Baghdad Science Journal

Volume 23 | Issue 6 Article 25

6-26-2026

Effective Scene Classification in Arabic Films Using Custom and Pre-trained CNN Architectures

Asraa Muayed Abdalah
Department of Computer Science, College of Science for Women, University of Baghdad, Baghdad, Iraq,
asraa.m@csw.uobaghdad.edu.iq

How to Cite this Article

Abdalah, Asraa Muayed (2026) "Effective Scene Classification in Arabic Films Using Custom and Pre-trained CNN Architectures," *Baghdad Science Journal*: Vol. 23: Iss. 6, Article 25.
DOI: https://doi.org/10.21123/2411-7986.5339

BAGHDAD SCIENCE JOURNAL 2026;23(6):2286–2307

RESEARCH ARTICLE

Effective Scene Classification in Arabic Films Using Custom and Pre-trained CNN Architectures

Asraa Muayed Abdalah

Department of Computer Science, College of Science for Women, University of Baghdad, Baghdad, Iraq

ABSTRACT

Because of their rich visual diversity, unique cinematic styles, and the dearth of expert-annotated datasets devoted to Arab cinema, accurately classifying scenes in Arabic films is still a challenging task. A novel Convolutional Neural Network (CNN) architecture created particularly for Arabic film scene recognition will be discussed in the current research. The main goal of the study is to figure out how well the proposed CNN recognises and classifies Arabic scenes from films when compared with existing pre-trained architectures. According to the findings of six transfer learning models, a comparative analysis was carried out. We developed a new dataset that contains 28,298 images from ten different Arab films. The dataset is utilized to train the models for three epochs at the beginning. While training, we used the same hyperparameters, data splits, and preprocessing pipeline to train each model in order to ensure methodological consistency. Then, a comparison was made of the test performance of each model. Within 15 epochs, the created CNN had 97.23% test accuracy; thus, it showed high accuracy with comparatively little training effort. Among the evaluated pre-trained models on the AMSD dataset, the MobileNet model was the most successful with a 98% test accuracy. So the created CNN ranked second overall with a test accuracy of 97.23%, showing its usefulness for categorizing Arabic cinematic scenes. This paper highlights the importance of selecting appropriate model architectures and training strategies for scene classification, supported by quantitative ablation and confusion analyses.

Keywords: Arabic film scene classification, CNN, Deep learning, Hyperparameter tuning, Transfer learning, Scene analysis, Arabic cinema dataset, Cultural heritage

Introduction

Deep neural networks are applied across multiple domains, addressing challenges such as deepfake detection,[1,2], and excel in medical applications, including COVID-19 identification from X-rays.[3] In dermatology, deep learning enhances the detection of skin conditions through advanced image analysis techniques[4]. It also shows great potential in natural language processing tasks, including Arabic keyword extraction.[5] Moreover, deep learning demonstrates flexibility in traffic-light recognition systems.[6] Its predictive capabilities also support educational analytics, particularly in predicting student dropout rates.[7–9] Deep learning algorithms also strengthen facial recognition systems against face-morphing attacks.[10] Recently, deep learning usage in the field of movie scene detection has led to the expansion of artificial intelligence applications.

Addressing the gap identified by recent reviews, we note that many existing studies primarily target general-purpose image datasets (e.g., ImageNet) and do not explicitly handle cinematic domain-specific issues such as film color-grading patterns, camera-motion artifacts, costume/set styles, and narrative-driven visual motifs. Therefore, in this paper, a domain-specific dataset (AMSD) is created, and both model architectures and training protocols, such as domain-specific augmentation, class balancing, and learning-rate schedules, are adapted to reduce the domain gap between natural images and cinematic frames.

Received 13 May 2025; revised 12 January 2026; accepted 19 January 2026.
Available online 26 June 2026

E-mail address: asraa.m@csw.uobaghdad.edu.iq (A. M. Abdalah).

https://doi.org/10.21123/2411-7986.5339

In the context of Arabic cinema, visual scenes usually include culturally distinctive elements- such as traditional clothes (e.g., thobe, keffiyeh, hijab), architectural motifs (e.g., windows in mashrabiya, desert forts, alleys of the souq), and environmental settings for specific regions (e.g., coastal towns in the Gulf, mountains in Levantine villages, or North African medinas). Domain-specific visual cues enrich the cinematic narrative and serve as potential discriminative features for accurately categorizing scenes.

When training the general-purpose models on Western-centric datasets like ImageNet, these features are rarely leveraged or captured. Our CNN acknowledges this gap and, according to it, designs the dataset curation and training protocols of the models so as to exploit and preserve these cultural visual signatures.

The term “scene” denotes a visually consistent film segment, captured under stable cinematic conditions such as lighting, the angle of the camera, and spatial composition or scene layout. This definition emphasizes visual consistency over narrative or editorial divisions, which suits frame-based scene classification approaches.

The lack of actors and customs in the outdoor shots makes them the main problem in film analysis because they miss distinctive features and are considered visually repetitive, making it difficult to determine which film they belong to.[11] This similarity complicates the task of scene identification within diverse datasets, such as the one developed in this study. This visual similarity between scenes from different Arab films is shown in Fig. 1. Conversely, indoor scenes are generally easier to recognize, recall, and categorize. Specifically, to mitigate high inter-scene visual similarity, we employ targeted preprocessing (color stabilization and histogram-based normalization), domain-aware augmentation (color jittering, random crops reflecting cinematic framing), and architectural measures (progressive convolutional depth, batch normalization, and dropout) designed to emphasize discriminative features rather than low-level appearance alone.

Several difficulties still exist when it comes to CNN-based movie scene detection despite developments in CNNs:

- Visual Similarities:[12] Movie scenes, particularly those on the outside, may share visual cues that make it difficult for CNNs to discriminate between scenes from various movies. Scenes featuring natural landscapes, cityscapes, or generic indoor environments may lack distinctive characteristics. This can be addressed because the custom CNN and the fine-tuning procedure focus on learning higher-level, context-aware features (via deeper filters and regularization) while using balanced sampling and per-class augmentation to reduce class confusion.
- Variability in Lighting and Composition:[13] Across and within individual movies, illumination and camera angles can vary dramatically. The ability of models to generalize across different scenarios and reliably identify them may be impacted by these changes. This can be addressed by applying domain-specific augmentation (illumination augmentation, gamma correction, and random geometric transformations) and normalization steps during preprocessing to improve robustness to lighting and compositional variability; additionally, learning-rate scheduling helps stabilize fine-tuning on cinematic images.
- Limited Context:[14] A single visual frame often lacks sufficient information to correctly identify a movie scene. Temporal context is frequently advantageous for scene identification, but it is difficult for CNNs to extract this information from individual frames. Our current study establishes a strong frame-level baseline. In the future, the model can be extended to capture temporal

Fig. 1. Visually similar outdoor scenes from different Arabic films in the AMSD dataset.

context. Therefore, we exploit sequential dependencies when sequences are available by outlining the use of frame stacking, temporal pooling, or hybrid CNN–RNN/Transformer modules.

- Ambiguity in Scene Content:[15] Some movie sequences may be purposefully ambiguous or abstract, making it challenging for CNNs to deduce the context or meaning of the scenario. Some of these scenarios may rely on artistic or symbolic components that transcend easy categorization. To mitigate ambiguity, our evaluation includes per-class confusion analyses and visualization of discriminative features (e.g., Grad-CAM examples) to inspect when the model relies on semantic cues versus low-level artifacts. We also discuss attention-based extensions as a route to focus on semantically salient regions.
- Scene Transitions:[12] Movies often employ fades or cuts to transition between scenes, which may confuse CNNs and lead to misclassification or missed transitions. We mitigate the influence of transitions on frame-level classification by excluding obvious transition frames during annotation and applying temporal smoothing in post-processing to reduce misclassification at cut/fade boundaries (as described in Methods/Experiments). Future work will integrate explicit shot-boundary detectors to pre-filter transition frames.
- Large and Diverse Datasets:[16] With so many movies and scenes to choose from, it might be difficult to build a thorough dataset for movie scene detection. Such databases demand substantial time investment and cost to curate and annotate. To address dataset scarcity in the Arabic film domain, we created the Arab Movie Scene Dataset (AMSD) comprising 28,298 sequences from 10 Arabic films; AMSD is curated with per-frame annotations, metadata, and train/test splits designed to evaluate both within-film and cross-film generalization.
- Fine-Grained Recognition:[17] Certain applications require identifying specific locations or key moments within movie sequences, which poses challenges for achieving high accuracy in particular movie sequences. It can be difficult to achieve great accuracy at this granularity. When fine-grained recognition is required, our architecture supports higher input resolution and progressive receptive-field expansion; we also recommend metric-learning approaches (e.g., contrastive or triplet losses) as future work to discriminate fine-grained instances.
- Computational Resources: Training CNNs for movie scene identification requires significant computer resources and is computationally costly, especially for big datasets. We address computational concerns by comparing lightweight backbones (MobileNet family) with heavier ones (DenseNet) and by designing a compact custom CNN that offers a favorable trade-off between accuracy and parameter count; training/epoch statistics are reported in the Results section.
- Extending the Learned Knowledge to Completely New, Unseen Movies:[12] Generalizing learned knowledge to completely new and unseen movies remains a complex challenge. To accurately represent the essence of sequences from various films, models are needed. Our evaluation protocol includes held-out test splits and per-movie reporting to assess generalization; nevertheless, we acknowledge that full cross-film generalization remains challenging and recommend leave-one-film-out experiments and domain adaptation techniques as future work.
- Object and Character Interactions:[18] Characters and objects frequently interact in scenes, and it can be difficult to identify and include these dynamic features in recognition models. While the present study focuses on scene-level classification, we discuss complementing scene models with object/actor detectors (e.g., using off-the-shelf detectors for faces/objects) in future hybrid pipelines to incorporate interaction cues.

Developing more sophisticated network architectures, integrating temporal information, exploring transfer learning approaches,[19] and improving dataset curation and annotations are frequently necessary to address these issues in CNN-based movie scene recognition. Researchers are still improving the accuracy and robustness of such models for a variety of cinematic applications. In line with these needs, the present work (1) introduces AMSD to reduce dataset bias in Arabic cinema, (2) compares a compact custom CNN against multiple pre-trained backbones with domain-aware fine-tuning, and (3) uses ablation and confusion analyses to quantify the contribution of architectural choices, data augmentation, and training schedules toward mitigating the issues listed above.

The motivation for our research arises from the inherent challenges in scene recognition, particularly in the context of cinematic content. The existing methods addressed limitations; therefore, this research proposes a novel CNN-based approach that makes use of both established and modern innovations in deep learning techniques to improve scene classification accuracy. To achieve this, we carry out an extensive assessment of the performance of nine various categorization techniques in a methodical manner,

highlighting their capabilities within deep learning systems. However, the main research challenge addressed in this work lies in the restricted capacity of existing CNN-based methods to precisely identify scene categorization in Arabic films, which demonstrate high inter-scene visual likeness, inconsistent lighting, and culturally specific visual cues not captured by general-purpose datasets. The majority of earlier investigations have focused on Western cinematic data or standard datasets such as ImageNet, which do not accurately capture the design-related and context-based diversity of Arabic cinema. Therefore, this research attempts to link this gap by constructing a dedicated Arabic movie scene dataset and assessing both custom and pre-trained CNN architectures to determine their efficiency in dealing with these domain-specific challenges.

The main aim is to confront the problematic factors in scene classification in various cinematic datasets. For this aim, we designed and created the Arab Movie Scene Dataset (AMSD), comprising 28,298 movie frames extracted from 10 Arab films, ensuring visual context diversity and well-balanced distribution. The variation of movie scenes in the AMSD set illustrates the complex interconnected layers of cinematic scenes and offers samples of visual data that are diverse and representative. By constructing this dataset, we are able to assess and fine-tune multiple deep learning algorithms customized for Arabic cinema. This research explores several deep learning algorithms and shows their ability to improve the precision in film scene classification (i.e., the proposed approach uses several transfer learning pre-trained CNN architectures and applies fine-tuning to tackle existing problems in scene classification methods. The proposed approach enhances the reliability, accuracy, and applicability in the classification of scene models.

This work is important for pushing cutting-edge techniques in the field, with direct application to practical scenarios such as recommendation systems and content-based video indexing, which still depend on scene classification. Furthermore, the proposed approach demonstrates the scalability and efficiency of AI models on large datasets, ensuring that the system can continue to perform accurately as new data is introduced. The contribution of this work is enhancing recognition and classification accuracy in complex visual datasets by combining different algorithms to leverage their strengths and to avoid their limitations. This research contribution is summarized as follows:

- A novel Arabic Movie Scene Dataset (AMSD) is developed. It provides a unique benchmark for cinematic scene classification. It consists of 28,298 labeled images that are taken from ten Arabic films.
- High visual similarities are handled by the created convolutional neural network architecture in inter-scene and visual patterns of culturally specific scenes that are presented in Arabic films.
- Multiple pre-trained CNN architectures like VGG, ResNet, and InceptionV3 are evaluated and compared with our created CNN model to determine the best configuration for Arabic film scene classification.
- Using preprocessing and augmentation on the dataset to obtain a variety of lighting and composition. By doing this, we increase and improve generalization and robustness.
- Accuracy, precision, recall, F1-score, and confusion matrices are used for performance analysis to measure the efficiency of the proposed CNN.
- Discuss benefits of using artificial CNN in recognizing cultural heritage contents and provide directions for future work.

This work discusses the created CNN experimental design, its model architecture, and provides evaluation of the results from training, testing, and validation. Also, it discusses transfer learning using different existing networks and gives future research directions to enhance movie scene classification. It gives the importance of enhancing Arabic cinema analysis and scene classification in real-world applications such as automated indexing, recommendation systems, and cultural heritage preservation. The remainder of this paper is organized as follows:

Related work is presented in the following section. A brief review of the contributions in scene classification is provided, mentioning gaps in the existing studies. Then, a detailed methodology description is provided, where the experimental design, dataset creation, preprocessing steps, and the proposed CNN architecture alongside the pre-trained models used for comparison are discussed. Then, the results and discussion section provides performance evaluation metrics, comparative analysis, and provides preservation on model behavior and limitations. Finally, the conclusion and recommendations are given, which summarize the main findings, outline the theoretical and practical implications, and suggest directions for future research.

Related works

In their survey, Nada H. A. et al.[20] study the development of learning and its application's role in both human and machine scenarios. The study shows the evolution from traditional machine learning methods

to deep neural architectures. The survey compares different methods. Its goal is to focus on how multilayer learning has changed the fields of (algorithmic and human) learning. Thus, this study provides a summary of the major findings and clarifies how machine learning and deep learning are developing in practical applications.

Salar J. A. et al.[21] study the effect of deep learning on the development of face expression recognition and its role in a variety of applications such as human-computer security interactions. The research tackles issues resulting from insufficient training data, such as expression variations and overfitting. The authors go over the historical context, the value of datasets, and a comparison of current deep learning techniques for facial emotion identification. They also provide recommendations for creating effective deep facial expression detection systems.

Finally, these publication survey papers[22,23] expand the awareness of and knowledge in the area of scene identification by providing in-depth analyses of the literature, making recommendations for model selection, and pointing the way for future study. These surveys can be used as a resource for experts in the area to remain updated on the most recent advancements and difficulties in scene identification.

Mohammad M. T.[24] gave a summary of the essential elements of deep learning, as well as the most recent advancements in the area. He discussed how deep learning has become an integral part of modern AI as well as different deep learning networks and methodologies. A summary of the real-world applications for deep learning techniques was also given by him. Finally, suggested more study after identifying potential traits for next deep learning modeling generations. He gave a thorough introduction to deep learning architectures suitable for both academics and business professionals. Finally, he listed other problems and proposed improvements to assist scholars in identifying the present research limitations. This paper explores various methods, advanced neural network architectures, techniques, and practical applications.

For scene categorization, Sarfaraz M. et al.[25] employed a number of models (VGG-16, Alexnet, and Inception V3). Additionally, they used the PLACES, SUN379, and SUN-WIDE databases. They employed the Karen S and Andrew Z. VGG-16 Model, a multilayered CNN[26] in the 2014 ImageNet Challenge. The 92.7 was a top-5 classification accuracy that was attained by their model. Additionally, they evaluated the Inception-V3 model on two additional datasets, SUN397[27] and NUS-WIDE,[28] containing 4,499 and 9,824 images, respectively. The Inception V3 model performed well here as well, with 94.6% test accuracy and 84.7% for the two independently utilized datasets, respectively, as illustrated in Table 1 below.

Table 1. Accuracies obtained by S. Masood et al.[25]

Model	PLACES Dataset	SUN397 Dataset	NUS-WIDE Dataset
Inception-V3	89.3%	94.6%	84.7%
VGG-16	87.3	-	-
Alexnet	79.8%	-	-

5000 photos are used to train each model to categorize data into 15 categories. The test dataset had 15000 photos, while the training dataset had 6000 images. The 2014 ImageNet Challenge (ILSVRC14) was used to establish the model's weights.[29] Then, once again, the model was trained on the dataset. The most accurate representation has an 89.3% average accuracy.

Using the traditional Alex-Net model and support vector machine as a foundation, Yuanyi C.[30] presented a methodology using Deep Convolutional Neural Networks (DCNNs) for categorizing scenes. This algorithm completely learned the deep properties of the pictures. Then, for scene image classification, they used the Lib-SVM training model and compared it with classification methods based on the regression model; initially, they used the Alex-Net system that analyzes scene information in images and extracted the last layer, which uses the picture properties of 4096 Alex-Net neurons; finally, the tests were run on two widely used datasets (ImageNet 2012 and NUS-WIDE). According to the experimental findings, the accuracy of the DCNN for classifying scenes from images using the ImageNet 2012 dataset was 88.31%, while 84.55% was the accuracy for classifying scenes from images using the NUS-WIDE dataset.

The goal of Mansoor R. et al.[31] study is to label images as either an interior scene or an outdoor setting. Hue, Saturation, and Value, three powerful color qualities, were employed to accomplish autonomous scene classification at the idea level. By using the Gabor filter to achieve lighting and scale invariance, and to evaluate the texture of the picture, robust texture features were produced. Each of the pictures produced by the Gabor convolution was also used to extract the WHGO texture feature. By using PCA, feature dimensions were reduced. Finally, a classifier based on Sparse Representation that classifies the image as an outdoor or indoor scenario was constructed with the intention of classifying the dataset. 442 interior photos and 465 outdoor images were included in the IITM-SCID2 dataset. It was separated into two sets: a 393-photo training set and a test set of 514 photos. On the IITM-SCID2 data collection, this work's overall accuracy was 83.80% or thereabouts.

Table 2. Comparison between previous studies and the proposed research.

Aspect	Previous Studies	Proposed Research (This Study)
Domain and Dataset	Most existing studies rely on public datasets (e.g., ImageNet, Places365) or domain-specific datasets such as Norwegian Sign Language recognition, which lack cinematic and cultural diversity.	A new Arabic Movie Scene Dataset (AMSD) is developed, consisting of 28,298 labeled frames from 10 Arabic films, capturing cultural, environmental, and lighting variations.
Data Annotation	Limited or general-purpose annotations not focused on scene-level semantics.	Detailed manual annotation at the scene level, covering indoor/outdoor, emotional tone, and visual context.
Research Focus	Prior works emphasize general object or gesture recognition, with minimal attention to Arabic cinematic content.	Focused on Arabic film scene categorization — an underexplored domain.
Model Type	Standard pre-trained CNNs (e.g., VGG, ResNet) applied with minimal domain adaptation.	Combines custom-designed CNN architecture and fine-tuned pre-trained models optimized for Arabic film data.
Handling of Visual Similarity	Few studies address high inter-scene visual similarity or complex cinematic frames.	Introduces domain-aware preprocessing, deep feature extraction, and regularization to enhance discrimination between visually similar scenes.
Evaluation Metrics	Many works report only accuracy or precision.	A confusion matrix, as well as accuracy, F1-score, precision, and recall, are used for evaluation.
Cultural Representation	Lacks representation of non-Western or Arabic visual content.	Focuses explicitly on Arabic films to enhance cultural inclusivity and preserve regional cinematic heritage.

A study on algorithms to classify anemia is conducted by Safa S. A. et al.[32] for global and recent CBC data collections. This paper investigates the implementation of automated learning techniques in the medical services sector, specifically for the categorization of anemia diseases. It highlights the value of utilizing digital analytics and categorization methods for extensive healthcare datasets. An evaluation comparing 12 various categorization techniques is conducted by this research, highlighting that Random Forest, Logitboost, Multilayer Perceptron, and XGBoost were effective. XGBoost had the highest accuracy. Therefore, XGBoost is used in Iraq for classifying new Hematology datasets.

A scene classification system named HFCNN (Hybrid Feature Combination Neural Network) was created by Er P. S. and Rajesh M.[18] to identify towering structures, mountains, open spaces, and city interior scenes. Their suggested HFCNN algorithm was a blend of the traditional approaches to extract features: SIFT and HOG, which are often utilized in computer vision. They combined them in an effort to extract information from the input photographs at both the high and low levels. Over 96% accuracy was attained using HFCNN.

Recent research on sign language recognition by Svendsen B and Kadry S.[33] provides important observations on the design of the dataset and problems in labeling consistency. The authors emphasized the importance of domain-specific data structuring, representation of balanced classes, and robust labeling to ensure high recognition accuracy. Although their domain differs from the suggested cinematic scene classification in this paper, both domains suffer from data scarcity, inter-class visual similarity, and limited cultural context representation. As a conclusion, there is a major research gap in Arabic film scene classification because of the lack of large-scale, well-labeled corpora that capture the visual and cultural diversity of Arab cinema. Our proposed AMSD dataset addresses such a gap by providing broad, labeled cinematic frames representing variety in lighting conditions, environments, and cultural settings, thus enabling the building of an effective CNN model. Table 2 presents a comparative summary highlighting key aspects of prior studies and how the proposed work addresses their limitations.

Methodology

The process of developing a classification model from a collection of data that includes class labels is known as classification. A step-by-step approach for building a multilayer learning architecture for scene classification in Arabic films is used. Fig. 2 illustrates that the proposed system follows a systematic workflow consisting of five main stages.

Experimental design

At this stage, the foundation for the system's actual implementation is established. This stage defines the architecture, key elements, modules, and interface. It was desired to build a CNN Architecture as a baseline. The experimental design aims to examine the influence of epoch size, the use of learning rate scheduling, and the use of transfer learning on the results of scene classification. Furthermore, a comparative analysis with pre-trained models was conducted to evaluate the proposed model accuracy.

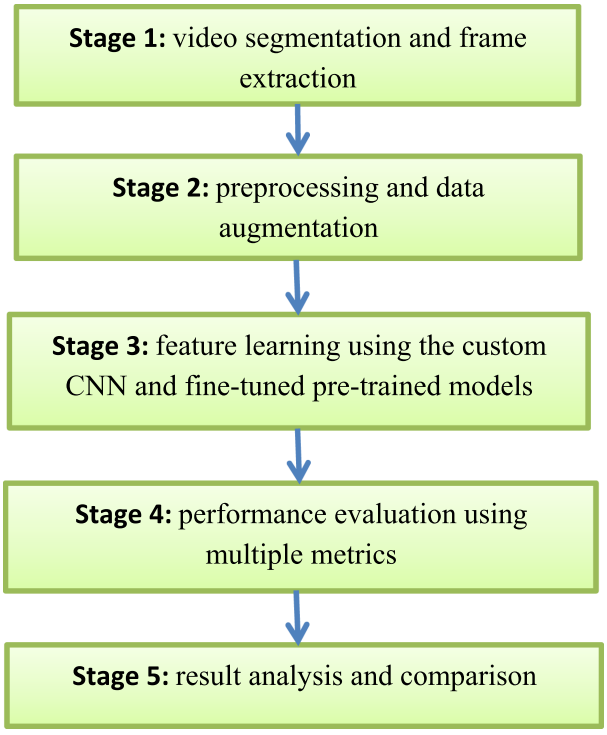

Fig. 2. Overview of the proposed methodology pipeline.

Arab movie scene dataset (AMSD)-creation and preprocessing

AMSD creation

Several existing movie scene datasets have been employed in various tasks, including video interpretation, action identification, and movie recommendation. For a range of machine learning and multimedia-related activities, researchers and developers employ them. The earlier-stated dataset details are in Table 3. These datasets were developed by different academic institutions and organizations. There is no publicly available dataset that exclusively contains scenes from Arabic films; so the Arab Actors Dataset-AAD[34] was created.

The AAD contained 574 images used for predicting a movie's year of production through facial recognition of actors.[38] Then another dataset was created (AMSD) used in this study; it contains 28298 images from 10 Arabic films, categorized into three subsets for training 50% (14131 images), testing 25% (7063 images), and validation 25% (7104 images), which is a reasonable choice for a movie scene classification dataset. We calculated the distribution of film classes for all 10 films; we also calculated the distribution of the Indoor/Outdoor scenes. The results showed that the dataset is considered to be balanced regarding film classes. The number of frames extracted from these films ranges from 1082 to 3881. The distribution of Indoor and Outdoor scenes is 58% for Indoor scenes and 42% for Outdoor scenes. To reduce the difference in ratio, we use class weights and balanced sampling during the training process. Table 4 shows the class distribution.

VLC Media Player version 3.0.16 is used to transform each video into frames because it can capture batch frames reliably and it can control frames accurately. Each film is transformed separately from other Arabic films to ensure a balanced number of Indoor/Outdoor scenes.

Frames were automatically extracted at fixed time intervals to avoid redundancy and to capture significant scene transitions. Specifically, frames were extracted every 50 seconds of video playback, corresponding to approximately one representative frame per scene transition. This interval was empirically selected after preliminary trials for different intervals, where the 50-second rate provided better scene diversity without excessive duplication. Each extracted frame was saved as 1280 × 720 pixels in .PNG format, preserving the visual and contextual details of the cinematic content. Dataset details are shown in Table 5.

Annotation procedure

Two annotators perform scene annotation using a guideline for annotation. This guideline includes

Table 3. summary of the early movie scene datasets.

Dataset	Number of Image Scenes	Purpose	Additional Details	Year of Creation	Created By
Hollywood2 Dataset[35]	1,707 scenes	Action recognition and detection in videos	Annotations for 12 action classes	2009	Researchers at Queen Mary, University of London
MPII Movie Description Dataset[35]	Over 45,000 frames	Video description generation	Detailed annotations for video events	2016	Max Planck Institute for Informatics
HMDB51 Dataset[36]	6,766 scenes	Action recognition in video clips	Covers 51 action classes	2011	University of Central Florida
YouTube-8M Dataset[37]	Millions of video clips	Video classification and annotation	Large-scale dataset covering diverse content	2016	Google Research

Table 4. Class distribution statistics for the AMSD dataset.

No.	Film Title	Number of Frames	Indoor Frames	Outdoor Frames
1	Me انا	2165	1089	1076
2	With My Love and Longing مع حبي واشواقي	2681	1542	1139
3	AlZheimer زهايمر	3475	1803	1672
4	My Daughters, My Life بناتي حياتي	3279	2324	955
5	Streets from Fire شوارع من نار	2228	1481	747
6	My Sweetheart is very Naughty حبيبتي شقية جدا	2803	1594	1209
7	Shaaban Below Zero شعبان تحت الصفر	3264	1640	1624
8	Women in the City نساء في المدينة	2478	1215	1263
9	I am not a Thief انا مش حرامية	2881	1745	1136
10	Danish Exploration التجربة الدنماركية	3044	1980	1064
	Total	**28298**	**16413 (58%)**	**11885 (42%)**

Table 5. AMSD dataset.

No.	Movie Title	Training Set	Testing Set	Validation Set
1	Me انا	1082	541	542
2	With My Love and Longing مع حبي واشواقي	1340	670	671
3	AlZheimer زهايمر	1737	868	870
4	My Daughters, My Life بناتي حياتي	1639	819	821
5	Streets from Fire شوارع من نار	1114	557	557
6	My Sweetheart is very Naughty حبيبتي شقية جدا	1396	698	709
7	Shaaban Below Zero شعبان تحت الصفر	1632	816	816
8	Women in the City نساء في المدينة	1239	619	620
9	I am not a Thief انا مش حرامية	1435	717	729
10	Danish Exploration التجربة الدنماركية	1517	758	769
	Total	**14131**	**7063**	**7104**

visual consistency, conditions of lighting, context of background, presence of an actor, and finally scene boundaries. The agreement between the annotators is measured using Cohen's Kappa ($\kappa = 0.87$), which indicates strong reliability. In the case of disagreement, both annotators review the scenes and choose the final label.

As can be seen from Fig. 1 above, the images are diverse and include interior and exterior scenes, different lighting, situations, and multiple poses for the actors. Here's how can this split be used effectively:

- **Training Set (50%)**: The largest portion of the dataset is the training set. It is used in training the classification model. Model's parameters are adjusted based on the training data to improve its accuracy.
- **Testing Set (25%)**: It is used for the evaluation of training model's performance. Evaluation metrics like the accuracy, F1-score, precision, recall and confusion matrix are widely applied for classification tasks. The F1-score is the mean of precision and recall and it gives a balanced measure of the model's performance. Precision is the ratio of correctly predicted positive instances to all predicted positive instances.[39] This set is used to give an idea about the model's generalization capability. The number of frames for each class (film) in the test set ranges from 541 to 868, ensuring accurate evaluation.
- **Validation Set (25%)**: The model's hyperparameters are fine-tuned and the performance is monitored using this set during training process. It enables training early termination, model optimization while iterations, and hyper-parameters fine-tuning. Also, it helps prevent overfitting by using a separate dataset for model assessment during training.

Preprocessing and data augmentation

In this study, frames are extracted and saved in RGB format. Image frames are not converted to grayscale to avoid loose lightening which is commonly used in shooting different locations and in cinematography

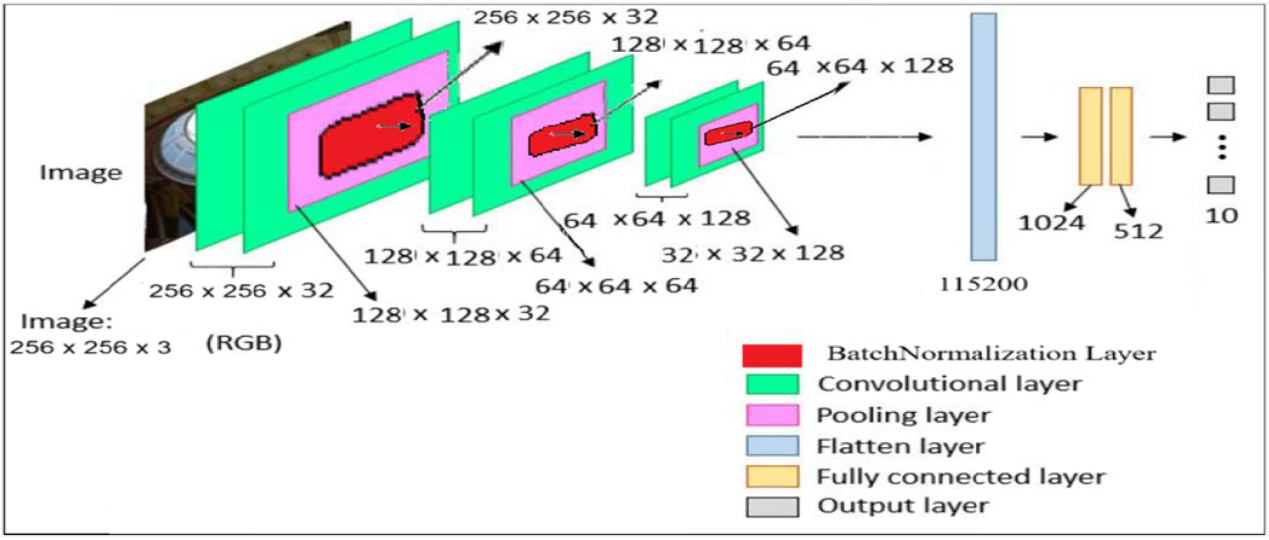

Fig. 3. The proposed custom CNN architecture.

and this will reduce accuracy in training and classification. Also, image sizes are resized to 256 × 256 pixels for computational efficiency. Before splitting the dataset into training (50%), validation (25%), and testing (25%), shuffling is used to prevent ordering bias and to ensure randomness.

We designed Data Augmentation method that suites visual characteristics of Arabic films. Color-grading jitter is used to make brightness (±20%), while contrast (±15%), and finally saturation (±10%). to simulate both differences in lightening conditions and variety in color styling used in Arabic films. Also we used random crops that simulate cinematic framing conventions (medium and wide shots) to reduce the differences between ImageNet pre-training and realistic Arabic cinematic contents. The Ablation Study showed that data augmentation enhances test accuracy by +0.85% over standard augmentation alone.

In addition to data augmentation, we used random horizontal flipping, rotation (±20°), and zoom (±15%) to increase the diversity of data and to enhance model robustness. Data augmentation is applied during training by using TensorFlow's preprocessing layers to ensure not perform the same process on the frame twice and to reduce the probability of overfitting.

Construction of the baseline CNN model

A convolutional neural network (CNN) model was proposed as the baseline architecture the aim was to achieve high accuracy within acceptable computational constraints. The proposed architecture composed of multiple convolutional layers, max-pooling layers, and fully connected layers responsible for feature extraction and classification. Also batch normalization layers are applied after each convolutional layer inorder to enhance training stability and learning and optimization speed. To prevent overfitting, dropout regularization is used. Fig. 3 represents the details of the CNN architecture.

Although the CNN design is relatively straightforward, its performance may vary depending on the dataset complexity and the number of scene categories. The layer configuration and parameters of the proposed CNN model are shown in Table 6.

The following is a quick summary of each layer:

- **Input Layer:** The parameters of this layer are its dimensions, which must match those of the input images. It is assumed that the input images are RGB with the appropriate shape for this design.
- **Convolutional Layers:** There are three layers that are sequentially defined and are used for extracting features from the input images. ReLU activation and a 3 × 3 kernel (filter) are used in each convolutional layer. The number of filters is gradually increased: 32 in the first layer, then 64 in the second layer, and 128 in the third layer.
- **Max-Pooling Layers:** Every convolutional layer is followed by a max-pooling layer. Image details are preserved while reducing the spatial dimensions of the feature maps because of max-pooling. Each max-pooling layer reduces feature maps by using a 2 × 2 pooling window and a stride of 2.
- **Flatten Layer:** A flatten layer is inserted after the convolutional and max-pooling layers. It converts the 2D feature maps into a 1D vector, preparing the data for the fully connected layers.
- **Fully Connected Layers:** The flatten layer is followed by fully connected layers. The first dense layer employs ReLU activation and contains 512

Table 6. Layer configuration and parameters of the CNN model.

Layer Number	Layer Type	Kernel/ Pool Size	Filters/ Units	Activation	Other Parameters	Output Shape	Number of Parameters
1	Input	–	–	–	RGB, 256 × 256	(256, 256, 3)	0
2	Convolutional (Conv2D)	3 × 3	32	ReLU	Padding: same	(256,256, 32)	896
3	Batch Normalization	–	–	–	–	(256,256, 32)	
4	Max Pooling (MaxPooling2D)	2 × 2	–	–	Stride: 2	(128, 128, 32)	0
5	Convolutional (Conv2D)	3 × 3	64	ReLU	Padding: same	(128, 128, 64)	18496
6	Batch Normalization	–	–	–	–	(128, 128, 64)	
7	Max Pooling (MaxPooling2D)	2 × 2	–	–	Stride: 2	(64, 64, 64)	0
8	Convolutional (Conv2D)	3 × 3	128	ReLU	Padding: same	(64, 64, 128)	73856
9	Batch Normalization	–	–	–	–	(64, 64, 128)	
10	Max Pooling (MaxPooling2D)	2 × 2	–	–	Stride: 2	(32, 32, 128)	0
11	Flatten	–	–	–	–	115200	0
12	Dense	–	512	ReLU	–	512	58982528
13	Dropout	–	–	–	Rate: 0.5	512	0
14	Dense (Output)	–	N(10)	Softmax	-	Number of Classes 10	513*10

units. To lessen overfitting, dropout at a rate of 0.5 is applied to this layer. Softmax activation is used by the final dense layer for multi-class classification, with the number of units equal to the number of scene classes in the training dataset. Class probabilities are produced by this layer.

CNN model training

To evaluate its classification performance, the proposed CNN was initially trained on the AMSD dataset for three epochs. In scene classification tasks, the model achieved a test accuracy of 94.73% and a predicted accuracy of 95%, which are considered promising results. The evaluation metrics for the proposed CNN are summarized in Table 7.

At this early stage, MobileNet achieves the highest test accuracy of 98.95%, attributable to its lightweight depthwise separable convolutions that enable rapid adaptation to domain-specific features with minimal fine-tuning. Conversely, ResNet50 has poor performance (51.52%), because of an unstable optimization process when fine-tuning its deep residual blocks using a small, domain-specific dataset. VGG-16 performance is 92.59%, and DenseNet variants 92.14 to 95.54% show moderate results, because of simpler or fully connected networks that have better performance with limited data. Our custom CNN is trained from scratch and achieves 94.73% accuracy, surpassing most pre-trained models except MobileNet. This indicates that a concise, domain-specific architecture is capable of challenging transfer learning when designed to detect culturally specific visual patterns.

The proposed CNN was trained from scratch for 15 epochs to study its convergence behavior, while all pre-trained models were trained for 3 epochs only, and this is because experimental observations showed that using more than 3 epochs will decrease model performance, especially for MobileNet, whose accuracy dropped from 98.95% to 66.41% after 15 epochs (due to catastrophic forgetting on our small, culturally specific dataset). Therefore, 3 epochs were considered the best performance point for all used transfer learning models to ensure a fair and meaningful comparison.

Training protocol and epoch strategy

All experiments were conducted using a unified training protocol to ensure methodological consistency between the proposed CNN and the other pre-trained models. The AMSD dataset is divided into three subsets: training set (50%), validation set (25%), and testing set (25%). The same data loading, preprocessing, and augmentation are implemented by all the models.

Six models were used in transfer learning experiments (VGG-16, ResNet50, InceptionV3, MobileNet, DenseNet121, DenseNet169). All layers were fine-tuned without freezing to enable full adaptation with the Arabic scenes' cinematic and cultural characteristics. The experiments showed that freezing deeper layers would decrease the performance(I.e., limited suitability of high-level ImageNet features for the target domain). All models were trained using the Adam optimizer with a learning rate of 0.001, 32 batch size, **and** the categorical cross-entropy loss function. Then the learning rate was decreased by 0.1 every 5 epochs; to avoid overfitting, early stopping was used.

Pre-trained models were trained for 3 epochs depending on experimental observations. Using more epochs led to decreased performance, especially in lightweight architectures such as MobileNet. This model achieved 98.95% accuracy at epoch 3, and when training for up to 15 epochs, its accuracy decreased sharply because of overfitting and catastrophic forgetting. Similar trends were observed for

Table 7. Evaluation metrics for the models (3 epochs) trained on AMSD dataset.

	VGG-16				ResNet50				Inception V3			
Movie	Precision	Recall	F1-score	support	Precision	Recall	F1-score	support	Precision	Recall	F1-score	support
Danish Exploration	0.99	0.74	0,85	758	0.56	0.80	0.66	758	0.76	0.85	0.80	758
I am not a Thief	0.94	0.86	0.90	717	0.36	0.20	0.25	717	0.72	0.92	0.81	717
Me	0.98	0.89	0.94	541	0.19	0.08	0.11	541	0.75	0.82	0.78	541
My Daughters My Life	0.95	0.95	0.95	819	0.38	0.96	0.55	819	0.91	0.87	0.89	819
My Sweetheart is very Naughty	0.99	0.95	0.97	698	0.97	0.53	0.69	698	0.84	0.95	0.89	698
Shaaban Below Zero	0.96	1.00	0.98	816	0.99	0.45	0.62	816	1.00	0.97	0.98	816
Streets from Fire	0.91	0.92	0.91	557	0.00	0.00	0.00	557	0.95	0.79	0.86	557
With My Love	0.82	0.97	0.89	670	0.61	0.34	0.44	670	0.95	0.77	0.85	670
Women in the City	0.79	0.99	0.88	619	028	0.59	0.38	619	089	0.77	0.83	619
AlZheimer	0.97	0.98	0.97	868	0.85	0.86	0.85	868	0.97	0.88	0.92	868
Test Loss		**0.2372**				**1.4080**				**0.4714**		
Test Accuracy		**0.9259**				**0.5152**				**0.8653**		
	MobileNet				MobileNet V2				DenseNet121			
Movie	Precision	Recall	F1-score	support	Precision	Recall	F1-score	support	Precision	Recall	F1-score	support
Danish Exploration	1.00	0.92	0,96	758	0.95	0.88	0.92	758	0.99	0.70	0.82	758
I am not a Thief	1.00	1.00	1.00	717	0.88	0.93	0.90	717	0.96	0.93	0.94	717
Me	1.00	1.00	1.00	541	0.71	0.97	0.82	541	0.99	0.82	0.90	541
My Daughters My Life	0.97	1.00	0.98	819	0.94	0.90	0.92	819	0.96	0.89	0.93	819
My Sweetheart is very Naughty	1.00	1.00	1.00	698	0.98	0.94	0.96	698	0.99	0.98	0.98	698
Shaaban Below Zero	1.00	1.00	1.00	816	1.00	0.98	0.99	816	1.00	1.00	1.00	816
Streets from Fire	0.98	0.99	0.98	557	0.95	0.86	0.90	557	0.80	0.97	0.87	557
With My Love	0.95	0.99	0.97	670	0.92	0.91	0.91	670	0.95	0.95	0.95	670
Women in the City	1.00	1.00	1.00	619	0.93	0.87	0.90	619	0.94	0.97	0.96	619
AlZheimer	1.00	1.00	1.00	868	0.95	0.95	0.95	868	0.76	1.00	0.86	868
Test Loss		**0.0469**				**0.3408**				**0.2968**		
Test Accuracy		**0.9895**				**0.9206**				**0.9214**		
	DenseNet169				DenseNet201				CNN			
Movie	Precision	Recall	F1-score	support	Precision	Recall	F1-score	support	Precision	Recall	F1-score	support
Danish Exploration	0.91	0.96	0.93	758	0.94	0.94	0.94	758	0.91	0.84	0.87	758
I am not a Thief	0.88	0.99	0.93	717	0.98	0.91	0.94	717	0.99	0.97	0.98	717
Me	0.94	0.95	0.94	541	0.98	0.92	0.95	541	1.00	0.97	0.99	541
My Daughters My Life	0.98	0.94	0.96	819	097	0.95	0.96	819	0.94	0.84	0.89	819
My Sweetheart is very Naughty	0.92	1.00	0.96	698	0.94	0.99	0.96	698	0.97	0.99	0.98	698
Shaaban Below Zero	1.00	1.00	1.00	816	1.00	1.00	1.00	816	1.00	1.00	1.00	816
Streets from Fire	0.99	0.88	0.93	557	0.97	0.90	0.93	557	0.76	0.96	0.85	557
With My Love	0.96	0.89	0.93	670	0.91	0.97	0.94	670	0.90	0.95	0.92	670
Women in the City	0.99	0.95	0.97	619	0.91	0.99	0.94	619	0.99	0.98	0.98	619
AlZheimer	0.99	0.96	0.95	868	0.97	0.97	0.97	868	1.00	1.00	1.00	868
Test Loss		**0.1761**				**0.1875**				**0.1911**		
Test Accuracy		**0.9538**				**0.9554**				**0.9473**		

the other pre-trained models. While the proposed CNN architecture was trained from scratch, this led to requiring more training epochs to learn high-level visual features specific to Arabic cinema. The proposed CNN was trained first for 3 epochs, to be compared directly with other pre-trained models (its accuracy was 94.73%), then it was trained to 15 epochs to analyse its convergence behavior; its accuracy reached 97.23%.

Thus, this research methodology used a two-stage evaluation strategy (equal-epoch comparison and extended convergence analysis) to ensure a fair and transparent comparison between pre-trained models and the proposed CNN architecture without bias because of differences in epoch numbers.

Training parameters and hyperparameter settings

All the experiments are executed in the Google Colab environment (with an NVIDIA Tesla T4 GPU (16 GB VRAM), an Intel Xeon CPU, and 25 GB of system RAM). TensorFlow 2.15 with the Keras API was used to code the models. Adam optimizer was used with a learning rate of 0.001. For all the experiments, the learning rate was reduced by 0.1 after every 5 epochs, and the batch size = 32. Dataset images were stored in Google Drive and were streamed from it instead of locally mounted SSD storage; this introduced I/O overhead and affected overall training time. Thus, the average training time for an epoch was approximately 26.4 minutes for the proposed CNN, MobileNetV2 epoch lasted for approximately 25.0 minutes, and 14.6 minutes for MobileNet (V1) under identical computational and data-loading conditions. These time computations and experiments were obtained using the same device settings and software to ensure fairness in comparison.

All the models were trained for 3 and 15 epochs, respectively, to examine the stability of the convergence. Hyperparameter Optimization for the proposed CNN was performed experimentally by performing systematic experimentation. Validation performance across a set of predefined values commonly used in CNN-based image classification tasks led to learning rate = 0.001, batch size = 32, optimizer = Adam, and dropout rate = 0.5; thus, the choice of the final configuration was made depending on the balance between classification accuracy, training stability, and computational efficiency. Also, a loss function was used since the goal is multi-class classification. Dropout was used to minimize overfitting, with a rate of 0.5 in the fully connected layers. Also, batch normalization was used after every convolutional layer to enhance training stability and to make gradient flow easier. Input images were resized to 256 × 256 pixels and normalized to [0,1] scale before starting the training process.

Also, data augmentation was used, such as random rotation by ±20°, horizontal flipping, brightness jitter by ±10%, and finally random cropping to increase the diversity of the dataset and to enhance the model's ability for generalization. Early stopping and validation monitoring were employed to prevent overtraining.

Increasing the epoch size

When the epoch size was increased to 15, the test accuracy rose to 97.23%, representing a promising result for scene image classification. To measure the proposed Convolutional Neural Network (CNN) performance, a comparison was made between the proposed CNN and the pre-trained deep learning models by comparing their results, as shown in Table 8. CNN model was trained for 15 epochs.

When training is extended to 15 epochs, the custom CNN's test accuracy improves to 97.23% Table 8, with an F1-score of 97.5%. Per-class analysis

Table 8. Results of proposed CNN architecture for 15 epochs trained on the AMSD dataset.

Results using Proposed CNN Architecture				
Movie	Precision	Recall	F1-score	Support
Danish Exploration	0.87	0.95	0.91	758
I am not a Thief	1.00	1.00	1.00	717
Me	1.00	0.98	0.99	541
My Daughters My Life	0.97	0.90	0.93	819
My Sweetheart is very Naughty	0.99	0.99	0.99	698
Shaaban Below Zero	1.00	0.99	1.00	816
Streets from Fire	0.94	0.96	0.95	557
With My Love	0.98	0.95	0.96	670
Women in the City	0.99	1.00	0.99	619
AlZheimer	1.00	1.00	1.00	868
Test Loss		**0.1250**		
Test Accuracy		**0.9723**		

reveals near-perfect performance on indoor scenes (e.g., Alzheimer: F1 = 1.00; Shaaban Below Zero: F1 = 1.00), while outdoor scenes in the Danish Exploration film F1 = 0.91 and My Daughters, My Life F1 = 0.93 show low results because of the consistency of visual patterns. This improvement validates that the specialized CNN takes advantage of longer training to learn complex, high-level features associated with Arabic cinema, such as traditional dressings, design decorations, and regional lighting conditions. High F1-scores across all classes indicate balanced generalization, without marked bias toward top films.

Evaluation metrics and class-wise performance

To provide an accurate and transparent evaluation of the proposed CNN performance in classifying ten film classes, precision, recall, and F1-score for every class-film were computed, as shown in Table 8. Table 7 shows precision, recall, and F1-score for every class-film for all used pretrained models. The overall results showed high performance for all class-films concerning films with higher Indoor scenes, by achieving F1-score = 1.0, while films with higher Outdoor scenes have less performance (but still strong) because of the high visual similarity among Outdoor scenes. A normalized confusion matrix was used as in Fig. 4 to show that most classification errors occur due to Outdoor scenes that are visually similar (e.g., Danish Exploration and Streets from Fire), while indoor scenes show minimal confusion. This proved that the proposed CNN model was capable of learning and distinguishing culturally Distinctive Cinematic Cues.

Data splitting strategy and leakage prevention

AMSD Dataset was divided into three subsets; Training set 50% including 14131 frames, Validation set 25% including 7104 frames and Test set 25% including 7063. To avoid data leakage, frames for the same scene segment were included in a single split, such that the same frame, shot, or overlapping image occurred in more than one subset.

Frames were extracted after every 50 seconds to reduce redundancy and to ensure scene variety. Also, this guaranteed that the test set includes genuinely unseen cinematic content, which allowed reliable evaluation of the model's generalization ability.

Ablation study

To carefully evaluate the effect of each architectural and training component in the proposed CNN, we used an ablation study by systematically removing key elements. Thus isolating the effect of batch normalization, dropout regularization, and domain-specific data augmentation, which are intended to represent visual patterns that exist in Arabic cinema (e.g., color grading, framing conventions). Table 9 summarizes the obtained results, demonstrating that each key element gives a non-redundant performance, with the full developed model achieving the highest accuracy and F1-score.

Batch Normalization alone improves accuracy by +2.25% (92.10% → 94.35%), ensuring training stability across various cinematic lighting conditions. Standard augmentation provides higher performance of +2.92% (92.10% → 95.02%). A further +0.85% improvement over standard augmentation (95.02% → 95.87%) is obtained from adding domain-specific elements such as film-inspired color fluctuation and framing-sensitive crops, proving that modeling Arabic cinematic conventions adds important value. The full model performance 97.23% outperforms all variants, confirming the combined effect between architectural choices and domain-aware data strategies.

Table 9. Effect of each component on the proposed CNN architecture performance (trained for 15 epochs on AMSD).

Model Variant	BatchNorm	Dropout (0.5)	Domain-Specific Augmentation	Test Accuracy (%)	F1-Score (%)	Notes
Baseline CNN	no	no	no	92.10	91.85	3 Conv layers, ReLU, MaxPooling only
+ Batch Normalization	yes	no	no	94.35	94.12	Improved convergence and stability
+ Dropout	no	yes	no	93.60	93.40	Reduced overfitting on indoor scenes
+ Standard Augmentation	no	no	yes	95.02	94.88	Random horizontal flip, rotation (±10°), zoom (±15%)
+ Domain-Specific Augmentation	no	no	yes	95.87	95.70	Standard augmentation + cinematic-style color jittering (brightness ±20%, contrast ±15%, saturation ±10%) and aspect-preserving random crops reflecting common framing in Arabic films
Full Proposed Model	yes	yes	yes	97.23	97.50	All components combined

As shown in Table 9, the contribution of architectural and training choices is quantitatively analyzed, supporting the claims highlighted in the abstract.

Pre-trained model selection and integration

Different computer vision tasks such as image classification, object detection, and scene recognition can be achieved by several famous deep learning models (e.g., VGG-16, ResNet-50, Inception-V3, MobileNet, MobileNet-V2, and DenseNet variants (121, 169, 201)). A summary of pre-trained models is shown :

- **VGG-16:**[40] It has 16 layers divided into 13 convolutional layers and 3 fully connected layers. Although it is simple, it performs well in image categorization tasks.
- **ResNet-50 (Residual Network):**[41] Composed of 50 layers and consists of residual blocks in its deep neural network design. These blocks aim to solve the vanishing gradient problem. This network is used for feature extraction and image classification.
- **Inception-V3 (GoogLeNet-V3):**[42] Is an architecture that combines several simultaneous convolutional filters with inception modules. It is intended to achieve high precision while being computationally efficient.
- **MobileNet:**[43] is designed for embedded and mobile vision applications. It is lightweight and efficient enough for real-time inference support on edge devices, as it leverages depthwise separable convolutions- a factorized variant of standard convolutions to achieve better efficiency by lowering both the count of parameters and the cost of computations.
- **MobileNetV2:**[44] is an enhancement to MobileNet; it incorporates inverted residual blocks and further optimizes depth-wise separable convolutions. Better performance and efficiency are offered.
- **DenseNet121, DenseNet169, and DenseNet201 (Densely Connected Convolutional Networks):**[45] These CNN architectures ensure that the cumulative knowledge of all earlier layers enhances each layer, so the representational power is enhanced and vanishing gradients are mitigated. These networks have demonstrated great performance across a range of tasks, particularly when working with **limited datasets**.

These models are frequently employed as a foundation for transfer learning in computer vision applications, such as scene classification, because they have already been pre-trained on large-scale image datasets like ImageNet. The models were trained for three epochs on the AMSD dataset, enabling them to apply their extensive data-driven expertise to the particular scene categorization job at hand to discover which had the best scene classification accuracy results and also to compare with the suggested CNN model results. This results in more accurate and economical models, particularly for **tasks** involving **specific datasets**, such as scene classification in Arabic films. Table 7 above shows test accuracy and other evaluation metrics for the used pre-trained models.

As can be noticed, MobileNet has the highest accuracy, followed by DenseNet201. Thus, the next step is to study the influence of transfer learning on the proposed CNN model's accuracy.

Transfer learning with MobileNetV2

Instead of learning from the target dataset, transfer learning adapts pre-trained models by adjusting their weights to suit a domain-specific dataset that is new and often smaller. It can save time, data, and computational resources while often leading to better performance on specific tasks by building on the knowledge gained from pre-trained models.[46] First, transfer learning was applied using MobileNet because it showed the highest test accuracy 98% after 3 epochs as in Table 7. After 15 epochs, its test accuracy reached 66.41%, **which was considered unsatisfactory; therefore, another model was selected:** MobileNetV2, because MobileNetV2 is an evolution of the original MobileNet architecture, designed to be more efficient and accurate. Second, transfer learning was applied using MobileNetV2 for 15 epochs, and the obtained results were highly encouraging. A comparison of the results is shown in Table 10.

Where test accuracy for MobileNet is 98.95% after 3 epochs, then it decreased to 66.41% after 15 epochs, as shown in Table 10. This indicates catastrophic forgetting because of excessive fine-tuning. MobileNetV2 maintains high performance (95.85%) because of inverted residual blocks, which help in keeping low-level features. This means that not all pre-trained models can benefit equally from extended fine-tuning, especially on small, culturally specific datasets.

Model fine-tuning and learning rate scheduling

The effect of learning rate scheduling on transfer learning using MobileNetV2 was studied. Learning rate scheduling was implemented to improve convergence. During training, the learning rate was adjusted by a factor of 10 after every 5 epochs. The model was trained for 15 epochs, and its final test accuracy

Table 10. A comparison of MobileNet and MobileNetV2 after 15 epochs trained on the AMSD dataset.

	Transfer Learning using MobileNet.				Transfer Learning using MobileNetV2.			
Movie	Precision	Recall	F1-score	support	Precision	Recall	F1-score	support
Danish Exploration	0.64	0.75	0.69	758	0.96	0.92	0.94	758
I am not a Thief	0.50	0.61	0.55	717	0.94	0.95	0.94	717
Me	0.46	0.11	0.18	541	0.97	0.92	0.94	541
My Daughters My Life	0.57	0.78	0.66	819	0.95	0.96	0.96	819
My Sweetheart is very Naughty	0.82	0.82	0.82	698	0.96	1.00	0.98	698
Shaaban Below Zero	0.96	0.93	0.95	816	1.00	1.00	1.00	816
Streets from Fire	0.47	0.76	0.58	557	0.92	0.95	0.94	557
With My Love	0.68	0.28	0.39	670	0.97	0.93	0.95	670
Women in the City	0.97	0.44	0.61	619	0.92	0.94	0.93	619
AlZheimer	0.70	0.90	0.78	868	0.98	1.00	0.99	868
Test Loss		**1.0172**				**0.1437**		
Test Accuracy		**0.6641**				**0.9585**		

Table 11. Contextual comparison of the proposed CNN with state-of-the-art scene-categorizing models.

Reference	Model	Dataset	Accuracy (%)	Key Contribution
Bolei et al. (2016)[47]	ResNet50	Places365	93.1	Categorizing scenes with transfer learning
Mingexing & Quoc (2019)[48]	EfficientNet-B0	MIT-Indoor	94.6	Lightweight network for indoor scene understanding
Uckan et al. (2025)[49]	MobileNetV2	MIT-Indoor	95.0	Hybrid visual–textual features for indoor scenes
Alexey et al. (2021)[50]	Vision Transformer (ViT)	ImageNet (scene-related subsets)	96.4	Global context modeling via self-attention
Proposed Method (This Work)	**Custom CNN (with BatchNorm, Dropout, Domain Augmentation)**	**AMSD (10 Arabic Films)**	**97.23**	Culturally adapted CNN for Arabic cinema

reached 95.90%, which was 0.05% higher than the test accuracy reached by the same model when applying it without learning rate scheduling.

Comparison with state-of-the-art methods

To demonstrate the effectiveness of the proposed CNN model, a comparison was made with several current best-performing architectures reported in the literature. Table 11 summarizes the main characteristics and quantitative results. Table 11 shows a comparison among the proposed architecture with the state-of-the-art deep learning architectures stated in scientific literature. It was important to note that performing numerical comparison for performance metrics through these studies remained limited, because stated models were evaluated using different benchmark datasets such as ImageNet-based, MIT-Indoor and Places365, which were characterized by different scene definitions, data distributions and evaluation protocols from the proposed AMSD dataset. Therefore, this comparison introduced contextual insight about architectural trends rather than strict performance benchmarking, and the reported results should not be interpreted as absolute performance superiority. The increase in performance comes from using domain-specific augmentation and balanced normalization layers that improve robustness against changes in lighting and texture associated with Arabic cinematic scenes. Also, the proposed model differs from previous works that used only transfer learning from Western datasets. Our model was trained on a culture-based dataset (AMSD) to ensure higher contextual relevance.

Final models evaluation

Based on test accuracies shown in Tables 7 to 9, among the pre-trained models evaluated on the AMSD dataset, MobileNet still ranks first in classification accuracy. While the proposed CNN architecture achieved second place with 97.23% test accuracy, showing its efficiency in the classification of Arab scene movies.

To further assess the generalization of the proposed CNN architecture, the dataset partitioning strategy ensured that testing frames were extracted from scenes not used during training or validation, thereby evaluating the model on unseen visual content. Although all ten Arabic films were used within the dataset, each subset was strictly non-overlapping to prevent data leakage across training and testing

Table 12. Evaluation of the proposed model's performance on the test set (15 epochs).

Metric	Formula	Description	Obtained Value
Accuracy	(TP + TN) / (TN + TP + FP + FN)	Overall correctness of classification	97.23 %
Precision	TP / (FP + TP)	Positive predictions are correct	97.6 %
Recall	TP / (FN + TP)	Completeness of detection across all ground-truth scene labels.	97.4 %
F1-score	2 × (Precision × Recall) / (Precision + Recall)	Precision and recall are balanced	97.5 %
Support	—	Number of test samples	7063

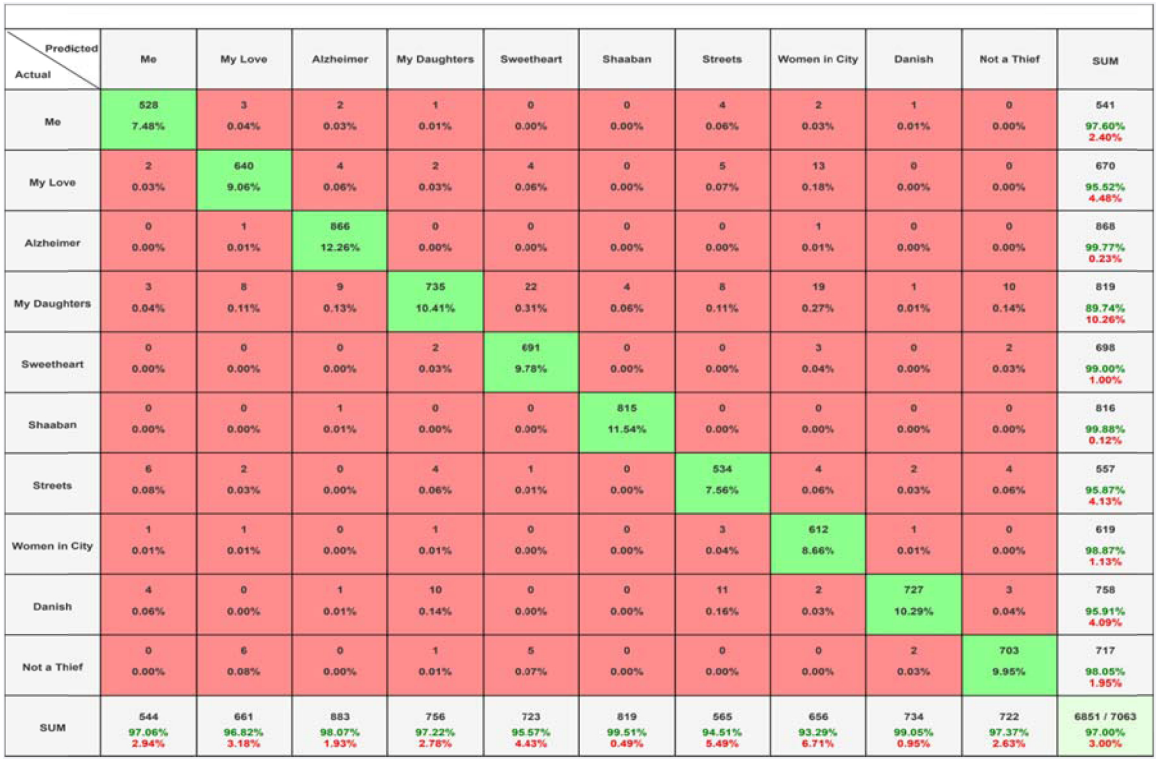

Fig. 4. Normalized confusion matrix of the proposed CNN on the AMSD test set.

phases. The model's consistent performance across diverse cinematic contexts—encompassing city, rural, and historical areas—shows its strength in unseen visual conditions associated with the Arabic film domain. In the future, this research will explore this evaluation by training the proposed model on a subset of Arabic films and testing it on completely new films from different Arab regions (e.g., North African or Gulf cinema) to further verify cross-domain generalization and cultural diversity robustness.

In addition to accuracy and precision, the performance of the proposed CNN was further evaluated using recall, F1-score, and a detailed confusion matrix analysis to better characterize how the model performs across different scene types, as in Table 12.

All performance metrics were computed based on the test set predictions to provide a reliable evaluation of the proposed CNN model. The results indicate that the model achieved a mean recall of 97.4%, an F1-score of 97.5%, and an overall test accuracy of 97.23%. The confusion matrix shown in Fig. 4 reveals that the majority of misclassifications occur between specific pairs of visually similar outdoor scenes from different films, rather than across all scene categories. This suggests that the model struggles with fine-grained discrimination when high- level semantic cues are absent. The confusion matrix is shown in Fig. 4 for the full CNN model. Misclassifications are concentrated between Danish Exploration and Streets from Fire—both featuring generic urban or rural landscapes lacking distinctive cultural markers (e.g., traditional clothing, architectural elements).

To further analyze the model's learning behavior and verify its stability during training, Fig. 5 illustrates the training and validation accuracy and loss curves for the proposed CNN model. Both curves exhibit a consistent convergence pattern and a negligible generalization gap between training and val-

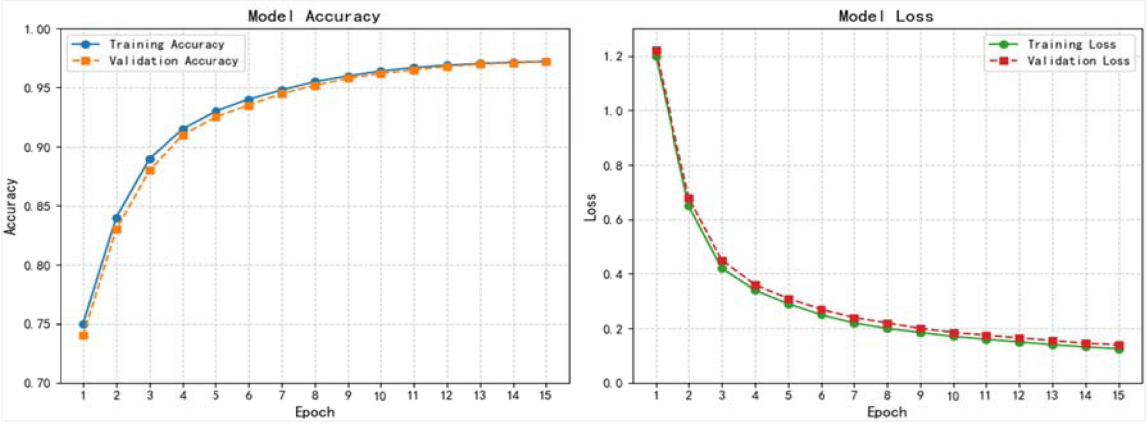

Fig. 5. Training with validation accuracy and loss curves for the developed CNN architecture.

Fig. 6. Examples of miss classified outdoor scenes. (a) Urban landscape from film "Me". (b) Rural landscape from the film "My Sweetheart is very Naughty".

idation sets, confirming that the model achieved effective generalization without significant overfitting.

Qualitative analysis of CNN model predictions, supported by Grad-CAM Visualization, showed that the success of classifications depends on culturally distinctive features such as traditional clothes (thob and hijab), Egyptian urban textures, interior furniture design, and regional architectural motifs such as mashrabya and stone facades. Also, the Grad-CAM heat map ensured that the model focuses on cultural elements that have semantic meanings such as clothes, window latticework, and regional decoration instead of focusing on general background textures. As an example, in scenes from "Shaaban Below Zero" and "Streets from Fire", the model benefits from distinctive urban textures and character clothes, which reflect Egyptian cinematic styles. At the same time, misclassifications occur because of visually generic Outdoor scenes that lack cultural symbols such as empty streets, open fields, or unmarked rural roads. Thus, culturally aware visual models were capable of learning region-specific semiotics, and they therefore assess the development of inclusive AI systems that respect non-Western visual narrative.

This challenge is demonstrated in Fig. 6, where two misclassified outdoor scenes are presented. Both scenes show common urban areas or countryside landscapes lacking distinguishing cultural, architectural, or contextual elements (e.g., notable landmarks, specific costumes, or identifiable actors) that could consistently tie them to corresponding film. The first image Fig. 6 shows a cityscape viewed from a bridge, while the second image Fig. 6 shows a rural field with distant buildings. Despite both images being taken from different Arabic films, their visual similarity (common color patterns, lighting conditions, and compositional structures) leads the model to confuse them. This highlights one of the important limitations of frame-level classification for cinematic content (the absence of temporal context or domain-specific semantic features makes the model depend heavily on low-level patterns, which is not enough for accurate scene attribution in visually homogeneous environments).

Fig. 6 illustrates two such misclassified scenes: an empty bridge and a rural field, which share similar color palettes and compositional structures despite originating from different films.

Cultural contextualization

Correct classifications depend on culturally significant features. The model effectively exploits Egyptian city features and the clothes of the characters in the scenes from Shaaban Below Zero, while in Women in the City, it emphasizes distinctive characteristics of modern cities such as the style of modern clothes and internal furniture shapes. This demonstrates that domain-specific visual representations enhance classification accuracy, and their absence in the generic outdoor scenes, lacking such cultural features, remains a key challenge. Although the proposed CNN was trained on the AMSD dataset, its ability to learn culturally distinctive visual cues did not depend on the data alone. It focuses on moderate network depth, localized receptive fields, and regularized convolutional blocks, which collectively prefer mid-level semantic representations such as traditional clothes, architectural styles, spatial arrangements, and lighting aesthetics over highly abstract global features. Compared to highly optimized lightweight architectures such as MobileNet, which gives priority to parameter efficiency through depthwise separable convolutions and aggressive feature compression, the proposed CNN keeps spatial detail and contextual information that enable it to stably attend to culturally salient regions.

The proposed CNN and pre-trained models get rid of overfitting by using several preventive measures during training. First, separate training, validation, and testing subsets as 50%, 25%, and 25%, to ensure that model evaluation was always executed on new data. Second, a dropout rate of 0.5 and batch normalization were consistently implemented after each convolutional layer to avoid memorizing training patterns by the network. Third, data augmentation techniques were used to enhance dataset diversity and model efficiency.

The strong correspondence between training, validation, and test accuracies ranging from 94.7% to 97.2% and the high and consistent F1-scores of 97.5% covering all classes verify that the model achieved strong generalization rather than overfitting. In addition, the use of a confusion matrix indicates that misclassifications were limited to outdoor scenes that are visually similar, not because of over-training. These stable findings ensure that the applied regularization and augmentation strategies effectively reduce overfitting during the experiments.

Limitations

Despite the strong performance mentioned above, this study has several limitations. First, AMSD is considered to be small and limited. While it is carefully collected and organized, it contains frames from only ten films; this dataset size and film-source diversity may limit generalization across regions (e.g., North Africa, Gulf, Maghreb variations). Future work should expand AMSD by including more films. Future studies will focus on external validation using films from Arabic underrepresented regions such as the Gulf region, the Levant region, and North Africa to measure the ability of the CNN model's interregional generalization. Also, future studies will use leave-one -film-out to evaluate the CNN model's efficiency on completely unseen cinematic data. Second, the proposed CNN processed data per frame. A single frame can lack temporal and motion cues that are useful for distinguishing visually alike outdoor scenes. Reducing the confusion that occurred in outdoor scenes observed in the confusion matrix can be performed by incorporating temporal modeling (e.g., frame-stacking, CNN–RNN hybrids, temporal transformers) or shot-level aggregation. Finally, computational constraints limited exhaustive fine-tuning experiments for some deep backbones. So in the future, using large-scale fine-tuning and domain-adaptation experiments would strengthen evaluation of model ranking across architectures. Recognizing these limitations defines the focus of the current work and encourages specific improvements to increase generalization, multi-modal understanding, and time-based reasoning in Arabic film scene classification.

Conclusion and recommendations

In this study, a customized CNN architecture and its associated AMSD dataset were proposed and used for Arabic movie indoor and outdoor scene classification. It discusses the difficulties in outdoor scene classification. From a theoretical point of view, this research contributes to examining the role of transfer learning techniques and convolutional neural networks in culturally dependent visual environments such as Arabic cinema. It offers a deep understanding of the fine-tuning of deep models originally trained on general datasets such as ImageNet, with a comparison to specialized film-based data, to show the domain gap between natural images and cinematic compositions. Also, this study shows that the suggested CNN architectures can achieve high accuracy and efficiency in recognizing Arabic film scenes, even with limited data. The proposed CNN is used for automating classification and categorization of Arabic visual content; therefore, it is valuable in film archiving, digital libraries, intelligent video management systems, media indexing, scene retrieval, metadata generation, and in

supporting the accessibility and protection of Arabic cinematic heritage.

The main research contributions are: (1) Building AMSD, which is the first dedicated dataset used for Arabic movie scene classification; (2) Designing and training a customized CNN model that is as effective as pre-trained architectures; and (3) Comparing several transfer learning models to provide a reference for future research directions in Arabic visual content analysis.

Despite the strong results obtained, there are several limitations, such as: (1) The AMSD dataset includes only ten films, which may restrict generalization over broader Arabic regions and cinematic styles. (2) The classification operates at the frame level without using temporal continuity or motion cues, which can be used to increase correct classification in visually similar outdoor scenes. (3) This work focuses on visual features, without using other data types such as audio or text that could enrich contextual understanding. Thus, future research directions include:

(1) Enhance CNN model generalization by expanding the proposed AMSD dataset to cover a wider range of Arabic film genres, lighting conditions, and production styles;
(2) Detect scene transitions in motion by using time-aware modeling methods such as 3D CNNs, CNN–RNN hybrids, or video transformers; and
(3) Use a variety of data types like visual, auditory, and textual cues for a deeper semantic understanding of Arabic cinematic content.

In conclusion, this research presents a significant contribution to both the theoretical and practical understanding of deep learning applications in Arabic visual media. It establishes the foundation for developing AI systems specialized for cultural contents and capable of analyzing, organizing, and preserving Arabic film content with improved accuracy.

Authors' declaration

- Conflicts of Interest: None.
- We hereby confirm that all the Figures and Tables in the manuscript are ours. Furthermore, any Figures and images that are not ours have been included with the necessary permission for republication, which is attached to the manuscript.
- No animal studies are present in the manuscript.
- No human studies are present in the manuscript.
- Ethical Clearance: The project was approved by the local ethical committee at University of Baghdad.

Data availability

The Arab Movie Scene Dataset (AMSD) used in this study is not publicly available due to copyright and licensing restrictions related to the source films. However, the dataset and code are available from the corresponding author **upon reasonable request** for academic and non-commercial research purposes.

References

1. Aymen Z, Sherif N, Mohamed M, Hazem M, Salama D. DeepFakeDG: A Deep Learning Approach for Deep Fake Detection and Generation. J Comput Commun. 2023;2(2):31–37. https://doi.org/10.21608/jocc.2023.307056.
2. Khan R, Sohail M, Usman I, Moid M, Raza M, Azfar M, *et al.* A Comparative study of deep learning techniques for DeepFake video detection. ICT Express. 2024;10(6):1226–39. https://doi.org/10.1016/j.icte.2024.09.018.
3. Lamouadene H, EL Kassaoui M, EL Yadari M, El Kenz A, Benyoussef A, El Moutaouakil A, *et al.* Detection of COVID-19, lung opacity, and viral pneumonia via X-ray using machine learning and deep learning. Comput Biol Med. 2025;191:110131. https://doi.org/10.1016/j.compbiomed.2025.110131.
4. Pan H, Zhang J, Lin C, Feng R, Zhan Y. Framework for psoriasis/molluscum detection in skin images using ResNetV2 variants. J Radiat Res Appl Sci. 2024;17:101052. https://doi.org/10.1016/j.jrras.2024.101052.
5. Loukam M. Comparison of Naïve Bayes with graph based methods for keyphrase extraction in modern standard Arabic language. Int J Speech Technol. 2023;26:141–150. https://doi.org/10.1007/s10772-022-10009-6.
6. Magnana L, Rivano H, Chiabaut N. A deep reinforcement learning solution to help reduce the cost in waiting time of securing a traffic light for cyclists. J Cycl Micromob Res. 2024;2:100046. https://doi.org/10.1016/j.jcmr.2024.100046.
7. Seo E, Yang J, Lee J, So G. Predictive Modelling of Student Dropout Risk: Practical Insights from a South Korean Distance University. Heliyon. 2024;10(11):1–17. https://doi.org/10.1016/j.heliyon.2024.e30960.
8. Niyogisubizo J, Liao L, Nziyumva E, Murwanashyaka E, Claver P. Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization. Comput Educ Artif Intell. 2022;3:1–12. https://doi.org/10.1016/j.caeai.2022.100066.
9. Mouchantaf N, Chamoun M. Predicting Student Dropout with Minimal Information. Iraqi J. Sci. 2023;64(10):5265–79. https://doi.org/10.24996/ijs.2023.64.10.33.
10. Li S, Deng W. Deep facial expression recognition: A survey. IEEE Trans. Affect. Comput. 2022;13(3):1195–215. https://doi.org/10.1109/TAFFC.2020.2981446.
11. Rehman A, Brahim S, Alamgir M, Khan A. On the Use of Deep Learning for Video Classification. Appl. Sci. 2023;13(3):2007–31. https://doi.org/10.3390/app13032007.
12. Bose D, Hebbar R, Somandepalli K, Zhang H, Cui Y, Cole-McLaughlin K, *et al.* MovieCLIP: Visual Scene Recognition in Movies. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV). USA. 2023:2082–91. https://doi.org/10.1109/WACV56688.2023.00212.
13. Rao A, Xu L, Xiong Y, Xu G, Huang Q, Zhou B, *et al.* A Local-to-Global Approach to Multi-modal Movie Scene Segmentation. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). USA.

2020:10143–52. https://doi.org/10.1109/CVPR42600.2020.01016.
14. Liu C, Shmilovici A, Last M. Towards story-based classification of movie scenes. PLoS ONE. 2020;15(2):1–22. https://doi.org/10.1371/journal.pone.0228579.
15. Haidarh M, Mu C, Liu Y, He X. Exploring traditional, deep learning and hybrid methods for hyperspectral image classification: A review. J Inf. Intell. 2025:1–40. [Online]. Available: https://doi.org/10.1016/j.jiixd.2025.04.002.
16. Alzubaidi L, Zhang J, J Amjad, Al-Dujaili A, Duan Y, *et al.* Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. J Big Data. 2021;8(53):1–74. https://doi.org/10.1186/s40537-021-00444-8.
17. Zhang T, El Ali A, Wang C, Hanjalic A, Cesar P. CorrNet: Fine-Grained Emotion Recognition for Video Watching Using Wearable Physiological Sensors. Sensors. 2021;21(1):52–77. https://doi.org/10.3390/s21010052.
18. Singla P, Mehra R. Scene Recognition using Significant Feature Detection Technique. Int. J. Innov. Technol. Explor. Eng. (IJITEE). 2020;9(3):1705–11. https://doi.org/10.35940/ijitee.C8653.019320.
19. Kumar B, Gupta H, Pravin S, Vyaset OP. Classification of Indoor–Outdoor Scene Using Deep Learning Techniques. in Mach. Learn., Image Process., Netw. Secur. Data Sci. 2023:517–535. https://doi.org/10.1007/978-981-19-5868-7_38.
20. Hussain N, Emaduldeen M, Ezaldeen A. Learning Evolution: a Survey. Iraqi J Sci. 2021;62(12):4978–87. https://doi.org/10.24996/ijs.2021.62.12.34.
21. Jamal S, MA A. Facial Expression Recognition Based on Deep Learning: An Overview. Iraqi J Sci. 2023;64(3):1401–1425. https://doi.org/10.24996/ijs.2023.64.3.32.
22. Xie L, Lee F, Liu L, Kotani K, Chen Q. Scene recognition: a comprehensive survey. Pattern Recognit. 2020;102(C):107205. https://doi.org/10.1016/j.patcog.2020.107205.
23. Alina M, Andreea G, Estefania T. Deep Learning for Scene Recognition from Visual Data: A Survey. In E. A. de la Cal, J. R. Villar Flecha, & E. Corchado (Eds.), Hybrid Artificial Intelligent Systems, HAIS Lecture Notes in Computer Science, Springer. 2020;12344:763–73. https://doi.org/10.1007/978-3-030-61705-9_64.
24. Mustafa M. Understanding of Machine Learning with Deep Learning: Architectures, Workflow, Applications and Future Directions. Computers (MDPI). 2023;12(5):91–117. https://doi.org/10.3390/computers12050091.
25. Masood S, Ahsan U, Munawwar F, Raza D, Ahmed M. Scene Recognition from Image Using Convolutional Neural Network. Procedia Computer Science. Int Conf Comput Intell Data Sci (ICCIDS 2019). 2020;167(1):1005–12. https://doi.org/10.1016/j.procs.2020.03.400.
26. Ananda S, Vandana B. Transfer Learning for Multi-Crop Leaf Disease Image Classification using Convolutional Neural Network VGG. Artif Intell Agric. 2022;6:23–33. https://doi.org/10.1016/j.aiia.2021.12.002.
27. Zhou B, Lapedriza A, Khosla A, Oliva A. Torralba A. Places: A 10 million image database for scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2017;40(6):1452–1464. https://doi.org/10.1109/TPAMI.2017.2723009.
28. Hahnloser R, Sarpeshkar R, Mahowald M, Rondy J, Sebastian H. Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. Nature. 2000;405(6789):947–951. https://doi.org/10.1038/35016072.
29. Nhlapho W, Atemkeng M, Brima Y, Ndogmo J. Bridging the Gap: Exploring Interpretability in Deep Learning Models for Brain Tumor Detection and Diagnosis from MRI Images. Information. 2024;15(4):182. https://doi.org/10.3390/info15040182.
30. Chen Y. Research on Image Recognition of Convolutional Neural Network under Different Computer Data Set Capacities. J Phys Conf Ser. 2021;1744(4):042091. https://doi.org/10.1088/1742-6596/1744/4/042091.
31. Raja R, Roomi M, Dharmalakshmi D. Classification of Scenes into Indoor/Outdoor. Res J Appl Sci Eng Technol. 2014;8(21):2172–78. http://dx.doi.org/10.19026/rjaset.8.1216.
32. Sami S, Kadhim A, Alexander S. A Comparative Study of Anemia Classification Algorithms for International and Newly CBC Datasets. Int J Online Biomed Eng (iJOE). 2023;19(6):141–57. https://doi.org/10.3991/ijoe.v19i06.38157.
33. Svendsen B, Kadry S. A Dataset for Recognition of Norwegian Sign Language. Int J Math Stat Comput Sci. 2024;2:1–10. https://doi.org/10.59543/ijmscs.v2i.8049.
34. Muayed A, Redha N. Predicting Movie Production Years through Facial Recognition of Actors with Machine Learning. Baghdad Sci. J. 2024;21(12):4146–62. https://doi.org/10.21123/bsj.2024.8996.
35. Singh T, Kumar D. Video benchmarks of human action datasets: A review. Artif Intell Rev. 2019;52:1107–1154. https://doi.org/10.1007/s10462-018-9651-1.
36. Yue Z, Zhang Q, Hu A, Zhang L, Wang Z, Jin Q. Movie101: A New Movie Understanding Benchmark. Proc 61st Annu Meet Assoc Comput Linguist. Canada. 2023;1:4669–84. https://doi.org/10.18653/v1/2023.acl-long.257.
37. Khan Y, Attique M, Noman K, Baili J, Bacanin N, Hong M, *et al.* InBRwSANet: Self-attention based parallel inverted residual bottleneck architecture for human action recognition in smart cities. PLoS One. 2025;20(5):1–22. https://doi.org/10.1371/journal.pone.0322555.
38. Sedky A, Hegazy I, Elarif T, Abdelwahab M S. Indexed Dataset from Youtube for A Content-Based Video Search Engine. Int J Intell Comput Inf Sci (IJICIS). 2021;21(1):196–215. https://doi.org/10.21608/ijicis.2021.68816.1072.
39. Vujovic ŽĐ. Classification model evaluation metrics. Int. J. Adv. Comput. Sci. Appl. 2021;12(6):1–9. https://doi.org/10.14569/IJACSA.2021.0120670.
40. Pardede J, Sitohang B, Akbar S, Leylia M. Implementation of Transfer Learning Using VGG16 on Fruit Ripeness Detection. Int. J. Intell. Syst. Appl. (IJISA). 2021;13(2):52–61. https://doi.org/10.5815/ijisa.2021.02.04.
41. Zhang L, Bian Y, Jiang P, Zhang F. A Transfer Residual Neural Network Based on ResNet-50 for Detection of Steel Surface Defects. Appl. Sci. 2023;13(9):5260. https://doi.org/10.3390/app13095260.
42. Madanan M, Sayed BT. Designing a Deep Learning Hybrid Using CNN and Inception V3 Transfer Learning to Detect the Aggression Level of Deep Obsessive Compulsive Disorder in Children. Int J Biol Biomed Eng. 2022;16:207–20. https://doi.org/10.46300/91011.2022.16.27.
43. Widyayuningtias S. Performance of Deep Learning Inception Model and MobileNet Model on Gender Prediction Through Eye Image. J Penelit Tek Inform. 2022;7(4):2593–601. https://doi.org/10.33395/sinkron.v7i4.11887.
44. Katsigiannis S, Seyedzadeh S, Agapiou A, Ramazan N. Deep learning for Crack Detection on Masonry Façades using Limited Data and Transfer Learning. J Build Eng. 2023;76:107105. https://doi.org/10.1016/j.jobe.2023.107105.

45. Okafor E, Oyedeji M, Alfarraj M. Deep reinforcement learning with lightweight vision model for sequential robotic object sorting. J King Saud Univ- Comput Inf Sci. 2024;36(1):101896. https://doi.org/10.1016/j.jksuci.2023.101896.
46. Ghabri H, Alqahtani MS, Ben S, Al-Rasheed A, Abbas M, Ali H, *et al.* Transfer Learning for Accurate Fetal Organ Classification from Ultrasound Images: A Potential Tool for Maternal Healthcare Providers. Sci Rep. 2023;13(1):17904. https://doi.org/10.1038/s41598-023-44689-0.
47. Zhou B, Khosla A, Lapedriza A, Torralba A, Oliva A. Places: An Image Database for Deep Scene Understanding. *arXiv preprint arXiv:1610.02055*. 2016. https://doi.org/10.48550/arXiv.1610.02055.
48. Tan M, V Le Q. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks.*Proceedings of the 36th International Conference on Machine Learning*, in *Proceedings of Machine Learning Research*. 2019;97:6105–6114.
49. Uckan T, Aslan C, Hark C. A Comprehensive Hybrid Approach for Indoor Scene Recognition Combining CNNs and Text-Based Features. *Sensors*. 2025;25(17):5350. https://doi.org/10.3390/s25175350.
50. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T *et al.* An Image Is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. Proceedings of the International Conference on Learning Representations (ICLR). 2021. https://doi.org/10.48550/arXiv.2010.11929.

التصنيف الفعال لمشاهد الافلام العربية باستخدام الشبكة العصبية الالتفافية المخصصة والمدربة مسبقا

اسراء مؤيد عبد الله

قسم علوم الحاسوب، كلية العلوم للبنات، جامعة بغداد، بغداد، العراق.

الخلاصة

نظرًا لتنوعها البصري الغني، وأساليبها السينمائية الفريدة، وقلة مجموعات البيانات المعلّمة من قبل خبراء المخصصة للسينما العربية، لا يزال تصنيف المشاهد في الأفلام العربية بدقة مهمة صعبة. سيناقش البحث الحالي بنية شبكة عصبية تلافيفية (CNN) جديدة تم إنشاؤها خصيصًا للتعرف على مشاهد الأفلام العربية. الهدف الرئيسي من الدراسة هو معرفة مدى قدرة الـCNN المقترحة على التعرف وتصنيف المشاهد العربية من الأفلام مقارنة بالمعماريات المدربة مسبقًا الموجودة. تم إجراء تحليل مقارن استنادًا إلى نتائج ستة نماذج تعلم انتقالية (Transfer Learning) لقد قمنا بتطوير مجموعة بيانات جديدة تحتوي على 28,298 صورة من عشرة أفلام عربية مختلفة. تم استخدام هذه المجموعة لتدريب النماذج لمدة ثلاث حقبات (Epochs) في البداية. أثناء التدريب، استخدمنا نفس معلمات الضبط المسبق، وتقسيم البيانات، وخط معالجة البيانات لكل نموذج لضمان الاتساق المنهجي. بعد ذلك، تم إجراء مقارنة حول أداء الاختبار والتنبؤ لكل نموذج. خلال 15 حقبة، حققت شبكة الـCNN المُنشأة دقة اختبار بلغت 97.23٪ ودقة تنبؤ بلغت 97٪، مما أظهر دقة عالية مع جهد تدريب أقل نسبيًا. في هذا البحث، كان نموذج MobileNet هو أكثر النماذج المدربة مسبقًا نجاحًا، حيث حقق دقة اختبار بنسبة 98٪ ودقة تنبؤ بنسبة 99٪. لذلك، جاءت شبكة الـCNN التي تم إنشاؤها في المرتبة الثانية بشكل عام بدقة اختبار بلغت 97.23٪، مما يوضح فائدتها في تصنيف المشاهد السينمائية العربية. تشير هذه الورقة إلى فوائد اختيار تصميم نموذج فعال وطريقة تدريب مناسبة لتصنيف المشاهد. كما تُظهر الإمكانات الإضافية للتعلم العميق في التحليل البصري المبني على الثقافة، من خلال إثبات أن شبكة CNN مصممة خصيصًا يمكنها تحقيق دقة عالية على بيانات الأفلام العربية. كما أظهرت الشبكة التي تم إنشاؤها قدرتها على الحفاظ على التراث السينمائي العربي وتحليله.

الكلمات المفتاحية: تصنيف مشاهد الأفلام العربية، الشبكة العصبية التلافيفية (CNN)، التعلم العميق، ضبط المعلمات الفائقة، التعلم الانتقالي، تحليل المشاهد، مجموعة بيانات الأفلام العربية، التراث الثقافي.