Exploring Artificial Intelligence Challenges for Monitoring Cyber Child Abuse

Vita Santa Barletta[1,*,†], Danilo Caivano[1,†], Giovanni Dimauro[1,†], Francesca Mantini[1,*,†] and Massimiliano Morga[2,†]

[1]*Department of Computer Science, University of Bari Aldo Moro*
[2]*SER&Practices, Spin-off of the University of Bari Aldo Moro*

Abstract

The growing phenomenon of online production and dissemination of child sexual abuse material (CSAM - Child Sexual Abuse Material) poses increasingly complex challenges for law enforcement agencies in the area of online safety and child protection. Online Child Sexual Abuse (OCSA) emerges as a major threat in an increasingly digitalized world. It is estimated that more than one billion children between the ages of 2 and 17 are sexually abused each year, a figure that probably underestimates the true extent of the phenomenon, as most violence goes unreported to the relevant authorities. The situation is further exacerbated by the proliferation of dark web platforms that, lacking moderation, provide fertile ground for these crimes, making the task of tracing the origin of abuse extremely difficult and demonstrating how new methods of detection are needed. Based on these premises, this paper aims to conduct an in-depth analysis of the literature regarding current machine learning models for the detection of images, videos and texts containing CSA, evaluating their effectiveness, limitations, and ethical implications.

Keywords

Child Sexual Abuse (CSA) Detection, Artificial Intelligence, Age Detection, Pornography detection, Cybercrime

1. Introduction

The advent of the so-called *digital revolution* and its rapid growth has radically transformed the way we live and interact with society, offering unprecedented opportunities in terms of connectivity, education, and economic development. However, this progress brings with it an extreme dichotomy. While it is undeniable how it has brought prosperity and growth, one cannot ignore how the same progress has amplified and complicated certain pre-existing criminal phenomena, creating a stark contrast between technological advancement and the ethical issues that have inherently arisen from it [1].

Among these, *Child Sexual Abuse* (CSA) represents one of the most serious violations of human rights[1]. This crime has devastating and long-term consequences on the psycho-physical health of its victims, leading to what is referred to in psychopathology as *complex post-traumatic stress disorder (C-PTSD)*[2]. The psychological impact can vary significantly depending on elements such as personal resilience, the levels of stress the victim is subjected to, the family environment and, finally, the quality, timeliness, and effectiveness of the support received. This phenomenon has taken on alarming dimensions, evolving into increasingly complex and hard-to-counteract manifestations, including Online Child Sexual Abuse (OCSA) and the digital production and distribution of child sexual abuse material (CSAM).

Online child sexual abuse has become a significant concern with the rise of the Internet and social networking sites [2]. This form of abuse can occur through online grooming, sexual solicitation, and the

Joint Proceedings of IS-EUD 2025: 10th International Symposium on End-User Development, 16-18 June 2025, Munich, Germany.

*Corresponding author.
†These authors contributed equally.
vita.barletta@uniba.it (V. S. Barletta); danilo.caivano@uniba.it (D. Caivano); giovanni.dimauro@uniba.it (G. Dimauro); f.mantini@studenti.uniba.it (F. Mantini); m.morga@serandp.com (M. Morga)

[1]Rapporto Clusit 2025, https://clusit.it/rapporto-clusit

[2]A psychological condition that may develop as a result of prolonged and repeated traumatic experiences, such as childhood abuse, domestic violence, imprisonment or torture.

distribution of child abuse images [3]. The distinction between online and offline abuse is increasingly blurred, as abuse can begin offline and transition online through filming or photography. Vulnerable children are at risk, and research focuses on identifying these children and fostering resilience [3]. While online abuse is not necessarily more serious than offline abuse, social media behaviors may present significant risk factors for some children [4]. To address this issue, experts recommend increased awareness of online risks among children, improved training for law enforcement (Martellozzo, 2012), and the adoption of a public health approach to prevention and management [2, 4].

According to INTERPOL's most recent data, in 2022 the phenomenon of CSA reached epidemic proportions, with a 29% increase in CSAM reports compared to the previous year and with over 4.3 million images and videos analysed by their International Child Sexual Exploitation (ICSE) image and video database[3]. A joint report by INTERPOL and ECPAT International [5], based on the analysis of the information recorded for more than one million data records in the ICSE database, revealed the existence of an inversely proportional correlation between the age of the victims and the severity of the abuse: the younger the victim, the more severe the abuse suffered. The analysis of the images revealed that 84% of the images contained explicit sexual acts and that 65% of the unidentified victims were female, while 92% of the visible rapists were male. It was also found that when the victims are male, the perpetrated abuses are more violent and more often involve paraphilic themes. Particularly worrying is the finding that over 60 per cent of unidentified victims are prepubescent, a group that includes infants and very young children. This figure dispels the myth that sexual abuse mainly affects adolescents, revealing instead a much starker reality, in which the youngest are the main targets of these crimes.

UNICEF and the World Health Organization (WHO) estimate that globally, about 73 million boys and 150 million girls under 18 have suffered some form of sexual violence, with around 90% being committed by a person known to the victim, often a family member or authority figure.

The rapid spread of digital technologies and the increasingly early access of children to the internet have created new vulnerabilities. Like INTERPOL, the National Center for Missing and Exploited Children (NCMEC) has seen an exponential increase in online CSAM reports from 1 million in 2014 to over 65 million in 2023. This dramatic increase is partly attributable to increased detection capacity and awareness, but also reflects a real expansion of the phenomenon.

Finally, the COVID-19 pandemic has further negatively affected the problem. Social isolation, increased online time, and reduced adult supervision have created favorable conditions for attackers. Europol reported a 30% increase in cases of online child grooming[4] during the first months of the pandemic in 2020. It is therefore evident that the contrast to CSA and CSAM comes up against numerous technological and legal difficulties. Moreover, given the sensitive nature of the subject and the sheer volume of data, the figures released may underestimate the true extent of the problem. In order to take a step towards improving CSAM detection techniques, it is necessary to analyse existing methods to identify common techniques and methodological approaches, to propose a set of guidelines for the development of such systems, covering all the challenges mentioned above.

The literature review revealed how interest in this area of research has only recently developed. The limited number of publications does not reflect the actual scope of the problem, but rather the existence of strong cultural, social, and psychological barriers that have long hindered, and still limit, in-depth scientific investigation. No literature emerged from this analysis that perfectly matched the goal of our research

Therefore, the aim of this paper is reviewing the current methodologies and technologies proposed for the detection of Child Sexual Abuse Material. The results will help to identify new methods to be tested in the fight against this type of cybercrime.

The rest of the paper is organized as follows. Section 2 explains the research methodology adopted and data extraction method. In Section 3, the results of the systematic review are presented and discussed. Finally, Section 4 presents the conclusions and introduces future research activities.

[3]https://www.interpol.int/How-we-work/Databases/International-Child-Sexual-Exploitation-database

[4]Internet solicitation of a child through psychological manipulation to overcome resistance and gain trust for sexual abuse.

2. Research Methodology

A systematic literature review was conducted following the approach proposed by Kitchenham [6].

To ensure transparency and reliability of the entire review process, the following quality parameters were chosen:

- The presence of possible publication bias was verified using all the standard search strategies suggested by [6]: scanning of conference proceedings, scanning of grey literature, and contacting experts and researchers working in the area asking if they are aware of unpublished results.
- Exclusion of grey literature, such as dissertations, theses, posters, and unpublished works; only peer-reviewed journals and conference proceedings were included, as they guarantee a certified level of quality of the results.
- Rigor in following Kitchenham's review process [6], except for the selection of studies and quality assessment. For these tasks, it was necessary to apply more stringent specifications, given the significant number of research articles dealing with topics related to the proposed research questions, but not containing relevant information for CSAM detection.

2.1. Data Sources and Search Strategy

The foundations for a good systematic literature review are the exhaustive collection of the largest possible number of publications relevant to answering the identified research questions and the use, in the research itself, of a rigorous and impartial investigation methodology [7]. In the context of this analysis, the search strategy was designed to maximize coverage of studies related to structured methods for detecting material containing child sexual abuse.

The construction of the search string required careful terminological selection, based on the mapping of key concepts. Primary terms and semantic alternatives were considered. This strategy allowed for collecting the largest number of documents, while ensuring significant precision in identifying the most relevant studies. The specific terms included in the investigation comprise linguistic variants referring to Minors, Sexual Abuse, and Detection.

In particular, the term "child" was associated with commonly used synonyms, e.g., "baby" and "kid," along with terms that are currently used on social media to refer to children with malicious undertones, such as "lolita" and "kiddie." Similarly, the term "Sexual Abuse" was matched with synonyms such as "sexual assault" and other related terms, including their respective acronyms, e.g., "CSAM." Finally, regarding terms related to "detection," general concepts such as "identification," "classification," and "recognition" were included, avoiding adding more specific terminologies in order to keep the research field as broad as possible. For the same reason, the objects of detection were not specified, such as "images," "video," or "text," which risk excluding other potentially relevant types of material.

Based on these considerations, the following search string was formulated, using the Boolean operators AND and OR: *("minor" OR "child" OR "baby" OR "boy" OR "girl" OR "juvenile" OR "infant" OR "underage" OR "kiddie" OR "lolita" OR "school" OR "kid")* ***AND*** *("sexual abuse" OR "NSFW" OR "sexual assault" OR "pornography" OR "pervert" OR "prostitution" OR "sexual exploitation" OR "molestation" OR "harassment" OR "rape" OR "CSA" OR "sexual aggression" OR "sexual violence" OR "sexploitation" OR "CSAM" OR "CSEM")* ***AND*** *("detection" OR "identification" OR "classification" OR "recognition").*

The resources used to analyze the results of the formulated search string are: Scopus[5], ACM Digital Library[6], IEEE Xplore Digital Library[7], Semantic Scholar[8], Google Scholar[9], ResearchGate [10].

[5]https://www.scopus.com
[6]https://dl.acm.org/
[7]https://ieeexplore.ieee.org/
[8]https://www.semanticscholar.org/
[9]https://scholar.google.com
[10]https://www.researchgate.net

These were selected as they are regularly used by other reviews in this field, as well as by systematic reviews in general, e.g., [8]. Scopus, a multidisciplinary bibliographic database, includes over 70 million peer-reviewed literature records. ACM Digital Library is configured as a comprehensive computational archive, containing full-text papers and bibliographic literature in the computing and information technology domain. IEEE Xplore, a specialized digital library, provides access to more than five million publications in the engineering and technology fields. Semantic Scholar, implemented with Artificial Intelligence technologies, indexes over 200 million scientific documents. Google Scholar, used as an integrative tool compared to the previous one, operates as a web search engine to intercept potentially unretrieved literature. ResearchGate is added as an academic networking platform, facilitating the sharing of and access to scientific publications.

This list has allowed access to a wide collection of relevant resources including computer science conferences and journals, such as Journal of Visual Communication and Image Representation, Journal of Cyber Security and Mobility, Journal of Applied Security Research, and International Journal of Cyber Criminology, as well as conferences and journals in the health and legal fields (for example, Journal of Forensic and Legal Medicine and Journal of Forensic Sciences).

2.2. Study Selection and Quality Assessment

In continuity with Kitchenham's analysis [6], once potential relevant research articles have been identified, they must be evaluated and selected based on their relevance. In relation to the identified research questions, relevance is demonstrated by defining a set of selection criteria. As mentioned previously, to improve quality assessment, this systematic review has adopted a more detailed set of selection criteria than those proposed by [6].

Table 1 and Table 2 list the selection criteria defined as inclusion and exclusion criteria, respectively. The research articles analyzed will be *all and only* those that satisfy each inclusion criterion, while studies that will not be analyzed will be those that satisfy *at least one* of the exclusion criteria.

Inclusion criteria from IN1 to IN4 and exclusion criteria from EX1 to EX5 are related to more general scientific arguments. For example, including all research articles published after 2014 (IN1) ensures that the results concern the current generation of technology for CSAM detection, in line with other reviews in the same field. Including only research articles written in English (IN2) ensures the highest possible quality of results, as it is considered the universal language of science. Similarly, excluding research articles without available full text (EX5) (for example: only abstracts or titles are available online) is necessary to ensure that the articles used for the systematic review have sufficient and consistent data.

On the other hand, inclusion criteria from IN5 to IN10 and exclusion criteria from EX6 to EX12 are derived directly from the research questions addressed by this systematic review. For example, the inclusion of criterion IN5 ensures that research articles focus exclusively on combating Child Sexual Abuse. The inclusion of criterion IN6 ensures that studies are focused on describing the design and development of methods for CSAM detection. Similarly, the exclusion of all research articles not related to the themes of this review (EX6) allows focusing only on articles relevant to the defined research questions (for example, articles clearly unrelated to the scope of the systematic review based on title and abstract).

The exclusion of articles that present interventions not implemented with detection models (EX9) ensures the elimination of articles that propose interventions based on other types of solutions, such as e.g., the recognition of abuse at the physical level. Other inclusion criteria (IN7, IN8, IN9) and exclusion criteria (EX11, EX12) aim to narrow down the thematic areas analyzed, reducing the number of potentially misleading documents.

For each phase, specific actions performed are reported, while in the lower part, some of the inclusion and exclusion criteria are specified.

The five phases that characterized the process are described below:

1. **Phase 1: Digital Resource Search** - The search string was applied to digital resources. In the case of Semantic Scholar, it was adapted to "Child" "Sexual Abuse" "Detection," as it is an AI-based

Table 1
Adopted Inclusion Criteria

Code	Inclusion Criteria
IN1	Research articles published between 2014 and 2024
IN2	Research articles written in English
IN3	Research articles published in peer-reviewed journals or conferences
IN4	Research articles with full text available (not just title and abstract)
IN5	Research articles focused on Child Sexual Abuse
IN6	Research articles that include the description of the design and development of methods for CSAM detection
IN7	Research articles in the thematic areas of Medicine, Nursing, Health Professions, Biochemistry, and Genetics
IN8	Research articles in the thematic areas of Psychology, Social Sciences, and Neuroscience
IN9	Research articles in the thematic areas of Computer Science, Engineering, Decision Sciences, and Multidisciplinary
IN10	Articles with any type of access

Table 2
Adopted Exclusion Criteria

Code	Exclusion Criteria
EX1	Research articles published before 2014
EX2	Research articles not written in English (e.g., Chinese)
EX3	Research articles of the following types: surveys, reviews, systematic reviews, meta-analyses, editorials, dissertations, technical reports, student reports, posters, and unpublished works
EX4	Research articles that have duplicates
EX5	Research articles whose full text is not available or obtainable after a specific request to the authors
EX6	Research articles that do not deal with the topics of the systematic review
EX7	Research articles that focus on other conditions that affect children
EX8	Research articles that present interventions for the caregivers of people who are victims of CSA and not for individuals who have suffered CSA
EX9	Research articles that present interventions not implemented as detection models
EX10	Research articles that present interventions regarding other types of violence
EX11	Research articles in the thematic areas of Physics, Materials Science, Dentistry, Agriculture, Economics, and Business
EX12	Research articles in the thematic areas of Art, Pharmacy, Immunology, Mathematics

resource that does not allow the use of Boolean operators. The output of this phase consists of 46,449 research articles.

2. **Phase 2: Digital Resource Filtering** - The filters reflecting the exclusion criteria in Table 2 were applied to the output of phase 1, for example, the year of publication (EX1) or the chosen language (EX2). Based on the functionalities of digital resources, selection criteria were applied appropriately From the application of the above filters, 9,998 research articles resulted.
3. **Phase 3: Additional Semi-Automatic Filtering** - The research articles obtained as output from Phase 2 were collected in a single document in .xlsx format, reporting the list of authors, title, year of publication, EID, source (for example, name of the journal or conference proceedings), document type, publication status, DOI, and access type. In case of missing information (for example, sources), these were retrieved manually and inserted into the document.
   Since many digital libraries do not provide automatic filters related to all the exclusion criteria listed in Table 2, in this phase they were applied semi-automatically. For example, many research articles that met the criteria established in the second phase were often published two or even three times, thus, in the current phase, these duplicates were removed (according to EX4). Additionally, both research articles not published in peer-reviewed journals or conference proceedings, and simple research or similar contributions were excluded (according to EX3). This activity was conducted by analyzing the titles and sources of the retrieved research articles. Finally, in this phase, the authors of articles not fully available were contacted (according to EX5). The overall output derived from this selection was 1,874 research articles.
4. **Phase 4: Title, Abstract, and Conclusion Filtering** - The research articles obtained from phase 3 were subjected to a manual filter, analyzing titles, abstracts, and conclusions. During this process, documents that, while dealing with the topic of CSAM, addressed other types of detection, such as the recognition of CSA at the physical level, were excluded. This filter allowed the selection of only articles relevant to the specific focus desired for the review. The output of the manual filtering for relevance consists of 248 research articles.
5. **Phase 5: Full Text Filtering** - In phase 5, a further manual filter was applied to the filtered research articles, namely the analysis of the full texts of the research articles. In this phase, relevant information was extracted and added to the document, such as *Research objective, Methodologies used, Resulting metrics* (Accuracy, Precision, Recall, F1-score), *Learning type* (Supervised/Unsupervised), and *Number of networks employed.*
   In this phase, it was also established that, in case of multiple publications related to the same structured method for CSAM detection, only one research article would be included. Specifically, the article that provided the most complete and relevant information for the research questions under examination was selected. This decision is in line with the guidelines of Kitchenham's review process [6], according to which the inclusion of duplicates in the synthesis of a systematic review would tend to significantly polarize the results.

At the end of the entire study selection process, the final output consists of 25 research articles, each concerning the design and development of structured methods for detecting material containing child sexual abuse (Table 3).

The extracted data were subsequently compared with the aim of resolving any discrepancies, obtaining a single document for each selected research article.

2.3. Data Synthesis

The information extracted from the 25 selected papers was subjected to descriptive analysis through frequency analysis, in order to identify trends, patterns, and recurring characteristics in the examined literature. The statistical analysis included:

- Analysis of the most frequently used methodologies
- Comparative evaluation of model performance
- Identification of evolutionary trends in research

Table 3
Resulting research articles

ID	Authors	Title
[9]	Al-Nabki M.W.; Fidalgo E.; Alegre E.; Alaiz-Rodriguez R.	Short text classification approach to identify child sexual exploitation material
[10]	Al-Nabki M.W.; Fidalgo E.; Alegre E.; Aláiz-Rodríguez R.	File name classification approach to identify child sexual abuse
[11]	Oronowicz-Jaśkowiak W.; Kozłowski T.; Polańska M.; Wojciechowski J.; Wasilewski P.; Ślęzak D.; Kowaluk M.	Using expert-reviewed CSAM to train CNNs and its anthropological analysis
[12]	Ngo V.M.; Gajula R.; Thorpe C.; Mckeever S.	Discovering child sexual abuse material creators' behaviors and preferences on the dark web
[13]	Ngo V.M.; McKeever S.; Thorpe C.	Identifying Online Child Sexual Texts in Dark Web through Machine Learning and Deep Learning Algorithms
[14]	Fauzi M.A.; Wolthusen S.; Yang B.; Bours P.; Yeng P.	Identifying Sexual Predators in Chats Using SVM and Feature Ensemble
[15]	Rondeau J.; Deslauriers D.; Howard III T.; Alvarez M.	A deep learning framework for finding illicit images/videos of children
[16]	Spalazzi L.; Paolanti M.; Frontoni E.	An offline parallel architecture for forensic multimedia classification
[17]	Gangwar A.; González-Castro V.; Alegre E.; Fidalgo E.	AttM-CNN: Attention and metric learning based CNN for pornography, age and Child Sexual Abuse (CSA) Detection in images
[18]	Vitorino P.; Avila S.; Perez M.; Rocha A.	Leveraging deep neural networks to fight child pornography in the age of social media
[19]	De Souza Viana T.S.; De Oliveira M.; Da Silva T.L.C.; Júnior M.S.R.F.; Gonçalves E.J.T.	Textual analysis for the protection of children & teenagers in social media: Classification of inappropriate messages for children & teenagers
[20]	Sae-Bae N.; Sun X.; Sencar H.T.; Memon N.D.	Towards automatic detection of child pornography
[21]	Pereira M.; Dodhia R.; Anderson H.; Brown R.	Metadata-Based Detection of Child Sexual Abuse Material
[22]	Bennabhaktula G.S.; Alegre E.; Karastoyanova D.; Azzopardi G.	Camera model identification based on forensic traces extracted from homogeneous patches
[23]	Reinders S.; Guan Y.; Ommen D.; Newman J.	Source-anchored, trace-anchored, and general match score-based likelihood ratios for camera device identification
[24]	Sarkar B.N.; Barman S.; Naskar R.	Blind Source Camera Identification of Online Social Network Images Using Adaptive Thresholding Technique
[25]	Timmerman D.; Bennabhaktula G.S.; Alegre E.; Azzopardi G.	Video Camera Identification from Sensor Pattern Noise with a Constrained ConvNet
[26]	Bennabhaktula G.S.; Alegre E.; Karastoyanova D.; Azzopardi G.	Device-based Image Matching with Similarity Learning by Convolutional Neural Networks that Exploit the Underlying Camera Sensor Pattern Noise
[27]	Meij C.; Geradts Z.	Source camera identification using Photo Response Non-Uniformity on WhatsApp
[28]	Steinebach M.; Berwanger T.; Liu H.	Image Hashing Robust Against Cropping and Rotation
[29]	Singh P.; Farid H.	Robust homomorphic image hashing
[30]	Garcia J.	A Fragment Hashing Approach for Scalable and Cloud-Aware Network File Detection
[31]	Kulshrestha A.; Mayer J.	Identifying harmful media in end-to-end encrypted communication: Efficient private membership computation
[32]	Guerra E.; Westlake B.G.	Detecting child sexual abuse images: Traits of child sexual exploitation hosting and displaying websites
[33]	Ibrahim A.; Valli C.	Image similarity using dynamic time warping of fractal features

3. Analysis and Discussion of the Results

In the following graph, the distribution of research conducted over a ten-year range is reported, consistent with the IN1 filter adopted in the previous chapter. The dashed line represents and demonstrates the growing interest, although still limited, in the search for structured methods for the detection of material containing Child Sexual Abuse. Below is the analysis and synthesis of the documents resulting from the previously described research.

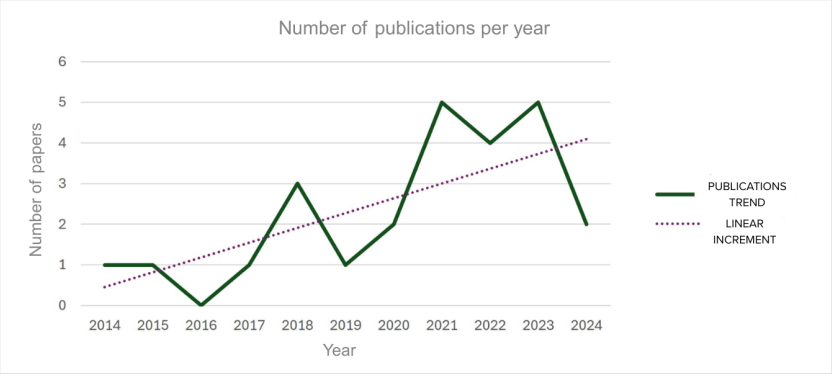

Figure 1: Distribution of 25 research studies over the ten years analyzed

The systematic analysis of the literature revealed three predominant macro-areas: Advanced Hashing Cryptographic techniques, Source Camera Identification (SCI) techniques, and methodologies based on Machine Learning algorithms.

Hashing is a cryptographic technique that represents a linear function capable of transforming an input of arbitrary size into an output of fixed and predetermined length, called a fingerprint or hash. In the context of modern cryptography, hashing algorithms play a fundamental role in computer security processes, as they make it possible to verify that a digital document is authentic, has not been modified, and can be traced back to its author with certainty. In the specific field of Content Security detection (as in the case of CSAM), hashing techniques allow for the rapid identification and classification of explicit content by comparing their fingerprints with specialised databases, guaranteeing a fast and effective screening process.

The *Source Camera Identification* (SCI) is a technique used in digital forensics that aims to identify the source of an image or video by tracing it back to the original camera or capture device. This technique takes advantage of the fact that devices such as cameras, smartphones, or tablets possess distinctive physical and electronic characteristics that generate recurring 'imperfections' or 'noise' in the images produced. It is precisely these imperfections that can be considered as identifying 'signatures' of the device. Typical analysis methodologies are based on techniques such as Pattern Noise Analysis, which studies the electronic noise of the image sensor, and the identification of Fixed Pattern Noise (FPN), i.e. constant structural defects in the sensor. Experts analyze the chrominance and luminance components, extracting features using machine learning algorithms. However, research in this field faces several challenges, such as the variability of acquisition conditions, image degradation, the complexity of identification algorithms, and the need for constantly updated reference databases.

Last but not least are the methodologies based on *Machine Learning (ML)*, a subset of artificial intelligence (AI) that deals with creating systems capable of learning or improving their performance based on the data used. Al-Nabki et al.[10] emerged as an alternative to the classification and identification of CSEM by LEAs, procedures which currently take place through manual inspection, which is why, in most cases, they are not feasible in the available time. One option for detection consists of analyzing the file names stored on the hard disk of the suspected person, searching the text for references to CSEM. However, due to the peculiarities of the names themselves, namely their length and their being deliberately distorted by owners through the use of obfuscated words and user-defined naming patterns, current file name classification methods suffer from a low recall rate, essential in counteracting this

problem.

This study was recently revisited by the same research group in [9] with the aim of further accelerating file identification, focusing not only on the analysis of names but also of their absolute paths, continuing to leave out their content.

The proposal by Pereira et al. [21] also addresses the ethical and legal challenges associated with acquiring images for training machine learning models, presenting a detection framework based on file metadata. This approach stands out because metadata does not constitute a record of a crime, which allows circumventing the legal restrictions that would normally hinder the collection of sensitive data.

The study by Oronowicz-Jaśkowiak et al.[11] stands out for a rigorous methodological approach in training convolutional neural networks, using explicit images previously examined by forensic experts in anthropology. This research fills important gaps in existing methodologies, addressing three fundamental limitations: the absence of expert annotations, the lack of models trained with real explicit content involving minors, and insufficient justification of classification decisions. Instead, Ngo et al.[13, 12] address the issue of sharing child sexual abuse material (CSAM) within dark web forums, an environment that offers a high degree of anonymity, making it difficult for law enforcement to identify the criminals involved. The research analyzed and manually labeled a massive dataset comprising over 353,000 posts generated by 35,400 distinct users, operating in 118 different languages across eight forums. The study by Fauzi et al.[14] focuses on creating an automatic system capable of detecting potentially predatory conversations and distinguishing between the predator and victim profiles. This resulted in the development of a two-phase approach that integrates linguistic and behavioral analysis techniques, analyzed in two different stages.

Spalazzi et al.[16] address the growing challenge represented by the volume of heterogeneous multimedia evidence presented for digital forensic analysis, highlighting the need to apply big data technologies, cloud-based forensic services, and Machine Learning (ML) techniques.

Regarding the tasks of automatic age estimation and nudity detection, Rondeau et al.[15] highlights how modern machine learning algorithms can predict with surprising accuracy the presence of a minor or explicit content. The research introduces an innovative framework to automatically identify sexually exploitative images and videos of minors, merging separate models for apparent age estimation and nudity detection. Specifically, two CNNs are tested: DenseNet-161, pre-trained on ImageNet and then further refined on specific datasets for age classification, and OpenNSFW, also pre-trained, for nudity detection.

As in the previous study, also in the research by Gangwar et al.[17] the problem of automatic CSA detection has been divided into two sub-problems, each with a specialized network: the detection of pornographic content and the age classification of a person as minor or adult. An innovative convolutional neural network (CNN) architecture, called AttM-CNN, that integrates an attention mechanism and metric learning, has been proposed. This architecture is designed to improve the model's ability to focus on the most relevant features of images, thus optimizing classification performance.

Sae-Bae et al.[20] present a system for detecting images containing CSA, structured around two fundamental modules: the first, dedicated to the detection of explicit images (Explicit Image Detection, EID), and the second, focused on the detection and classification of child faces. The methodology adopted for classification is based on LIBSVM, a powerful support tool for vector machines, which allows optimizing the system's performance in the recognition and classification of images. The novelty of the proposed technique lies in the adoption of a rapid and robust filter to discriminate skin color from the rest of the photo, along with a new set of facial features that significantly improve the identification of child faces. Vitorino et al.[18] analyze the evolution of the phenomenon of child sexual abuse over the last two decades, highlighting how the modes of generating, distributing, and possessing images have radically changed, moving from reserved and offline exchanges to a massive network of contacts and data sharing. In the process, a convolutional neural network (CNN) architecture known as GoogLeNet is used, chosen for its consolidated performance in image classification tasks, which includes several processing modules called "inception", which allow extracting features at different scales and levels of complexity.

De Oliveira et al.[19] address the risks of Internet use by minors, highlighting how privacy and pro-

tection from indiscriminate exposure are frequently overlooked issues, leaving young people vulnerable to potential sexual predators. The research is based on a previous study that proposed a tool to detect potentially dangerous conversations, which was based on the analysis of minors' behavior. The authors, recognizing that such an approach did not comprehensively address the textual analysis of exchanged messages, limiting itself to a superficial analysis, have proposed a new version.

4. Conclusions and Future Work

This paper stems from the desire and need to explore a complex and dramatic phenomenon, Child Sexual Abuse, which leads to the uncontrolled spread of Child Sexual Abuse Material (CSAM) on the internet, as well as cases of Grooming. Despite the growing attention dedicated to research in this field today, it is undeniable that much remains to be done both in the state of the art and in practice, given the proportions of the problem, and there are many challenges associated with it from technical, legal, and ethical perspectives.

This systematic review seeks to outline some primary guidelines for designing methods for the detection of material containing CSA, offering a broad overview of current methodologies and technologies proposed for the detection of Child Sexual Abuse Material at the state of the art.

Machine learning methods prove to be fast, effective, and capable of analyzing large amounts of data while maintaining high performance. On the other hand, however, they raise important issues such as respecting the privacy of victims and being strictly dependent on the use of explicit material during training, a constraint that binds them closely to the need for a joint project with law enforcement. Valid solutions to "circumvent" the legal restrictions that would normally hinder the collection of sensitive data are those based on the analysis of file metadata that do not constitute a record of a crime in themselves. Examples of studies on detection through metadata are those proposed in [9], [10], and [21], which focus on the study of file names (FNC) and file paths (FPC) stored on the suspect's hard drive, looking for references to CSEM in the text.

In conclusion, it would be appropriate to emphasize the added value of the results of this work, both in terms of identifying the main gaps in research on this topic, and outlining the future research agenda to fill these gaps. Subsequent research in this field should focus on developing methodological solutions based on the proposed guidelines, which ensure tangible and practical support for designing these systems more effectively, ethically, safely, and sustainably. It is important to emphasize how these advances should not be limited to the sterile academic sphere but should also extend to public dynamics, promoting collaboration and the implementation of innovative approaches.

Acknowledgments

This work was partially supported by the following projects: SERICS - "Security and Rights In the CyberSpace - SERICS" (PE00000014) under the MUR National Recovery and Resilience Plan funded by the European Union - NextGenerationEU; Accordo Quadro CrASte - "Cyber Academy for Security and Intelligence".

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] M. T. Baldassarre, V. S. Barletta, D. Caivano, D. Raguseo, M. Scalera, Teaching cyber security: The hack-space integrated model, volume 2315, 2019. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85061370504&partnerID=40&md5=e8da8bde8df7b4a276e5517e34136832.

[2] E. Martellozzo, Chapter 4 - online child sexual abuse, in: I. Bryce, Y. Robinson, W. Petherick (Eds.), Child Abuse and Neglect, Academic Press, 2019, pp. 63–77. URL: https://www.sciencedirect.com/science/article/pii/B9780128153444000040. doi:https://doi.org/10.1016/B978-0-12-815344-4.00004-0.

[3] C. May-Chahal, E. Kelly, 1introduction, in: Online Child Sexual Victimisation, Policy Press, 2020. URL: https://doi.org/10.1332/policypress/9781447354505.003.0001. doi:10.1332/policypress/9781447354505.003.0001. arXiv:https://academic.oup.com/policy-press-scholarship-online/book/0/chapter/264465636/chapter-ag-pdf/44543872/book_31694_section_264465636.ag.pdf.

[4] E. Quayle, N. Koukopoulos, Deterrence of online child sexual abuse and exploitation, Policing: A Journal of Policy and Practice 13 (2018) 345–362. URL: https://doi.org/10.1093/police/pay028. doi:10.1093/police/pay028. arXiv:https://academic.oup.com/policing/article-pdf/13/3/345/29198223/pay028.pdf.

[5] I. ECPAT International, Towards a global indicator on unidentified victims in child sexual exploitation material, Technical Report, INTERPOL, ECPAT, 2018.

[6] B. Kitchenham, S. Charters, Guidelines for performing systematic literature reviews in software engineering 2 (2007).

[7] V. S. Barletta, F. Caruso, T. Di Mascio, A. Piccinno, Serious games for autism based on immersive virtual reality: A lens on methodological and technological challenges, in: M. Temperini, V. Scarano, I. Marenzi, M. Kravcik, E. Popescu, R. Lanzilotti, R. Gennari, F. De La Prieta, T. Di Mascio, P. Vittorini (Eds.), Methodologies and Intelligent Systems for Technology Enhanced Learning, 12th International Conference, Springer International Publishing, Cham, 2023, pp. 181–195.

[8] M. Gusenbauer, N. Haddaway, Which academic search systems are suitable for systematic reviews or meta-analyses? evaluating retrieval qualities of google scholar, pubmed and 26 other resources [open access], Research Synthesis Methods 11 (2020) 181–217. doi:10.1002/jrsm.1378.

[9] W. Al Nabki, E. Fidalgo, E. Alegre, R. Alaiz, Short text classification approach to identify child sexual exploitation material, Scientific Reports 13 (2023). doi:10.1038/s41598-023-42902-8.

[10] W. Al Nabki, E. Fidalgo, E. Alegre, R. Alaiz, File name classification approach to identify child sexual abuse, 2020, pp. 228–234. doi:10.5220/0009154802280234.

[11] W. Oronowicz-Jaśkowiak, T. Kozłowski, M. Polanska, J. Wojciechowski, P. Wasilewski, D. Ślęzak, M. Kowaluk, Using expert-reviewed csam to train cnns and its anthropological analysis, Journal of Forensic and Legal Medicine 101 (2023) 102619. doi:10.1016/j.jflm.2023.102619.

[12] V. Ngo, R. Gajula, C. Thorpe, S. Mckeever, Discovering child sexual abuse material creators' behaviors and preferences on the dark web, Child Abuse Neglect 147 (2023) 106558. doi:10.1016/j.chiabu.2023.106558.

[13] V. Ngo, S. Mckeever, C. Thorpe, Identifying online child sexual texts in dark web through machine learning and deep learning algorithms, 2024.

[14] M. Fauzi, S. Wolthusen, B. Yang, P. Bours, P. Yeng, Identifying sexual predators in chats using svm and feature ensemble, 2023, pp. 1–6. doi:10.1109/ETNCC59188.2023.10284950.

[15] J. Rondeau, D. Deslauriers, T. Howard, M. Alvarez, A deep learning framework for finding illicit images/videos of children, Machine Vision and Applications 33 (2022). doi:10.1007/s00138-022-01318-6.

[16] L. Spalazzi, M. Paolanti, E. Frontoni, An offline parallel architecture for forensic multimedia classification, Multimedia Tools and Applications 81 (2022). doi:10.1007/s11042-021-10819-x.

[17] A. Gangwar, V. González-Castro, E. Alegre, E. Fidalgo, Attm-cnn: Attention and metric learning based cnn for pornography, age and child sexual abuse (csa) detection in images, Neurocomputing 445 (2021). doi:10.1016/j.neucom.2021.02.056.

[18] P. Vitorino, S. Avila, M. Perez, A. Rocha, Leveraging deep neural networks to fight child pornography in the age of social media, Journal of Visual Communication and Image Representation 50 (2017). doi:10.1016/j.jvcir.2017.12.005.

[19] M. De Oliveira, T. Viana, T. Coelho da Silva, E. Gonçalves, M. Jr, Textual analysis for the protection of children and teenagers in social media: Classification of inappropriate messages for children

and teenagers, 2017. doi:10.5220/0006370606560662.
[20] N. Sae-Bae, X. Sun, T. Sencar, N. Memon, Towards automatic detection of child pornography, 2014 IEEE International Conference on Image Processing, ICIP 2014 (2015) 5332–5336. doi:10.1109/ICIP.2014.7026079.
[21] M. Pereira, R. Dodhia, H. Anderson, R. Brown, Metadata-based detection of child sexual abuse material, IEEE Transactions on Dependable and Secure Computing PP (2023) 1–13. doi:10.1109/TDSC.2023.3324275.
[22] G. Bennabhaktula, E. Alegre, D. Karastoyanova, G. Azzopardi, Camera model identification based on forensic traces extracted from homogeneous patches, Expert Systems with Applications 206 (2022) 117769. doi:10.1016/j.eswa.2022.117769.
[23] S. Reinders, Y. Guan, D. Ommen, J. Davidson, Source-anchored, trace-anchored, and general match score-based likelihood ratios for camera device identification, Journal of Forensic Sciences 67 (2022). doi:10.1111/1556-4029.14991.
[24] B. Sarkar, S. Barman, R. Naskar, Blind Source Camera Identification of Online Social Network Images Using Adaptive Thresholding Technique, 2021, pp. 637–648. doi:10.1007/978-981-15-7834-2_59.
[25] D. Timmerman, S. Bennabhaktula, E. Alegre, G. Azzopardi, Video camera identification from sensor pattern noise with a constrained convnet, 2020. doi:10.48550/arXiv.2012.06277.
[26] G. Bennabhaktula, E. Alegre, D. Karastoyanova, G. Azzopardi, Device-based image matching with similarity learning by convolutional neural networks that exploit the underlying camera sensor pattern noise (2020). doi:10.48550/arXiv.2004.11443.
[27] C. Meij, Z. Geradts, Source camera identification using photo response non-uniformity on whatsapp, Digital Investigation 24 (2018). doi:10.1016/j.diin.2018.02.005.
[28] M. Steinebach, T. Berwanger, H. Liu, Image hashing robust against cropping and rotation, Journal of Cyber Security and Mobility (2023). doi:10.13052/jcsm2245-1439.1221.
[29] P. Singh, H. Farid, Robust homomorphic image hashing, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2019. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85113838951&partnerID=40&md5=024473edc4f5619b1f642d58f48ff490.
[30] J. Garcia, A fragment hashing approach for scalable and cloud-aware network file detection, 2018, pp. 1–5. doi:10.1109/NTMS.2018.8328746.
[31] A. Kulshrestha, J. Mayer, Identifying harmful media in end-to-end encrypted communication: Efficient private membership computation, 2021, p. 893 – 910. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85114489270&partnerID=40&md5=707478615c4762e128ebc2f49c1f41f9.
[32] E. Guerra, B. Westlake, Detecting child sexual abuse images: Traits of child sexual exploitation hosting and displaying websites, Child Abuse Neglect 122 (2021) 105336. doi:10.1016/j.chiabu.2021.105336.
[33] A. Ibrahim, C. Valli, Image similarity using dynamic time warping of fractal features (2015). doi:10.4225/75/57b3fe44fb890.