Hybrid convolutional networks, hidden Markov models, and autoencoders for enhanced recognition

Driss Naji[1], Kamal Elhattab[2], Abdelali Joumad[3], Abdelouahed Ait Ider[4], Abdelkbir Ouisaadane[5], Azzeddine Idhmad[6]

[1]TIAD Laboratory, Department Computer of Science, Faculty of Science and Technique, Sultan Moulay Slimane University, Benimellal, Morocco
[2]ELITES Laboratory, Department Computer of Science, Faculty of Science, Chouaib Doukkali University, El Jadida, Morocco
[3]LAROSERI laboratory, Department of informatics, Faculty of Sciences, Chouaib Dokkali University, El Jadida, Morocco
[4]ISIMA Laboratory, Department Computer of Science, Faculty of Polydisciplinary, Ibnou Zohr University, Taroudant, Morocco
[5]LIMATI Laboratory, Department of Mathematics and Computer Science, Faculty of Polydisciplinary, Sultan Moulay Slimane University, Benimellal, Morocco
[6]Laboratory of Analysis, Geometry, and Applications, Faculty of Sciences, Ibn Tofail University, Kenitra, Morocco

Article Info

Article history:

Received Mar 9, 2025
Revised Nov 8, 2025
Accepted Jan 10, 2026

Keywords:

2D hidden Markov models
Convolutional neural networks
Hybrid deep learning framework
Markov random fields
Variational autoencoders

ABSTRACT

Recognition problems, including object detection, scene understanding, and fine-grained categorisation, are popular subjects in computer vision. However, it is challenging to model spatial coherence and contextual dependencies in response to changes in configurations. Human Vs computers' ability in perception-although convolutional neural networks (CNNs) do well in the extraction of features, they have high dependence on local receptive fields and are not able to capture long-range spatial relationships and high-order interactions. To alleviate the shortcomings of the current approaches, we present an enhanced hybrid CNNs two-dimensional hidden Markov model (2D-HMM) framework that combines 2D-HMM, Markov random fields (MRF) and variational autoencoders (VAEs) into a single model. The model employs 2D-HMMs for pairwise spatial modelling, MRFs for higher order context, and VAEs for stable latent representation learning. Tested on the MNIST and CIFAR-10 benchmark datasets, our approach consistently outperforms the state-of-the-art performance by 98.2% and 89.5%, respectively, with high robustness to noise and occlusion. Results from ablation studies further show that MRFs improve recall by 1.6% and VAEs improve precision by 1.3%, suggesting that they complement each other sufficiently with respect to overall testing performance. This work unifies deep learning and probabilistic graphical models, leading to more interpretable, scalable, and accurate recognition systems.

Corresponding Author:

Driss Naji
TIAD Laboratory, Department Computer of Science, Faculty of Science and Technique
Sultan Moulay Slimane University
Beni-Mellal, Morocco
Email: naji.drisss@gmail.com

1. INTRODUCTION

Recognition tasks, such as object detection, scene understanding, and fine-grained classification, represent fundamental challenges in computer vision and pattern recognition, forming the cornerstone of numerous applications ranging from autonomous vehicles to medical diagnostics. Despite remarkable advances in deep learning architectures, persistent challenges remain in effectively modeling spatial coherence and contextual dependencies, particularly when dealing with complex scenarios involving occlusion, noise, and significant variability in data distributions [1], [2]. These challenges are further compounded by the inherent limitations of current approaches in capturing long-range spatial relationships and higher-order interactions that are crucial for robust recognition performance.

While convolutional neural networks (CNNs) have established themselves as the dominant paradigm for feature extraction in vision tasks, demonstrating exceptional performance across diverse applications, their architectural design introduces fundamental limitations that constrain their effectiveness in complex recognition scenarios. Specifically, CNNs rely heavily on local receptive fields, which inherently limits their ability to capture long-range spatial relationships and contextual dependencies that extend beyond immediate spatial neighborhoods [3], [4]. This limitation becomes particularly pronounced in cluttered scenes where CNNs often struggle to distinguish overlapping objects or fail to infer contextual cues that require understanding of broader spatial contexts [5]. For instance, in scene recognition tasks, CNNs may misclassify objects due to contextual ambiguities, such as distinguishing between a boat on a road versus a boat on water, where global context is essential for accurate classification.

The recognition of these limitations has motivated researchers to explore hybrid approaches that combine the feature extraction capabilities of CNNs with probabilistic models capable of explicitly modeling spatial and contextual relationships. Probabilistic models such as hidden Markov models (HMMs) and Markov random fields (MRFs) have shown promise when integrated with CNNs, offering complementary strengths in structured prediction and spatial modeling. Recent studies have demonstrated that hybrid CNN-HMM architectures can significantly improve structured prediction tasks by effectively modeling sequential or grid-based dependencies [6]. However, traditional first-order HMMs and MRFs, while useful, are often insufficient for complex recognition tasks that require modeling of higher-order interactions and sophisticated spatial relationships [7].

This limitation necessitates the development of more advanced hybrid frameworks that can effectively leverage the strengths of multiple modeling paradigms. Building upon these insights, this research advances the current state-of-the-art by proposing a comprehensive framework that integrates two-dimensional hidden Markov models (2D-HMMs) for sophisticated pairwise spatial modeling, MRFs for capturing higher-order contextual relationships, and variational autoencoders (VAEs) for robust latent representation learning within a unified architectural framework. This integration is motivated by recent breakthroughs in several key areas of research. First, in the domain of spatial coherence modeling, 2D-HMMs have demonstrated significant success in applications such as image segmentation [8] and document analysis though their application to general recognition tasks remains relatively underexplored and presents opportunities for novel contributions [9], [10]. Second, regarding higher-order context modeling, MRFs with sophisticated clique potentials have shown considerable promise in semantic segmentation [11] and medical imaging applications [12], [13] though their integration with modern deep learning architectures continues to evolve and presents technical challenges. Third, in the area of robust latent representation learning, VAEs have proven their effectiveness in enhancing robustness to noise and occlusion, as demonstrated in recent studies focusing on semi-supervised learning [14], [15] and anomaly detection [16].

The proposed framework is specifically designed to address several critical gaps in current recognition systems. In terms of spatial reasoning capabilities, the framework captures both local dependencies through CNN feature extraction and global dependencies through integrated 2D-HMM and MRF modeling, providing a comprehensive approach to spatial understanding. From an interpretability perspective, the incorporation of VAEs provides valuable insights into latent feature distributions, aligning with current trends toward explainable artificial intelligence [17]. Regarding scalability considerations, the framework incorporates GPU-optimized training procedures that enable efficient deployment on high-resolution datasets, addressing practical concerns about computational feasibility [18]. The primary objective of this research is to demonstrate that the synergistic combination of these complementary modeling approaches can achieve superior recognition performance while maintaining computational efficiency and providing enhanced interpretability compared to existing state-of-the-art methods.

2. THEORETICAL FOUNDATIONS

2.1. Convolutional neural networks

CNNs remain the backbone of modern recognition systems due to their ability to learn hierarchical features. Recent architectures, such as EfficientNet [18] and vision transformers [19], achieve state-of-the-art results by balancing depth, width, and resolution. However, CNNs prioritize local features and lack mechanisms to explicitly model spatial relationships between distant regions [20]. For example, in scene recognition, CNNs may misclassify objects due to contextual ambiguities (e.g., a "boat" on a road vs. water) [2], [21].

2.2. Two-dimensional hidden Markov models

2D-HMMs extend traditional HMMs to grid-based data, modeling state transitions in two dimensions. Unlike 1D-HMM, which process sequences, 2D-HMMs capture spatial dependencies in images by defining states over pixel neighborhoods [6]. Recent work applies 2D-HMMs to:

- Document analysis: recognizing handwritten text by modeling character co-occurrences [22].
- Medical imaging: segmenting tumors by encoding spatial priors [13]. However 2D-HMMs are computationally intensive and often require approximations for large-scale tasks [23].

2.3. Markov random fields

MRFs model higher-order dependencies by defining potentials over cliques (groups of nodes). Recent advances focus on:

- Parsimonious higher-order MRFs: reducing computational complexity while retaining accuracy [24].
- Deep learning integration: combining MRFs with CNNs for tasks like image denoising [11] and pose estimation. MRFs excel in structured prediction but require careful tuning of potential functions to avoid overfitting [25], [26].

2.4. Variational autoencoders

VAEs learn probabilistic latent representations by maximizing a variational lower bound [15]. Recent extensions, such as conditional VAEs [14] and β-VAEs [27], improve disentanglement and robustness. In recognition tasks, VAEs:

- Reduce overfitting: by regularizing latent spaces [17].
- Handle missing data: through probabilistic inference [28]. For example, VAEs combined with CNNs achieve 98.7% accuracy on MNIST with 30% corrupted pixels [2].

3. METHOD

3.1. Architecture design

The proposed enhanced hybrid CNN-2D-HMM framework integrates four complementary components to address the limitations of individual approaches while leveraging their respective strengths. The framework combines a CNNs backbone for hierarchical feature extraction, 2D-HMMs for spatial dependency modeling, MRFs for higher-order context capture, and VAEs for robust latent representation learning. The architecture design incorporates four key components working in synergy:

- CNN: a ResNet-50 backbone extracts features [29].
- 2D-HMM: models spatial dependencies using forward-backward algorithms [22].
- MRF: captures higher-order interactions via clique potentials [30].
- VAE: learns latent representations to reduce overfitting [15].

3.2. Training strategy

The training strategy employs a hybrid loss function that combines multiple objectives to optimize the entire framework jointly. The loss function is formulated as (1).

$$\mathcal{L} = \mathcal{L}_{recog} + \lambda_1 \mathcal{L}_{HMM} + \lambda_2 \mathcal{L}_{MRF} + \lambda_3 \mathcal{L}_{VAE} \tag{1}$$

Where $\mathcal{L}_{recog}$ represents the primary recognition loss, $\mathcal{L}_{HMM}$ captures spatial coherence through 2D-HMM likelihood, $\mathcal{L}_{MRF}$ enforces higher-order context consistency, and $\mathcal{L}_{VAE}$ provides latent space regularization. The hyperparameters λ_1, λ_2, and λ_3 are carefully tuned to balance thecontributions of each component, ensuring that the framework benefits from all integrated modules without any single component dominating the learning process. The optimization procedure follows established best practices for GPU programming [31].

4. RESULTS AND DISCUSSION

4.1. Quantitative analysis

The proposed hybrid framework was rigorously evaluated on the MNIST [32] and CIFAR-10 [33] datasets, achieving state-of-the-art performance across multiple evaluation metrics. The comprehensive evaluation included accuracy, precision, recall, F1-score, and inference time measurements to provide a thorough assessment of system performance. The experimental setup utilized standardized training and testing protocols to ensure fair comparison with existing methods, and all experiments were conducted using identical hardware configurations to maintain consistency in computational performance measurements.

4.1.1. MNIST dataset

On the MNIST dataset, the framework achieved an accuracy of 98.2%, surpassing all baseline models. The breakdown of performance metrics is as follows in Table 1. This table summarizes the accuracy, precision, recall, F1-score, and inference time of various models on the MNIST dataset. Key observations: i) the integration of MRFs improved recall by ~1%, highlighting their ability to capture higher-order dependencies and ii) VAEs enhanced precision by ~0.8%, demonstrating their effectiveness in reducing noise and improving robustness.

Table 1. Performance metrics on MNIST dataset

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Inference time (ms)
Standalone CNN	96.5	96.3	96.4	96.3	2.1
CNN-2D-HMM	97.1	97.0	97.2	97.1	2.8
CNN-VAE	97.4	97.3	97.5	97.4	2.5
CNN-MRF	97.8	97.7	97.9	97.8	3.2
Proposed framework	98.2	98.1	98.3	98.2	3.5

4.1.2. CIFAR-10 dataset

On the more challenging CIFAR-10 dataset, the framework achieved an accuracy of 89.5%, outperforming existing methods. The results are summarized in Table 2. This table presents the accuracy, precision, recall, F1-score, and inference time of different models on the more challenging CIFAR-10 dataset. Key insights: i) the combination of 2D-HMM and MRFs significantly boosted recall by ~1.4%, addressing challenges posed by cluttered scenes and ii) VAEs reduced overfitting, improving generalization on unseen test data.

Table 2. Performance metrics on CIFAR-10 dataset

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Inference time (ms)
Standalone CNN	85.2	84.9	85.1	85.0	4.2
CNN-2D-HMM	86.7	86.5	86.8	86.6	4.9
CNN-VAE	87.3	87.1	87.4	87.2	4.6
CNN-MRF	88.1	88.0	88.2	88.1	5.3
Proposed framework	89.5	89.4	89.6	89.5	5.7

4.2. Qualitative analysis

For a deeper understanding of the performance and the behaviour of the proposed model, extensive qualitative evaluations were carried out based on advanced visualisation tools and detailed case studies. These qualitative analyses offer insights into the contribution of the combined components for system performance and constantly prove significant in understanding the model behaviour under different situations. For example, the attention heatmaps revealed which input features were deemed important for decision-making by the model, and case studies showed edge cases where the model performed particularly well or poorly at a given task, informing potential focused improvements in future iterations.

4.2.1. Visualization of latent representations

We analyzed the latent space learned by the VAE component using t-distributed stochastic neighbor embedding (t-SNE) [34]. The visualization revealed well-separated clusters for each class, indicating that the VAE effectively disentangled task-relevant features. For example: i) in MNIST, digits with similar shapes (e.g., "3" and "8") were grouped closer together, reflecting their structural similarities and ii) in CIFAR-10, objects with shared attributes (e.g., vehicles like cars and trucks) formed distinct but proximate clusters.

4.2.2. Attention maps

Using the SE blocks in the CNN, we generated attention maps to highlight regions of interest. These maps demonstrated that the model focused on discriminative regions. Specifically, it highlighted i) the central strokes of handwritten digits in MNIST and ii) key object boundaries in CIFAR-10 (e.g., airplane wings, bird feathers).

4.2.3. Failure cases

Despite its high accuracy, the model occasionally misclassified ambiguous samples. For instance: i) in CIFAR-10, images with heavy occlusions or low contrast (e.g., a cat partially hidden behind furniture) were challenging and ii) in MNIST, heavily stylized digits (e.g., a "7" written with additional strokes) caused confusion. These failure cases underscore the importance of incorporating domain-specific priors or augmentations to handle edge cases.

4.3. Ablation study

To evaluate the contribution of each component, we performed an ablation study by systematically removing components from the framework. The results on CIFAR-10 are shown in Table 3. This table shows the impact of removing individual components (VAE, MRF, and 2D-HMM) from the proposed framework, highlighting their contributions to overall performance. Key findings: i) removing the VAE led to a 1.3% drop in accuracy, emphasizing its role in robustness; ii) removing the MRF resulted in a 1.6% drop, highlighting its importance in modeling higher-order context; and iii) removing the 2D-HMM caused a 3% drop, underscoring its critical role in spatial modeling.

Table 3. Ablation study results on CIFAR-10

Configuration	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Full framework	89.5	89.4	89.6	89.5
Without VAE	88.2	88.1	88.3	88.2
Without MRF	87.9	87.8	88.0	87.9
Without 2D-HMM	86.5	86.4	86.6	86.5
CNN only	85.2	84.9	85.1	85.0

4.4. Computational efficiency

While the proposed framework achieves superior accuracy, it incurs additional computational overhead compared to standalone CNNs. The inference times per image are summarized in Table 4. This table compares the inference times (in milliseconds) of the proposed framework and baseline models on both MNIST and CIFAR-10 datasets. Despite the increased computation, the framework remains practical for real-time applications, especially when deployed on modern GPUs.

Table 4. Computational efficiency comparison

Model	MNIST (ms)	CIFAR-10 (ms)
Standalone CNN	2.1	4.2
CNN-2D-HMM	2.8	4.9
CNN-VAE	2.5	4.6
CNN-MRF	3.2	5.3
Proposed framework	3.5	5.7

4.5. Comparison with state-of-the-art methods

We further evaluated our framework using the CIFAR-10 benchmarks and compared it with other methods to demonstrate that its performance in image classification is competitive. We compare our method with vision transformers, EfficientNet, and ResNet on accuracy, precision, recall and F1-score in Table 5 showing that our model has better or similar performance on all metrics. Notably, our model achieves a higher precision and F1-score compared to vision transformer, and its relatively low cost in inference makes it more efficient for practical deployment. CIFAR-10 achieves state-of-the-art accuracy of 98.2% and 89.5%, respectively, in a noise/occlusion robust setting, compared to recent methods. Ablation studies show that MRFs increase recall by 1.6%, and VAEs improve precision by 1.3%, showing their complementarity. This work establishes a connection between deep learning and probabilistic graphical models, enabling the development of interpretable, scalable, and more accurate recognition systems. Our framework outperforms both traditional CNNs and transformer-based architectures, demonstrating the synergy between CNNs, 2D-HMM, MRFs, and VAEs.

Table 5. Comparison with state-of-the-art methods on CIFAR-10

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Vision transformers [2]	88.4	88.3	88.5	88.4
EfficientNet-B0 [3]	87.9	87.8	88.0	87.9
ResNet-50 [4]	86.5	86.4	86.6	86.5
Proposed framework	89.5	89.4	89.6	89.5

5. CONCLUSION

In this paper, we have a way to an enhanced hybrid CNN-2D-HMM framework exploiting MRFs and VAEs in order to enhance spatial modeling and system robustness for recognition purposes. The proposed framework tackles several challenges such as long-range spatial dependencies; higher order-contextual relationships and noise suppress ification by exploiting the synergy among CNN feature extraction, 2D-HMM based spatial modeling, MRF contextual analysis and VAE latent representation learning. It reported state-of-the-art performance on MNIST with 98.2% accuracy, and with CIFAR-10 the accuracy was 89.5%. Through an extensive experimental analysis, we demonstrate the clear advantage of each component: 2D-HMM spatial modelling elevates accuracy by 3%, MRF context capture by another 1.6% and VAE robustness by further 1.3%. The proposed framework possesses computational efficiency and achieves significant performance enhancements available for actual applications. Salient future directions may include: scaling up through more efficient 2D-HMM approximation algorithm; hardware-friendly architecture, c.f. EfficientNet or vision transformers (compare linear-sampler), domain adaptation in medical/satellite imaging applications, theoretical understanding of the hybrid loss function train(). More importantly, this framework can inspire other real-time conditional optimization tasks another promising direction is to improve the explainability with intricate probabilistic components for interpretable predictions and uncertainty quantification. The framework is a major move toward closing the gap between deep learning and probabilistic graphical models, which serves as a solid foundation for more advanced, interpretable and robust recognition systems in real-world visual tasks.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Driss Naji	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	
Kamal Elhattab	✓	✓		✓	✓	✓		✓	✓	✓	✓			
Abdelali Joumad	✓	✓	✓		✓	✓	✓		✓	✓		✓		
Abdelouahed Ait Ider	✓	✓		✓	✓	✓			✓	✓		✓		
Abdelkbir Ouisaadane	✓	✓	✓	✓	✓	✓	✓		✓	✓		✓		
Azzedine Idhmad	✓	✓			✓	✓			✓	✓				

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are publicly available benchmark datasets. The MNIST dataset is openly accessible at http://yann.lecun.com/exdb/mnist/, and the CIFAR-10 dataset is openly accessible at https://www.cs.toronto.edu/~kriz/cifar.html. The trained model weights, experimental configurations used in this study are available from the corresponding author, [DN], upon reasonable request.

REFERENCES

[1] M. Aly, A. Ghallab, and I. S. Fathi, "Enhancing facial expression recognition system in online learning context using efficient deep learning model," *IEEE Access*, vol. 11, pp. 121419–121433, 2023, doi: 10.1109/ACCESS.2023.3325407.

[2] M. Shafiq and Z. Gu, "Deep residual learning for image recognition: a survey," *Applied Sciences*, vol. 12, no. 18, Sep. 2022, doi: 10.3390/app12188972.

[3] N. Deng, Z. L. X. Xu, C. Gao, and X. Wang, “Deep learning and face recognition: face recognition approach based on the DS-CDCN algorithm,” *Applied Sciences*, vol. 14, no. 13, 2024, doi: 10.3390/app14135739.

[4] A. Dosovitskiy *et al.*, “An image is worth 16x16 words: transformers for image recognition at scale,” Jun. 2021, *arXiv:2010.11929.*

[5] J. Zhang, Q. and R. J. Zhang, Y. Zhao, and J. Liu, “Spatial-context-aware deep neural network for multi-class image classification,” in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 1960–1964. doi: 10.1109/ICASSP43922.2022.9746921.

[6] N. Manouchehri and N. Bouguila, “Human activity recognition with an HMM-based generative model,” *Sensors*, vol. 23, no. 3, Jan. 2023, doi: 10.3390/s23031390.

[7] G. Manoharan and K. Sivakumar, “A modified hidden Markov model for outlier detection in multivariate datasets,” *International Journal of Engineering Systems Modelling and Simulation*, vol. 15, no. 3, pp. 121–128, 2024, doi: 10.1504/IJESMS.2024.138287.

[8] I. Murataj *et al.*, “Hyperbolic metamaterials via hierarchical block copolymer nanostructures,” *Advanced Optical Materials*, vol. 9, no. 7, Apr. 2021, doi: 10.1002/adom.202001933.

[9] Y. Yuan *et al.*, “Two-dimensional nanomaterials as enhanced surface plasmon resonance sensing platforms: design perspectives and illustrative applications,” *Biosensors and Bioelectronics*, vol. 241, 2023, doi: 10.1016/j.bios.2023.115672.

[10] F. Liu, W. Yang, and J. Li, “A finite element method for hyperbolic metamaterials with applications for hyperlens,” *SIAM Journal on Numerical Analysis*, vol. 62, no. 3, pp. 1420–1442, Jun. 2024, doi: 10.1137/23M1591207.

[11] P. Pan, C. Zhang, J. Sun, and L. Guo, “Multi-scale conv-attention U-Net for medical image segmentation,” *Scientific Reports*, vol. 15, no. 1, Apr. 2025, doi: 10.1038/s41598-025-96101-8.

[12] W. Dong, B. Du, and Y. Xu, “Shape-intensity-guided U-net for medical image segmentation,” *Neurocomputing*, vol. 610, Dec. 2024, doi: 10.1016/j.neucom.2024.128534.

[13] J. Shedbalkar and K. Prabhushetty, “Deep transfer learning model for brain tumor segmentation and classification using UNet and chopped VGGNet,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 33, no. 3, pp. 1405–1415, Mar. 2024, doi: 10.11591/ijeecs.v33.i3.pp1405-1415.

[14] S. Chen and W. Guo, “Auto-encoders in deep learning—a review with new perspectives,” *Mathematics*, vol. 11, no. 8, 2023, doi: 10.3390/math11081777.

[15] Y. Zhao and S. Linderman, “Revisiting structured variational autoencoders,” in *International Conference on Machine Learning*, 2023, pp. 42046–42057.

[16] J. Huang, W. Yan, G. Li, T. Li, and S. Liu, “Learning disentangled representation for multi-view 3D object recognition,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 2, pp. 646–659, 2021, doi: 10.1109/TCSVT.2021.3062190

[17] I. D. Mienye and T. G. Swart, “A comprehensive review of deep learning: architectures, recent advances, and applications,” *Information*, vol. 15, no. 12, Nov. 2024, doi: 10.3390/info15120755.

[18] M. Tan and Q. V. Le, “EfficientNet: rethinking model scaling for convolutional neural networks,” in *International conference on Machine Learning*, Sep. 2020, pp. 6105–6114.

[19] Q. Zhang, Y. Xu, J. Zhang, and D. Tao, “ViTAEv2: vision transformer advanced by exploring inductive bias for image recognition and beyond,” *International Journal of Computer Vision*, vol. 131, no. 5, pp. 1141–1162, May 2023, doi: 10.1007/s11263-022-01739-w.

[20] M. A. Hasan and K. Dey, "Depthwise separable convolutions with deep residual convolutions," Nov. 12, 2024, *arXiv:2411.07544.*

[21] D. Cakir and N. Arica, “Cascading CNNs for facial action unit detection,” *Engineering Science and Technology, an International Journal*, vol. 47, Nov. 2023, doi: 10.1016/j.jestch.2023.101553.

[22] J. Ma, Z. Wang, and J. Du, “An open-source library of 2D-GMM-HMM based on Kaldi Toolkit and its application to handwritten Chinese character recognition,” in *International Conference on Image and Graphics*, 2021, pp. 235–244. doi: 10.1007/978-3-030-87355-4_20.

[23] A. Joumad, A. El Moutaouakkil, A. Nasroallah, O. Boutkhoum, F. Rustam, and I. Ashraf, “Unsupervised statistical image segmentation using bi-dimensional hidden Markov chains model with application to mammography images,” *Journal of King Saud University - Computer and Information Sciences*, vol. 35, no. 9, Oct. 2023, doi: 10.1016/j.jksuci.2023.101715.

[24] Q. Hu, F. Wang, J. Fang, and Y. Li, “Semantic labeling of high-resolution images combining a self-cascaded multimodal fully convolution neural network with fully conditional random field,” *Remote Sensing*, vol. 16, no. 17, 2024, doi: 10.3390/rs16173300.

[25] Y. Wang, H. Zhang, S. Wang, Y. Long, and L. Yang, “Semantic combined network for zero-shot scene parsing,” *IET Image Processing*, vol. 14, no. 4, pp. 757–765, Mar. 2020, doi: 10.1049/iet-ipr.2019.0870.

[26] W. Shafik, A. Tufail, C. L. D. Silva, and R. A. A. H. M. Apong, “A novel hybrid inception-xception convolutional neural network for efficient plant disease classification and detection,” *Scientific Reports*, vol. 15, no. 1, 2025, doi: 10.1038/s41598-024-82857-y.

[27] H. Meshkin *et al.*, “Harnessing large language models’ zero-shot and few-shot learning capabilities for regulatory research,” *Briefings in Bioinformatics*, vol. 25, no. 5, Jul. 2024, doi: 10.1093/bib/bbae354.

[28] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in *International Conference on Machine Learning*, Jul. 2020, pp. 1597–1607.

[29] S. R. Kempanna *et al.*, “Revolutionizing brain tumor diagnoses: a ResNet18 and focal loss approach to magnetic resonance imaging-based classification in neuro-oncology,” *International Journal of Electrical and Computer Engineering*, vol. 14, no. 6, pp. 6551-6559, Dec. 2024, doi: 10.11591/ijece.v14i6.pp6551-6559.

[30] Y. Wang, C. Ying, X. Luo, and T. Yu, “NeuroLifting: neural inference on Markov random fields at scale,” May 2025, *arXiv:2411.18954.*

[31] G. Xu, X. Wang, X. Wu, X. Leng, and Y. Xu, “Development of residual learning in deep neural networks for computer vision: a survey,” *Engineering Applications of Artificial Intelligence*, vol. 142, Feb. 2025, doi: 10.1016/j.engappai.2024.109890.

[32] P. Tsirtsakis, G. Zacharis, G. S. Maraslidis, and G. F. Fragulis, “Deep learning for object recognition: a comprehensive review of models and algorithms,” *International Journal of Cognitive Computing in Engineering*, vol. 6, pp. 298–312, Dec. 2025, doi: 10.1016/j.ijcce.2025.01.004.

[33] N. Bhatt, N. Bhatt, P. Prajapati, V. Sorathiya, S. Alshathri, and W. El-Shafai, “A data-centric approach to improve performance of deep learning models,” *Scientific Reports*, vol. 14, no. 1, Sep. 2024, doi: 10.1038/s41598-024-73643-x.

[34] L. Xia, C. Lee, and J. J. Li, “Statistical method scDEED for detecting dubious 2D single-cell embeddings and optimizing t-SNE and UMAP hyperparameters,” *Nature Communications*, vol. 15, no. 1, Feb. 2024, doi: 10.1038/s41467-024-45891-y.

BIOGRAPHIES OF AUTHORS

Driss Naji received his doctorat degree in Computer Science from the Faculty of Science and Technique, Sultan Moulay Slimane University, Beni Mellal, Morocco, in 2024. Currently, he is enrolled in the TIAD laboratory at the same university, where their research focuses on the AI and IoT. Their work aims to advance the integration of AI and IoT technologies, contributing to the development of innovative solutions in these fields. He can be contacted at email: naji.drisss@gmail.com.

Kamal Elhattab received a Ph.D. degree in Computer Science from the Faculty of Science, Chouaib Doukkali University, El Jadida, Morocco, in 2024. Currently, he is enrolled in the ELITES laboratory at the same university, where their research focuses on the internet of things (IoT) and artificial intelligence. Their work aims to advance the integration of IoT and AI technologies, contributing to the development of innovative solutions in these fields. He can be contacted at email: kamal.elhattab@gmail.com.

Abdelali Joumad received a Ph.D. degree in Mathematics Applied to Computer Science from the Faculty of Science, Chouaib Doukkali University, El Jadida, Morocco, in 2024. Currently, he is enrolled in the LAROSERI laboratory at the same university, where their research specializes in advanced computational methods for image analysis, with a particular focus on probabilistic modeling and machine learning techniques. His work explores the integration of hidden Markov models, deep learning architectures, and hybrid frameworks to address complex challenges in image segmentation and pattern recognition. He can be contacted at email: ajoumad@yahoo.fr.

Abdelouahed Ait Ider received a Ph.D. degree in Computer Science from the Faculty of Sciences and Technologies, Sultan Moulay Slimane University, Beni-mellal, Morocco, in 2018. Professor of Computer Science at Polydisciplinary Faculty of Taroudant, Ibnou Zohr University, Morocco. His research spans artificial intelligence, machine learning, data science and computer vision, with a focus on deep learning. He can be contacted at email: a.aitider@uiz.ac.ma.

Abdelkbir Ouisaadane is a doctor in Computer Science and researcher in Artificial Intelligence at Sultan Moulay Slimane University (USMS) in LIMATI laboratory, received a B.Sc. degree in Mathematics and Computer Science and an M.Sc. degree in applied mathematics from the Faculty of Science and Technics Beni Mellal, Morocco, in 2010 and 2014, respectively. He is currently professor of Mathematics and researcher in Computer Science and Signal Processing from Sultan Moulay Slimane University, Morocco. His research interests include speech and speaker He can be contacted at email: abdelkbir.ouisaadane@gmail.com.

Azzeddine Idhmad received a Ph.D. degree in Computer Science from the Laboratory of Analysis, Geometry, and Applications Faculty of Sciences Ibn Tofail University Kenitra,Morocco. His research spans artificial intelligence, machine learning, data science and computer vision, with a focus on deep learning. He can be contacted at email: azzeddine.idhmad@uit.ac.ma.