Scientific Journal of Informatics
Vol. 13, No. 3, August 2026
p-ISSN 2407-7658 https://journal.unnes.ac.id/journals/sji e-ISSN 2460-0040

Virtual vs Human Influencers: A Comparative Analysis Of Campaign Effectiveness Using Deepfake Detection Technology From The Perspective Of Ethics And State Responsibility Under The Rule Of Law

Safira Hasna Setiyani[1]*, Nabila Aliffia[2], Fanny Agustina Sri Rahayu[3]
[1]Department of Informatics, Faculty of Law, Management and Informatic, Universitas Karya Husada Semarang, Indonesia
[2]Department of Management, Faculty of Law, Management and Informatic, Universitas Karya Husada Semarang, Indonesia
[3]Department of Law, Faculty of Law, Management and Informatic, Universitas Karya Husada Semarang, Indonesia

Abstract.
Purpose: This study aims to examine the effectiveness of AI-driven influencer marketing, develop an Xception-based Convolutional Neural Network (CNN) for deepfake detection, and analyze ethical and legal responsibility for artificial intelligence use from a rule-of-law perspective.
Methods: A mixed-methods approach was employed by integrating Computer Science, Management, and Law perspectives. Quantitative analysis used Partial Least Squares Structural Equation Modeling (PLS-SEM) to examine the relationships among AI Personalization, AI Interaction, User Experience, and Trust. The Xception-based CNN was evaluated using standard classification metrics. Qualitative analysis involved expert interviews and examination of legal principles concerning AI governance, transparency, accountability, and consumer protection.
Result: The Xception-based CNN achieved 97.12% accuracy and an AUC of 0.9920 in distinguishing authentic from AI-generated content. PLS-SEM results indicate that AI Interaction significantly influences User Experience and Trust, while User Experience has the strongest effect on Trust ($\beta = 0.520$; $p < 0.001$). The model explains 53.3% of the variance in Trust, and User Experience significantly mediates the relationship between AI Interaction and Trust. The legal analysis highlights the need for stronger AI governance addressing transparency, disclosure of AI-generated content, consent, accountability, and consumer protection.
Novelty: The novelty of this study lies in integrating deepfake detection, AI-driven influencer marketing, consumer trust analysis, and rule-of-law perspectives into a unified framework for responsible digital marketing. This integrated approach provides a multidisciplinary perspective for balancing technological innovation with ethical responsibility, legal accountability, and consumer protection.

Keywords: Virtual Influencer, CNN, Deepfake Technology, Detection, Legal Ethics

Received July 2026 / **Revised** August 2026 / **Accepted** August 2026

INTRODUCTION

The rapid advancement of Generative Artificial Intelligence (AI) has fundamentally transformed digital communication by challenging conventional perceptions of authenticity[1]. Through increasingly sophisticated generative models, digital systems can produce highly realistic images, videos, voices, and human-like representations, making the distinction between authentic and synthetic content increasingly difficult[2]. This transformation has significant implications for digital marketing because consumer responses are strongly influenced by perceptions of credibility and trustworthiness [3]. When consumers cannot confidently determine whether the content they encounter is authentic, uncertainty may weaken trust, engagement, and purchase intention [4]. Therefore, authenticity has become an increasingly important determinant of effective AI-mediated marketing communication [5].

[1]*Corresponding author.
Email addresses: safirahasna@unkaha.ac.id(Setiyani), nabila@unkaha.ac.id(Aliffia), fanny@unkaha.ac.id(Rahayu).
DOI: 10.15294/sji.v13i3.59356

This issue is particularly relevant to the growing adoption of virtual influencers as brand representatives [6]. Virtual influencers provide strategic advantages, including consistent visual identity, flexible content production, continuous availability, and greater control over brand representation [7]. However, their digitally constructed identities also create challenges concerning authenticity and transparency [8]. Unlike human influencers, whose credibility is often associated with perceived personal experience and human interaction, virtual influencers require consumers to evaluate the credibility of communication that is fundamentally mediated by AI [9]. Consequently, the effectiveness of virtual influencer marketing depends not only on content attractiveness but also on consumers' confidence in the authenticity and transparency of the underlying technology [10]. This condition extends conventional Source Credibility Theory by introducing technological authenticity as an important consideration in AI-mediated communication [11].

The authenticity challenge becomes more critical with the development of deepfake technology, which enables highly realistic manipulation of facial appearance, expressions, voices, and body movements [12]. In influencer marketing, such technology can blur the distinction between legitimate AI-generated representations and manipulated content [5], creating potential risks of deception, identity misuse, misinformation, and reduced consumer trust [13]. Thus, the marketing problem of authenticity creates a corresponding technological requirement: digital content must be capable of being objectively verified [14]. From an informatics perspective, deepfake detection using computer vision and deep learning provides a mechanism for addressing this requirement [15]. Convolutional Neural Networks (CNNs) have demonstrated strong capability in identifying subtle manipulation patterns [16], while the Xception architecture, through its depthwise separable convolution, offers an efficient approach for extracting discriminative spatial features from manipulated content [17].

Nevertheless, deepfake detection should not be viewed solely as a technical classification problem [18]. Although metrics such as accuracy, precision, recall, F1-score, and AUC are essential for evaluating model performance, the practical value of detection lies in how its results support trustworthy digital communication [19]. Authenticity verification can serve as an intermediary mechanism connecting technological analysis with consumer-oriented marketing evaluation by ensuring that content used in marketing campaigns can be assessed according to its authenticity [20]. In this sense, deepfake detection becomes not only a computer vision solution but also a decision-support mechanism for strengthening transparency and consumer confidence [21].

The technological verification of authenticity consequently leads to broader ethical and legal considerations [22]. AI-generated influencer content raises questions concerning disclosure [23], informed consumer judgment, consent, identity rights, accountability, and responsible commercial communication [24]. A detection model can identify potential manipulation, but it cannot independently determine whether the use of AI is ethically justified or legally permissible [25]. These issues therefore require appropriate governance mechanisms that connect technological evidence with organizational responsibility and legal accountability [26]. Within a rule-of-law framework, state regulation is needed to balance technological innovation with consumer protection, individual rights, transparency, and public interest [22].

Accordingly, authenticity represents the central connection among the technological, marketing, and legal dimensions of AI-driven influencer communication [27]. At the consumer level, authenticity influences trust and marketing effectiveness; at the technological level, it requires reliable deepfake detection; and at the governance level, it requires transparency, accountability, and regulatory protection. Based on this integrated perspective, this study combines an Xception-based CNN for deepfake detection, an analysis of AI-driven influencer marketing through AI Personalization, AI Interaction, User Experience, and Trust, and an examination of ethical and legal responsibility under the rule of law. This interdisciplinary framework positions deepfake detection as not merely a technical solution but as an enabling mechanism for trustworthy, transparent, and responsible AI-mediated digital marketing.

METHODS

This study was conducted in Semarang City, Indonesia, in 2026, using a mixed-methods approach with a sequential explanatory design to integrate quantitative and qualitative analyses in evaluating virtual influencer and human influencer marketing campaigns. The quantitative component examined consumer responses toward influencer marketing, particularly Trust, Customer Engagement, Purchase Intention, Perceived Authenticity, and AI Transparency, while the technical component employed an Xception-based Convolutional Neural Network (CNN) to examine content authenticity. The qualitative component

investigated the ethical, legal, and governance implications of AI and deepfake technology. The overall research procedure is presented in Figure 1.

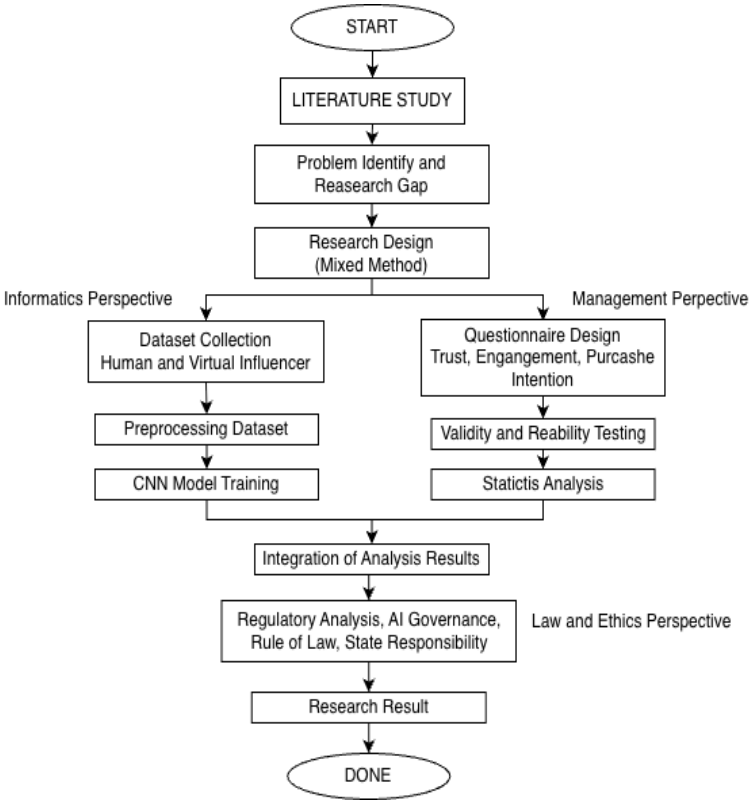

Figure 1. Research flowchart

The CNN dataset consisted of 15.000 images/video frames, comprising 8.500 authentic human-content samples (56.67%) and 6.500 AI-generated/manipulated samples (43.33%). The dataset combined publicly available and privately collected data. Public datasets obtained from Kaggle, Hugging Face, and other publicly available deepfake datasets contributed 90% of the total dataset, equivalent to approximately 13.500 samples, while the private dataset contributed 10%, equivalent to approximately 1.500 samples. The dataset therefore provided two classes, namely authentic human content and AI-generated/manipulated content, with a distribution of 56.67% and 43.33%, respectively.

The research procedure began with data collection followed by preprocessing, including face detection, face alignment, resizing to 299 × 299 pixels, pixel normalization, and data augmentation. The dataset was divided into 80% training and 20% validation data. The training and validation datasets were subsequently used to develop and evaluate the Xception-based CNN model. The deepfake detection model employed an Xception-based Convolutional Neural Network (CNN) because of its ability to extract discriminative spatial features through depthwise separable convolution. The model was trained using the Adam optimizer and Binary Cross-Entropy loss function.

To reduce the risk of overfitting, data augmentation was applied to the training dataset and model performance was monitored using the validation dataset during training. A formal k-fold cross-validation procedure was not employed; therefore, the reported validation performance should be interpreted within the limitations of the single 80:20 train-validation split. Hyperparameter selection was based on validation performance, and the final model was evaluated using accuracy, precision, recall, F1-score, and Area Under the Receiver Operating Characteristic Curve (AUC).

Consumer perception data were collected through a survey of 100 respondents in Semarang City. A purposive sampling technique was employed, with respondents required to be at least 17 years old, actively use social media, and have accessed influencer content within the previous six months. The survey

measured Trust, Customer Engagement, Purchase Intention, Perceived Authenticity, and AI Transparency using a five-point Likert scale. Before hypothesis testing, the measurement instrument was evaluated using indicator loading, Cronbach's Alpha, Composite Reliability, Average Variance Extracted (AVE), and discriminant validity.

The quantitative data were analyzed using Partial Least Squares Structural Equation Modeling (PLS-SEM). The analysis consisted of measurement-model and structural-model assessment. The structural model was evaluated using R^2, Q^2, path coefficients, bootstrapping, and mediation analysis. The total sample consisted of 100 valid respondents, corresponding to the effective observations used in the PLS-SEM analysis. Accordingly, the SSO values reported in the predictive-relevance analysis reflect the number of observations available for the respective constructs and should be interpreted in relation to the 100 respondents included in the survey.

The legal and ethical component involved two legal experts from academic institutions in Semarang. To protect their identity, the experts were anonymized as Expert 1 and Expert 2. Semi-structured interviews focused on four themes: the benefits and risks of AI and deepfake technology, transparency and disclosure, regulatory and accountability gaps, and state responsibility and human-rights protection. The interview data were analyzed using thematic analysis, involving data familiarization, coding of relevant statements, categorization of related codes, and identification of recurring themes. The perspectives of the two experts were then compared to identify convergent findings. The analysis was integrated with a normative juridical approach to examine relevant In
donesian regulations concerning AI, deepfakes, electronic information, personal data protection, consumer protection, and freedom of expression.

RESULT AND DISCUSSION

Performance of the Xception model in deepfake detection

The performance of the proposed Xception-based Convolutional Neural Network (CNN) model was evaluated using standard classification metrics, including accuracy**,** precision**,** recall**,** F1-score**,** and the Area Under the Receiver Operating Characteristic Curve (AUC) in Table 1. Experimental results demonstrate that the proposed model achieved an overall accuracy of 97.12%, indicating excellent capability in distinguishing authentic human-generated content from AI-generated (deepfake) content.

Table 1. Performance evaluation of the proposed Xception-base CNN model

Metric	Human	AI (Deepfake)	Overall
Precision (%)	96.85	97.39	97.12
Recall (%)	97.46	96.78	97.12
F1-Score (%)	97.15	97.08	97.11
Accuracy (%)	-	-	97.12
AUC	-	-	0.992

The classification report in Table 2 further reveals consistently high performance across all evaluation metrics. The Human class (8.500 image) achieved a precision of 96.85%, a recall of 97.46%, and an F1-score of 97.15%, while the AI/deepfake class (6.500 image) obtained a precision of 97.39%, a recall of 96.78% and an F1-score of 97.08%. The macro-average F1-score reached 97.11%, demonstrating balanced classification performance across both classes and confirming the robustness of the proposed model in identifying both authentic and manipulated content.

Table 2. Classification report of the proposed Xception CNN

Class	Precision	Recall	F1-Score	Support
Human (8.500 image)	0.9685	0.9746	0.9715	2,000
AI (6.500 image)	0.9739	0.9678	0.9708	2,000
Accuracy	-	-	0.9712	4,000
Macro Average	0.9712	0.9712	0.9711	4,000
Weighted Average	0.9712	0.9712	0.9712	4,000

The high classification accuracy can be attributed to the effectiveness of the Xception architecture, which employs depthwise separable convolution to efficiently learn discriminative spatial features while reducing computational complexity. This architectural design enables the network to capture subtle manipulation artifacts, including inconsistencies in facial textures, illumination patterns, compression artifacts, and boundary transitions that are often imperceptible to the human eye. Consequently, the proposed model

maintained consistently high performance even when evaluated on datasets containing variations in lighting conditions, facial expressions, head poses, image resolutions, and video compression levels.

The obtained results demonstrate that the proposed approach is highly suitable for practical deepfake detection applications, particularly in digital marketing environments where content authenticity plays a critical role in maintaining consumer trust and communication credibility.

A confusion matrix in Figure 2. analysis was conducted to provide a more comprehensive evaluation of the classification performance. The results indicate that the majority of samples were correctly classified into their respective categories, producing a high proportion of True Positive (TP) and True Negative (TN) predictions while maintaining only a limited number of False Positive (FP) and False Negative (FN) cases. This finding confirms the reliability of the proposed model in verifying the authenticity of digital content prior to subsequent marketing analysis.

Figure 2. Confusion matrix

The False Positive cases primarily occurred when authentic human-generated content exhibited characteristics commonly associated with synthetic media. Factors such as aggressive video compression, motion blur during live-streaming sessions, strong facial beauty filters, inconsistent illumination, and low image resolution introduced visual patterns that closely resembled manipulation artifacts learned during the training process. As a result, a small number of genuine images were incorrectly classified as AI-generated content.

Conversely, False Negative cases were mainly associated with high-quality deepfake content generated using advanced synthesis techniques. These manipulated samples exhibited highly realistic facial textures, consistent lighting, natural facial expressions, and seamless boundary transitions, making them visually indistinguishable from authentic content. The substantial improvements in modern generative AI models have significantly reduced traditional visual artifacts, thereby increasing the difficulty of detecting sophisticated deepfakes using spatial features alone.

Furthermore, the classification performance was further evaluated using the Receiver Operating Characteristic (ROC) curve, as illustrated in Figure 3. The ROC curve provides a comprehensive assessment of the model's discriminative capability by illustrating the trade-off between the True Positive Rate (TPR) and the False Positive Rate (FPR) across different classification thresholds. The proposed Xception-based CNN achieved an Area Under the Curve (AUC) value of 0.9920, indicating excellent discriminative performance in distinguishing authentic content from AI-generated content. An AUC value

approaching 1.0 suggests that the classifier consistently assigns higher confidence scores to the correct class while maintaining a relatively low false positive rate. Moreover, the ROC curve rises steeply toward the upper-left corner of the graph, demonstrating that the model achieves high sensitivity without substantially sacrificing specificity. These findings complement the confusion matrix analysis and confirm that the proposed model is not only accurate at a single decision threshold but also remains robust across a wide range of classification thresholds. Consequently, the high AUC score further validates the suitability of the proposed Xception-based CNN for real-world deepfake detection applications, particularly in digital marketing environments where reliable authenticity verification is essential for supporting consumer trust and transparent AI-driven communication.

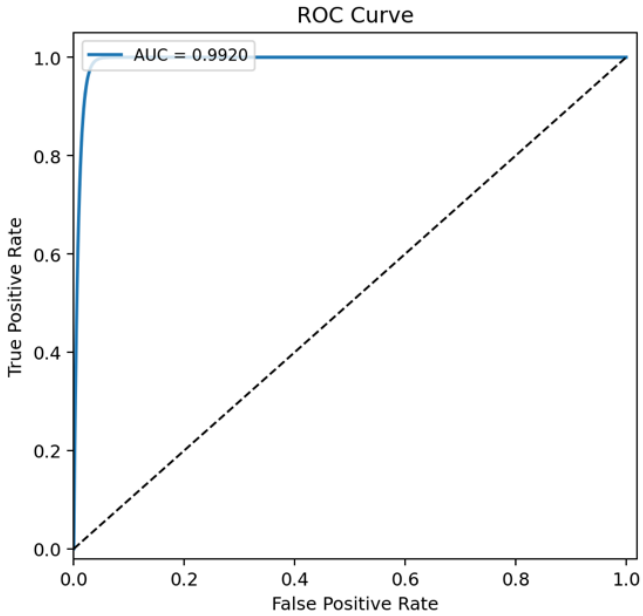

Figure 3. ROC curve

The experimental results demonstrate that the proposed Xception-based CNN achieved a high level of classification performance in distinguishing authentic and AI-generated content. Nevertheless, the practical deployment of deepfake detection systems extends beyond achieving high classification accuracy. In real-world digital marketing environments, influencer content is generated under highly dynamic conditions, including varying camera angles, lighting conditions, video compression, motion blur, image filters, and platform-specific encoding algorithms. These factors introduce additional complexity that may significantly influence model performance.

The confusion matrix analysis revealed that the majority of prediction errors occurred in samples exhibiting substantial visual degradation or highly realistic synthetic characteristics. False Positive predictions were generally associated with authentic live-streaming screenshots containing excessive compression artifacts, aggressive beauty filters, or unstable lighting conditions. These visual characteristics closely resemble synthetic patterns learned during model training, leading the classifier to incorrectly identify genuine content as AI-generated. Conversely, False Negative cases predominantly involved high-quality deepfake content generated using state-of-the-art generative models, where manipulated facial regions exhibited highly consistent illumination, natural facial expressions, and seamless facial boundaries.

Despite these challenging conditions, the relatively low number of misclassified samples indicates that the proposed Xception architecture possesses strong generalization capability across diverse visual environments. This robustness is particularly important because digital marketing campaigns increasingly rely on multimedia content collected from heterogeneous online platforms rather than standardized benchmark datasets. Consequently, the model demonstrates considerable potential for supporting automated authenticity verification in practical marketing applications.

Beyond technical performance, the obtained results also reinforce the importance of integrating deepfake detection into digital marketing workflows. The ability to automatically verify content authenticity before marketing evaluation reduces the risk of analyzing manipulated promotional materials that could bias consumer perception studies. Therefore, the proposed model functions not only as a computer vision classifier but also as a decision-support mechanism that enhances transparency and credibility in AI-driven marketing communication.

To demonstrate the practical applicability of the proposed model, the trained Xception CNN was deployed as a web-based application using the Streamlit framework. The deployment transforms the deep learning model into an interactive graphical user interface (GUI), allowing users to perform real-time authenticity verification of screenshots captured from live-streaming sessions or promotional videos involving both human influencers and virtual influencers.

The developed application provides a user-friendly interface consisting of three primary components: model selection**,** image upload, and prediction visualization. Users begin by uploading a facial image extracted from a live-streaming session in common image formats such as JPG, JPEG, or PNG. After the image is uploaded, the system automatically executes the preprocessing pipeline, including facial image normalization and resizing to the input dimension required by the Xception network. The processed image is then forwarded to the trained CNN model to generate a prediction.

The prediction results are presented through an intuitive visualization interface that displays the predicted class (Human or AI Generated), the associated confidence score, the probability distribution for each class, and the inference time required for classification. In the example shown in Figure 4, the uploaded screenshot was classified as AI Generated with a confidence score of 97.64%, while the remaining probability (2.36%) was assigned to the Human class. Furthermore, the application completed the entire inference process in approximately 557.62 ms, demonstrating its capability to perform near real-time deepfake detection.

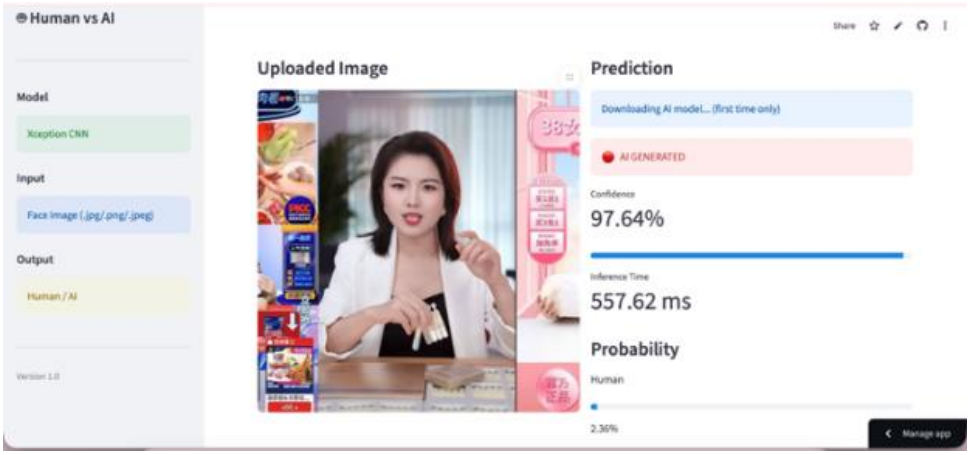

Figure 4. AI generated

In Figure 5, screenshot was classified as Human with a confidence score of 90.25%, because the live streamer using active effects in their live streaming activities, thus destroying some of the natural details in human images such as pores, real eye color and others, it appears as noise. The application completed the entire inference process in approximately 605.80 ms, demonstrating its capability to perform near real-time deepfake detection.

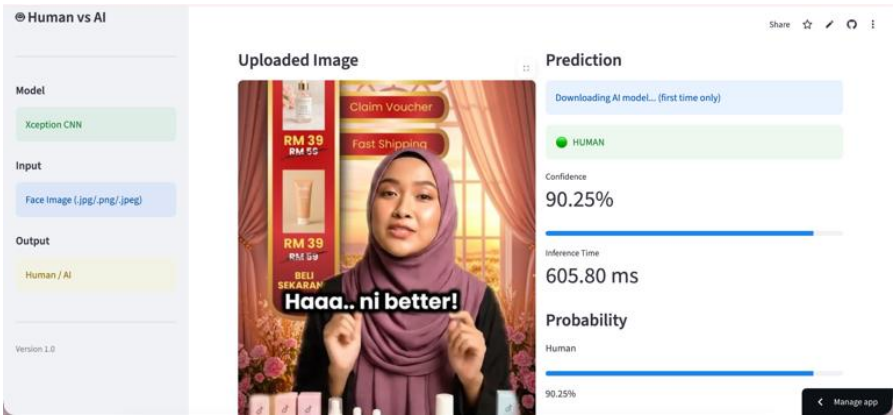

Figure 5. Human

The confidence score serves as an important indicator of prediction certainty. Rather than providing only a binary classification result, the system presents probabilistic outputs that enable users to interpret the reliability of each prediction. This probabilistic representation is particularly valuable in practical scenarios where screenshots exhibit ambiguous visual characteristics or contain partial facial occlusions, allowing human evaluators to make more informed verification decisions.

From an implementation perspective, deploying the model through a web-based GUI substantially enhances its accessibility compared with conventional offline deep learning models. Users without programming expertise can perform authenticity verification directly through a standard web browser without installing specialized machine learning software or configuring computational environments. Such accessibility broadens the potential adoption of the proposed system among researchers, digital marketing practitioners, e-commerce platform operators, and regulatory institutions.

More importantly, the developed application demonstrates that deepfake detection can be seamlessly integrated into digital marketing monitoring workflows. Before evaluating influencer performance or measuring consumer engagement, organizations can verify whether promotional content originates from authentic human influencers or AI-generated virtual influencers. This additional verification stage contributes to greater transparency in digital advertising and reduces the possibility of misleading promotional practices involving undisclosed AI-generated content.

Analysis of the effectiveness of virtual campaigns and human influencers

The structural model was evaluated using Partial Least Squares Structural Equation Modeling (PLS-SEM) to examine the relationships among AI Personalization, AI Interaction, User Experience, and Trust in AI-driven influencer marketing. The analysis consisted of two stages: evaluation of the measurement model and assessment of the structural model. The measurement model was examined through indicator loadings, internal consistency reliability, convergent validity, and discriminant validity, while the structural model was assessed using the coefficient of determination (R^2), predictive relevance (Q^2), path coefficients, bootstrapping results, and mediation analysis.

In Table 3, the measurement model was initially evaluated using indicator loadings. The outer loadings of all indicators exceeded the recommended threshold of 0.70, ranging from 0.858 to 0.927 for the multi-item constructs. These results indicate that the indicators adequately represent their respective constructs and provide an initial indication of convergent validity.

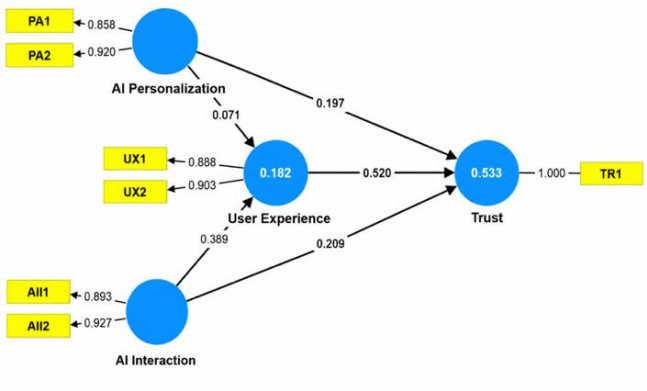

Figure 6. PLS algorithm

Table 3. Outer Loading

Variable	Indicator	Loading Factor	Description
AI Personalization	PA1	0,858	Valid
	PA2	0,920	Valid
AI Interaction	AII1	0,893	Valid
	AII2	0,927	Valid
User Experience	UX1	0,888	Valid
	UX2	0,903	Valid
Trust	TR1	1,000	Valid

However, outer loading alone is insufficient to establish the overall reliability and validity of the measurement model. Therefore, additional assessment was conducted using Cronbach's Alpha, Composite Reliability (CR), Average Variance Extracted (AVE), and discriminant validity.

Table 4. Construct reliability and convergent validity

Variable	Cronbach's Alpha	Composite Reliability (ρa)	Composite Reliability (ρc)	AVE
AI Personalization	0.741	0.779	0.884	0.792
AI Interaction	0.795	0.814	0.906	0.828
User Experience	0.754	0.756	0.890	0.802
Trust	N/A	N/A	N/A	N/A

The results demonstrate satisfactory internal consistency and convergent validity for AI Personalization, AI Interaction, and User Experience. Cronbach's Alpha values range from 0.741 to 0.795, while Composite Reliability (ρc) values range from 0.884 to 0.906, exceeding the recommended threshold of 0.70. The AVE values range from 0.792 to 0.828, substantially exceeding the minimum threshold of 0.50. These findings confirm that the multi-item constructs demonstrate adequate internal consistency and convergent validity.

Trust was operationalized using a single indicator (TR1) with an outer loading of 1.000. Because Trust was specified as a single-item construct, Cronbach's Alpha, Composite Reliability, and AVE are not applicable in the same manner as for multi-item constructs. Therefore, the value of 1.000 should not be interpreted as evidence of perfect reliability, but rather as the standardized loading of the single-indicator construct. This operationalization was intended to capture respondents' overall assessment of trust toward AI-driven influencer communication. Nevertheless, the use of a single-item measure represents a methodological limitation, and future research should employ multiple indicators to provide a more comprehensive measurement of Trust.

Discriminant validity was subsequently assessed using the Heterotrait-Monotrait Ratio (HTMT) and the Fornell-Larcker criterion.

Table 5. HTMT discriminant validity

Construct Pair	HTMT
Trust ↔ AI Interaction	0.574
User Experience ↔ AI Interaction	0.542
User Experience ↔ Trust	0.757
AI Personalization ↔ AI Interaction	0.573
AI Personalization ↔ Trust	0.478
AI Personalization ↔ User Experience	0.336

In Table 5, the HTMT values range from 0.336 to 0.757, with all values below the conservative threshold of 0.85, indicating adequate discriminant validity. The Fornell-Larcker criterion provides additional evidence that the constructs are empirically distinct, as the square root of AVE for each multi-item construct exceeds its correlations with the other constructs. Thus, the measurement model demonstrates satisfactory reliability, convergent validity, and discriminant validity.

The coefficient of determination (R^2) in Table 6 indicates that User Experience achieved an R^2 value of 0.182, implying that 18.2% of its variance is explained by AI Personalization and AI Interaction. This value represents relatively weak explanatory power, suggesting that User Experience is influenced not only by AI-related characteristics but also by other factors that were not incorporated into the present research model. These factors may include interface usability, perceived usefulness, system quality, individual technological literacy, and users' previous experiences with artificial intelligence applications.

Table 6. Coefficient of determination (R2)

Variable	R-square	R-square adjusted
Trust	0,533	0,518
User Experience	0,182	0,165

The R^2 value of 0.533 demonstrates that approximately 53.3% of the variance in consumer Trust is explained by the combined effects of AI Personalization, AI Interaction, and User Experience. This represents a moderate level of explanatory power, suggesting that the proposed research model captures important determinants influencing consumer Trust toward AI-driven influencer communication. The remaining 46.7% of unexplained variance may be associated with external factors not included in the present model, such as perceived authenticity, brand credibility, privacy concerns, perceived risk, or consumers' previous experiences with artificial intelligence technologies.

The strength and direction of the relationships among the latent constructs were subsequently examined through the standardized path coefficients (β). As presented in Table 7, the structural paths demonstrate positive relationships, although their magnitude and statistical significance vary across the proposed relationships.

Table 7. Path coefficient (β)

	AI Interaction	Trust	User Experience	AI Personalization
AI Interaction	-	0,209	0,389	-
Trust	-	-	-	-
User Experience	-	0,520	-	-
AI Personalization	-	0,197	0,071	-

Among the investigated relationships, User Experience has the strongest direct effect on Trust ($\beta = 0.520$), indicating that positive user experiences are the primary determinant of consumer trust within the proposed model. AI Interaction also positively enhances User Experience ($\beta = 0.389$), suggesting that responsive and engaging AI features contribute to improved interaction quality. Moreover, AI Interaction ($\beta = 0.209$) and AI Personalization ($\beta = 0.197$) both positively influence Trust, indicating that interactive and personalized AI characteristics contribute to consumers' confidence in AI-driven influencer communication. In contrast, AI Personalization has only a weak effect on User Experience ($\beta = 0.071$).

To assess the model's predictive capability, the Stone–Geisser predictive relevance (Q^2) was evaluated using the blindfolding procedure, as presented in Table 8. Unlike R^2, which evaluates explanatory power, Q^2 assesses the model's predictive relevance for endogenous constructs.

Table 8. Predictive Relevance (Q2)

	SSO	SSE	Q^2 (=1-SSE/SSO)
AI Interaction	200,000	200,000	-
Trust	100,000	49,519	0,505
User Experience	200,000	178,653	0,107
AI Personalization	200,000	200,000	-

The results show that Trust achieved a Q^2 value of 0.505, while User Experience obtained 0.107, indicating positive predictive relevance for both endogenous constructs. The higher Q^2 value for Trust demonstrates stronger predictive capability, whereas the lower value for User Experience suggests that additional factors beyond the proposed model may contribute to users' experiences.

The proposed hypotheses were then tested using the bootstrapping procedure in Table 9, with significance evaluated based on the standardized path coefficient (β), t-statistic, and p-value.

Table 9. Hypothesis Testing

	Original sample (O)	Sample mean (M)	Standard deviation (STDEV)	T statistics (|O/STDEV|)	P values
AI Interaction -> Trust	0,209	0,208	0,072	2,909	0,004
AI Interaction -> User Experience	0,389	0,385	0,129	3,017	0,003
AI Personalization -> Trust	0,197	0,200	0,075	2,634	0,008
AI Personalization -> User Experience	0,071	0,078	0,113	0,634	0,526
User Experience -> Trust	0,520	0,517	0,065	8,051	0,000

The results indicate that AI Interaction positively affects Trust (β = 0.209; p = 0.004) and User Experience (β = 0.389; p = 0.003), supporting H1 and H2. Likewise, AI Personalization positively influences Trust (β = 0.197; p = 0.008), supporting H3. However, AI Personalization does not significantly affect User Experience (β = 0.071; p = 0.526); therefore, H4 is not supported. Finally, User Experience has the strongest positive effect on Trust (β = 0.520; p < 0.001), confirming that a positive user experience is the primary driver of consumer Trust within the proposed model.

The mediation analysis in Table 10 was conducted to examine whether User Experience mediates the relationships between AI capabilities and consumer Trust.

Table 10. Mediation analysis

	Original sample (O)	Sample mean (M)	Standard deviation (STDEV)	T statistics (|O/STDEV|)	P values
AI Interaction -> User Experience -> Trust	0,202	0,201	0,076	2,653	0,008
AI Personalization -> User Experience -> Trust	0,037	0,040	0,058	0,638	0,524

The results show that User Experience significantly mediates the relationship between AI Interaction and Trust (β = 0.202; p = 0.008). Because the direct effect of AI Interaction on Trust is also significant, the result indicates partial mediation. This finding suggests that AI Interaction strengthens consumer Trust both directly and indirectly through enhanced User Experience. In contrast, User Experience does not significantly mediate the relationship between AI Personalization and Trust (β = 0.037; p = 0.524), indicating that the influence of AI Personalization on Trust primarily occurs through its direct pathway.

Overall, the PLS-SEM findings demonstrate that User Experience is the strongest determinant of Trust, while AI Interaction contributes both directly and indirectly to Trust through User Experience. AI Personalization contributes significantly to Trust but does not significantly affect User Experience. These findings indicate that the effectiveness of AI-mediated influencer communication depends substantially on interaction quality and user experience.

Ethic and responsibility in a state governed by the rule of law

The qualitative data were obtained through interviews with two legal experts, hereafter referred to as Expert 1 and Expert 2, and analyzed using thematic analysis. The analysis consisted of familiarization with the interview data, identification and coding of relevant statements, categorization of related codes, and interpretation of recurring patterns. The perspectives of both experts were subsequently compared to identify convergent findings. Four major themes emerged from the analysis: the benefits and risks of AI, transparency and disclosure, regulatory and accountability gaps, and proportional state intervention and human rights protection.

The first theme concerns the benefits and risks of AI, virtual influencers, and deepfake technology. Expert 1 viewed AI, including virtual influencers and deepfake technology, as a technological instrument that can support human activities, provided that its use remains within the framework of the rule of law and does not replace human responsibility in decision-making. Expert 2 similarly emphasized that AI can provide benefits in education, creative industries, and entertainment, while also creating risks related to fraud, information manipulation, and violations of individual rights. The findings indicate that AI should function as a supporting technology rather than replacing human responsibility in decision-making. Thus, the legal

and ethical concern is not the existence of AI technology itself, but the manner in which it is developed, implemented, and controlled.

The second theme concerns transparency and disclosure in AI-mediated communication. Expert 1 emphasized that the use of AI-generated virtual influencers and deepfake content may undermine honesty and transparency in public communication and digital marketing when consumers are not informed that the content has been generated or manipulated using AI. Consistent with this view, Expert 2 highlighted the importance of disclosure or labeling of AI-generated content to maintain public trust, prevent misleading information, and strengthen consumer protection. Disclosure or labeling was therefore identified as an important mechanism for maintaining transparency in AI-mediated communication. This finding establishes a direct connection between technological authenticity and digital marketing effectiveness, as consumers require sufficient information to assess whether the content and representation presented by an influencer are authentic or synthetically generated.

The third theme relates to regulatory gaps and legal accountability. Expert 1 emphasized that Indonesia does not yet have specific regulations comprehensively governing generative AI, virtual influencers, and deepfake technology. The expert also identified limitations in the existing Electronic Information and Transactions (ITE) framework and Personal Data Protection Law, particularly in addressing emerging forms of AI-generated content and deepfake manipulation. Expert 2 similarly identified a potential regulatory vacuum (rechtvacuum), particularly concerning visual identities, synthetic voices, consent, transparency, and legal accountability. The perspectives of both experts indicate that the existing regulatory framework requires further development to address emerging issues involving synthetic identities, visual likenesses, synthetic voices, AI-generated content, and deepfake manipulation. Accordingly, the findings support the development of specific and adaptive AI regulations that provide legal certainty while addressing technological developments.

The fourth theme concerns state responsibility and proportional intervention under the rule of law. Expert 1 emphasized that the state has a responsibility to establish regulations, conduct supervision, and provide accountability mechanisms for AI misuse, while maintaining technological innovation and freedom of expression. Expert 2 similarly argued that state intervention is justified when necessary to protect the public interest and prevent misuse of AI, provided that human rights remain respected. Both experts therefore emphasized that state intervention must remain proportionate and should balance public protection with technological innovation and freedom of expression. Within the Indonesian constitutional framework, freedom of expression is guaranteed under Article 28E paragraph (3) and Article 28F of the 1945 Constitution, while Article 28J provides the basis for limitations to protect the rights of others, morality, security, and public order. Therefore, AI regulation should not be directed solely toward restricting technology but should establish predictable and proportionate safeguards that protect human rights and the public interest.

The analysis further demonstrates that effective AI governance depends not only on the formulation of legal rules but also on law enforcement and legal culture. Expert 2 emphasized that the effectiveness of AI regulation depends on the quality of law enforcement and the legal culture of society. This perspective is consistent with Lawrence M. Friedman's legal-system theory, in which the effectiveness of regulation requires alignment among legal substance, legal structure, and legal culture. Regulatory provisions concerning AI and deepfakes therefore need to be accompanied by institutional capacity, competent law enforcement, public digital literacy, and ethical awareness. This is particularly important because technological regulation may become ineffective if legal institutions and society lack the capacity to understand and respond to rapidly evolving AI applications.

Overall, the qualitative findings from Expert 1 and Expert 2 demonstrate that technological authenticity, consumer trust, and legal responsibility constitute interconnected dimensions of AI-driven influencer marketing. Deepfake detection provides a technical mechanism for identifying potentially manipulated content, while transparency and disclosure enable consumers to make informed judgments about AI-generated communication. Nevertheless, technical detection alone cannot determine whether the use of AI is ethically or legally appropriate. Therefore, responsible AI-mediated marketing requires the integration of authenticity verification, transparent disclosure, consumer protection, consent, accountability, and proportionate state regulation. This integrated approach enables technological innovation to develop within

the boundaries of the rule of law while preserving public trust, human rights, and the integrity of digital communication.

Research implication

The findings of this study provide several important implications from technological, managerial, and regulatory perspectives regarding the growing adoption of AI-generated virtual influencers in digital marketing. By integrating deepfake detection technology, consumer behavior analysis, and legal assessment within a mixed-methods framework, this research contributes to a more comprehensive understanding of how AI-generated promotional content can be managed responsibly while maintaining consumer trust.

1) Technological implications

From the technological perspective, this study demonstrates that the proposed Xception-based CNN is capable of accurately distinguishing authentic human-generated content from AI-generated content, achieving an overall accuracy of 97.12% and an AUC of 0.9920. These results indicate that deepfake detection systems have reached a level of performance suitable for practical implementation in digital content verification. The deployment of the model through a Streamlit-based web application further illustrates that deepfake detection can be integrated into real-time marketing workflows without requiring specialized technical expertise. Rather than functioning solely as a computer vision classifier, the proposed system can serve as an automated authenticity verification layer before influencer performance is evaluated. This capability reduces the likelihood of biased marketing analysis caused by undisclosed AI-generated promotional content and supports greater transparency in digital communication. Moreover, the findings suggest that future AI-driven marketing platforms may incorporate authenticity verification as a standard component alongside engagement analytics, sentiment analysis, and campaign performance measurement. Such integration would improve the reliability of marketing intelligence systems operating in increasingly AI-generated digital environments.

2) Managerial implications

From the managerial perspective, the structural model demonstrates that User Experience is the strongest determinant of consumer Trust ($\beta = 0.520$), followed by AI Interaction ($\beta = 0.209$) and AI Personalization ($\beta = 0.197$). These findings indicate that organizations should prioritize creating interactive and engaging consumer experiences rather than relying solely on personalization algorithms. For businesses employing virtual influencers, the results imply that technological sophistication alone is insufficient to establish long-term consumer trust. Instead, trust is developed through transparent interactions, responsive communication, and positive user experiences throughout the customer journey. Consequently, marketing managers should carefully balance automation with authenticity by ensuring that AI-generated influencers communicate naturally while clearly disclosing their artificial nature. The study also highlights the importance of integrating deepfake verification into influencer marketing strategies. Companies can utilize authenticity verification before launching campaigns, evaluating influencer effectiveness, or measuring consumer engagement. Such practices reduce reputational risks associated with manipulated promotional content and strengthen brand credibility in increasingly AI-driven markets.

3) Legal and ethical implications

The qualitative findings emphasize that technological innovation should be accompanied by adaptive legal governance. Interviews with constitutional law experts indicate that Indonesia currently faces a regulatory gap regarding artificial intelligence, virtual influencers, and deepfake technology. Existing regulations, including the Electronic Information and Transactions Law and the Personal Data Protection Law, provide only partial legal protection and do not explicitly regulate AI-generated identities or synthetic media. The findings therefore support the need for a dedicated regulatory framework governing AI-generated content. Such regulations should establish clear principles concerning transparency, mandatory disclosure of AI-generated promotional content, accountability for misuse, protection of personal identity, and ethical standards for commercial communication. Implementing mandatory labeling for AI-generated influencer content would enhance consumer awareness while reducing the potential for deception and misinformation. Furthermore, the results reinforce the principle that technological innovation should remain consistent with constitutional values, including legal certainty, consumer protection, proportional state intervention, and respect for fundamental rights. Rather than restricting technological advancement, regulatory frameworks should facilitate responsible AI innovation while minimizing potential social and economic harms.

4) Theoretical implications

From a theoretical perspective, this research extends previous studies on influencer marketing by integrating three disciplines that are rarely examined simultaneously: artificial intelligence, consumer behavior, and legal governance. Previous studies generally investigate virtual influencers from marketing or computer vision perspectives independently. In contrast, this study proposes an interdisciplinary

framework that combines deepfake detection technology, AI-driven consumer trust, and regulatory considerations into a unified research model. The findings also demonstrate that technological authenticity and consumer trust should not be viewed as independent constructs. Instead, accurate authenticity verification contributes indirectly to trustworthy digital marketing ecosystems by ensuring that marketing evaluations are conducted using verified content. This interdisciplinary perspective enriches existing theories of AI adoption, digital trust, and responsible AI governance while providing a foundation for future research on ethical AI implementation in marketing. Overall, this study contributes practical evidence that the future success of AI-generated influencer marketing depends not only on advances in artificial intelligence but also on transparent technological verification, positive consumer experiences, and adaptive legal frameworks that collectively promote trustworthy and responsible digital ecosystems.

CONCLUSION

This study integrates deepfake detection, consumer behavior, and legal perspectives to examine AI-mediated influencer marketing in Semarang City, Indonesia. The Xception-based CNN achieved 97.12% accuracy and an AUC of 0.9920, demonstrating strong capability in distinguishing authentic and AI-generated content. The PLS-SEM results show that User Experience is the strongest determinant of Trust ($\beta = 0.520$; $p < 0.001$), while AI Interaction significantly influences User Experience and Trust. However, this study does not establish a direct statistical difference between virtual and human influencers, as no multi-group comparison was performed. The study is limited to a single-city context, which may restrict the external validity and generalizability of the findings to other populations and cultural contexts. The involvement of only two legal experts ($n = 2$), the relatively weak explanatory power of User Experience ($R^2 = 0.182$), and the potential limited generalizability of the CNN to unseen deepfake generation techniques further constrain the study. Future research should employ multi-group SEM to directly compare virtual and human influencers, expand sampling beyond Semarang, and conduct cross-dataset and unseen-technique validation of the CNN. These findings emphasize that technological authenticity, consumer trust, and adaptive AI governance should be developed together to support responsible and sustainable digital marketing.

REFERENCES

[1] L. Zhang and C. Hur, "The Impact of Generative AI Images on Consumer Attitudes in Advertising," *Adm. Sci.*, vol. 15, no. 10, p. 395, Oct. 2025, doi: 10.3390/admsci15100395.

[2] J. Landgrebe, "Rise of Digital Influencer Marketing," *HERMES - Journal of Language and Communication in Business*, no. 64, pp. 75–102, Sep. 2024, doi: 10.7146/hjlcb.vi64.143879.

[3] J. Arsenyan and A. Mirowska, "Almost human? A comparative case study on the social media presence of virtual influencers," *Int. J. Hum. Comput. Stud.*, vol. 155, p. 102694, Nov. 2021, doi: 10.1016/j.ijhcs.2021.102694.

[4] G. Zafar *et al.*, "IMPACT OF DEEP-FAKE ADVERTISING DISCLOSURE ON PURCHASE INTENTION WITH MEDIATING ROLES OF PERCEIVED REALITY, TRUST, PERCEIVED ETHICALITY, AND IRRITATION," vol. 3, p. 2025, 2025, doi: 10.5281/zenodo.15266445.

[5] S. L. T. Mouritzen, V. Penttinen, and S. Pedersen, "Virtual influencer marketing: the good, the bad and the unreal," *Eur. J. Mark.*, vol. 58, no. 2, pp. 410–440, Feb. 2024, doi: 10.1108/EJM-12-2022-0915.

[6] Y. Diao, M. Liang, C. Jin, and H. Woo, "Virtual Influencers and Sustainable Brand Relationships: Understanding Consumer Commitment and Behavioral Intentions in Digital Marketing for Environmental Stewardship," *Sustainability*, vol. 17, no. 13, p. 6187, Jul. 2025, doi: 10.3390/su17136187.

[7] L. Zhang, L. Mo, X. Sun, Z. Zhou, and J. Ren, "How Visual and Mental Human-Likeness of Virtual Influencers Affects Customer–Brand Relationship on E-Commerce Platform," *Journal of Theoretical and Applied Electronic Commerce Research*, vol. 20, no. 3, p. 200, Aug. 2025, doi: 10.3390/jtaer20030200.

[8] I. Kim, C.-W. Ki, H. Lee, and Y.-K. Kim, "Virtual influencer marketing: Evaluating the influence of virtual influencers' form realism and behavioral realism on consumer ambivalence and marketing performance," *J. Bus. Res.*, vol. 176, p. 114611, Apr. 2024, doi: 10.1016/j.jbusres.2024.114611.

[9] J. Peemanee, T. Udomlarp, P. Weber, and R. Weerarathna, "The Role of Virtual and Human Influencer Characteristics in Shaping Gen Z Purchases on TikTok: Hybrid SEM-ANN Approach," *Journal of Theoretical and Applied Electronic Commerce Research*, vol. 21, no. 5, p. 150, May 2026, doi: 10.3390/jtaer21050150.

[10] M. Gerlich, "The Power of Virtual Influencers: Impact on Consumer Behaviour and Attitudes in the Age of AI," *Adm. Sci.*, vol. 13, no. 8, p. 178, Aug. 2023, doi: 10.3390/admsci13080178.

[11] M.-H. Wu, "From Influencer Credibility to E-Loyalty Intentions in Social Commerce: Digital Promotional Signals, Brand Authenticity, and the Trust–Engagement Pathway," *Journal of Theoretical and Applied Electronic Commerce Research*, vol. 21, no. 7, p. 204, Jun. 2026, doi: 10.3390/jtaer21070204.

[12] J. Yi, C. Wang, J. Tao, X. Zhang, C. Y. Zhang, and Y. Zhao, "Audio Deepfake Detection: A Survey," Aug. 2023.

[13] Y. Vyas and P. Vishwakarma, "Artificial Intelligence in Brand Communication: A Comprehensive Literature Review and Research Outlook," *Int. J. Consum. Stud.*, vol. 50, no. 1, Jan. 2026, doi: 10.1111/ijcs.70170.

[14] A. R. Ananda, M. N. T. Putra, and M. I. Y. Azhar, "Virtual Influencers, Real Impact: A Narrative Review on Credibility, Generational Trust, and Purchase Intention in Digital Marketing," *Sinergi International Journal of Communication Sciences*, vol. 3, no. 2, pp. 99–112, Feb. 2025, doi: 10.61194/ijcs.v3i2.690.

[15] V. L. L. Thing, "Deepfake Detection with Deep Learning: Convolutional Neural Networks versus Transformers," Apr. 2023.

[16] S. H. Setiyani, E. Noersasongko, and A. Affandy, "Classification of Breast Cancer Histopathology Images with Attention-Based Multiple Instance Learning Method," *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, Oct. 2025, doi: 10.22219/kinetik.v10i4.2310.

[17] R. I. Bendjillali, M. Beladgham, and K. Merit, "Face recognition based on DWT feature for CNN," in *ACM International Conference Proceeding Series*, Association for Computing Machinery, Mar. 2019. doi: 10.1145/3361570.3361584.

[18] F. A. Twince Tobing, A. Kusnadi, I. Z. Pane, and R. Winantyo, "Deepfake detection using convolutional neural networks: a deep learning approach for digital security," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 39, no. 2, p. 1092, Aug. 2025, doi: 10.11591/ijeecs.v39.i2.pp1092-1099.

[19] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017, doi: 10.1109/TPAMI.2016.2577031.

[20] T. T. A. NGO, K. B. NGUYEN, M. T. PHAM, N. T. TRAN, H. T. TRAN, and C. T. DAO, "The influence of online customer reviews (OCRs) on the online-shopping intention for domestic fashion: A case study of Vietnam," *Acta Psychol. (Amst).*, vol. 260, p. 105639, Oct. 2025, doi: 10.1016/j.actpsy.2025.105639.

[21] A. Raza *et al.*, "A Comprehensive Review of Deepfake Detection Techniques: From Traditional Machine Learning to Advanced Deep Learning Architectures," *AI*, vol. 7, no. 2, p. 68, Feb. 2026, doi: 10.3390/ai7020068.

[22] A. H. H. Hasibuan, David Novan Setyawan, and Yanuriansyah Ar Rasyid, "Legal Protection For Consumers Against The Potential Harm Caused By Promotions Using Deepfakes," *JIHK*, vol. 6, no. 2, pp. 116–126, Jan. 2025, doi: 10.46924/jihk.v6i2.232.

[23] E. Yıldırım and F. Dursun, "Disclosure Matters: Perceived Manipulation, Perceived Ethics, and Purchase Intention Toward AI Influencers in Social Media Marketing," *Journal of Theoretical and Applied Electronic Commerce Research*, vol. 21, no. 6, p. 194, Jun. 2026, doi: 10.3390/jtaer21060194.

[24] X. Qiu, Y. Wang, Y. Zeng, and R. Cong, "Artificial Intelligence Disclosure in Cause-Related Marketing: A Persuasion Knowledge Perspective," *Journal of Theoretical and Applied Electronic Commerce Research*, vol. 20, no. 3, p. 193, Aug. 2025, doi: 10.3390/jtaer20030193.

[25] P. Das Deep, W. D. Edgington, N. Ghosh, and Md. S. Rahaman, "Evaluating the Effectiveness and Ethical Implications of AI Detection Tools in Higher Education," *Information*, vol. 16, no. 10, p. 905, Oct. 2025, doi: 10.3390/info16100905.

[26] N. Nahum, U. Larsson-Olaison, T. Uman, and L. Achtenhagen, "Corporate governance for digital transformation: The role of ownership and the board of directors," *Technol. Forecast. Soc. Change*, vol. 223, p. 124453, Feb. 2026, doi: 10.1016/j.techfore.2025.124453.

[27] E. P. Boediman, "Exploring the impact of deepfake technology on public trust and media manipulation: A scoping review," *Jurnal Komunikasi*, vol. 19, no. 2, pp. 313–334, Apr. 2025, doi: 10.20885/komunikasi.vol19.iss2.art8.