Journal of Applied Informatics and Computing (JAIC)
Vol.10, No.1, February 2026, pp. 336~347
e-ISSN: 2548-6861

Evaluating Image Recognition Accuracy in Explicit Content Detection: A Comparative Study with Indonesian Perceptions

Rauhil Fahmi[1*],Deni Utama[2*], Muhammad Ridho Kurniawan Pratama[3*], Fathan Bainal Kaffi[4], Senoaji Pamungkas[5*]**
*Sistem dan Teknologi Informasi, Fakultas Teknik, Universitas Negeri Jakarta
**Ilmu Komputer Fakultas Matematikan dan Ilmu Pengetahuan Alam, Universitas Negeri Jakarta
rauhilfahmi@unj.ac.id[1], deniutama@unj.ac.id[2], muhammadridho@unj.ac.id[3], fathanbainalkaffi_1313621037@mhs.unj.ac.id[4], senoaji.pamungkas@mhs .unj.ac.id[5]

Article Info

Article history:

Received 2025-12-04
Revised 2026-01-13
Accepted 2026-01-20

Keyword:

Image Recognition,
Explicit Content,
Google Vision SafeSearch,
Comparative Analysis,
Indonesian Perceptions.

ABSTRACT

This study evaluates image recognition accuracy in explicit content detection by using the Indonesian social context as a comparative reference. Google Vision SafeSearch is employed as a representative automated image recognition system widely used in online content moderation. Although such systems provide efficiency in detecting adult, violent, or racy content, challenges arise when their detection outputs must align with more conservative cultural and religious norms, such as those in Indonesia. A quantitative descriptive-comparative method was applied by testing six representative images based on SafeSearch explicit content categories (adult, racy, violence, medical, and spoof) and comparing the automated detections with Indonesian respondents' perceptions collected through a Likert-scale questionnaire. Statistical analysis shows a significant difference between the system's explicit content classifications and human perceptions, with respondents consistently rating explicitness higher than Google Vision API. Despite this difference, a strong Spearman rank correlation indicates that Google Vision SafeSearch is consistent in ranking explicit content levels, although still limited in capturing emotional intensity and cultural sensitivity. These findings highlight how Indonesian social and cultural norms shape the perception of explicit imagery, emphasizing the need for image recognition systems that incorporate local contextual factors.

I. INTRODUCTION

The internet has greatly increased convenience in modern society by serving as a platform for communication, self-expression, and a wide range of activities. Applications such as social media, video-sharing platforms, discussion forums, and blogs facilitate these interactions. Users can access and upload image and video content with few restrictions, regardless of whether the content is original or sourced from other online locations. Although these platforms provide significant benefits to diverse demographic groups, the presence of explicit or harmful material uploaded by some users poses substantial challenges [1], [2]

Explicit content encompasses various types of material considered inappropriate or harmful. Examples include images or videos depicting sexual activities, extreme physical violence, promotion of hatred, or the use and production of illegal drugs and other dangerous substances [3], [4]. This phenomenon has created new problems, as such content is often visible to all users, negatively affecting individuals with trauma, phobias, or teenagers exposed to such material [3]. These issues often arise because the descriptions provided by uploaders frequently do not match the actual content [1].

Given that internet users have distinct preferences regarding content avoidance, numerous websites and applications have developed systems to filter explicit material. Examples include the Mute Words feature on X, Not Interested on YouTube, Restricted Mode on TikTok, and Sensitive Content Filters on Instagram. Furthermore, platforms such as Facebook and Reddit employ artificial

intelligence (AI) algorithms to automatically detect and flag explicit content [5], [6].

Despite these advancements, current filtering systems still fall short in protecting users from explicit content. Explicit material may still bypass filters when uploaders provide inaccurate descriptions or intentionally mislead the system, thereby reducing filtering effectiveness. Developers continue to update and enhance detection capabilities [1], [2] to address the evolving characteristics of harmful content [1], [3].

Therefore, relying solely on descriptions provided by the uploader is insufficient for filtering media content. The internet requires additional systems capable of analyzing media content itself, such as images and videos. These systems must classify content into specific categories; the more precise the classification performed by the system, the more effectively explicit content can be filtered [6], [7].

Image recognition is a technology that enables the creation of systems to filter explicit content. Through these systems, image recognition plays an important role in selecting and categorizing explicit media content. Thus, the filtering process does not only rely on descriptions provided by users but also directly analyses the content itself. Image recognition has become a crucial technology for identifying objects in visual media, as it mimics human recognition processes by analysing patterns, lines, colours, and relationships between objects, resulting in increasingly precise object identification [5], [6].

With the rapid development of AI technology, image recognition systems such as Google Vision SafeSearch enable digital platforms to automatically detect explicit content on a large scale [1]. These systems assign probabilities to content categories such as “adult”, “spoof”, “medical”, “violence”, or “racy,” allowing platforms to adjust their moderation policies accordingly. However, recent studies show that although this technology offers efficiency and consistency, automated systems often fail to understand the cultural context that influences people’s perceptions of explicit content [2], [3]. In Indonesia, cultural and religious norms tend to be more conservative than in Western countries, so content considered “safe” by automated systems may still be deemed inappropriate by local communities [2].

In addition, this study is aligned with the United Nations Sustainable Development Goals (SDGs) [8], particularly SDG 16 (Peace, Justice, and Strong Institutions), which emphasizes the protection of individuals from exposure to violence, exploitation, and harmful content online, and SDG 9 (Industry, Innovation, and Infrastructure), which promotes the advancement of safe and responsible digital innovations.

This study examines the effectiveness of image recognition systems in detecting explicit content by comparing automated classification results with the perceptions of Indonesian respondents. This comparative approach aims to reveal the extent to which the technical capabilities of automated systems align with Indonesia’s social and cultural norms. The research also seeks to contribute to the development of content filtering technology that is more sensitive to local values, thereby supporting a safer and more comfortable internet experience for Indonesian society.

The objective of this research is to evaluate the alignment between automated image recognition assessments and Indonesian users’ perceptions of explicit visual content. The findings aim to provide practical benefits by helping Indonesian society reduce exposure to explicit online content, thereby enhancing the safety and comfort of internet use. Additionally, the study aims to offer guidance for technology developers in creating or optimizing visual content filtering systems that are more effective and efficient, particularly within the Indonesian context.

II. Literature Review

A. Image Recognition

Recent advances in artificial intelligence (AI) and machine learning (ML) have dramatically improved the capabilities of image recognition systems. Convolutional neural networks (CNNs), in particular, have enabled models to learn complex visual patterns directly from pixel data without requiring manual feature extraction. Previous studies [9], [10]emphasize that CNNs revolutionized image recognition performance by allowing networks to capture low-level details and high-level semantic features hierarchically. These capabilities have paved the way for image recognition to handle challenging tasks such as object detection, facial recognition, and content moderation at unprecedented levels of accuracy.

Building on this progress, researchers like Yousaf *et al.* [11]have shown how transfer learning and deeper network architecture extend CNN applications to specialized domains with limited labelled data. These techniques significantly lower the barrier to entry for applying image recognition to new problems, including detecting explicit content in online images. As a result, platforms can deploy powerful image classifiers capable of evaluating billions of images daily, automating what would otherwise be an impossible human task.

However, critical voices argue that the effectiveness of image recognition systems is undermined by their lack of cultural and contextual awareness. Shahid and Vashistha [12] note that AI moderation tools trained on Western-centric datasets often misclassify content when deployed in non-Western contexts, where definitions of explicitness can differ markedly. Similarly, Mohammadi *et al.* [13] highlight that cultural symbols, traditional clothing, or modesty norms specific to regions like Indonesia are frequently overlooked by generic AI models. These critiques point out that a purely technical focus on accuracy is insufficient when content moderation requires sensitivity to cultural meaning and context.

A synthesis of these perspectives suggests that while CNN-based image recognition systems offer significant technical capabilities for detecting explicit content, their effectiveness in diverse cultural contexts depends on careful calibration to local norms. Combining technical models with human

feedback or retraining on culturally specific datasets could improve moderation accuracy and acceptability [12]. This conclusion underscores the importance of studies that evaluate whether AI outputs align with user perceptions in particular cultural environments, such as Indonesia, to bridge the gap between technical performance and social expectations.

B. Image Recognition Capabilities

Modern image recognition systems possess a range of sophisticated capabilities critical for detecting explicit content. By analysing colour distributions, textures, shapes, and spatial relationships between objects, these systems can identify patterns indicative of nudity, sexual acts, or violent scenes [10]. Deep CNNs enable fine-grained detection, distinguishing subtle differences between explicit and non-explicit images even under challenging conditions such as cluttered backgrounds or inconsistent lighting.

Advanced models assign probability scores reflecting the likelihood of different content types, such as adult, violent, or medical imagery, which allows platforms to customize moderation thresholds. Tools such as Google Vision SafeSearch provide multi-level ratings, offering nuanced classifications ranging from "very unlikely" to "very likely" explicit. This granularity permits flexible moderation policies tailored to platform guidelines or local legal requirements.

Some models incorporate object detection to capture context, recognizing interactions for example, two people in a sexual pose rather than isolated features. However, these systems still struggle with cultural, artistic, or situational contexts. They may misclassify art, satire, or culturally specific attire because they cannot interpret meaning beyond visual patterns [12]. Additionally, the performance of image recognition models often declines when applied in settings that differ significantly from their training data, such as non-Western cultural contexts, which can increase false positives or false negatives when detecting explicitness in region-specific imagery [13].

C. Explicit Content Detection

Scholars broadly agree that automated image recognition offers powerful solutions for moderating explicit content on digital platforms. According study in. [10], [13], [14], AI-based moderation systems enable rapid, large-scale detection of inappropriate material, protecting users from exposure to sexual or violent images. These systems use features extracted from CNNs to assign likelihood scores indicating the presence of explicit content categories, giving platforms control over thresholds for automatic blocking or review.

Research supporting AI moderation emphasizes efficiency and consistency. Automated systems can process massive volumes of images far faster than human reviewers, ensuring explicit material is removed before reaching vulnerable users. This ability is particularly valuable in social media environments where content is uploaded continuously, requiring real-time monitoring to comply with regulations and maintain community standards [1].

However, scholars such as Shahid and Vashistha [12] and Gorwa *et al.* [1] challenge the adequacy of fully automated moderation, arguing that these systems often misclassify images when they fail to understand cultural or contextual nuances. False positives, in which non-explicit images are mistakenly flagged, can stifle freedom of expression, while false negatives, in which explicit content is missed, undermine user safety. These limitations are exacerbated in cross-cultural deployments, where content considered explicit in one society may be entirely acceptable elsewhere. For example, certain traditional garments or performances may be flagged by AI systems unfamiliar with their cultural context.

The synthesis of these perspectives highlights that while automated explicit content detection systems provide scalability, their outputs should not be trusted in isolation. Integrating AI tools with human oversight or culturally adapted training data can improve accuracy and legitimacy [13]. By comparing automated detection outputs with perceptions from local users, researchers can assess whether these tools are aligned with community expectations, offering insights into how AI moderation can be refined to balance efficiency and cultural sensitivity.

D. Indonesian Perspective on Explicit Content

Indonesia's legal, religious, and cultural frameworks create stricter standards for explicit content than those in many Western societies. Alhakim [15] describes how national laws, rooted in Islamic values and moral codes, criminalize the production, distribution, and possession of pornographic or sexually explicit material. Meilani *et al.* [16] further emphasize that Indonesian norms extend beyond legal definitions of pornography to encompass more subtle forms of explicitness, such as provocative clothing or suggestive behaviour, which may also be socially or legally unacceptable.

These cultural expectations place unique demands on platforms operating in Indonesia. Unlike Western audiences, Indonesian users may consider images that show partial nudity or intimacy as explicit, even if such images do not meet the threshold for explicitness in international AI models [17]. This discrepancy creates a risk of either failing to filter offensive content or censoring benign material that reflects local customs or traditions.

Critics argue that global AI moderation systems are ill-suited to Indonesia's stricter standards because these systems typically reflect Western norms of explicitness. Shahid and Vashistha [12] note that without retraining or local adaptation, AI systems may produce inconsistent moderation outcomes that fail to respect Indonesian values. For example, cultural events or traditional dances involving exposed skin might trigger automated filters not calibrated to local contexts, frustrating users and creating potential legal challenges for platforms.

As a synthesis, scholars advocate for moderation systems that integrate local knowledge and user perceptions to ensure culturally appropriate detection of explicit content. This approach aligns AI capabilities with the expectations of Indonesian society, balancing technological efficiency with cultural legitimacy. Consequently, comparative research such as studies matching AI-based SafeSearch results with Indonesian respondents' perceptions is essential to validate whether automated systems align with national values, providing actionable insights for refining moderation strategies in Indonesia.

E. Comparative Approach to Evaluating Content Moderation Systems

Recent studies emphasize the value of combining automated tools and human judgment to evaluate the performance of content moderation systems. In Studies [10], [18] highlight that automated systems, such as Google Vision SafeSearch, can process large volumes of visual content and assign probability scores indicating the likelihood of explicit or violent material. This capability makes automated moderation indispensable for platforms handling billions of images, offering consistent and scalable classification aligned with predefined categories such as "adult" or "violence."

However, other researchers question whether automated outputs alone are sufficient for evaluating the appropriateness of images, especially in culturally diverse societies. Shahid and Vashistha [12] argue that AI-based moderation systems often reflect Western norms, which may not align with local cultural or moral standards. For example, images flagged as "safe" by automated tools may still be perceived as explicit or offensive in more conservative societies, including Indonesia. A study [13] suggests that failing to incorporate local perspectives can lead to significant mismatches between technical classifications and user expectations, resulting in either under- or over-moderation.

Synthesizing these positions, recent comparative research demonstrates that combining AI outputs with human perceptions can provide a more accurate and culturally relevant assessment of moderation systems. Bao *et al.* [18] showed that comparing machine-generated labels with user ratings helps identify biases or cultural blind spots in AI systems, enabling refinement for specific contexts. Consequently, studies evaluating tools such as Google Vision SafeSearch against responses from local users offer a promising method for assessing both the technical capabilities of image recognition and its social acceptability. This comparative approach ensures that moderation systems not only detect explicit content accurately but also align with cultural and community standards.

III. METHOD

A. Research Approach

The study focuses on the five primary categories of the SafeSearch system: *adult*, *racy*, *violence*, *medical*, and *spoof*. A total of six images were selected to represent these categories, with one image assigned to each category to ensure balanced representation.

The use of a limited sample was guided by several methodological considerations. First, as an exploratory study, this research aims to establish an initial understanding of the perception gap between AI-based classification and human judgment. Such objectives prioritize focused comparative analysis over large-scale sampling[19], [20]. Second, a representative sampling strategy enables more in-depth statistical examination—such as mean differences, correlation analysis, and consensus measurement—while remaining practically feasible. Third, the primary objective of this study is hypothesis testing to determine whether systematic differences exist between AI and human perceptions, rather than developing predictive models that would require larger datasets [21], [22].

Each image was carefully selected as a clear exemplar of its respective SafeSearch category based on preliminary testing using the Google Vision API. Although the sample size is intentionally limited, it is considered appropriate for exploratory analysis. The methodological implications and limitations of this sampling approach are discussed in detail in Section V.E.

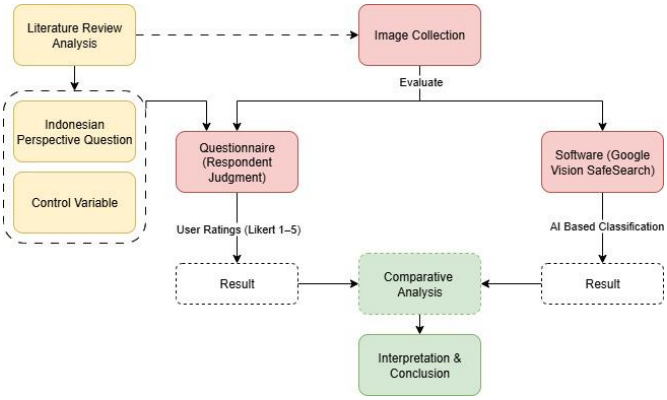

Figure 1. Research Procedure

The research procedure consists of two main stages. First, images are tested using the Google Vision "Try It" web tool. Second, public perceptions of these images are collected through a questionnaire containing Likert-scale items [23]. An overview of the research procedure is presented in Figure 1.

B. Population and Sampling

The population in this study consists of Indonesian internet users aged 18 years and above, as this age group is legally recognized as adults and is actively exposed to digital media content. A purposive sampling technique [19], [20] was employed to select respondents who met specific criteria, including being 18 years or older, residing in Indonesia, having experience using the internet and social media, and providing informed consent to participate in the study. Respondents under 18 years old, those who do not reside in Indonesia, or those who decline participation are excluded

from the sample. The sample size is determined based on the feasibility of data collection while ensuring sufficient representation for statistical analysis. A minimum sample size of 30 respondents is considered adequate for basic statistical analysis and non-parametric testing, as suggested by Sekaran and Bougie [20]. The detailed demographic profile of the respondents is presented in Section IV.A.

C. Research Hypothesis

Based on the research objectives and theoretical framework, this study aims to examine whether there is a significant difference between the evaluations produced by automated image recognition systems and the perceptions of Indonesian users regarding explicit visual content. The formulation of the hypothesis is grounded on the assumption that artificial intelligence systems, which rely on algorithmic training data, may interpret visual explicitness differently from human perception that is shaped by cultural and contextual factors.

Therefore, the hypotheses proposed in this study are as follows:

- H_0 (Null Hypothesis): There is no significant difference between automated image recognition assessments and Indonesian users' perceptions of explicit visual content.
- H_1 (Alternative Hypothesis): There is a significant difference between automated image recognition assessments and Indonesian users' perceptions of explicit visual content.

Testing these hypotheses allows the researcher to determine whether the automated image recognition results align statistically with human judgment, or whether perceptual and algorithmic evaluations diverge significantly in identifying explicit content.

D. Data Analysis Technique

The data collected from both sources, namely Google Vision API outputs and user perception questionnaires, were analysed quantitatively. The AI-generated "SafeSearch" scores were converted into numerical values from 1 to 5, corresponding to the likelihood scale ranging from *Very Unlikely* to *Very Likely*. Descriptive statistics were first calculated to summarize the distribution and overall trends of both datasets. Subsequently, normality tests were performed to determine the appropriate inferential approach [21]. Given the ordinal nature of the data, the relationship between AI assessments and human perceptions was analysed using the Spearman correlation test [24] In addition, the Wilcoxon signed-rank test was employed to examine whether there were significant differences between the automated detection scores and the respondent's perception ratings [22]. The results of these analyses were intended to determine whether automated image recognition aligns statistically with Indonesian viewers' perceptions of explicitness.

These analytical procedures were applied to examine whether automated image recognition outputs align with Indonesian cultural perceptions of visual explicitness. The results of these analyses are presented in the following section.

IV. RESULT

A. Respondent Demographic Profile.

A total of 38 respondents participated in this study. As shown in Table 1, the gender distribution was relatively balanced with 55.3% (21) female and 44.7% (17) male respondents. The majority of respondents (86.8%, 33 individuals) were aged 26–35 years, with smaller representations from the 18–25 age group (7.9%, 3 respondents) and those above 35 years (5.3%, 2 respondents). Regarding educational attainment, 84.2% (32 respondents) held a master's degree or higher, while 10.5% (4) held a bachelor's degree and 5.3% (2) had diploma qualifications. This demographic profile suggests the sample represents digitally active young adults with high educational attainment and strong media literacy [25]. The implications of this demographic composition for the generalizability of findings are discussed in Section V.E."

TABLE I
DEMOGRAPHIC PROFILE OF RESPONDENTS (N=38)

Demographic Variable	Category	Frequency (%)
Gender	Male	17 (44.7%)
	Female	21 (55.3%)
Age Group	18–25 years	3 (7.9%)
	26–35 years	33 (86.8%)
	Above 35 years	2 (5.3%)
Educational Level	Diploma	2 (5.3%)
	Bachelor's Degree	4 (10.5%)
	Master's Degree or Higher	32 (84.2%)

B. Google Vision Assessment

The first sample received an average score of 2.4. The highest contributor to this score was the Racy category, rated Very Likely. This indicates that Google Vision successfully identified the content as provocative or potentially suggestive. The high score in Racy contrasts with the low scores in Adult, Violence, and Medical, which were all rated Unlikely or Very Unlikely, meaning the tool did not detect explicit adult material, physical violence, or medical imagery.

The second sample received an average score of 2.0. The highest scores, both rated Possible, were found in the Medical and Racy categories. Google Vision identifies the image as having a Possible medical context due to the visible tissue, but the raw nature of the image also contributed to a Possible score in Racy, suggesting it could be disturbing or potentially offensive to some viewers. The low scores in Adult and Spoof confirm the image is neither explicit adult content nor a parody.

The third sample, featuring a magazine cover, received an average score of 2.4. Similar to the first sample, the highest score was in the Racy category, which was rated Very Likely. This high rating confirms that Google Vision detected the

image as highly suggestive due to the model's pose and minimal attire. The low scores of Unlikely or Very Unlikely across the Adult, Spoof, Medical, and Violence categories confirm that the API considers the image provocative but not outright explicit adult content, violent, or a parody.

The fourth sample received a low average score of 1.6. The highest score was Possible in the Racy category, reflecting the image's sensitive nature as the subjects are wearing underwear. However, the Possible rating suggests the API recognizes that the image is not overtly offensive or explicit. Crucially, all other categories Adult, Spoof, Medical, and Violence received scores of Unlikely or Very Unlikely. This confirms that Google Vision identified the image as general advertising or body positivity content, with minimal concern regarding explicit adult material or violence.

The fifth sample received an average score of 2.2. Similar to the first sample, the highest score was in the Racy category, which was rated Very Likely. This high rating confirms that Google Vision detected the image as highly suggestive due to the minimal and revealing nature of the subject's outfit. The low scores in categories like Adult, Violence, and Medical confirm the image is not classified as explicit adult material, violent, or medical. The score for Spoof was slightly higher at Unlikely, likely reflecting the possibility of the image being promotional or an intentional staged photo.

The sixth sample received a high average score of 3.6. The highest scores were in the Medical and Violence categories, both rated Very Likely, indicating the API recognized the image as containing severe physical harm and being explicitly graphic. The Racy category also received a high rating of Likely (4), likely due to the graphic nature of the wound. The low scores in Adult and Spoof confirm the image is not classified as explicit adult material nor is it identified as a hoax or parody. This image is deemed highly sensitive across multiple categories, with Violence and Medical being the primary concerns.

C. Respondent Assessment

The first sample received an average score of 3.47 from respondents, which is significantly higher than the 2.4 average score provided by the Google Vision API. The difference between the two ratings is 1.07. The high score from the respondents indicates they perceived the image as moderately to highly disturbing or inappropriate (with 57.9% scoring it 4 or 5), likely due to the highly suggestive Racy content. This contrast suggests that while the API successfully identified the Racy element (rating it Very Likely), human judgment assigned a stronger negative impact to the image, leading to a much higher overall score compared to the API's calculation

The second sample image received a significantly higher average score of 3.68 from respondents compared to the Google Vision API's score of 2.0. The substantial difference between the two ratings is 1.68. While the API categorized the image as generally safe, only rating the Medical and Racy categories as Possible, human respondents found the image highly disturbing, with 60.5% scoring it 4 or 5. This high human score indicates that the explicit, graphic nature of the exposed tissue strongly triggered negative responses, a severity which the API failed to fully capture, resulting in a large difference in perception.

The third sample received an average score of 3.97 from human respondents, which is considerably higher than Google Vision API's score of 2.4. The significant difference between these two ratings is 1.57. The high respondent score indicates strong discomfort, as 76.3% of respondents rated the image 4 or 5, suggesting they perceived the minimal attire and pose as highly inappropriate or disturbing. Although the API correctly identified the image as Very Likely to be Racy, its mathematical average was pulled down by the low scores in other categories, resulting in the API severely underestimating the level of disturbance caused by the highly suggestive visual content among human users.

The fourth sample image received an average score of 2.76 from human respondents, which is moderately higher than Google Vision API's score of 1.6. The difference between the two ratings is 1.16. The human score indicates that the image caused a noticeable level of discomfort, pulling the average past the midpoint (3). The API, however, heavily discounted the image, rating the Racy content only as Possible and giving low scores to all other categories. This disparity shows that while the API considered the image largely benign, human judgment applied a stricter standard due to the semi-exposed nature of the subjects, leading the API to underestimate the perceived offensiveness by over one full point.

The fifth sample received an average score of 3.34 from human respondents, which is moderately higher than Google Vision API's score of 2.2. The difference between the two ratings is 1.14. The human score indicates that the image caused a moderate level of discomfort, with 52.7% of respondents rating it 4 or 5. Although the API correctly identified the image as Very Likely to be Racy, its overall average remained low because it gave minimal concern to other categories. This disparity shows that the API underestimated the overall perceived disturbance, as human judgment applied a stricter standard due to the combined impact of the revealing costume and the context of the image.

The sixth sample received an extremely high average score of 4.79 from human respondents, which is significantly higher than the Google Vision API's score of 3.6. The difference between the two ratings is 1.19. The human score demonstrates near-universal condemnation, as 97.4% of respondents rated the image 4 or 5, indicating it was perceived as profoundly disturbing or highly inappropriate. While the API correctly rated both Medical and Violence as Very Likely, its resulting average of 3.6 severely underestimated the emotional impact and level of disturbance caused by the graphic content, failing to match the near-maximum severity assigned by human judgment.

D. Data Processing

Table 2 compares the mean explicitness scores generated by Google Vision API (A) and those provided by respondents (R) for six image samples. Across all samples, respondents consistently assigned higher scores than the API, with differences (R – A) ranging from 1.07 to 1.68. Sample 2, representing a medical-related image, had the largest difference of 1.68, while Sample 1, associated with spoof content, had the smallest difference of 1.07. The API's highest explicitness categories varied by sample, predominantly within Racy, Medical, and Violence classifications

TABLE II
COMPARISON OF MEAN EXPLICITNESS SCORES

Sample	A	R	(R - A)	Keyword Search	API Highest Category
1	2.4	3.47	1.07	Spoof	Racy (5)
2	2.0	3.68	1.68	Medical	Medical (3) / Racy (3)
3	2.4	3.97	1.57	Adult	Racy (5)
4	1.6	2.76	1.16	Racy	Racy (3)
5	2.2	3.34	1.14	Racy	Racy (5)
6	3.6	4.79	1.19	Violence	Medical (5) / Violence (5)

Table 3 presents the level of agreement among respondents for each image sample, measured by the standard deviation (SD). Samples 1 and 2 showed moderate consensus (SD = 1.27 and 1.21), while Samples 3, 4, and 5 exhibited high consensus (SDs between 1.04 and 1.14). Sample 6 demonstrated very high consensus (SD = 0.42), indicating minimal variation among respondents' ratings.

TABLE III
RESPONDENT RATING CONSENSUS (STANDARD DEVIATION)

Sample	Respondent (μ)	Standard Deviation (SD)	Interpretation
1	3.47	1.27	Moderate
2	3.68	1.21	Moderate
3	3.97	1.04	High
4	2.76	1.11	High
5	3.34	1.14	High
6	4.76	0.42	Very High

Table 4 shows the results of the paired sample t-test comparing API and respondent scores. The mean difference between the two sets was 1.268 (SD = 0.278). The t-test yielded $t(6) = 11.166$, $p < 0.001$, indicating a statistically significant difference between the scores from the Google Vision API and respondents [21].

TABLE IV
HYPOTHESIS TESTING: SIGNIFICANCE OF DIFFERENCES (PAIRED SAMPLE T-TEST)

Paired Samples Statistics	Value
N (Paired Samples)	6
Mean Difference ($\bar{d}$)	1.268
Standard Deviation of Difference (sd)	0.278
Paired Samples T-Test	
t-Statistic	11.166
Degrees of Freedom (df)	5
Sig. (2-tailed) p-value	< 0.001

Table 5 presents the Spearman correlation analysis examining the rank consistency between API and respondent ratings. The correlation coefficient ($\rho = 0.929$, $p = 0.007$) indicates a strong positive relationship between the two ranking systems, suggesting that while absolute scores differ, the relative ordering of explicit content remains consistent across AI and human assessments.

TABLE V
RELATIONSHIP ANALYSIS: RANK VALIDITY (SPEARMAN CORRELATION)

Correlation Statistic	Value
$\sum d^2$ (Sum of Squared Differences in Rank)	2.5
Spearman's Rho (ρ)	0.929
Sig. (2-tailed) p-value	0.007

E. Indonesian Perception

Table 6 summarizes respondent consensus and sensitivity levels toward disturbing content. Respondents most strongly agreed or were annoyed by images showing blood or open wounds (86.8%, 33 respondents), followed by negative impact on morals (84.2%, 32 respondents). High agreement was also observed for explicit sexual/minimal clothing (76.3%, 29 respondents) and medical procedures violating norms (73.6%, 28 respondents). Moderate agreement was recorded for sensual content and sexual/body spoof content, each with 68.4% (26 respondents).

TABLE VI
RESPONDENT CONSENSUS AND SENSITIVITY LEVELS TOWARD DISTURBING CONTENT

Content Category	Question	Respondents Agree/Annoyed (Score 4 & 5)
Blood/Open Wounds	Images showing blood or open wounds should not be displayed on public social media.	86.8% (33 Respondents)
Negative Impact on Morals	Do you agree that explicit content (adult, spoof, violence, medical, racy) on social media negatively impacts the moral values of Indonesian society?	84.2% (32 Respondents)
Explicit Sexual/Minimal Clothing	How disturbing do you find it to see sexually explicit content (e.g., nudity or very revealing clothing) in public places or in the media?	76.3% (29 Respondents)
Medical Procedures (Violating Norms)	In your opinion, does displaying images of medical procedures (e.g., surgeries, open wounds, or blood) in public media violate the norms of	73.6% (28 Respondents)

	decency in Indonesian society?	
Sensual Content	To what extent do you feel disturbed by content displaying sensuality, such as erotic dancing, very revealing clothing, or seductive expressions, even if it does not contain explicit pornography?	68.4% (26 Respondents)
Sexual/Body Spoof Content	How disturbing do you find spoof content that exposes the body or depicts sexual acts for comedic purposes?	68.4% (26 Respondents)

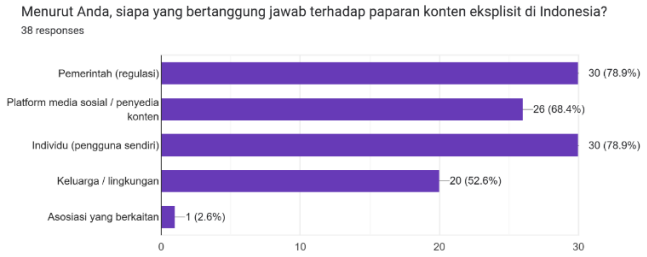

Figure 2. Respondent Views on Responsibility for Explicit Content Exposure in Indonesia

Respondents were asked to indicate which parties they considered responsible for exposure to explicit content on social media. The results are summarized in Figure 2 in terms of the number and percentage of respondents selecting each category. Both government (regulation) and individual users were identified by the highest number of respondents, with 30 individuals each (78.9%) assigning responsibility to these parties. Social media platforms or content providers were selected by 26 respondents (68.4%), while family or the surrounding environment was considered responsible by 20 respondents (52.6%).

The findings presented in this Results section reveal systematic differences between automated assessments and human perceptions, as well as clear patterns in Indonesian cultural sensitivity toward different types of explicit content. The following Discussion section interprets these findings and examines their broader implications.

V. DISCUSSION

The statistical analyses presented in the Results section reveal a significant and consistent gap between automated AI assessments and Indonesian users' perceptions of explicit visual content. This gap extends beyond purely technical or statistical concerns, carrying important implications across multiple dimensions: the practical operation of content moderation systems, the ethical governance of AI in culturally diverse contexts, and the technical approaches needed to develop more culturally intelligent systems. This section interprets the main findings, examines their practical, ethical, and technical implications, and acknowledges the limitations of the current study.

A. Interpretation of Main Findings

Based on the results of the study, there is a significant difference between algorithmic assessments and human evaluations, indicating that the null hypothesis (H_0) can be rejected, and the alternative hypothesis (H_1) is accepted. This finding is supported by the paired-sample t-test results, showing $t(6) = 11.166$ with $p < 0.001$, as well as mean score differences (R – A) ranging from 1.07 to 1.68, indicating that respondents consistently rated the explicitness of content higher than Google Vision API.

Although there is an absolute difference in scores, rank analysis using Spearman's rho shows a strong correlation between API and respondent ratings ($\rho = 0.929$, $p = 0.007$). This indicates that Google Vision API is able to recognize consistent visual patterns when determining the ranking of explicit content, even though the intensity or emotional impact of the content is not fully reflected in the system's evaluation. In other words, AI is effective at detecting categories such as Racy, Medical, and Violence, but it is still limited in capturing human perception influenced by cultural context and social norms.

The method used by Google Vision API, namely Convolutional Neural Networks (CNNs), is effective in recognizing visual patterns and detecting objects but has limitations under conditions of uneven lighting, varying postures, large object scales, or occlusion, which can lead to detection errors. Additionally, high image complexity can reduce object recognition accuracy, aligning with findings that CNNs still experience limitations in accurately assessing complex visual elements [26]. Furthermore, previous studies have shown that the system is sensitive to image rotation and noise. Image rotation between 45°–340° can cause interpretation errors of up to 80–100% [27], while even 10% noise can alter object interpretation [28]. These technical limitations partly explain the differences between API scores and human ratings, as respondents evaluate content based on visual perception, context, and emotional impact, whereas Google Vision API only assesses pure visual attributes.

Analysis of respondent consensus reveals variation in perception, with Samples 1 and 2 exhibiting moderate consensus (SD = 1.27 and 1.21), Samples 3 to 5 showing high consensus (SD = 1.04–1.14), and Sample 6 showing very high consensus (SD = 0.42). This confirms that humans have uniform judgment for clearly disturbing content, while the system tends to be more rigid, considering only visual attributes. Analysis of content categories deemed disturbing indicates that images of blood or open wounds cause the highest level of discomfort (86.8%), followed by content with a perceived negative impact on public morals (84.2%), sexual/minimal clothing content (76.3%), and medical procedures violating norms (73.6%). Sensual and spoof content received lower agreement levels (68.4%). These

findings demonstrate that human perception of explicit content is influenced by content category, social norms, and cultural sensitivity, which are not fully captured by the API.

Respondents indicated that the responsibility for the dissemination of explicit content primarily lies with the government and individual users (78.9%), followed by social media platforms (68.4%) and family or surrounding environment (52.6%). This emphasizes that although AI technology can assist in content identification, human oversight and social regulation remain necessary to ensure that distributed content aligns with local norms and cultural values.

B. Practical Implications for Content Moderation

The significant gap between API assessments and Indonesian perceptions has important practical implications for content moderation. The consistent pattern of respondents rating content as more explicit (mean differences 1.07–1.68) suggests two primary risks: underblocking and overblocking.

Underblocking occurs when AI systems classify content as less explicit than Indonesian users perceive it. For instance, Sample 2 (medical imagery) received a system score of 2.0 but a human rating of 3.68. If platforms rely solely on automated filtering with Western-calibrated thresholds, disturbing content such as graphic medical procedures or blood may remain visible, resulting in: (a) exposure of vulnerable populations to psychologically distressing content [3]; (b) user dissatisfaction and platform abandonment; (c) potential violations of Indonesian regulations criminalizing harmful content [15]; and (d) erosion of trust in platform safety mechanisms.

Conversely, overblocking arises if platforms apply overly strict thresholds to compensate for AI limitations. This could lead to excessive removal of culturally appropriate content such as traditional ceremonies or medical education materials, resulting in: (a) censorship of legitimate cultural and educational content [11], [12]; (b) disproportionate impact on creators and educators; (c) user frustration and circumvention behaviors; and (d) competitive disadvantages for compliant platforms.

These findings suggest that platforms operating in Indonesia should implement culturally adaptive moderation combining automated AI detection with human oversight by local moderators familiar with Indonesian norms. Indonesia-specific moderation thresholds could automatically flag borderline cases for human review rather than immediate publication or blocking [1], [18]. Additionally, tiered user controls allowing customization of filtering preferences across different content categories would empower users to align platform filtering with personal values while reducing platform liability.

Beyond these practical considerations, the findings raise fundamental ethical questions about algorithmic governance in culturally diverse societies.

C. Ethical and Regulatory Considerations

The AI-human perception gap raises important ethical concerns regarding algorithmic bias, censorship, and cultural self-determination. First, the systematic underestimation of explicitness by Google Vision SafeSearch reflects algorithmic bias rooted in Western training data, which may not reflect Indonesian norms [12], [13]. This imposes a form of cultural hegemony [29] where Western frameworks for assessing explicitness are privileged over local values. Ethically, technology companies have a responsibility to ensure AI systems are culturally responsive, acknowledging diverse moral frameworks of global users [4], [12]. Second, both underblocking and overblocking raise concerns about the balance between user protection and freedom of expression. Underblocking fails to protect users from psychological harm, particularly vulnerable groups [3], while overblocking risks suppressing legitimate cultural, educational, or artistic expression [1]. This balance is particularly delicate in Indonesia, where constitutional free expression guarantees coexist with laws criminalizing content harmful to public morals [15]. Automated systems lacking cultural nuance may either fail to protect users adequately or suppress legitimate expression, both constituting ethical failures. Third, the imposition of Western-trained AI systems without local adaptation represents technological colonialism, where digital infrastructure is exported globally without meaningful consultation with local communities [12]. Indonesian users have a right to digital environments that respect their cultural values. This principle suggests content moderation systems should be co-designed with local stakeholders, including users, civil society, religious leaders, and regulators [4].

From a regulatory standpoint, Indonesia could establish mandatory standards for AI content moderation requiring platforms to demonstrate compliance with Indonesian norms through periodic audits comparing AI outputs with local evaluator assessments. Regulations could require algorithmic transparency, obligating platforms to disclose training data sources and cultural adaptation mechanisms. Human review processes for borderline cases and co-regulatory models involving government, platforms, civil society, and users could ensure governance evolves with changing norms [1], [4]. Additionally, ethical governance requires proportionality in moderation—distinguishing between content warranting strict blocking (e.g., child exploitation) versus content requiring human oversight (e.g., medical imagery, cultural practices)—and robust accountability mechanisms including transparent criteria, user notification, accessible appeals, and regular reporting on moderation outcomes [11], [12].

While ethical frameworks provide necessary guidance, realizing culturally adaptive content moderation requires concrete technical approaches capable of integrating emotional and cultural factors into AI systems.

D. Technical Integration of Emotional and Cultural Factors

Several computational approaches could enable AI systems to better capture emotional intensity and cultural sensitivity. First, culturally specific training datasets compiled by Indonesian annotators applying local standards could calibrate AI models to Indonesian norms. Transfer learning techniques enable fine-tuning existing models on culturally specific datasets without complete retraining [11], [12], [13]. Second, multimodal analysis integrating visual content with textual descriptions, metadata, and cultural context markers could enable systems to assess explicitness differently based on context—applying stricter standards to ambiguous posts while being permissive for educational or cultural content [6]. Third, emotion recognition and affective computing techniques could assess emotional intensity by detecting facial expressions, body language, and compositional elements associated with distress. Training on Indonesian emotional response data would enable models to flag content based on predicted emotional impact, addressing the 86.8% disturbance rate for medical graphic content observed in this study [11].

Fourth, human-in-the-loop (HITL) systems using AI as an initial filter with validation by Indonesian moderators provide a practical hybrid approach. HITL creates feedback loops where moderator decisions continuously improve AI models through active learning [18]. Fifth, ensemble models combining multiple AI systems with different cultural calibrations and culturally weighted voting could acknowledge that no single model universally captures all perspectives [18]. Sixth, explainable AI (XAI) techniques [30] such as attention visualization could identify whether AI focuses on culturally relevant features, guiding model refinement and supporting user trust through transparency. Finally, adaptive threshold mechanisms [31] allowing dynamic calibration based on user preferences or community standards could accommodate diversity within Indonesian society.

Implementing these approaches faces challenges including dataset compilation costs, computational complexity, privacy concerns, HITL scalability, and system design complexity [32]. Despite these challenges, the gap documented in this study makes clear that culturally adaptive approaches are necessary for ethical and effective content moderation. Moving beyond purely visual pattern recognition toward holistic approaches incorporating cultural knowledge, contextual reasoning, emotional impact assessment, and human judgment represents a promising direction for developing AI systems that are both technically sophisticated and culturally intelligent.

Overall, this study demonstrates that Google Vision is effective in recognizing categories of explicit content visually, but its absolute scores are often lower than human assessments. The consistency of content ranking remains high, indicating valid visual recognition patterns. Human sensitivity to explicit content is influenced by content category, cultural context, and social norms, particularly for medical, violent, or sexual content. The findings underscore the need for combining AI technology with human oversight, culturally adaptive thresholds, transparent regulatory frameworks, and robust accountability mechanisms to ensure content moderation systems serve both user safety and freedom of expression within Indonesian cultural values.

The implementation challenges discussed above underscore the importance of acknowledging the limitations of the current study, which inform both the interpretation of findings and directions for future research.

E. Study Limitations

Despite providing valuable empirical evidence of the gap between automated systems and human perceptions in the Indonesian context, several limitations should be considered when interpreting the findings and planning future research. First, the sample size was limited to 38 respondents with demographic homogeneity concentrated in the 26–35 years age group (86.8%) and individuals holding master's degrees or higher (84.2%). This composition limits generalizability to younger respondents (18–25 years), older adults (above 35 years), and individuals with lower educational attainment, who may perceive explicit content differently due to variations in media literacy and adherence to conservative norms [2], [16]. Future research should employ stratified sampling across age, education, geographic regions, and socioeconomic backgrounds.

Second, the study examined only six images representing five SafeSearch categories. This limited sample may not capture the full complexity and variability of explicit content in real-world contexts, including visual ambiguity, cultural contexts, and technical conditions such as lighting, rotation, or noise that affect AI accuracy[27], [28]. Future studies should expand the dataset to include larger and more diverse image samples across multiple categories and conditions.

Third, while this study identifies significant differences between AI assessments and human perceptions, it does not deeply explore the practical consequences for content moderation policies, such as overblocking (false positives) or underblocking (false negatives) in the Indonesian [1], [12]. Future research should investigate real-world implications including user experience impacts, freedom of expression concerns, and platform compliance with local regulations.

Fourth, although this study recognizes that Google Vision API does not capture emotional intensity and cultural sensitivity, it does not propose specific technical mechanisms for integration. Future research should explore culturally specific training datasets, transfer learning techniques, multimodal analysis combining visual and contextual features, human-in-the-loop validation systems, and hybrid models integrating rule-based cultural filters with deep learning approaches [13], [18].

Fifth, this study focused exclusively on Google Vision SafeSearch without comparing performance with other AI systems such as Amazon Recognition, Microsoft Azure Computer Vision, or open-source models. Comparative

analysis could determine whether observed discrepancies are system-specific or represent broader limitations of current AI technologies. Future research should include multi-system comparisons to identify optimal tools for Indonesian contexts.

Finally, Indonesia's cultural diversity across regions, religions, and urban-rural settings was not systematically analyzed. Future research should investigate how specific cultural dimensions and regional contexts moderate perceptions of explicit content through comparative studies across different provinces or religious communities.

Despite these limitations, this study provides valuable empirical evidence of the discrepancy between automated systems and human perceptions in the Indonesian context, highlighting clear directions for future research to improve culturally adaptive content moderation technologies.

VI. CONCLUSION

This study evaluated the effectiveness of Google Vision API in detecting explicit content by comparing algorithmic assessments with Indonesian respondents' perceptions. The results demonstrate a significant difference between AI and human evaluations, with paired t-test results ($t(6) = 11.166$, $p < 0.001$) and mean score differences ranging from 1.07 to 1.68, confirming that respondents consistently rated explicitness higher than Google Vision API. This leads to rejection of the null hypothesis (H_0) and acceptance of the alternative hypothesis (H_1). Despite these absolute differences, Spearman's rank correlation ($\rho = 0.929$, $p = 0.007$) revealed that Google Vision maintains consistency in ranking explicit content, though it remains limited in capturing emotional intensity and cultural context that shape Indonesian perceptions.

The practical implications of these findings are significant for content moderation in Indonesia and similar culturally conservative contexts. The systematic underestimation of explicitness by AI systems creates risks of both underblocking (exposing users to disturbing content) and overblocking (censoring legitimate cultural expression) if platforms rely solely on automated filtering or apply overly strict compensatory thresholds. Effective content moderation in Indonesia requires hybrid approaches combining automated AI detection with human oversight by local moderators, culturally calibrated thresholds, and tiered user controls that empower individuals to customize filtering preferences.

From an ethical and regulatory standpoint, this study reveals how Western-trained AI systems can perpetuate cultural hegemony by imposing Western norms of explicitness on diverse global contexts. The gap between automated assessments and Indonesian perceptions raises concerns about algorithmic bias, the balance between user protection and freedom of expression, and the need for technological systems that respect cultural self-determination. Indonesia could establish mandatory standards for AI content moderation requiring platforms to demonstrate compliance with local norms through periodic audits, algorithmic transparency, human review processes, and co-regulatory models involving government, platforms, civil society, and users.

Technically, several computational approaches could enable AI systems to better capture emotional intensity and cultural sensitivity. These include training on culturally specific datasets compiled by Indonesian annotators, multimodal analysis integrating visual content with textual and contextual information, emotion recognition techniques trained on Indonesian emotional response data, human-in-the-loop systems combining automated filtering with validation by local moderators, ensemble models with culturally weighted voting, explainable AI techniques for transparency, and adaptive threshold mechanisms allowing dynamic calibration based on user preferences or community standards.

Several limitations should be acknowledged. The sample was limited to 38 respondents with demographic homogeneity (86.8% aged 26–35, 84.2% with master's degrees), restricting generalizability across age groups and educational backgrounds. Only six images were examined, potentially limiting representation of real-world content complexity and variability. The study did not analyze practical consequences such as overblocking or underblocking in depth, did not propose specific technical mechanisms for cultural sensitivity integration beyond conceptual discussion, did not compare Google Vision with other AI systems, and did not systematically examine regional cultural diversity within Indonesia.

Future research should employ stratified sampling across diverse demographics and regions, expand image datasets with varied contexts and conditions, investigate practical consequences for content moderation policies including user experience and regulatory compliance, explore technical integration of cultural sensitivity through culturally specific training data and human-in-the-loop systems, conduct comparative analysis across multiple AI platforms, and systematically analyze regional and religious variations in Indonesian perceptions. Additionally, longitudinal studies could examine how perceptions of explicitness evolve over time and how AI systems adapt to changing cultural norms.

Despite these limitations, this study provides critical empirical evidence that globally deployed image recognition systems may not align with local cultural norms, particularly in conservative societies such as Indonesia. The primary contribution is evaluative and reflective rather than algorithmic or technical, providing a critical assessment of how existing global AI systems perform when deployed in culturally diverse contexts. By systematically documenting the gap between AI assessments and Indonesian user perceptions, the study contributes empirical evidence to ongoing discussions in AI ethics, algorithmic fairness, and technology policy regarding cultural biases embedded in globally deployed systems. These findings are valuable for informing platform governance decisions, regulatory policy

development, and the design of more culturally responsive content moderation frameworks, contributing to ongoing efforts to develop AI technologies that are technically sophisticated, socially responsible, and culturally sensitive.

REFERENCES

[1] R. Gorwa, R. Binns, and C. Katzenbach, "Algorithmic content moderation: Technical and political challenges in the automation of platform governance," *Big Data Soc*, vol. 7, no. 1, Jan. 2020, doi: 10.1177/2053951719897945.

[2] M. Ruckenstein and L. L. M. Turunen, "Re-humanizing the platform: Content moderators and the logic of care," *New Media Soc*, vol. 22, no. 6, pp. 1026–1042, Jun. 2020, doi: 10.1177/1461444819875990.

[3] N. Sambasivan *et al.*, "'They don't leave us alone anywhere we go': Gender and digital abuse in South Asia," in *Conference on Human Factors in Computing Systems - Proceedings*, Association for Computing Machinery, May 2019. doi: 10.1145/3290605.3300232.

[4] N. Sambasivan, E. Arnesen, B. Hutchinson, T. Doshi, and V. Prabhakaran, "Re-imagining algorithmic fairness in India and beyond," in *FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, Association for Computing Machinery, Inc, Mar. 2021, pp. 315–328. doi: 10.1145/3442188.3445896.

[5] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive Angular Margin Loss for Deep Face Recognition." [Online]. Available: https://github.com/

[6] M. D. Zakir Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga, "A comprehensive survey of deep learning for image captioning," Nov. 30, 2019, *Association for Computing Machinery*. doi: 10.1145/3295748.

[7] "Detect explicit content (SafeSearch)," Cloud Vision API Documentation."

[8] "The 17 goals – Sustainable development."

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun ACM*, vol. 60, no. 6, pp. 84–90, Jun. 2017, doi: 10.1145/3065386.

[10] J. S. Lee, Y. M. Kuo, P. C. Chung, and E. L. Chen, "Naked image detection based on adaptive and extensible skin color model," *Pattern Recognit*, vol. 40, no. 8, pp. 2261–2270, Aug. 2007, doi: 10.1016/j.patcog.2006.11.016.

[11] K. Yousaf and T. Nawaz, "A Deep Learning-Based Approach for Inappropriate Content Detection and Classification of YouTube Videos," *IEEE Access*, vol. 10, pp. 16283–16298, 2022, doi: 10.1109/ACCESS.2022.3147519.

[12] F. Shahid, M. Elswah, and A. Vashistha, "Think Outside the Data: Colonial Biases and Systemic Issues in Automated Moderation Pipelines for Low-Resource Languages," Aug. 2025, [Online]. Available: http://arxiv.org/abs/2501.13836

[13] E. Mohammadi, Y. Cai, A. Novin, V. Vera, and E. Soltanmohammadi, "Who is a scientist? Gender and racial biases in google vision AI," *AI and Ethics*, vol. 5, no. 5, pp. 4993–5010, Oct. 2025, doi: 10.1007/s43681-025-00742-4.

[14] "International Journal of Artificial Intelligence and Machine Learning in Engineering 763|p AI in Automated Content Moderation on Social Media."

[15] A. Alhakim, "Criminal Control for the Distribution of Pornographic Content on the Internet: An Indonesian Experience," 2021, [Online]. Available: https://ejournal.undiksha.ac.id/index.php/jkh

[16] N. Meilani, S. S. Hariadi, and F. T. Haryadi, "Social media and pornography access behavior among adolescents," *Int J Publ Health Sci*, vol. 12, no. 2, pp. 536–544, Jun. 2023, doi: 10.11591/ijphs.v12i2.22513.

[17] M. P. F. Purwaningtyas and C. K. A. Wibowo, "Negotiating Sexuality: Indonesian Female Audience towards Pornographic Media Content," *IKAT: The Indonesian Journal of Southeast Asian Studies*, vol. 5, no. 2, Apr. 2022, doi: 10.22146/ikat.v5i2.70077.

[18] H. Bao *et al.*, "VModA: An Effective Framework for Adaptive NSFW Image Moderation," May 2025, [Online]. Available: http://arxiv.org/abs/2505.23386

[19] J. W. Creswell, *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*, 4th ed. Thousand Oaks, CA, USA: SAGE, 2014.

[20] U. Sekaran, *Research Methods for Business: A Skill-Building Approach*, 7th ed. Hoboken, NJ, USA: Wiley, 2016.

[21] A. Field, *Discovering Statistics Using IBM SPSS Statistics*, 5th ed. London, UK: SAGE, 2018.

[22] F. Wilcoxon, "Individual Comparisons by Ranking Methods," 1945. [Online]. Available: https://www.jstor.org/stable/3001968

[23] A. Joshi, S. Kale, S. Chandel, and D. Pal, "Likert Scale: Explored and Explained," *Br J Appl Sci Technol*, vol. 7, no. 4, pp. 396–403, Jan. 2015, doi: 10.9734/bjast/2015/14975.

[24] C. Spearman, "The Proof and Measurement of Association between Two Things," Autumn-Winter, 1987.

[25] W. J. Potter, "The state of media literacy," *J Broadcast Electron Media*, vol. 54, no. 4, pp. 675–696, Oct. 2010, doi: 10.1080/08838151.2011.521462.

[26] H. Hosseini, B. Xiao, M. Jaiswal, and R. Poovendran, "On the Limitation of Convolutional Neural Networks in Recognizing Negative Images," Aug. 2017, [Online]. Available: http://arxiv.org/abs/1703.06857

[27] A. Apte *et al.*, "Countering Inconsistent Labelling by Google's Vision API for Rotated Images."

[28] H. Hosseini, B. Xiao, and R. Poovendran, "Google's Cloud Vision API Is Not Robust To Noise," Jul. 2017, [Online]. Available: http://arxiv.org/abs/1704.05051

[29] P. Ricaurte, "Data Epistemologies, The Coloniality of Power, and Resistance," *Television and New Media*, vol. 20, no. 4, pp. 350–365, 2019, doi: 10.1177/1527476419831640.

[30] A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, Sep. 2018, doi: 10.1109/ACCESS.2018.2870052.

[31] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," Nov. 01, 2016, *Springer Verlag*. doi: 10.1007/s13748-016-0094-0.

[32] K. Holstein, J. W. Vaughan, H. Daumé, M. Dudík, and H. Wallach, "Improving fairness in machine learning systems: What do industry practitioners need?," in *Conference on Human Factors in Computing Systems - Proceedings*, Association for Computing Machinery, May 2019. doi: 10.1145/3290605.3300830.