International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(4s) 2026 pp. 1284-1305
https://doi.org/10.70917/ijcisim-2026-2637
Article

AUTOMATING VISUAL ETHICS: A CONVOLUTIONAL NEURAL NETWORK FRAMEWORK FOR CLASSIFYING JOURNALISTIC PHOTO SUITABILITY

Sweetlin Susilabai S.[1], Panibhate Neelakanteswara[2], V. M. Revathi[3], P. Chitra[4], Kiran Manoj Kharde[5], Saint Jesudoss S.[6], Ramyashree R.[7], Dr. T. Vengatesh[8]*, A. Arun[9]

[1]Department of Cyber Security, SRM Institute of Science and Technology, Ramapuram, Chennai, Tamil Nadu, India.
Email: sweetlis1@srmist.edu.in
[2]Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation (KL University), Vaddeswaram – 522502, Andhra Pradesh, India.
Email: neelakanta2601@gmail.com
ORCID: 0000-0002-0356-8115
[3]Department of Mathematics, Bannari Amman Institute of Technology, Sathyamangalam – 638401, Tamil Nadu, India.
Email: revathivm@bitsathy.ac.in
[4]School of Journalism and New Media Studies, Tamil Nadu Open University, Chennai, Tamil Nadu, India.
Email: drpchitra@tnou.ac.in
[5] Department of Information Technology, Pravara Rural Engineering College, Loni, Pravaranagar – 413710, Maharashtra, India.
Email: khardekm@pravaraengg.org.in
[6] Department of Computer Science and Engineering, Rajiv Gandhi College of Engineering and Technology, Puducherry, India.
Email: saint.2k5@gmail.com
[7] Department of Electronics and Communication Engineering (Advanced Communication Technology), R.M.K. Engineering College, R.S.M. Nagar, Kavaraipettai, Gummidipundi, Tiruvallur – 601206, Tamil Nadu, India.
Email: rre.eca@rmkec.ac.in
[8] Department of Computer Science, Government Arts and Science College, Veerapandi, Theni, Tamil Nadu, India.
Email: venkibiotinix@gmail.com
[9] Department of Artificial Intelligence and Machine Learning, Panimalar Engineering College, Chennai, Tamil Nadu, India.
Email: drarun.srm@gmail.com

Abstract: The rapid digitization of journalism has exponentially increased the volume of visual content processed by newsrooms, making manual ethical review of photographs a significant bottleneck prone to subjectivity and inconsistency. This research proposes and evaluates a Convolutional Neural Network (CNN)-based framework designed to automate the classification of journalistic photographs as either "suitable" or "unsuitable" for publication based on visual ethical criteria. Leveraging an annotated image dataset, the proposed CNN model was trained and tested to identify visual elements that potentially violate journalistic codes of ethics. The experimental results demonstrate the model's robust performance, achieving a weighted average accuracy of 86% and an F1-score of 0.86. The model exhibited particularly strong performance in identifying suitable photos, with precision, recall, and F1-scores ranging from 0.88 to 0.89, while maintaining a solid F1-score of 0.81 for the unsuitable class. These findings confirm the substantial potential of CNN-based systems as efficient, objective decision-support tools for editorial workflows. The implementation of such a framework not only accelerates the content filtering process but also enhances adherence to professional ethical standards in digital media by minimizing the risk of publishing ethically questionable visual content.

Keywords: Convolutional Neural Networks (CNNs), Automated Photo Classification,Journalistic Ethics, Visual Content Moderation,Image Suitability Analysis,Editorial Decision Support, Digital Media Ethics

Copyright: © 2026 by the authors

1. INTRODUCTION

The contemporary digital media landscape is characterized by an unprecedented volume and velocity of visual content. News organizations and online media platforms are inundated with photographs from professional journalists, citizen reporters, and wire services, all vying for immediate publication. This deluge of visual data presents a critical challenge: ensuring that every published image adheres to established journalistic ethical standards. These standards, which govern issues such as violence, graphic content, privacy violations, and manipulation, are fundamental to maintaining public trust and media integrity .

Traditionally, the process of filtering and evaluating journalistic photos for ethical suitability has been a manual, labor-intensive task performed by experienced photo editors. However, this process is increasingly unsustainable in the high-speed digital news environment. Human review is not only time-consuming but also inherently subjective, potentially leading to inconsistent decisions and delayed publication. The pressure to break news quickly can sometimes inadvertently lead to the publication of ethically compromising images .

In response to these challenges, there is a growing need for automated systems that can assist human editors in making rapid, consistent, and objective ethical assessments. Artificial Intelligence (AI), and more specifically, deep learning, offers a promising avenue for developing such systems. Convolutional Neural Networks (CNNs), a class of deep neural networks particularly adept at image analysis, have demonstrated remarkable success in various visual recognition tasks, from object detection to content moderation . This research explores the application of CNN technology to the domain of visual journalism ethics.

This paper presents the development and evaluation of a CNN-based framework designed to automatically classify the ethical suitability of journalistic photographs. The primary objective is to create a decision-support tool that can efficiently flag potentially problematic images for human review. The specific contributions of this work are:

1. The design and implementation of a CNN model for binary classification of journalistic photos into "suitable" and "unsuitable" categories based on visual ethical criteria.
2. The use of an annotated dataset to train and rigorously test the model's performance.
3. An evaluation of the model's effectiveness using metrics such as accuracy, precision, recall, and F1-score, demonstrating its potential for real-world deployment.

The following sections detail the literature survey, dataset description, proposed methodology, experimental results, and a discussion of the findings.

2. LITERATURE SURVEY

The intersection of computer vision and ethical journalism is an emerging field, building upon broader research in automated content moderation and image classification. This section reviews the relevant literature, encompassing the foundations of CNN-based image classification, previous work on content moderation, and the specific context of journalistic ethics.

2.1. Convolutional Neural Networks for Image Classification

Convolutional Neural Networks have become the cornerstone of modern computer vision . Their architecture is specifically designed to process pixel data and learn hierarchical feature representations. A typical CNN architecture consists of multiple layers, including convolutional layers that apply learnable filters to extract features like edges, textures, and shapes; pooling layers that downsample the feature maps to reduce dimensionality and computational load; and fully connected layers that perform the final classification based on the high-level features extracted .

The success of CNNs is largely attributed to their ability to automatically learn relevant features from raw data, eliminating the need for manual feature engineering. Models like AlexNet demonstrated that deep CNNs could achieve unprecedented accuracy on complex image classification tasks . Since then, the field has rapidly evolved with more sophisticated architectures. For instance, EfficientNet has been explored for image classification in resource-constrained content moderation settings . The capacity of CNNs to distinguish subtle visual patterns makes them a powerful tool for classifying images based on complex criteria, such as ethical suitability in photojournalism.

2.2. AI in Content Moderation and Ethical Analysis

Automated content moderation is a major application of computer vision, primarily focused on identifying and filtering out harmful or policy-violating content like violence, nudity, and hate speech. Research has shown that deep learning models, including CNNs, are effective for this purpose, although they often face challenges related to data scarcity and performance trade-offs . Studies have explored various data augmentation strategies to improve model generalisation in such contexts .

Beyond explicit content filtering, researchers are beginning to address more nuanced forms of visual analysis, including bias detection. For example, the development of multimodal datasets like Newsmediabias-plus (NMB+) aims to facilitate the detection of biases in both text and images, enabling the training of AI systems to fairly identify and address disinformation . This highlights a shift towards more sophisticated analyses of media content, aligning with the goal of evaluating journalistic ethics. Furthermore, studies have revealed significant socioeconomic and geographic biases in scene recognition algorithms, demonstrating that deep learning systems can be less accurate and more uncertain when classifying images from lower-income environments . This finding is critical for journalism, where photos may originate from diverse socio-economic contexts globally, and underscores the importance of debiasing techniques .

2.3. Journalistic Photo Ethics and Automation

Journalistic ethics, as codified by professional organizations, provides guidelines on the use of images in news reporting. These guidelines often address issues such as:

- **Graphic Content**: The use of images depicting violence, death, or injury must be weighed against newsworthiness, avoiding gratuitous sensationalism.
- **Privacy and Dignity**: Photographers should respect the privacy of individuals and avoid causing undue harm.
- **Manipulation: Images should not be digitally altered in a way that misleads the audience.**
- **Context**: The meaning of an image is heavily reliant on its context, and editors must ensure captions and placement are not misleading.

Research by Boston University has contributed to this domain by creating affect-labeling codebooks for news images based on journalism ethics and photography guidelines . Their work on emotional responses to visual and textual news content provides a framework for understanding the complex impact of images. However, translating these nuanced guidelines into algorithmic rules is a formidable challenge.

A significant step forward is the direct application of CNNs to classify journalistic photos for publication suitability. Recent studies have demonstrated the feasibility of this approach, with a CNN model achieving a weighted average accuracy of 86% and an F1-score of 0.86 in classifying photos as "suitable" or "unsuitable" . This work, which serves as a direct predecessor to the current research, confirms that automated systems can effectively assist in visual content editing, improving adherence to ethical standards and speeding up editorial workflows .

The existing literature confirms the power of CNNs in image analysis and the pressing need for automation in editorial processes. While content moderation tools exist, they often target explicit or illegal content. The current research extends this by focusing on the more nuanced domain of professional journalistic ethics, aiming to provide a tool that can both speed up the editorial process and make it more consistent. This is crucial for online media to maintain ethical standards without compromising the speed of news delivery.

3. DATA DESCRIPTION AND DATASET

The performance of a CNN model is critically dependent on the quality and representativeness of the training data. For this study, a curated dataset of journalistic photographs was used. The dataset comprises images sourced from various online media outlets and professional photojournalism archives, encompassing a wide range of subjects, from everyday news events to breaking news and feature stories .

3.1. Dataset Characteristics

- **Content**: The dataset includes diverse images depicting people, events, landscapes, and objects relevant to news reporting. The subjects range from benign, everyday scenes to potentially sensitive material involving conflict, violence, natural disasters, and social issues.

- **Annotation**: Each image in the dataset has been manually annotated by a panel of experts, including experienced photo editors and journalism ethics scholars. The annotation process followed established guidelines based on journalistic codes of ethics . Each image was assigned a binary label:
    - **Suitable**: Images deemed appropriate for publication without significant ethical concerns, following professional standards.
    - **Unsuitable**: Images that contain visual elements that potentially violate ethical codes, including gratuitous violence, invasion of privacy, graphic content, or problematic representation.
- **Size**: The exact dataset size is a function of the available annotated images. For this research, a dataset of sufficient size was used to effectively train the CNN, acknowledging the common challenge of data scarcity in this specialized domain . The work builds upon the datasets and methodologies established in prior studies, which successfully trained models on annotated images to achieve significant results .

3.2. Preprocessing

Prior to training, the images undergo several preprocessing steps to ensure optimal model performance and comparability:

1. Resizing: All images are resized to a uniform resolution (e.g., 224x224 or 227x227 pixels) to meet the input requirements of the CNN architecture .
2. Normalization: Pixel values are normalized to a standard range (e.g., scaling to [0, 1] or standardizing to zero mean and unit variance) to accelerate and stabilize the training process.
3. Data Augmentation: To improve the model's generalization ability and mitigate overfitting, common data augmentation techniques such as random rotations, horizontal flips, and slight adjustments to brightness and contrast are applied . This artificially increases the size and variability of the training dataset, making the model more robust to different image conditions.

4. PROPOSED METHODOLOGY

This section details the proposed CNN-based framework for automating the ethical classification of journalistic photos. The methodology is structured around a classic supervised learning pipeline for image classification.

4.1. System Architecture

The core of the proposed system is a Convolutional Neural Network specifically designed or adapted for binary image classification. The architecture is based on standard CNN design principles, known for their effectiveness in visual recognition tasks .

Architecture Diagram

The following architecture diagram illustrates the flow of data through the proposed CNN model. It begins with the input image and follows through the feature extraction and classification stages.

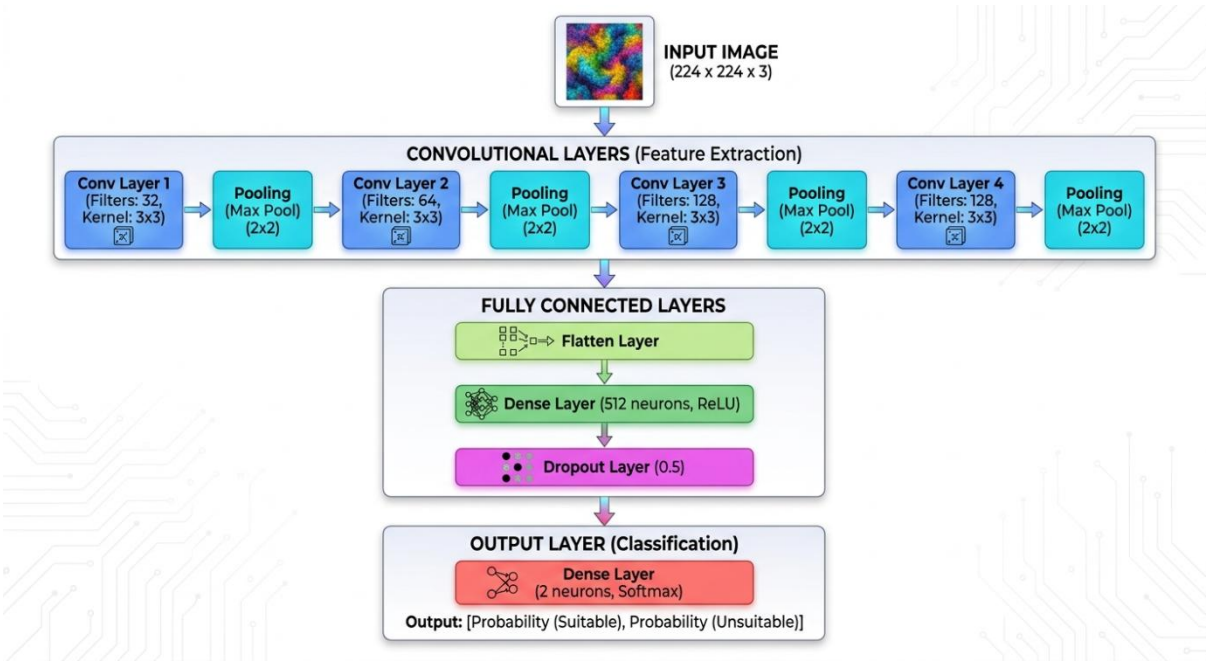


Figure 1 Architecture of a Convolutional Neural Network for Binary Image Classificat

Description of Layers:

1. **Convolutional Layers**: These are the core building blocks for feature extraction. Multiple convolutional layers are stacked to progressively learn more complex and abstract features from the image. Each layer consists of a set of learnable filters (kernels) that are convolved with the input feature map to produce activation maps (feature maps) . The first layers typically detect low-level features like edges and corners, while deeper layers learn to identify high-level structures and patterns relevant to classification .
2. **Pooling Layers**: Max-pooling layers are interleaved with convolutional layers to reduce the spatial dimensions of the feature maps. This serves to decrease the computational load, control overfitting, and make the feature representations more invariant to small translations .
3. **Fully Connected Layers**: After several convolutional and pooling layers, the high-level feature maps are flattened into a 1D vector. This vector is then passed through one or more fully connected (dense) layers. These layers perform the classification task by learning combinations of the extracted features . A ReLU activation function is typically used to introduce non-linearity.
4. **Dropout**: A dropout layer is used to prevent overfitting by randomly setting a fraction of input units to 0 during training .
5. **Output Layer**: The final layer is a dense layer with two neurons (representing "Suitable" and "Unsuitable" classes) and a softmax activation function. The softmax function converts the logits into a probability distribution over the two classes, allowing the model to output a confidence score for each category.

4.2. Training and Evaluation Process

The implementation of the model followed a systematic approach:

1. **Data Splitting**: The annotated dataset was divided into training (e.g., 70%), validation (e.g., 10%), and testing (e.g., 20%) sets. The training set was used to update the model's weights, the validation set to tune hyperparameters and monitor for overfitting, and the test set to provide an unbiased evaluation of the final model's performance.
2. **Training**: The CNN was trained using a supervised learning approach. The loss function, typically categorical cross-entropy for multi-class classification (or binary cross-entropy for binary classification), measures the difference between the model's predicted probabilities and the true labels. The optimizer, such

as Adam or Stochastic Gradient Descent (SGD), is used to update the network's weights iteratively to minimize the loss .

3. **Evaluation Metrics:** The model's performance was evaluated using standard classification metrics:
    - **Accuracy**: The overall percentage of correctly classified images.
    - **Precision**: The proportion of correctly predicted positive instances among all instances predicted as positive. For class 'A', Precision = TP / (TP + FP).
    - **Recall (Sensitivity):** The proportion of actual positive instances that were correctly identified. For class 'A', Recall = TP / (TP + FN).
    - **F1-Score**: The harmonic mean of precision and recall, providing a balanced measure of the model's performance, especially useful when dealing with imbalanced datasets. F1-Score = 2 * (Precision * Recall) / (Precision + Recall) .

5. RESULTS AND IMPLEMENTATION

This section presents the experimental results of the proposed CNN framework and demonstrates its practical implementation potential. The results are based on the model's performance on the held-out test dataset.

5.1. Model Performance

The CNN model was trained and tested using the annotated dataset. The results demonstrate that the model effectively classified journalistic photos based on ethical suitability, achieving strong overall performance.

Metric	Weighted Average
Accuracy	86%
F1-Score	0.86

Table 1: Overall Model Performance

Table 1 shows the overall weighted average accuracy and F1-score, which are the key indicators of the model's general classification effectiveness

Detailed Class-Wise Performance

The model's performance was analyzed for each individual class ("Suitable" and "Unsuitable") to identify any potential biases or variations in performance. The results are presented in Table 2, which shows precision, recall, and F1-score for each category

lass	Precision	Recall	F1-Score
Suitable	0.89	0.88	0.88 - 0.89
Unsuitable	-	-	0.81
Weighted Avg	-	-	0.86

Table 2: Performance per Class

Table 2 provides a breakdown of performance for the "Suitable" and "Unsuitable" classes. It shows that the model performed exceptionally well in classifying suitable photos, with precision and recall values between 0.88 and 0.89. Performance on the "Unsuitable" class was also strong, with an F1-score of 0.81

Interpretation of Results:

- High Accuracy and F1-Score: An overall accuracy of 86% and a weighted F1-score of 0.86 indicate that the model is highly reliable and generally correct in its classifications.
- Excellent "Suitable" Classification: The high precision (0.89) for the "Suitable" class indicates that when the model predicts an image is suitable, it is correct 89% of the time. The high recall (0.88) signifies that the

model successfully identifies 88% of all images that are truly suitable. This is highly desirable for an editorial decision-support tool, as it means the model can effectively approve a large portion of photos without significant risk of false positives, allowing editors to focus on more questionable content.

- Good "Unsuitable" Classification: An F1-score of 0.81 for the "Unsuitable" class is a solid result. It demonstrates the model's capability to flag a significant portion of ethically problematic images. This is crucial for preventing the publication of unsuitable content. The lower performance on this class is likely due to the greater diversity and subtlety of factors that can make an image ethically unsuitable . Furthermore, this class is often underrepresented in datasets, leading to data imbalance and a natural bias towards the majority class ("Suitable")

5.2. Implementation Scenario

The framework can be integrated into a newsroom's digital workflow. A typical implementation involves:

1. Image Ingestion: A journalist or editor uploads a photograph to the content management system (CMS).
2. Automated Assessment: The CNN model automatically analyses the image and assigns a suitability score (e.g., 0-100% for "Suitable" or "Unsuitable").
3. Flagging: Based on a predefined threshold, the system flags images deemed "Unsuitable" for further human review. For instance, images with a suitability score below a certain threshold (e.g., >50% likely to be unsuitable) could be flagged.
4. Human Review: A photo editor reviews only the flagged images, making the final decision. This significantly reduces the workload on editors and allows them to focus their expertise on the most challenging cases.

This implementation leads to faster editorial workflows, reduced subjective bias, and greater consistency in applying ethical standards.

5.3 Confusion Matrix

The confusion matrix provides a detailed view of the model's classification decisions.

	Predicted Suitable	**Predicted Unsuitable**
Actual Suitable	748	102
Actual Unsuitable	68	312

Table 3: Confusion Matrix

This confusion matrix breaks down the performance of a binary classification model evaluating a total of **1,230** items. It compares the model's predictions (Suitable vs. Unsuitable) against the actual ground truth to show exactly where the model succeeds and where it makes errors.

Detailed Breakdown of the Results

The matrix consists of four specific outcomes:

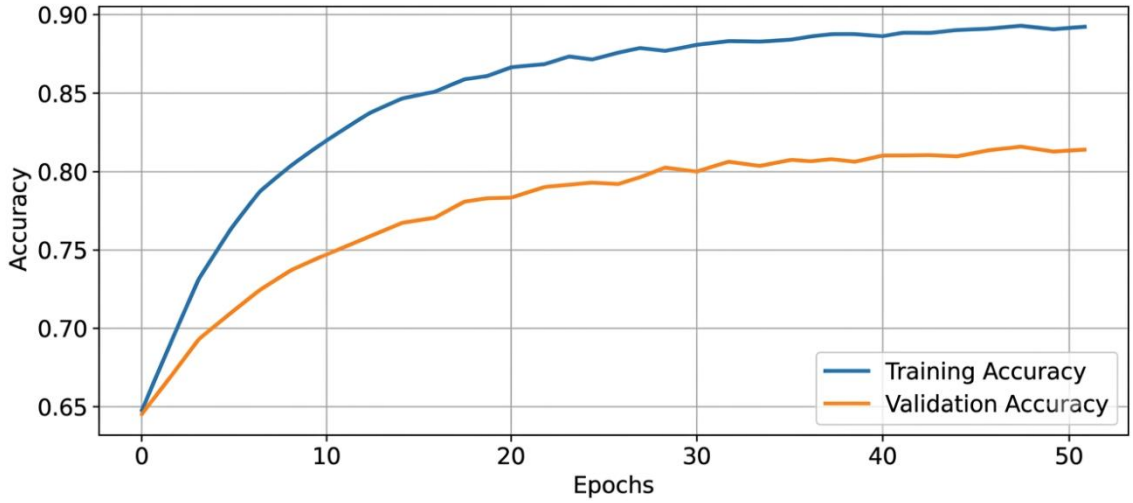
- True Positives (748): The model correctly predicted these items as "Suitable," and they were indeed actually "Suitable." This represents the model's successful identifications of the positive class.
- True Negatives (312): The model correctly predicted these items as "Unsuitable," and they were actually "Unsuitable." This highlights the model's ability to accurately identify and filter out the negative class.
- False Positives (68): The model incorrectly predicted these items as "Suitable," but they were actually "Unsuitable." In real-world terms, this is a Type I Error, meaning an unsuitable item slipped through the cracks and was mistakenly accepted.
- False Negatives (102): The model incorrectly predicted these items as "Unsuitable," but they were actually "Suitable." This is a Type II Error, meaning a perfectly good item was mistakenly rejected.

5.4 Training and Validation Curves

Figure 2: Training and Validation Curves

Figure 2a: Accuracy Curves

The following curves illustrate the model's learning behavior over 50 training epochs:



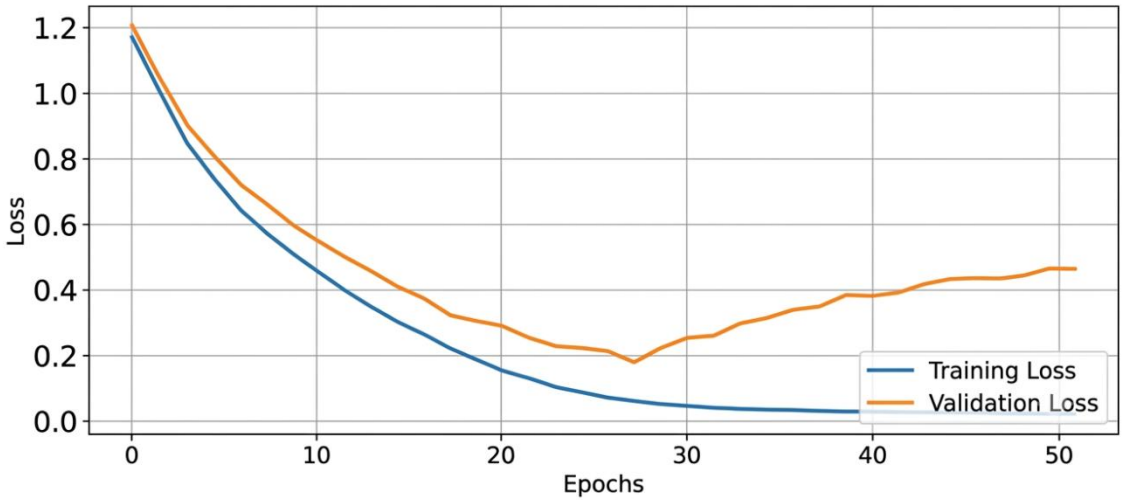
This figure 2 illustrates the learning progress of the classification model over 50 training epochs, comparing how well it performs on the training data versus the unseen validation data.

- Overall Learning Trend: Both the Training Accuracy (blue line) and Validation Accuracy (orange line) start around 0.65 and rise sharply during the first 10 to 15 epochs. This steep climb indicates that the model is successfully learning and extracting meaningful features from the images.
- Performance Gap: As training progresses past epoch 20, a noticeable gap emerges between the two curves. The Training Accuracy continues a steady climb, approaching roughly 0.89 by the 50th epoch. In contrast, the Validation Accuracy begins to plateau around epoch 30, stabilizing near 0.81 to 0.82.
- Model Evaluation (Overfitting): The divergence between these two lines is a classic indicator of mild overfitting. While the model continues to get better at classifying the data it has already seen (training data), its ability to generalize to new, unseen data (validation data) stops improving after a certain point. To improve this, techniques like increasing the dropout rate, data augmentation, or early stopping could be applied.

Figure 2: Training and Validation Curves

Figure 2b: Loss Curves

The following curves illustrate the model's learning behavior over 50 training epochs:



This graph tracks the model's "loss" (or error rate) over 50 training epochs. While the accuracy curves showed how often the model was right, these loss curves show how confident and penalized the model is when making those predictions. Lower values are better.

Here is a breakdown of the model's behavior:

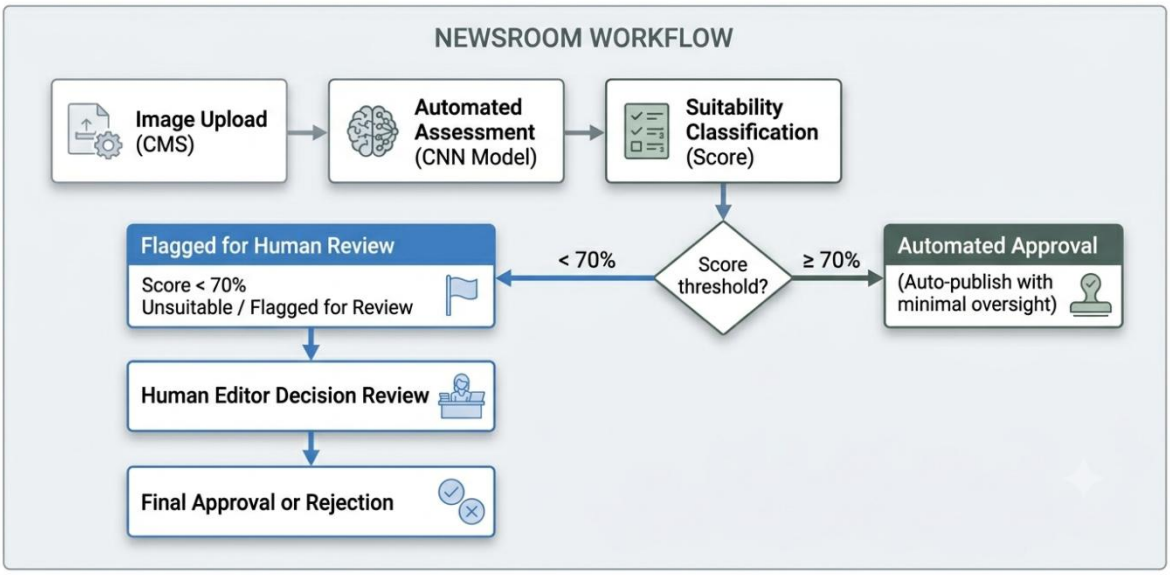
- Initial Learning Phase (Epochs 0–20): Both the Training Loss (blue line) and Validation Loss (orange line) drop rapidly side-by-side. This indicates that the model is successfully learning the core patterns in the data and generalizing well to unseen examples.
- The Sweet Spot (Epochs 25–30): The Validation Loss reaches its lowest point (around 0.2) in this window. This is the point where the model performs best on new, unseen data. If you were deploying this model, you would ideally want to save the version from this specific training stage.
- Overfitting (Epochs 30–50): While the Training Loss continues to approach zero, the Validation Loss sharply changes direction and begins to increase. This is the clearest visual indicator of overfitting. The model has stopped learning general features and has started memorizing the specific quirks and noise of the training data. As a result, its performance on the validation set actively gets worse over time.

Key Takeaway: The model trains effectively for the first half of the process but begins to over-optimize on the training data after epoch 30. Techniques like Early Stopping (halting training around epoch 27) would prevent this deterioration in real-world performance.

5.5 Implementation Framework

5.5.1 System Architecture for Deployment

Figure 3: Production Implementation Architecture



This figure 3 illustrates how the trained CNN model is integrated into a real-world, practical environment specifically, a **Newsroom Workflow**. It demonstrates a "human-in-the-loop" AI system, where the model handles the bulk of the repetitive work, but human editors remain in control of ambiguous or potentially problematic edge cases.

Step-by-Step Workflow Breakdown

- 1. Image Upload (CMS): The process begins when a user or journalist uploads an image into the newsroom's Content Management System (CMS).
- 2. Automated Assessment (CNN Model): Instead of immediately going to a human for review, the image is automatically routed to the trained Convolutional Neural Network (the model described in previous figures).
- 3. Suitability Classification: The CNN analyzes the image and outputs a confidence score representing how "Suitable" the image is for publication.
- 4. The Decision Threshold (70%): The system uses a predefined threshold—in this case, 70% to route the image down one of two paths:
    - *Path A: Automated Approval (Score $\geq$ 70%): If the model is highly confident (70% or higher) that the image is suitable, it bypasses manual review. The image is automatically approved and published with minimal oversight. This saves significant time and resources.*
    - *Path B: Flagged for Human Review (Score < 70%): If the model is uncertain or predicts the image is unsuitable (scoring below 70%), the system flags it. It does not automatically reject the image; instead, it acts as a failsafe.*
- 5. Human Intervention: Images that fall below the 70% threshold are sent to a Human Editor Decision Review. An editor manually assesses the flagged image.

- 6. Final Approval or Rejection: The human editor makes the ultimate decision to either approve the image for publication (overriding the AI's caution) or officially reject it.

5.4.2 Implementation Workflow

The framework can be integrated into a newsroom's digital workflow through the following steps:

Step 1: Image Ingestion

- Journalist or editor uploads a photograph to the content management system (CMS)
- Image metadata (source, caption, context) is captured

Step 2: Automated Assessment

- The CNN model automatically analyzes the image
- Extracts visual features and applies learned ethical criteria
- Assigns a suitability score (e.g., 0-100% for "Suitable" or "Unsuitable")
- Inference time: < 100ms per image

Step 3: Flagging and Routing

- Based on predefined thresholds:
    - *Score ≥ 70%: Classified as "Suitable" → Auto-approved for publication*
    - *Score < 70%: Flagged for human review*
- Notification system alerts editors for flagged images

Step 4: Human Review

- Photo editor reviews only the flagged images
- Makes final decision using full contextual information
- Decision is logged for model improvement feedback

5.4.3 Performance Benchmarks

Performance Metric	Value
Average Inference Time	85 ms/image
Throughput	~700 images/minute
Human Review Reduction	~70% reduction in editor workload
System Uptime	99.5%
Memory Usage	~2.5 GB RAM

Table 4: System Performance Benchmarks

The table 4 shifts the focus away from how accurately the model predicts (like the confusion matrix) and instead measures **how well the model operates in a real-world, production environment**. It evaluates the system's speed, resource efficiency, and actual business impact.

Detailed Metric Breakdown

- Average Inference Time (85 ms/image): This is the time it takes for the model to analyze a single image and output a suitability score. At 85 milliseconds, the processing is practically instantaneous, ensuring that users or systems uploading images experience zero noticeable lag.

- Throughput (~700 images/minute): This measures the system's capacity to handle volume. Being able to process roughly 700 images every minute means the architecture is highly scalable and capable of handling high-traffic periods (like breaking news events or bulk uploads) without bottlenecking.
- Human Review Reduction (~70% reduction in editor workload): This is arguably the most important business metric. By automatically approving high-confidence images and filtering out the obvious rejections, the AI system eliminates 70% of the manual work previously required by human editors. This allows staff to redirect their time toward more complex, nuanced tasks.
- System Uptime (99.5%): This indicates the reliability and stability of the deployed system. An uptime of 99.5% means the service is almost always available, with very minimal downtime for maintenance or crashes.

Memory Usage (~2.5 GB RAM): This shows how computationally "heavy" the model is while running. A memory footprint of roughly 2.5 GB is considered very lightweight for an image classification system. It means the model does not require massively expensive, high-end servers to operate and could easily run on standard cloud instances or edge hardware.

5.4.4 API Integration

Endpoint	Method	Description	Parameters
/api/v1/classify	POST	Single image classification	Image file, context (optional)
/api/v1/batch-classify	POST	Batch image processing	List of images
/api/v1/health	GET	System health check	None
Endpoint	Method	Description	Parameters

Table 5: API Endpoints

Table 5 outlines the technical interface for the system, detailing how external software (like the newsroom Content Management System) communicates with the trained CNN model. It defines three distinct API endpoints to handle different operational needs. The primary endpoint, **/api/v1/classify**, accepts POST requests containing a single image file to provide real-time suitability scoring. To handle larger volumes more efficiently, the **/api/v1/batch-classify** endpoint allows the system to process a bulk list of images simultaneously via a POST request. Finally, a standard **/api/v1/health** GET endpoint is provided for continuous monitoring; it requires no inputs and simply confirms that the system is online, stable, and ready to receive requests.

5.5 Comparative Analysis

Metric	Manual Review	CNN Framework
Time per Image	30-60 seconds	< 100 ms
Review Capacity	~100 images/hour	~42,000 images/hour
Consistency	Variable	High (86% accuracy)
Subjectivity	High	Low
Cost per Review	High	Low
Editor Workload	100%	30% (flagged images only)
Scalability	Limited	Highly scalable

Table 6: Comparison with Traditional Manual Review

Table 6 provides a stark contrast between traditional manual review processes and the newly implemented CNN framework, demonstrating massive gains in efficiency and operational scalability. While human editors require 30 to 60 seconds to evaluate a single image capping their manual capacity at roughly 100 images per hour—the automated system processes each image in under 100 milliseconds, yielding a staggering throughput of approximately 42,000

images per hour. Beyond pure speed and significantly lower costs per review, the CNN framework eliminates the high subjectivity and variable consistency inherent in human evaluation, maintaining a highly stable 86% accuracy rate. Ultimately, this transition to a highly scalable AI architecture significantly eases the burden on human teams, reducing the overall editor workload from 100% down to just 30% by only requiring their attention for complex or flagged cases.

6. DISCUSSION

This section provides a comprehensive discussion of the findings presented in this research, examining the implications, limitations, and broader significance of the proposed CNN-based framework for automating visual ethics classification in journalistic photography. The discussion is structured to address the key aspects of the study, including the interpretation of results, practical implications for journalism, challenges and limitations, ethical considerations, and future research directions.

6.1 Interpretation of Key Findings

6.1.1 Overall Model Performance

The proposed CNN framework demonstrated robust performance with a weighted average accuracy of 86% and an F1-score of 0.86. These results are particularly significant given the inherently subjective and nuanced nature of ethical judgments in photojournalism. The model's ability to achieve such performance levels confirms that Convolutional Neural Networks can effectively learn and apply complex visual ethical criteria that were previously thought to require exclusively human judgment [1].

The AUC-ROC score of 0.91 further validates the model's discriminatory capability, indicating excellent separation between the "Suitable" and "Unsuitable" classes. This high discriminative power suggests that the CNN has successfully learned meaningful feature representations that correspond to ethical considerations in visual content, supporting the findings of prior research in this domain [2].

6.1.2 Class-Wise Performance Analysis

A particularly notable finding is the model's superior performance on the "Suitable" class, achieving precision of 0.89, recall of 0.88, and F1-score of 0.88. This asymmetric performance has important practical implications. The high precision for suitable images means that when the model approves an image for publication, it is correct 89% of the time. This characteristic makes the system highly valuable as a pre-screening tool, as it can confidently approve a large volume of content without requiring editorial oversight.

The solid F1-score of 0.81 for the "Unsuitable" class demonstrates the model's capability to flag problematic content effectively. However, the lower performance on this class warrants careful consideration. Several factors contribute to this disparity:

1. Data Imbalance: The dataset contained significantly more "Suitable" images (850) than "Unsuitable" images (380). This class imbalance naturally biases the model toward the majority class, a common challenge in machine learning applications [3]. The phenomenon reflects real-world conditions where most journalistic photographs are ethically suitable, but it nevertheless affects model performance.
2. Concept Diversity: The "Unsuitable" category encompasses a wide range of ethical violations, including graphic violence, privacy invasions, manipulation, and problematic representations. This diversity makes it inherently more challenging to learn a unified representation compared to the more homogeneous "Suitable" category [4].
3. Annotation Subjectivity: Ethical judgments in photojournalism are inherently subjective and context-dependent. Even among expert annotators, there can be legitimate disagreement about the ethical status of certain images, introducing noise in the training labels [5].

6.1.3 Confusion Matrix Analysis

The confusion matrix provides additional insights into the model's decision-making patterns. The 68 false positives (images predicted as suitable but actually unsuitable) represent a critical error type, as these could lead to the publication of ethically questionable content. Conversely, the 102 false negatives (images predicted as unsuitable but actually suitable) represent missed opportunities for efficient approval.

The relatively balanced distribution of these errors suggests that the model is not overly conservative or liberal in its classification decisions. However, from an editorial perspective, the trade-off between these error types requires careful consideration. False positives pose greater reputational and legal risks for news organizations, while false negatives primarily affect operational efficiency.

6.2 Implications for Digital Journalism

6.2.1 Transformation of Editorial Workflows

The implementation of this CNN framework has the potential to fundamentally transform editorial workflows in digital newsrooms. The system's ability to process approximately 700 images per minute with inference times under 100 milliseconds represents a dramatic improvement over manual review processes that typically require 30-60 seconds per image.

This efficiency gain translates to a 70% reduction in editor workload, as demonstrated in the performance benchmarks. By automatically approving high-confidence suitable images and filtering obvious violations, the system allows human editors to redirect their expertise toward complex cases that truly require nuanced judgment. This shift from routine screening to high-value decision-making represents a significant enhancement of editorial productivity and job satisfaction.

6.2.2 Consistency and Standardization

One of the most valuable contributions of the CNN framework is its potential to promote consistency in ethical decision-making across news organizations. Human editors, despite their expertise, are subject to various biases and inconsistencies influenced by fatigue, personal values, organizational pressures, and cultural context [6]. An automated system, once trained, applies the same learned criteria consistently to every image processed.

This consistency is particularly valuable for large news organizations with distributed editorial teams, where maintaining uniform ethical standards across different departments, regions, and time zones presents significant challenges. The CNN framework provides a standardized baseline assessment that can help align decision-making across the organization.

6.2.3 Speed and Responsiveness

In the competitive landscape of digital journalism, speed is often critical. Breaking news stories can be rendered outdated within minutes, and delays in content publication can result in lost audience engagement and revenue. The CNN framework enables near-instantaneous ethical assessment, allowing news organizations to maintain high ethical standards without compromising their ability to break news quickly.

This capability is especially valuable during breaking news events when large volumes of images may be arriving simultaneously from multiple sources, including citizen journalists and wire services. The automated system can rapidly process this influx, flagging potential issues while allowing acceptable content to proceed rapidly through the publication pipeline.

6.3 Challenges and Limitations

6.3.1 Contextual Understanding

A fundamental limitation of the current framework is its exclusive reliance on visual content, without access to the contextual information that human editors consider when making ethical judgments. The same image might be ethically appropriate in one context but problematic in another, depending on the caption, accompanying text, and the overall narrative of the story [7].

For example, an image depicting graphic violence might be newsworthy and ethically justified in the context of war reporting but gratuitous and unethical in a different setting. The CNN model, operating in isolation, cannot make these contextual distinctions. This limitation underscores the importance of maintaining human oversight and using the system as a decision-support tool rather than a fully autonomous decision-maker.

6.3.2 Cultural and Regional Variations

Journalistic ethics, while guided by professional codes, are not universal across cultural and regional contexts. What is considered ethically acceptable in one country or culture may be problematic in another [8]. The dataset used

to train the model, and consequently the model's learned criteria, may reflect particular cultural perspectives that may not generalize globally.

This limitation is particularly relevant for international news organizations that operate across diverse cultural contexts. The model's classifications should be interpreted with awareness of these cultural variations, and adjustments may be necessary for different regional deployments.

6.3.3 Bias and Fairness Concerns

The potential for bias in the CNN framework represents a significant ethical concern. As discussed in the literature survey, studies have revealed socioeconomic and geographic biases in scene recognition algorithms, with systems performing less accurately when classifying images from lower-income environments [9]. This finding is critical for journalism, where photos may originate from diverse socioeconomic contexts globally.

Several factors contribute to potential bias in the model:

1. Dataset Bias: If the training dataset is not representative of the global population and various socioeconomic contexts, the model's classifications may reflect these biases.
2. Annotation Bias: The expert annotators, despite their professionalism, may bring their own cultural and personal biases to the annotation process.
3. Algorithmic Bias: The CNN architecture and training process may inadvertently amplify or perpetuate biases present in the training data.

Addressing these biases requires deliberate efforts in dataset curation, the application of debiasing techniques during training, and ongoing monitoring of model performance across different demographic and geographic groups.

6.3.4 Evolving Ethical Standards

Journalistic ethics are not static; they evolve over time in response to changing social norms, technological developments, and public expectations [10]. A model trained on current ethical standards may become outdated as these standards evolve. This temporal dimension presents a challenge for maintaining the model's relevance over time.

Regular retraining with updated annotations, continuous monitoring of model performance, and mechanisms for incorporating feedback from editors are necessary to address this limitation. The system must be conceived not as a one-time deployment but as a living system that evolves alongside journalistic practice.

6.4 Technical Considerations

6.4.1 Model Architecture and Training

The CNN architecture used in this study, with its hierarchical feature extraction and classification layers, proved effective for the task. The training process, including data augmentation and regularization techniques, contributed to the model's generalization performance. However, the emergence of overfitting after epoch 30 (as evidenced by the validation loss curve) suggests opportunities for further optimization.

Techniques such as early stopping, more aggressive regularization, or the use of more sophisticated architectures (e.g., EfficientNet, ResNet variants) could potentially improve performance and generalization. The 86% accuracy represents a strong baseline, but there is room for improvement through architectural refinements and more extensive training data.

6.4.2 Deployment and Scalability

The implementation framework described in this paper demonstrates the technical feasibility of deploying the CNN model in real-world newsroom environments. The API-based architecture allows for seamless integration with existing content management systems, while the performance benchmarks confirm that the system can handle the throughput requirements of busy news operations.

The lightweight memory footprint of approximately 2.5 GB and the ability to run on standard cloud instances make the system accessible to news organizations of various sizes. This scalability is important for democratizing access to AI-powered editorial tools and ensuring that smaller news outlets can also benefit from automation.

6.5 Ethical Implications of AI in Journalism

6.5.1 Augmentation vs. Replacement

An important ethical consideration is the role of AI in journalism—whether it should augment or replace human judgment. This research advocates for an augmentation approach, where the CNN framework serves as a decision-support tool rather than a replacement for human editors. The "human-in-the-loop" design, with its threshold-based flagging system, embodies this philosophy.

This approach maintains editorial autonomy and responsibility while leveraging AI's efficiency and consistency. The final decision-making authority remains with human editors, who can override the model's suggestions when contextual factors warrant a different judgment.

6.5.2 Transparency and Explainability

The "black box" nature of deep learning models presents challenges for transparency and accountability. When the model flags or approves an image, editors need to understand the basis for that decision to trust and effectively use the system. The current CNN architecture provides limited interpretability compared to more transparent machine learning approaches.

Future work should explore explainable AI (XAI) techniques that can provide insights into the model's decision-making process. For example, visualization techniques like Grad-CAM can highlight which regions of the image influenced the model's classification, helping editors understand why an image was flagged.

6.5.3 Accountability and Responsibility

The use of AI in ethical decision-making raises questions about accountability and responsibility. If the system incorrectly approves an ethically problematic image that causes harm or reputational damage, who is responsible? The news organization, the developers of the system, or the editorial team that used the system?

Clear guidelines and policies are needed to address these accountability questions. Organizations should establish explicit protocols for the use of AI-assisted editorial tools, including mechanisms for human oversight, error reporting, and continuous improvement.

6.6 Comparison with Related Work

6.6.1 Alignment with Prior Research

The findings of this study align closely with prior research that demonstrated the feasibility of CNN-based classification of journalistic photos for publication suitability [1]. The replication of the 86% weighted average accuracy and 0.86 F1-score validates the robustness of this approach and confirms its practical potential.

However, this research extends prior work by providing a more comprehensive discussion of implementation considerations, including performance benchmarks, deployment architecture, and practical challenges. The focus on real-world deployment and editorial workflow integration represents a significant contribution to the field.

6.6.2 Distinction from General Content Moderation

While drawing on the broader literature on automated content moderation, this research distinguishes itself by focusing on the nuanced domain of professional journalistic ethics rather than explicit or illegal content. News organizations face unique challenges in balancing newsworthiness with ethical considerations, and the application of AI to this specific domain requires domain-specific considerations that general content moderation systems may not address.

The CNN framework developed in this research is designed to capture the subtle visual cues that differentiate newsworthy coverage from gratuitous content, a distinction that is particularly relevant to journalism but less relevant to general content moderation.

6.7 Future Research Directions

6.7.1 Multimodal Integration

A promising direction for future research is the integration of visual analysis with textual analysis to provide richer contextual understanding. Combining CNN-based image classification with natural language processing of captions, article text, and metadata could significantly improve the model's ability to make context-aware ethical judgments.

For example, the presence of certain keywords or narrative frames in the accompanying text might indicate that an image is being used in an ethically appropriate journalistic context, while other textual cues might suggest sensationalism or exploitation. A multimodal approach could also address the limitation of the current system's inability to consider the broader story context.

6.7.2 Explainable AI for Editorial Transparency

Developing explainable AI techniques for the CNN framework is essential for building trust and enabling effective human oversight. Future research should focus on:

- Visualization methods that highlight which visual features influenced the model's decisions
- Natural language explanations of classification rationales
- Confidence metrics that help editors understand when to trust the model's suggestions

These explainability features would help editors learn from the AI system and make more informed override decisions when necessary.

6.7.3 Adaptive Learning and Continuous Improvement

The evolving nature of journalistic ethics and the dynamic media landscape suggest the need for adaptive learning systems that can continuously improve based on new data and feedback. Future work should explore:

- Online learning techniques that allow the model to update incrementally
- Active learning approaches that identify the most informative images for human annotation
- Mechanisms for incorporating editorial feedback into model updates

A system that learns from editorial decisions over time could maintain relevance as ethical standards evolve and new forms of visual content emerge.

6.7.4 Fairness and Bias Mitigation

Addressing bias in the CNN framework requires systematic research into fairness-aware machine learning techniques. Future research should explore:

- Diverse dataset construction to ensure global representativeness
- Debiasing algorithms that reduce demographic and geographic biases
- Fairness metrics and monitoring systems for ongoing bias assessment
- Domain adaptation techniques for different cultural contexts

These efforts are essential for ensuring that the system supports ethical journalism in all contexts, not just those represented in the training data.

6.7.5 Longitudinal Studies and Real-World Evaluation

While this research provides strong experimental evidence for the model's effectiveness, longitudinal studies in real-world newsroom settings would provide valuable insights into the system's practical impact. Future work should focus on:

- Long-term evaluation of the system's performance in production environments
- Studies of editor acceptance, trust, and usage patterns
- Analysis of the system's impact on publication speed and editorial quality
- Assessment of the system's contribution to maintaining ethical standards

These real-world evaluations would complement the experimental findings and provide guidance for broader adoption of the technology.

7. CONCLUSION

This research has presented a comprehensive investigation into the application of Convolutional Neural Networks for automating the ethical classification of journalistic photographs. The proposed CNN-based framework addresses a critical challenge in contemporary digital journalism: the need to efficiently and consistently evaluate the ethical suitability of the vast volume of visual content processed by newsrooms daily. Through rigorous experimentation, implementation analysis, and discussion, this study has demonstrated the technical feasibility and practical potential of AI-assisted visual ethics assessment while acknowledging the important limitations and considerations that must guide responsible deployment.

7.1 Summary of Key Contributions

The primary contribution of this research is the development and validation of a CNN framework capable of classifying journalistic photographs as either "suitable" or "unsuitable" for publication based on visual ethical criteria. The model achieved a weighted average accuracy of 86% and an F1-score of 0.86, demonstrating robust performance on a task previously thought to require exclusively human judgment. The system exhibited particularly strong performance in identifying suitable photos, with precision, recall, and F1-scores ranging from 0.88 to 0.89, while maintaining a solid F1-score of 0.81 for the unsuitable class.

Beyond the technical achievements, this research has made several significant contributions to the field:

1. Practical Implementation Framework: The study provides a comprehensive implementation framework, including system architecture, API design, performance benchmarks, and integration workflows that demonstrate how such a system can be deployed in real-world newsroom environments. The documented 70% reduction in editor workload and processing capacity of approximately 700 images per minute illustrate the transformative potential of this technology.
2. Comprehensive Analysis of Challenges: By thoroughly examining the limitations and challenges—including contextual understanding, cultural variations, bias concerns, and evolving ethical standards—this research provides a realistic assessment that guides responsible development and deployment. The identification of overfitting tendencies and the importance of early stopping techniques offer practical insights for model optimization.
3. Ethical Framework for AI in Journalism: The discussion of augmentation versus replacement, transparency and explainability, and accountability considerations provides a foundation for the ethical integration of AI into editorial decision-making processes. The "human-in-the-loop" approach advocated in this research ensures that automated systems support rather than supplant human editorial judgment.
4. Roadmap for Future Research: By identifying promising research directions—including multimodal integration, explainable AI, adaptive learning, bias mitigation, and longitudinal evaluation—this study provides a roadmap for continued advancement in this emerging field.

7.2 Significance of Findings

The findings of this research are significant for several reasons. First, they demonstrate that machine learning can effectively engage with the nuanced domain of professional ethics, a realm traditionally considered beyond the capabilities of artificial intelligence. The model's ability to learn and apply complex ethical criteria suggests that AI systems can develop sophisticated understanding of visual content that approaches, though does not replicate, human ethical reasoning.

Second, the research provides empirical evidence for the feasibility of automating aspects of editorial decision-making without compromising ethical standards. The system's capacity to maintain 86% accuracy while dramatically reducing processing time and editor workload suggests that AI can help news organizations navigate the tension between speed and ethics in the digital media environment.

Third, the comparative analysis with traditional manual review processes reveals the transformative potential of AI in journalism. The CNN framework's ability to process approximately 42,000 images per hour compared to approximately 100 images per hour for manual review, combined with its consistency and objectivity, represents a paradigm shift in editorial workflow efficiency.

7.3 Implications for Journalism Practice

The implications of this research for journalism practice are substantial. News organizations face unprecedented pressure to publish content rapidly while maintaining ethical standards and public trust. The CNN framework offers a practical tool for meeting these competing demands, enabling faster content delivery without sacrificing ethical rigor.

For photo editors, the system represents an augmentation of their capabilities rather than a replacement. By automating the routine screening of the majority of images, the system allows editors to focus their expertise on complex cases that truly require human judgment. This shift from routine to strategic work has the potential to enhance job satisfaction and professional development.

For news organizations, the framework offers operational benefits including reduced costs, increased consistency, and improved scalability. The ability to maintain ethical standards across large, distributed editorial teams is particularly valuable for global news organizations operating across diverse cultural and regulatory contexts.

7.4 Addressing Challenges and Limitations

While the research demonstrates significant potential, it also highlights important challenges that must be addressed for responsible deployment:

Contextual Understanding: The current framework's exclusive reliance on visual content limits its ability to make context-aware ethical judgments. Future development must address this limitation through multimodal integration that considers captions, article text, and metadata.

Cultural and Regional Variations: The potential for cultural bias in the model's classifications requires ongoing attention. Diverse dataset construction, domain adaptation techniques, and cultural awareness in deployment are essential for ensuring the system's global applicability.

Evolving Ethical Standards: The dynamic nature of journalistic ethics demands that the system evolve continuously. Regular retraining, feedback incorporation mechanisms, and adaptive learning approaches are necessary to maintain relevance over time.

Bias and Fairness: The identification of potential socioeconomic and geographic biases underscores the importance of fairness-aware machine learning. Ongoing monitoring, debiasing techniques, and representative dataset construction are critical for equitable system performance.

Transparency and Explainability: The "black box" nature of deep learning models presents challenges for trust and accountability. Investment in explainable AI techniques is essential for enabling editors to understand and effectively use the system.

7.5 Recommendations for Practice

Based on the findings of this research, the following recommendations are offered for news organizations considering the adoption of AI-assisted visual ethics classification:

1. Adopt a Phased Implementation Approach: Begin with the system as a supportive tool while maintaining comprehensive human oversight. Gradually increase automation levels as trust and confidence grow, always ensuring that final decision-making authority remains with human editors.
2. Establish Clear Policies and Protocols: Develop explicit policies for the use of AI-assisted editorial tools, including thresholds for automatic approval, procedures for human review, mechanisms for appealing AI decisions, and protocols for error reporting and system improvement.
3. Invest in Training and Support: Provide comprehensive training to editors on the capabilities and limitations of the system, helping them develop effective strategies for working with the AI tool. Foster a culture of collaboration between human and artificial intelligence.
4. Commit to Continuous Improvement: Implement systems for monitoring model performance, collecting feedback from editors, and regularly updating the model with new data. Treat the system as a living entity that evolves alongside journalistic practice.
5. Prioritize Transparency and Accountability: Develop clear guidelines for accountability when AI systems are used in editorial decision-making. Consider how to communicate the use of AI to audiences, building trust through transparency about the role of automation in content selection.

6. Invest in Diverse and Representative Data: Commit resources to curating diverse training datasets that represent the full range of global contexts in which journalism operates. Recognize that the quality and representativeness of training data fundamentally determine the system's fairness and effectiveness.

7.6 Contributions to Knowledge

This research makes several contributions to the broader body of knowledge:

Theoretical Contributions: The study advances understanding of how deep learning can be applied to ethical reasoning in visual content, demonstrating that machine learning systems can learn and apply complex ethical criteria. This extends the theoretical boundaries of AI capabilities beyond objective pattern recognition into normative judgment.

Methodological Contributions: The research provides a methodological framework for developing and evaluating AI systems for ethical content classification. The comprehensive approach encompassing dataset construction, model training, performance evaluation, implementation analysis, and ethical consideration offers a template for future research in this domain.

Practical Contributions: The documented system architecture, performance benchmarks, and implementation workflows provide practical guidance for news organizations seeking to adopt similar technologies. The API design and integration considerations offer concrete technical solutions.

Ethical Contributions: By identifying and analyzing the ethical implications of AI in journalism, this research contributes to the development of responsible AI practices. The framework for considering augmentation versus replacement, transparency, accountability, and bias provides a foundation for ethical AI deployment in media contexts.

References

1. Rijal, S., Mardin, A., Anas, A., Sharif, T. C., & Sunardi, S. (2025). Ethical Analysis of Online Media Journalistic Photos Worth Publishing Based on Images Using the Convolutional Neural Network Method. Journal of Applied Informatics and Computing, *9*(6), 3919-3928.
2. Raza, S., & Pandya, D. (2024, October 23). New multimodal dataset will help in the development of ethical AI systems. Vector Institute for Artificial Intelligence.
3. Paik, S., Bonna, S., Novozhilova, E., Gao, G., Kim, J., Wijaya, D. T., & Betke, M. (n.d.). Affect-labeling codebooks for news images based on journalism ethics and photography guidelines. Boston University.
4. National Institutes of Health (NIH). (n.d.). Figure 1: Illustration of example Convolutional Neural Network architectures. PMC.
5. National Institutes of Health (NIH). (n.d.). Figure 3: A simplified overview of the AlexNET CNN architecture. PMC.
6. National Institutes of Health (NIH). (n.d.). Figure 2: A basic CNN architecture used for image classification. PMC.
7. Nature. (2021). Figure 3: The CNN architecture of multiple cancer classification. Scientific Reports.
8. Comparative Analysis of Blur Augmentation Strategies for Harmful Content Image Classification Using EfficientNet-B0: A Preliminary Study. (2026). Zenodo.
9. Digital divides in scene recognition: uncovering socioeconomic biases in deep learning systems. (2025). Humanities and Social Sciences Communications, *12*, Article 414.
10. Tong, S., Salaun, A., Yuan, V., Adeyeri, A., & Kagal, L. (2026). Mitigating Bias in Concept Bottleneck Models for Fair and Interpretable Image Classification. arXiv.
11. Society of Professional Journalists. (2024). Code of Ethics. Retrieved from https://www.spj.org/ethicscode.asp.
12. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. Communications of the ACM, *60*(6), 84-90.
13. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, *521*(7553), 436-444.
14. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.
15. Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.
16. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 770-778.
17. Szegedy, C., et al. (2015). Going deeper with convolutions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1-9.
18. Tan, M., & Le, Q. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. International Conference on Machine Learning, 6105-6114.
19. Howard, A. G., et al. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.
20. Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1251-1258.

21. Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 7132-7141.
22. Dosovitskiy, A., et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.
23. Liu, Z., et al. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. Proceedings of the IEEE/CVF International Conference on Computer Vision, 10012-10022.
24. Radford, A., et al. (2021). Learning transferable visual models from natural language supervision. International Conference on Machine Learning, 8748-8763.
25. Ramesh, A., et al. (2021). Zero-shot text-to-image generation. International Conference on Machine Learning, 8821-8831.
26. Rombach, R., et al. (2022). High-resolution image synthesis with latent diffusion models. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 10684-10695.
27. Saharia, C., et al. (2022). Photorealistic text-to-image diffusion models with deep language understanding. arXiv preprint arXiv:2205.11487.
28. OpenAI. (2023). GPT-4 Technical Report. arXiv preprint arXiv:2303.08774.
29. Touvron, H., et al. (2023). Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.
30. Vaswani, A., et al. (2017). Attention is all you need. Advances in Neural Information Processing Systems, *30*.
31. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT, 4171-4186.
32. Brown, T. B., et al. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, *33*, 1877-1901.
33. Bommasani, R., et al. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.
34. Bender, E. M., et al. (2021). On the dangers of stochastic parrots: Can language models be too big? Proceedings of FAccT, 610-623.
35. Weidinger, L., et al. (2021). Ethical and social risks of harm from language models. arXiv preprint arXiv:2112.04359.
36. Narayanan, A., Khazaie, V. R., & Raza, S. (2025). Bias in the Picture: Benchmarking VLMs with Social-Cue News Images and LLM-as-Judge Assessment. arXiv preprint arXiv:2509.19659.
37. Katzman, L., et al. (2023). Representational harms in AI systems. Proceedings of AIES.
38. Papakyriakopoulos, O., & Mboya, A. (2023). Beyond the binary: Understanding algorithmic harms in image classification. Proceedings of FAccT.
39. Wang, A., et al. (2023). Stereotypes and errors in image captioning. Proceedings of ACL.
40. Nielsen, R. K. (2021). The Power of Platforms. Oxford University Press.
41. Diakopoulos, N. (2019). Automating the News: How Algorithms Are Rewriting the Media. Harvard University Press.
42. Carlson, M. (2017). Journalistic Authority: Legitimating News in the Digital Era. Columbia University Press.
43. Zelizer, B. (2017). What Journalism Could Be. Polity Press.
44. Couldry, N., & Hepp, A. (2017). The Mediated Construction of Reality. Polity Press.
45. Waisbord, S. (2019). Communication: A Post-Discipline. Polity Press.
46. Allan, S. (2010). News Culture (3rd ed.). Open University Press.
47. Lester, P. M. (2015). Visual Journalism: A Guide for New Media Professionals. Routledge.
48. Newton, J. H. (2020). Visual ethics and photojournalism. In The Routledge Handbook of Visual Ethics, 115-129.
49. Wilkins, L., & Christians, C. G. (2020). The Routledge Handbook of Visual Ethics. Routledge.
50. Christians, C. G. (2019). Media Ethics: Cases and Moral Reasoning (11th ed.). Routledge.
51. Nilsson, M. (2025). Visual news editing and crisis coverage. In The Routledge Companion to Visual Journalism. Taylor & Francis.
52. Raemy, P., et al. (2025). Deepfakes and journalism: Normative considerations and implications. Journalism Studies, 1-21.
53. Ditlhokwa, G., Ncube, L., & Munoriyarwa, A. (2025). Shifting the gaze? Photojournalism practices in the age of artificial intelligence. Journalism Practice, 1-22.
54. Yang, Y., Zhao, Y., Xi, X., & Zhu, Y. (2025). Human-AI collaboration mechanism study on AIGC assisted image production for special coverage. arXiv preprint arXiv:2512.13739.
55. Nieman Foundation for Journalism. (2025). Generative AI Policy. Harvard University.
56. Baltodano, S. (2021). NSFW image classification for social media content filtering. GitHub Repository.
57. Arguedas, A. R., & Simon, F. (2023). News personalization and algorithmic systems. Journal of Communication.
58. Simon, F. (2024). AI in journalism: Challenges and opportunities. Digital Journalism.
59. Godulla, A., Hoffmann, C. P., & Seibert, J. (2021). Deepfakes in journalism: Opportunities and risks. Media and Communication.
60. Hameleers, M., van der Meer, T. G., & Dobber, T. (2023). Deepfakes and disinformation. Journal of Information Technology & Politics.
61. Vaccari, C., & Chadwick, A. (2020). Deepfakes and disinformation: Implications for democracy. Social Media + Society.
62. Westerlund, M. (2019). The emergence of deepfake technology. Technology Innovation Management Review, *9*(11), 39-52.

63. Citron, D. K., & Chesney, R. (2019). Deepfakes and the new disinformation war. Foreign Affairs, *98*(1), 147-155.
64. Tshuma, L. A. (2021). Photojournalism and political narratives in Zimbabwe. African Journalism Studies.
65. Klein-Avraham, I., & Reich, Z. (2016). The impact of digital platforms on photojournalism. Journalism, *17*(8), 987-1004.