explanation as a separate post-processing step.

The relevance of this research lies in addressing the critical need for AI systems that are both accurate and inherently interpretable, particularly in the context of few-shot learning where minimal training data is available. Current approaches to explainable AI predominantly rely on post-hoc techniques that cannot guarantee fidelity to model reasoning, while few-shot learning methods typically depend on opaque learned embeddings through meta-learning frameworks. The convergence of these two challenges creates an opportunity for fundamentally rethinking how AI systems represent and manipulate knowledge.

This article presents a graph-based approach to concept formation that addresses both interpretability and few-shot learning through explicit structural representation. The core innovation is representing visual patterns as attributed graphs where nodes capture critical structural points and connecting segments, edges encode spatial adjacency, and attributes preserve geometric and semantic properties directly within graph elements. Multiple training samples are reduced to stable concept attractors through iterative structural composition, eliminating instance-specific variations while preserving shared topological patterns. This representation enables transparent classification through graph matching, where decision paths are navigable objects rather than implicit weight activations.

2 Analysis of Latest Achievements and Publications

Current research in explainable artificial intelligence has developed along several distinct but related trajectories, each addressing different aspects of the interpretability challenge with varying degrees of success and fundamental limitations.

Post-hoc explanation methods represent the dominant paradigm in contemporary XAI research. LIME generates local approximations of model behavior by perturbing inputs and training simple interpretable models on the perturbed samples, while SHAP computes Shapley values from cooperative game theory to attribute prediction importance to input features. Both methods have achieved widespread adoption due to their model-agnostic nature and ability to generate human-readable explanations for arbitrary classifiers. However, Slack (2020) demonstrated that both LIME and SHAP can be adversarially manipulated through a two-stage training process that maintains accuracy on normal inputs while hiding discriminatory behavior from explanation queries. Hooshyar and Yang (2024) further documented high sensitivity to hyperparameter selection and feature collinearity, showing that explanation stability varies dramatically with relatively minor configuration changes. These vulnerabilities stem from a fundamental architectural limitation: post-hoc methods operate independently from the model’s training process, making it impossible to verify whether generated explanations accurately reflect the model’s actual reasoning.

Graph-based representations for visual reasoning offer an alternative approach where structure carries semantic information explicitly. Vision GNN architectures treat images as graphs with content-based connectivity, enabling interpretable spatial reasoning through message passing on graph structures. Contour and skeleton representations have long been studied for capturing boundaries and topological structure in shape analysis, with comparative studies demonstrating advantages for tasks requiring structural invariance. Graph Edit Distance provides a principled framework for comparing such representations through sequences of node and edge operations, with recent hybrid approaches combining GED with learned embeddings to balance computational efficiency and interpretability. However, existing graph-based vision methods typically rely on learned graph neural networks that encode patterns in opaque weight matrices rather than explicit structural primitives, limiting their interpretability advantages.

Few-shot learning research addresses concept formation from minimal data through meta-learning frameworks that learn to adapt quickly across tasks. Prototypical Networks learn metric spaces where classification reduces to nearest-centroid matching, while Model-Agnostic Meta-Learning optimizes initialization parameters for rapid fine-tuning. Meta-Transfer Learning combines meta-learning with transfer learning to leverage pre-trained representations. These approaches demonstrate impressive data efficiency but fundamentally depend on learned embeddings whose semantics remain implicit and uninterpretable. The meta-learning paradigm requires episodic training across numerous tasks to learn adaptation strategies, creating dependencies on large meta-training sets even when individual task supervision is minimal.

Neuro-symbolic AI integrates explicit knowledge graphs and symbolic reasoning with learned representations, providing introspectable decision paths through the combination of neural perception modules with symbolic reasoning engines. Knowledge-graph-based explainable AI demonstrates that explicit graph structures improve interpretability over weight matrices by making relationships queryable and verifiable. Recent work on commonsense knowledge graphs for neural network interpretability shows promise for grounding learned representations in human-understandable concepts. However, neuro-symbolic approaches typically treat perception and reasoning as separate stages, with learned perception modules remaining opaque and requiring post-hoc explanation.

Despite these advances, no existing approach combines structural explainability with few-shot capability through graph-based concept formation where the representation itself is inherently interpretable. The theoretical foundation for structural reduction as a learning mechanism has been developed in recent work on energy-based neural network formulations and architectural principles of information representation, but practical implementations and empirical validation on standard benchmarks remain limited.

3 Earlier Unresolved Parts of the General Problem and Study Aim

The analysis of existing research reveals three fundamental gaps that limit progress toward truly explainable few-shot learning systems.

First, current explainable AI methods remain architecturally disconnected from the learning process. Post-hoc explanation techniques operate independently from model training, generating attributions by perturbing inputs or analyzing gradients after parameters have been fixed. This separation creates an inherent validation problem: there is no mechanism to verify that explanations accurately reflect the model’s actual reasoning rather than artifacts of the approximation process. Even when explanations appear plausible to humans, adversarial examples demonstrate that they can be systematically misleading. The fundamental issue is that interpretability is treated as a post-processing step rather than an intrinsic property of the representation itself.

Second, few-shot learning approaches achieve data efficiency through opaque learned embeddings rather than explicit structural primitives. Meta-learning frameworks optimize embedding spaces or initialization parameters to enable rapid adaptation, but the resulting representations encode patterns implicitly in weight matrices that are difficult to inspect, validate, or transfer across domains. While these methods demonstrate impressive performance on benchmark tasks, they provide no mechanism for understanding what patterns have been captured or why a particular test instance matches a learned concept. The embedding space may capture discriminative features effectively but remains semantically opaque.

Third, existing graph-based vision methods employ learned graph neural networks that propagate information through message passing operations, encoding patterns in edge weights and node embeddings that are learned through gradient descent. While these approaches represent images as graphs rather than tensors, the actual pattern recognition still depends on opaque parameters. The structural representation provides spatial inductive bias but not interpretability, because the semantics of recognition are encoded in learned weights rather than explicit graph properties.

The aim of this research is to develop and validate a graph-based representation for visual concept formation that makes structure the carrier of both representation and explanation, enabling few-shot learning through explicit structural reduction without gradient-based optimization. Specifically, the research aims to: (1) formulate a graph representation for contour images where nodes represent critical structural points and connecting segments with semantic attributes that preserve geometric and topological properties; (2) develop structural reduction operations that iteratively compose multiple training graphs into stable concept attractors by eliminating instance-specific variations while preserving shared topological patterns; (3) design parametric generalization strategies that transform point estimates into range representations capturing acceptable variation while maintaining interpretability; (4) validate the approach on standard benchmark data to demonstrate that few-shot learning from minimal training samples (2-6 instances per concept) can achieve meaningful classification accuracy through explicit graph matching; (5) analyze concept properties and confusion patterns to confirm that decision boundaries are queryable, verifiable, and trace directly to structural and geometric properties encoded in graph elements.

4 Materials and Methods

4.1 Graph Representation of Contour Images

The proposed approach represents image contours as bipartite attributed graphs where structure and geometry are encoded as first-class queryable entities rather than implicit learned features. This representation consists of three complementary components: node types with topological roles, attributes capturing geometric and directional properties, and normalization procedures ensuring invariance to scale and translation.

Node types and bipartite structure. The graph alternates between Point nodes representing critical structural locations and Line nodes representing connecting segments. This bipartite design makes line segments first-class nodes with their own attributes rather than treating them as mere connectivity markers between points. Point nodes are classified into four topological types based on connectivity: EndPoint nodes mark terminals of open contours (degree 1), CornerPoint nodes indicate sharp directional changes (degree 2, storing angle attributes), IntersectionPoint nodes represent junctions where multiple segments meet (degree ≥ 3), and StartPoint nodes serve as anchors for consistent graph traversal. Each node carries multiple type labels (e.g., ["Point", "CornerPoint"]), enabling both type-specific queries and generic structural operations. Bidirectional CONNECTED_TO relationships link Point and Line nodes, creating traversal patterns of the form Point \rightarrow Line \rightarrow Point \rightarrow Line.

Attributes and semantic properties. Each node type carries attributes encoding geometric, directional, and topological information. Point nodes store Cartesian coordinates (x, y) in both absolute and normalized forms, as well as

angle values between connected lines. Line nodes store endpoint pairs (x_1, y_1, x_2, y_2) and length attributes, enabling direct access to segment geometry without recomputation. Directional attributes capture spatial orientation: each Line is assigned a quadrant (I, II, III, IV) based on displacement vectors, plus horizontal direction (LEFT, RIGHT, NONE) and vertical direction (TOP, BOTTOM, NONE) encoding relative positioning. These attributes support pattern comparison across different contours by encoding how shapes develop spatially.

Coordinate normalization. To achieve scale and translation invariance, all spatial measurements are normalized to a centered coordinate system. Point coordinates are transformed via $\text{normalized}_x = (x - \text{center}_x) / \text{center}_x$ and similarly for y , producing values in the range $[-1, 1]$. This transformation ensures that concept attractors remain stable across variations in image size, position, and proportions, allowing shapes to be recognized despite changes in viewing conditions while preserving relative geometric relationships.

4.2 Concept Formation via Structural Reduction

Concept formation proceeds through iterative structural composition that reduces multiple training graphs to a single canonical attractor representing shared topological and parametric patterns. This process follows principles of architectural information representation where learning occurs through structural reduction rather than gradient descent.

Iterative composition algorithm. Given training samples G_1, G_2, \dots, G_n , the algorithm initializes the concept with the first graph ($C_0 = G_1$) and iteratively refines it through custom reduction operations: $C_{i+1} = \text{CRO}(C_i, G_{i+1})$ for $i = 1, 2, \dots, n - 1$. Each integration step coordinates five operations: (1) align start points across both graphs using spatial clustering to establish consistent traversal origins; (2) preprocess critical points through reduction strategies to achieve structural compatibility; (3) generate synchronized traversal paths maintaining correspondence between critical points; (4) identify common structure along matched segments by comparing all simple paths between consecutive critical points and selecting best matches via node similarity; (5) merge geometric and semantic properties using type-specific integration strategies.

Structural reduction rules. Three complementary strategies align critical point structures before path-level comparison, operating within a type hierarchy where $\text{IntersectionPoint} \rightarrow \text{CornerPoint} \rightarrow \text{Point}$, with the constraint that StartPoint and EndPoint types define structural boundaries and cannot be arbitrarily reduced. Endpoint removal eliminates terminal nodes lacking correspondence across samples by computing similarity matrices between concept and image endpoints, removing those with maximum similarity below threshold, and traversing from removed endpoints toward nearest critical points to eliminate connecting paths. Intersection point merging consolidates junction nodes through semantic reduction (relabeling IntersectionPoint nodes with degree ≤ 2 as CornerPoint or EndPoint) followed by excess intersection removal via similarity comparison. Path pruning normalizes segments between aligned critical points by identifying all simple paths, filtering those containing intermediate critical points, computing node similarity matrices, and constructing result paths containing only matched positions using the shorter path as template.

Parametric generalization. While structural reduction determines which graph elements persist, parametric generalization determines how attributes are merged. Numeric properties transform into range representations: for values v_1, v_2, \dots, v_n , the merged property becomes $\{\min : \min_i v_i, \max : \max_i v_i, \text{center} : \frac{1}{n} \sum_i v_i\}$. Categorical properties preserve only values consistent across all samples: for strings s_1, s_2, \dots, s_n , the merged property is s_1 if all values match, otherwise the property is excluded. List properties retain only elements present in all samples via set intersection: $L_1 \cap L_2 \cap \dots \cap L_n$. These strategies ensure concepts encode distributions of acceptable values rather than point estimates, enabling flexible matching while maintaining interpretability through explicit parameter bounds.

4.3 Classification via Graph Edit Distance

Classification compares each test graph against all concept attractors using Graph Edit Distance as a structural similarity metric, then selects the concept with highest similarity score. GED quantifies dissimilarity as the minimum-cost sequence of edit operations (node/edge insertion, deletion, substitution) required to transform one graph into another.

Custom cost functions. Node substitution costs enforce label compatibility (concept labels must form a subset of test graph labels) and compute weighted property differences across shared attributes. Range-based cost functions enable matching against parametric distributions: for test value v and concept range $[\min, \max]$ with center c , the cost is zero if $v \in [\min, \max]$, otherwise proportional to the minimum distance to the range boundary. Node deletion and insertion carry unit costs, while edge operations incur reduced costs (0.1) to prioritize topological over connectivity differences.

Similarity scoring. The optimal edit path is computed via dynamic programming with a 60-second timeout per comparison to prevent excessive computation on large graphs. Raw GED values are converted to normalized similarity scores via $\text{similarity} = 1.0 - (\text{GED} / \text{max_cost})$, where $\text{max_cost} = |V_{\text{test}}| + |V_{\text{concept}}|$ represents the theoretical upper bound for complete graph replacement. The test instance is assigned to the concept with maximum similarity score, providing a nearest-neighbor classification rule in GED-based metric space.

4.4 Experimental Setup

Validation employs a subset of the MNIST handwritten digit dataset to evaluate few-shot learning capability with minimal training data and interpretable graph structures.

Dataset and classes. Six digit classes (1, 2, 3, 6, 7, 9) were selected to represent diverse structural complexity. These classes exhibit varying topological characteristics: digit 1 represents simple linear structures, digits 2 and 3 exhibit curved open contours with multiple corner points, digits 6 and 9 contain closed loops with intersection points, and digit 7 demonstrates angular transitions.

Training configuration. Each digit class was divided into structural subclasses capturing distinct writing styles. Training data consisted of 2-4 manually selected samples per subclass, with 8 concepts total across the six classes. This few-shot configuration constrains the algorithm to learn from minimal supervision without gradient-based optimization.

Data augmentation. Each original training sample was augmented by generating 10 variants through random rotation (uniformly sampled within $\pm 10^\circ$) and spatial translation (up to 10% of image dimensions). Images were rendered at 100×100 pixels in grayscale with transformations applied on black background to maintain boundary clarity.

Graph extraction pipeline. Each image underwent binary thresholding, morphological skeletonization via medial axis transformation, Growing Neural Gas topology fitting, Ramer-Douglas-Peucker simplification, graph construction with Point and Line nodes, and coordinate normalization to centered $[-1, 1]$ range.

Test set. Classification performance was evaluated on 5467 test images from standard MNIST test split, with approximately 750-1000 samples per digit class. No test images were used during concept formation, ensuring unbiased generalization evaluation.

5 Study Results and Their Discussion

5.1 Concept Attractor Characteristics

The iterative reduction process transforms multiple training graphs into single concept attractors characterized by explicit structural metrics. Structural complexity correlates systematically with digit topology. Simple linear structures (digit 1) yield minimal graphs with only 3 nodes and mean degree 1.33, reflecting unbranched paths from start to endpoint. Curved open contours (digits 2, 3, 7) produce moderate-sized graphs with 5-12 nodes and mean degrees between 1.60 and 2.00, where corner points encode directional transitions. Closed-loop digits (6, 9) exhibit higher mean degrees (2.00) due to intersection points creating cyclic paths.

Critical point distributions reveal topological distinctions: all concepts except closed-loop digits contain exactly one endpoint, signifying open contours with terminal segments. Concepts for digits 6 and 9 lack endpoints but contain intersection points where loops close, a topological signature differentiating circular from linear structures. Corner point counts range from 1 to 3, reflecting the number of directional changes required to trace each digit’s canonical form.

Parametric generalization encodes acceptable variation while preserving structural identity. For digit 3, normalized x-coordinates span $[-0.7, 0.2]$ with center -0.33 ; normalized y-coordinates span $[0.3, 0.9]$ with center 0.63 . Structural counts encode variation: endpoint counts range $[2, 4]$ with center 2.67 ; intersection point counts range $[0, 2]$ with center 0.67 . Categorical properties consistent across samples persist as exact values, while inconsistent properties are excluded. This range-based encoding is interpretable: attribute centers indicate typical values, while min-max bounds reveal acceptable deviation extent, making decision boundaries explicit and queryable.

5.2 Classification Performance

Across 5467 test images, the system achieved 82.35% accuracy with 83.28% precision, 82.35% recall, and 82.16% F1 score. Pipeline success rate reached 100%, with only 10 images (0.18%) failing due to skeletonization errors producing disconnected graphs. These results demonstrate that few-shot concept formation from 2-6 training samples per subclass achieves meaningful classification without gradient-based optimization, relying solely on explicit structural comparison.

Per-class performance reveals substantial variation correlated with structural distinctiveness. Digits 6 and 9 achieve highest precision (94.23%, 91.55%) and F1 scores (85.40%, 90.55%), likely due to unique topological signatures with closed loops and intersection points distinguishing them from open-contour digits. Digit 7 exhibits lowest precision (74.38%) and F1 score (78.06%), suffering from structural ambiguity with digit 1 where both are angular open contours with similar orientations. Digits 2 and 3 demonstrate moderate performance (precision 84.17% and 78.21%), with digit 2 achieving higher precision at the cost of lower recall (60.02%).

Confusion matrix analysis reveals systematic misclassification patterns. Primary confusion occurs between digits 2 and 3 (152 misclassifications of 2 as 3, 28 misclassifications of 3 as 2), reflecting overlapping curved morphology where corner positions and line orientations vary subtly. Secondary confusions appear between digits 7 and 1 (118 misclassifications of 7 as 1), attributable to shared angular open-contour structure. Digits 6 and 9 exhibit minimal mutual confusion (48 and 4 misclassifications), confirming that closed-loop topology provides discriminative power. Errors concentrate along structurally similar digit pairs, validating that graph-based representation captures meaningful topological distinctions while remaining sensitive to ambiguous boundary cases.

5.3 Explainability Through Structure

The graph representation makes structure the carrier of explanations, providing inherent interpretability without post-hoc approximation. Graph cycles directly encode topological properties: closed-loop digits exhibit mean degree 2.0 and contain intersection points, while open-contour digits exhibit lower mean degrees and terminal endpoints. Parametric generalization encodes decision boundaries as explicit attribute ranges rather than implicit weight matrices: normalized coordinate spans define acceptable variation transparently, while categorical properties persist only when consistent across training samples.

This contrasts fundamentally with post-hoc explanation methods. LIME and SHAP generate attributions by perturbing inputs and approximating model behavior locally, yielding explanations vulnerable to adversarial manipulation. The present system embeds semantics directly in graph elements: explanations are navigable objects derived from structure itself. Confusion patterns validate this interpretability, as misclassifications trace directly to queryable topological and geometric properties.

5.4 Comparison with Existing Approaches

The experimental results demonstrate competitive classification performance using only 2-6 training samples per concept, orders of magnitude fewer than typical supervised learning baselines. Traditional convolutional networks require hundreds to thousands of labeled examples per class to achieve comparable accuracy on MNIST, while prototypical networks and meta-learning approaches necessitate task-level training across numerous episodes. The present approach forms concepts through direct structural composition without gradient descent, episodic sampling, or auxiliary training tasks. The 82.35% accuracy represents meaningful few-shot capability, particularly given that the method operates on explicitly interpretable graph structures rather than opaque embeddings.

Limitations include computational complexity of Graph Edit Distance (NP-hard in general case, requiring 60-second timeout per comparison), dependency on preprocessing pipeline quality (0.18% skeletonization failures), limited rotation invariance (bounded by $\pm 10^\circ$ augmentation range), and discarding of texture information (contour-only representation). These constraints suggest directions for future optimization while confirming the viability of structural reduction as a learning mechanism.

6 Conclusions and Perspectives for Further Development

This research presents a graph-based approach to concept formation that makes structure the carrier of explanations, advancing explainable artificial intelligence through explicit structural encoding rather than opaque weight matrices. The key contributions and findings are as follows.

First, the bipartite graph representation with Point and Line nodes as first-class entities enables transparent encoding of visual patterns where topological roles, geometric properties, and spatial relationships are directly queryable. Coordinate normalization ensures scale and translation invariance, while semantic attributes preserve meaning within graph elements rather than requiring post-hoc interpretation of learned features.

Second, structural reduction through iterative composition of training graphs demonstrates that concept attractors can be formed from minimal supervision (2-6 samples per concept) without gradient-based optimization. The three reduction strategies (endpoint removal, intersection point merging, path pruning) eliminate instance-specific variations while preserving shared topological patterns, producing compact concept graphs (3-12 nodes) with interpretable structural signatures.

Third, parametric generalization transforms point estimates into range representations that capture acceptable variation while maintaining transparency. Numeric attributes encode min-max bounds with central tendencies, categorical properties preserve only consistent values, and list properties retain universal elements. This encoding makes decision boundaries explicit and verifiable rather than implicit in learned weights.

Fourth, validation on MNIST handwritten digits confirms that few-shot learning through explicit graph matching achieves 82.35% classification accuracy competitive with traditional approaches requiring vastly more training data. Confusion patterns trace directly to structural similarity, demonstrating that the representation captures meaningful topological distinctions. Closed-loop digits with intersection points achieve highest discrimination (precision 91-94%), while structurally ambiguous pairs (digits 2/3, digits 7/1) exhibit expected confusions corresponding to overlapping geometric properties.

Fifth, the approach provides inherent explainability where decision paths are navigable graph elements rather than gradient-based attributions vulnerable to adversarial manipulation. Graph cycles encode topology directly, parametric ranges define acceptable variation transparently, and classification scores derive from explicit structural edit operations. This contrasts fundamentally with post-hoc methods that approximate opaque model behavior after training.

Perspectives for further development include several promising directions. Computational efficiency can be improved through approximate Graph Edit Distance algorithms, hierarchical concept organization to prune comparison space, and learned cost functions that preserve interpretability while accelerating matching. Representation capacity can

be extended to handle rotation invariance beyond augmentation ranges through learned alignment procedures, incorporate texture and gradient information through additional node attributes while maintaining structural primacy, and generalize to three-dimensional contours for volumetric data. Theoretical foundations merit deeper investigation of convergence properties for structural reduction operators, formal characterization of concept attractor stability, and relationships between graph complexity and generalization capacity. Application domains should explore medical image analysis where explainability is critical, industrial quality control requiring transparent defect classification, and educational systems where concept transparency supports learning. Integration with neuro-symbolic reasoning frameworks could combine structural concept formation with symbolic manipulation, enabling compositional generalization and abstract reasoning while maintaining transparency throughout the inference pipeline.

The demonstrated viability of graph-based structural reduction for few-shot explainable learning establishes a foundation for AI systems where interpretability is an intrinsic property of representation rather than a post-processing step. By making structure the carrier of both pattern encoding and decision justification, this approach advances toward artificial intelligence systems that are simultaneously accurate, data-efficient, and inherently comprehensible to human stakeholders.

7 References

Bunke, H., Allermann, G. (1983), "Inexact graph matching for structural pattern recognition", *Pattern Recognition Letters*, Vol. 1, No. 4, P. 245-253.

Chatbri, H., Kameyama, K., Kwan, P. (2016), "A comparative study using contours and skeletons as shape representations for binary image matching", *Pattern Recognition Letters*, Vol. 76, P. 59-66.

Finn, C., Abbeel, P., Levine, S. (2017), "Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks", *Proceedings of the 34th International Conference on Machine Learning*, P. 1126-1135.

Gharoun, Y., Chadli, M., Douik, A. (2023), "Meta-learning Approaches for Few-Shot Learning: A Survey", *ACM Computing Surveys*, Vol. 55, No. 11, P. 1-38.

Han, K., Wang, Y., Guo, J., Tang, Y., Wu, E. (2022), "Vision GNN: An Image is Worth Graph of Nodes", *Advances in Neural Information Processing Systems*, Vol. 35, P. 8305-8319.

Hooshyar, D., Yang, Y. (2024), "Problems With SHAP and LIME in Interpretable AI for Education: A Comparative Study of Post-Hoc Explanations and Neural-Symbolic Rule Extraction", *IEEE Access*, Vol. 12, P. 137472-137490. DOI: <https://doi.org/10.1109/ACCESS.2024.3463948>

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P. (2002), "Gradient-based learning applied to document recognition", *Proceedings of the IEEE*, Vol. 86, No. 11, P. 2278-2324.

Liu, C. (2023), "MATA: Multi-Attribute Time Series Anomaly Detection via Meta-learning and Attentive Adaptive Framework", *Proceedings of the VLDB Endowment*, Vol. 16, No. 12, P. 3983-3995.

Lundberg, S. M., Lee, S. (2017), "A Unified Approach to Interpreting Model Predictions", *Advances in Neural Information Processing Systems*, Vol. 30.

Nawaz, U., Anees-ur-Rahaman, M., Saeed, Z. (2025), "A review of neuro-symbolic AI integrating reasoning and learning for advanced cognitive systems", *Intelligent Systems with Applications*, P. 200541.

Parzhyn, Y. (2025), "Architecture of Information", *arXiv preprint arXiv:2503.21794*. DOI: <https://doi.org/10.48550/arXiv.2503.21794>

Parzhyn, Y., Lapin, M., Bokhan, K. (2025), "A NEW APPROACH TO BUILDING ENERGY MODELS OF NEURAL NETWORKS", *Advanced Information Systems*, Vol. 9, No. 4, P. 100-119.

Parzhin, Y. (2013), "Principles of modal and vector theory of formal intelligence systems", *arXiv preprint arXiv:1302.1334*, available at: <https://arxiv.org/abs/1302.1334>

Parzhin, Y., Galkyn, S., Sobol, M. (2022), "Method For Binary Contour Images Vectorization Of Handwritten Characters For Recognition By Detector Neural Networks", *2022 IEEE 3rd KhPI Week on Advanced Technology (KhPIWeek)*, P. 1-6. DOI: <https://doi.org/10.1109/KhPIWeek57572.2022.9916331>

Rajabi, E., Etminani, K. (2024), "Knowledge-graph-based explainable AI: A systematic review", *Journal of Information Science*, Vol. 50, No. 4, P. 1019-1029. DOI: <https://doi.org/10.1177/01655515221112844>

Ribeiro, M. T., Singh, S., Guestrin, C. (2016), "Why should i trust you? Explaining the predictions of any classifier", *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, P. 1135-1144.

Shen, W., Wang, Y., Bai, X., Wang, H., Zhang, L. J. (2016), "Shape Recognition by Bag of Skeleton-associated Contour Parts", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, P. 2125-2133.

Slack, D., Hilgard, S., Jia, E., Singh, S., Lakkaraju, H. (2020), "Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods", *arXiv preprint arXiv:1911.02508*, available at: <https://arxiv.org/abs/1911.02508>

Snell, J., Swersky, K., Zemel, R. (2017), "Prototypical Networks for Few-Shot Learning", *Proceedings of NeurIPS*.

Wang, F., Jiang, P., Riesen, K., Zhang, J. (2021), "Combinatorial Learning of Graph Edit Distance via Dynamic Embedding", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, P. 5184-5193.

About the Authors

Lapin Mykyta - PhD Student, National Technical University "Kharkiv Polytechnic Institute", Department of System Analysis and Information and Analytical Technologies, Kharkiv, Ukraine; email: Mykyta.Lapin@cit.khpi.edu.ua; ORCID: <https://orcid.org/0000-0003-2037-5587>

Bokhan Kostiantyn - PhD, Associate Professor, National Technical University "Kharkiv Polytechnic Institute", Department of System Analysis and Information and Analytical Technologies, Kharkiv, Ukraine; email: kostiantyn.bokhan@khpi.edu.ua; ORCID: <https://orcid.org/0000-0002-9861-8911>

Parzhyn Yuriy - Doctor of Sciences (Engineering), Postdoctoral Fellow, Augusta University, School of Computer and Cyber Sciences, Augusta, USA; email: yparzhyn@augusta.edu; ORCID: <https://orcid.org/0000-0002-5007-4076>

Perevoznyk Kyrylo - PhD Student, National Technical University "Kharkiv Polytechnic Institute", Department of System Analysis and Information and Analytical Technologies, Kharkiv, Ukraine; email: kyrylo.perevoznyk@cs.khpi.edu.ua; ORCID: <https://orcid.org/0000-0002-4668-6870>