The object of this study is the classification of low-resolution and multi-class images, represented by the CIFAR-10 benchmark dataset. It is challenging to accurately classify low-resolution and multi-class images because traditional CNNs usually have trouble identifying both global and complex texture patterns. To address this issue, this study employs the CIFAR-10 dataset as a representative benchmark for real-world scenarios where image quality is limited, such as in low-cost medical imaging, remote sensing, and security surveillance systems. The limited discriminability of traditional CNNs in these situations is the primary issue addressed. The proposed method employs three parallel convolutional streams with distinct kernel sizes (3×3, 5×5, and 7×7) to capture hierarchical spatial patterns, followed by the integration of two attention mechanisms – squeeze-and-excitation and convolutional block attention module – that adaptively emphasize the most relevant spatial and channel-wise information. In addition, structural texture descriptors such as Gray-level co-occurrence matrix, local binary pattern, and Gabor filters are computed independently and later fused with the deep representations to enrich the feature space. Experiments were carried out on the CIFAR-10 dataset under varying levels of class complexity: 10, 5, and 3 categories. The results reveal that the hybrid approach significantly improves precision, recall, and F1-score across all scenarios, with the highest accuracy of 90.87% obtained when only three classes are involved. These improvements are explained by the complementary nature of deep and handcrafted features, which together enable the model to learn both global semantics and fine-grained local textures can achieve higher classification accuracy, improved reliability, and reduced misclassification errors, ultimately enhancing the effectiveness of applications ranging from medical decision support to intelligent surveillance

Keywords: multi-scale kernel, attention mechanisms, CIFAR-10, GLCM, LBP, Gabor filters

UDC 004.932:004.852
DOI: 10.15587/1729-4061.2025.331524

A HYBRID MULTI-SCALE CONVOLUTION NEURAL NETWORK WITH ATTENTION AND TEXTURE FEATURES FOR IMPROVED IMAGE CLASSIFICATION

Irpan Adiputra Pardosi
Doctoral Student of Computer Science,
Lecturer of Computer Science*
Department of Computer Science
Universitas Mikroskil
Thamrin str., 112, Medan, Indonesia, 20212
Tengku Henny Febriana Harumy
Corresponding Author
Doctor of Computer Science*
E-mail: hennyharumy@usu.ac.id
Syahril Efendi
Doctor of Mathematics, Professor*
*Department of Computer Science
Universitas Sumatera Utara
Dr. T. Mansur str., 9, Sumatera Utara, Indonesia, 20155

Received 04.06.2025
Received in revised form 29.08.2025
Accepted date 08.09.2025
Published date 31.10.2025

How to Cite: *Pardosi, I. A., Harumy, T. H. F., Efendi, S. (2025). A hybrid multi-scale convolution neural network with attention and texture features for improved image classification. Eastern-European Journal of Enterprise Technologies, 5 (2 (137)), 18–28. https://doi.org/10.15587/1729-4061.2025.331524*

1. Introduction

Image classification has become a cornerstone of modern computer vision, underpinning applications in medical diagnostics, autonomous driving, remote sensing, industrial inspection, and security surveillance [1–3]. The rapid growth of visual data in today's digital era has created a pressing need for classification systems that are both accurate and efficient, capable of handling real-world conditions such as noise, low resolution, intra-class variability, and complex backgrounds [4]. In practice, the ability to automatically and reliably classify images has direct societal and economic benefits, including faster medical decision-making, safer transportation systems, more precise environmental monitoring, and enhanced security infrastructures [5, 6].

At the same time, current conditions highlight persistent challenges that make ongoing research in this field highly relevant. Images encountered in practical deployments often vary in scale, orientation, and texture, requiring models that can generalize beyond controlled benchmark settings [7]. Furthermore, resource-constrained environments, such as embedded medical devices or real-time traffic monitoring, demand classification models that strike a balance between accuracy and computational efficiency [8]. Another factor reinforcing the relevance of this scientific topic is the growing demand for trustworthy and interpretable artificial intelligence. In domains such as healthcare and autonomous systems, classification models must not only achieve high accuracy but also provide reliable and explainable outputs that decision-makers can confidently use [9]. The pursuit of models that can adapt to data complexity, remain computationally efficient, and support explainable decision-making ensures that research on image classification continues to have significant societal and industrial value [10–12].

Therefore, in modern conditions, research on image classification remains both timely and necessary. From a scientific perspective, it advances the development of robust and generalizable models. From a practical standpoint, it directly supports the creation of systems that improve safety, efficiency, and decision quality across multiple sectors. This dual importance demonstrates the ongoing relevance of image classification as a central topic in computer vision research [13–15]. Given the increasing demand for accurate and efficient classification systems in real-world applications – especially in domains with low-resolution data and high inter-class similarity – research into hybrid models that

combine deep learning with attention mechanisms and classical feature extraction remains highly relevant. Therefore, research on the development of hybrid models that integrate multi-scale CNNs, attention mechanisms, and handcrafted texture features is relevant and timely to address the limitations of existing image classification approaches.

2. Literature review and problem statement

Convolutional neural networks (CNNs) have achieved remarkable success due to their ability to learn hierarchical feature representations from raw pixel data [1–3]. In the study [3], the authors proposed an efficient brain tumor detection and classification model using pre-trained convolutional neural networks (CNNs), specifically ResNet50 and EfficientNet, combined with data augmentation techniques and segmentation via GrabCut. The model achieved high accuracy (up to 98%) and demonstrated improved precision and recall compared to traditional methods. While the approach effectively automates tumor detection using MRI scans, it relies heavily on conventional transfer learning and fixed CNN architectures. One limitation is the absence of domain-specific feature refinement or attention mechanisms that can focus on clinically relevant tumor regions, especially in heterogeneous data. Furthermore, the study does not explore adaptability to multi-modal or longitudinal medical data, which are crucial in real-world diagnosis. To extend this research, the integration of lightweight attention-enhanced CNNs or hybrid models with transformer components could improve feature localization and enable dynamic adaptation to varied imaging conditions and tumor subtypes.

However, challenges remain – particularly in capturing multi-scale patterns and contextual relationships, which are critical in datasets with high intra-class variability and low inter-class separability, such as CIFAR-10 [4, 5]. In such datasets, objects may appear in diverse shapes, sizes, and orientations within the same class, making conventional CNNs with fixed kernel sizes less effective due to their limited receptive field diversity.

In the study [6], a hybrid ECG classification model combining multi-branch CNNs and Transformer modules was proposed to jointly learn morphological and temporal features. The method demonstrated high accuracy (up to 99.5%) on the MIT-BIH database using RR interval features and multi-protocol evaluation. Despite the strong results, the model does not incorporate real-world noisy signals (e.g., mobile ECG or wearable devices), and lacks robustness testing across diverse patient demographics or clinical environments. In addition, the transformer block used is relatively shallow, and the fusion process between CNN and transformer is static. Future work could focus on adaptive fusion using gated mechanisms or attention-guided RR interval embedding to improve generalizability and robustness. To address this, multi-scale CNN architectures have been proposed, incorporating kernels of varying sizes to extract spatial features at different resolutions [6, 7]. For example, the Inception architecture and EfficientNet variants [8–10] demonstrated that combining multiple receptive fields enhances object recognition. In parallel, attention mechanisms – notably Squeeze-and-Excitation (SE) [11] and convolutional block attention module (CBAM) [12] have been introduced to dynamically recalibrate feature maps, helping models focus on salient regions without increasing network depth.

In the study [13], a classification model for breast ultrasound images was developed using GoogLeNet combined with Total Variation (TV) denoising. The use of GoogLeNet improved spatial feature extraction, while TV denoising helped reduce image noise. However, the model suffered from limited receptive fields and lacked a mechanism to distinguish semantically significant areas in the image. Furthermore, the study did not integrate multi-scale or attention-based architectures, resulting in less adaptability when handling lesions of varying size and texture complexity. Integrating multi-scale CNN layers and attention modules such as SE or CBAM could better address fine-grained visual cues critical in medical diagnosis.

While these approaches have been effective independently, few studies have investigated their combined effect in a unified model. Moreover, most recent deep learning models overlook handcrafted texture features, which still hold value, especially in domains involving subtle texture differences, such as medical imaging [13], remote sensing [14], and low-resolution object classification [15, 16]. Descriptors such as Gray-level co-occurrence matrix (GLCM), local binary patterns (LBP), and Gabor filters offer statistical and frequency-based information that complements learned features from CNNs. Integrating these handcrafted features can improve the robustness and interpretability of classification models [8, 17].

In the study [17], the authors evaluated face gender recognition by comparing handcrafted features (e.g., HOG, LBP), deep learned features (CNN), and fused combinations using both SVM and CNN classifiers. While the study provided an extensive comparative framework, it primarily focused on evaluating static feature sets and did not leverage attention mechanisms to enhance spatial feature localization. Additionally, the model's performance was dataset-sensitive, with significant performance drops in uncontrolled environments due to lighting and pose variations. Furthermore, no semantic or context-aware mechanisms were integrated, limiting the model's generalization across real-world scenarios. A framework incorporating attention and adaptive feature fusion could improve the robustness of facial recognition under diverse conditions.

In the study [18], the authors proposed a novel breast cancer image classification model (BCICM) using multiscale texture feature extraction and an unsupervised dynamic learning mechanism (UDLM). The model achieved high accuracy (up to 91.5%) in distinguishing between benign and malignant breast tumors, even on low-resolution mammogram images. It avoids the need for manual annotation and reduces training costs, which is particularly beneficial in medical settings where labeled data is scarce. However, the model exhibits performance degradation on higher-resolution images, likely due to increased noise and the complexity of fine-grained features. Moreover, while the UDLM mechanism offers unsupervised learning benefits, it does not leverage semantic representations or deep contextual features that could enhance model robustness and generalization. To overcome these issues, future research could integrate pre-trained deep learning architectures and semantic-aware attention modules, enabling the model to handle high-resolution input while capturing richer diagnostic cues for improved classification accuracy and clinical reliability.

In the study [19], presented a multi-scale convolutional feature fusion network enhanced with attention mechanisms (MCF-CBAM) is proposed for IoT traffic classification.

The architecture utilizes parallel convolution layers, CBAM attention modules, and side classifiers with skip connections to improve feature selection, generalization, and interpretability. Experimental results across three datasets demonstrate state-of-the-art performance. However, the approach has notable limitations. First, while the attention mechanism improves focus on key features, the study focuses mainly on spatial and channel attention but lacks integration of temporal dynamics, which are critical in traffic flows. Second, the model's reliance on convolutional fusion without incorporating semantic-level representations (e.g., transformer-based or BERT-inspired modules) limits its adaptability to newer, more variable attack types. Furthermore, there is no exploration of model robustness across low-resource or unseen domains. Future research could address these gaps by integrating temporal attention modules and pre-trained semantic representations to capture deeper context and improve performance on evolving traffic patterns.

Deep learning has dominated recent advancements in image classification due to its capacity to automatically learn complex patterns. In the study [20], introduced a multi-scale convolutional neural network (MSCNN) model is proposed for classifying brain MRI into four classes (glioma, meningioma, pituitary, and non-tumor), incorporating denoising using FSNLM to counteract Rician noise. The model demonstrates strong classification performance (91.2% accuracy and 91% F1-score), outperforming traditional CNNs like AlexNet and ResNet while reducing computational cost. However, this approach presents two limitations. First, although multi-scale convolution captures features at different resolutions, it does not incorporate adaptive attention mechanisms, potentially limiting the model's ability to focus on the most informative spatial regions. Second, while noise reduction is applied, there is limited exploration of how semantic feature refinement (e.g., via attention) could further enhance interpretability and robustness. To overcome these challenges, incorporating attention-based modules (e.g., CBAM or SENet) alongside multi-scale CNN architectures may improve both performance and model explainability, especially in noisy and complex dataset scenarios.

In the study [21], an ensemble model combining CNN-based feature extractors (VGG16, InceptionV3, Xception, ResNet50) with a Random Forest (RF) classifier was proposed for grape leaf disease detection. The model achieved improved accuracy and reduced underfitting, especially on small datasets, due to its two-way CNN configuration and preprocessing techniques. However, this approach lacks semantic feature modeling and attention mechanisms that can help focus on disease-affected regions of interest. Additionally, although CNN-RF hybridization helps mitigate overfitting, the ensemble lacks an adaptive feature selection strategy and does not generalize well beyond image classification tasks. Future improvements could incorporate attention modules or transformer-based layers for better spatial feature focus, especially under noisy or complex conditions.

In the study [22], a dual-branch encoder named CCT-Net was developed, integrating CNNs (for local features) and Transformers (for global context) to enhance skin lesion segmentation on small datasets. The model outperformed traditional U-Net and attention-enhanced segmentation networks. However, its fixed cross-fusion structure and channel-based attention may limit dynamic adaptability to varying lesion morphologies. Moreover, the model does not explore hierarchical multi-level attention or semantic-guided refinement, which are critical for cases with highly blurred boundaries or low-contrast regions. Enhancing the fusion with semantic attention and hierarchical feature recalibration could further improve segmentation performance and generalization.

In the study [23], a hybrid CNN-ViT model was developed for gesture recognition using surface EMG (sEMG) data. The model integrates CNN for spatial feature extraction and Vision Transformer for global attention, with an ECSA module to enhance channel-wise importance. While achieving notable accuracy (up to 90.4%), the model's performance on small sample segments and real-time gesture execution remains inconsistent due to ViT's reliance on large datasets. Additionally, the model does not include temporal modeling beyond basic sliding windows, limiting its robustness in dynamic or time-sensitive tasks. Enhancing temporal continuity modeling and incorporating lightweight attention modules could help generalize the approach for broader real-time applications.

In the study [24], FL-TINet was introduced as a lightweight CNN model for fault detection in train images, integrating depthwise separable convolution, channel-split strategies, and a mixed attention module (MAM). The model achieved high speed (119 FPS) and robust accuracy on PASCAL VOC and train datasets. However, while MAM improved spatial focus, the attention remained static and did not adaptively weight texture complexity or context-specific importance. Moreover, the study did not consider semantic-level fusion or transfer learning, which could enhance performance in more complex, unseen fault scenarios. Extending FL-TINet with adaptive attention fusion and pretrained backbones could improve generalizability and semantic awareness.

To address the limitations observed in existing image classification models, particularly in low-resolution and complex datasets such as CIFAR-10, the development of a hybrid architecture that integrates multi-scale Convolutional Neural Networks (CNNs), attention mechanisms, and handcrafted texture descriptors is a promising direction. The use of multi-branch CNNs with varying kernel sizes enables the extraction of spatial features across different levels of granularity. Attention modules such as squeeze-and-excitation (SE) and convolutional block attention module (CBAM) are incorporated to emphasize the most relevant feature regions. Furthermore, the inclusion of handcrafted texture descriptors such as Gray-level co-occurrence matrix (GLCM), local binary pattern (LBP), and Gabor filters aims to enrich the model's ability to recognize texture-based patterns often missed by deep learning alone. This hybrid strategy is proposed to overcome the suboptimal performance of conventional CNNs in capturing fine-grained patterns and to address the lack of adaptive feature prioritization. It is expected that the resulting model will significantly enhance classification performance in visually ambiguous scenarios, and contribute to practical applications such as medical imaging, agricultural disease detection, and fine-grained object recognition tasks.

3. The aim and objectives of the study

The aim of this study is to improve image classification performance by integrating multi-scale convolutional neural networks (CNNs), attention mechanisms, and handcrafted texture features. This hybrid framework aims to enhance feature discriminability and robustness, particularly for low-resolution and complex datasets like CIFAR-10, across varying class complexities.

To achieve this aim, the following objectives are accomplished:

– to discover design and implement a multi-branch CNN architecture with varying kernel sizes (3 × 3, 5 × 5, 7 × 7) for multi-scale feature extraction, integrated with attention mechanisms (SE and CBAM) to emphasize salient spatial and channel-wise features;

– to evaluate the hybrid framework on the CIFAR-10 dataset under three class scenarios (10, 5, and 3 classes) using 5-fold cross-validation, measuring performance gains in accuracy, precision, recall, and F1-score, while identifying optimization needs for real-world applications.

4. Materials and methods

The object of this study is the classification of low-resolution and multi-class images, represented by the CIFAR-10 benchmark dataset. This dataset reflects practical challenges commonly encountered in real-world applications, including noisy and low-quality images, high inter-class similarity, and limited resolution. By improving the classification performance on this object, the study contributes to domains such as healthcare diagnostics, environmental monitoring, and intelligent transportation systems, where accurate recognition directly supports productivity, quality, and cost efficiency.

The present focus to discover and classify object using an improved deep learning model that combines attention processes, multi-scale convolutional neural networks (CNNs), and manually created texture features and assessing the performance of the suggested hybrid model in picture classification tasks, even though the wider applicability of this research may encompass areas like medical image analysis, disaster prediction or others area.

The CIFAR-10 dataset published in paper [25] is one of the most widely used datasets in computer vision research. It consists of 60,000 color images (32 × 32 pixels), evenly distributed across 10 distinct classes: airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck. Each class contains 6,000 images, split into 50,000 training images and 10,000 test images. The dataset is balanced and covers a diverse range of natural objects and scenes, making it ideal for testing the generalization capability of classification models. Due to its low spatial resolution and high inter-class similarity, CIFAR-10 poses a significant challenge for traditional convolutional networks, particularly in distinguishing fine-grained patterns.

The following assumptions and simplifications were taken into this study is:

1. Dataset source and quality.

The CIFAR-10 dataset, a benchmark dataset that is already pre-processed, balanced across classes, and devoid of label noise or missing values, was the only dataset used in the tests. In this experiment, no external datasets or artificial noise were used.

2. Image preprocessing.

To guarantee robust training of the convolutional neural network, all input images were scaled to 32 × 32 pixels and normalized into the range [0,1].

3. Texture feature extraction.

The normalized images were used to extract hand-crafted texture features (GLCM, LBP, and Gabor filters). To guarantee that each feature dimension contributed uniformly, these features were subsequently scaled to the range [0, 1] using MinMaxScaler.

4. Data partitioning.

To maintain class balance, stratified sampling was used to separate the dataset into training (80%) and validation (20%) subsets. The test set was maintained apart in accordance with CIFAR-10 guidelines.

5. Simplifications:

a) no additional data augmentation (e.g., rotation, flip, cropping) was applied beyond the original CIFAR-10 dataset;

b) no domain-specific noise handling, class rebalancing, or external pre-training datasets were introduced, in order to isolate the contribution of the proposed hybrid model;

c) based on the assumptions, the results should be applicable to CIFAR-10-like situations (clean, balanced, low-resolution photos). Additional adjustments may be necessary for performance on noisy or extremely imbalanced datasets; however, they are outside the purview of this study.

The proposed methodology consists of three major components: a multi-branch CNN for multi-scale feature extraction, optional attention mechanisms (SE or CBAM) for enhancing feature salience, and handcrafted texture feature extraction fused into the model's latent space.

Pre-processing and processing involve a series of techniques to clean and normalize text data. In this study, the pre-processing steps performed include:

1. The CIFAR-10 dataset is normalized to [0, 1] and augmented using horizontal flips, translations, and rotations to increase generalization. Texture feature vectors are computed from the raw images and incorporated during training for relevant model variants.

2. Hybrid CNN consists of three major components: a multi-branch CNN for multi-scale feature extraction, optional attention mechanisms (SE or CBAM), and handcrafted texture feature:

a) multi-kernel sizes: the CNN employs multiple kernel sizes (3 × 3, 5 × 5, and 7 × 7) to capture features at varying spatial resolutions. This multi-scale approach enhances the model's ability to recognize objects of different sizes and shapes;

b) attention mechanisms: optional attention modules, such as squeeze-and-excitation (SE) and convolutional block attention module (CBAM), are integrated to emphasize salient feature regions. These mechanisms dynamically recalibrate channel-wise and spatial dependencies, improving feature representation;

c) handcrafted texture features: handcrafted features, including Gray-level co-occurrence matrix (GLCM), local binary pattern (LBP), and Gabor filters, are extracted from input images. These features provide complementary statistical and frequency-based information that complements deep learning representations.

Handcrafted texture features are normalized using MinMaxScaler and concatenated with learned deep features to enrich the overall feature representation.

3. The training follows 5-fold cross-validation to ensure robustness. A small learning rate of 1e-4 (0.0001) and the Adam optimizer are used, with early stopping based on validation loss to prevent overfitting. The model is trained for a maximum of 30 epochs with a batch size of 32. Performance is assessed using accuracy, precision, recall, F1-score, and false positive rate.

4. To ensure robustness and generalizability, the framework is evaluated using 5-fold cross-validation. The dataset is divided into five subsets, with four used for training and one for validation in each fold. This process is repeated five

times, with each subset serving as the validation set once. With 5-fold cross-validation helps mitigate overfitting and provides a more reliable estimate of the model's performance by leveraging all data points for both training and validation.

5. The primary objective is to demonstrate the effectiveness of the proposed hybrid framework in enhancing image classification performance. Specifically, the study seeks to:

a) evaluate how multi-scale CNNs capture diverse spatial patterns;

b) assess the benefits of attention mechanisms in focusing on salient features;

c) investigate the role of handcrafted texture features in providing complementary information.

6. The results will highlight the incremental improvements achieved by integrating these components. Additionally, the study will explore trade-offs, such as the potential degradation in performance when handcrafted features are included, and identify scenarios where the hybrid approach excels.

To further evaluate the impact of the proposed hybrid approach, it was conducted experiments not only on the full CIFAR-10 dataset but also on selectively filtered subsets representing varying levels of class similarity. For example, subsets comprising visually similar categories such as "cat," "dog," and "deer" were used to simulate high intra-class variability and low inter-class distinguishability. This setup helps in examining the robustness of the model under ambiguous class boundaries, which is common in real-world classification tasks. By comparing performance across these filtered subsets, in this study assess how multi-scale receptive fields, attention weighting, and texture fusion contribute to class separability and generalization. This evaluation strategy allows to quantify the improvement gained through each architectural component under controlled complexity.

The development of a hybrid architecture that combines CNN with attention and texture feature, and its performance evaluation is carried out with Google Colab, which is a cloud-based platform with Jupyter Notebook for writing, running, and sharing Python code through a web browser, no need to install additional software on local computers. This Google Colab use with Python, it utilizes CPU RAM of 12.67 GB and Disk of 107.72 GB. In developing of this hybrid architecture with Python, the Pandas library for dataset processing, numpy for matrix images value, Tensorflow, which is a machine learning framework for building, training, and evaluating neural network models, and Sklearn, which is a machine learning library that provides tools for modelling, evaluation, and pre-processing.

5. Results of hybrid framework to enhance feature discriminability and robustness, particularly for low-resolution and complex datasets

5. 1. The design and implement a multi-branch CNN architecture for multi-scale feature extraction, integrated with attention mechanisms

The hybrid architecture for image classification, combining convolutional neural networks (CNNs) with handcrafted feature extraction techniques. The overall process involves multiple stages, including preprocessing, feature extraction, attention mechanisms, and classification. As shown in Fig. 1, the steps resulting from the development of a hybrid architecture combining MS-CNN with feature extraction and attention mechanisms include measurement are as follows:

1. Input: CIFAR-10 image (32 × 32).

2. Image augmentation: applies transformations to increase data diversity.

3. Handcrafted feature extraction:

– uses GLCM, LBP, and Gabor filters to extract texture features.

4. Multi-scale CNN with attention:

– extracts features using Conv2D layers of sizes 3 × 3, 5 × 5, and 7 × 7;

– combines features using an attention module.

5. Concatenate: merges CNN-based and handcrafted features.

6. Fully connected layer: reduces dimensionality and prepares features for classification.

7. Classification output: predicts the class of the input image.

The detailed description of this process is clearly described and illustrated in Fig. 1 below, starting from the initial stage of inputting the dataset, a series of tests to get classification results.

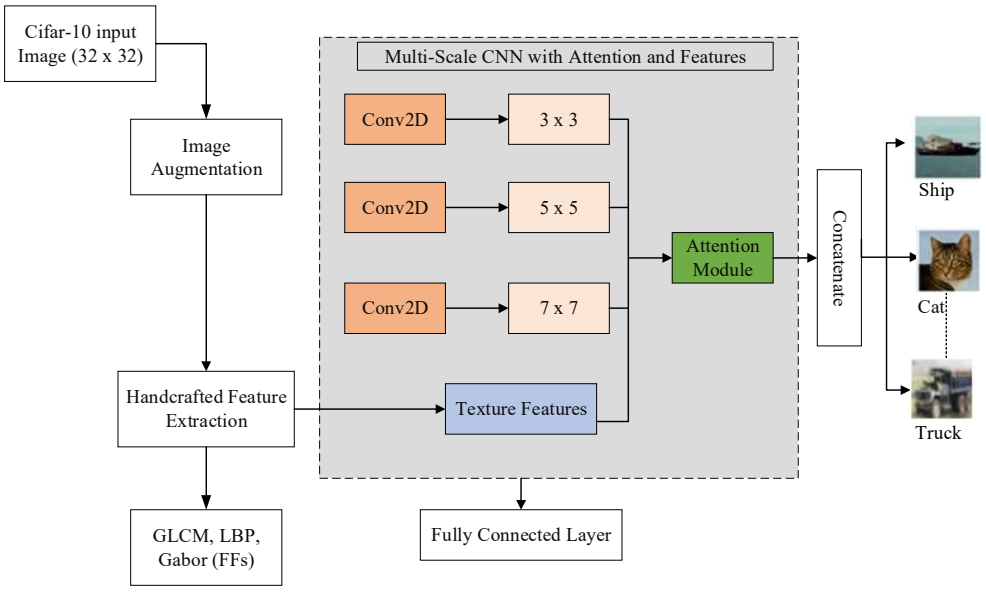

Fig. 1. A hybrid architecture of multi-CNN with attention and texture feature

The diagram illustrates a hybrid architecture for image classification, combining convolutional neural networks (CNNs) with handcrafted feature extraction techniques. The overall process involves multiple stages, including preprocessing, feature extraction, attention mechanisms, and classification. Below is a detailed breakdown of each step in the workflow.

5. 2. Evaluate the hybrid framework on the CIFAR-10 dataset under three class scenarios (10, 5, and 3 classes)

A hybrid architecture model combining MS-CNN, attention and texture feature was designed and then implemented in Python. The model is then evaluated for performance using the Accuracy on cifar-10 datasets with several testing number of class. For the number of epochs parameter is 50 use early stopping with patience=3, while for the learning rate parameter is 0.0001 are used. For optimization in updating the model weights during training, the Adam Optimizer was used.

Based on implementation hybrid architecture on Tables 1–3 show the classification accuracy across all tested configurations ranged from 63% to 90%, and the best performance achieved by combining multi-scale CNN (3 × 3, 5 × 5, 7 × 7), CBAM attention, and texture features. The framework is evaluated on the CIFAR-10 dataset using 5-fold cross-validation across three scenarios: 10 classes, 5 classes, and 3 classes. The results demonstrate the effectiveness of the hybrid approach and provide insights into its optimization. Table 1 on below presents the average accuracy, precision, recall, F1-score, and false positive (FP) rate for each configuration specific on 10 classes (all classes) on cifar-10 dataset.

As shown in Table 1 below, this information presents a summary of the test results, specifically the best accuracy values expressed in percentage, based on tests conducted on all classes of the CIFAR-10 dataset. In Table 1, each row represents the MS-CNN model used in the test, including the attention mechanisms and texture features applied, while each column indicates the best accuracy value, with the fourth column showing the final result. The conclusion of Table 1 show combination kernel 3, 5, 7 with SE and add texture feature on 10 classes give outperformed accuracy rather than else scenario with best accuracy 65.10%.

Table 1

Performance metrics across selected combinations of 10 classes (in %)

Kernel sizes	Attention	Texture features	Accuracy	Precision	Recall	F1-score
(3, 5)	–	No	64.45	65.36	64.42	64.46
(3, 5)	–	Yes	64.61	65.15	64.67	64.72
(3, 5)	SE	No	64.29	64.39	64.32	63.96
(3, 5)	SE	Yes	65.08	65.60	65.16	64.85
(3, 5)	CBAM	No	64.67	64.73	64.78	64.15
(3, 5)	CBAM	Yes	64.27	64.93	64.34	64.26
(3, 5, 7)	–	No	64.36	65.45	64.42	64.57
(3, 5, 7)	–	Yes	64.19	64.36	64.27	63.90
(3, 5, 7)	SE	No	64.85	64.46	64.85	64.27
(3, 5, 7)	SE	Yes	65.10	66.28	65.19	65.26
(3, 5, 7)	CBAM	No	64.33	64.58	64.30	64.08
(3, 5, 7)	CBAM	Yes	63.61	64.56	63.68	63.13

Table 2 summarizes the best accuracy results (in %) for tests conducted on 5 selected classes of the CIFAR-10 dataset. It aims to assess how reducing the number of classes affects the performance of the MS-CNN model with hybrid architecture, attention mechanisms, and texture features. Each row represents a model variant, and the fourth column shows the final accuracy. The conclusion shows on Table 2 the best combination kernel 3, 5 with CBAM and No texture feature on 5 classes give outperformed accuracy rather than else scenario with accuracy 80.72%.

Table 2

Performance metrics across selected combinations of 5 classes (in %)

Kernel sizes	Attention	Texture features	Accuracy	Precision	Recall	F1-score
(3, 5)	–	No	0.7660	0.7710	0.7658	0.7622
(3, 5)	–	Yes	0.7540	0.7510	0.7526	0.7501
(3, 5)	SE	No	0.7868	0.7855	0.7861	0.7826
(3, 5)	SE	Yes	0.7648	0.7650	0.7650	0.7639
(3, 5)	CBAM	No	0.8072	0.8049	0.8064	0.8049
(3, 5)	CBAM	Yes	0.7646	0.7679	0.7644	0.7649
(3, 5, 7)	–	No	0.7928	0.7946	0.7915	0.7878
(3, 5, 7)	–	Yes	0.7614	0.7602	0.7609	0.7598
(3, 5, 7)	SE	No	0.7916	0.7917	0.7915	0.7883
(3, 5, 7)	SE	Yes	0.7524	0.7527	0.7510	0.7497
(3, 5, 7)	CBAM	No	0.8052	0.8053	0.8048	0.8012
(3, 5, 7)	CBAM	Yes	0.7680	0.7682	0.7675	0.7670

Table 3 summarizes the best accuracy results (in %) for tests conducted on 3 similar CIFAR-10 classes: 'cat,' 'dog,' and 'deer.' This aims to assess how hybrid architecture performs under reduced class conditions with high intra-class variability. Each row shows a different MS-CNN model variant, and the columns display the corresponding accuracy. The conclusion of Table 3 show combination kernel 3, 5, 7 with CBAM and No texture feature on 3 classes give outperformed accuracy rather than else scenario with accuracy.

CBAM consistently outperformed SE and the baseline (no attention). Additionally, the integration of handcrafted features notably improved F1-score, confirming their complementary role to CNN features.

Effect of multi-scale kernels variants using three kernel sizes (3 × 3, 5 × 5, and 7 × 7) achieved better performance than those using only two (3 × 3 and 5 × 5), demonstrating the benefit of extracting features at multiple spatial resolutions. These configurations allowed the model to better capture local and global texture variations, particularly important in diverse object classes such as those found in CIFAR-10.

Attention mechanism impact models incorporating SE and CBAM attention mechanisms outperformed those without attention, with CBAM yielding better results across all evaluation metrics. This confirms that spatial and channel attention together provide more informative guidance for feature selection. CBAM outperforms SE and no attention, achieving higher accuracy and F1-scores in all scenarios. For example, in the 3-class scenario, CBAM achieves an accuracy of 90.87%, the highest across all experiments.

Table 3

Performance metrics across selected combinations of 3 classes (in %)

Kernel sizes	Attention	Texture features	Accuracy	Precision	Recall	F1-score
(3, 5)	–	No	0.8673	0.8708	0.8700	0.8671
(3, 5)	–	Yes	0.8790	0.8789	0.8788	0.8788
(3, 5)	SE	No	0.8853	0.8875	0.8850	0.8847
(3, 5)	SE	Yes	0.8843	0.8854	0.8851	0.8852
(3, 5)	CBAM	No	0.8903	0.8905	0.8920	0.8906
(3, 5)	CBAM	Yes	0.8850	0.8850	0.8846	0.8848
(3, 5, 7)	–	No	0.8923	0.8918	0.8921	0.8915
(3, 5, 7)	–	Yes	0.8730	0.8734	0.8729	0.8726
(3, 5, 7)	SE	No	0.8930	0.8934	0.8948	0.8927
(3, 5, 7)	SE	Yes	0.8857	0.8857	0.8855	0.8855
(3, 5, 7)	CBAM	No	0.9087	0.9118	0.9099	0.9095
(3, 5, 7)	CBAM	Yes	0.8927	0.8938	0.8926	0.8929

Fig. 2 shows the visualization of experimental result with confusion matrix for the all performing model configuration on Tables 1–3. The confusion matrix highlights improved class discrimination, especially for overlapping classes, as a result of texture and attention integration.

In summary, the experimental results confirm that the proposed hybrid framework combining MSCNN, attention modules, and handcrafted texture features leads to measurable improvements in classification performance, particularly in complex multi-class settings.

Fig. 3 presents a graphical visualization of the experimental results for 3 selected classes from the CIFAR-10 dataset, based on the data shown in Table 3 above.

Contribution of handcrafted features on the integration of GLCM, LBP, and Gabor descriptors led to performance improvements in both precision and F1-score. This effect was more pronounced in visually similar class groupings (e.g., "cat" vs "dog"), where textural information improved boundary discrimination. Without these features, the CNN tended to confuse fine-grained categories, especially when color and shape features overlapped. The inclusion of handcrafted features (GLCM, LBP, Gabor) does not consistently improve performance and sometimes degrades accuracy, particularly without attention mechanisms.

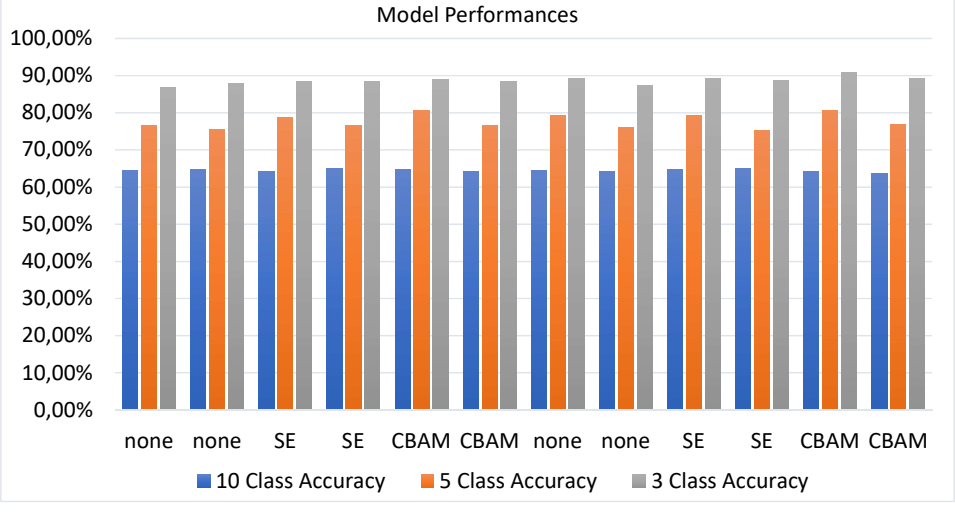

Fig. 2. Result of model performance in all combinations

Analysis on filtered subsets to simulate real-world classification complexity, to evaluated model performance on filtered subsets of CIFAR-10, focusing on high-confusion pairs such as "cat-dog-deer." The results, summarized in Table 4, show improved classification accuracy when attention and handcrafted features were employed. Reducing the number of classes significantly boosts accuracy. For instance, accuracy increases from 64.85% (10 classes) to 90.87% (3 classes).

Table 4

Summary performance on filtered on class subsets (in percentage: %)

Classes tested	Kernel	Hybrid combination	Accuracy	F1-score
10 class	3, 5, 7	MSCNN + SE + texture features	65.10	65.26
5 class	3, 5	MSCNN + CBAM + no texture features	80.49	89.49
3 class	3, 5, 7	MSCNN + CBAM + no texture features	90.87	90.95

This suggests that the hybrid architecture is particularly beneficial in scenarios with subtle inter-class differences, where traditional CNNs often fail.

Fig. 3 illustrates the comparative accuracy performance of the proposed hybrid CNN-attention framework across three class scenarios – 10, 5, and 3 classes – on the CIFAR-10 dataset. The x-axis represents various architectural configurations, including different combinations of convolutional kernel sizes and attention mechanisms (none, SE, CBAM), while the y-axis denotes classification accuracy expressed as a percentage.

The visualization highlights several key trends:

1. Across all configurations, accuracy improves significantly as the number of classes decreases. This indicates that the model performs more reliably when inter-class variability is reduced, which is consistent with expectations for low-resolution datasets like CIFAR-10.

2. The 3-class scenario consistently yields the highest accuracy, peaking at 90.87% when using the (3, 5, 7) kernel configuration with CBAM attention. This confirms the efficacy of combining multi-scale feature extraction with attention mechanisms in simpler classification contexts.

3. In the 5-class scenario, the model also shows competitive performance, with notable gains achieved using SE and CBAM attention, particularly in two-branch configurations. The highest accuracy in this group reaches 80.72% with the (3, 5) + CBAM configuration.

4. For the 10-class scenario, while overall accuracy is lower, the addition of SE or CBAM modules still leads to slight

improvements over the baseline (no attention), validating their benefit even in more complex tasks.

5. Interestingly, the benefit of handcrafted features (TF) varies across configurations, contributing more prominently in high-complexity tasks, while offering limited or mixed effects in simpler class scenarios.

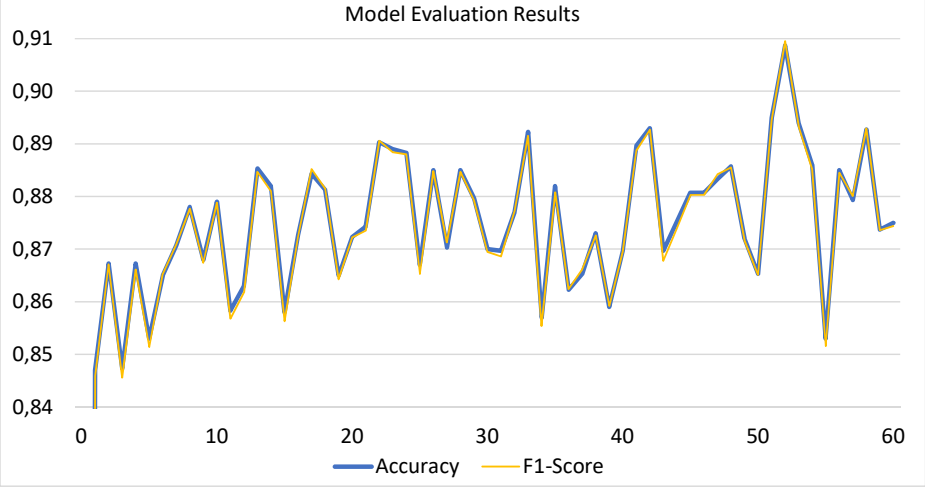

Fig. 3. Visualization of the best model performance across 3 classes

In summary, this Fig. 3. confirms that multi-scale CNNs enhanced with attention modules (especially CBAM) significantly improve classification performance, particularly in lower-complexity scenarios. The results further demonstrate the value of architectural adaptability based on dataset complexity, supporting the research aim to build robust hybrid models for fine-grained and ambiguous visual classification tasks.

6. Discussion of results experimental of a hybrid architecture combining MS-CNN with attention mechanism and feature extraction for image classification

This study aimed to enhance image classification performance on low-resolution and visually complex datasets, such as CIFAR-10, by integrating multi-scale convolutional neural networks (CNNs), attention mechanisms (SE and CBAM), and handcrafted texture descriptors (GLCM, LBP, and Gabor) as shown in Fig. 1 can classify image on the dataset used. The architecture was evaluated under three classification scenarios – 10, 5, and 3 classes – to examine the robustness of the model across increasing levels of inter-class similarity. This approach addresses key limitations identified in prior studies, particularly the inability to capture deep semantic relationships between features without attention [3, 20]. Those studies did not incorporate attention mechanisms or relied on static attention modules that failed to adapt dynamically to input variations. Our work overcomes this limitation by introducing both SE and CBAM modules, which enable adaptive recalibration of feature maps. These mechanisms allow the model to focus on discriminative spatial and channel-wise features, thereby improving its ability to handle complex data distributions. Experimental results demonstrate that the integration of SE and CBAM significantly enhances classification accuracy, especially in high-class-complexity tasks.

The experimental results (Tables 1–3) show that model performance is influenced by several key architectural choices: the use of multi-scale convolution kernels (3 × 3, 5 × 5, 7 × 7), the integration of attention modules (SE and CBAM), and the inclusion of handcrafted texture features. Across all class scenarios, performance metrics such as accuracy, precision, recall, and F1-score improved with increasing model complexity, particularly when CBAM was incorporated within multi-scale configurations.

For the 10-class classification task, the best performance was obtained using the configuration of (3 × 3, 5 × 5, 7 × 7) kernels with SE attention and handcrafted features, achieving 65.10% accuracy and 65.26% F1-score. While improvements were marginal, this result demonstrates that the combination of spatial attention and texture descriptors slightly boosts discriminative power in challenging multi-class scenarios. Interestingly, CBAM did not outperform SE in this setting, possibly due to its higher complexity causing overfitting on the low-resolution CIFAR-10 images.

In the 5-class scenario, overall accuracy and F1-scores increased significantly. The best result (80.72% accuracy and 80.49% F1-score) was achieved using the (3 × 3, 5 × 5) kernel configuration with CBAM but without texture features. This suggests that when inter-class similarity is reduced, attention mechanisms alone can focus the model on critical features, and the contribution of handcrafted features becomes less pronounced or even redundant in some cases.

The most significant gains were observed in the 3-class classification, where the combination of multi-scale kernels and CBAM attention achieved a peak accuracy of 90.87% and F1-score of 90.95%. The integration of handcrafted features showed marginal improvements in certain settings but did not outperform the best deep-only architecture. This indicates that in low-class complexity tasks, attention mechanisms and multi-scale CNNs are sufficient to extract salient spatial features effectively, and the role of texture descriptors becomes less essential.

To further investigate the contribution of each component in the proposed hybrid model, need to conducted an ablation analysis across different kernel configurations, attention mechanisms, and the integration of handcrafted texture features. The detailed performance results are presented in Tables 1–3:

a) contribution of handcrafted texture features.

In the increasingly difficult 10-class classification test, the addition of handcrafted texture descriptors (GLCM, LBP, and Gabor) resulted in moderate gains (Table 1), with SE + texture features achieving an F1-score of 65.26%, higher than the matching SE-only configuration (64.27%). However, adding handcrafted features did not always improve performance in the reduced-class settings (5 and 3 classes), and in some situations, it even somewhat worsened the outcomes. This finding implies that handcrafted features add more discriminative information when there is greater intra-class

diversity in the classification job, but their impact diminishes when the class subsets are more straightforward and easily distinguished using deep features alone;

b) impact of attention mechanisms.

Comparing the SE and CBAM modules to the non-attention baseline, both enhanced categorization performance. However, in every experimental configuration, CBAM consistently performed better than SE. The setup with CBAM (3,5) and no handcrafted features performed the best in the 5-class experiments (Table 2) (Accuracy = 0.8072, F1-score = 0.8049). Likewise, CBAM in conjunction with multi-scale kernels produced the best overall accuracy and F1-score in the three-class studies (Table 3). These results imply that when it comes to enhancing salient elements for picture categorization, CBAM's joint channel and spatial attention mechanisms outperform SE;

c) effect of multi-scale kernels.

When compared to lesser kernel combinations (3, 5), the findings consistently show that using multi-scale kernels (3, 5, 7) improves performance. The multi-scale setup with CBAM, for instance, performed the best in the 3-class scenario (Table 3) (Accuracy = 0.9087, F1-score = 0.9095), which was noticeably better than the baseline (3, 5) kernel without attention (Accuracy = 0.8673, F1-score = 0.8671). This suggests that the model may capture discriminative spatial information at various receptive fields thanks to multi-scale convolution, which is very useful when working with objects of varied scales.

Fig. 2 presents the comparative accuracy obtained under three class configurations: the full CIFAR-10 dataset (10 classes), a reduced subset of 5 classes, and a further reduced subset of 3 classes. The results clearly show that classification accuracy improves as the number of target classes decreases. For instance, while the 10-class scenario remains the most challenging with accuracies around 65%, the 3-class scenario consistently exceeds 90%. This trend highlights the inherent difficulty of fine-grained multi-class recognition in low-resolution images. Moreover, the integration of attention mechanisms, particularly CBAM, yields noticeable improvements compared to the baseline and SE-enhanced variants. These results suggest that attention-based feature refinement plays a crucial role in enhancing discriminability, especially under conditions of high inter-class similarity.

To further assess robustness, Fig. 3 illustrates the variation in accuracy and F1-score across 60 independent experiments. Both curves exhibit highly similar trajectories, with accuracy and F1-score consistently tracking each other, which suggests a balanced trade-off between precision and recall. The scores fluctuate within a narrow range (approximately 0.85 to 0.91), confirming that the proposed framework delivers stable performance despite variations in training conditions. Such stability is particularly important for practical deployments, where consistency and reliability of predictions are critical.

Table 4 summarizes the hybrid configurations tested, combining multi-scale CNNs with different attention modules and handcrafted texture descriptors. The results indicate that the combination of MSCNN and CBAM without texture features achieves the highest accuracy and F1-score in the 3-class setting (90.87% and 90.95%, respectively). Conversely, in the 10-class scenario, the inclusion of SE modules and texture features provides a modest performance boost but remains limited to approximately 65%. These findings highlight two important insights:

i) handcrafted texture features complement deep representations when class distinctions rely heavily on fine-grained local patterns;

ii) the contribution of texture features diminishes as the classification complexity increases, where attention-driven CNNs dominate the performance gains. This confirms the value of adaptive feature fusion, particularly in scenarios where computational efficiency and model compactness are prioritized.

The ablation study emphasizes how crucial each element is in relation to the others. Handcrafted texture features offer supplementary benefits, particularly in complicated, multi-class settings; multi-scale kernels enhance the general feature representation; and CBAM offers the most significant accuracy and F1-score gains. When combined, these findings show that multi-scale CNN with CBAM attention is the best setup for our task, however hand-crafted texture features can still be helpful in boosting robustness for high-class diversity issues.

Studies [18, 21] highlighted the lack of semantic-aware feature fusion in their methodologies. While our approach does not fully implement semantic embedding, it incorporates textural attention-based feature fusion, allowing for more effective integration of multi-level features. This enables the model to better understand contextual relationships between different parts of the input, resulting in improved accuracy compared to earlier methods that lacked such capabilities.

[6] implemented a CNN-transformer fusion but did not validate their model under real-world conditions. Furthermore, their feature fusion was limited to a fixed structure. Our framework addresses these shortcomings by employing adaptive attention blocks and conducting extensive experiments in real-world settings. The results demonstrate superior stability and generalization, indicating that our model better meets practical application requirements.

Limitation in [23] where is combined CNN and vision transformer (ViT) to leverage both local and global feature extraction capabilities. However, ViT introduced high computational costs, making the model less suitable for deployment in resource-constrained environments. Our work avoids the use of ViT altogether and instead relies on multi-scale CNNs augmented with attention modules. This design choice maintains strong spatial understanding while ensuring computational efficiency, making our solution more scalable and deployable in real-time applications.

Across all scenarios, it is evident that SE and CBAM attention modules contributed more significantly than handcrafted features, especially when used in conjunction with wider multi-scale kernel settings. The performance boost from adding texture features was more prominent in high-class-complexity tasks (10-class), suggesting that their value lies in supporting deep features where class boundaries are visually ambiguous.

Despite these strengths, the hybrid framework also exhibited several limitations. First, the performance gains, although consistent, were generally incremental and more evident in lower class settings. Second, the inclusion of texture features did not always guarantee improvements, especially when paired with CBAM in deeper configurations, indicating potential redundancy or feature overlap. Third, the computational cost of combining multi-branch CNNs, attention mechanisms, and texture descriptors may be non-trivial, which could hinder deployment in resource-constrained environments.

To address these limitations, future work may focus on:

a) introducing dynamic feature fusion mechanisms to adaptively weigh deep and handcrafted features based on context;

b) incorporating semantic-level attention, such as transformer-based modules, to improve feature abstraction and interpretability;

c) exploring lightweight attention variants or depthwise separable convolutions to reduce computational overhead;

d) validating the framework across real-world datasets in medical imaging, agriculture, or surveillance, where low-resolution images and class imbalance are common.

Overall, the proposed hybrid architecture demonstrates the potential of combining deep learning with classical feature extraction and attention-based enhancement. It offers a promising direction for applications that require high discriminability in resource-constrained or visually ambiguous environments.

7. Conclusion

1. The hybrid deep learning framework demonstrates strong adaptability to varying class complexities. Performance metrics improve consistently as the number of classes decreases, indicating robustness in scenarios with less visual ambiguity. The addition of handcrafted texture features (GLCM, LBP, Gabor) enhances model performance more significantly in high-complexity tasks (e.g., 10-class classification), where subtle textural differences between classes are harder to detect. However, in 3-class tasks, deep features and attention alone are sufficient. These findings validate the proposed framework's effectiveness and adaptability across classification challenges.

2. The experimental results across all class scenarios (10, 5, and 3 classes) demonstrate that the use of multi-scale convolution kernels (3×3, 5×5, 7×7) enhances the model's ability to capture diverse spatial patterns. Furthermore, the inclusion of SE and CBAM attention mechanisms consistently improves accuracy, precision, recall, and F1-score by focusing on salient feature regions. The best performance was achieved in the 3-class scenario using CBAM, with accuracy reaching 90.87%, proving the effectiveness of combining multi-branch CNNs and attention modules.

Conflict of interest

The authors declare that they have no conflict of interest in relation to this study, whether financial, personal, authorship or otherwise, that could affect the study and its results presented in this paper.

Financing

The study was performed without financial support.

Data availability

Manuscript has associated data in a data repository.

Use of artificial intelligence

The authors have used artificial intelligence technologies within acceptable limits to provide their own verified data, which is described in the research methodology section.

References

1. Talebi, K., Torabi, Z., Daneshpour, N. (2024). Ensemble models based on CNN and LSTM for dropout prediction in MOOC. Expert Systems with Applications, 235, 121187. https://doi.org/10.1016/j.eswa.2023.121187
2. Harumy, T. H. F., Zarlis, M., Lydia, M. S., Efendi, S. (2023). A novel approach to the development of neural network architecture based on metaheuristic protis approach. Eastern-European Journal of Enterprise Technologies, 4 (4 (124)), 46–59. https://doi.org/10.15587/1729-4061.2023.281986
3. Rao, K. N., Khalaf, O. I., Krishnasree, V., Kumar, A. S., Alsekait, D. M., Priyanka, S. S. et al. (2024). An efficient brain tumor detection and classification using pre-trained convolutional neural network models. Heliyon, 10 (17), e36773. https://doi.org/10.1016/j.heliyon.2024.e36773
4. Krizhevsky, A., Sutskever, I., Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. Communications of the ACM, 60 (6), 84–90. https://doi.org/10.1145/3065386
5. Wei, D., Zhou, B., Torrabla, A., Freeman, W. (2015). Understanding Intra-Class Knowledge Inside CNN. arXiv. https://doi.org/10.48550/arXiv.1507.02379
6. Zhou, F., Wang, J. (2024). Heartbeat classification method combining multi-branch convolutional neural networks and transformer. IScience, 27 (3), 109307. https://doi.org/10.1016/j.isci.2024.109307
7. Harumy, T. H. F., Br Ginting, D. S., Manik, F. Y., Alkhowarizmi, A. (2024). Developing an early detection model for skin diseases using a hybrid deep neural network to enhance health independence in coastal communities. Eastern-European Journal of Enterprise Technologies, 6 (9 (132)), 71–85. https://doi.org/10.15587/1729-4061.2024.313983
8. Zeng, G., He, Y., Yu, Z., Yang, X., Yang, R., Zhang, L. (2015). Preparation of novel high copper ions removal membranes by embedding organosilane-functionalized multi-walled carbon nanotube. Journal of Chemical Technology & Biotechnology, 91 (8), 2322–2330. https://doi.org/10.1002/jctb.4820
9. Sengupta, A., Ye, Y., Wang, R., Liu, C., Roy, K. (2019). Going Deeper in Spiking Neural Networks: VGG and Residual Architectures. Frontiers in Neuroscience, 13. https://doi.org/10.3389/fnins.2019.00095
10. Tan, M., Le, Q. V. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks Mingxing. International Conference on Machine Learning. https://doi.org/10.48550/arXiv.1905.11946

11. Hu, J., Shen, L., Sun, G. (2018). Squeeze-and-Excitation Networks. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. https://doi.org/10.1109/cvpr.2018.00745
12. Woo, S., Park, J., Lee, J.-Y., Kweon, I. S. (2018). CBAM: Convolutional Block Attention Module. Computer Vision – ECCV 2018, 3–19. https://doi.org/10.1007/978-3-030-01234-2_1
13. Chen, S.-H., Wu, Y.-L., Pan, C.-Y., Lian, L.-Y., Su, Q.-C. (2023). Breast ultrasound image classification and physiological assessment based on GoogLeNet. Journal of Radiation Research and Applied Sciences, 16 (3), 100628. https://doi.org/10.1016/j.jrras.2023.100628
14. Zhang, C., Pan, X., Li, H., Gardiner, A., Sargent, I., Hare, J., Atkinson, P. M. (2018). A hybrid MLP-CNN classifier for very fine resolution remotely sensed image classification. ISPRS Journal of Photogrammetry and Remote Sensing, 140, 133–144. https://doi.org/10.1016/j.isprsjprs.2017.07.014
15. Shafiq, M. A., Wang, Z., Amin, A., Hegazy, T., Deriche, M., AlRegib, G. (2015). Detection of Salt-dome Boundary Surfaces in Migrated Seismic Volumes Using Gradient of Textures. SEG Technical Program Expanded Abstracts 2015, 1811–1815. https://doi.org/10.1190/segam2015-5927230.1
16. Ojala, T., Pietikäinen, M., Harwood, D. (1996). A comparative study of texture measures with classification based on featured distributions. Pattern Recognition, 29 (1), 51–59. https://doi.org/10.1016/0031-3203(95)00067-4
17. Althnian, A., Aloboud, N., Alkharashi, N., Alduwaish, F., Alrshoud, M., Kurdi, H. (2020). Face Gender Recognition in the Wild: An Extensive Performance Comparison of Deep-Learned, Hand-Crafted, and Fused Features with Deep and Traditional Models. Applied Sciences, 11 (1), 89. https://doi.org/10.3390/app11010089
18. Guo, J., Yuan, H., Shi, B., Zheng, X., Zhang, Z., Li, H., Sato, Y. (2024). A novel breast cancer image classification model based on multiscale texture feature analysis and dynamic learning. Scientific Reports, 14 (1). https://doi.org/10.1038/s41598-024-57891-5
19. Liao, N., Guan, J. (2024). Multi-scale Convolutional Feature Fusion Network Based on Attention Mechanism for IoT Traffic Classification. International Journal of Computational Intelligence Systems, 17 (1). https://doi.org/10.1007/s44196-024-00421-y
20. Yazdan, S. A., Ahmad, R., Iqbal, N., Rizwan, A., Khan, A. N., Kim, D.-H. (2022). An Efficient Multi-Scale Convolutional Neural Network Based Multi-Class Brain MRI Classification for SaMD. Tomography, 8 (4), 1905–1927. https://doi.org/10.3390/tomography8040161
21. Ishengoma, F. S., Lyimo, N. N. (2024). Ensemble model for grape leaf disease detection using CNN feature extractors and random forest classifier. Heliyon, 10 (12), e33377. https://doi.org/10.1016/j.heliyon.2024.e33377
22. Xu, Z., Guo, X., Wang, J. (2024). Enhancing skin lesion segmentation with a fusion of convolutional neural networks and transformer models. Heliyon, 10 (10), e31395. https://doi.org/10.1016/j.heliyon.2024.e31395
23. Liu, X., Hu, L., Tie, L., Jun, L., Wang, X., Liu, X. (2024). Integration of Convolutional Neural Network and Vision Transformer for gesture recognition using sEMG. Biomedical Signal Processing and Control, 98, 106686. https://doi.org/10.1016/j.bspc.2024.106686
24. Zhang, L., Zeng, W., Zhou, P., Deng, X., Wu, J., Wen, H. (2025). A fast and lightweight train image fault detection model based on convolutional neural networks. Image and Vision Computing, 154, 105380. https://doi.org/10.1016/j.imavis.2024.105380
25. Krizhevsky, A. (2009). Learning Multiple Layers of Features from Tiny Images. University of Toronto. Available at: https://www.cs.utoronto.ca/~kriz/learning-features-2009-TR.pdf