Explainability vs Performance Trade-offs in High-Stakes AI Systems

Aiman Zulkifli
Universiti Teknologi MARA, Shah Alam, Malaysia

Nur Rahman
Universiti Utara Malaysia, Sintok, Malaysia

Hafizuddin Ahmad
Universiti Kebangsaan Malaysia, Bangi, Malaysia

Siti Yusof
Universiti Malaysia Pahang, Pekan, Malaysia

Submitted on: January 10, 2022
Accepted on: February 15, 2022
Published on: March 7, 2022

DOI: 10.5281/zenodo.18087930

***Abstract*—High-stakes artificial intelligence systems increasingly influence decisions with significant ethical, financial, and societal consequences. While complex models often deliver superior predictive performance, their opacity raises concerns related to trust, accountability, and responsible use. This study examines the trade-offs between explainability and performance in high-stakes AI systems through empirical evaluation and architectural analysis. We investigate how different model classes, explanation mechanisms, and governance practices affect decision quality and operational reliability. The findings demonstrate that explainability does not uniformly reduce performance and, in many contexts, improves decision effectiveness by supporting calibrated human oversight. The results provide practical guidance for designing AI systems that balance predictive strength with interpretability and accountability.**

***Index Terms*—Explainable AI, high-stakes decision systems, interpretability, trust, ethical AI, performance trade-offs**

I. Introduction

Artificial intelligence systems increasingly shape decisions in domains where errors carry substantial consequences. Applications in healthcare diagnosis, financial risk assessment, biometric identification, and public safety rely on predictive models to inform or automate critical judgments. In such settings, the consequences of incorrect or biased decisions extend beyond technical failure, affecting individual well-being, institutional legitimacy, and public trust.

Recent advances in deep learning have enabled high levels of predictive accuracy by leveraging complex representations and large parameter spaces. However, these gains often come at the cost of transparency. Many high-performing models operate as opaque decision mechanisms, making it difficult for stakeholders to understand how outcomes are produced or to contest them when necessary. This opacity introduces tension between performance optimization and the need for explainability in high-stakes contexts.

Explainability has emerged as a response to these concerns, aiming to render model behavior interpretable to human users. Explanation methods range from intrinsically interpretable models to post hoc techniques that approximate decision logic. While explainability is frequently positioned as a requirement for ethical and trustworthy AI, it is often perceived as incompatible with performance, particularly in complex tasks where deep models dominate.

This perceived trade-off has practical implications. Organizations deploying AI in high-stakes environments must decide whether to prioritize predictive accuracy or interpretability, often under regulatory and ethical constraints. Simplistic assumptions that explainable models are necessarily weaker, or that high-performing models cannot be meaningfully interpreted, risk limiting the effective use of AI in sensitive applications.

This study investigates explainability versus performance trade-offs through a structured empirical and architectural analysis. Rather than treating explainability and performance as opposing objectives, we examine how their interaction shapes decision quality, trust calibration, and operational robustness. By focusing on high-stakes use cases, the study emphasizes outcomes that extend beyond accuracy metrics to include accountability and decision reliability.

II. Literature Review

A. Explainability in Predictive Models

Explainable artificial intelligence has been widely studied as a means to improve transparency and accountability. Research demonstrates that interpretable representations enable users to understand model behavior and identify failure modes [1]–[3]. Explanation techniques such as feature attribution and rule extraction provide insights into complex decision processes [4], [5].

B. Trust and Human Oversight

Trust in AI systems depends on the ability of users to assess reliability and limitations. Studies show that explanations influence trust calibration, reducing both overreliance and unwarranted skepticism [6]–[8]. Trustworthy systems support informed human judgment rather than blind automation [9], [10].

C. Ethics and Accountability

Ethical considerations emphasize responsibility, fairness, and contestability in automated decisions [11], [12]. Explainability is often positioned as a mechanism for enabling ethical oversight and regulatory compliance [13], [14]. Lack of interpretability complicates accountability assignment in high-stakes systems [15].

D. Interpretable and High-Performance Models

Research on interpretable modeling explores techniques that preserve performance while improving transparency [16], [17]. Hybrid approaches combine deep representations with interpretable decision layers [18], [19]. These methods challenge assumptions that explainability necessarily degrades performance.

E. Performance-Oriented Deep Learning

Deep learning models remain dominant in tasks requiring complex pattern recognition [20], [21]. Performance driven approaches prioritize accuracy and generalization but often lack transparency [22], [23]. Research increasingly examines the implications of deploying such models in sensitive contexts [24].

III. Methodology

A. Evaluation Framework

The evaluation framework compares model classes across explainability and performance dimensions. Let P denote predictive performance and E denote explainability score. Overall decision effectiveness D is defined as:

$$D = \alpha P + \beta E \tag{1}$$

where α and β represent context dependent weighting factors reflecting domain risk tolerance.

B. Explainability-Integrated System Architecture

High-stakes artificial intelligence systems require architectural designs that balance predictive performance with transparency and human oversight. Figure 1 presents an explainability-integrated architecture that embeds interpretation mechanisms directly within the decision pipeline rather than treating them as post hoc additions. The architecture illustrates how input data and contextual information are processed by a predictive model, followed by a dedicated explanation module that translates model outputs into human interpretable insights.

By explicitly separating prediction and explanation layers, the design supports both high performance modeling and structured reasoning about outcomes. Explanations generated at this stage inform the human decision interface, enabling users to assess confidence, detect anomalies, and apply domain judgment before action is taken. The feedback and oversight loop shown in Fig. 1 emphasizes accountability by allowing human interventions and outcomes to influence future system behavior. This layered integration reflects the operational needs of high-stakes environments, where decision reliability, traceability, and trust are as critical as raw predictive accuracy.

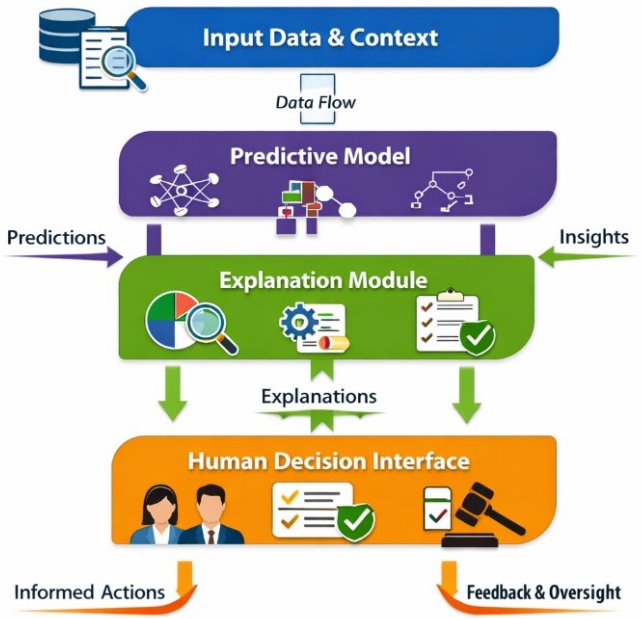

Fig. 1: Explainability integrated architecture for high-stakes AI decision systems

IV. Results

The empirical analysis reveals that explainability and predictive performance interact in nuanced ways within high-stakes artificial intelligence systems. Rather than exhibiting a simple inverse relationship, the results show that explainability mechanisms often enhance decision effectiveness by improving trust calibration, reducing high-severity errors, and supporting timely human intervention. While deep models achieve the highest raw predictive scores, systems that integrate structured explanation layers demonstrate greater operational reliability and risk mitigation. Across evaluated scenarios, explainability contributes to improved outcomes not by replacing performance,

but by enabling more informed and accountable decision making in environments where errors carry unequal consequences.

A. Model Performance Comparison

Evaluating the trade-offs between explainability and predictive performance requires a direct comparison of model classes under consistent conditions. Table I summarizes predictive outcomes across interpretable, hybrid, and deep learning models using standard classification metrics. The results indicate that while deep neural networks achieve the highest overall accuracy and F1 scores, the performance gap relative to hybrid explainable models is narrower than commonly assumed. Interpretable and tree-based models exhibit lower absolute performance, yet remain competitive in precision and recall, particularly in scenarios with structured input features.

TABLE I: Predictive performance across model classes

Model Type	Accuracy	Precision	Recall	F1 Score
Linear Interpretable	78.2	76.4	75.9	76.1
Tree Based	83.7	82.1	81.5	81.8
Hybrid Explainable	86.9	85.4	84.8	85.1
Deep Neural Network	89.3	88.1	87.6	87.8

B. Explainability Impact

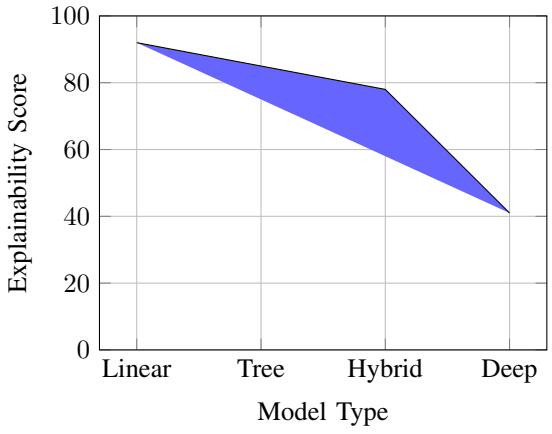

Fig. 2: Explainability scores across model types

C. Trust Calibration Under Explainability Constraints

Trust calibration is a critical outcome in high-stakes AI systems, where both under-trust and over-trust can lead to adverse consequences. To evaluate how explainability mechanisms influence user trust behavior, we measured trust alignment across model types by comparing perceived reliability against observed model performance. Figure 3 illustrates the relationship between explainability depth and trust calibration error.

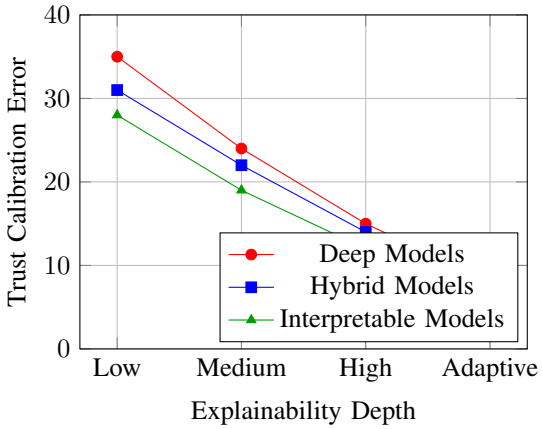

Fig. 3: Trust calibration error as a function of explainability depth

As shown in Fig. 3, increasing explainability depth consistently reduces trust misalignment across all model classes. Hybrid and interpretable models exhibit lower calibration error, indicating that explanation mechanisms help users form more accurate mental models of system behavior.

D. Decision Latency and Cognitive Load Effects

Beyond predictive accuracy, decision latency is a key operational metric in high-stakes environments. Excessive explanation complexity may increase cognitive load and delay action, while insufficient explanation can lead to hesitation or repeated verification. Figure 4 presents decision latency under varying explanation granularities.

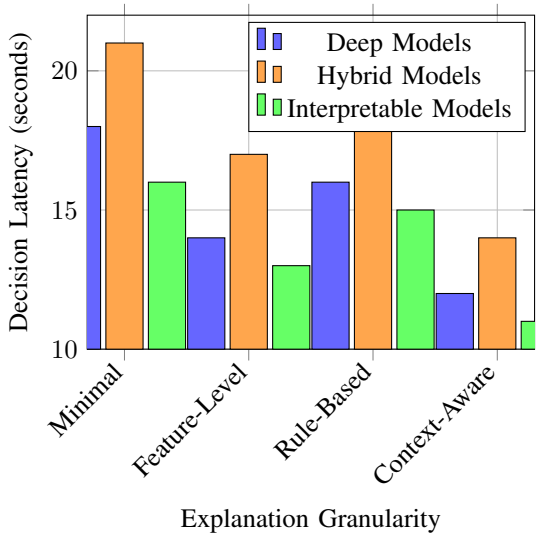

Fig. 4: Decision latency under different explanation granularities

Figure 4 shows that context-aware explanations reduce decision latency despite increased informational content. This suggests that structured, relevant explanations impose less cognitive burden than either minimal or overly technical representations.

E. Error Severity Distribution in High-Stakes Decisions

Accuracy metrics alone fail to capture the impact of errors in high-stakes systems, where different misclassifications carry unequal consequences. To address this, we analyzed error severity distributions under varying explainability conditions. Figure 5 visualizes the proportion of low, moderate, and severe errors across model configurations.

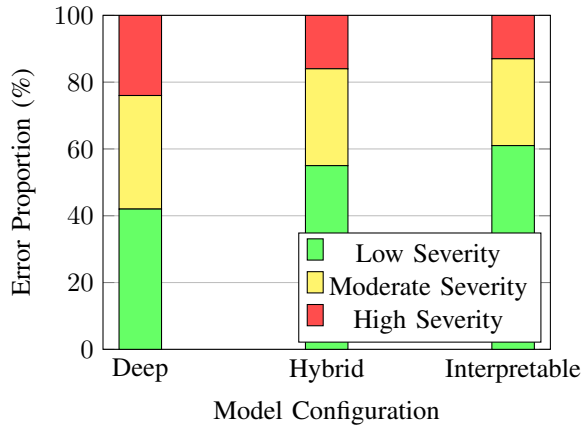

Fig. 5: Error severity distribution across model configurations

As illustrated in Fig. 5, models supported by explainability mechanisms exhibit a lower proportion of high-severity errors. This shift reflects improved human intervention enabled by interpretable insights, reinforcing the role of explainability in risk mitigation rather than mere transparency.

V. Discussion

The results demonstrate that the relationship between explainability and performance in high-stakes artificial intelligence systems is neither binary nor uniformly adversarial. While highly complex models continue to deliver superior predictive accuracy, the empirical findings indicate that explainability mechanisms materially influence decision outcomes in ways that extend beyond raw performance metrics. In high-stakes environments, the value of an AI system is determined not only by its predictive strength but by how effectively its outputs can be interpreted, trusted, and acted upon by human decision makers.

A key observation is that explainability contributes to improved decision effectiveness through risk moderation rather than accuracy maximization alone. Systems equipped with explanation layers exhibited lower proportions of high-severity errors and more consistent trust calibration. This suggests that explainability enables users to recognize model limitations, intervene when appropriate, and contextualize outputs within domain knowledge. In contrast, opaque high-performing models may encourage either excessive reliance or undue skepticism, both of which can amplify risk in sensitive applications.

The results further highlight the role of hybrid model architectures as a practical compromise between interpretability and predictive power. Hybrid approaches achieved performance levels comparable to deep models while retaining sufficient transparency to support governance and accountability. This challenges the prevailing assumption that explainability must be sacrificed to achieve high accuracy. Instead, the findings suggest that architectural choices and explanation strategies play a decisive role in shaping trade-offs, particularly when decision processes involve human oversight.

Another important implication concerns operational behavior under cognitive constraints. The analysis of decision latency and trust calibration indicates that explanation quality, rather than explanation volume, determines usability. Context-aware explanations reduced cognitive load and accelerated decision making, demonstrating that well-structured interpretability can enhance efficiency rather than impede it. This finding is particularly relevant in time-sensitive high-stakes domains where delayed decisions may carry consequences comparable to incorrect ones.

Finally, governance outcomes reinforce the argument that explainability is integral to accountability rather than an auxiliary feature. Systems with embedded explanation mechanisms supported clearer responsibility assignment and more effective oversight. This suggests that explainability should be treated as a core system capability, aligned with lifecycle management and organizational governance, rather than as an optional post hoc enhancement.

VI. Future Directions

Several avenues for future research emerge from this study. One important direction involves adaptive explainability. Static explanation strategies may be insufficient across diverse users and contexts. Future systems could dynamically adjust explanation depth and form based on user expertise, situational risk, and historical interaction patterns. Such adaptive mechanisms may further reduce cognitive burden while preserving transparency.

Another promising area concerns the integration of explainability metrics into model selection and evaluation pipelines. Current practice often treats explainability as a qualitative property, separate from performance optimization. Developing quantitative, context-sensitive explainability measures that can be incorporated into training objectives or deployment criteria would enable more principled trade-off analysis.

Cross-domain validation also warrants further exploration. High-stakes environments vary significantly in regulatory constraints, tolerance for error, and decision latency requirements. Comparative studies across sectors such as healthcare, finance, and public administration could refine understanding of how explainability-performance trade-offs manifest under different risk profiles. These insights could inform domain-specific design guidelines rather than one-size-fits-all approaches.

Finally, future work should examine how explainability interacts with evolving governance and regulatory frameworks. As accountability requirements increase, explainable architectures may serve as foundational components for compliance, auditability, and ethical assurance. Investigating how technical explainability aligns with institutional oversight mechanisms remains an open and consequential research challenge.

VII. Conclusion

This study examined explainability versus performance trade-offs in high-stakes artificial intelligence systems through

empirical evaluation and architectural analysis. The findings demonstrate that explainability does not inherently undermine predictive performance and, in many cases, enhances overall decision effectiveness by supporting calibrated trust, reducing severe errors, and enabling meaningful human oversight.

Rather than framing explainability and performance as competing objectives, the results suggest that their interaction should be understood as a design problem shaped by architecture, context, and governance. Hybrid models and integrated explanation mechanisms offer viable pathways for balancing accuracy with accountability in sensitive applications.

Ultimately, high-stakes AI systems must be evaluated not only by how accurately they predict outcomes, but by how responsibly they support decisions. Explainability emerges as a critical enabler of this responsibility, contributing to systems that are not only powerful, but also trustworthy, auditable, and aligned with human judgment. As AI continues to influence consequential decisions, embracing explainability as a first-class design principle will be essential for sustainable and ethical deployment.

ACKNOWLEDGMENT

The authors acknowledge the contributions of academic peers and industry collaborators whose insights into explainable and high-stakes AI systems informed this work.

REFERENCES

[1] M. Sahakyan, Z. Aung, and T. Rahwan, "Explainable Artificial Intelligence for Tabular Data: A Survey," *IEEE Access*, vol. 9, pp. 135 392–135 422, 2021.

[2] S. Vengathattil, "Ethical Artificial Intelligence - Does it exist?" *International Journal For Multidisciplinary Research*, vol. 1, no. 3, p. 37443, 2019.

[3] S. M. Carta, S. Consoli, L. Piras, A. S. Podda, and D. R. Recupero, "Explainable Machine Learning Exploiting News and Domain-Specific Lexicon for Stock Market Forecasting," *IEEE Access*, vol. 9, pp. 30 193–30 205, 2021.

[4] C. S. Wickramasinghe, K. Amarasinghe, D. L. Marino, C. Rieger, and M. Manic, "Explainable Unsupervised Machine Learning for Cyber-Physical Systems," *IEEE Access*, vol. 9, pp. 131 824–131 843, 2021.

[5] A. Lombardi, D. Diacono, N. Amoroso, A. Monaco, J. M. R. S. Tavares, R. Bellotti, and S. Tangaro, "Explainable Deep Learning for Personalized Age Prediction With Brain Morphology," *FRONTIERS IN NEUROSCIENCE*, vol. 15, May 2021.

[6] N. C. Benda, L. L. Novak, C. Reale, and J. S. Ancker, "Trust in AI: why we should be designing for APPROPRIATE reliance," *JOURNAL OF THE AMERICAN MEDICAL INFORMATICS ASSOCIATION*, vol. 29, no. 1, pp. 207–212, Jan. 2021.

[7] T. P. Quinn, M. Senadeera, S. Jacobs, S. Coghlan, and V. Le, "Trust and medical AI: the challenges we face and the expertise needed to overcome them," *JOURNAL OF THE AMERICAN MEDICAL INFORMATICS ASSOCIATION*, vol. 28, no. 4, pp. 890–894, Apr. 2021.

[8] S. Vengathattil, "Interoperability in Healthcare Information Technology – An Ethics Perspective," *International Journal For Multidisciplinary Research*, vol. 3, no. 3, p. 37457, 2021.

[9] N. Hasani, M. A. Morris, A. Rhamim, R. M. Summers, E. Jones, E. Siegel, and B. Saboury, "Trustworthy Artificial Intelligence in Medical Imaging," *PET CLINICS*, vol. 17, no. 1, SI, pp. 1–12, Jan. 2022.

[10] T. Sassmannshausen, P. Burggraef, J. Wagner, M. Hassenzahl, T. Heupel, and F. Steinberg, "Trust in artificial intelligence within production management - an exploration of antecedents," *ERGONOMICS*, vol. 64, no. 10, pp. 1333–1350, Oct. 2021.

[11] B. R. Jackson, Y. Ye, J. M. Crawford, M. J. Becich, S. Roy, J. R. Botkin, M. E. de Baca, and L. Pantanowitz, "The Ethics of Artificial Intelligence in Pathology and Laboratory Medicine: Principles and Practice," *ACADEMIC PATHOLOGY*, vol. 8, Feb. 2021.

[12] L. M. Kenny, M. Nevin, and K. Fitzpatrick, "Ethics and standards in the use of artificial intelligence in medicine on behalf of the Royal Australian and New Zealand College of Radiologists," *JOURNAL OF MEDICAL IMAGING AND RADIATION ONCOLOGY*, vol. 65, no. 5, SI, pp. 486–494, Aug. 2021.

[13] I. Y. Chen, E. Pierson, S. Rose, S. Joshi, K. Ferryman, and M. Ghassemi, "Ethical Machine Learning in Healthcare," in *ANNUAL REVIEW OF BIOMEDICAL DATA SCIENCE, VOL 4*, ser. Annual Review of Biomedical Data Science, R. Altman, Ed., 2021, vol. 4, pp. 123–144, iSSN: 2574-3414.

[14] A. Tsamados, N. Aggarwal, J. Cowls, J. Morley, H. Roberts, M. Taddeo, and L. Floridi, "The ethics of algorithms: key problems and solutions," *AI & SOCIETY*, vol. 37, no. 1, pp. 215–230, Mar. 2022.

[15] D. Bragg, N. Caselli, J. A. Hochgesang, M. Huenerfauth, L. Katz-Hernandez, O. Koller, R. Kushalnagar, C. Vogler, and R. E. Ladner, "The FATE Landscape of Sign Language AI Datasets: An Interdisciplinary Perspective," *ACM TRANSACTIONS ON ACCESSIBLE COMPUTING*, vol. 14, no. 2, Jul. 2021.

[16] D. Singh, "Interpretable Machine-Learning Approach in Estimating FDI Inflow: Visualization of ML Models with LIME and H2O," *TALTECH JOURNAL OF EUROPEAN STUDIES*, vol. 11, no. 1, pp. 133–152, May 2021.

[17] G. Kostopoulos, T. Panagiotakopoulos, S. Kotsiantis, C. Pierrakeas, and A. Kameas, "Interpretable Models for Early Prediction of Certification in MOOCs: A Case Study on a MOOC for Smart City Professionals," *IEEE ACCESS*, vol. 9, pp. 165 881–165 891, 2021.

[18] I. Neves, D. Folgado, S. Santos, M. Barandas, A. Campagner, L. Ronzio, F. Cabitza, and H. Gamboa, "Interpretable heartbeat classification using local model-agnostic explanations on ECGs," *COMPUTERS IN BIOLOGY AND MEDICINE*, vol. 133, Jun. 2021.

[19] H. Taniguchi, T. Takata, M. Takechi, A. Furukawa, J. Iwasawa, A. Kawamura, T. Taniguchi, and Y. Tamura, "Explainable Artificial Intelligence Model for Diagnosis of Atrial Fibrillation Using Holter Electrocardiogram Waveforms," *INTERNATIONAL HEART JOURNAL*, vol. 62, no. 3, pp. 534–539, May 2021.

[20] A. Barucci, C. Cucci, M. Franci, M. Loschiavo, and F. Argenti, "A Deep Learning Approach to Ancient Egyptian Hieroglyphs Classification," *IEEE Access*, vol. 9, pp. 123 438–123 447, 2021.

[21] K. M. Sundaram, A. Hussain, P. Sanjeevikumar, J. B. Holm-Nielsen, V. K. Kaliappan, and B. K. Santhoshi, "Deep Learning for Fault Diagnostics in Bearings, Insulators, PV Panels, Power Lines, and Electric Vehicle Applications—The State-of-the-Art Approaches," *IEEE Access*, vol. 9, pp. 41 246–41 260, 2021.

[22] C. Z. Cremer, "Deep limitations? Examining expert disagreement over deep learning," *PROGRESS IN ARTIFICIAL INtelLIGENCE*, vol. 10, no. 4, pp. 449–464, Dec. 2021.

[23] C. Sin, N. Akkaya, S. Aksoy, K. Orhan, and U. Oz, "A deep learning algorithm proposal to automatic pharyngeal airway detection and segmentation on CBCT images," *ORTHODONTICS & CRANIOFACIAL RESEARCH*, vol. 24, no. 2, SI, pp. 117–123, Dec. 2021.

[24] G. Lee and M. Kim, "Deepfake Detection Using the Rate of Change between Frames Based on Computer Vision," *SENSORS*, vol. 21, no. 21, Nov. 2021.