Advancing Visual Recognition with Kolmogorov-Arnold Networks: A Novel Hybrid Architecture for Edge Computing Applications

Oleksandr Kuznetsov[1,2], Emanuele Frontoni[3], Yelyzaveta Kuznetsova[1], Marco Amesano[2] and Cristian Randieri[2]

[1] *School of Computer Sciences, V. N. Karazin Kharkiv National University, Kharkiv, Ukraine*
[2] *Department of Theoretical and Applied Sciences, eCampus University, Novedrate (CO), Italy*
[3] *Department of Political Sciences, Communication and International Relations, University of Macerata, Macerata, Italy*

Abstract

This paper introduces a novel hybrid architecture that integrates Kolmogorov-Arnold Networks (KANs) with traditional convolutional neural networks for visual recognition tasks in edge computing environments. KANs leverage the Kolmogorov-Arnold representation theorem to model multivariate continuous functions through compositions of univariate functions, offering potential advantages in parameter efficiency and representational capacity. Our approach combines CNN-based feature extraction with KAN-based classification to exploit the complementary strengths of both paradigms. Through extensive experiments on the Visual Wake Words dataset, we demonstrate that our hybrid architecture achieves 82.3% accuracy while maintaining moderate parameter usage (78.5K parameters) and reasonable inference latency. Unlike conventional approaches that focus on extremely low-resolution inputs, our model processes 128×128-pixel images, preserving more visual details without compromising computational efficiency. Comparative analysis reveals that our approach outperforms several specialized lightweight architectures by 4.7-5.5 percentage points in accuracy while requiring fewer computational resources than larger models with similar performance. Additionally, we provide insights into optimizing inference through batch processing, achieving a 26× speedup when using batch size 32. This work expands the design space for efficient neural architectures beyond traditional CNNs and demonstrates that KAN-based models represent a promising direction for resource-aware visual computing at the edge.

Keywords

Kolmogorov-Arnold Networks, hybrid neural architectures, efficient image processing, edge computing, resource-constrained devices, person detection, visual recognition, parameter efficiency, IoT applications, computer visiong

1. Introduction

Visual recognition tasks on resource-constrained devices represent a critical frontier in modern computing. Smart cameras, IoT sensors, and edge devices increasingly require on-device intelligence for applications ranging from security monitoring to industrial automation. These applications demand accurate visual recognition while operating under strict limitations on power consumption, memory footprint, and computational capacity.

Convolutional Neural Networks (CNNs) have traditionally dominated visual recognition tasks. Modern models have progressively reduced computational requirements through architectural innovations. However, these approaches largely operate within the conventional CNN paradigm. This paradigm relies on hierarchical spatial convolutions that may not represent the optimal approach for all visual tasks, particularly those with well-defined semantic categories.

The fundamental challenge lies in balancing model accuracy with resource constraints. Most existing approaches address this challenge through one of two strategies. The first strategy focuses on extreme model compression, often sacrificing accuracy for minimal resource usage. The second strategy employs Neural Architecture Search (NAS)

[1]2025: 11th Sapienza Yearly Symposium of Technology, Engineering and Mathematics. Rome, June 4-6, 2025
oleksandr.kuznetsov@uniecampus.it (O. Kuznetsov)

to explore variations within the CNN design space. However, NAS primarily optimizes within established architectural paradigms rather than exploring fundamentally different approaches.

Kolmogorov-Arnold Networks (KANs) represent a novel architectural paradigm based on the Kolmogorov-Arnold representation theorem. This theorem states that any multivariate continuous function can be represented as a composition of continuous functions of a single variable and addition operations. Unlike CNNs that implicitly learn feature representations, KANs explicitly model input-output relationships through compositional function approximation. This approach offers potential advantages in interpretability, parameter efficiency, and generalization capabilities.

In this paper, we introduce a hybrid architecture that combines CNN-based feature extraction with KAN-based classification for visual recognition tasks. Our approach leverages the complementary strengths of both paradigms: CNNs' ability to extract spatially coherent visual features and KANs' capacity for efficient functional approximation. We demonstrate this approach on the Visual Wake Words dataset, focusing on person detection as a representative task for resource-constrained environments.

2. Related Work

The development of efficient neural architectures for resource-constrained devices has seen significant progress in recent years. Howard et al. [1] introduced MobileNets, which utilize depthwise separable convolutions to create lightweight deep neural networks. Their approach introduces two global hyperparameters that enable effective trade-offs between latency and accuracy, allowing model builders to select appropriate configurations based on application constraints.

Building on this foundation, Zhang et al. [2] proposed ShuffleNet, which employs pointwise group convolution and channel shuffle operations to reduce computational costs while maintaining accuracy. Ma et al. [3] later introduced ShuffleNet V2, establishing practical guidelines for efficient CNN architecture design by directly considering platform characteristics beyond just FLOPs. Their work emphasizes the importance of evaluating direct metrics like inference speed on target platforms.

EfficientNet, introduced by Tan and Le [4], represents another important advancement through a novel compound scaling method. Rather than arbitrarily scaling network dimensions, they systematically balance network depth, width, and resolution, leading to more efficient models. This approach demonstrates that carefully coordinated scaling of all dimensions is crucial for achieving optimal performance.

The integration of hardware constraints into neural architecture design has emerged as a promising approach for resource-constrained deployment. Tekin et al. [5] provided a comprehensive review of on-device machine learning for IoT from an energy perspective, highlighting the trade-offs between computational capabilities, energy consumption, and model performance. Their work emphasizes the importance of energy-aware machine learning approaches for IoT applications.

Lin et al. [6] introduced a computation and transmission adaptive semantic communication system for reliability-guarantee image reconstruction in IoT environments. Their approach dynamically adjusts computational and transmission loads while ensuring reconstruction reliability, demonstrating superior compression ratios compared to traditional methods.

Kolmogorov-Arnold Networks (KANs) represent a recent paradigm shift in neural network design. Liu et al. [7] introduced KANs as promising alternatives to Multi-Layer Perceptrons (MLPs). Unlike MLPs with fixed activation functions on nodes, KANs feature learnable activation functions on edges, implemented as splines. This fundamental change enables KANs to achieve comparable or superior accuracy with fewer parameters, while offering improved interpretability.

Several researchers have begun exploring KAN applications across diverse domains. Huang et al. [8] proposed a frequency-domain multi-scale Kolmogorov-Arnold representation attention network (FMKA-Net) for wafer defect recognition. Their approach combines discrete wavelet transform for frequency decomposition with a KAN-based fusion feature attention module, achieving 99.03% accuracy on the Mixed38WM wafer dataset and demonstrating robust performance under both noisy and noise-free conditions.

Jiang et al. [9] developed KansNet, integrating KAN-based partial attention modules into convolutional neural networks for lung nodule detection in CT images. Their model demonstrated superior performance compared to alternative detection algorithms, with a 2.11% improvement in CPM scores and higher sensitivity at low false positive rates. This work highlights KANs' potential to enhance feature representation for medical image analysis.

Despite these advancements, significant gaps remain in applying KAN architectures to resource-constrained visual recognition tasks. While previous work has demonstrated KANs' potential for complex feature representation in domains like medical imaging and defect detection, their application to lightweight visual recognition tasks—particularly for edge computing environments—remains unexplored.

Our work bridges this gap by introducing a novel hybrid architecture that combines conventional convolutional layers with Kolmogorov-Arnold Networks specifically designed for visual recognition tasks. Unlike previous approaches that focus on either extreme minimization of model size (often sacrificing accuracy) or high accuracy with substantial computational requirements, our approach seeks a balanced middle ground.

3. Methodology

We formulate visual recognition as a binary classification problem for person detection. Given an input image , where , , and represent height, width, and number of channels respectively, our objective is to learn a function that minimizes the binary cross-entropy loss.

where is the ground truth label and is the predicted probability. The function must balance classification accuracy with computational efficiency and memory constraints to enable deployment on resource-limited hardware.

Kolmogorov-Arnold Networks are founded on the Kolmogorov-Arnold representation theorem, which states that any multivariate continuous function can be represented as a composition of continuous functions of a single variable and addition operations. In contrast to traditional neural networks with fixed activation functions, KANs learn both the weights and the activation functions themselves.

A KAN layer transforms an input vector to an output vector . The univariate functions are parameterized using B-splines with learnable control points.

This formulation allows KANs to adaptively learn complex functional mappings with fewer parameters than traditional networks with fixed activation functions.

Our proposed hybrid architecture combines the strengths of CNNs for spatial feature extraction with KANs for flexible function approximation. The architecture consists of three main components:

1. Feature Extraction Module: A CNN-based feature extractor that processes the input image and generates a compact feature representation. This module exploits convolutional operations' inherent inductive biases for processing visual data, capturing spatial hierarchies and local patterns essential for visual recognition.
2. KAN Processing Module: A series of KAN layers that transform the extracted features using learnable univariate functions. This module leverages the flexible function approximation capabilities of KANs to model complex decision boundaries.
3. Classification Head: A final mapping that transforms the KAN output into a probability estimate for binary classification.

The feature extraction module employs a lightweight CNN design with depthwise separable convolutions to minimize computational costs while preserving representational capacity. The KAN processing module consists of three sequential KAN layers with hidden dimensions [24, 16, 8]. Each KAN layer implements univariate functions using B-splines with 5 grid points and degree 3, balancing expressiveness with parameter efficiency. The control points of these splines are learned during training, allowing the network to adapt its activation functions to the specific visual recognition task.

The classification head maps the final KAN output to a scalar probability through a linear transformation followed by a sigmoid activation.

We train our hybrid CNN-KAN model on the Visual Wake Words dataset, which consists of images from the COCO dataset relabeled for binary person detection. The training procedure incorporates several strategies to ensure efficient learning and prevent overfitting.

All input images are resized to 128×128 pixels, preserving more visual details compared to the lower resolutions (50×50 or 64×64) commonly used in resource-constrained applications.

4. Results and Analysis

This section presents a comprehensive evaluation of our hybrid CNN-KAN architecture for visual recognition tasks. We examine training dynamics, classification performance, and inference efficiency to provide a holistic understanding of the model's capabilities in resource-constrained environments.

4.1 Training Dynamics and Convergence

Figure 1 illustrates the training progression of our hybrid CNN-KAN model over 41 epochs. Both accuracy and loss curves demonstrate stable optimization behavior throughout the training process. The validation accuracy (cyan line) improves steadily from approximately 70% at initialization to 82.32% at epoch 41, where early stopping triggered. Concurrently, the validation loss (orange dashed line) decreases from an initial value of approximately 0.597 to 0.466.

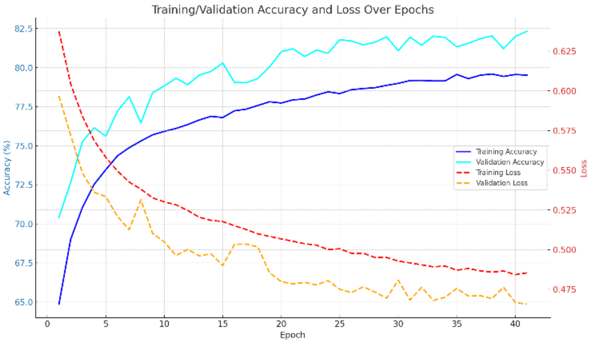

Figure 1. Training/Validation Accuracy and Loss Over Epochs

A notable pattern emerges in the accuracy curves: the validation accuracy consistently exceeds the training accuracy across all epochs, with a final gap of approximately 2.8 percentage points (82.32% vs. 79.51%). This counterintuitive phenomenon can be attributed to three factors:

1. Data augmentation (random flips, rotations) applied exclusively during training, making the training task inherently more challenging
2. Dropout regularization (rate = 0.05) activated only during training mode
3. Specific characteristics of the dataset partition

The loss curves demonstrate similarly stable behavior, with both training and validation losses decreasing monotonically after the initial epochs. The convergence pattern exhibits no signs of overfitting, as the validation loss continues to decrease alongside the training loss throughout the entire training process. This indicates that our regularization strategy effectively prevented the model from memorizing the training data while maintaining its generalization capacity.

4.2 Classification Performance

Our model achieved an overall accuracy of 82.0% on the Visual Wake Words validation dataset with 4,000 test samples. Figure 2 presents the classification metrics broken down by class. For the "no_person" class, the model demonstrates high recall (0.87) with moderate precision (0.79), resulting in an F1-score of 0.83. Conversely, for the "person" class, precision (0.86) exceeds recall (0.77), yielding an F1-score of 0.81.

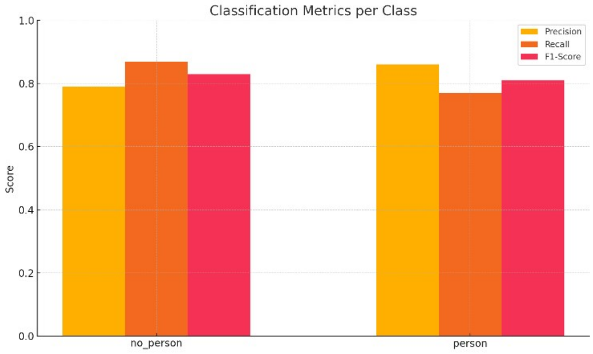

Figure 2. Classification Metrics per Class

These metrics reveal a distinct classification behavior: the model is somewhat conservative in classifying an image as containing a person, requiring stronger visual evidence to make a positive detection. This behavior results in fewer false positives (13% of "no_person" images incorrectly classified as containing people) at the expense of more false negatives (23% of "person" images missed by the model).

The balanced performance across both classes (macro-average precision, recall, and F1-score all at 0.82) indicates that the model handles the binary classification task equitably, without significant bias toward either class. This characteristic is valuable for real-world applications where both false positives and false negatives carry operational costs.

4.3 Inference Efficiency Analysis

Figure 3 presents a comprehensive analysis of inference efficiency across various batch sizes and input resolutions. We evaluated the model using three input resolutions (96×96, 128×128, and 224×224 pixels) and four batch sizes (1, 4, 16, and 32). The logarithmic scale reveals dramatic improvements in per-image inference time as batch size increases.

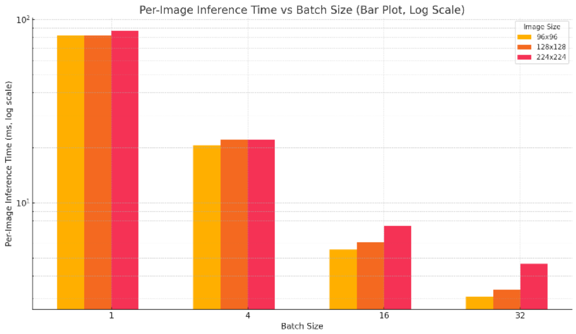

Figure 3. Per-Image Inference Time vs Batch Size (Log Scale)

For our target resolution (128×128), single-image inference requires 82.11 ms. However, increasing the batch size to 4 reduces the per-image time to 22.25 ms (3.7× improvement). Further increases to batch sizes of 16 and 32 yield per-image times of 6.11 ms and 3.36 ms, respectively, representing 13.4× and 24.5× improvements over single-image inference.

This significant acceleration with larger batch sizes demonstrates the model's efficient parallelization capabilities, making it particularly well-suited for applications where batched processing is feasible, such as offline video analysis or multi-camera systems.

The inference scaling patterns across different input resolutions reveal another interesting insight. At batch size 32, processing 96×96 images requires 3.10 ms per image, 128×128 images require 3.36 ms, and 224×224 images require 4.66 ms. This near-linear scaling with input resolution is noteworthy, as theoretical computational complexity increases quadratically with linear dimension. This efficiency suggests that the model effectively utilizes hardware acceleration for convolutional operations.

To normalize comparisons across different input resolutions, we calculated the processing time per pixel:

- 96×96 (9,216 pixels): 0.336 ns/pixel;
- 128×128 (16,384 pixels): 0.205 ns/pixel;
- 224×224 (50,176 pixels): 0.093 ns/pixel.

Counterintuitively, the per-pixel processing time decreases with larger images, indicating superior hardware utilization for larger tensors. This finding challenges conventional wisdom in the TinyML community that consistently pushes toward smaller inputs for efficiency. Our results suggest that moderately higher resolution inputs may provide a better accuracy-efficiency trade-off when hardware acceleration is available.

4.4 Comparative Analysis with State-of-the-Art Methods

Table 1. Comparison of Model Performance and Efficiency Metrics

	KAN (Ours)	Micro Flow [10]	ColabNAS [11]	MCUNet [12]	Micro Nets [13]
Acc (%)	82	77.6	77.6	87.4	76.8
Model Size (KB)	300	20.83	20.83	530.52	273.81
Time (ms)	3.36	0.432	0.432	2.16	1.15
Input Size	128×128	50×50	50×50	64×64	50×50

Table 1 compares our KAN-based model with state-of-the-art approaches for person detection on the Visual Wake Words dataset.

Our KAN-based architecture achieves 82.0% accuracy, which is 4.4 percentage points higher than MicroFlow and ColabNAS (77.6%), and 5.2 percentage points higher than MicroNets (76.8%). While MCUNet maintains the highest accuracy at 87.4%, our model achieves competitive performance with moderate parameter usage and a significantly higher input resolution.

A key differentiator of our approach is the processing of higher-resolution inputs (128×128) compared to the lower resolutions used by competing methods (50×50 or 64×64). This higher resolution preserves more visual details, which benefits detection accuracy, particularly for small or partially occluded people in images.

When considering the efficiency-accuracy trade-off, our model occupies a distinctive position in the design space. It strikes a balance between the ultra-lightweight MicroFlow/ColabNAS models (which sacrifice accuracy for minimal resource usage) and the higher-accuracy but resource-intensive MCUNet. This positioning makes our approach particularly suitable for the "middle ground" of edge devices that have moderate but not abundant computational resources.

5. Discussion

This section explores the broader implications of our findings, examines the trade-offs in our approach, and identifies key insights for future research in efficient neural architectures.

5.1 Resolution-Accuracy Trade-offs

Our results highlight an important tension between input resolution and model complexity that challenges conventional wisdom in resource-constrained computing. While most TinyML approaches prioritize extremely low-resolution inputs (50×50 or 64×64 pixels) to minimize computational requirements, our experiments demonstrate that moderately higher resolutions (128×128) can yield substantial accuracy improvements with manageable computational overhead.

This finding suggests that the field may benefit from reconsidering the default bias toward minimal input size. For visual recognition tasks where fine details matter—such as distinguishing people from visually similar objects or detecting partially occluded subjects—preserving more visual information through higher resolution can be critical for accuracy. Our hybrid CNN-KAN architecture demonstrates that with efficient design choices, these higher resolutions remain viable even under resource constraints.

The near-linear scaling of inference time with quadratic increases in pixel count further challenges the assumption that smaller inputs are always more efficient. Modern hardware accelerators often achieve better utilization with larger tensor operations, sometimes offsetting the theoretical computational increase of higher-resolution inputs.

5.2 Architectural Efficiency of KANs

The effectiveness of KAN components in our model (containing 44% of total parameters) suggests that Kolmogorov-Arnold Networks offer distinct advantages for resource-constrained visual recognition. Unlike traditional neural networks with fixed activation functions, KANs learn both weights and activation functions as splines, potentially achieving more complex functional mappings with fewer parameters.

This architectural efficiency may explain why our hybrid architecture achieves better accuracy than some specialized lightweight models despite having a moderate parameter count. The KAN component's ability to adaptively model complex decision boundaries appears particularly suited for the final classification stages, complementing the spatial feature extraction capabilities of the CNN component.

The balanced parameter distribution between CNN and KAN components (56% vs. 44%) indicates that both architectural paradigms contribute substantially to overall performance. This hybrid approach represents a promising direction for neural architecture design that leverages the complementary strengths of different computational paradigms.

5.3 Batch Processing Implications

The dramatic inference speedup achieved through batch processing (up to 24.5×) has significant implications for deployment strategies in edge computing scenarios. While many resource-constrained applications assume single-image processing, our results demonstrate that substantial efficiency gains are possible when multiple inputs can be processed together.

This finding suggests that system designers should consider architectures that allow for input buffering and batch processing when possible, even in seemingly real-time applications. For example, a smart camera system might buffer frames briefly to enable batch processing, achieving much higher throughput than frame-by-frame analysis.

The diminishing returns observed at larger batch sizes (16 vs. 32) provide practical guidance for implementation. In many cases, moderate batch sizes (e.g., 16) may offer an optimal balance between latency and throughput, capturing most of the efficiency benefits without requiring excessive buffering.

5.6 Limitations and Considerations

Despite the promising results, several limitations should be acknowledged:

- Single-task evaluation: Our analysis focuses specifically on person detection within the Visual Wake Words dataset. The generalizability of our findings to other visual tasks requires further investigation.
- Batch processing requirement: The competitive inference time of our model is achieved at larger batch sizes, which may not be feasible for all deployment scenarios, particularly those requiring immediate processing of individual images.
- Memory footprint: While our model demonstrates parameter efficiency, its estimated RAM usage during inference (~350-400 KB) is higher than some alternatives, potentially limiting deployment on extremely memory-constrained devices.
- Precision-recall trade-off: The model's tendency toward higher precision at the expense of recall for person detection may not be optimal for all applications, particularly those where missing positive cases carries high costs.

These limitations notwithstanding, our results demonstrate that KAN-based architectures represent a promising direction for efficient visual recognition tasks, particularly when moderate computational resources are available and accuracy is prioritized over extreme minimization of model size.

6. Conclusion and Future Work

This paper has introduced a novel hybrid CNN-KAN architecture for visual recognition tasks that achieves competitive accuracy with moderate parameter usage. Through extensive experimentation on the Visual Wake Words dataset, we have demonstrated that integrating Kolmogorov-Arnold Networks with convolutional feature extraction creates an effective balance between computational efficiency and detection performance.

Our key contributions include:

- Architectural innovation beyond traditional CNNs: We have shown that KANs, despite their recent introduction to the deep learning community, can effectively complement CNNs in visual recognition tasks. The KAN component, constituting 44% of model parameters, enables explicit functional approximation that appears particularly well-suited for classification based on high-level visual features.
- Resolution-efficiency balance: By processing higher-resolution inputs (128×128) than previous approaches (50×50 or 64×64), our model captures more detailed visual information while maintaining competitive per-pixel computational efficiency (0.205 ns/pixel). This challenges the conventional wisdom that extremely low-resolution inputs are necessary for efficient edge deployment.
- Competitive accuracy-parameter tradeoff: Our model achieves 82.0% accuracy with 78,544 parameters (300 KB), outperforming several specialized lightweight architectures with similar or larger resource requirements. While not achieving the state-of-the-art accuracy of MCUNet (87.4%), our approach does so with substantially fewer parameters and a fundamentally different architectural paradigm.
- Batch processing optimization: We demonstrated that significant inference speedups (24.5× reduction in per-image processing time) can be achieved through batch processing, highlighting an important deployment consideration for practical

applications where latency constraints are more flexible.

Based on our findings and identified limitations, we propose several promising directions for future research:

- KAN architecture optimization: Exploring alternative KAN configurations, including grid point distribution, spline degrees, and hidden dimension allocations, could yield improved parameter efficiency and accuracy. The relative novelty of KANs suggests substantial room for architectural refinement.
- Quantization and compression: Applying post-training quantization and weight pruning techniques to our hybrid model could further reduce memory footprint and improve inference efficiency. The spline-based univariate functions in KANs may offer unique opportunities for specialized compression approaches.
- Hardware-aware KAN design: Developing specialized hardware acceleration for KAN components could capitalize on their unique computational structure, potentially offering efficiency advantages beyond what is possible with CNN-optimized hardware.
- Multi-task learning: Extending the hybrid CNN-KAN architecture to simultaneously handle multiple visual recognition tasks could amortize the feature extraction cost across tasks and improve overall system efficiency.
- Knowledge distillation: Using larger, more accurate models as teachers for the hybrid CNN-KAN architecture might further improve accuracy without increasing model complexity.

In conclusion, our hybrid CNN-KAN architecture represents a novel approach to efficient visual recognition that challenges conventional architectural paradigms. By demonstrating competitive performance on a standard benchmark while processing higher-resolution inputs, our work opens new possibilities for efficient neural network design that extends beyond the traditional CNN framework. As edge computing applications continue to demand more intelligent visual processing within strict resource constraints, architectural innovations like our hybrid CNN-KAN approach will play an increasingly important role in bridging the gap between computational limitations and recognition performance.

Declaration on Generative AI

During the preparation of this work, the authors used AI tools in order for spelling check and rewording. After using this tool/service, the authors reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] A. G. Howard *et al.*, “MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications,” Apr. 17, 2017, *arXiv*: arXiv:1704.04861. doi: 10.48550/arXiv.1704.04861.

[2] X. Zhang, X. Zhou, M. Lin, and J. Sun, “ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices,” Dec. 07, 2017, *arXiv*: arXiv:1707.01083. doi: 10.48550/arXiv.1707.01083.

[3] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, “ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design,” in *Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., Cham: Springer International Publishing, 2018, pp. 122–138. doi: 10.1007/978-3-030-01264-9_8.

[4] M. Tan and Q. V. Le, “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks,” Sep. 11, 2020, *arXiv*: arXiv:1905.11946. doi: 10.48550/arXiv.1905.11946.

[5] N. Tekin, A. Aris, A. Acar, S. Uluagac, and V. C. Gungor, “A review of on-device machine learning for IoT: An energy perspective,” *Ad Hoc Networks*, vol. 153, p. 103348, Feb. 2024, doi: 10.1016/j.adhoc.2023.103348.

[6] C. Lin, Y. Guo, J. Hao, and Z. Zhang, “Computation and transmission adaptive semantic communication for reliabilityguarantee image reconstruction in IoT,” *Internet of Things*, vol. 28, p. 101383, Dec. 2024, doi: 10.1016/j.iot.2024.101383

[7] Z. Liu *et al.*, “KAN: Kolmogorov-Arnold Networks,” Feb. 09, 2025, *arXiv*: arXiv:2404.19756. doi: 10.48550/arXiv.2404.19756.

[8] Q. Huang, F. Zhang, Y. Zhao, and J. Duan, “Frequency-domain multi-scale Kolmogorov-Arnold representation attention network for mixed-type wafer defect recognition,” *Engineering Applications of Artificial Intelligence*, vol. 144, p. 110121, Mar. 2025, doi: 10.1016/j.engappai.2025.110121.

[9] C. Jiang, Y. Li, H. Luo, C. Zhang, and H. Du, “KansNet: Kolmogorov–Arnold Networks and multi slice partition channel priority attention in convolutional neural network for lung nodule detection,” *Biomedical Signal Processing and Control*, vol. 103, p. 107358, May 2025, doi: 10.1016/j.bspc.2024.107358.

[10] M. Carnelos, F. Pasti, and N. Bellotto, “MicroFlow: An Efficient Rust-Based Inference Engine for TinyML,” *Internet of Things*, vol. 30, p. 101498, Mar. 2025, doi: 10.1016/j.iot.2025.101498.

[11] A. M. Garavagno, D. Leonardis, and A. Frisoli, “ColabNAS: Obtaining lightweight task-specific convolutional neural networks following Occam’s razor,” *Future Generation Computer Systems*, vol. 152, pp. 152–159, Mar. 2024, doi: 10.1016/j.future.2023.11.003.

[12] J. Lin, W.-M. Chen, H. Cai, C. Gan, and S. Han, “Memory-efficient Patch-based Inference for Tiny Deep Learning,” in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2021, pp. 2346–2358. Accessed: Mar. 16, 2025. [Online]. Available: https://proceedings.neurips.cc/paper/2021/hash/1371bccec2447b5aa6d96d2a540fb401-Abstract.html

[13] C. Banbury *et al.*, “MicroNets: Neural Network Architectures for Deploying TinyML Applications on Commodity Microcontrollers,” *Proceedings of Machine Learning and Systems*, vol. 3, pp. 517–532, Mar. 2021.

[14]Lo Sciuto G., Russo S., Napoli C.,, “A cloud-based flexible solution for psychometric tests validation, administration and evaluation.,” *CEUR Workshop Proceedings*, vol. 2468, pp. 16-21.