An Interpretable Multimodal AI Framework to Support Psychological Follow-Up in Child Maltreatment Cases

1st Ana Evelyn Vázquez Pérez
Department of Artificial Intelligence
Escuela Superior de Cómputo
Instituto Politécnico Nacional
Mexico City, Mexico
ORCID: 0009-0009-0018-6769
evelyn.vazquez070898@gmail.com

2nd Andrés Emiliano Arenas Jiménez
Department of Artificial Intelligence
Escuela Superior de Cómputo
Instituto Politécnico Nacional
Mexico City, Mexico
ORCID: 0009-0004-9209-6229
emilianoarenasj17@gmail.com

3rd Daniel Molina Pérez
Department of Computation
CINVESTAV-IPN
Department of Computer Science
ESCOM-IPN
Mexico City, Mexico
dmolinap@ipn.mx

***Abstract*—Child maltreatment presents a persistent challenge for psychological follow-up, as its manifestations are indirect, heterogeneous, and often conveyed through subtle verbal, vocal, and non-verbal cues. This study introduces an interpretable multimodal Artificial Intelligence (AI) framework designed to support clinical follow-up in child maltreatment cases by integrating three complementary modules: Speech-based Emotion Recognition (SER), lexicon-driven Natural Language Processing (NLP), and vision-based non-verbal behavior analysis. Each module operates independently to produce transparent, auditable outputs, which are subsequently combined through a late-fusion strategy that evaluates cross-modal coherence rather than issuing diagnostic decisions. Experimental results show that the NLP module achieved the highest standalone performance (85.0% accuracy, 83.0% Macro-F1), whereas speech and vision-based modules yielded lower results (70.1% and 52.5% accuracy, respectively), reflecting the ambiguity and context-dependence of paralinguistic and facial signals in child-centered domains. Multimodal analysis revealed strong cross-modal agreement in 15% of cases and partial alignment between linguistic and vocal cues in 30%, illustrating the complementary nature of indirect modalities.**

***Index Terms*—child maltreatment, natural language processing, speech emotion recognition, computer vision, multimodal AI**

I. Introduction

Child maltreatment is a complex global public health and social problem with severe and long-lasting consequences. The World Health Organization (WHO) reports that approximately six in ten children under the age of five experience physical punishment and/or psychological aggression from caregivers, and that one in five girls has been subjected to sexual abuse during childhood [1]. Child maltreatment includes all forms of physical abuse, emotional abuse, sexual abuse, and neglect experienced by individuals under 18 years of age [2]. These forms of abuse may occur independently or, more frequently, in combination, leading to actual or potential harm to the child's health, survival, development, or dignity within a relationship of responsibility, trust, or power [3].

One of the main challenges in the psychological follow-up of child maltreatment cases lies in the indirect and multifaceted nature of its manifestations [4]. In many cases, indicators of abuse do not emerge through explicit disclosure but are instead reflected in subtle verbal and non-verbal cues. These may include variations in language use, prosody, and vocal affect, as well as atypical facial expressions, defensive body postures, or behaviors incongruent with the child's developmental stage [5]. Such signals are often ambiguous and difficult to detect during clinical interviews. Consequently, psychological assessment and follow-up remain highly subjective processes that rely heavily on professional expertise, contextual interpretation, and observational sensitivity [6].

In recent years, advances in Artificial Intelligence (AI) have enabled the automated analysis of large-scale multimodal data, facilitating the identification of complex behavioral and affective patterns. Techniques such as speech processing, computer vision, and Natural Language Processing (NLP) have shown strong potential for capturing acoustic, visual, and temporal features associated with human emotional expression and behavior [7]–[9]. In particular, multimodal learning approaches integrate complementary verbal and non-verbal information streams, thereby enhancing the sensitivity of analysis to indirect or subtle manifestations that may be overlooked by unimodal methods [10], [11].

This paper presents an interpretable multimodal AI framework designed to support psychological follow-up in cases of child maltreatment. The framework integrates three complementary modules: (i) Speech-based Emotion Recognition (SER), (ii) lexicon-driven NLP, and (iii) and vision-based non-verbal behavior analysis. This approach focuses on generating traceable evidence to assist psychologists during follow-up sessions with children who have been victims of maltreatment. The framework is conceived as a decision-support tool that complements clinical expertise.

II. Related Work

A. Psychological and Behavioral Foundations of Child Maltreatment

From a psychotraumatological perspective, child maltreatment is understood not as a single event but as a cumulative process embedded in relational, developmental, and sociocultural contexts [12]. Within an ecological framework, children's emotional expressions and coping strategies are shaped by interactions with caregivers, family dynamics, and broader environments [13]. Consequently, trauma-related distress rarely follows a uniform trajectory and often manifests through heterogeneous, indirect, and predominantly non-verbal behaviors that vary across individuals and developmental stages.

Children exposed to chronic abuse or neglect may exhibit distress through disorganized narratives, affective incongruence, altered language patterns, hypervigilance, withdrawal, or defensive body language rather than explicit verbal disclosure [6], [14]. In clinical practice, assessment and follow-up rely largely on the interpretation of these subtle verbal and non-verbal cues observed longitudinally, particularly in cases of chronic trauma [4], [15], [16]. These psychological foundations highlight the need for multimodal approaches capable of integrating diverse behavioral and emotional indicators to support systematic and context-aware interpretation.

B. Artificial Intelligence Approaches to Child Maltreatment Analysis

NLP methods have been used to examine children's narratives, interview transcripts, and clinical documentation to identify linguistic patterns associated with emotional distress, trauma, or abuse-related experiences [17]–[19]. Parallel advances in speech processing have leveraged acoustic and paralinguistic features to model affective states such as fear, anxiety, or emotional dysregulation [8], [20].

On the other hand, computer vision has been employed to analyze facial expressions, body posture, and non-verbal behaviors of children to evaluate potential indicators of psychological distress, thereby contributing complementary observational evidence to psychological assessment [10].

Finally, research in affective computing and mental health analysis has explored multimodal learning frameworks that integrate textual, acoustic, and visual information to model human emotions and social interactions [11], [21], [22]. These approaches demonstrate that combining complementary verbal and non-verbal modalities can improve the detection of subtle affective cues and complex emotional states.

However, such multimodal methods have not yet been applied to the specific context of child maltreatment or trauma-related assessment. Existing work remains focused on adult populations and general affective recognition tasks, leaving a significant research gap in the development of interpretable, multimodal AI systems tailored to the psychological follow-up of children exposed to abuse or neglect.

C. AI-Based Decision Support Systems in Child Protection

Several real-world AI-based decision support systems have been developed to assist child protection agencies in managing large volumes of referrals and prioritizing cases based on estimated risk. For instance, the Allegheny Family Screening Tool (AFST), implemented in the United States, to support child welfare call screening decisions by generating risk scores derived from large-scale administrative and social service records [23]. The system was explicitly designed to assist, rather than replace, human decision-making, emphasizing transparency at the level of input variables while relying on population-level historical data.

Similarly, in Sweden, the municipality of Norrtälje has deployed AI-based tools to analyze early warning referrals and identify cases at higher risk of future maltreatment, with the aim of supporting preventive intervention strategies [24]. These systems likewise operate on structured institutional data and are intended to improve consistency and efficiency in case assessment.

While such decision support tools demonstrate the growing interest for AI applied to child protection, they offer limited insight into the child's current psychological state or non-verbal behavior observed during clinical follow-up. This limitation highlights the gap between population-level risk prediction systems and clinically interpretable, multimodal approaches tailored to psychological assessment.

D. Limitations of Current AI-Based Approaches

Most existing AI systems for child protection focus on population-level risk prediction based on administrative or historical data, offering limited sensitivity to a child's current emotional state, communicative behavior, or non-verbal expression during clinical interactions. As a result, these models generate broad risk scores rather than session-level indicators that can guide individualized psychological follow-up. Research explicitly addressing child maltreatment remains scarce and is typically centered on detection or prevention, with little attention to post-diagnostic intervention or longitudinal assessment. The lack of close collaboration between AI researchers and clinical practitioners further restricts the development of tools aligned with therapeutic processes.

In addition, current approaches rarely integrate linguistic, vocal, and visual information within a unified analytical framework. Most systems process each modality separately, missing the cross-modal and temporal relationships—such as inconsistencies between verbal content and affective tone that characterize trauma-related distress. The absence of multimodal architectures for child maltreatment cases limits the applicability in real-world psychological contexts.

III. Methods and Materials

A. Data Sources

The evaluation of the proposed framework relies on the publicly available *Multimodal Child Emotion for Learning* dataset [25] and the LIRIS-Children Spontaneous Facial Expressions

dataset [26], which provide synchronized audio, visual recordings, and annotations from children engaged in learning-related activities and spontaneous facial behavior in children, respectively. These datasets include speech recordings and facial expression data acquired under controlled conditions, together with emotion labels spanning multiple affective states.

In addition, the NLP module was evaluated using a curated set of 100 short sentences designed to reflect child maltreatment–related language patterns across five classes (neutral, physical abuse, psychological abuse, sexual abuse, and neglect). The sentences were reviewed and validated by three licensed psychologists from a non-governmental organization specializing in the psychological follow-up of child maltreatment cases [4]. All data were used exclusively for methodological experimentation, without any attempt to infer clinical diagnoses or real-world maltreatment cases.

B. SenseVoice

SenseVoice is a large-scale pretrained speech representation model designed to encode semantic, emotional, and prosodic information by leveraging multilingual and multitask training on diverse speech corpora. The use of such pretrained embeddings enables the transfer of knowledge learned from large-scale data, improving robustness to speaker variability, recording conditions, and expressive differences, while reducing dependence on limited task-specific datasets [27].

C. Mel-Frequency Cepstral Coefficients

MFCCs model the short-term spectral envelope of speech by applying a perceptually motivated Mel-scale filter bank followed by cepstral decorrelation, effectively capturing information related to vocal tract configuration, articulation, and timbral characteristics of speech signals. MFCC-based representations have demonstrated robustness and computational efficiency across a wide range of speech-related applications, including emotion recognition, speaker characterization, and behavioral analysis, particularly in scenarios with limited training data or constrained computational resources [28].

D. Leave-One-Speaker-Out

LOSO cross-validation is an evaluation protocol designed to assess speaker-independent generalization by ensuring that all samples from a given speaker are excluded from training and used exclusively for testing. This strategy addresses inter-speaker variability, which is a critical challenge in speech-based affective and behavioral analysis. LOSO-style evaluation provides a realistic estimate of real-world performance, particularly when models are expected to operate on previously unseen individuals [29].

E. Facial landmark-based representation

An approach for modeling facial structure and dynamics in affective computing. Rather than operating directly on raw pixel intensities, this paradigm represents the face through a set of anatomically meaningful key points corresponding to regions such as the eyes, eyebrows, nose, mouth, and facial contour. By capturing geometric configurations and their temporal evolution, facial landmarks provide a compact description of facial movement that is robust to variations in illumination, background, and individual appearance [30].

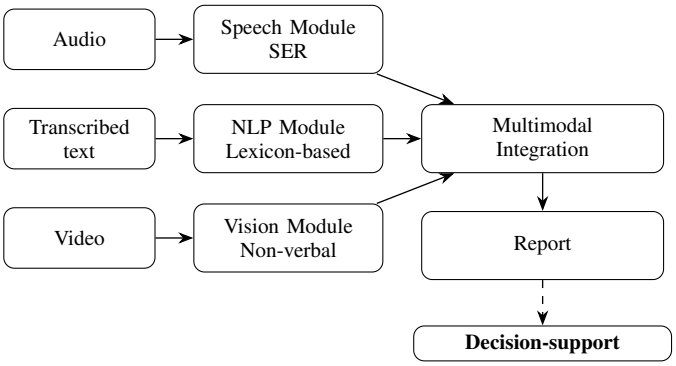

Fig. 1. Multimodal architecture designed for psychological child maltreatment follow-up.

This representation has been extensively used for facial expression analysis and behavioral modeling, particularly in contexts where transparency and sensitivity are critical, as it enables the analysis of expressive patterns without relying on identity-dependent visual features. As such, landmark-based methods offer a theoretically grounded alternative to pixel-based convolutional approaches, supporting interpretable analysis of non-verbal behavior in child-centered and ethically sensitive applications.

IV. Multimodal Framework for Child Maltreatment Psychological Follow-Up

We propose a modular multimodal architecture designed to support psychological follow-up through the integration of three complementary modules: (i) SER, (ii) lexicon-driven NLP, and (iii) and vision-based non-verbal behavior analysis. Figure 1 illustrates the workflow of the proposed system, showing the input data streams, the internal processing modules, and the final stage that generates structured multimodal reports.

Given the substantial variability in how children manifest maltreatment-related distress, this framework is designed to accommodate heterogeneous expressive patterns. Some children may exhibit heightened gestural or emotional expressiveness, whereas others may display emotional blunting or apparent tranquility as a result of normalization or coping mechanisms. Accordingly, each module generates interpretable outputs that can be examined independently or jointly, facilitating transparency and clinical interpretability.

The proposed framework does not aim to produce diagnostic decisions. Instead, it provides psychologists with structured multimodal evidence to support professional judgment during psychological follow-up sessions.

A. Speech-based Emotion Recognition Module

This module focuses on identifying emotional patterns expressed through children's vocal behavior. Audio segments

are processed using SenseVoice, which extracts high-level acoustic embeddings capturing paralinguistic and affective features. These representations are used to classify seven emotional categories: anger, disgust, fear, happiness, neutrality, sadness, and surprise. In parallel, MFCCs are extracted to enable comparative evaluation between learned and traditional acoustic features.

Emotion classification is performed following a LOSO cross-validation protocol to ensure speaker-independent generalization. This protocol is particularly relevant in child-centered emotion recognition, where inter-speaker variability in prosody, pitch range, and articulation is pronounced and can substantially affect generalization.

B. Lexicon-driven NLP Risk Scoring Module

The module operates on textual input obtained either directly or via automatic speech recognition and performs sentence-level analysis to preserve local linguistic context. A curated lexicon defines risk-related lexical items, each associated with a maltreatment category (physical, psychological, sexual or neglect), a base weight, and one or more regular expression patterns used for robust matching across linguistic variations.

For each sentence, matched lexical items contribute to category-specific risk scores, which are contextually adjusted through explicit handling rules, including negation detection within a token window and multiplicative down-weighting to avoid false positives from negated or third-person statements. The module outputs normalized risk scores for the four maltreatment categories along with severity labels and a detailed evidence list specifying matched text spans, contextual modifiers, and final contribution weights.

C. Vision-based Non-verbal Behavior Analysis Module

The vision module focuses on modeling facial dynamics using geometric representations instead of raw pixel data. Facial landmarks are extracted from each video frame using MediaPipe Face Mesh, yielding 468 fiducial points (x,y,z) per frame. The (x,y) coordinates are concatenated into a 936-dimensional feature vector that captures the relative configuration and movement of facial regions over time. To ensure robustness across recording conditions, a geometric normalization procedure is applied at the frame level: coordinates are re-centered on the nose tip to achieve translation invariance and scaled by facial width to ensure scale invariance.

Variable-length recordings are segmented into fixed-length sequences of 30 frames using an overlapping sliding window to capture the onset, apex, and offset of expressions. Temporal modeling is performed using a bidirectional GRU network with an integrated self-attention mechanism, which emphasizes frames exhibiting stronger expressive content. As a result, the module classifies each sequence into one of seven discrete emotion categories: anger, disgust, fear, happiness, neutral, sadness, and surprise.

D. Multimodal Integration

The combination of modules is implemented through a late fusion strategy at the decision level, in which each modality performs independent inference prior to a centralized aggregation stage. This design preserves modular independence, allowing each component to operate autonomously and generate preliminary outputs based on its own signal-specific evidence. Subsequently, a centralized logical aggregation layer evaluates the coherence and complementarity of the outputs produced by the individual modules.

The processing pipeline handles raw video recordings through an automated preprocessing workflow implemented in Python. Video files are decomposed into audio and visual streams, where the audio is converted into standardized mono-channel waveforms for speech analysis and SenseVoice-based transcription. Long recordings are segmented into discrete temporal units corresponding to individual interaction windows, ensuring that each data instance represents a coherent behavioral episode across modalities.

Finally, outputs from all modules are consolidated into a unified data structure, enabling systematic comparison across unimodal, bimodal, and trimodal evaluation scenarios. This structure supports the generation of structured, interpretable multimodal reports that can be directly utilized for psychological follow-up and expert clinical review.

V. RESULTS

Figure 2 summarizes the classification performance of the vision, speech, and NLP modules evaluated independently. The NLP module achieved the highest performance, with an accuracy of 85.0% and a Macro-F1 score of 83.0%, reflecting the effectiveness of the lexicon-driven scoring approach in capturing explicit verbal indicators associated with maltreatment-related content. The speech-based emotion recognition module obtained intermediate performance (70.1% accuracy, 68.0% Macro-F1), indicating that vocal affect provides informative but inherently noisy cues in child-centered emotion analysis.

In contrast, the vision-based module yielded lower performance (52.5% accuracy, 49.0% Macro-F1). While these values may appear modest compared to pixel-based deep learning approaches, they must be interpreted in light of the characteristics of the task and dataset. The LIRIS-Children Spontaneous Facial Expressions dataset comprises only 208 videos of spontaneous, non-acted facial behavior, where emotional expressions are subtle, heterogeneous, and often temporally fragmented. In this context, performance is inherently constrained by limited data availability and high intra-class variability.

A. Multimodal Consistency Analysis

To assess the degree of agreement across modalities, Figure 4 illustrates the distribution of multimodal consistency patterns observed in the evaluated samples. A majority of instances (55%) exhibited signals detected exclusively by the NLP module, corresponding to weak or isolated verbal

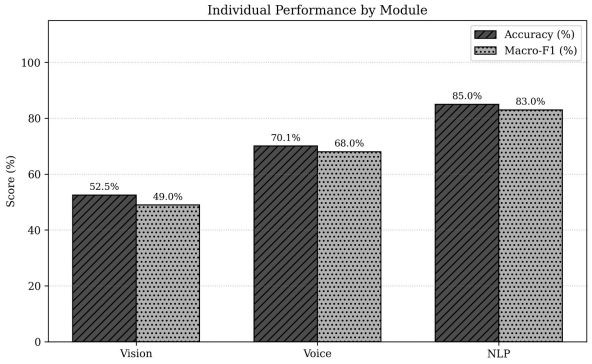

Fig. 2. Individual performance of modules.

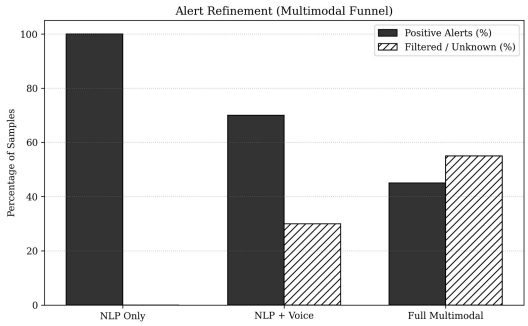

Fig. 3. Distribution of multimodal consistency patterns across evaluated samples.

indicators without corroboration from vocal or facial cues. Approximately 30% of the samples showed agreement between the NLP and speech modules, representing medium-strength multimodal evidence where verbal indicators co-occurred with affective vocal patterns. Only 15% of the samples demonstrated a strong triple-modality match, in which verbal, vocal, and non-verbal cues converged simultaneously.

Figure 3 presents the effect of the proposed multimodal fusion strategy on alert refinement. When relying solely on the NLP module, all detected cases are flagged as positive alerts, resulting in high sensitivity but limited specificity. The inclusion of speech-based emotion information reduces the proportion of positive alerts to approximately 70%, with 30% of cases reclassified as filtered or uncertain due to the absence of supporting vocal evidence. Incorporating the full multimodal configuration further refines the output, yielding 45% positive alerts and 55% filtered or uncertain cases.

This progressive reduction illustrates how multimodal integration functions as a funnel mechanism that increases evidential rigor by requiring cross-modal consistency. Rather than suppressing potential signals, the framework explicitly distinguishes between weak, medium, and strong evidence configurations, enabling more nuanced and interpretable reporting for psychological follow-up.

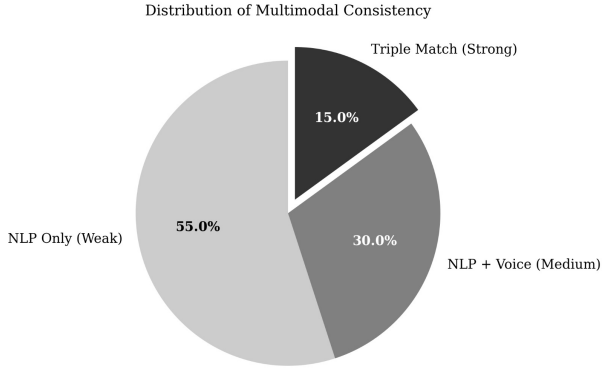

Fig. 4. Alert refinement through progressive multimodal integration.

VI. Discussion

The experimental results indicate that the signals associated with child maltreatment are heterogeneous, context-dependent, and rarely expressed consistently across a single modality. The NLP module achieved the highest performance metrics, suggesting that, although children may have normalized the abuse and may not exhibit explicit gesticulation or vocal inflection, they often verbalize experiences of maltreatment. However, reliance on NLP alone may overlook relevant affective or behavioral cues in cases where children minimize or avoid explicit disclosure.

In contrast, the speech-based and vision-based modules demonstrated lower standalone performance, which can be attributed to intrinsic signal ambiguity, dataset limitations, and the conservative evaluation metrics adopted to ensure methodological rigor. Unlike the NLP component, which captures explicit semantic indicators of maltreatment, vocal and facial modalities convey affective and behavioral information in an indirect and highly context-dependent manner.

Both modules, therefore, could be better oriented toward capturing descriptive non-verbal behavioral tendencies over time. Their outputs should be interpreted as inherently less separable within a traditional classification framework, reinforcing their role as complementary behavioral indicators for specific cases rather than as generalized primary decision-making components.

Furthermore, a key limitation across all modules is the restricted availability of ethically sourced, large-scale datasets involving child participants. Ethical constraints understandably preclude the collection of extensive, naturalistic speech and video data depicting real cases of maltreatment. As a result, current experiments rely on proxy datasets and small child-centered multimodal corpora that do not contain authentic

abuse-related material. These constraints reflect the broader ethical and data accessibility challenges inherent to child maltreatment research, while simultaneously underscoring the need for ethically curated, representative datasets and responsible model development in future work.

VII. Conclusion

This work proposes a multimodal AI framework designed to support psychological follow-up in cases of child maltreatment through the integration of linguistic, vocal, and non-verbal behavioral indicators. The combination of computer vision, NLP, and SER modules enables multimodal corroboration across otherwise isolated indicators.

The results demonstrate that while linguistic content provides the most explicit and reliable standalone signals, vocal and facial modalities contribute complementary, context-sensitive evidence that represents a key research opportunity for future studies.

At the same time, the study underscores the structural limitations inherent to AI research in child-centered domains. The modest standalone performance of the speech and vision modules is largely attributable to the use of limited, proxy datasets and small numbers of child participants. These constraints highlight the scarcity of ethically curated and representative datasets for child maltreatment research. This limitation also presents an opportunity for longitudinal data collection, deeper collaboration with clinical practitioners, and the refinement of multimodal fusion strategies that further align computational outputs with psychological interpretability.

Acknowledgment

The authors acknowledge the use of AI-based language tools for English editing and consistency checking. All ideas, results, and analyses were entirely conceived and conducted by the authors.

References

[1] World Health Organization, "Child maltreatment," World Health Organization, Tech. Rep., 2016. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/child-maltreatment

[2] ——, "Report of the consultation on child abuse prevention," World Health Organization, Geneva, Switzerland, Tech. Rep., 1999, consultation held 29–31 March 1999.

[3] L. Wang, H. Cheng, Y. Qu, Y. Zhang, Q. Cui, and H. Zou, "The prevalence of child maltreatment among chinese primary and middle school students: a systematic review and meta-analysis," *Social Psychiatry and Psychiatric Epidemiology*, vol. 55, no. 9, pp. 1105–1119, 2020.

[4] K. B. Andrade Rodríguez and V. Villa Aguirre, "Personal interview on psychosocial indicators of child maltreatment," Interview conducted by the authors, Reinserta A.C., 2025, psychologists at Reinserta Foundation.

[5] J. Overton, "Child abuse: Recognition, reporting, and response," *Journal of Christian Nursing*, vol. 39, no. 2, pp. 104–111, April–June 2022. [Online]. Available: https://doi.org/10.1097/CNJ.0000000000000942

[6] D. Cicchetti and S. L. Toth, "Child maltreatment and developmental psychopathology: A multilevel perspective," *Development and Psychopathology*, vol. 17, no. 3, pp. 533–536, 2006. [Online]. Available: https://doi.org/10.1017/S0954579405050279

[7] K. Jiang, N. Zeng, and J. Lin, "Multimodal emotion recognition based on deep learning: A comprehensive survey," *IEEE Transactions on Affective Computing*, 2022.

[8] G. Shani, D. B. Shalom, and A. M. Bronstein, "Automatic recognition of emotional distress from speech signals," *IEEE Transactions on Affective Computing*, vol. 12, no. 4, pp. 1091–1104, 2021.

[9] J. Barboza and J. J. Salerno, "Deep multimodal fusion for emotion recognition in naturalistic interactions," *Frontiers in Artificial Intelligence*, vol. 7, p. 1359821, 2024.

[10] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[11] A. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*, 2018.

[12] J. Barudy Labrín, *The Invisible Pain of Childhood: An Ecosystemic Perspective on Child Maltreatment*. Paidós, 1998.

[13] J. E. Korbin and R. D. Krugman, Eds., *Handbook of Child Maltreatment*. Springer, 2014.

[14] K. Kim, F. E. Mennen, and P. K. Trickett, "Patterns and correlates of co-occurrence among multiple types of child maltreatment," *Child & Family Social Work*, vol. 22, no. 2, pp. 492–502, 2016.

[15] A. Ampudia Rueda, G. B. Santaella Hidalgo, and S. Eguía Malo, *Clinical Guide for the Assessment and Diagnosis of Child Maltreatment*. National Autonomous University of Mexico, Faculty of Psychology / El Manual Moderno Publishing, 2009.

[16] M. D. DiLauro, "Psychosocial factors associated with types of child maltreatment," Ph.D. dissertation, Fordham University, 2001, uMI Microform 3005917.

[17] A. Y. Landau, S. Ferrarello, A. Blanchard, K. Cato, N. Atkins, S. Salazar, D. U. Patton, and M. Topaz, "Developing machine learning-based models to help identify child abuse and neglect: Key ethical challenges and recommended solutions," *Journal of the American Medical Informatics Association*, vol. 29, no. 3, pp. 576–580, March 2022. [Online]. Available: https://doi.org/10.1093/jamia/ocab286

[18] F. Lupariello, G. Di Vella, B. Gualco, and E. D'Aloja, "Artificial intelligence in the detection of child maltreatment: Opportunities and ethical challenges," *Children*, vol. 10, no. 9, p. 1659, 2023.

[19] J. I. Sorensen, R. M. Nikam, and A. K. Choudhary, "Artificial intelligence in child abuse imaging," *Pediatric Radiology*, 2021, received June 2020; accepted March 2021.

[20] N. Haines, M. W. Southward, J. S. Cheavens, T. P. Beauchaine, and W.-Y. Ahn, "Using computer vision and machine learning to automate facial coding of positive and negative affect intensity," *PLOS ONE*, vol. 14, no. 2, p. e0211735, 2019.

[21] Y. Jiang, X. Li, H. Luo, S. Yin, and O. Kaynak, "Artificial intelligence: Research, applications, and industrial outlook," *Discover Artificial Intelligence*, vol. 2, no. 4, p. 22, 2022.

[22] C. Zucco, B. Calabrese, and M. Cannataro, "Sentiment analysis and affective computing for depression monitoring," in *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. Kansas City, MO, USA: IEEE, 2017, pp. 1988–1995.

[23] Allegheny County Department of Human Services, "Allegheny family screening tool (afst): Frequently asked questions," Allegheny County Department of Human Services, Allegheny County, PA, USA, Tech. Rep., Aug. 2018, updated August 2018.

[24] A. Kaun, M. Männiste, and A. Liminga, "Mapping the automated decision-making landscape in swedish and estonian welfare state," CHANSE-funded research consortium "Automating Welfare – Algorithmic Infrastructures for Human Flourishing in Europe" (AUTO-WELF), Stockholm, Sweden, Tech. Rep., Aug. 2023, working Paper #02.

[25] Programmer3, "Multimodal child emotion for learning," https://www.kaggle.com/datasets/programmer3/multimodal-child-emotion-for-learning, 2020, accessed: 2025-01-28.

[26] R. A. Khan, A. Crenn, A. Meyer, and S. Bouakaz, "A novel database of children's spontaneous facial expressions (liris-cse)," 2018, preprint. [Online]. Available: https://arxiv.org/abs/1812.01555

[27] FunAudioLLM Team, "Sensevoice: Multilingual and multitask speech representation model," https://github.com/FunAudioLLM/SenseVoice, 2024, accessed: 2026-01-29.

[28] M. Sahidullah and G. Saha, "Mel-frequency cepstral coefficient and its applications: A review," *Digital Signal Processing*, vol. 69, pp. 1–21, 2017.

[29] R. Peri, S. O. Sadjadi, and D. Garcia-Romero, "Voxwatch: An open-set speaker recognition benchmark on voxceleb," *arXiv preprint arXiv:2307.00169*, 2023.

[30] Y. Wu, Y. Zhang, J. Song, and Z. Zhang, "Facial expression recognition based on facial landmark detection," *IEEE Access*, vol. 6, pp. 72 007–72 016, 2018.