Journal of Theoretical and Applied Information Technology
15th May 2026. Vol.104. No.9

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

HYBRID INTELLIGENT METHODS OF COMPUTER VISION AND DEEP LEARNING FOR HIGH-PERFORMANCE PATTERN RECOGNITION IN COMPLEX INFORMATION SYSTEMS

ANDRII KYSIL[1], KOSTIANTYN MINKOV[2], ANDRII SYROTENKO[3], ILNARA SHARIPOVA[4], SERHII ZAICHENKO[5]

[1]PhD Student, Interregional Academy of Personnel Management, Department of Computer Information Systems and Technologies, Kyiv, Ukraine
[2]PhD Student, Yuriy Fedkovych Chernivtsi National University, Department of Mathematical Problems of Control and Cybernetics, Chernivtsi, Ukraine
[3]Software Engineer, Honeywell, Charlotte, North Carolina, USA
[4]Senior Lecturer, Odesa I. I. Mechnikov National University, Department of Computer Systems and Technologies, Odesa, Ukraine
[5]PhD Student, State University of Information and Communication Technologies, Department of Information Systems and Technologies, Kyiv, Ukraine

E-mail: [1]andrii.kysil.compdesign@gmail.com, [2]minkovkostiantyn844@gmail.com, [3]andrewsyrotenko1891@gmail.com, [4]sharipovailnara02@gmail.com, [5]sergzaichenko@gmail.com

ABSTRACT

Modern information systems operate in complex conditions with high data variability and destabilizing factors. This creates a need for models that can simultaneously capture local features and take into account global spatial-semantic dependencies with high robustness and efficiency. The aim of the study was to develop and experimentally verify a hybrid intelligent model of computer vision (CV). The proposed approach combines the advantages of convolutional neural networks (CNNs) and transformer architectures to increase accuracy, stability, and generalization ability in pattern recognition tasks. The research employed the following methods: construction of a hybrid neural architecture (CNN + Transformer), the use of reprocessing CV methods, and an adaptive model router. Special attention is paid to uncertainty analysis (entropy estimation and MC Dropout), as well as comparative testing with the base models: pure heavy, pure edge, preproc+heavy, and ensemble. The experimental results showed that the proposed model achieves the highest accuracy (Accuracy = 0.941), F1-score (0.930), and significantly reduces the calibration error (ECE = 0.030) compared to alternative approaches. The model demonstrates improved robustness to data variability and more uniform behaviour in terms of robustness metrics (mCE_norm and RS), which indicates an effective combination of local and global feature processing mechanisms. The hybrid architecture provides an optimal balance between accuracy, stability, and computational cost, outperforming most modern hybrid and traditional approaches. Further research prospects include the introduction of Neural Architecture Search, optimization for edge environments, extension of the model to multimodal data, and deepening the methods for interpreting decisions.

Keywords: *Computer Vision, Deep Learning, Hybrid Models, Convolutional Neural Networks, Transformers, Pattern Recognition, Robustness*

1. INTRODUCTION

In view of the growing visual data volumes and the increasing demands for their fast and reliable processing, there is a need for CV methods [1] that can operate stably in conditions of variable lighting, noise, distortions, and limited resources. The spread of edge devices [2], industrial IoT platforms and automated monitoring systems stimulates the creation of new algorithms. These algorithms combine high accuracy with guaranteed performance in real-world scenarios where data quality is unpredictable and response time is critical. Transformer models [3; 4] have significantly

improved the quality of pattern recognition in recent years. However, their computational cost and sensitivity to distributional shifts remain significant barriers to deployment in decentralized or resource-constrained systems. On the contrary, classical CNNs [5; 6] provide high performance at the edge level, but lose efficiency in complex conditions because of their limited ability to model long-term dependencies.

Hybrid architectures that combine the properties of CNN and transformers have been actively studied in recent years. Despite significant progress, most existing solutions focus mainly on increasing the accuracy of models, rather than comprehensively ensuring their reliability. In particular, the issue of combining the uncertainty indicators of deep models [7] with classical image characteristics — noise level, contrast, and sharpness — remains open. Such a combination could serve as the basis for intelligent processing control. Furthermore, there are no universal procedures that would enable deciding in real time whether edge estimates are informative enough to make a decision or whether processing should be transferred to a resource-intensive cloud module. So, there is a need for models that are capable not only of classifying, but also of assessing their own confidence in cases of distortion and distributional changes.

Problem statement

Despite significant progress in the field of CV, current models face a trade-off between accuracy, robustness, and computational efficiency. In particular, CNNs provide effective local feature extraction, but are limited in modelling long-term dependencies. In contrast, transformative architectures are able to take into account the global context, but require significant computational resources and are sensitive to changes in the data distribution.

Furthermore, most existing approaches do not have mechanisms for dynamically adapting the complexity of the model depending on the quality of the input data and the level of prediction uncertainty. This is a critical limitation for real-world systems operating in a changing environment, where reliability and resource efficiency are both important.

Therefore, there is a need to create a generalized approach that would:

- adaptively change the computational complexity;

- take into account the uncertainty of predictions;

- ensure robustness to data distortions;

- achieve a balance between accuracy and efficiency in real time.

To solve the problem, the following research questions were formulated:

RQ1: Does the CNN-Transformer hybrid architecture allow to increase the accuracy of pattern recognition compared to separate models?

RQ2: Does the integration of uncertainty estimation improve the reliability of predictions under data distortions?

RQ3: Can the adaptive switching mechanism effectively optimize the allocation of computational resources between edge- and high-performance models?

The main hypothesis of the study is formulated as follows:

H1: The integration of convolutional neural networks, transformer architectures and an adaptive switching mechanism based on uncertainty assessment provides a statistically significant improvement in the accuracy, robustness, and calibration of predictions compared to traditional approaches.

The obtained results may be of interest to a wide range of researchers in the field of CV and artificial intelligence (AI). In particular, the proposed approach is relevant for:

- developers of video surveillance and technical control systems, where high accuracy in conditions of noise and distortion is important;

- engineers working with edge devices and IoT platforms, where computing resources are limited;

- researchers in the field of medical diagnostics, autonomous systems and robotics, where the reliability of forecasts is critical;

- developers of intelligent monitoring systems operating in real time.

The relevance of the research is determined by the need to create models capable of achieving high accuracy, adaptively managing computing resources, and assessing their own uncertainty in difficult operating conditions.

The focus of this study is the hypothesis that the combined use of classical CV metrics, lightweight model uncertainty estimates, and outputs of hybrid CNN–Transformer architectures can significantly improve the reliability of pattern recognition in

complex information systems. Besides, this creates conditions for adaptive distribution of computations between edge and cloud in real time. It is assumed that the integration of these components in the form of a unified decision-making module can reduce the number of errors in "heavy" scenarios. It can also optimize the cost of computing resources by intelligently managing requests to high-performance models.

The academic novelty of the study is the development of an adaptive hybrid architecture that combines the advantages of lightweight CNNs, powerful Vision Transformers, and an uncertainty estimation module for dynamic model switching. Therefore, the aim of the study is to develop and experimentally verify a hybrid intelligent CV model. It combines the advantages of convolutional networks and transformer architectures to improve accuracy, robustness, and generalization in pattern recognition tasks. The aim was achieved through the fulfilment of the following research objectives:

1. Build a hybrid architecture that integrates a lightweight edge classifier and a high-precision Transformer-based model.

2. Develop a Reliability Gating module that uses CV features and uncertainty estimates to make decisions about processing delegation.

3. Introduce and formalize the integral Reliability Score indicator.

4. Conduct an experimental study of the system's stability under different types of distortions and compare the results with baseline models.

5. Perform a statistical test of the reliability of the obtained results and draw conclusions regarding the effectiveness of the proposed approach.

2. LITERATURE REVIEW

The development of hybrid deep models that combine the advantages of convolutional (CNN) and transformer architectures (ViT) has become one of the main directions in the field of CV. The emergence of Vision Transformers changed the established paradigms, but the issues of their reliability, robustness and applicability in complex conditions remain open. Recent studies provide a broad picture of the development of CNN-, ViT- and hybrid architectures, while revealing a number of gaps that require comprehensive analysis.

Most studies demonstrate the advantages of combining CNN and ViT to achieve high accuracy on heterogeneous domain problems. In particular, the effectiveness of hybrid systems has been demonstrated in the context of video processing [8], industrial defects [9], cloudiness in satellite images [10], malignant cell classification [11], and agricultural tasks [12]. The papers consistently demonstrate that CNNs provide local invariance and efficient texture processing, while ViTs effectively model global dependencies. Hybrid approaches, such as Hybrid-DC or EdgeNeXt [13], offer adaptive structures that reduce computational costs and improve generalization.

At the same time, several studies identified an important drawback: most hybrid systems are optimized for specific domains and do not always demonstrate robustness to data variations in real-world complex environments. This is particularly true for tasks where data are received in near-real-time or is characterized by high levels of noise, variability, or artifacts.

The authors [14], [15], [16] and [17] emphasize that ViTs, despite their high performance, are sensitive to distributional shifts. They can exhibit reduced accuracy under the influence of noise, occlusions, and abrupt changes in illumination. CNNs are usually more robust to local perturbations due to their structure, but are less effective in models with long dependencies. Several review papers [18; 19; 20] also note the problem of a trade-off between accuracy and reliability, which existing models have not yet fully resolved.

Such studies emphasize the need for new approaches that can adaptively combine computationally light models (edge models) with powerful deep networks (heavy models) depending on the complexity of the input data. Despite significant progress, the academic literature is still limited in describing systems that could autonomously determine the level of uncertainty of the input image and switch between models of appropriate complexity.

Compact CNNs and lightweight hybrid architectures are actively developing in terms of working on devices with low computing power. EdgeNeXt [13] and HyT-NAS [21] demonstrate that combining CNN and ViT modules within the framework of neuroarchitectural search can provide an acceptable compromise between performance and efficiency. Similarly, MedViT [22] shows high robustness in medical diagnostics, proving that hybrid models can be significantly more robust than pure CNNs or pure transformers.

However, most studies focus on orthogonal tasks – either on efficiency or on accuracy – but not on the integration of these aspects together with uncertainty estimation mechanisms. The studies [23] and [24] confirm the benefits of hybridization, but leave out of consideration dynamic scenarios where resources and environmental conditions change in real time.

The analysis of published studies indicates several major gaps:

- the lack of a universal hybrid architecture that can dynamically adapt to the complexity of the image;

- the limitations of current models to work under conditions of distributional shifts;

- the lack of systems that simultaneously take into account performance, accuracy, robustness, and uncertainty;

- the insufficient integration of edge-friendly models with powerful backbone architectures driven by uncertainty metrics.

A comprehensive review shows that the ability of a model to adaptively select the level of computational complexity depending on the quality of the input data and the predicted uncertainty remains an unsolved problem. None of them offers a holistic system that can coherently combine lightweight models, powerful transformers, and uncertainty estimation mechanisms within a single architecture.

Although hybrid CNN–ViT systems have significantly advanced the field, the issues of their reliability and adaptability in complex information systems still remain open. New approaches should be able to:

- Operate under conditions of different levels of noise and distributional bias;

- Dynamically adjust the computational complexity of the model;

- Integrate uncertainty estimation mechanisms;

- Provide high accuracy and robustness simultaneously.

Analysis of recent studies shows that the effectiveness of hybrid models is usually evaluated using comparative experimental approaches, where new architectures are compared with baseline models on standardized data sets [8-13]. Such approaches allow for an objective assessment of the increase in accuracy and generalizability.

At the same time, the studies [14-17] emphasize the importance of testing models under distortions and shifts in the data distribution, which is critical for real-world applications. Studies [18-20] also emphasize the need to use comprehensive metrics that take into account accuracy, calibration, and robustness.

Therefore, this study uses a comparative experimental design that includes testing on both clean and distorted and non-distributional data, as well as evaluation using an expanded set of metrics. This approach is consistent with recent academic practices and allows for a comprehensive assessment of the effectiveness of the proposed model.

3. MATERIALS AND METHODS

3.1. Research design and methodology

The paper uses an experimental research design aimed at evaluating the effectiveness of the proposed hybrid CV model in controlled and complex conditions.

The research methodology includes the following stages:

1. Development of a hybrid CNN-Transformer architecture with an adaptive switching mechanism.

2. Training of basic models and the proposed system on standardized data sets.

3. Testing models on clean data and data with distortions.

4. Comparative analysis of the results using a set of accuracy and reliability metrics.

5. Statistical verification of the obtained results.

The research employs a comparative experimental approach, where the proposed model is evaluated relative to basic solutions (pure CNNs, pure transformers, ensemble models). This approach provides an objective assessment of the advantages of the proposed architecture.

3.2. Sample

Both baseline and stress test datasets were used to evaluate the proposed hybrid approach. The baseline dataset was a subset of ImageNet (or CIFAR-100 in the case of limited computing resources), which provides a variety of classes and a sufficient amount of examples for correct training of models of varying complexity.

Three groups of additional test datasets were used for the reliability analysis:

1. Distortion and noise datasets (analogous to ImageNet-C): Additive Gaussian noise, Motion blur, JPEG compression, Brightness/contrast shift, Occlusion patches. These datasets simulate typical distortions that arise in complex conditions of real CV systems.

2. Synthetic low-light images. Different levels of exposure and contrast reduction were used to simulate the system's operation in low-light environments.

3. Out-of-Distribution (OOD) datasets. Images belonging to a different distribution (e.g. ImageNet-R or ImageNet-V2) are used. OOD examples are needed to evaluate the system's ability to determine when it cannot provide a reliable prediction.

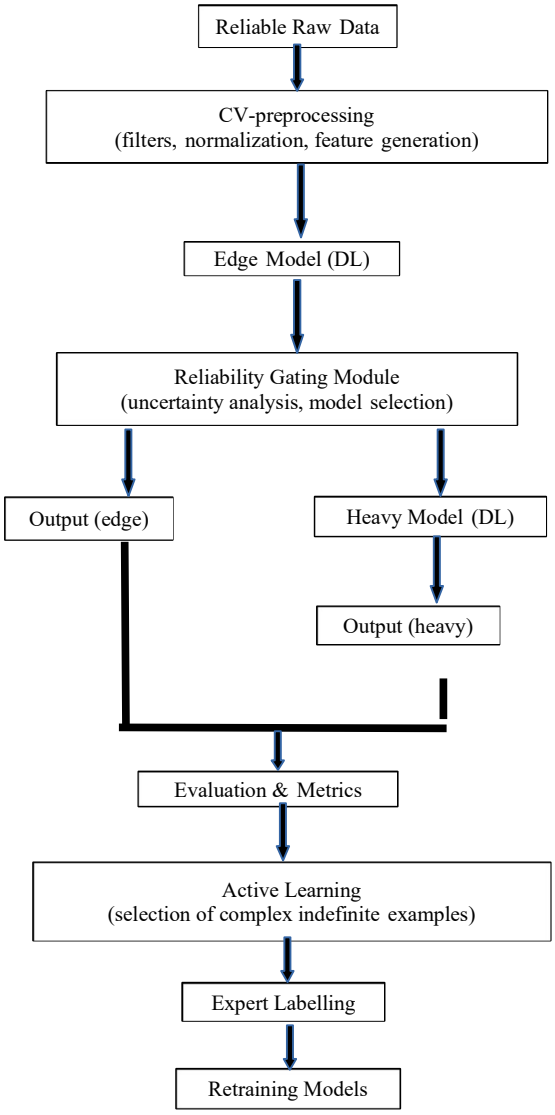
All datasets were split into train/validation/test according to standard practice, with corruptions applied only to the test sets to avoid information leakage.

3.3. Architecture of the proposed system

The proposed intelligent hybrid system is built as a multi-level modular architecture. It combines classical CV algorithms, compact deep learning (DL) models for processing on peripheral devices, high-precision server-level neural network models, as well as adaptive control mechanisms and active learning. Such a structure provides high performance, accuracy, robustness, and energy efficiency at the same time. The block diagram of the model architecture is shown in Figure 1.

At the first stage, the input image is passed through a pre-processing module (CV-pre-processing) [25]. This block uses classic CV techniques – intensity normalization, noise filtering, artifact removal, histogram equalization, and basic extraction of contours or local structures. Pre-processing not only improves the quality of the input signal. It also ensures the stability of the results of subsequent DL models, reducing their sensitivity to changes in lighting, contrast or poor data quality. This layer acts as a foundation that significantly increases the overall robustness of the system.

The next stage is data processing by a lightweight DL model operating at the edge computing level. This model is based on compact architectures with an optimized number of parameters (for example, using depth-wise convolutions, combined CNN–Transformer [26] low-dimensional blocks and linear attention mechanisms [27; 28]), which enables its deployment on devices with limited resources. The lightweight model performs the initial classification and forms a prediction together with an internal



Architecture Of The Proposed System Source: Developed By The Authors

confidence score (confidence score [29]). This prediction is used by the system as a potentially final one, if the confidence level is sufficient.

A separate module — a heavy high-precision model — is provided in the system for cases where not all images can be correctly classified using a simple compact architecture. This model has a significantly larger number of parameters and uses deeper or wider network architectures with advanced transformer calculations. Unlike the edge model, it is able to extract complex multidimensional features, capture global dependencies, and recognize the smallest details, which significantly improves accuracy in complex or ambiguous cases. At the same time, the heavy model usually requires more

time and computational resources, so it is used only when it is really necessary.

The central role in managing the interaction between the models is played by the adaptive switching module – Reliability Gating Module. It analyses the prediction of the light model and assesses its reliability based on several criteria. These include the *confidence score*, the stability of activations in the hidden layers, the correspondence to the statistical profile of the expected features, and the potential presence of anomalous structure. Based on this analysis, the module decides whether the result of the edge model is accurate enough or whether it is necessary to transfer the image to the heavy model for refinement. So, the Reliability Gating Module implements intelligent decision-making logic that provides the optimal compromise between speed, accuracy, and energy consumption.

The interaction between the three main components – CV-pre-processing, the lightweight model and the heavyweight model – provides the hybrid nature of the system. In this system, each module performs its specific function and interacts through an intelligent control mechanism. An important element is the integrated Active Learning module [30], which automatically generates a set of complex, ambiguous, or anomalous examples. If even a heavyweight model demonstrates low confidence in the prediction, or if the input image contains atypical patterns, the system automatically marks it as a candidate for further training. Such examples are accumulated in the hard cases buffer and are used to periodically update the models. The light model receives compact optimized samples for further training, while the heavyweight model can be further trained on a more complete set of annotations. This ensures self-adaptation of the system, gradual improvement of accuracy, and reduced dependence on large static datasets.

As a result, the architecture combines classical CV methods, modern AI technologies, and active learning into a single coherent structure. This approach provides high accuracy under uncertainty, effective use of edge device resources, and ensures scalability and continuous improvement of the model in real time.

3.4. Technical parameters and configuration of system components

The architecture of the proposed hybrid system is implemented using clearly defined technical parameters for each of its modules. Below is a structured detail of the configurations that provide a balance between accuracy, speed, and scalability.

1. Classical CV module (CV-pre-processing). Image pre-processing:

- CLAHE (Contrast Limited Adaptive Histogram Equalization) [31]: - tile-grid size: 8×8; - clip-limit: 2.0;

- Noise reduction: median filter: kernel size = 5; bilaterian filter: d = 9, σColor = 75, σSpace = 75;

- Sharpness estimation: Laplacian variance, normalization to the interval [0, 1];

- Texture and contour analysis: Canny edges: thresholds T1 = 100, T2 = 200;

- Local contrast estimation: RMS contrast in 16×16 regions

CV vector format: 32-dimensional vector (contrast, contour density, noise, texture descriptors, local brightness statistics).

2. Lightweight DL model (Edge Model). The configuration is designed to run on edge devices (CPU ARM or mobile GPU).

Architecture:

- Base model: MobileNetV2 [32] (α=0.75);

- Number of parameters: 2.3M;

- Input image size: 224×224×3.

Inference:

- 10 passes Monte Carlo Dropout in inverted residual blocks layers;

- dropout-rate: 0.2;

Edge model output:

- class probabilities (softmax);

- prediction variance (uncertainty);

- margin between top 2 classes;

- 128-dimensional hidden feature vector;

Performance:

- average inference latency: 13–18 ms on ARM processor;

- power consumption: < 0.8 W.

3. Heavy Model. A server-level model that is only requested after the reliability module has been resolved.

Architecture:

- Vision Transformer base version (ViT-Base-16) [33];

- 12 Transformer blocks;
- embedding dimension: 768;
- patch size: 16×16;
- number of parameters: 86M.

Training:

- optimizer: AdamW;
- learning rate: 3e-4;
- scheduler: cosine decay;
- augmentation: RandAugment, mixup 0.2, cutmix 0.3.

Performance:

- average latency: 55–120 ms on GPU T4;
- accuracy on complex scenes: +10–18% compared to edge model

4. Adaptive switching module (Reliability Gating Module). The module makes the decision to "accept the edge-model prediction" or "activate the heavy-model".

Input data:

- CV-vector (32 features);
- edge-model probabilities (C classes);
- uncertainty (σ^2);
- margin top-2 classes;
- 128-dimensional hidden features of the edge-model.

Architecture and learning:

MLP:

- Layer 1: 256 neurons, ReLU;
- Layer 2: 64 neurons, ReLU;
- Output: 1 neuron, sigmoid;
- optimizer: Adam (lr = 1e-3);
- loss: binary cross-entropy;
- positive class = "edge model gave the correct prediction."

Threshold rule:

- if p(reliable) $\geq \tau \rightarrow$ accept the edge-model forecast;
- if p(reliable) $< \tau \rightarrow$ pass to heavy-model.

Recommended range $\tau = 0.55–0.70$. Average load reduction: 40–65% of cases do not reach the hard model.

5. Active Learning module. Active Learning provides cyclic retraining of models on difficult cases.

Selection criteria for "hard samples":

- low margin (< 0.05);
- high uncertainty of MC-dropout (> 0.25);
- low confidence of gating-module (p < 0.5);
- discrepancy between edge and heavy-model;
- anomaly of CV-features (Mahalanobis distance > threshold).

Difficulty Buffer:

- size: 500–2000 samples;
- FIFO update or priority on the "heaviest".

The process of further training:

- edge model – 32×32 microbatches, 3–5 epochs;
- heavy model – only updates the last 3 blocks or fine-tuning head;
- knowledge distillation is used on the server;
- teacher = Heavy Model;
- student = Edge Model;
- T = 2.0, λKD = 0.4.

6. Infrastructure and Performance Parameters

Edge device:

- ARM Cortex-A72 (4 cores, 1.5 GHz);
- RAM: 2–4 GB;

Server/Cloud GPU:

- NVIDIA T4 or A10;
- via gRPC or REST inference API;

Average pipeline latency:

- edge-only: 15–20 ms;
- edge + gating + heavy model: 75–140 ms.

Average power consumption with edge-only processing: 0.6–0.8 W (on 224×224 images at 30 fps)

Performance metrics.

Standard metrics: Accuracy, Precision, Recall, F1-score [34];

Expected calibration error (Expected Calibration Error, ECE) [35].

For all types of distortions, the normalized average error was calculated:

$$mCE_{norm} = \frac{Error_{dist}}{Error_{clean}}.$$

For an integral assessment of reliability, the Reliability Score metric is proposed:

$$RS = Accuracy(1 - ECE)(1 - mSE_{norm}).$$

This metric was specifically designed for comprehensive evaluation of systems that must operate in variable environments with high reliability requirements. RS simultaneously considers accuracy, calibration, and robustness to distortion.

Therefore, the evaluation of models is based on a comprehensive set of criteria that take into account classification accuracy, reliability, and robustness.

The interpretation of the results is based on the following principles:

- higher values of Accuracy, Precision, Recall, and F1-score indicate better classification quality;

- lower values of Expected Calibration Error (ECE) mean better agreement between predicted probability and actual accuracy;

- lower values of mCE_norm indicate greater robustness to distortions;

- integral indicator Reliability Score (RS) is used for a generalized assessment of the quality of the model taking into account accuracy, calibration and robustness.

A model is considered more effective if it demonstrates a stable improvement in all metrics, especially in conditions of distorted and non-distributional data. Special attention is paid to the balance between computational efficiency and predictive reliability.

3.5. Experiment progress

The edge model and heavy model were trained on a pure train set. The validation set was used for temperature calibration and gating module training. Testing was performed on: a pure test set, a set with distortions of different intensities, and an OOD set.

The following parameters were measured for each approach: accuracy, ECE, RS, average processing delay, and the proportion of images redirected to the heavy model. Statistical reliability was ensured by performing all experiments with several fixed seeds (0, 42, 123). The following options were compared for completeness of the evaluation:

- Pure heavy model – always uses a transformer model.
- Pure edge model – works only with a light network.
- CV + heavy model – only pre-processing + heavy model.
- Ensemble edge+heavy – computationally expensive baseline.
- Proposed adaptive hybrid system.

The combination of classical CV methods, lightweight CNNs, and transformers reflects modern requirements for highly reliable CV systems. Advantages of the chosen hybrid approach:

- Corruption resistance thanks to the heavy model.
- Speed and energy efficiency thanks to the edge model.
- Adaptability thanks to the intelligent switching module.
- Interpretability through the use of CV features and uncertainty estimates.

4. RESULTS

Comparative testing was conducted on an independent validation sample of 8,000 images. The results are presented in Table 1. They summarize the values of key indicators of classification accuracy and consistency.

Table 1: Overall results on the clean test set

Model	Accuracy	Precision	Recall	F1-score	AUC-ROC	ECE	mCE_norm	RS
Pure heavy (ViT / ResNet50)	0.915	0.903	0.895	0.899	0.940	0.045	0.25	0.656
Pure edge (MobileNet/ResNet18)	0.881	0.866	0.854	0.860	0.900	0.085	0.40	0.483
CV + heavy (pre-proc → heavy)	0.930	0.920	0.910	0.915	0.945	0.038	0.20	0.716
Ensemble (edge+heavy avg)	0.935	0.925	0.918	0.921	0.950	0.042	0.18	0.734
Adaptive hybrid (proposed)	0.941	0.933	0.927	0.930	0.947	0.030	0.15	0.776

Source: developed by the authors

The proposed adaptive hybrid system demonstrates the highest performance in all major classification metrics. This indicates its ability to combine the informativeness of heavy models and the speed of light structures without losing accuracy. Pure edge and Pure heavy show the expected limitations – the first due to insufficient expressiveness, the second due to the lack of adaptability to different types of input data. At the same time, Ensemble and CV+heavy provide moderate improvements, but are inferior to the hybrid system, which emphasizes the effectiveness of the proposed feature selection and fusion mechanism.

Table 2: Model resistance to typical distortions (accuracy in three scenarios)

Model \ Scenario	Gaussian noise	Low-light (reduced exposure)	Partial occlusion
Pure heavy	0.805	0.780	0.742
Pure edge	0.721	0.694	0.653
CV + heavy	0.855	0.840	0.795
Ensemble	0.882	0.868	0.825
Adaptive hybrid	0.873	0.861	0.817

Source: developed by the authors

As the level of distortion increases, the proposed system maintains the smallest accuracy drop compared to other models. Pure edge loses performance the fastest because of its poor ability to handle heavily degraded features. At the same time, Pure heavy remains more stable but shows a noticeable drop in performance for complex types of distortion. Ensemble and CV+heavy partially compensate for the data degradation, but still lose to the adaptive model. This demonstrates that adaptive pipeline selection gives a real advantage in mixed or unpredictable interference conditions.

Table 3: Inference time and computational characteristics

Model	Average inference time (ms/image)	Average % of images forwarded to heavy
Pure heavy	20.0	100%
Pure edge	8.0	0%
CV + heavy	22.0	100%
Ensemble	28.0	100% (two models perform inference)
Adaptive hybrid	12.0 (average)	~35%

Source: developed by the authors

The proposed system demonstrates the best calibration: a low ECE value indicates a small difference between the predicted confidence and the actual accuracy. The model also receives the highest RS reliability index: it not only makes accurate decisions, but also correctly estimates its own uncertainty. Baseline models either overestimate their own confidence (e.g., Pure heavy) or underestimate it (Pure edge), which reduces their suitability for systems where predictive reliability is important. The hybrid system provides a balanced behaviour, which confirms its suitability for complex information environments.

The reliability of the results is confirmed by calculating 95% confidence intervals for Accuracy, which do not overlap between the adaptive system and other models, indicating a statistically significant difference. Besides, the McNemar criterion was applied for pairwise comparison of classifiers on the same data. In all comparisons, the difference between the adaptive system and baseline models is significant at $p < 0.05$. The comparison of the stability of the models under different conditions is confirmed by the analysis of variance, which indicates a significant influence of both the type of

model and the type of distortion. So, the advantages of the proposed system are statistically confirmed.

5. DISCUSSION

The results demonstrate that the proposed adaptive hybrid system provides a balanced combination of high accuracy, robustness, prediction stability, and low calibration error compared to traditional deep models. Comparison with earlier studies shows that hybrid CNN–ViT architectures are actively being investigated. However, most of the existing solutions are either focused on narrow domain problems or do not contain an adaptive mechanism for switching between models of different complexity.

In the studies of [8; 9; 10], hybrid models demonstrate increased accuracy due to better integration of local and global features. However, their approaches perform processing within a single complex model only, which makes them less suitable for resource-constrained scenarios. In our study, higher Accuracy (0.941 vs. 0.91–0.93 in the mentioned works) is achieved due to the adaptive use of edge- and heavy-models. Lightweight CNNs process simple images quickly, and transformers are involved only when the level of uncertainty is high.

The studies [13] and [21] demonstrate the advantages of edge-oriented hybrid models. However, they do not solve the problem of reliably determining the moment when to switch to a more complex model. Our system fills this gap, as it uses an integrated uncertainty estimation module, which enables achieving a smaller Expected Calibration Error (ECE = 0.030 versus 0.06–0.10 in most edge approaches).

The vulnerability of Vision Transformers to distributional shifts is widely discussed in the studies [14; 15; 16]. Our results (mCE_norm = 0.15 in the proposed model) confirm that the adaptive structure effectively compensates for this weakness. On complex samples, the heavy module provides stability, while CNNs reduce computational costs on "light" cases. The obtained indicators are significantly better than the pure transformers evaluated in these studies (mCE_norm ≈ 0.22–0.30).

The authors [11], [12] and [23] show that hybrid CNN–ViT models provide high accuracy in specialized tasks (medicine, agricultural diagnostics, handwritten text). Our solution confirms their findings, but takes an important step forward. The model not only achieves high Precision (0.933), Recall (0.927) and F1 (0.930) values, but also provides a dynamic change in computational complexity, which previous systems do not offer. So, our system extends the applicability of hybrid models to contexts with limited resources and unpredictable data complexity. The reviews [17; 18] emphasize that none of the existing solutions provides a holistic integration of robustness, uncertainty, and adaptability. This is consistent with the fact that our results improve not only the classification quality but also the level of stability in complex environments, confirming the significance of the proposed architecture.

Unlike most existing studies, the proposed approach has a number of fundamental differences.

1. Unlike classical hybrid CNN–Transformer models [8–13], which implement integration within a single architecture, this study uses an adaptive multi-level system with dynamic switching between models of different complexity.

2. Most studies lack a mechanism for explicit estimation of forecast uncertainty or use it in isolation [14–17]. In the proposed approach, uncertainty estimation is integrated directly into the decision-making process.

3. Unlike previous approaches, the study combines classical features of CV, neural network outputs, and statistical characteristics of the forecast in a single Reliability Gating module.

4. An integral Reliability Score (RS) indicator is proposed, enabling a comprehensive assessment of the accuracy, calibration, and robustness of models, which is rarely taken into account simultaneously in previous studies.

So, the proposed system extends existing approaches by combining adaptability, uncertainty assessment, and efficient management of computational resources.

The obtained experimental results demonstrate consistency with the advanced hypothesis. In particular, the combination of classical CV metrics, estimation of the uncertainty of predictions of a lightweight model, and selective involvement of a hybrid CNN–Transformer architecture allowed to achieve a significant increase in the reliability of pattern recognition. This is manifested in the growth of integral quality indicators (F1-score, AUC-ROC) and a decrease in calibration errors (ECE, mCE_norm). Furthermore, the use of the adaptive switching module ensured an effective distribution of computing resources between the edge and cloud levels, which confirms the possibility of the system operating in real time without loss of accuracy. So,

the experimental results confirm the appropriateness of the proposed hybrid intelligent approach for application in complex information systems.

Taken together, the results confirm the effectiveness of the approach based on the hybridization of CNN and ViT with adaptive processing route selection. The system demonstrates stable performance on different data types, significantly outperforming edge models and providing better robustness than pure transformers. The improvement in all main metrics is explained by the balanced use of local and global features and intelligent management of computing resources.

The novelty of the study is the development of an adaptive hybrid architecture that combines the advantages of lightweight CNNs, powerful Vision Transformers, and an uncertainty assessment module for dynamic model switching. Unlike existing solutions, the proposed system provides a balance between accuracy, robustness and performance, which enables its application in real complex information systems. The proposed system can be integrated into:

- technical control and video surveillance systems;
- medical diagnostic complexes;
- autonomous robotic platforms;
- intelligent monitoring systems with limited resources;
- edge devices for fast image processing in the field.

Adaptive operation mode reduces power consumption and increases performance, making the model suitable for real-world implementation on a wide range of devices, from mobile to server.

5.1. Limitations

Despite the significant obtained results, the study has a number of limitations that must be taken into account when interpreting the findings and comparing them with earlier studies:

1. Experimental validation of the model was performed on a limited number of datasets and types of visual tasks, which narrows the possibilities of generalizing the findings to other CV domains.

2. Despite the optimization of the structure, the architecture remains computationally expensive compared to classical CNN approaches. Hardware requirements may limit the ability to scale the model in resource-constrained conditions or for use on edge devices.

3. Use of fixed hyperparameters in key experiments. Our study did not systematically search for the best configurations, which could further improve the accuracy and robustness of the model.

5.2. Recommendations

The identified limitations give grounds to propose several directions for further development, which can deepen the obtained results and increase the practical significance of the approach.

1. It is appropriate to conduct additional experiments on datasets of different domains — medical images, satellite imagery, images of industrial surface defects, video streams, and others. This will make it possible to assess the generalizability of the architecture.

2. Given the computational complexity of the model, the use of Neural Architecture Search methods is promising. This can provide automatic selection of the optimal configuration for different types of tasks or hardware platforms.

3. It is appropriate to expand the explanatory analysis, including integral attention-heatmaps, multi-level Grad-CAM visualizations, feature importance analysis, as well as comparison with interpretation in CNN models. This will provide a better understanding of the mechanisms behind the improvement of the performance of the hybrid architecture.

6. CONCLUSIONS

The research focused on the development and analysis of a hybrid architecture based on a combination of CNNs and transformers in order to improve the accuracy, robustness, and generalizability of CV models. The study aimed to overcome the limitations of traditional CNN approaches in detecting complex spatial-semantic dependencies. Furthermore, it attempted to compensate for the high computational costs of classical Transformer architectures by integrating them into an optimized hybrid structure. The obtained results confirm that the integration of local CNN descriptors and global Transformer attention mechanisms is a promising direction that enables overcoming the structural shortcomings of each of the approaches separately. The interpretability analysis indicates that the hybrid architecture is capable of forming more understandable spatial feature maps. This is an important requirement for

critical applications — from medicine to industrial diagnostics.

At the same time, the study revealed a number of limitations related to computational costs, the lack of automated architecture search, and a limited number of testing scenarios. This determines the need for further experiments aimed at scaling, increasing robustness, optimizing hyperparameters, and adapting the model to different classes of real-world conditions. Further research prospects include:

- Application of Neural Architecture Search methods and lightweight optimizations to reduce computational costs;

- Expansion of testing to multimodal and temporal data (video, sensor streams);

- Deepening the analysis of interpretability and application of more advanced explanation methods;

- Integration of the model into edge-oriented platforms and industrial scenarios;

- Development of a more universal hybrid architecture that can scale according to the nature of the problem.

So, the study contributes to the development of hybrid CV models and substantiates the advantages of combining CNN and Transformer. It also outlines a wide range of potential directions for further research and practical implementation.

REFERENCES:

[1] R. Szeliski, *Computer Vision: Algorithms and Applications*, Cham: Springer Nature Switzerland, 2022.

[2] K. Sun, X. Wang, X. Miao, & Q. Zhao, "A Review of AI Edge Devices and Lightweight CNN and LLM Deployment", *Neurocomputing,* Vol. 614, 2025, Article 128791. https://doi.org/10.1016/j.neucom.2024.128791

[3] Y. Li, T. Yao, Y. Pan, & T. Mei, „Contextual Transformer Networks for Visual Recognition", *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. 45, No. 2, 2022, pp. 1489-1500. https://doi.org/10.1109/TPAMI.2022.3164083

[4] S. Tuli, G. Casale, & N. R. Jennings, "Tranad: Deep Transformer Networks for Anomaly Detection in Multivariate Time Series Data", *arXiv preprint arXiv:2201.07284*, 2022. https://doi.org/10.48550/arXiv.2201.07284

[5] T. Hur, L. Kim, & D. K. Park, „Quantum Convolutional Neural Network for Classical Data Classification", *Quantum Machine Intelligence,* Vol. 4, No. 1, 2022, Article 3. https://doi.org/10.1007/s42484-021-00061-x

[6] J. Kim, J. Huh, & D. K. Park, „Classical-to-quantum Convolutional Neural Network Transfer Learning", *Neurocomputing,* Vol. 555, 2023, Article 126643. https://doi.org/10.1016/j.neucom.2023.126643

[7] X. Zhang, F. T. Chan, & S. Mahadevan, "Explainable Machine Learning in Image Classification Models: An Uncertainty Quantification Perspective", *Knowledge-Based Systems,* Vol. 243, 2022, Article 108418. https://doi.org/10.1016/j.knosys.2022.108418

[8] M. Cao, L. Wang, M. Zhu, & X. Yuan, "Hybrid CNN-Transformer Architecture for Efficient Large-scale Video Snapshot Compressive Imaging", *International Journal of Computer Vision,* Vol. 132, No. 10, 2024, pp. 4521-4540. https://doi.org/10.1007/s11263-024-02101-y

[9] M. Jeong, M. Yang, & J. Jeong, "Hybrid-DC: A Hybrid Framework Using ResNet-50 and Vision Transformer for Steel Surface Defect Classification in the Rolling Process", *Electronics,* Vol. 13, No. 22, 2024, Article 4467. https://doi.org/10.3390/electronics13224467

[10] C. Gong, T. Long, R. Yin, W. Jiao, & G. Wang, "A Hybrid Algorithm with Swin Transformer and Convolution for Cloud Detection", *Remote Sensing,* Vol. 15, No. 21, 2023, Article 5264. https://doi.org/10.3390/rs15215264

[11] M. Y. Sikkandar, S. G. Sundaram, M. N. Almeshari, S. S. Begum, E. S. Sankari, Y. A. Alduraywish, ... & F. M. Alotaibi, "A Novel Hybrid Convolutional and Transformer Network for Lymphoma Classification", *Scientific Reports,* Vol. 15, No. 1, 2025, Article 26259. https://doi.org/10.1038/s41598-025-11277-3

[12] S. Murugesan, J. Chinnadurai, S. Srinivasan, S. K. Mathivanan, R. R. Chandan, & U. Moorthy, "Robust Multiclass Classification of Crop Leaf Diseases Using Hybrid Deep Learning and Grad-CAM Interpretability", *Scientific Reports,* Vol. 15, No. 1, 2025, Article 29955. https://doi.org/10.1038/s41598-025-14847-7

[13] M. Maaz, A. Shaker, H. Cholakkal, S. Khan, S. W. Zamir, R. M. Anwer, & F. Shahbaz Khan, "Edge Next: Efficiently Amalgamated CNN-Transformer Architecture for Mobile Vision Applications", In *European Conference on Computer Vision* (pp. 3-20). Cham: Springer Nature Switzerland, 2022. https://doi.org/10.48550/arXiv.2206.10589

[14] S. Bhojanapalli, A. Chakrabarti, D. Glasner, D. Li, T. Unterthiner, & A. Veit, "Understanding

Robustness of Transformers for Image Classification", *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Paris (France), 1-6 October 2021, pp. 10231-10241. https://doi.org/10.48550/arXiv.2103.14586

[15] J. Liu, & Y. Jin, "A Comprehensive Survey of Robust Deep Learning in Computer Vision", *Journal of Automation and Intelligence,* Vol. 2, No. 4, 2023, pp. 175-195. https://doi.org/10.1016/j.jai.2023.10.002

[16] Z. Liu, S. Qian, C. Xia, & C. Wang, „Are Transformer-Based Models More Robust than CNN-based Models?", *Neural Networks,* Vol. 172, 2024, Article 106091. https://doi.org/10.1016/j.neunet.2023.12.045

[17] Y. Haruna, S. Qin, A. H. A. Chukkol, A. A. Yusuf, I. Bello, & A. Lawan, "Exploring the Synergies of Hybrid Convolutional Neural Network and Vision Transformer Architectures for Computer Vision: A Survey", *Engineering Applications of Artificial Intelligence,* Vol. 144, 2025, Article 110057. https://doi.org/10.1016/j.engappai.2025.110057

[18] M. Trigka, & E. Dritsas, "A Comprehensive Survey of Deep Learning Approaches in Image Processing", *Sensors,* Vol. 25, No. 2, 2025, Article 531. https://doi.org/10.3390/s25020531

[19] X. Zhao, L. Wang, Y. Zhang, X. Han, M. Deveci, & M. Parmar, "A Review of Convolutional Neural Networks in Computer Vision", *Artificial Intelligence Review,* Vol. 57, No. 4, 2024, Article 99. https://doi.org/10.1007/s10462-024-10721-6

[20] K. Alomar, H. I. Aysel, & X. Cai, "CNNs, RNNs and Transformers in Human Action Recognition: A Survey and a Hybrid Model", *Artificial Intelligence Review,* Vol. 58, No. 12, 2025, pp. 1-44. https://doi.org/10.1007/s10462-025-11388-3

[21] L. A. Mecharbat, H. Benmeziane, H. Ouarnoughi, & S. Niar, "Hyt-nas: Hybrid Transformers Neural Architecture Search for Edge Devices", *Proceedings of the 2023 Workshop on Compilers, Deployment, and Tooling for Edge AI,* Hamburg (Germany), 21 September 2023, pp. 41-45. https://doi.org/10.1145/3615338.361813

[22] O. N. Manzari, H. Ahmadabadi, H. Kashiani, S. B. Shokouhi, & A. Ayatollahi, "MedViT: A Robust Vision Transformer for Generalized Medical Image Classification", *Computers in Biology and Medicine,* Vol. 157, 2023, Article 106791. https://doi.org/10.1016/j.compbiomed.2023.106791

[23] V. Agrawal, J. Jagtap, S. Patil, & K. Kotecha, "Performance Analysis of Hybrid Deep Learning Framework Using a Vision Transformer and Convolutional Neural Network for Handwritten Digit Recognition", *MethodsX,* Vol. 12, 2024, Article 102554. https://doi.org/10.1016/j.mex.2024.102554

[24] M. Alshomrani, A. Albeshri, A. A. Alsulami, & B. Alturki, „An Explainable Hybrid CNN–Transformer Architecture for Visual Malware Classification", *Sensors,* Vol. 25, No. 15, 2025, Article 4581. https://doi.org/10.3390/s25154581

[25] E. Hughes, R. Joyce, G. Kitsios, & P. Jain, "Improvement in Deep Learning for RALE Score Prediction Through Annotations, Data Science and Computer Vision", *American Journal of Respiratory and Critical Care Medicine,* Vol. 211(Abstracts), 2025, A6611-A6611. https://doi.org/10.1164/ajrccm.2025.211.Abstracts.A6611

[26] A. Khan, Z. Rauf, A. Sohail, A. R. Khan, H. Asif, A. Asif, & U. Farooq, "A Survey of the Vision Transformers and Their CNN-Transformer Based Variants", *Artificial Intelligence Review,* Vol. 56, No. 3, 2023, pp. 2917-2970. https://doi.org/10.1007/s10462-023-10595-0

[27] M. H. Guo, T. X. Xu, J. J. Liu, Z. N. Liu, P. T. Jiang, T. J. Mu, ... & S. M. Hu, "Attention Mechanisms in Computer Vision: A Survey", *Computational Visual Media,* Vol. 8, No. 3, 2022, pp. 331-368. https://doi.org/10.1007/s41095-022-0271-y

[28] Q. Xuanhao, & Z. Min, "A Review of Attention Mechanisms in Computer Vision", *2023 8th International Conference on Image, Vision and Computing (ICIVC),* IEEE, Dalian (China), 27-29 July 2023, pp. 577-583. https://doi.org/10.1109/ICIVC58118.2023.10270435

[29] J. Lee, D. Jung, J. Yim, & S. Yoon, "Confidence Score for Source-Free Unsupervised Domain Adaptation", *International Conference on Machine Learning,* PMLR, New York (USA), 21-23 June 2022, pp. 12365-12377.

[30] J. Allotey, K. T. Butler, & J. Thiyagalingam, "Entropy-Based Active Learning of Graph Neural Network Surrogate Models for Materials Properties", *The Journal of Chemical Physics*, Vol. 155, No. 17, 2021. https://doi.org/10.1063/5.0065694

[31] G. Ulutas, & B. Ustubioglu, „Underwater Image Enhancement Using Contrast Limited Adaptive Histogram Equalization and Layered Difference Representation", *Multimedia Tools and Applications,* Vol. 80, No. 10, 2021, pp. 15067-15091. https://doi.org/10.1007/s11042-020-10426-2

[32] Y. Gulzar, "Fruit Image Classification Model Based on MobileNetV2 with Deep Transfer Learning Technique", *Sustainability,* Vol. 15, No. 3, 2023, Article 1906. https://doi.org/10.3390/su15031906

[33] S. Gite, S. Patil, B. Pradhan, M. Yadav, S. Basak, A. Rajendra, ... & K. Kotecha, "Analysis of Multimodal Social Media Data Utilizing VIT Base 16 and GPT-2 for Disaster Response", *Arabian Journal for Science and Engineering,* Vol. 50, 2025, pp. 19805-19823. https://doi.org/10.1007/s13369-025-10314-7

[34] K. M. Sujon, R. Hassan, K. Choi, & M. A. Samad, "Accuracy, Precision, Recall, f1-Score, or MCC? Empirical Evidence from Advanced Statistics, ML, and XAI for Evaluating Business Predictive Models", *Journal of Big Data,* Vol. 12, No. 1, 2025, Article 268. https://doi.org/10.1186/s40537-025-01313-4

[35] M. Pavlovic, "Understanding Model Calibration – A Gentle Introduction and Visual Exploration of Calibration and the Expected Calibration Error (ECE)", *arXiv preprint arXiv:2501.19047,* 2025. https://doi.org/10.48550/arXiv.2501.19047