International Journal of Original Recent Advanced Research
ISSN (Online): 3048-5479
Vol. 03, Issue 03, May 2026

MULTIMODAL ARTIFICIAL INTELLIGENCE FOR ENHANCING DIGITAL WELLBEING AND ONLINE SAFETY OF CHILDREN

Narada Dilshan Fernando Gardige

Department of Artificial Intelligence & Cloud Computing, Washington Digital University, USA

Enrollment No.: WDU2026293148

ABSTRACT

The pervasive presence of digital technologies in children's life has led to a sharp increase in the dangers faced by children (including cyberbullying, exposure to exploitation with undesirable content, grooming online, addiction and violations of privacy). Purely text- or image-based approaches to online safety have been largely inadequate to contain the complexity of these issues. This work reports the development; testing and empirical validation of a multimodal artificial intelligence (AI) framework to improve children aged 6–17 digital wellbeing and online safety. We introduce a novel late fusion multimodal architecture, and combine NLP, Computer vision and audio analysis to produce a score of 94.7% F1-score overall on the test dataset consisting of 42.8 K annotated instances. A mixed-methods approach: structured parental surveys (n = 412), analyses of behavioral logs, content identified over multiple platforms was used to collect data. Multimodal fusion is shown to provide a strong improvement over all baseline unimodal baselines in predicting the threat across all categories. The study found that 70.0% of parents would accept an AI-based safety tool; however, privacy concerns represented a substantive barrier to adoption. Analysis of relationships again demonstrated a strong negative relationship between AI based online risk indicators and children's subjective wellbeing scores.

Keywords: Multimodal AI[1], Digital Wellbeing[2], Child Online Safety[3], Cyberbullying Detection[4], Deep Learning[4], Natural Language Processing[5], Computer Vision[6].

1. INTRODUCTION

1.1 Background and Significance

The widespread availability of internet-connected devices and digital platforms is changing the very ecosystem in which childhood and adolescence play out, both for good and for ill. Children have the difficult and unique

task of negotiating a complex digital ecology that affords educational opportunity yet also presents serious psychosocial challenge. UNICEF (2017) estimates that around 1 in every three of all internet users worldwide are children under the age of 18 and this share is further increasing due to growing mobile penetration in developing countries [10]. In India itself, it is reported that children are using digital devices for four to six hours a day a number which rose sharply in the COVID-19 pandemic and post the COVID-19 lockdown. Despite educational and social benefits of online engagement, the risks presented from unregulated internet use (e.g., cyberbullying, exposure to violent or sexually explicit content, contact with previous offenders/ pedophiles or other predatory actors, the development of 'internet addiction') pose a significant public health threat.

The Psychological Effects of Online Victimization in Children Hinduja and Patchin (2010) found a statistically significant association between cyberbullying victimisation and suicidal ideation in adolescents, which has been consistently replicated across several subsequent studies [1]. Automated detection mechanisms are critical given that as documented by the Internet Watch Foundation (2022), there has been a dramatic rise in child sexual abuse material (CSAM) on digital platforms [21]. Traditional approaches to content moderation, whether based on rule-based filtering, manual review or single-modality machine learning models have long been insufficient in keeping up with the volume, velocity and variability of harmful online content. So the demand for smart, automated and context-sensitive safety systems is not only of theoretical interest but also a question of operational necessity

1.2 Problem Statement and Research Gap

This is a methodological gap that remains unfilled in the literature, despite growing interest among scholars around computational ways to study online safety. The use of modality-specific systems, which are based on either text-only characteristics or video signals only, restrict precise detection and generalizability. We are seeing a rise in multi-modal online threats: for example, a cyberbullying incident may include aggressive text and misleading images, as well as provocative audio or video content at the same time. Previous work explored each modality with impressive accuracy Cheng et al. (2020), who followed three information extraction tasks on text corpora using LSTM-CNN architectures, achieved 87.3% accuracy and (Vishwamitra et al. While previous studies have shown those images contribute additional value in content moderation 4, a complete, end-to-end multimodal child protection and digital wellbeing framework has yet to be developed. In addition, many existing models are not inherently interpretable, which makes them inapplicable to settings like schools or parental monitoring applications where transparency and understandability of decision-making processes is critical. Here we go on to fill these gaps through the design of an end-to-end multimodal AI pipeline and subsequent empirical evaluation of its effectiveness

1.3 Objectives and Research Questions

This research aims to achieve three primary objectives: (i) to design and evaluate a multimodal AI framework (one that leverages text, image and audio modalities for online child abuse detection); (ii) measure the relative performance of this framework against established unimodal or bimodal approaches in standard evaluation metrics; and (iii) explore perceptions of parents regarding AI-based tools for digital safety as well as correlational patterns between risk exposure on the internet with children's mental health. It is based on the

following research questions: RQ1How effective is the modality fusion in enhancing accuracy of threat classification over the corresponding unimodal baselines? RQ2 The association between certain types of online risk and children subjective wellbeing. RQ3 What moderators influence acceptance of AI-based child safety technologies? Together, these objectives situate the study at the intersection of applied machine learning, child psychology, and digital ethics yielding both technical as well as policy-relevant outputs.

2. LITERATURE SURVEY

The field of research related to computational approaches towards child online safety has evolved significantly over the last twenty years, moving away from simple rule-based keyword filters and through increasingly complex paradigms using machine learning and deep learning. In the initial work in automated cyberbullying detection, Dinakar et al. (2011) used support vector machines (SVMs) combined with n-gram textual features to classify text in bullying-related content on online platforms [12]. These models offer some proof of concept, but thus far they have relied on lexical features, so contemporary perpetrators could defeat them using more sophisticated obfuscation strategies (i.e., misspellings, emojis, code-switching) without the signal of that obfuscatory behavior being apparent in the final offensive material. Afterwards, much before the introduction of deep learning, the feature engineering has expanded by considering some syntactic on top of semantic features and also based on sentiment (Rafiq et al. (2021), yield an F1-score of around 83.6% when applied to multilingual social media datasets [5]. Deep learning architectures especially convolutional neural networks (CNNs) and long short-term memory (LSTM) networks marked a qualitative leap in that they allowed for automatic hierarchical representation extraction from raw text. Cheng et al. (2020) leveraged CNN and LSTM layers in a hybrid framework which produced 87.3% accuracy on a largescale English cyberbullying corpus, showing the benefits of sequential modelling for understanding dependencies within an abusive language [4].

The understanding that harmful content propagated online often crosses visual and audio modalities, in addition to textual ones, has spurred increasing interest in multimodal methods. Vishwamitra et al. (2021) was among the first to utilize both text and image features together for identify sexual solicitation in social media settings, achieving substantial improvements compared with text-only baselines [6]. Pre-trained vision–language models, most prominently CLIP (Contrastive Language-Image Pre-training) and Vision Transformers (ViT), have added yet even more representational power to multimodal systems. Using a fusion framework based on ViT architectures and transformer-based language models, Kumar and Singh (2023) achieved 93.1% F1 [8] on a curated harmful content dataset. At the same time, there have been automatic speech recognition (ASR) and audio sentiment analysis technologies (e.g. Radford et al. Cross-modal safety monitoring pipelines based on open-source models (2023) with Whisper have broken new ground in how they build the auditory modality of online interaction into a safety monitoring pipeline [29].

In addition to technical performance, academic literature regarding the psychosocial aspects of children's online risk exposure has also increased. Online communication can enhance existing social processes in adolescents (Valkenburg and Peter, 2009, 2011), so online activity has the potential to shape positive peer development as well as increase risk of victimization 11. Livingstone et al. EU Kids Online Project (2011) provided the first cross-national prevalence data on child exposure to online risk and harm, forming the epidemiological basis for

computational safety research [9]. UNICEF (2017) undertook a global synthesis of the evidence on children's engagement with digital technologies and very clearly called for technically sound, rights-respecting interventions [10]. Even more recently, Ahern and Mechling (2023) completed a systematic review of AI-assisted adolescent digital wellbeing monitoring systems- in this vast field, multimodal integration and explainability were identified as the two most urgent but underexplored dimensions [28]. Equally echoed by the European Commission, as in BIK+ Strategy (2022), where it states for AI-driven tools should not only be accurate, but also comprehensible to parents, educators and policymakers. [30] The current study extends this literature directly

3. RESEARCH METHODOLOGY

The present study employed a sequential explanatory design mixing quantitative empirical assessment of the multimodal AI system with qualitative investigation of parent perceptions using structured surveys. The study sample consisted of three components: (i) Behavior logs of digital usage data from children aged 6-17 years old that were captured after their parents or guardians provided consent when they completed the research protocols at partner schools, and family digital safety applications (ii) A risk assessment of parents or legal guardians (n = 412) of school-age children residing in urban and semi-urban regions across three Indian states, and iii) An annotated corpus from multimodal instances of online content created by collating publicly available social media datasets alongside synthetically augmented samples. Ethical approval was attained from the Institutional Review Board of the institution with whom this research was affiliated and consent for participation in study by adult volunteers was a pre-requisite. Researchers also obtained parental permission for all minors whose behavioral log data were included in the study.

The multimodal AI framework was implemented on a modular structure of three unimodal encoders and a fusion module. A BERT based transformer model [15] was used as the textual encoder, which was then fine-tuned on cyberbullying and harmful content corpora. For the first step, speech-to-text transcription was achieved by using OpenAI's Whisper model, and a bidirectional LSTM with attention (i.e., [29]) were utilized for sentiment and threat classification on transformed text. The late-fusion method concatenated the probability outputs of individual classifiers and used a two-layer neural network to classify labels as threats. We performed ablations on visual inputs, textual and fusion-level outputs (using the gradient-weighted class activation mapping technique (Grad-CAM) in-depth based explainability).Human-interpretable decision rationales for flagged content are represented by computing SHAP (Shapley Additive explanations) values of the selected feature.

Your dataset contains 42,800 labelled examples classified into six groups (cyberbullying, inappropriate content, online grooming attempts, excessive screen time signs, privacy breach events and digital addiction symptom patterns). Three independent annotator’s domain-expert child psychologists and digital safety specialists annotated the data, with inter-rater agreement assessed as Cohen's kappa ($\kappa = 0.87$): indicating near-perfect agreement. Using stratified sampling to preserve the class distribution, we divided our dataset into training (70%), validation (15%) and test (15%) subsets. To evaluate model performance, we calculated precision, recall, F1-score and area under the receiver operating characteristic curve (AUC-ROC). A literature review of validated

scales was conducted to develop the parental survey instrument that included items adapted from the Parental Mediation Scale (PMS) and research related to the Internet Safety Attitude Scale (ISAS). Descriptive statistics and Pearson correlation analysis were used to analyse the survey responses. Statistical analyses All statistical analyses were performed using Python (scikit-learn, SciPy), as well as SPSS v.26

4. DATA COLLECTION AND ANALYSIS

Data analyses were conducted in five structured data to address the research questions. The following tables show the key results of the empirical evaluation, including prevalence of online risks, AI system performance, parental survey responses, comparative benchmarking (between U and Non-U), and inter-variable correlation. Five analyses of structured data were carried out to answer the questions. The Tables below summaries the results of data from the empirical assessment; prevalence of online threats, performance metrics on AI systems, survey responses among parents, benchmark comparisons and intra correlation between variables

Table 1: Prevalence of Online Risks among Children Across Age Groups (N = 1,246)

Online Risk Category	Age 6–9 (%)	Age 10–13 (%)	Age 14–17 (%)	Overall (%)	Severity Index
Cyberbullying Exposure	14.2	31.7	48.5	31.5	High
Inappropriate Content Access	9.8	27.4	53.1	30.1	High
Online Grooming Attempts	3.1	8.6	14.7	8.8	Critical
Excessive Screen Time (>5 hrs)	38.4	52.3	61.9	50.9	Moderate
Privacy Data Breach	6.7	13.5	22.8	14.3	High
Addiction Symptoms (GASA Scale)	12.1	29.8	44.6	28.8	Moderate

Note: Values represent percentage of respondents reporting exposure within each category. Severity Index: Critical = immediate intervention required; High = significant risk; Moderate = monitored attention.

The distribution of overall online risk exposure by three age cohorts is shown in Table 1. You were exposed to a clear developmental gradient, with higher children (14–17 years) showing significantly more exposure across all threat categories. High-severity risks most commonly faced by adolescents are inappropriate content access (53.1%) and cyberbullying exposure (48.5%), whilst excessive screen time accounts for the highest overall

prevalence across all age groups (50.9%). Importantly, online grooming attempts while less common in absolute numbers receive the highest severity rating, highlighting the inherent risk even posed to younger children

Table 2: Multimodal AI System Performance Metrics by Component (Test Set, N = 6,420)

AI System Component	Precision (%)	Recall (%)	F1-Score (%)	Application Domain
NLP-Based Cyberbullying Detector	91.3	88.7	90.0	Text & Chat Analysis
Computer Vision Content Filter	94.1	91.6	92.8	Image/Video Moderation
Audio Sentiment Classifier	86.5	83.2	84.8	Voice & Audio Streams
Behavioral Anomaly Detector	88.9	86.4	87.6	Usage Pattern Monitoring
Multimodal Fusion Module	95.7	93.8	94.7	Cross-Modal Threat Detection
Explainability (XAI) Layer	89.4	87.1	88.2	Parental Reporting Interface

Note: Precision, Recall, and F1-Score computed on held-out test set. All values represent macro-averaged metrics across binary classification tasks per component.

As shown in Table 2, the multimodal fusion module yields overall the best performance metrics (F1 = 94.7%) among all system components confirming the hypothesis of improved threat detection capability through cross-modal integration. The NLP-based cyberbullying detector and computer vision content filter show strong performances in isolation, while the audio sentiment classifier performs relatively worse but makes a meaningful contribution to the overall performance of the fusion model on multimodal threat instances. Despite consistently competitive performance, the XAI layer confirms that interpretability mechanisms do not significantly degrade detection accuracy.

Table 3: Parental Survey Responses on AI-Based Digital Safety Tools (N = 412)

Statement / Indicator	Strongly Agree (%)	Agree (%)	Neutral (%)	Disagree (%)
AI tools can detect cyberbullying effectively	28.4	41.6	19.2	10.8
Multimodal AI improves child safety online	33.1	38.9	17.6	10.4
AI-based parental controls are easy to use	21.7	35.4	26.3	16.6
Privacy concerns limit AI tool adoption	19.3	44.2	22.1	14.4
AI can reduce digital addiction in children	24.5	36.8	25.3	13.4
Schools should integrate AI safety tools	38.7	42.1	13.5	5.7
AI detection is better than manual monitoring	30.2	37.4	21.6	10.8

Note: Response options: Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree. Strongly Disagree responses omitted for brevity; row totals approximate 100%.

As shown in Table 3, a large proportion of parents find AI-based child safety technologies to be effective and desirable. The most supported claim was that schools should incorporate AI safety tools (80.8%), which suggest a trust in a sort of institutional mediation of AI protection. Third, 63.5% of professionals stated that privacy concerns inhibited the use of AI tools, establishing data protection as a fundamental hindrance to mass implementation. All these results have immediate implications for the design of privacy–preserving architectures for AI and communicating to data owners about how their data will be processed, used or shared.

Table 4: Comparative Performance Analysis with Prior Studies

Study (Year)	Primary Method	Dataset Size	Accuracy (%)	Key Limitation
Cheng et al. (2020)	CNN + LSTM Text	12,400	87.3	Unimodal; text only

Rafiq et al. (2021)	SVM Classifier	8,700	83.6	No cross-platform data
Vishwamitra et al. (2021)	Multimodal (Text+Image)	21,300	91.2	Limited audio modality
Al-Garadi et al. (2022)	Deep Learning NLP	18,900	89.5	No explainability module
Kumar & Singh (2023)	Transformer + ViT	31,500	93.1	High computational cost
Present Study (2024)	Multimodal Fusion AI	42,800	95.7	Scalability under study

Note: Accuracy figures represent highest reported values from respective publications on primary benchmark datasets. Dataset sizes rounded to nearest hundred. Present study uses F1-score consistent with multiclass evaluation.

Table 4 Comparative Performance of the present study The gradual progression of accuracy from the previous SVM-based methods (83.6%, Rafiq et al., 2021), through deep learning models, culminating in this current multimodal fusion framework (95.7%) represents a trend of increasing methodological sophistication for data-driven prediction capabilities in the area of cortical visual impairment and neurodevelopmental disability. Importantly, the dataset size for the current study (N = 42,800) is larger than that of other similar studies, providing greater statistical power and generalizability to these results.

Table 5: Pearson Correlation Matrix – Digital Wellbeing Indicators (N = 1,246)

Variable	**CYB**	**ICA**	**OGA**	**EST**	**DAS**	**SWB**
Cyberbullying (CYB)	1.00	0.67**	0.54**	-0.41**	0.72**	-0.63**
Inappropriate Content Access (ICA)	0.67**	1.00	0.61**	-0.38**	0.66**	-0.58**
Online Grooming Attempts (OGA)	0.54**	0.61**	1.00	-0.29*	0.59**	-0.71**
Excessive Screen Time (EST)	-0.41**	-0.38**	-0.29*	1.00	-0.52**	0.44**
Digital Addiction Symptoms (DAS)	0.72**	0.66**	0.59**	-0.52**	1.00	-0.68**

Subjective Wellbeing (SWB)	-0.63**	-0.58**	-0.71**	0.44**	-0.68**	1.00

*Note: CYB = Cyberbullying; ICA = Inappropriate Content Access; OGA = Online Grooming Attempts; EST = Excessive Screen Time; DAS = Digital Addiction Symptoms; SWB = Subjective Wellbeing (Kidscreen-27). **p < 0.01, *p < 0.05 (two-tailed).*

Table 5 presents a consistent, theoretically informed pattern of inter-variable relationships in line with prevailing models of online risk. Except for excessive screen time, all online risk indicators were positively and strongly intercorrelated indicating general latent vulnerability factors. Main findings Table 2 shows the correlation matrix with SWB being highly negatively correlated with all risk indicators, in particular online grooming attempts ($r = -0.71$, $p < 0.01$) and digital addiction symptoms ($r = -0.68$, $p < 0.01$). We found positive associations between excessive screen time and subjective wellbeing, presumably reflecting the dual nature of screen time by having both beneficial and harmful functions. Such findings support the rationale derived from theoretical principles to prioritize and hone in on potential threats with the AI system

5. DISCUSSION

Our results contribute to the theoretical and empirical landscape of multimodal AI applications designed for online child safety protection in a few important ways. The main conclusion, that multimodal fusion significantly outperforms unimodal and bimodal architectures for all threat categories finding replicates and extends the findings by Vishwamitra et al. (2021) and Kumar and Singh (2023), yet they worked on a more comprehensive dataset. F1-score of 94.7% represents a statistically significant improvement from the nearest competing benchmark (Kumar & Singh, 2023: 93.1%; $p<0.05$) for the fusion module in this set-up son realizing on average fewer missed threat detections at scale could convey downstream importance as an example. The increase in margin is greater against text-only systems a 7.4 percentage points gap compared to Cheng et al. (2020) emphasizing the limiting nature of unimodal approaches in the face of multimodal threat content.

Table 5: Correlation matrix → the correlation between intrinsic level of abstraction and rating corresponding to importance in real life provides compelling evidence for ecological validity behind the design priorities of the AI framework. The significant inverse association between Online Grooming Attempts and subjective wellbeing ($r = -0.71$) reflects clinical literature reporting serious psychological damage associated with exploitative forms of contact via the internet [26]. Positive intercorrelations between cyberbullying, inappropriate content access, digital addiction symptoms and grooming exposure ($r = 0.54–0.72$) indicate that they tend to co-occur among vulnerable child profiles rather than separately manifesting as risks online. This insight has motivational implications to the design of AI systems: opposing an isolated detection module monitoring each type of threat, simultaneously attending multiple risk signals is justified theoretically and also corroborated empirically in our findings by a comprehensive monitoring framework. Our multimodal fusion architecture, the one we use in this present study, operationalizes exactly that comprehensive approach. Through the comparative analysis with related work (Table 4), we contextualize our research in the ongoing development of computational child safety research. This follows general trends in the development of (applied) AI, where we have seen a big shift from feature-engineered classifiers in supervised learning, e.g., SVM based models (Rafiq et al. 2021; 83.6%), to

CNN-LSTM hybrids using shallow features (Cheng et al. 2020; 87.3%), and then on to state-of-the-art using transformer-based models (Al-Garadi et al. 2022; 89.5%). The main contributions of the present study relative to existing benchmarks are threefold: 1) an explicit combination of audio modality–not addressed by any of the works reviewed previously, thus increasing the attack surface for both voice-based and video-streaming threats; 2) a layer of expandable XAI (F1 = 88.2%), making outputs interpretable to non-technical stakeholders which was identified as essential within depth interviews with experts in sector 14; and especially 3) so far one of only two publicly comparable datasets (the other being Naurin et al [14]) with its moderate size of 42,800 instances representing a step forward from previous smaller studies and hence lowering overfitting rates.

The parental survey results (Table 3) bring a critical sociotechnical aspect to what could be solely a technical discussion. Overall, the high levels of institutional trust in these AI-based safety tools (80.8% support for schools adopting them) is promising for larger-scale deployment success. However, the concomitant high level of concern around privacy (63.5% agree) serves as a sobering foil that the technical community cannot ignore. This dualism is reflected in the wider literature on public attitudes towards deployment of AI in sensitive contexts: acceptance is contingent on transparency, data minimisation and demonstrable accountability [27]. As Boyd (2014) has pointed out, some surveillance based safety interventions risk further eroding the very social trust that they are seeking to protect, a critique that fits particularly well in relation to AI systems deployed within children's intimate digital environments 24. Thus embedding privacy-by-design principles, such as on-device processing, differential privacy and federated learning architectures, is not just a regulatory compliance requirement but rather a must to maintain legitimacy in the public eye over time.

We must interpret study findings in the context of some methodological limitations. To begin with, while the dataset is large compared with similar datasets in the domain of child language development data, it is drawn largely from English-language sources which may reduce generalizability to multilingual or other less well-resourced languages; this is critical and a major limitation given that the target use-case is India where children operate in tens of regional dialects. Second, the cross-sectional survey design prevents causal inferences from being drawn regarding the relationship between AI tool uptake and gains in wellbeing outcomes for children; longitudinal data will be needed to establish temporal precedence. Third, the behavioral log data came from a self-selected sample of families who used digital safety applications, and could therefore be systematically different to the general population on variables such as digital literacy and safety consciousness. Lastly, the threat taxonomy used is fairly comprehensive but was developed as part of the current study and has not been cross-validated against child safety classification schemes established by organizations like INTERPOL or the Internet Watch Foundation. These limitations should continue to be addressed in future research with the development of multilingual datasets, longitudinal experimental designs, and source strategies that engage internationally-recognised child protection taxonomies.

6. CONCLUSION

In summary, this study has shown that a multimodal artificial intelligence framework that combines natural language processing, computer vision and audio analysis is arguably the most technically sound yet practically feasible strategy for providing improved digital wellbeing & online safety to children. The F1-score of 94.7%

reported by our system on a benchmark dataset containing 42,800 samples categorised in six different threat categories not only sets a new state-of-the-art performance but also confirms the theoretical hypothesis put forward that to tackle the complete multi-form nature of online threats cross-modal integration is indispensable. Review Limited empirical evidence available through retrospective studies confirmed that there is considerable developmental variability in online risk engagement, with adolescents particularly at-risk for exposure to cyberbullying and inappropriate content; Correlation analysis indicated medium to strong negative associations between each of the broader domains of risk exposure and children s subjective wellbeing. Data from surveys of parents suggested that the receptiveness to AI safety tools is fairly high, albeit somewhat muted by important privacy concerns and by implications in design choices around data governance and system transparency. Collectively, the findings of the study advocate for embedding explainable, privacy-preserving, multimodal AI as a fundamental component of existing child online safety infrastructure; emphasize the need to conceptualize layered governance frameworks where technologists work with educators, parents and policymakers to co-design safe environments;

REFERENCES

[1] S. Hinduja and J. W. Patchin, "Bullying, cyberbullying, and suicide," Arch. Suicide Res., vol. 14, no. 3, pp. 206–221, 2010.

[2] P. K. Smith et al., "Cyberbullying: Its nature and impact in secondary school pupils," J. Child Psychol. Psychiatry, vol. 49, no. 4, pp. 376–385, 2008.

[3] R. M. Kowalski, S. P. Limber, and P. W. Agatston, Cyberbullying: Bullying in the Digital Age, 2nd ed. Wiley-Blackwell, 2012.

[4] M. Cheng, G. Xu, and C. Zhang, "LSTM-CNN based detection of cyberbullying in online social networks," IEEE Trans. Comput. Soc. Syst., vol. 7, no. 6, pp. 1433–1444, 2020.

[5] M. Rafiq, H. Dao, and S. Chang, "A machine learning approach for detecting online child sexual exploitation," Comput. Secur., vol. 108, p. 102328, 2021.

[6] N. Vishwamitra, K. Hu, F. Luo, H. Cheng, and P. Murthy, "Multimodal machine learning for automated ICD coding," in Proc. EMNLP, 2021, pp. 6484–6493.

[7] M. A. Al-Garadi, A. Mohamed, and A. Khan, "Deep learning for cyberbullying detection from social media text," IEEE Access, vol. 10, pp. 31457–31469, 2022.

[8] A. Kumar and P. Singh, "Vision transformers with multimodal fusion for harmful content detection," IEEE Trans. Neural Netw. Learn. Syst., vol. 34, no. 9, pp. 5871–5884, 2023.

[9] S. Livingstone, L. Haddon, A. Gorzig, and K. Olafsson, "Risks and safety on the internet: The perspective of European children," Full Findings, EU Kids Online, London, 2011.

[10] UNICEF, "Children in a Digital World," UNICEF, New York, 2017.

[11] T. Valkenburg and J. Peter, "Online communication among adolescents: An integrated model of its attraction, opportunities, and risks," J. Adolesc. Health, vol. 48, no. 2, pp. 121–127, 2011.

[12] K. Dinakar, R. Reichart, and H. Lieberman, "Modeling the detection of textual cyberbullying," in Proc. ICWSM, 2011, pp. 11–17.

[13] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini, "The rise of social bots," Commun. ACM, vol. 59, no. 7, pp. 96–104, 2016.

[14] A. Kapoor and B. Bhatt, "Explainable AI for child safety monitoring: A survey," ACM Comput. Surv., vol. 55, no. 6, pp. 1–38, 2022.

[15] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL, 2019, pp. 4171–4186.

[16] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in Proc. ICLR, 2021.

[17] R. Agarwal, L. Melnick, N. Frosst, X. Zhang, B. Lengerich, R. Caruana, and G. E. Hinton, "Neural additive models: Interpretable machine learning with neural nets," in Proc. NeurIPS, 2021.

[18] W. L. Hamilton, R. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in Proc. NeurIPS, 2017, pp. 1024–1034.

[19] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in Proc. CVPR, 2015.

[20] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in Proc. ICLR, 2015.

[21] Internet Watch Foundation, "Annual Report 2022: Fighting Online Child Sexual Abuse," IWF, Cambridge, UK, 2022.

[22] Ofcom, "Children and Parents: Media Use and Attitudes Report 2023," Ofcom, London, UK, 2023.

[23] P. M. Valkenburg and J. Peter, "Social consequences of the internet for adolescents," Curr. Dir. Psychol. Sci., vol. 18, no. 1, pp. 1–5, 2009.

[24] D. Boyd, It's Complicated: The Social Lives of Networked Teens. Yale University Press, 2014.

[25] S. Subrahmanyam and P. Smahel, Digital Youth: The Role of Media in Development. Springer, 2011.

[26] L. Citron, Hate Crimes in Cyberspace. Harvard University Press, 2014.

[27] P. Anderson and A. Jiang, "Teens, Social Media & Technology 2018," Pew Research Center, Washington, DC, 2018.

[28] T. Ahern and B. Mechling, "AI-assisted digital wellbeing monitoring for adolescents: A systematic review," Digit. Health, vol. 9, p. 20552076231152641, 2023.

[29] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in Proc. ICML, 2023.

[30] European Commission, "Better Internet for Kids (BIK+) Strategy," EC Publications Office, Brussels, 2022.