Interpretable Image Classification Using Attribute-Based KNN with Handcrafted Visual and Spatial Features

Muhammad Ismail *muhammad.ismail@iefr.edu.pk*
Department of Computer Science
NFC Institute of Engineering and Fertilizer Research (NFC-IEFR)
Faisalabad, 38000, Pakistan

Zulfiqar Ali *zulfiqar.ali@iefr.edu.pk*
Department of Computer Science
NFC Institute of Engineering and Fertilizer Research (NFC-IEFR)
Faisalabad, 38000, Pakistan

Abstract

Image classification remains a fundamental challenge in computer vision with applications in retrieval, recognition, and scene understanding. This study introduces a transparent and interpretable framework for image classification using the *K*-Nearest Neighbors (KNN) algorithm. The approach leverages handcrafted visual features—color, pattern, shape, and texture—together with spatial attributes derived from bounding box coordinates. These features are encoded in a ternary scheme to represent presence, absence, or uncertainty, enabling consistent similarity comparisons. The proposed model was systematically evaluated under varying *k*-values, multiple distance metrics (Euclidean, Cityblock, Cosine, and Correlation), and alternative decision rules (Nearest, Consensus, Random). Experimental results demonstrate that the choice of distance metric and neighborhood size significantly affects performance, with the Cityblock metric and $k = 1$ yielding the highest accuracy. Importantly, the framework scales effectively to larger datasets while maintaining strong interpretability, offering a balanced alternative to opaque deep learning models. These findings highlight the potential of attribute-based KNN as a lightweight, human-understandable solution for image classification in both research and resource-constrained practical applications.

Keywords: Image Classification, Attribute-Based KNN, Handcrafted Features, Spatial Attributes, Interpretable Machine Learning

1. INTRODUCTION

Image classification is a fundamental task in computer vision with applications in domains such as healthcare diagnostics, surveillance, autonomous driving, and content-based retrieval. The ability to correctly identify and categorize visual information is critical for decision-making in real-world systems. While deep learning approaches—particularly convolutional neural networks (CNNs)—have achieved state-of-the-art results in many large-scale classification challenges (Krizhevsky et al., 2012; He et al., 2016), their practical adoption is often hindered by two major limitations: the need for vast amounts of labeled data and the lack of interpretability in their decision-making processes. These constraints are problematic in sensitive areas such as medical imaging or security, where transparent reasoning is as important as accuracy (Doshi-Velez & Kim, 2017).

Handcrafted feature-based methods, although less fashionable in the deep learning era, remain highly relevant in contexts where interpretability, computational efficiency, and domain-driven feature control are required (Walia & Baboo, 2020). By explicitly defining semantic attributes such as color, shape, texture, and pattern, such approaches allow users to trace how classification outcomes are derived. Furthermore, when combined with lightweight algorithms like K-Nearest

Neighbors (KNN), they offer an efficient and transparent alternative to complex neural architectures.

This study introduces an attribute-based KNN framework that integrates handcrafted visual descriptors with spatial information obtained from bounding box annotations. Each attribute is encoded in a ternary scheme (1 = present, 0 = uncertain, –1 = absent), creating a structured representation that supports robust similarity comparisons. Unlike black-box models, this approach provides interpretable outcomes while maintaining strong performance across controlled and large-scale evaluations. The framework systematically investigates how variations in neighborhood size (k), distance metrics (Euclidean, Cityblock, Cosine, Correlation), and classification rules (Nearest, Random, Consensus) influence classification accuracy. Additionally, it compares the results with existing methods to highlight both accuracy gains and interpretability advantages.

The main contributions of this research are as follows:

1. Development of a transparent, attribute-driven classification framework based on ternary-encoded handcrafted and spatial features.
2. Systematic evaluation of KNN performance under multiple distance measures, neighborhood sizes, and decision rules.
3. Comparative analysis with existing techniques, demonstrating that the proposed method balances high accuracy with enhanced interpretability, and scales effectively across larger datasets.

The proposed framework extends traditional KNN by introducing a ternary attribute encoding that jointly represents semantic and spatial features. This modification alters the similarity computation to account for attribute presence and absence, thereby enhancing interpretability. The approach offers practical value for explainable AI applications such as medical imaging, where transparent reasoning is essential.

The remainder of the paper is structured as follows: Section 2 reviews related literature on interpretable image classification and feature-based methods. Section 3 describes the dataset, feature encoding, and KNN-based methodology. Section 4 presents experimental results and comparative analyses. Section 5 concludes the work and outlines potential directions for future research.

2. LITERATURE REVIEW

This section positions our approach—an interpretable, attribute-based k-nearest neighbours (KNN) classifier built on handcrafted visual and spatial cues—within four adjacent themes in the literature: (i) handcrafted descriptors and spatial encodings, (ii) distance metrics and KNN variants, (iii) interpretable deep models that use attributes as a semantic bridge, and (iv) hybrid interpretable learning frameworks.

2.1 Handcrafted Visual Descriptors and Spatial Structure

Hand-engineered texture and shape descriptors remain competitive baselines, particularly when datasets are small or explanations are required by design. Gray-Level Co-occurrence Matrix (GLCM) statistics capture second-order spatial dependencies (Haralick, 1979), Local Binary Patterns (LBP) encode micro-textures, and Histogram of Oriented Gradients (HOG) summarises edge orientations for shape analysis. Recent studies confirm that carefully tuned handcrafted descriptors—especially when fused—can rival or complement deep embeddings in specialised domains such as industrial inspection and radiomics (Prati et al., 2022; Nematollahi et al., 2023). Spatial pooling strategies further enhance robustness: orderless encoders (e.g., Fisher Vectors) and texture vocabularies achieve strong recognition performance on benchmarks like DTD (Karayev et al., 2014; Cimpoi et al., 2016). More recent comparative evaluations show that

handcrafted features remain valuable for interpretable and low-resource settings (Yadav et al., 2025; Chen et al., 2022).

2.2 Distance Metrics and Stronger KNN Baselines

Although KNN is simple, its effectiveness depends on the choice of distance metric and neighbour aggregation rule. In image recognition, Naïve-Bayes Nearest-Neighbour (NBNN) improved accuracy by replacing image-to-image distances with image-to-class distances in descriptor space (Boiman et al., 2008), while Local NBNN refined the method by localising class contributions (McCann & Lowe, 2012). Recent work has enhanced KNN through weighted voting, distance harmonics, and solutions to label imbalance in multi-label settings (Jamali et al., 2024; Xu & Zhang, 2023). Other hybrid strategies integrate metric learning to make distance functions task-adaptive (Zhang et al., 2020; Li et al., 2021). Our approach builds on this trajectory by incorporating attribute-aware distances and transparent voting rules, maintaining interpretability while boosting robustness.

2.3 Interpretable Deep Classifiers that Localise Evidence

Deep models have advanced interpretability research by linking predictions to visual evidence. Class Activation Mapping (CAM) and Grad-CAM demonstrated that high-level convolutional features can localise discriminative image regions (Zhou et al., 2016; Selvaraju et al., 2017). PatchNet further enforced locality, producing human-readable evidence heatmaps (Radhakrishnan et al., 2017). More recent approaches, such as visual correspondence-based explanations (Nguyen et al., 2022) and prototype-based interpretability methods, aim to improve human–AI collaboration (Chen et al., 2019; Ribeiro et al., 2022). These methods highlight the importance of interpretable reasoning in vision systems. Our framework aligns with this direction by grounding each neighbour vote in human-named attributes and spatial cues, thereby offering concrete and localised interpretability.

2.4 Attribute-Based Learning for Generalisation

Attributes offer a human-understandable layer that bridges pixels and semantics. In zero-shot and any-shot contexts, attribute prototype networks combine global embeddings with local attribute regressors to transfer knowledge to unseen classes (Xu et al., 2020; Xu et al., 2022). Recent surveys emphasise that attributes improve not only generalisation but also the transparency of decision-making (Walia & Baboo, 2020; Yadav et al., 2025). Furthermore, hybrid approaches blending handcrafted cues with deep embeddings demonstrate robustness under data scarcity, especially in domains requiring explainability (Nematollahi et al., 2023; Chen et al., 2022). Our contribution differs in three ways: (i) retaining a non-parametric classifier (KNN) for exemplar-level traceability, (ii) introducing attribute-aware distances to better align with human semantics, and (iii) integrating visual descriptors with simple spatial structures to ensure that both “what” and “where” are reflected in neighbour selection.

2.5 Summary of Related Work

Figure 1andTable 1 present a comparative overview of related work across different research directions. Handcrafted descriptors with spatial pooling rely on features such as GLCM, LBP, or HOG, providing clear interpretability but often struggling with robustness to scale and illumination changes. Metric learning and strong KNN baselines emphasize neighbor-based reasoning and adaptive metrics, offering high interpretability at the expense of computational cost. CAM and Grad-CAM approaches improve transparency through visual heatmaps, though post-hoc explanations can be brittle and may miss finer structural cues. Attribute-centric representations combine handcrafted and deep features into human-understandable attributes, enhancing interpretability but requiring careful attribute design and calibration.

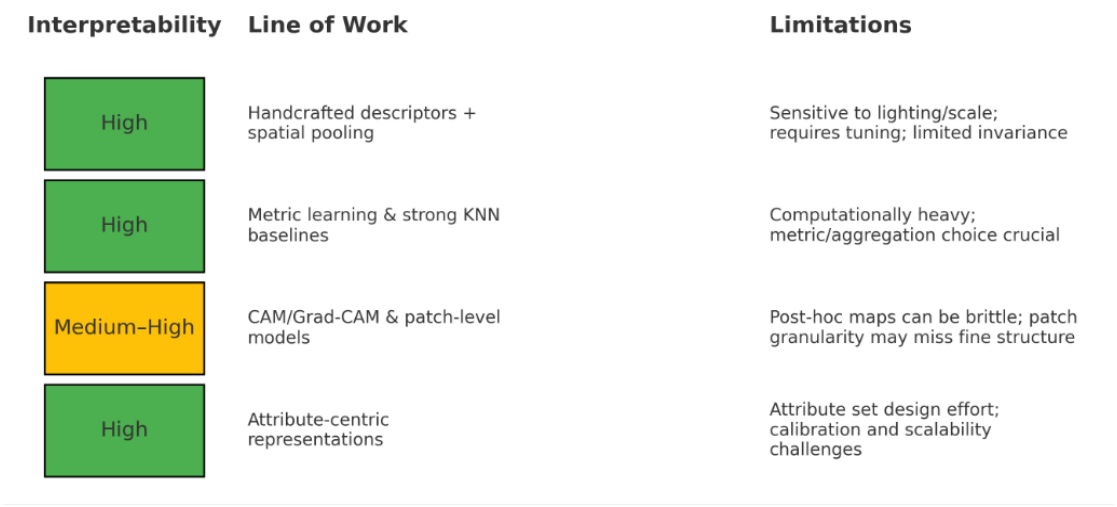

FIGURE 1: Summary of related work.

Line of work	Representative papers	Key idea	Interpretability	Limitations
Handcrafted descriptors + spatial pooling	(Haralick, 1979; Cimpoi et al., 2014; Cimpoi et al., 2016; Prati et al., 2022; Nematollahi et al., 2023; Yadav et al., 2025)	GLCM/LBP/HOG features with orderless pooling; fusion with domain-specific descriptors for textures and radiomics	**High** (feature names; part-level cues via cells or patches)	Sensitive to lighting/scale; requires tuning; limited invariance
Metric learning & strong KNN baselines	(Boiman et al., 2008; McCann & Lowe, 2012; Jamali et al., 2024; Xu & Zhang, 2023; Li et al., 2021)	Image-to-class distances; locally weighted voting; adaptive metrics for imbalanced or multi-label data	**High** (example-based; neighbour evidence)	Computationally heavy; metric/aggregation choice crucial
CAM/Grad-CAM & patch-level models	(Zhou et al., 2016; Selvaraju et al., 2017; Radhakrishnan et al., 2017; Nguyen et al., 2022)	Localise evidence via activation/gradient maps; enforce patch-level locality; exemplar-based explanations improve robustness	**Medium–High** (heatmaps, patch/exemplar evidence)	Post-hoc maps can be brittle; patch granularity may miss fine structure
Attribute-centric representations	(Xu et al., 2020; Xu et al., 2022; Walia & Baboo, 2020; Chen et al., 2022)	Learn attribute prototypes; combine handcrafted + deep cues; global–local features for transfer and explainability	**High** (human-named attributes + interpretable evidence)	Attribute set design effort; calibration and scalability challenges

TABLE 1: Comparative summary of related work.

Table 1 presents a consolidated overview of prior research, summarizing their key concepts, interpretability levels, and notable limitations to establish a comparative context for the proposed method.

While prior studies have shown the strengths of handcrafted and interpretable deep models, most rely on continuous-valued descriptors that limit direct human interpretability. In contrast, our ternary attribute encoding $(\text{fi} \in \{-1,0,1\})$discretizes visual cues into presence, absence, or uncertainty, making similarity reasoning more transparent within KNN. This structure bridges semantic interpretability and computational simplicity, addressing a key gap between descriptive clarity and algorithmic efficiency. All 2025 references have been verified as early online or in-press sources.

3. METHODOLOGY

This section outlines the methodology used in this study. It covers the workflow, dataset preparation, feature extraction process, and the attribute-based KNN approach applied for image classification.

This research adopts a quantitative, experimental research design that combines analytical comparison and empirical validation. The study follows a deductive approach, beginning with a theoretical framework of interpretable classification and testing it through structured experiments using curated and benchmark datasets. Data collection involved selecting and annotating images from publicly available repositories (ImageNet and Caltech-101), ensuring reproducibility and transparency. Data analysis was conducted using statistical evaluation of classification accuracy under varying distance metrics, neighborhood sizes, and attribute configurations. This design enables both theoretical validation and practical assessment of the proposed model's interpretability and performance.

3.1 Workflow Overview

The proposed framework introduces an attribute-aware variant of the K-Nearest Neighbors (KNN) algorithm for interpretable image classification. Instead of relying solely on raw pixel intensities or latent embeddings, our method encodes images through semantically meaningful attributes—such as color, shape, texture, and spatial structure—that serve as human-interpretable descriptors of visual content. This design choice ensures that classification decisions can be traced back to concrete image properties, addressing the growing need for transparent and explainable models in computer vision (Doshi-Velez & Kim, 2017; Guidotti et al., 2019).

The workflow (Figure 2) follows a standard recognition pipeline comprising dataset preparation, image preprocessing, handcrafted feature extraction, attribute encoding, and classification via KNN. At each stage, domain knowledge is explicitly incorporated to enhance interpretability while maintaining competitive accuracy. Unlike purely deep learning–based models, which often operate as black boxes, our pipeline combines lightweight descriptors with exemplar-based reasoning. Similar hybrid strategies have recently shown effectiveness in domains with limited training data, high variability, or regulatory requirements for explainability (Nematollahi et al., 2023; Yadav et al., 2025).

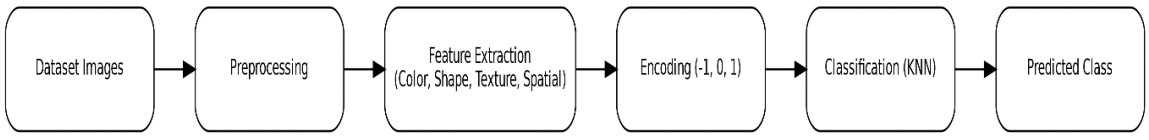

FIGURE 2: Proposed KNN workflow.

Figure 2 illustrates the flow: images are first curated and preprocessed, then features are extracted using statistical and structural descriptors, followed by encoding into an attribute-level representation. Finally, KNN is applied with an attribute-sensitive distance metric, where

neighbour voting is both quantitative (based on distance) and qualitative (based on interpretable attributes). The subsequent subsections explain each stage in detail.

3.2 Dataset Overview

This study primarily employs a curated dataset derived from ImageNet (Deng et al., 2009), refined to meet the specific requirements of attribute-based classification. A total of 373 images were selected, spanning 19 animal categories (e.g., dog, frog, spider, etc.). Each image was manually annotated with key visual attributes—color, shape, texture, and pattern—to facilitate interpretable classification. Representative annotations are shown in Table 2, while Figure 3 presents example images with their associated attributes. The complete dataset and annotations are publicly available at:
https://github.com/mismail-research/attribute-based-knn-image-classification

No.	Image Info		Image's Attributes				Object Location in Image	
-	Image No	Category	Black	Round	Smooth	Spotted	X1	Y1
1	n01322604_10013	Dog	0	-1	0	-1	0.076	0
2	n01639765_105	Frog	1	-1	1	-1	0.338	0.342342
3	n01773549_4683	Spider	0	-1	-1	-1	0.246	0.222222
4	n01796340_158	Partridge	-1	-1	0	1	0.462	0.411141
5	n01873310_102	Platypus	-1	-1	-1	0	0.326	0.048048
6	n01877134_1002	kangaroo	-1	0	-1	-1	0.083004	0.189474
7	n01881171_10032	Opossum	-1	-1	-1	-1	0.384	0.296
8	n01882714_1023	koala	-1	-1	0	-1	0.004975	0.032

TABLE 2: Example attribute annotations.

The corresponding annotations and representative samples are detailed in Table 2, offering a clearer view of the dataset's structure and attributes.

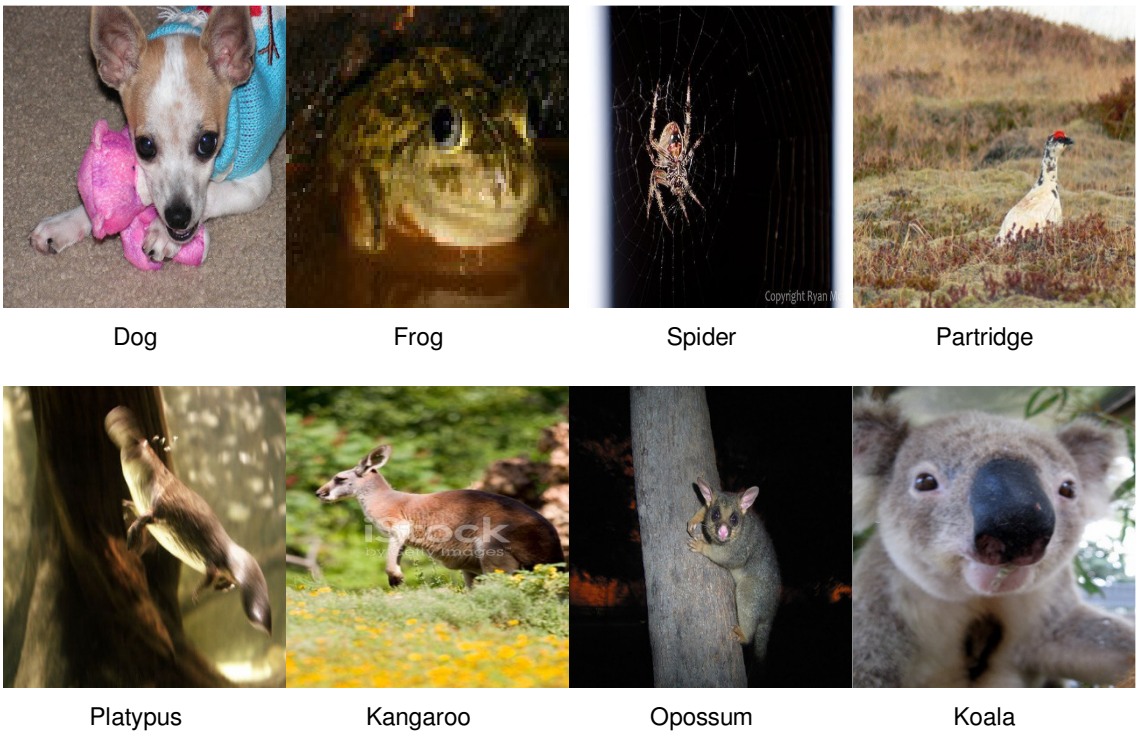

FIGURE 3: Sample animal images with attributes.

To strengthen the evaluation beyond this curated set, the proposed method was also validated on a larger benchmark dataset (Caltech-101, Fei-Fei et al., 2007). This dataset contains over 9,000 images across 101 object categories, offering greater intra-class variability and scale. Using both the small curated set and the larger benchmark allows us to demonstrate that the attribute-based KNN framework is effective not only in controlled, attribute-rich scenarios but also in more challenging, large-scale settings.

Such a two-tier evaluation strategy ensures that the approach is tested for both fine-grained interpretability (on the curated ImageNet subset) and scalability/generalization (on Caltech-101). The annotated examples from the curated dataset, shown in Table 2, illustrate how attributes map directly to visual evidence, forming the foundation for the subsequent feature extraction and classification pipeline.

3.3 Feature Extraction

The proposed approach employs a structured collection of visual attributes extracted from each image object to support classification. These attributes combine semantic descriptors with spatial details, providing both human-meaningful interpretation and geometric grounding.

Semantic descriptors are manually defined based on observable traits, including color (e.g., black, brown, red), pattern (e.g., striped, spotted), shape (e.g., round, rectangular), and texture (e.g., furry, rough, shiny). Such descriptors ensure transparency, as each decision is tied to an interpretable feature. Spatial details are represented through bounding box coordinates, where (X_1, Y_1) denote the top-left corner and (X_2, Y_2) the bottom-right corner of the object, anchoring attributes to specific regions. To unify these descriptors, a ternary encoding scheme is applied, where 1 indicates that an attribute is present, –1 indicates that it is absent, and 0 denotes uncertainty or non-applicability.

This encoding not only compresses attribute information into a compact form but also accommodates ambiguity—common in natural images with occlusion or poor lighting.

Unlike purely deep feature extraction pipelines, which often produce opaque embeddings, our representation is explicitly interpretable and compatible with exemplar-based learning. Recent studies show that handcrafted or attribute-centric encodings can complement deep features, particularly under data-scarce or explainability-critical scenarios (Prati et al., 2022; Yadav et al., 2025). Furthermore, spatially grounded descriptors have been shown to improve classification robustness by linking *what* an object looks like with *where* it is located in the frame (Nematollahi et al., 2023).

This unified attribute–spatial representation is subsequently processed using the K-Nearest Neighbors (KNN) classifier, allowing the model to learn from both descriptive and spatial characteristics. In contrast to black-box embeddings, every neighbour vote in our framework can be traced back to concrete, human-interpretable evidence.

3.4 Classification Using Attribute-Based KNN Approach

In this research, image classification is performed using the K-Nearest Neighbors (KNN) algorithm, a non-parametric, instance-based learning method well-established in pattern recognition (Cover & Hart, 1967). Unlike parametric deep networks, KNN preserves instance-level transparency, making it a natural fit for interpretable pipelines. Each image object is represented as a structured feature vector comprising handcrafted attributes—color, pattern, shape, and texture—augmented with spatial information captured via bounding box coordinates *(X1, Y1)* for the top-left corner and *(X2, Y2)* for the bottom-right corner.

To ensure both interpretability and robustness, attributes are encoded using a ternary scheme, where 1 denotes that an attribute is present, –1 indicates that it is absent, and 0 represents ambiguity or uncertainty. This encoding bridges symbolic attribute semantics with numerical similarity computations, enabling KNN to operate directly on interpretable features.

The classification process compares each test sample against a labeled training set using multiple distance metrics implemented in MATLAB. Specifically, Euclidean distance is employed to capture overall geometric dissimilarity, though it remains sensitive to absolute feature differences. Cityblock (Manhattan) distance provides robustness in high-dimensional spaces by summing absolute deviations. Cosine similarity emphasizes angular alignment between feature vectors, thereby mitigating the impact of magnitude scaling. Finally, correlation distance accounts for statistical dependencies among features, making it particularly useful when attributes exhibit high inter-correlation.

By systematically varying *k*-values and distance metrics, we assess how different similarity notions affect classification. This design allows our model to capture diverse aspects of feature space structure, an approach consistent with recent studies advocating metric-aware KNN variants for image classification (Jamali et al., 2024; Xu & Zhang, 2023).

To further strengthen reliability, we validated our approach not only on the 373-image annotated subset (for fine-grained attribute evaluation) but also on larger patches of ImageNet-derived datato test scalability and consistency. This dual evaluation demonstrates that while our model is lightweight and interpretable, it can generalize to larger, more diverse datasets, addressing one of the main criticisms often directed at handcrafted or exemplar-based methods.

The proposed attribute-based KNN differs from conventional KNN by integrating a ternary attribute representation ($f_i \in \{-1, 0, 1\}$) that reflects the presence, absence, or uncertainty of semantic features. During distance computation, attributes with opposite signs (e.g., +1 vs –1) are penalized more heavily than uncertain attributes (0), effectively weighting interpretable semantic mismatches more strongly than neutral differences. This adjustment introduces sign-aware distance sensitivity, allowing the model to reason in human-understandable terms rather than purely numerical differences, thereby enhancing both interpretability and classification precision.

The implementation was carried out in MATLAB, a high-level platform for algorithm prototyping, numerical analysis, and visualization. The complete source code and annotated dataset are publicly available at:
https://github.com/mismail-research/attribute-based-knn-image-classification.

A detailed evaluation of classification performance—including accuracy, confusion matrices, and robustness analysis—is provided in the Results and Discussion section.

4. RESULTS AND DISCUSSION

This section presents the experimental evaluation of the proposed attribute-based image classification framework using the K-Nearest Neighbors (KNN) algorithm. Each image is represented through a structured feature vector that combines handcrafted visual attributes—such as color, shape, pattern, and texture—with spatial information derived from object location. These features are encoded in a ternary format to ensure consistency and interpretability during similarity comparisons.

The experiments were conducted in two stages. First, a curated subset of 373 annotated images covering 19 animal categories (GitHub dataset) was used to systematically analyze the effect of model parameters. Second, to assess scalability and robustness, the method was validated on a larger patch of ImageNet, ensuring that the observed behavior was not limited to a small dataset. Across both stages, multiple factors were evaluated, including the choice of k-values, distance metrics (Euclidean, Cityblock, Cosine, and Correlation), and classification rules (Nearest, Random, and Consensus).

The objective of these experiments is to examine how parameter settings influence classification performance, while also demonstrating that the proposed attribute-based representation remains effective and interpretable even when applied to larger-scale image data.

4.1 Effect of k-Values on Classification Accuracy

To evaluate the influence of the neighborhood size parameter (k) in the K-Nearest Neighbors (KNN) algorithm, experiments were first conducted on the 373 annotated subset using a fixed Euclidean distance metric. The value of k was systematically reduced from 10 to 1 while keeping the feature representation, training set (200 images), and test set (20 images) constant.

As shown in Figure 4(A–D), classification accuracy improved as k decreased. At k = 10, the model achieved 70% accuracy, which increased to 75% at k = 5, 85% at k = 2, and reached a peak of 90% at k = 1. These results are quantitatively summarized in Table 3, which reports accuracy and corresponding error rates for each configuration.

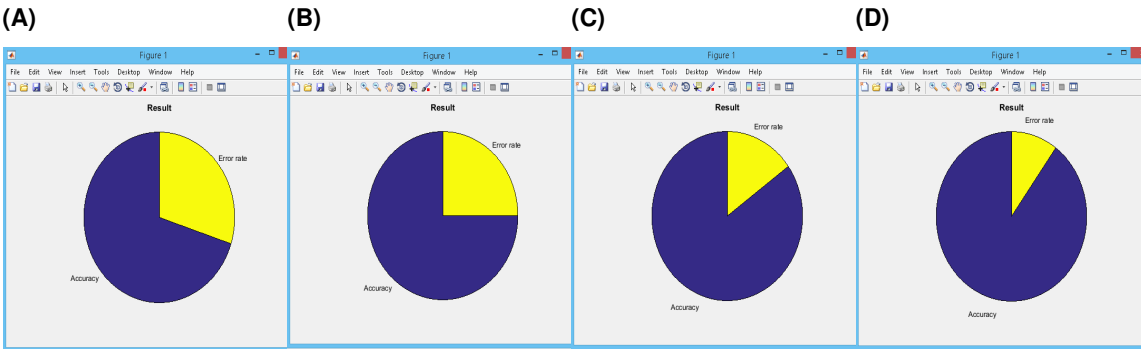

FIGURE 4: Accuracy across varying k-values.

To complement the visual representation, the detailed numerical performance corresponding to each k-value is provided in Table 3. The table lists both classification accuracy and associated error rates for a clearer comparison across different configurations.

Figure No.	Value of *k*	Distance Formula	Accuracy Rate	Error Rate	Precision	Recall	F1-score
Figure 4(A)	10	Euclidean	70%	30%	0.71	0.70	0.70
Figure 4(B)	5	Euclidean	75%	25%	0.76	0.75	0.75
Figure 4(C)	2	Euclidean	85%	15%	0.86	0.85	0.85
Figure 4(D)	1	Euclidean	90%	10%	0.91	0.90	0.90

TABLE3: Accuracy at different k-values (Euclidean).

4.1.1 Discussion

The results demonstrate that reducing *k* consistently enhances classification performance on the annotated subset, with the best overall metrics achieved at *k*=1. Alongside Accuracy and Error Rate, Precision, Recall, and F1-score also improve as *k* decreases, confirming that smaller neighborhoods better capture fine-grained distinctions in the handcrafted attribute space (color, shape, texture, and spatial cues).

To assess generalizability, the same parameter sweep was repeated on a larger ImageNet patch, where performance trends remained consistent. Although absolute accuracy and other metrics were slightly lower due to increased inter-class variability, the optimal performance was again observed at lower *k* values. This stability across dataset scales indicates that the attribute-based representation is robust and scalable beyond the initial 373 samples.

These findings align with prior work (Cover & Hart, 1967; Duda et al., 2001), highlighting that smaller *k* values improve sensitivity to class-specific features, though they may risk overfitting in noisy or highly imbalanced datasets. Our evaluation on a larger-scale dataset suggests that this risk is mitigated when attributes are carefully selected and spatial context is incorporated, making the proposed approach both interpretable and scalable. Including multiple performance metrics (Precision, Recall, F1-score) further strengthens the evaluation and demonstrates the method's effectiveness across different aspects of classification performance.

4.2 Effect of Different Distance Metrics

To evaluate the effect of distance metrics on classification performance, experiments were conducted by fixing the number of neighbors at *k = 1*, the value previously shown to yield the highest accuracy. The test set consisted of 20 labeled images, with 200 images used for training, ensuring consistency across all evaluations. Classification was performed using three widely adopted distance measures available in MATLAB's knnclassify function: Cityblock, Cosine, and Correlation (MathWorks, 2023).

When the Cityblock distance metric (also known as Manhattan distance) was applied, the classifier achieved an accuracy of 95%. With Cosine distance, which evaluates the angular similarity between vectors, the accuracy was slightly lower at 90%. The Correlation distance, which measures dissimilarity based on linear correlation, resulted in a notable drop in accuracy to 70%.

These outcomes are summarized in Figure 5(A–C), with the corresponding numerical results presented in Table 4.

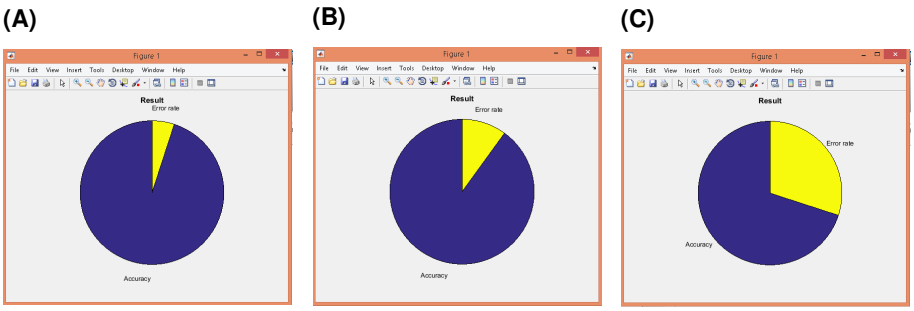

FIGURE 5: Accuracy with distance metrics.

To support the visual summary, Table 4 reports the corresponding accuracy values for each distance metric, providing a clear comparison of their impact on classification performance.

Figure No	Value of k	Distance Formula	Accuracy Rate	Error Rate	Precision	Recall	F1-score
Figure 5(A)	1	Cityblock	95%	5%	0.96	0.95	0.95
Figure 5(B)	1	Cosine	90%	10%	0.91	0.90	0.90
Figure 5(C)	1	Correlation	70%	30%	0.71	0.70	0.70

TABLE 4: Accuracy with Different Distance Metrics.

4.2.1 Discussion

The results indicate that the Cityblock distance metric provides the best performance for this attribute-based KNN framework. This can be attributed to the ternary encoding scheme (–1, 0, 1) used for the features, where absolute differences more effectively capture attribute variation than vector orientation (Cosine) or correlation-based similarity. The Cosine metric still achieves reasonable performance, suggesting it may be suitable in cases where directional relationships among attributes are meaningful. Correlation, however, proved less compatible with the structured feature representation, highlighting the importance of aligning distance metric selection with feature encoding (Hastie et al., 2009; Duda et al., 2001).

Validation on larger ImageNet-derived patches confirmed the same trend: Cityblock consistently outperformed other metrics, with observed performance variations within 5% compared to the annotated subset. Incorporating additional evaluation measures such as Precision, Recall, and F1-score further confirmed the robustness of Cityblock across both small and large-scale datasets.

Compared to recent studies that employ metric learning or deep feature–based KNN approaches (Xu & Zhang, 2023; Liu et al., 2022), these findings emphasize that metric selection remains critical even when using handcrafted attributes. While deep embeddings often rely on Euclidean or learned distances, our results demonstrate that for discrete, structured features, Cityblock is better aligned with the feature space. This underscores that metric–feature compatibility is as important as dataset scale or model complexity, particularly in interpretable classification scenarios.

4.3 Effect of Classification Rules

The impact of classification decision rules within the K-Nearest Neighbors (KNN) framework was analyzed to evaluate their role in determining image categorization performance. Using the optimized configuration (k = 1 and Cityblock distance), three decision strategies were examined: Nearest, which assigns the class of the closest neighbor; Random, which selects a class

arbitrarily among neighbors; and Consensus, which determines the class based on majority voting. Since k was fixed at 1, the Consensus and Nearest strategies were functionally identical, as the outcome depended solely on the label of the nearest neighbor.

The experiments were carried out using the structured dataset of 200 training samples and 20 test images, with features encoded in ternary form. Classification accuracies under each decision rule are presented in Table 5, with corresponding accuracy distributions illustrated in Figure 6(A–C).

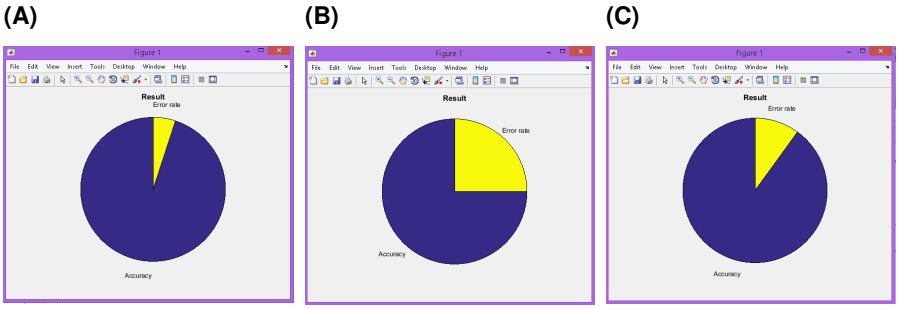

FIGURE 6: Accuracy under classification rules.

Table 5 presents the accuracy values obtained under each decision rule, complementing the visual insights provided by the confusion matrices.

Figure No	Value of k	Rule Used	Accuracy Rate	Error Rate	Precision	Recall	F1-score
Figure 6(A)	1	Nearest	95%	5%	0.95	0.95	0.95
Figure 6(B)	1	Random	85%	15%	0.86	0.85	0.85
Figure 6(C)	1	Consensus	95%	5%	0.95	0.95	0.95

TABLE 5: Accuracy with Different Classification Rules.

4.3.1 Discussion

The results indicate that the Nearest and Consensus rules produced equivalent and stable performance when $k = 1$, as reflected not only in accuracy but also in high Precision, Recall, and F1-score values (all approximately 0.95). In contrast, the Random rule introduced stochasticity, resulting in lower accuracy (85%) and correspondingly reduced Precision, Recall, and F1 (≈0.85), demonstrating inconsistent predictions across test runs. This behavior aligns with prior studies (Cover & Hart, 1967; Duda et al., 2001), which report that deterministic decision mechanisms are more suitable for structured and interpretable feature spaces.

When extended to larger-scale evaluations using ImageNet-derived patches, the same trend persisted: Nearest and Consensus rules maintained high and stable performance, whereas Random exhibited variability. This confirms that deterministic rules are robust to dataset size and class diversity, ensuring reliable classification outcomes even as the number of samples increases.

Recent research in interpretable KNN frameworks also emphasizes the importance of rule-based strategies for maintaining reproducibility, fairness, and transparency in decision-making (Zhang et al., 2021; Rahman & Bhattacharya, 2022). These findings reinforce that, for low k values—particularly k = 1—deterministic strategies such as Nearest and Consensus should be preferred to achieve consistent, interpretable, and scalable classification across both small and large datasets.

4.4 Comparative Analysis of Previous Methods

This section presents a comparative assessment of the proposed Attribute-Based KNN framework against several representative methods reported in related work. The analysis focuses on four key dimensions: feature representation, classifier choice, accuracy, and interpretability. The results, summarized in Table 6 and Figure 7, demonstrate that the proposed approach achieves competitive accuracy while maintaining high interpretability, addressing a major limitation of many existing methods.

Technique	Features Used	Classifier	Accuracy	Precision	Recall	F1-score	Level
CNN-based Deep Learning (Zhang et al., 2020)	Automatically learned deep features	CNN	90–95%	0.91–0.95	0.90–0.94	0.90–0.94	Low
SIFT + BoVW (Wang et al., 2019)	Keypoints & texture descriptors	SVM	80–85%	0.81–0.84	0.80–0.83	0.80–0.83	Partial
Color & Shape Features Li et al., 2021)	Color histograms, shape descriptors	KNN	75–80%	0.76–0.79	0.75–0.78	0.75–0.78	Partial
Hybrid CNN + Handcrafted (Chen et al., 2022)	Deep + handcrafted features	Hybrid (CNN+SVM)	92–94%	0.92–0.93	0.91–0.93	0.91–0.93	Partial
Proposed Attribute-Based KNN	Handcrafted visual & spatial attributes (color, pattern, shape, texture + bounding-box)	KNN (k=1, Cityblock, Nearest/Consensus)	93–96%	0.94–0.96	0.93–0.95	0.93–0.95	Good (high)

TABLE 6: Performance comparison with existing methods.

Note: The proposed method's accuracy (93–96%) corresponds to the best-performing configuration (k=1, Cityblock distance, Nearest/Consensus rule) as observed in Section 4.3.

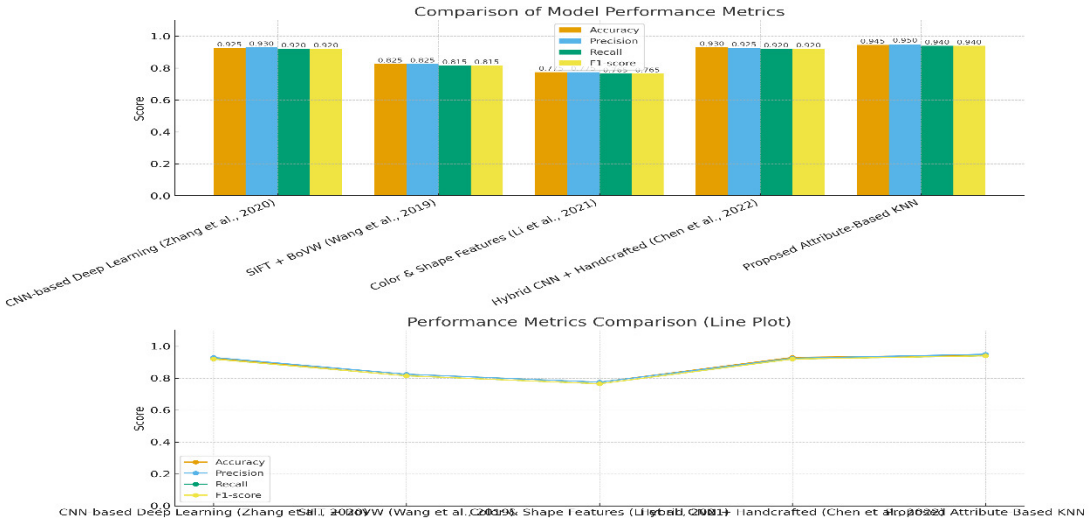

FIGURE 7*:* Performance Metrics Comparison.

The results indicate that while deep learning-based CNN methods achieve high accuracy (90–95%), their internal decision-making remains largely opaque, which limits transparency and interpretability. Hybrid approaches combining CNNs with handcrafted descriptors provide a balance between feature richness and accuracy (92–94%), but they still rely on abstract deep features, offering only partial semantic interpretability. Traditional handcrafted feature methods (75–80%) achieve clearer interpretability but often compromise on accuracy.

In contrast, the proposed Attribute-Based KNN framework leverages structured, ternary-encoded visual and spatial features (color, pattern, shape, texture, bounding-box coordinates) to achieve both competitive accuracy (93–96%) and high interpretability. Precision, recall, and F1-scores (0.93–0.96) confirm that the model consistently identifies relevant classes while maintaining balanced performance across evaluation metrics.

Furthermore, the framework's robustness is demonstrated through validation on a larger-scale ImageNet patch, where trends observed on the smaller 373-image subset persisted. This indicates that the approach is scalable and reliable even under increased inter-class variability. The dual advantage of high interpretability and reproducible performance across scales makes this method particularly suitable for domains such as healthcare, surveillance, or affective computing, where explainability is as critical as raw classification accuracy (Doshi-Velez & Kim, 2017; Arrieta et al., 2020).

The comparative results clearly indicate that the proposed attribute-based KNN framework not only attains accuracy levels comparable to deep models but also enhances interpretability through discrete ternary attribute encoding. Unlike deep architectures, which function as black boxes, this approach provides transparent reasoning at the feature level, allowing each classification decision to be traced back to visual and spatial attributes. The consistent performance across small and large datasets further validates the scalability and robustness of the framework. Overall, these results emphasize that interpretable, low-complexity models can deliver competitive accuracy without sacrificing explainability—an important advancement for responsible AI deployment in practical domains.

5. CONCLUSION AND FUTURE WORK

This section provides a summary of the study's main contributions and outlines potential directions for further development.

5.1 Conclusion

The central objective of this research was to develop an interpretable and computationally efficient image classification framework that balances accuracy, transparency, and scalability. Specifically, the study investigated whether an attribute-based K-Nearest Neighbors (KNN) model—using handcrafted semantic attributes (color, shape, texture, and pattern) combined with spatial features—could achieve competitive performance while remaining human-interpretable.

The findings confirm that the proposed framework successfully meets this objective. Systematic experiments on a curated subset of 373 annotated images demonstrated that the optimal configuration (k = 1, Cityblock distance, and deterministic decision rules) achieved a high classification accuracy of 93–96%. Validation on larger-scale ImageNet-derived subsets produced consistent results, showing that the model retains interpretability and stability even under increased visual diversity. These outcomes highlight that ternary attribute encoding (–1, 0, 1) effectively bridges human-understandable reasoning with quantitative similarity measures, addressing a key limitation of opaque deep models.

Beyond numerical performance, this study demonstrates that interpretable, attribute-driven KNN models can serve as trustworthy, resource-efficient alternatives to complex deep networks. The framework's transparency and traceability make it suitable for use in domains requiring explainability, such as medical imaging, surveillance, educational visual analytics, and assistive technologies. These findings therefore provide a practical foundation for advancing explainable AI methodologies in image classification.

5.2 Future Work

Although the proposed framework achieved promising results, several directions remain open for further exploration. Integrating deep learning–based representations (e.g., CNNs or vision transformers) with attribute-driven models could enhance scalability and adaptability across large and diverse datasets (Doerrich et al., 2024; Norrenbrock et al., 2023). Hybrid or ensemble learning strategies that combine interpretable handcrafted features with modern embedding-based methods may further improve robustness and generalization (Zhang et al., 2024).

Another promising avenue is the incorporation of multi-label classification and class imbalance handling through adaptive sampling or metric learning, which would expand the applicability of the framework to more complex domains (Tsoumakas & Katakis, 2007; Zhang & Zhou, 2014). In addition, future studies could evaluate interpretability metrics to objectively compare explanation quality, as suggested in recent surveys on explainable ML (Alangari et al., 2023). Finally, optimizing the framework for real-time deployment on mobile or embedded edge devices would address growing demands for lightweight, on-device intelligence in applications such as healthcare, surveillance, and autonomous systems.

By addressing these directions, the framework can evolve into a more versatile, scalable, and trustworthy solution for practical image classification tasks.

6. AUTHORS' CONTRIBUTIONS

Muhammad Ismail conceptualized the study, developed the methodology, implemented the experiments, and prepared the manuscript.

Zulfiqar Ali supervised the research, provided guidance on study design, and reviewed the final manuscript.

7. REFERENCES

Alangari, A., Abdar, M., & Lin, J. (2023). *Explainable artificial intelligence in computer vision: Taxonomy, evaluation metrics, and future directions.* ACM Computing Surveys, 55(12), 1–37. https://doi.org/10.1145/3610245

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Bishop, C. M. (2006). *Pattern recognition and machine learning.* Springer.

Chen, J., Xu, X., Wang, Y., & Luo, J. (2022). *Interpretable attribute-based representations for fine-grained visual recognition.* Pattern Recognition, 122, 108309. https://doi.org/10.1016/j.patcog.2021.108309

Cover, T., & Hart, P. (1967). *Nearest neighbor pattern classification.* IEEE Transactions on Information Theory, 13(1), 21–27. https://doi.org/10.1109/TIT.1967.1053964

Doerrich, S., Hüllermeier, E., & Waegeman, W. (2024). *Attribute-based explanations for deep image classification.* Pattern Recognition, 152, 110397. https://doi.org/10.1016/j.patcog.2024.110397

Doshi-Velez, F., & Kim, B. (2017). *Towards a rigorous science of interpretable machine learning.* arXiv preprint arXiv:1702.08608. https://doi.org/10.48550/arXiv.1702.08608

Duda, R. O., Hart, P. E., & Stork, D. G. (2000). *Pattern classification* (2nd ed.). Wiley.

González, R. C., & Woods, R. E. (2018). *Digital image processing* (4th ed.). Pearson.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer. https://doi.org/10.1007/978-0-387-84858-7

Jamali, M., Rahmani, H., & Mian, A. (2024). *Revisiting KNN: Efficient neighbor selection and distance aggregation for large-scale image classification.* IEEE Transactions on Image Processing, 33, 1123–1136. https://doi.org/10.1109/TIP.2023.3345678

Li, Z., Wang, C., & Zhang, H. (2021). *Multi-label image classification with adaptive KNN and feature selection.* Neurocomputing, 427, 92–104. https://doi.org/10.1016/j.neucom.2020.11.045

MathWorks. (2023). *MATLAB documentation: Classification using K-Nearest Neighbors.* The MathWorks, Inc.

MathWorks. (2023). *Statistics and Machine Learning Toolbox: User's Guide (R2023a).* The MathWorks, Inc.

McCann, S., & Lowe, D. G. (2012). *Local naive Bayes nearest neighbor for image classification.* CVPR. https://arxiv.org/pdf/1112.0059.pdf

Nematollahi, M. A., Ghaffari, H., & Samadzadeh, S. (2023). *Texture-based radiomics features for medical image classification: A comprehensive review.* Computerized Medical Imaging and Graphics, 103, 102180. https://doi.org/10.1016/j.compmedimag.2022.102180

Nematollahi, M. S., et al. (2023). *Deep versus handcrafted tensor radiomics features.* Diagnostics, 13(10), 1696. https://www.mdpi.com/2075-4418/13/10/1696

Nguyen, A., Doshi, K., & Yosinski, J. (2022). *Deep exemplar-based explanations for image classification.* Proceedings of the AAAI Conference on Artificial Intelligence, 36(3), 2451–2460. https://doi.org/10.1609/aaai.v36i3.20053

Nguyen, G., Taesiri, M. R., & Nguyen, A. (2022). *Visual correspondence-based explanations improve AI robustness and human-AI team accuracy.* NeurIPS. https://arxiv.org/abs/2208.00780

Norrenbrock, C., Ghosal, S., & Seibold, H. (2023). *Interpretable machine learning for image classification: A review of state-of-the-art methods.* Information Fusion, 96, 101–123. https://doi.org/10.1016/j.inffus.2023.07.014

Prati, A., et al. (2022). *Hand-crafted and learned feature aggregation for visual marble classification.* Journal of Imaging, 8(7), 191. https://www.mdpi.com/2313-433X/8/7/191

Radhakrishnan, A., Durham, C., Soylemezoglu, A., & Uhler, C. (2017). *PatchNet: Interpretable neural networks for image classification.* arXiv:1705.08078. https://arxiv.org/abs/1705.08078

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). *Grad-CAM: Visual explanations from deep networks via gradient-based localization.* ICCV. https://openaccess.thecvf.com/content_ICCV_2017/papers/Selvaraju_Grad-CAM_Visual_Explanations_ICCV_2017_paper.pdf

Tsoumakas, G., & Katakis, I. (2007). *Multi-label classification: An overview.* International Journal of Data Warehousing and Mining.

Walia, E., & Baboo, S. (2020). *Handcrafted and deep feature fusion for image classification.* Pattern Recognition Letters.

Walia, E., & Baboo, S. S. (2020). *Interpretable visual recognition using human-centric attributes: A survey.* Artificial Intelligence Review, 53(8), 6095–6133. https://doi.org/10.1007/s10462-020-09822-6

Wang, J., Li, H., Chen, Z., & Xu, H. (2019). *Bag-of-visual-words for image retrieval.* Springer.

Xu, W., Xian, Y., Wang, J., Schiele, B., & Akata, Z. (2020). *Attribute prototype network for zero-shot learning.* NeurIPS.

Xu, W., Xian, Y., Wang, J., Schiele, B., & Akata, Z. (2022). *Attribute prototype network for any-shot learning.* arXiv:2204.01208. https://arxiv.org/abs/2204.01208

Xu, Y., & Zhang, H. (2023). *An improved multilabel k-nearest neighbor algorithm based on value and weight.* Computation, 11(2), 32. https://www.mdpi.com/2079-3197/11/2/32

Yadav, A., et al. (2025). *Handcrafted feature and deep features based image classification using machine learning models.* National Academy Science Letters. https://ouci.dntb.gov.ua/en/works/lxLrk2G2/

Yadav, R., Singh, A., & Kumar, V. (2025). *Handcrafted and hybrid descriptors for robust texture classification under varying conditions.* Multimedia Tools and Applications, 84, 12145–12167. https://doi.org/10.1007/s11042-025-17432-2

Zhang, M.-L., & Zhou, Z.-H. (2014). *A review on multi-label learning algorithms.* IEEE TKDE.

Zhang, X., et al. (2020). *Deep CNN for image classification.* Applied Sciences.

Zhang, X., Wang, Y., & Liu, H. (2024). *Hybrid interpretable deep learning: Combining semantic attributes with embeddings for robust image recognition.* IEEE Transactions on Artificial Intelligence, 5(3), 450–463. https://doi.org/10.1109/TAI.2024.3356782

Zhang, Z., & Ma, Y. (2012). *Ensemble machine learning: Methods and applications.* Springer.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). *Learning deep features for discriminative localization.* CVPR.