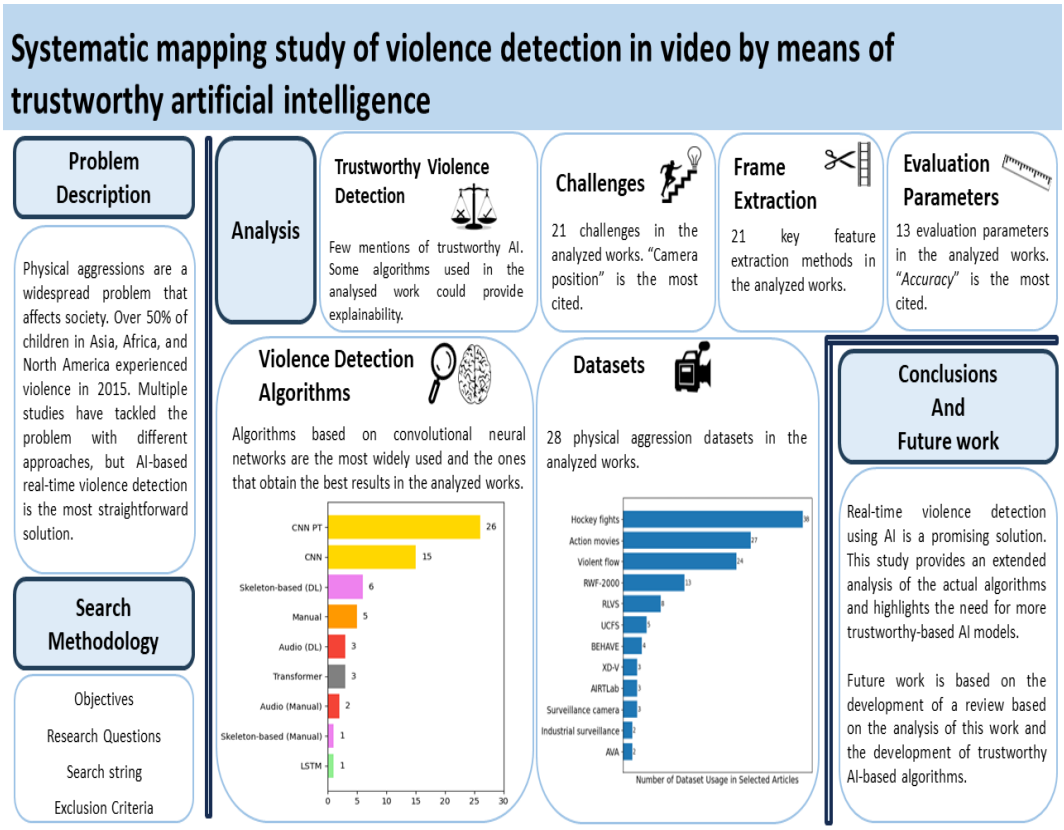
Graphical Abstract

Systematic mapping study on violence detection in video by means of trustworthy artificial intelligence

Pablo Negre, Ricardo S. Alonso, Javier Prieto, Nhan Cach Dang, Juan M. Corchado



Highlights

Systematic mapping study on violence detection in video by means of trustworthy artificial intelligence

Pablo Negre, Ricardo S. Alonso, Javier Prieto, Nhan Cach Dang, Juan M. Corchado

- There is a limited body of work on trustworthy AI, nevertheless, there are algorithms that provide explainability.
- Robustness is the most mentioned trustworthy artificial intelligence pillar in violence detection.
- Algorithms based on CNN and a combination of CNN + LSTM achieve the best results in violence detection.
- Since 2018, 22 new datasets have been created for the detection of physical aggression.
- 21 violence detection challenges have been gathered, *camera position* is the most referenced.

Systematic mapping study on violence detection in video by means of trustworthy artificial intelligence

Pablo Negre[a], Ricardo S. Alonso[b,c], Javier Prieto[a], Nhan Cach Dang[d], Juan M. Corchado[a,b]

[a]*BISITE Research Group, Universidad de Salamanca, Patio de Escuelas, 1, Salamanca, 37008, Castilla y León, Spain*
[b]*AIR Institute, Paseo de Belen 9A, Valladolid, 47011, Castilla y León, Spain*
[c]*UNIR (International University of La Rioja), Av. de la Paz, 137, Logroño, 37008, La Rioja, Spain*
[d]*Department of Information Technology, Ho Chi Minh City University of Transport (UTH-HCMC), Ho Chi Minh, 70000, Vietnam*

Abstract

Physical aggression is a serious and widespread problem in society. Studies show that in 2015 alone, at least 50% of children living in Asia, Africa and North America experienced violence. Real-time violence detection, powered by artificial intelligence, offers a direct and efficient solution that can save lives whilst reducing the need for extensive human supervision. There is a growing concern regarding the development of trustworthy artificial intelligence and this is reflected in the reports of the organizations of the European Union. Thus, this systematic mapping study delves into violence detection in videos by means of trustworthy artificial intelligence, with a specific focus on physical aggression. An extensive analysis of 63 selected articles published between June 2020 and June 2023 has been carried out and the findings have been compiled, including the 21 challenges associated with violence detection, 28 datasets on physical aggression, and 13 evaluation methods. Furthermore,

Email addresses: pablo.negre@usal.es (Pablo Negre), ralonso@air-institute.com (Ricardo S. Alonso), ricardoserafin.alonso@unir.net (Ricardo S. Alonso), javierp@usal.es (Javier Prieto), tucach@hcmutrans.edu.vn (Nhan Cach Dang), corchado@usal.es (Juan M. Corchado), corchado@air-institute.com (Juan M. Corchado)

URL: ORCiD (Pablo Negre), ORCiD (Ricardo S. Alonso), ORCiD (Javier Prieto), ORCiD (Nhan Cach Dang), ORCiD (Juan M. Corchado)

the optimal results of algorithms based on convolutional neural networks are exemplified. Finally, it is evidenced that the trustworthiness of violence detection algorithms is still limited, despite the fact that robust and explainable algorithms are available.

Keywords: Violence detection, Physical aggression, Trustworthy artificial intelligence, Video surveillance, Explainable artificial intelligence

---

1. Introduction

Physical assault is a serious and widespread problem in society, affecting people all over the world. It is a problem that affects all of us; not only the direct victims, but also the families, and society as a whole. Aggression hinders development on a national scale and even impedes economic growth (shopping, travel, tourism, etc.) [1] [2]. Moreover, the mental health of citizens suffers because of a persistent sense of insecurity [3]. In the American Academy of Pediatrics, Hillis et al. [4] state that 50% or more, of children in Asia, Africa and North America experienced violence in 2015, and that globally more than a half of all children, that is, 1 billion children, aged between 2 and 17 years old, experienced violence. An ActionAid study reported that 79% of Indian women, 86% of Thai women, 89% of Brazilian women and 75% of London women experienced harassment or violence in public [5].
To address the existence of violence, some studies have tried to comprehend the origins of and motives for physical assault [6] [7] [1]. Other studies have approached this problem by trying to relate street population density and the urban landscape to crime rates [8] [9]. However, both approaches are indirect and offer no immediate solutions. On the other hand, the increasing use of artificial intelligence-based systems has led citizens and large organisations to raise the alarm. Several major organisations such as the European Commission [10], the International Organization for Standardization (ISO) [11] and many others have strived to create definitions, standards and legislation in this regard. In particular, the European Union has published a report on artificial intelligence [10] which defines trustworthy AI, among many other concepts. In its definition, trustworthy AI is underpinned by three components:

- Lawful: ensuring compliance with all relevant laws and regulations.

- Ethical: Prioritizing principles and values that uphold moral considerations and standards.
- Robust: both technically and socially, to ensure that AI systems, even if well-intentioned, do not cause accidental harm.

The report further defines how each of these components is necessary but not sufficient for AI to be trustworthy. The inherent opacity (also known as the black-box concept) of most artificial intelligence algorithms hinders the comprehensibility of their processes and results, and therefore of their trustworthiness [12]. Explainable artificial intelligence has received much attention from the research community in recent years. Explainability of Artificial Intelligence (XAI) refers to the ability to understand and explain how an artificial intelligence model makes decisions; where lack of transparency in AI models can be a barrier to their acceptance and adoption, especially in critical applications where it is important to understand the reasoning behind automated decisions[10]. Speith [13] exemplified why aggression detection systems must use explainable artificial intelligence (XAI), stating that otherwise there would be a risk of an algorithm taking biased decisions, on the basis of the person's skin colour, for instance.
The use of artificial intelligence makes it possible to address the problem of aggression directly, enabling the real-time detection of citizens who are being physically assaulted. This is made possible by: the increased use of security images and videos [14] [15], the development of big data platforms [16] [17] [18] [19] and the development of algorithms that analyse images and video through the use of AI [20] [21] [22]. Moreover, against the backdrop of continually advancing AI capabilities, there is a growing concern regarding the lack of explainability and trustworthiness of AI. Thus, it is essential that a system for the detection of aggression be trustworthy, especially because of its role in the protection of citizens.
In an effort to contribute to the advancement of this field, this article presents a systematic mapping study on physical aggression detection in video, by means of trustworthy artificial intelligence. The rest of this article is organized as follows: Section 2 compiles state-of-the-art review papers from the last 2 years dealing with violence detection in video. To date, a relatively small number of review papers have been published on this topic, and many did not consider key aspects in the aggression detection process. This justifies the need for an updated systematic mapping study that is as complete as

possible. Section 3 describes the research methodology followed throughout the conducted research, which includes: the description of objectives, motivation and relevance of the study, research questions as well as the whole process of article selection in several databases, as well as the definition of the exclusion filters for the selected articles; among other sections. Section 4 describes the mapping process of the selected articles as well as the number of articles filtered after applying each exclusion filter to the results obtained in the selected databases. Section 5 analyzes each part of the violence detection process, where Section 5.9 covers the use of trustworthy artificial intelligence techniques in selected articles. Finally, Section 6 draws conclusions from the conducted systematic mapping study and outlines future lines of research.

2. Related Work

This section provides an overview of related review articles on physical aggression detection in video using AI, specifically whether they have taken into account the use of trustworthy AI techniques. Only reviews published between 2021 and 2023 have been taken into account. The works found and brief information about its contents is shown at shown in Table 1 and Table 2; where the contents of each of the reviews are shown.

As can be seen, the reviews do not address all parts of violence detection. From among all the summarised reviews, the most comprehensive review article, which is also the only systematic review among the identified literature, is Omarov et al. [23]. However, the articles collected are from between 2015 and 2021 and none of the identified reviews mention the use of explainable artificial intelligence (XAI) or the importance of trustworthiness in physical assault detection. Therefore, it is considered relevant and necessary to produce an updated SMS with recent articles and comprehensive content on the whole field of AI video violence detection.

Table 1: Summary of related review articles addressing physical aggression. Part 1.

Cite	Review Title	Year	SMS	Related Work	Research Methodology	Challenges	Evaluation Parameters	Dataset						
								Name	Ref.	Year	Clips Number	Type	Quality	Citations
[24]	A survey of video violence detection.	2021				✓		✓		✓	✓	✓		
[25]	State-of-the-Art Violence Detection Techniques: A review.	2022			✓			✓	✓	✓			✓	
[26]	Violence Detection in Videos Using Deep Learning: A Survey.	2022		✓		✓								
[27]	A review on video violence detection approaches.	2022				✓		✓		✓	✓	✓	✓	
[23]	State-of-the-art violence detection techniques in video surveillance security systems: a systematic review.	2022	✓	✓	✓	✓	✓	✓	✓	✓			✓	✓

Table 2: Summary of related review articles addressing physical aggression. Part 2.

Cite	Trad. Methods					DL Methods					F.E. Focus
	Pre-processing	Object Detection	Feature Extraction	Classification	Dataset	Pre-processing	Object Detection	Feature Extraction	Classification	Dataset	
[24]			✓	✓				✓	✓		
[25]		✓	✓				✓	✓			✓
[26]			✓					✓	✓	✓	✓
[27]	✓		✓		✓	✓		✓		✓	
[23]		✓	✓	✓	✓		✓	✓	✓	✓	✓

3. Research Methodology

This section describes the methodology that has been employed to carry out this systematic mapping study (SMS). Systematic mapping studies (SMS) have gained in importance in recent years due to their vital contribution to the field of scientific research. Kitchenham et al. [28] provided an overview of systematic review practices in software engineering, identifying common challenges and offering recommendations for improving the quality of systematic reviews. The improved understanding of systematic review practices has contributed to the wider adoption of SMS as an initial approach to mapping and ranking existing knowledge in specific research areas. On the other hand, Petersen, et al. [29] established specific methodological guidelines for conducting SMS in software engineering, which has allowed for greater standardisation and rigour in the conduct of these studies.
An SMS is a research methodology used to classify, analyse and identify a field of interest [30]. The general objective of an SMS is to establish the range of study of a given field of research by analysing the frequency of publications according to various categories in a schematic way. The particular objectives of an SMS are as follows:

- To identify the main areas of research and relevant topics in a specific field and classify existing scientific publications on the topic of study.
- To scrutinize the temporal distribution of research and its evolution over time.
- To identify the methodologies and approaches most commonly used in research on the topic.
- To uncover possible gaps in knowledge and under-researched areas.
- To provide an overview of the current state of knowledge in the field of study.
- To detect possible trends and emerging themes that could be the subject of future research.

These particular objectives have been pursued in a methodical fashion in this systematic review, supported with classification schemes, mapping schemas and other graphs. Following the methodology proposed by Kitchenham et al. [28], This section is divided into three subsections: Planning, Study Development, and Report. Each section also contains its respective subsections.

3.1. Planning

Section 3.1 aims to establish the basis for the SMS; it is divided into three subsections: Motivation and Relevance, Objective and Research Questions.

3.1.1. Motivation and Relevance

The motivation for this SMS is to compile the latest studies on the use of artificial intelligence techniques in the detection of physical aggression in video. It is examined whether these studies addressed trustworthy artificial intelligence. Given the scarcity of research on trustworthy artificial intelligence, in contrast to the relatively broader studies on explainable artificial intelligence (XAI). Thus, XAI is emphasised as one of the pillars of ethical AI which is in turn one of the pillars of trustworthy AI. The relevance of this systematic review lies in the fact that in the last two years, to the best of our knowledge, there have only been 5 reviews on the detection of physical aggression in video [24] [25] [26] [27] [23] of which only one is an SMS [23]. Moreover, none of systematic reviews address the trustworthiness of the studied algorithms or their explainability. For all these reasons, this SMS is considered to be of relevance in assessing the latest studies published in the growing field of physical aggression detection in video [23], with a special focus on the application of another growing field, namely, that of trustworthy or explainable artificial intelligence [10] [11].

3.1.2. Objective

The objective of this SMS is to compile the latest studies addressing the detection of physical aggression in video, with a special emphasis on the use or mention of trustworthy and/or explainable artificial intelligence. This systematic review provides an updated perspective on the advances in this field of research. It compiles and describes in detail the datasets that have been used, as well as the algorithms used in each part of aggression detection. This review is to make it possible to conclude which combination of techniques is the most optimal in each case. Finally, this SMS tries to identify which trustworthy and/or explainable artificial intelligence techniques have been mentioned or used in the selected works.

3.1.3. Research Questions

Research questions in a systematic mapping study (SMS) help to focus and guide the research, enabling the identification and classification of relevant studies to map and understand the current state of knowledge in a

specific research field. Two key research questions have been selected, which are presented below:

- What are the latest state-of-the-art techniques applied in the detection of physical aggression in video?
- Which trustworthy and/or explainable artificial intelligence techniques have been used or discussed in the state of the art literature on aggression detection?

3.2. Study Development

This subsection develops the fundamental research question and it is divided into two subsections, under the headlines of "Search Strategy" and "Inclusion/Exclusion Criteria". The search strategy describes how relevant information is collected, using specific databases and terms. The set out inclusion/exclusion criteria serve as guidelines for selecting relevant studies and discarding unsuitable ones, ensuring the quality and trustworthiness of the results obtained in this SMS.

3.2.1. Search Strategy

A key part of an SMS is the development of a search strategy. Both Kitchenham et al. [28] and Petersen, et al. [29] proposed the PICO (Population, Intervention, Comparison, and Outcomes) strategy for conducting SMS, which consists of constructing search strings from keywords in the research questions. Below we define what the terms that make up PICO mean in our SMS.

- Population: In this SMS, the population comprises the collection of the research papers being analyzed.
- Intervention: Refers to the techniques, models or solutions dealing with the detection of physical aggression in video.
- Comparison: The selected research is contrasted according to different categories.
- Outcomes: Not applicable to this SMS.

Given that there are numerous alternative terms for the detection of physical aggression, the created string chain is as complete as possible, including

all the possible terminologies of studies that deal with this subject. The search string from Siddique et al. [25] was used as a starting point, as it was very complete and belonged to one of only two SMS that had been published in the last two years. In the search string, asterisks (*) denote the derivatives and variants of key terms, allowing for a more complete and exhaustive search:

```
(violen* OR fight* OR anomal*) AND
(activity OR event OR scene OR sequence) AND
(detect* OR recogni* AND from) AND
(surveillance OR cctv*) AND
(vi* OR motion) AND
(using OR through OR by OR via) AND
(machine learning OR computer vision OR deep learning)
OR artificial intelligence)
```

A number of changes have been made to the original search string, that improve the search criteria as well as make it more specific to the particularities of our study, creating the most complete search string possible. The term `anomal*` has been replaced with `physical aggression` and `physical attacks`, including the word "physical" so that it cannot be confused with verbal or other types of aggression or attacks. The connectors "from", "using", "through", "by" and "via" have been removed, as they do not add value to the search string and may limit relevant results. The term `camera` has been added to to the terms `video` or `motion`. The acronyms `artificial intelligence (AI)`, `machine learning (ML)`, `deep learning (DL)` and `computer vision (CV)` have been added to further extend the search coverage. Finally, a new AND statement has been added at the end of the string with terms related to trustworthy artificial intelligence and explainable artificial intelligence techniques: `trustworth*/explainab*/XAI/ethic*/robust*`. Therefore, the search string to be used is:

```
(physical aggression OR physical attack OR
violence OR fight) AND
(activity OR event OR scene OR sequence) AND
(detect* OR recogni*) AND
(surveillance OR CCTV) AND
(vi* OR motion OR camera) AND
(artificial intelligence OR AI OR machine learning OR ML OR
```

Table 3: Distribution of search string terms.

Action Description	*1*	physical aggression	physical attack	violence	fight
	2	activity	event	scene	sequence
	3	detect*	recogni*		
Tool	*4*	surveillance	cctv		
	5	vi*	motion	camera	
Technology	*6*	artificial intelligence, AI	machine learning, ML	deep learning, DL	computer vision, CV
	7	trustworth*	explainab*, XAI	ethic*	robust*

Table 4: Keyword possible derivative words of the terms included in the complete search string of the systematic mapping study.

Keywords	**Possible derivative words**
detect*	detect, detection, detectable, etc.
recogni*	recognition, recognizing, recognizer, etc.
vi*	video, visual, visualization, etc.
trustworth*	trustworthy, trustworthiness, trustworthily
explainab*	explainable, explainability, explainably
ethic*	ethical, ethics, ethically, etc.
robust*	robust, robustness, robustly, etc.

```
deep learning OR DL OR computer vision OR CV) AND
(trustworth* OR explainab* OR XAI OR ethic* OR robust*)
```

Table 3 makes it possible to easily visualise and analyse how the search string terms are distributed across the different categories, each AND term in the search string is found in a different row. In addition, the AND terms have been grouped into three main groups: "Action description" groups the first three items, "Tool" groups items 4-5 and "Technology" groups items 6-7.

As discussed above, the use of asterisks to cover the derivatives of key terms allows for a more complete and comprehensive search. In Table 4 the words used in search strings containing asterisks are shown together with some of their possible derivative words.

Three databases have been selected for the article search: Web of Sci-

ence, Scopus and Springer. These databases are good choices because of their broad interdisciplinary coverage, peer-reviewed content ensuring high quality, specialised focus on relevant areas, access to up-to-date research, and availability to many researchers, which ensures a comprehensive and robust review of the existing literature in the field. Choosing the right text string is difficult, as the search must contain a number of articles that is neither too low, nor too high. The search string shown in Table 4 is very comprehensive and encompasses all the terms that define our research topic, but certain terms may narrow the search too much or introduce articles that correspond to another topic of study.
All in all, it is an iterative process of trial and error. After testing multiple combinations, adding and removing elements, it was decided to remove two AND statements. The first is the statement that includes the elements: ("surveillance" or "CCTV"), as it has been observed that they limited the search for articles related to violence detection that simply did not use these terms. On the other hand, the sentence including the elements related to trustworthy artificial intelligence ("trustworth*" or "explainab*" or "XAI" or "ethic*" or "robust*") had very low results. This was something expected, since in the related work on physical assault detection, there was no mention of trustworthy AI elements at any point. Therefore, the final search string is as follows:

```
(physical aggression OR physical attack OR
violence OR fight) AND
(activity OR event OR scene OR sequence) AND
(detect* OR recogni*) AND
(vi* OR motion OR camera) AND
(artificial intelligence OR AI OR machine learning OR ML OR
deep learning OR DL OR computer vision OR CV)
```

The exact search string that applies to each database (as each database requires a specific format) is as follows:

- **Scopus database**

```
ABS(
((physical AND aggression) OR (physical AND attack) OR
   violence OR fight) AND
(activity OR event OR scene OR sequence) AND
```

```
(detect* OR recogni*) AND
(vi* OR motion OR camera) AND
((artificial AND intelligence) OR (ai) OR (machine AND
   learning) OR (ml) OR (deep AND learning) OR (dl) OR (
   computer AND vision) OR (cv))
)
```

- **Web of Science database**

```
AB = (
(((physical) AND (aggression)) OR ((physical) AND (
   attack)) OR (violence) OR (fight)) AND
((activity) OR (event) OR (scene) OR (sequence)) AND ((
   detect*) OR (recogni*)) AND
((vi*) OR (motion) OR (camera)) AND
(((artificial) AND (intelligence)) OR (ai) OR ((machine)
    AND (learning)) OR (ml) OR ((deep) AND (learning))
   OR (dl) OR ((computer) AND (vision)) OR (cv))
)
```

- **Springer database**

```
("physical aggression" OR "physical attack" OR "violence
   " OR "fight") AND
("activity" OR "event" OR "scene" OR "sequence") AND
("detect*" OR "recogni*") AND
("vi*" OR "motion" OR "camera") AND
("artificial intelligence" OR "ai" OR "machine learning"
    OR "ml" OR "deep learning" OR "dl" OR "computer
   vision" OR "cv ")
```

In the case of Scopus and Web of Science, it can be seen how the search string has been made for the Abstract, however, in the case of Springer this is not possible, so the search has been made for the full text. Although this significantly increased the number of articles to be discarded, Springer is a database with a very high number of published articles, so it was decided to continue with its selection.

3.2.2. Inclusion/Exclusion Criteria

The definition of inclusion/exclusion criteria for the collected articles is of great importance, as it ensures the quality of the results obtained in the SMS. [28]. In this case, we have defined exclusion criteria to filter the papers that had been selected through the created search strings.

- Articles that do not appear as a result of applying the selected search string, searching in the abstract.
- Papers other than journal articles or conference papers.
- Papers which were not published within the defined publication date, from June 2020 to June 2023.
- Papers which are not written in English.
- Duplicated papers.
- Papers that do not address the detection of physical aggression in video.

The adoption of the above-listed exclusion criteria must be justified. Firstly, limiting the publication date to the last three years (June 2020 to June 2023) ensures the inclusion of recent and up-to-date research. Secondly, excluding articles that are not written in English removes language barriers and promotes a comprehensive understanding of the SMS. Thirdly, removing duplicates prevents redundancy and biases. Focusing on studies which specifically address the detection of physical aggression in video ensures alignment with the research objectives. Thus, the SMS provides valuable insights on the topic within the established timeframe and scope.

3.3. Mapping Report

The mapping report involves reflecting the process followed in the selection of articles and their classification. Therefore, it is divided into three subsections: Filtering Studies, Classification Process and Validation of the Systematic Mapping Study.

3.3.1. Filtering Studies

From the articles selected in the search strings shown for each database, the exclusion criteria listed in Section 3.2.2 are applied. As mentioned above, the Springer database does not allow to filter by abstract. Thus, the results

had to be filtered according to the complete body of the article. Although this decision implied analyzing a much larger number of articles, we proceeded with the search on this database.
It is important to note that the term "violence detection" is ambiguous, and can include shootings, vandalism, robberies, etc. In this paper, violence detection has been termed as physical aggression, being in our opinion a much more specific term, although the most commonly used term is "violence detection". Those articles that have utilized at least one dataframe focused on physical aggression detection, rather than on violence in general (firearms, explosions, etc.), have been selected. All in all, after the initial search a total of 13137 journal articles and conference articles were retrieved (393 from Scopus, 240 from Web Of Science and 12,504 from Springer); a total of 63 articles have been selected following the application of the exclusion criteria.

3.3.2. Classification Process

Once the selection of the items was completed, the classification process was carried out. 27 of the selected works are journal articles and 36 are conference papers. All papers present a model for detecting violence using AI, 12 of them include a new dataset that is presented in the paper and 2 of them present an architecture for a system for detecting physical aggression in video was developed [31] [32].

3.3.3. Validation of the Systematic Mapping Study

Petersen et al. [29] introduced a framework for the validation and assessment of systematic mapping studies (SMS). This assessment guideline comprises 26 tasks that are to be carried out during the execution of an SMS. According to the authors, a proficient SMS should encompass a minimum of 33% of these tasks. Consequently, a minimum of 9 tasks needs to be executed to attain a mid-level quality SMS. As delineated in Petersen's Recommendations, in order to gauge the legitimacy of an SMS, it is imperative to compute the ratio between the number of tasks executed within an SMS and the 26 tasks delineated in the guidelines.

Table 5 exhibits the three distinct phases into which this mapping has been segmented: planning, development, and reporting. Additionally, a roster of tasks has been undertaken to guarantee the credibility of our study, all of which are grounded in the evaluation framework presented by Petersen and colleagues. Similarly, the SMS has been fashioned by incorporating 16 out of the 26 tasks prescribed by the Petersen framework. The resulting

ratio is 54%. Hence, the SMS surpasses the median quality as established by Montalvillo and Díaz.

4. Mapping

This section presents the systematic mapping study report, based on the methodology described in Section 3. First, a complete combination of character strings is shown to describe our topic of study, which is presented in Table 6. Several search trials had been carried out on the selected databases (Scopus, Web of Science and Springer) and it was decided to apply the search string with the best results, which is presented in the appropriate format for each database in Section 3.2.1. Finally, based on the articles obtained in each database for the same search string, the exclusion conditions set out in Section 3.2.2 are applied. Table 6 shows the number of selected articles after applying each exclusion criterion; the first column from the left being the first filter applied and the last column from the left the last filter applied. Each database is separated into two A for journal articles and C for conference articles.

The column *Search Results* indicates the number of articles obtained after applying the search string and filtering by journal articles and conference articles. *Article Type* contains the number of articles according to the row in journal articles or conference articles. *Year & Month* contains the number of articles after applying the exclusion criterion in which articles can only be published between June 2020 and June 2023; in the case of conference articles the date of the conference is selected using the date of the conference. *Physical Aggression* contains the number of articles after applying the exclusion criterion that articles must deal with video footage of physical aggression. Finally, *Papers with doubts* contains the number of articles selected after deciding not to select certain articles for diverse reasons like: not being focused on the detection of physical aggression violence, that information that is relevant to include in this study is not explained or does not appear, or that the quality of the study is not considered sufficient for inclusion in this work.

As mentioned above, the number of articles in Springer is high, as it does not allow searching by Abstract and it was decided to search by full body of the article. Finally, a total of 64 articles (27 journal articles and 36 conference articles) are selected.

Table 5: Activities from Petersen et al. [29] evaluation rubric conducted in this systematic mapping study (SMS).

			Applied	Section
Planning	**Need for the map**	Motivate the need and relevance	Yes	Sections 1, 2, 3
		Define objectives and questions	Yes	Objetive (3.1.2), Research Questions (3.1.3)
		Consult target audience to define questions	No	
Development	**Study identification/ Choosing search strategy**	Snowballing	No	
		Manual	Yes	3.2, 3.2.1
		Conduct database search	Yes	3.2.1
	Development search	PICO	Yes	3.2.1
		Consult librarian or experts	No	
		Iteratively try finding more relevant papers	Yes	3.2.1
		Keywords from know papers	Yes	
		Use standards, encyclopedias and thesaurus	No	
	Evaluate the search	Test-set of known papers	Yes	
		Expert evaluates results	No	
		Search web pages of key authors	No	
		Test-retest	No	
	Inclusion-Exclusion	Identify objective criteria for decision	Yes	3.2.2
		Add additional reviewer, resolve disagreements between then when needed	No	
		Decision rules	Yes	3.2.2
Report	**Extraction process**	Identify objective criteria for decision	Yes	3.2.2
		Obscuring information that could bias	No	
		Add additional reviewer, resolve disagreements between then when needed	No	
		Test-retest	No	
	Classification scheme	Research type	Yes	3
		Research method	No	
		Venue type	Yes	Table 6
	Validity discussion	Validity discussions / Limitations provided	Yes	3.3.3

Table 6: Number of selected articles after each exclusion criterion.

		Search Results	Article Type	Year & Month	Physical Aggression	Papers with Doubts
Scopus	**A**	393	162	114	23	16
	C	393	144	102	37	30
Web of Science	**A**	240	164	106	14	14
	C	240	65	63	7	6
Springer	**A**	43734	9637	2679	13	12
	C	43734	2867	1167	13	12

5. Analysis of Selected Articles

Section 5 contains the analysis of the selected violence detection articles. Section 5.1 explains the basic steps in the implementation of a video violence detection algorithm. Section 5.2 compiles the different challenges in video violence detection by grouping them into categories and organizing them according to the number of citations in the selected articles. Section 5.3 collects all the datasets used in the selected articles, providing all the information about their content, type of video, average duration, etc. Section 5.4 compiles all the evaluation parameters used to assess the success of violence detection in the selected papers. It is explained how these parameters are calculated and the usage frequency of those parameters is assessed across the selected articles. Section 5.5 presents all the methods for selecting relevant frames that may contain violence and assesses their importance in reducing unnecessary computational and processing costs In Section 5.6 the algorithm inputs used for violence detection algorithms are analyzed. Section 5.7 collects the different types of algorithms and classifiers that have been used in violence detection. Section 5.8 analyzes the accuracy obtained in the selected items, according to the dataframe used. Finally, Section 5.9 looks at the use of trustworthy artificial intelligence on the selected articles.

5.1. Basic Steps in Video Violence Detection Algorithm

Section 5.1 explains the basic steps in the implementation of a video violence detection algorithm. A graphic scheme has been made to help illustrate these steps in Figure 1. The first step starts with the input, for which a dataset of videos containing violent and non-violent scenes is necessary.

The second step is key frame extraction, although this step is not used in all algorithms. It consists in the selection of frames that may potentially contain violence; this selection is done to not have to process large amounts of video, thereby reducing the amount of computation being performed. The third step consists in transforming the data to serve as input to the violence detection algorithm, thus, the type of input depends on the features it is to extract. The fourth step is feature extraction and training the algorithm from these features; as discussed in Section 5.7, there are different combinations of algorithms and how the process is performed. Finally, a classifier decides whether the scene is violent or non-violent.

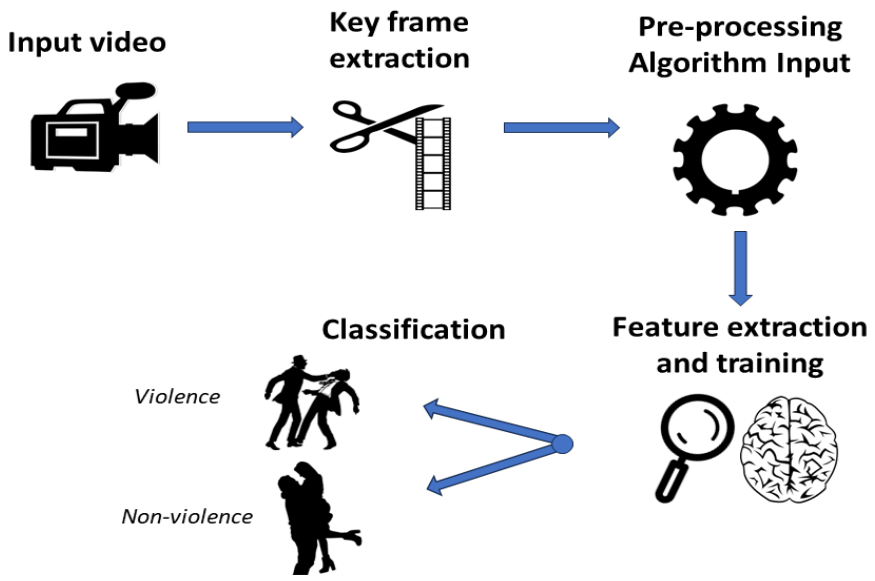


Figure 1: Basic steps in the implementation of a video violence detection algorithm.

5.2. Physical Aggression Challenges

Subsection 5.2 presents in Table 7 the challenges and limitations of physical aggression detection by compiling all the mentions that appear in any section of the selected articles. The aim is to provide a complete and updated view of the challenges encountered in this field, although only 12 of the 63 articles (20%) explicitly mention these challenges throughout their papers.

Table 7 presents the challenges compiled from the selected articles. The *Challenge Name* column contains the name of the challenge. It should be noted that the papers did not always use the same phrasing to describe the same type of challenge, thus, the most homogeneous terms have been chosen for each challenge. The *Challenge Type* column contains four different values in which the challenges have been grouped, since several of them are related to each other; these groups are: *Problems of Detection and Monitoring*, *Image Quality and Lighting Conditions*, *Hardware and Real-Time Considerations* and *Changes in the Scene and Environment*. *camera position* is the most addressed challenge, with a total of 6 articles mentioning it; then, with four citations, *occlusion of elements* and *non-stationary background*. In summary, the challenges grouped in the table point to the present challenges faced by the field of physical aggression detection in video.

5.3. Physical Aggression Datasets

Violence, in particular physical aggression, is a difficult event to detect in video due to a series of challenges which are catalogued below. One of these challenges is undoubtedly that, While physical assaults occur worldwide, they are not as commonly witnessed or recorded as other everyday activities such as people engaging in sports, taking strolls, or simply conversing. Therefore, the collection of quality public datasets is essential for the development of optimal physical assault detection algorithms. Section 5.3 analyzes and classifies all the datasets used in the selected papers, some of them being created and tested by the authors of papers in which aggression detection models are developed. Violence detection algorithms follow a series of steps to fulfill their task, and a multitude of methods can be used to carry out each step.

To evaluate the results obtained in different studies, datasets created several years ago have been used as training and testing data for the developed algorithms. Some of these datasets contain scenes of aggression from non-real scenarios, which poses a significant limitation to the algorithms' accuracy when applied to real-life situations. The most prominent examples in the selected articles are (in descending order of use): hockey fights [51], action movies [51], violent flow [52], RWF-2000 [53] and real life violence situations (RLVS) [54]. To evidence this, the number of the selected articles that have used the datasets is shown in Figure 2. Since there are many that have only been used once, those with two or more citations are shown. The extensive use of the hockey fight and action movie datasets, which were created in 2011

Table 7: A categorization of the challenges mentioned in the selected articles.

Challenge Name	**Challenge Type**	**Article Citation Count**	**Article Citation**
Occlusion of elements	Problems of Detection and Monitoring	4	[33], [34], [35], [36]
Scale variation		3	[33], [37], [35]
Different action depending on person		1	[38]
Similar appearances		1	[34]
Variation in illumination	Image Quality and Lighting Conditions	4	[32], [39], [37], [35]
Low-light video		4	[40], [41], [42], [36]
Low video resolution		4	[43], [44], [34], [45]
Motion blur		3	[32], [37], [35]
Weather conditions		2	[41], [42]
Illumination effects		1	[34]
Camera position	Hardware and Real-Time Considerations	6	[39], [46], [38], [36], [45], [37]
Real-time processing cost		4	[47], [48], [39], [49]
Limited availability of videos		2	[31], [49]
Camera movement		2	[44], [45]
Low person size		1	[44]
Non-stationary background	Changes in the Scene and Environment	4	[32], [33], [39], [43]
Crowded scene		3	[32], [39], [43]
Abrupt changes in motion		2	[50], [34]
Changes in object appearance		2	[34], [34]
Complex background		1	[35]
Scene clutter		1	[34]

and contain scenes of violence from field hockey games and action movies. While their use for comparison of results with other state-of-the-art items is understandable, these recordings are far from real situations for many reasons such as: camera movement, lighting conditions, focused approach to aggression, costumes of the people involved, image quality, etc.). This is why they should be treated with academic knowledge in view of the actual training of models.

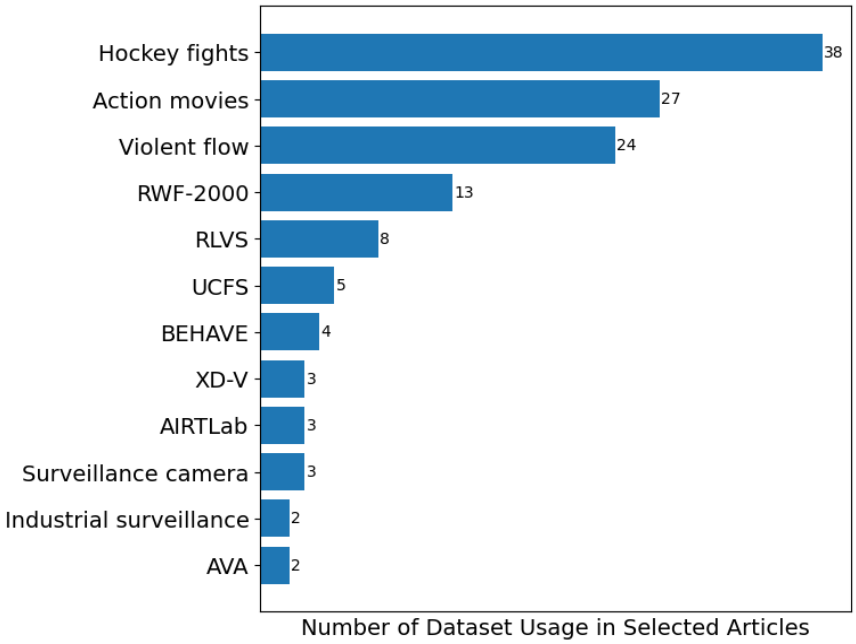


Figure 2: Count of datasets used in selected articles.

One of the exclusion criteria for the selection of articles were articles which addressed other types of aggression as opposed to physical violence. However, articles which employed datasets which contained dataframes with other types of violence such as shootings, sexual abuse, appearance of weapons, etc., in addition to dataframes with physical aggression, have been included in the selection. This decision stays in line with the objective of the present SMS while preventing the exclusion of widely used datasets. In addition,

it should be noted that in most cases, the authors filtered the dataframes, keeping only the videos that contained physical aggression and not other types of violence like fire. For example, two widely used algorithms that contain multiple types of violence are *UCF-Crime Selected (UCFS)* [55] and *XD-Violence Selected (XD-V)* [56]. On the other hand, datasets containing only non-violent actions, such as the *KTH dataset* [57] or datasets such as the *Sexual Harrasment Video (SHV)* [58] that contain scenes of sexual abuse, have not been considered in this SMS.

Table 8 summarises all the datasets (28 in total) used in the selected articles. Some of those datasets had been compiled as part of the proposal of the selected article. Value *N.S* in the cells, means that the information is *not specified* in the original article. The *Name* column contains the name of the dataframe; in cases where the creators of the dataframe did not give it a specific name it has been named as *untitled + article cite*. The *Clip No.* column contains the number of videos in the dataframe. The *Violent Videos* column indicates the number of violent videos in the dataframe and the *Non-Violent Videos* column indicates the number of non-violent videos in the dataframe. There is a wide variety of datasets from different sources, and of varying sizes and content. This is very positive because it enables the study of physical aggression detection from different angles.

Table 8: Physical Aggression Datasets used in selected articles.

Name	Origin Cite	Year	Only P.A.	Clip No.	Mean Frames or Duration	Frame Rate (FPS)	Video Quality	Violent Videos	Non-Violent Videos	Article Citations
Hockey Fights	[51]	2011	✓	1000	50 frames	20-30	720×576	500	500	[59], [32], [60], [33], [61], [31], [39], [62], [49], [63], [64], [43], [46], [65], [50], [66], [67], [44], [68], [38], [69], [70], [71], [72], [73], [74], [75], [47], [35], [76], [45], [58], [77], [78], [36], [79], [80], [81]
Action Movies	[51]	2011	✓	200	49 frames	20-30	515×720	100	100	[59], [60], [33], [61], [31], [49], [63], [64], [43], [46], [65], [66], [67], [44], [68], [38], [69], [71], [72], [73], [74], [76], [45], [58], [79], [80], [81]
Violent Flow/Violent Crowd	[52]	2012	✓	246	N.S.	25	320 × 240	123	123	[59], [59], [32], [60], [33], [61], [31], [43], [65], [50], [66], [67], [68], [38], [75], [75], [48], [35], [45], [77], [78], [79], [80], [81]
Real World Fight-2000 (RWF-2000)	[53]	2021	✓	2000	5 seconds	30	Multiple	1000	1000	[59], [61], [31], [39], [62], [49], [71], [72], [82], [42], [36], [81], [83]

Real Life Violence Situations (RLVS)	[54]	2019	✓	2000	5 seconds	20-30	480-720	1000	1000	[31], [62], [70], [84], [76], [77], [78], [83]
UCF-Crime Selected (UCFS)	[55]	2018		1900	7247 frames	N.S.	N.S.	950	950	[31], [64], [85], [72], [42]
BEHAVE	[86]	2010	✓	4	76800 Frames	25	640X480	Continuous	Continuous	[65], [67], [87], [72]
Surveillance Camera	[88]	2019	✓	300	2 seconds	Multiple	Multiple	150	150	[32], [39], [80]
XD-Violence Selected (XD-V)	[56]	2020		4754	N.S.	N.S.	N.S.	2405	2349	[31], [56], [89]
AIRTLab	[90]	2020		350	5.63 seconds	30	1920 ×1080	120	230	[50], [44], [91]
Industrial Surveillance	[39]	2021	✓	300	5 seconds	20-30	Multiple	150	150	[59], [39]
Human Violence	[92]	2021		1930	30-100 seconds	30	1,280 × 720	1930	0	[92]
Target	[93]	2022	✓	150	5-10 Seconds	60	1080p	N.S.	N.S.	[93]
Untitled [94]	[94]	2021	✓	273	N.S.	10	N.S.	180	93	[94]
Untitled [37]	[37]	2021	✓	N.S.	610 seconds	60	1080p	N.S.	N.S.	[37]
HD (High Definition)	[37]	2022	✓	2	2 hours	N.S.	1280×720	N.S.	N.S.	[87]
Conflict Event	[95]	2021	✓	6849	N.S.	N.S.	N.S.	N.S.	N.S.	[95]
Untitled [85]	[85]	2023	✓	20	N.S.	N.S.	N.S.	10	10	[85]
Untitled [70]	[70]	2023	✓	1112	N.S.	N.S.	N.S.	556	556	[70]

VSD2015	[96]	2015	✓	10900	8-12 seconds	N.S.	N.S.	505	10395	[97]
AVA-Kinectics Dataset	[98]	2020		430/143	15 minutes	N.S.	N.S.	35	83	[34], [99]
Media Eval VSD-2014	[100]	2014		N.S.	N.S.	25	N.S.	N.S.	N.S.	[101]
Untitled [102]	[102]	2021		60	30 seconds	Audio	Audio	not-specified	not-specified	[101]
Social Media Fight Images (SMFI)	[80]	2022		5691	N.S.	N.S.	N.S.	2739	2952	[80]
Untitled [103]	[103]	2021		2093	34.41 seconds	N.S.	N.S.	2093	0	[103]
Untitled [104]	[104]	2022		106	N.S.	N.S.	N.S.	66	40	[104]
Violent Clip Dataset (VCD)	[89]	2022		7279	8-12 seconds	N.S.	N.S.	3677	3609	[89]
Untitled [105]	[105]	2021		1500	3-6 seconds	13-15 seconds	N.S.	900	600	[105]

5.4. Physical Aggression Evaluation Parameters

Although it is possible to quantify the level of violence contained in a scene, violence detection is a problem with a binary result, i.e., physical aggression either is or is not present in a video; there are no intermediate values. Since this is a binary problem, the datasets with which the algorithms are trained basically have two classes, as classified in Table 5.3 in Section 8: *violent* and *non-violent* or other terms such as *fight* and *no fight*, although the type of violence that occurs or its intensity can be subdivided. When analyzing the results of a model, the confusion matrix is widely used. The selection of evaluation parameters is vital to correctly quantify and evaluate the results. The use of metrics is very useful, since it allows to have concise one-value information on the obtained results. Section 5.4 presents the evaluation parameters used in the selected articles, quantifies them according to usage frequency. All the evaluation parameters that had appeared in the selected articles are listed in Table 9. *Accuracy* is clearly the most used metric, as 54 articles used accuracy, followed by *recall*, which was applied in 18 articles.

5.5. Selection of Relevant Frame Methods

Although the datasets presented in Section 5.3 are mostly cropped and labeled, so that they only contain scenes of violence and non-violence at the precise moments when the action occurs, the ultimate goal of these systems is real-time detection. Given that physical aggression is an anomalous (rare) event, violence detection would be null in most of the video recording time [63]. This is a high computational cost, since in most cases, consecutive frames are duplicated [65]. Section 5.5 presents the methods for relevant frame extraction, that is, frames that contain physical aggression. In addition, this section presents methods to reduce frame processing in continuous recordings. These methods can be classified into two types. One type of method reduces the number of frames to be processed while continuously passing information to the violence detection algorithms. The other type of method does not pass information to the detection algorithms until a change is recorded in the video which could potentially be a violent situation. Despite the fact that relevant frame selection is vital to reducing the economic and computational cost of a real physical aggression detection system, not all the articles presented methods to address this issue, focusing solely on the level of accuracy of their proposed violence detection method. In fact, from among the 63 selected articles only 21 present relevant frame extraction

Table 9: Evaluation parameters used in the selected articles.

Evaluation Parameters	**Article Citations Count**	**Articles Citations**
Accuracy	54	[59], [32], [60], [33], [61], [31], [39], [49], [63], [64], [46], [65], [50], [92], [66], [67], [93], [94], [68], [37], [87], [95], [40], [38], [91], [69], [85], [70], [71], [72], [84], [34], [73], [82], [74], [101], [41], [75], [47], [48], [35], [76], [45], [58], [77], [78], [36], [79], [80], [103], [104], [81], [83], [89]
Recall	18	[32], [61], [39], [49], [63], [43], [50], [44], [93], [37], [38], [91], [73], [101], [41], [58], [103], [106]
F1 Score	16	[32], [61], [39], [49], [43], [50], [44], [93], [37], [38], [91], [73], [101], [41], [58], [103]
Precision	14	[61], [39], [49], [63], [43], [44], [93], [37], [38], [73], [41], [58], [103], [106]
Confusion Matrix	12	[59], [32], [61], [39], [43], [67], [44], [94], [85], [101], [58], [105]
Specificity	8	[61], [50], [44], [93], [37], [91], [73], [101]
AUC	7	[31], [39], [67], [44], [93], [35], [58]
ROC	6	[67], [93], [94], [37], [91], [35]
Computational Time	5	[63], [44], [93], [37], [71]
AP	4	[56], [99], [89], [106]
MAP	3	[97], [99], [106]
G-Mean	2	[93], [37]
Author-created parameters	1	[62]

methods in their papers i.e. 33%. Table 10 contains all the selected articles that use a relevant frame extraction method. The *Cite* column contains the reference of the article. The *Key Frame Extraction* column contains three categories grouping the different approaches to frame extraction, namely: *image variation*, *systematic sampling* and *object detection*. *Image variation* refers to the methods that rely on image variation for relevant frame extraction, such as the use of: optical flow, spatio-temporal variation and variation in brightness or textures. *Systematic sampling* refers to the methods that select a certain number of frames, independently of the video, while *object detection* methods are based on the detection of persons for selected frames. Finally, the *Pre-Processing Method* column lists the method used for relevant frame extraction.

5.6. Feature Extraction

Once the key frame extraction method has been selected, it is necessary to decide what information is introduced to the algorithm so that it can extract the corresponding features. This step is also known as *pre-processing*. It is a decisive step as it enables the algorithm to extract the features for the classification of violence or non-violence on the basis of the incoming information. There are a multitude of ways to feed video information to the algorithm: RGB video, grayscale video, optical flow, etc. Section 5.6 presents the types of input that are introduced into the algorithm and classifies them. Table 11 lists the different algorithm inputs used in the selected articles. The "Algorithm Focus" column contains three created categories to classify the different types of existing input, namely: image, movement and audio. It has been decided to count as a different type of input those that use image smoothing techniques such as Gaussian blur, since it implies an alteration of the original video. The following table provides a brief description of each of the algorithm inputs grouped by Cite Count in descending order. The most used input is *RGB video*, followed by *optical flow* with almost 5 times less citations.

5.7. Violence Detection Algorithm

Given the input types seen in Section 5.6, Section 5.7 presents the algorithms used in the selected articles for feature extraction and training, respectively. Section 5.7 categorises the types of algorithms used in the selected

Table 10: Categorization of the key frame extraction methods used in the selected articles.

Cite	Key Frame Extraction	Pre-Processing Method
[60]	Image variation	Structural Similarity Index Measure (SSIM)
[61]		Gunner Farneback's technique
[62]		Frame different method
[36]		Own video segmentation based on brightness difference of difference between image frames
[65]		Difference between two frames matrix with threshold
[63]		Mean Square Error (MSE)
[43]	Systematic sampling	Random Frame Selection
[73]		Random Frame Selection (20 frames/video)
[67]		Select frames in regular intervals
[70]		Select F number of frames each second with equidistant interval
[97]		Sparse sampling strategy
[99]	Object detection	YoloV3
[66]		YoloV4, Deep Sort & SiamRPN+
[104]		YoloV5, Deepsort
[106], [41]		YoloV5
[44]		Yolo + Farneback
[105]		OpenPose
[74]		OpenCV
[94]		Mask-RCNN single person detection
[32]		Mask-RCNN
[39]		Light-weight CNN

Table 11: Algorithm input in selected articles.

Algorithm Input	Algorithm Focus	Cite Count	Cite
RGB	Image	48	[59], [32], [60], [31], [39], [62], [49], [64], [46], [65], [50], [92], [66], [67], [44], [93], [94], [68], [37], [87], [95], [40], [91], [85], [84], [97], [34], [82], [41], [75], [47], [48], [35], [56], [58], [77], [78], [36], [36], [99], [79], [103], [104], [81], [83], [89], [106], [105]
Grey		5	[63], [38], [69], [70], [73]
Gaussian blur		1	[42]
Image patches		1	[80]
Laplacian operator		1	[36]
Median blur		1	[42]
Range 3 matrix (xy/xt/yt)		1	[33]
Boundary calculation		1	[42]
Optical flow	Movement	10	[61], [63], [92], [38], [69], [72], [42], [45], [36], [89]
Frame difference		2	[43], [71]
Separate motion energy picture		2	[38], [69]
Background supressed frames		1	[71]
Dynamic image		1	[76]
Gaussian blur optical flow image		1	[74]
Gaussian blur RGB frame difference		1	[74]
Audio	Audio	5	[95], [97], [101], [56], [89]

works and attention is paid to usage frequency, common algorithm combinations and the stage of the violence detection process these algorithms are applied to.

Traditionally, algorithms based on manual feature extraction differentiated the violence detection process into two phases: feature extraction and feature training. However, as seen below, deep learning-based algorithms combine training and feature extraction in a single process. Additionally, several works have divided the process into two phases when using deep learning methods, for example, through the combined use of CNN and LSTM. For these reasons, it has been decided to organize the violence detection process into two phases: Phase 1 and Phase 2. As mentioned, in the case of algorithms that perform manual feature extraction, Phase 1 is feature extraction, and Phase 2 is the training of these features. In the case of combinations such as CNN and LSTM, it does not necessarily mean that one is for feature extraction and the other is for feature training but that two different algorithms are used in the process. There are other types of combinations, and all of this is discussed in more detail below.

5.7.1. Types of Algorithms Used for Violence Detection

In this section, the different types of algorithms used in the detection of violence are presented. The algorithms are grouped into different categories, some of them with their own subcategories. The tables and figures in Section 5.7 are grouped and different categories are color-coded to facilitate the reader's comprehension. The categories are listed as follows:

- **Manual algorithms** usually focus on feature extraction from factors such as color and motion (analysis of abrupt changes in color and motion in video frames), textures and patterns [33]. They require a feature extraction process and a training process of those extracted features [45].

- **CNN**. CNNs are commonly used in violence detection due to their ability to capture spatial features in video frames, making them effective at recognizing patterns associated with violent actions [45]. Pre-trained convolutional neural networks (CNNs) are also widely used in the selected works [66]. The subgroups into which the CNNs are divided are: *CNN* and pre-trained CNNs (*CNN PT*).

- **LSTM**. LSTMs are the most widely used RNNs in violence detection problems. They allow to learn and control the flow of information along temporal sequences more effectively than traditional RNNs because they solve the problem of gradient fading that causes a slowdown in the learning process [68] [43].

- **Skeleton based**. Skeleton-based algorithms are based on the tracking and analysis of human postures and movements represented as skeletons in videos. These algorithms use information from joints and connections between body parts to identify abnormal or violent movement patterns. Skeleton-based algorithms are divided into two subcategories: skeleton-based deep-learning techniques *skeleton-based (DL)* [82] [105] and skeleton-based manual techniques *skeleton-based (manual)*.

- **Transformer** algorithms are another type of algorithms used for the detection of violence in video [80] [44] [70]. These divide the image into patches, add position information, compute attention between patches, apply layered transformations, and produce outputs for classification tasks, object detection, and more. These models are an effective alternative to convolutional networks in image processing applications.

- **Audio-based** algorithms in violence detection analyze sound features, such as pitch, rhythm, and violent sound events (such as screams or blows). Violence detection can be performed using only audio from videos (which results in much lighter processing than using images) or by combining audio and video analysis (which results in much lighter processing than using images). Audio-based algorithms are divided into two subcategories: audio-based manual feature extraction techniques *audio (manual)* [101] or audio-based deep learning techniques *audio (DL)* [97].

Once the types of algorithms used for violence detection in the selected articles have been presented, Table 12 shows for each selected article, the algorithms used in Phase 1 and Phase 2, the category to which they belong and the type of classifier used.

Table 12: Algorithms and classifiers used for violence detection in the selected articles.

Cite	Ph1 Type	Ph2 Type	Ph1 Algorithm	Ph2 Algorithm	Classifier
[65]	CNN	N	CNN-V4	N	FCL(2; Sigmoid)
[45]	CNN	N	CNN	N	FCL(2; SoftMax)
[84]	CNN	CNN	CNN-2D	CNN-3D	FCL(2; SoftMax)
[41]	CNN	N	CNN-3D	N	UNK
[60]	CNN	N	CNN-3D	N	FCL (2; SoftMax)
[34]	CNN	N	SlowfastNetwork	SlowfastNetwork + X3D Network	FCL (2; Sigmoid)
[106]	CNN	N	CNN-3D	N	FCL (2; Sigmoid)
[69]	CNN	N	Optimized CNN	N	FCL (2; SoftMax)
[92]	CNN	ML	Two Stream CNN	Rank Learning Machine	Ranking Score
[85]	CNN PT	N	3D-CNN	N	FCL (2; SoftMax)
[62]	CNN PT	N	ResNet-18 pre-trained	N	FCL (2; UNK)
[36]	CNN PT	N	ResNet + Inception-V1 combination)	N	FCL (1; SoftMax)
[66]	CNN PT	N	Pre-trained Two-stream inflated 3D ConvNets	N	Logistic Regression(LR); XGBoost; Linear Support Vector Machine(LSVM)
[48]	CNN PT	N	CNN(3D-ResNet)	N	FCL (UNK; UNK)
[31]	CNN PT	N	Transfer-Learned X3D-M pre-trained on ImageNet	N	FCL (2; Sigmoid)
[76]	CNN PT	N	ResNet-V2 pre-trained with ImageNet	ResNetV2 & Fine-Tunning	FCL (1; SoftMax)
[99]	CNN PT	N	Two-chanel [Slow and Fast Path] ResNet-50	N	FCL (1; UNK)

[46]	CNN PT	CNN PT	Deep Multi-Net (Combination of pre-trained GoogleNet + AlexNet)	DMN	UNK
[87]	CNN PT	CNN	C3D pre-trained with THUMOS2014 dataset	DC3D	FCL (UNK; SoftMax)
[83]	CNN PT	N	X3D Neural Network pre-trained with Kinetics-400	X3D Neural Network pre-trained with Kinetics-400 + Fine tunning	FCL (UNK; UNK)
[39]	LSTM	LSTM	Conv-LSTM	GRU Network	FCL (UNK; UNK)
[79]	CNN	LSTM	CNN	LSTM	FCL (2; UNK)
[68]	CNN	LSTM	CNN	Bi-LSTM	FCL(UNK; UNK)
[42]	CNN	LSTM	Conv-2D	LSTM	FCL (UNK; UNK)
[32]	CNN	LSTM	Darknet + Residual Optical Flow CNN	M-LSTM	FCL (UNK; SoftMax)
[49]	CNN	LSTM	Time-distributed U-Net	LSTM	FCL (2; UNK)
[71]	CNN	LSTM	MobileNet	Separable Convolutional LSTM	FCL (2; Binary cross-entropy)
[50]	CNN PT	LSTM	CNN-3D trained with Sports 1M	Conv-LSTM	SVM; Fully Connected Layers (SoftMax layers); Fully Connected Layers (SoftMax layers)
[72]	CNN PT	LSTM	VGG-16 pre-trained with ImageNet	Bi-ConvLSTM	FCL (UNK; SoftMax)
[77]	CNN PT	LSTM	VGG-16 pre-trained on INRA person dataset	Bi-GRU	FCL (3; SoftMax)
[47]	CNN PT	LSTM	VGG-16 pre-trained	LSTM	FCL (3; Sigmoid)
[67]	CNN PT	LSTM	VGG-16	Wide Dense Residual Blocks (WDRB) + LSTM	FCL (1; SoftMax)

[91]	CNN PT	LSTM	MobileNet-V2	Bi-LSTM; ConvLSTM	FCL (2; ReLU/ Simgmoid)
[75]	CNN PT	LSTM	VGG-16	LSTM/Bi-LSTM	Binary cross entropy
[58]	CNN PT	LSTM	VGG-16 pre-trained/VGG-19 pre-trained	LSTM	FCL (1; SoftMax)
[43]	CNN PT	LSTM	Mobile Neural Architecture Search (MNAS) Pre-trained Automated	Conv-LSTM	Random Forest; SVM; K nearest neighbour
[59]	CNN PT	LSTM	VGG-19	Bi-LSTM	FCL (UNK; SoftMax)
[64]	CNN PT	LSTM	Xception pre-trained with ImageNet	LSTM	FCL (UNK; SoftMax)
[74]	CNN PT	LSTM	Two channel DarkNet19 pre-trained on ImageNet	LSTM	FCL (2; UNK)
[93]	CNN PT	LSTM	Pre-trained CNN's: VGG16/ VGG19/ InceptionV3/ DenseNet201/ ResNet101/ MobileNet/ NASNet-Large/ VGG16+VGG19/ ResNet50+ ResNet152V2/ InceptionV3/ ResNet101V2	LSTM	FCL (3; SoftMax)
[78]	CNN PT	LSTM	Two channel EfficienNet-B0 pre-trained on ImageNet (one for optical flow and another for RGB video)	Bi-LSTM	FCL (1; Sigmoid)
[33]	Manual	Manual; CNN	TOP-ALCM	HandCraftedFeatures/CNN	SVM; CNN(light-CNN)
[63]	Manual	ML	Local Binary Pattern (LBP); Fuzzy Histogram of Optical Flow Orientations	AdaBoost (Adaptive Boosting)	Ensemble RobustBoost aggregation
[35]	Manual	ML	Local Orientation Pattern (LOOP)	SVM	SVM

[38]	Manual	CNN	Grayscale; Optical flow vectors; Motion Energy Image (MEI),	3-stream CNN	FCL (2; SoftMax)
[73]	Manual	ML	Principal Component Analysis (PCA)/Discrete Wavelet Transform (DWT)	SVM	SVM
[82]	Skeleton-based	Skeleton-based	CNN-3D + TokenPose	CNN-3D	FCL (1; Sigmoid)
[105]	Skeleton-based (DL)	ML	Number of skeletons; Distance between skeletons; changes in human skeleton hand acceleration	SVM	SVM
[94]	Skeleton-based (DL)	LSTM	DNN(DeepPose)	LSTM	FCL (UNK; SoftMax)
[103]	Skeleton-based (DL)	Skeleton-based (DL)	PoseNet pre-trained	PoseNet pre-trained	FCL (UNK; Softmax)
[104]	Skeleton-based (DL)	Manual	PoseNet pre-trained	Dynamic Time Warping (DTW)	K-Nearest Neighbors; Random Forest; Naive Bayes
[37]	Skeleton-based (DL)		CNN(two-branch multistage CNN)	N	SVM
[81]	Skeleton-based (Manual)	CNN	Own skeleton-based algorithm	SPIL	FCL (UNK; UNK)
[80]	Transformer	N	ViT	N	Fully connected layer(UNK; UNK)
[44]	Transformer	N	STAT Network (Generator	Discriminator)	STAT Network (Generator) + VGG + Cosine Similarity
[70]	Transformer	N	Multi-Headed Self Attention Layer (MSA)	N	FCL (1; Sigmoid)

[101]	Audio (Manual)	ML	Zero Crossing Rate (ZCR), Amplitude Envelope (AE), Short Time Energy (STE), Root Mean Square Energy (RMS), Spectral Flux (SF), Bandwidth (BW), Band Energy Ratio (VER), Mel-Frequency Cepstral Coefficient (MFCC)	ELM	ELM
[56]	Audio (Manual)	ML	Own formulas	Holistic Localized Network (Holistic Branch; Localized Branch; Dynamic Score Branch)	FCL (2; ReLU)
[89]	Audio (DL) + CNN PT	ML	TSM-50 (Temporal Shift Module 50); ResNet-50; PANNs (Pre-trained Audio Neural Networks)	MAF-Net	FCL (2; Sigmoid)
[97]	Audio (DL) + CNN PT	LSTM	TEA network, VGGish	LSTM	FCL (1; UNK)
[95]	Audio (DL) + CNN PT	CNN + P3D (Pseudo-3D Convolutional Networks pre-trained)	AudioNet pre-trained	Reasoning Network (2 FCL; SoftMax) + Predicted Network (3 FCL; SoftMax)	Reasoning Network (2 FCL; SoftMax) + Predicted Network (3 FCL; SoftMax)

The *Cite* column contains the article reference. The *Ph1 Type* column contains the category to which the Phase 1 algorithm belongs. The *Ph2 Type* column contains the category to which the Phase 2 algorithm belongs. The *Ph1 Algorithm* column contains the algorithm used in Phase 1. The *Ph2 Algorithm* column contains the algorithm used in Phase 2. The *Classifier* column lists the classifier that had been used, where terms with a structure similar to *FCL (2, Sigmoid)* mean that the classifier was a fully connected layer, consisting of two layers and that the activation function used in the last layer was the Sigmoid function. In the *Ph1 Type* and *Ph2 Type* columns, color codes have been used to facilitate the identification of the type of algorithms and their combinations. If term *N* appears in the Table in both the *Ph2 Type* and *Ph2 Algorithm columns*, this means that the article does not use a second algorithm to perform violence detection and everything is done in the first phase.

Overall, Section 5.7.1 presented the different types of algorithms used for violence detection in videos, explaining their operation, their subcategories and variants. In addition, Table 12 includes the algorithms and classifiers used in the selected papers, as well as the category to which these algorithms correspond.

5.7.2. Analysis of the Types of Algorithms Used in the Selected Articles

Section 5.7.2 presents graphs based on Table 12 to facilitate the reader's comprehension of the most used types of algorithms.

Figure 3 counts and categorises the types of algorithms used in the selected articles. The same color coding has been kept for the categories as in Table 12. There are also subcategories, namely, *CNN* is divided into *CNN* and *CNN PT* (CNN pre-training). Audio is divided into *audio (DL)* and *audio (manual)* while *skeleton-based* is divided into *Skeleton-based (DL)* and *skeleton-based (manual)* according to the type of algorithm used. Most of the algorithms that had been used in the first phase were pre-trained CNNs and CNNs, followed by skeleton-based deep learning techniques and manual feature extraction, with less than half of the citations using them. It can also be observed that most skeleton-based algorithms are based on DL techniques. Figure 4 shows the count of the algorithms used in phase 2. There is a large difference between this graph and the graph in Figure 3. In this case, LSTMs are evidently the most used type of algorithm. CNNs are used much less. Machine Learning (ML) algorithms that did not appear in Phase 1 are the second most used algorithm in Phase 2. In this case, the skeleton-based algorithms do not appear, since they are solely related to the process of feature extraction in which they detect joint points on the people appearing in the videos.

The following graph depicts the combinations used in the first and second phases. It has been created to facilitate the reader's comprehension of the most common combinations of algorithms. Figure 5 represents the count of the most used algorithm combinations grouped by algorithm subgroups. It evidences that the most used combination are pre-trained CNNs with and without LSTMs, as well as the use of only CNNs or CNNs and LSTMs. The fifth most common combination is the use of a traditional violence detection algorithm with a manual feature extraction and a classical machine learning algorithm for training.

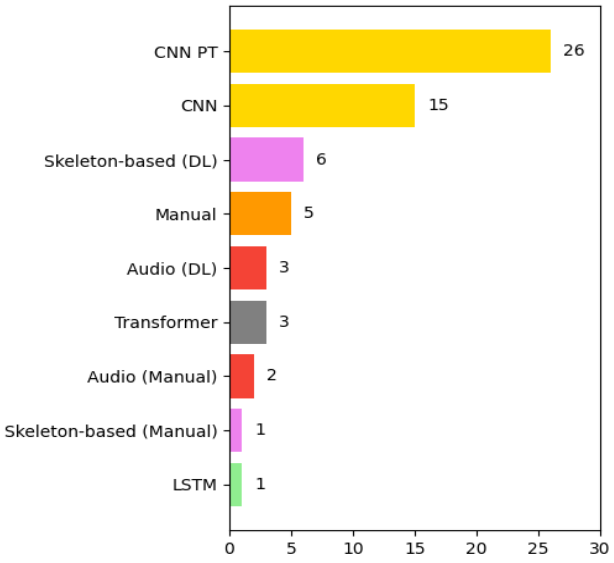


Figure 3: Count and categorisation of algorithm types used in phase 1 in the selected articles.

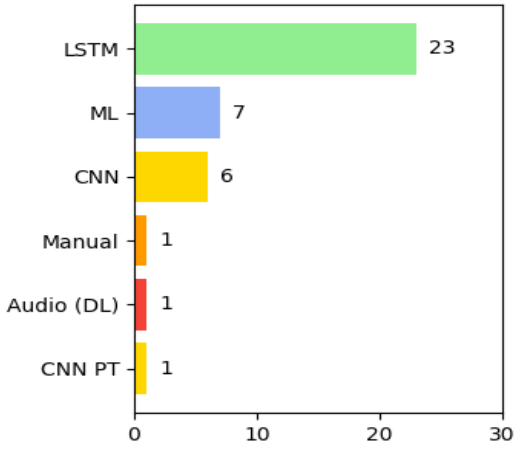


Figure 4: Count of the algorithm types used in phase 2, grouped by subcategories.

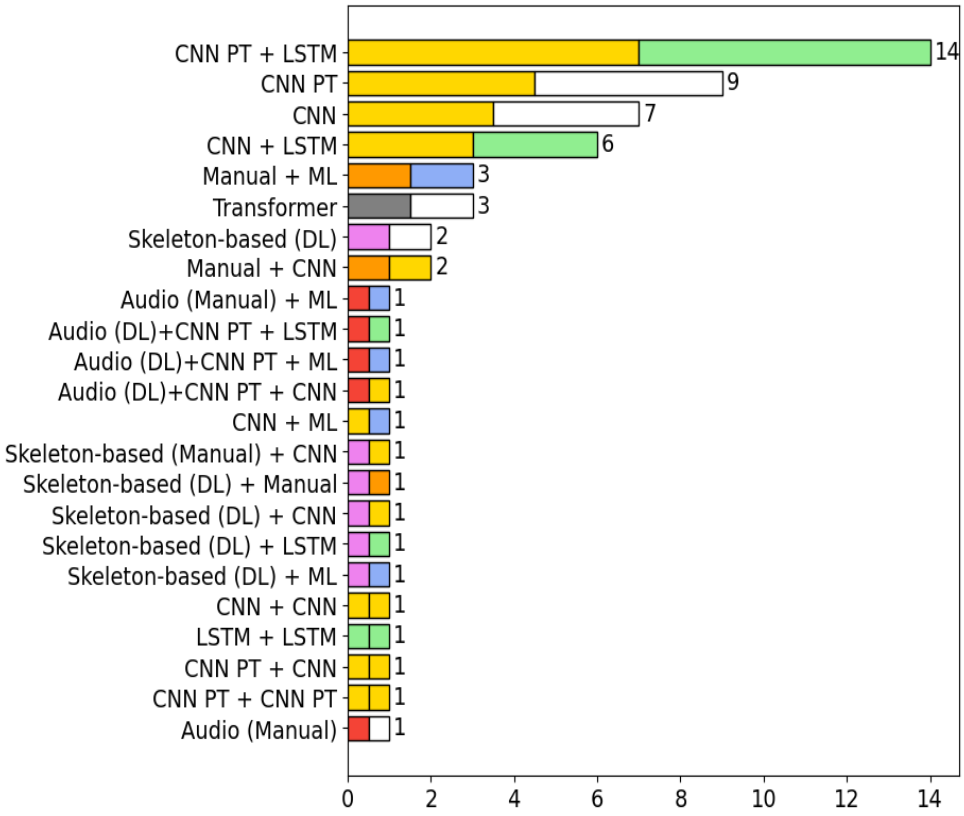


Figure 5: Count of algorithm type combinations, grouped by subcategories.

Overall, it has been possible to observe the most used algorithms in phases 1 and 2 of the process of detecting violence in videos, as well as the most used combinations.

5.7.3. Classifiers in Selected Articles

Following the feature extraction and training tasks outlined in Sections 5.7.1 and 5.7.2, the process ends with a classifier that determines whether the video is violent or not; this is therefore a very important process. Figure 6 shows the count of the classifiers used in the selected articles. A color code has been used to differentiate those that are based on deep learning techniques, and those that use traditional methods. As traditional methods we refer to those classifiers that are of mathematical origin or are based on machine learning techniques. The classification of classifiers is divided between those based on deep learning techniques and those based on traditional methods, which include classical Machine Learning algorithms and statistical methods. It can be seen in Figure 6 that the most used classifier technique is densely connected layers. This is logical given that this classifier is integrated in the output of deep learning methods and, as seen in Sections 5.7.1 and 5.7.2, a large number of articles are based on the use of CNNs and combinations

between CNNs and LSTMs.

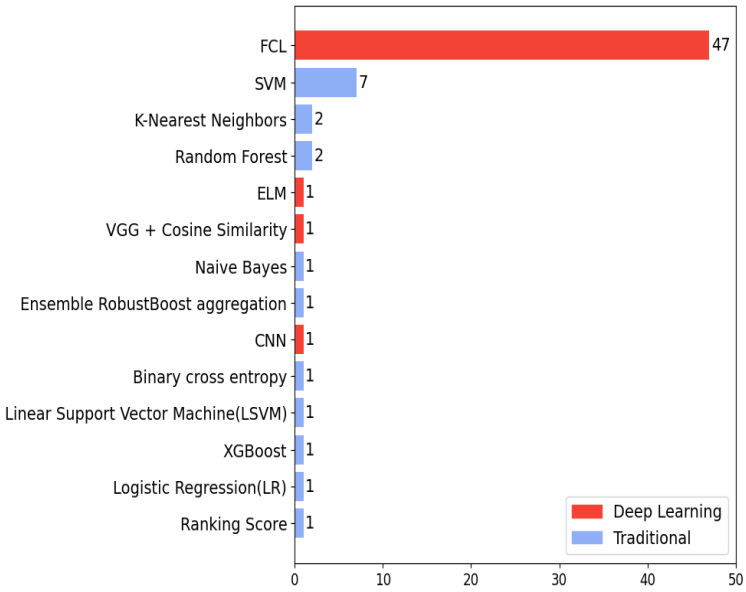


Figure 6: Count of the classifiers used in the selected articles.

There are different activation functions for the fully connected layers, as shown in Table 12. Sigmoid, ReLU and SoftMax activation functions were used in the selected articles. SoftMax does not give a binary output, thus, to obtain a binary classification of either violence or non-violence, a threshold value must be set at which violence is considered to occur.

5.8. Accuracy Obtained in the Detection of Violence by the Selected Articles

Section 5.8 compiles all the accuracy values obtained by the selected articles in the detection of violence. Some of the selected items do not use the accuracy metric, as discussed in Section 5.4; however, it is by far the most used metric. The results obtained from the selected articles that do not use accuracy as a metric, do not appear in this compilation.

Table 13 contains the accuracy results achieved in the selected articles, sorted by the dataset that had been employed and by the accuracy result obtained in descending order. Table 13 does not show those datasets that contain only one citation; therefore, only the results whose papers have used the datasets shown in Table 8 are considered. The *Dataset* column contains the dataset that had been used. The *Cite* column contains the reference of the article. The *Algorithm Combination* column contains the subcategories of each algorithm from phases 1 and 2, as well as the citation with the names of the authors. The *Ph1 Type* and *Ph2 Type* columns contain the subcategories to which the used algorithms

belong. The *Ph1* and *Ph2* columns contain the algorithms used in Phase 1 and Phase 2. Finally, the *Acc.* column contains the accuracy obtained by a given article, using the corresponding algorithm.

Table 13: Accuracy achieved in violence detection in the selected articles.

Dataset	Cite	Algorithm's Combined Name	Ph1 Type	Ph2 Type	Ph1	Ph2	Acc.
	[69]	CNN [Appavu]	CNN	N	Optimized CNN	N	100
	[38]	Manual + CNN [Mohtavipour et al.]	Manual	CNN	Grayscale + Optical flow vectors + Motion Energy Image (MEI)	3-stream CNN	100
	[33]	Manual + Manual/CNN [Hu et al.]	Manual	Manual/CNN	TOP-ALCM	HandCraftedFeatures + CNN	99.9
	[46]	CNN PT + CNN PT [Mumtaz et al.]	CNN PT	CNN PT	Deep Multi-Net	DMN	99.82
	[71]	CNN + LSTM [Islam et al.]	CNN	LSTM	MobileNet	Separable Convolutional LSTM	99.5
	[49]	CNN + LSTM [Vijeikis et al.]	CNN	LSTM	Time-distributed U-Net	LSTM	99.5
	[66]	CNN PT [Freire-Obregón et al.]	CNN PT	N	Pre-trained Two-stream inflated 3D ConvNets	N	99.45
	[60]	CNN [Mahmoodi et al.]	CNN	N	CNN-3D	N	99.4
	[68]	CNN + LSTM [Halder et al.]	CNN	LSTM	CNN	Bi-LSTM	99.27
	[47]	CNN PT + LSTM [Aarthy et al.]	CNN PT	LSTM	VGG-16 pre-trained	LSTM	99.1
	[72]	CNN PT + LSTM [Mugunga et al.]	CNN PT	LSTM	VGG-16 pre-trained with ImageNet	Bi-ConvLSTM	99.1
	[78]	CNN PT + LSTM [Traoré et al.]	CNN PT	LSTM	Two channel EfficienNet-B0 pre-trained on ImageNet	Bi-LSTM	99

[43]	CNN PT + LSTM [Jahlan et al.]	CNN PT	LSTM	Mobile Neural Architecture Search (MNAS) Pre-trained Automated	Conv-LSTM	99
[75]	CNN PT + LSTM [Gupta et al.]	CNN PT	LSTM	VGG-16	Bi-LSTM	98.9
[67]	CNN PT + LSTM [Asad et al.]	CNN PT	LSTM	VGG-16	Wide Dense Residual Blocks (WDRB) + LSTM	98.8
[50]	CNN PT + LSTM [Sernani et al.]	CNN PT	LSTM	CNN-3D trained with Sports 1M	Conv-LSTM	98.76
[39]	LSTM + LSTM [Fath et al.]	LSTM	LSTM	Conv-LSTM	GRU Network	98.5
[62]	CNN PT [Bi et al.]	CNN PT	N	ResNet-18 pre-trained	N	98.5
[65]	CNN [Ahmed et al.]	CNN	N	CNN-V4	N	98.11
[80]	Transformer [Aktı et al.]	Transformer	N	ViT	N	98
[81]	Skeleton-based (Manual) + CNN [Su et al.]	Skeleton-based (Manual)	CNN	Own skeleton-based algorithm	SPIL	98
[58]	CNN PT + LSTM [Islam et al.]	CNN PT	LSTM	VGG-19 pre-trained	LSTM	98
[77]	CNN PT + LSTM [Traoré et al.]	CNN PT	LSTM	VGG-16 pre-trained on INRA person dataset	Bi-GRU	98
[45]	CNN [Ehsan et al.]	CNN	N	CNN	N	98
[32]	CNN + LSTM [Ullah et al.]	CNN	LSTM	Darknet+Residual Optical Flow CNN	M-LSTM	98
[75]	CNN PT + LSTM [Gupta et al.]	CNN PT	LSTM	VGG-16	LSTM	97.6
[31]	CNN PT [Huszár et al.]	CNN PT	N	Transfer-Learned X3D-M pre-trained on ImageNet	N	97.5
[64]	CNN PT + LSTM [Sharma et al.]	CNN PT	LSTM	Xception pre-trained with ImageNet	LSTM	96.55

Hockey fights	[74]	CNN PT + LSTM [Singh et al.]	CNN PT	LSTM	Two channel DarkNet19 pretrained on ImageNet	LSTM	96.2
	[70]	Transformer [Kumar et al.]	Transformer	N	MSA	N	95.15
	[73]	Manual + ML [Wintarti et al.]	Manual	ML	Discrete Wavelet Transform (DWT)	SVM	95
	[79]	CNN + LSTM [Talha et al.]	CNN	LSTM	CNN	LSTM	94.9
	[36]	CNN PT [Chen et al.]	CNN PT	N	ResNet+Inception-V1 combination	N	94.1
	[101]	Audio (Manual) [Mahalle et al.]	Audio (Manual)	N	ZCR + AE + STE + RMS + SF + BW + VER + MFCC	ELM	93.5
	[76]	CNN PT [Jain et al.]	CNN PT	N	ResNet-V2 pre-trained with ImageNet	ResNetV2 + Fine-Tunning	93.33
	[35]	Manual + ML [Lohithashva et al.]	Manual	ML	Local Orientation Pattern (LOOP)	SVM	92.25
	[59]	CNN PT + LSTM [Mumtaz et al.]	CNN PT	LSTM	VGG-19	Bi-LSTM	91.29
	[63]	Manual + ML [Jaiswal et al.]	Manual	ML	Local Binary Pattern (LBP) + Fuzzy Histogram of Optical Flow Orientations	AdaBoost	88.66
	[60]	CNN [Mahmoodi et al.]	CNN	N	CNN-3D	N	100
	[69]	CNN [Appavu]	CNN	N	Optimized CNN	N	100
	[66]	CNN PT [Freire-Obregón et al.]	CNN PT	N	Pre-trained Two-stream inflated 3D ConvNets	N	100
	[31]	CNN PT [Huszár et al.]	CNN PT	N	Transfer-Learned X3D-M pre-trained on ImageNet	N	100
	[76]	CNN PT [Jain et al.]	CNN PT	N	ResNet-V2 pre-trained with ImageNet	ResNetV2 + Fine-Tunning	100

[46]	CNN PT + CNN PT [Mumtaz et al.]	CNN PT	CNN PT	Deep Multi-Net	DMN	100
[68]	CNN + LSTM [Halder et al.]	CNN	LSTM	CNN	Bi-LSTM	100
[71]	CNN + LSTM [Islam et al.]	CNN	LSTM	MobileNet	Separable Convolutional LSTM	100
[72]	CNN PT + LSTM [Mugunga et al.]	CNN PT	LSTM	VGG-16 pre-trained with ImageNet	Bi-ConvLSTM	100
[33]	Manual + Manual/CNN [Hu et al.]	Manual	Manual/CNN	TOP-ALCM	HandCraftedFeatures+CNN	100
[38]	Manual + CNN [Mohtavipour et al.]	Manual	CNN	Grayscale + Optical flow vectors + Motion Energy Image (MEI)	3-stream CNN	100
[73]	Manual + ML [Wintarti et al.]	Manual	ML	Discrete Wavelet Transform (DWT)	SVM	100
[80]	Transformer [Aktı et al.]	Transformer	N	ViT	N	100
[58]	CNN PT + LSTM [Islam et al.]	CNN PT	LSTM	VGG-19 pre-trained	LSTM	99.9
[65]	CNN [Ahmed et al.]	CNN	N	CNN-V4	N	99.03
[45]	CNN [Ehsan et al.]	CNN	N	CNN	N	99
[67]	CNN PT + LSTM [Asad et al.]	CNN PT	LSTM	VGG-16	Wide Dense Residual Blocks (WDRB) + LSTM	98.99
[81]	Skeleton-based (Manual) + CNN [Su et al.]	Skeleton-based (Manual)	CNN	Own skeleton-based algorithm	SPIL	98.5
[64]	CNN PT + LSTM [Sharma et al.]	CNN PT	LSTM	Xception pre-trained with ImageNet	LSTM	98.32
[74]	CNN PT + LSTM [Singh et al.]	CNN PT	LSTM	Two channel DarkNet19 pretrained on ImageNet	LSTM	97.41
[49]	CNN + LSTM [Vijeikis et al.]	CNN	LSTM	Time-distributed U-Net	LSTM	96.1

Action movies	[43]	CNN PT + LSTM [Jahlan et al.]	CNN PT	LSTM	Mobile Neural Architecture Search (MNAS) Pre-trained Automated	Conv-LSTM	96
	[101]	Audio (Manual) [Mahalle et al.]	Audio (Manual)	N	ZCR + AE + STE + RMS + SF + BW + VER + MFCC	ELM	94.32
	[79]	CNN + LSTM [Talha et al.]	CNN	LSTM	CNN	LSTM	92.2
	[50]	CNN PT + LSTM [Sernani et al.]	CNN PT	LSTM	CNN-3D trained with Sports 1M	Conv-LSTM	100
	[43]	CNN PT + LSTM [Jahlan et al.]	CNN PT	LSTM	Mobile Neural Architecture Search (MNAS) Pre-trained Automated	Conv-LSTM	100
	[38]	Manual + CNN [Mohtavipour et al.]	Manual	CNN	Grayscale+Optical flow vectors+Motion Energy Image (MEI)	3-stream CNN	100
	[69]	CNN [Appavu]	CNN	N	Optimized CNN	N	99.68
	[66]	CNN PT [Freire-Obregón et al.]	CNN PT	N	Pre-trained Two-stream inflated 3D ConvNets	N	99.45
	[48]	CNN PT [Gkountakos et al.]	CNN PT	N	ResNet-3D	N	99.31
	[68]	CNN + LSTM [Halder et al.]	CNN	LSTM	CNN	Bi-LSTM	98.64
	[32]	CNN + LSTM [Ullah et al.]	CNN	LSTM	Darknet + Residual Optical Flow CNN	M-LSTM	98.21
	[80]	Transformer [Aktı et al.]	Transformer	N	ViT	N	98
	[65]	CNN [Ahmed et al.]	CNN	N	CNN-V4	N	97.65
	[60]	CNN [Mahmoodi et al.]	CNN	N	CNN-3D	N	97.49
	[67]	CNN PT + LSTM [Asad et al.]	CNN PT	LSTM	VGG-16	Wide Dense Residual Blocks (WDRB) + LSTM	97.1

Violent flow	[67]	CNN PT + LSTM [Asad et al.]	CNN PT	LSTM	VGG-16	Wide Dense Residual Blocks (WDRB) + LSTM	97.1
	[67]	CNN PT + LSTM [Asad et al.]	CNN PT	LSTM	VGG-16	Wide Dense Residual Blocks (WDRB) + LSTM	97.1
	[77]	CNN PT + LSTM [Traoré et al.]	CNN PT	LSTM	VGG-16 pre-trained on INRA person dataset	Bi-GRU	95.5
	[75]	CNN PT + LSTM [Gupta et al.]	CNN PT	LSTM	VGG-16	Bi-LSTM	95.4
	[81]	Skeleton-based (Manual) + CNN [Su et al.]	Skeleton-based (Manual)	CNN	Own skeleton-based algorithm	SPIL	94.5
	[45]	CNN [Ehsan et al.]	CNN	N	CNN		94
	[78]	CNN PT + LSTM [Traoré et al.]	CNN PT	LSTM	Two channel EfficienNet-B0 pre-trained on ImageNet	Bi-LSTM	93.75
	[75]	CNN PT + LSTM [Gupta et al.]	CNN PT	LSTM	VGG-16	Bi-LSTM	92.2
	[31]	CNN PT [Huszár et al.]	CNN PT	N	Transfer-Learned X3D-M pre-trained on ImageNet	N	92
	[35]	Manual + ML [Lohithashva et al.]	Manual	ML	Local Orientation Pattern (LOOP)	SVM	91.54
	[33]	Manual + Manual/CNN [Hu et al.]	Manual	Manual/CNN	TOP-ALCM	HandCraftedFeatures + CNN	91
	[59]	CNN PT + LSTM [Mumtaz et al.]	CNN PT	LSTM	VGG-19	Bi-LSTM	89.63
	[79]	CNN + LSTM [Talha et al.]	CNN	LSTM	CNN	LSTM	77.31
	[62]	CNN PT [Bi et al.]	CNN PT	N	ResNet-18 pre-trained	N	94.6
	[72]	CNN PT + LSTM [Mugunga et al.]	CNN PT	LSTM	VGG-16 pre-trained with ImageNet	Bi-ConvLSTM	92.4

RWF-2000	[59]	CNN PT + LSTM [Mumtaz et al.]	CNN PT	LSTM	VGG-19	Bi-LSTM	90.47
	[71]	CNN + LSTM [Islam et al.]	CNN	LSTM	MobileNet	Separable Convolutional LSTM	89.75
	[82]	Skeleton-based (DL) + CNN [Zhou et al.]	Skeleton-based (DL)	CNN	TokenPose+HRNetW32	CNN-3D	89.45
	[81]	Skeleton-based (Manual) + CNN [Su et al.]	Skeleton-based (Manual)	CNN	Own skeleton-based algorithm	SPIL	89.3
	[39]	LSTM + LSTM [Fath et al.]	LSTM	LSTM	Conv-LSTM	GRU Network	88.2
	[31]	CNN PT [Huszár et al.]	CNN PT	N	Transfer-Learned X3D-M pre-trained on ImageNet	N	85
	[83]	CNN PT [Santos et al.]	CNN PT	N	X3D Neural Network pre-trained with Kinetics-400	X3D Neural Network pre-trained with Kinetics-400 + Fine tunning	84.75
	[49]	CNN + LSTM [Vijeikis et al.]	CNN	LSTM	Time-distributed U-Net	LSTM	82.0
	[36]	CNN PT [Chen et al.]	CNN PT	N	ResNet+Inception-V1 combination	N	72
	[83]	CNN PT [Santos et al.]	CNN PT	N	X3D Neural Network pre-trained with Kinetics-400	X3D Neural Network pre-trained with Kinetics-400 + Fine tunning	98
	[78]	CNN PT + LSTM [Traoré et al.]	CNN PT	LSTM	Two channel EfficienNet-B0 pre-trained on ImageNet	Bi-LSTM	96.74
	[31]	CNN PT [Huszár et al.]	CNN PT	N	Transfer-Learned X3D-M pre-trained on ImageNet	N	95.2

RLVS	[70]	Transformer [Kumar et al.]	Transformer	N	MSA	N	95.12
	[62]	CNN PT [Bi et al.]	CNN PT	N	ResNet-18 pre-trained	N	95
	[77]	CNN PT + LSTM [Traoré et al.]	CNN PT	LSTM	VGG-16 pre-trained on INRA person dataset	Bi-GRU	90.25
	[76]	CNN PT [Jain et al.]	CNN PT	N	ResNet-V2 pre-trained with ImageNet	ResNetV2 + Fine-Tunning	86.79
	[84]	CNN + CNN [Jayasimhan et al.]	CNN	CNN	CNN-2D	CNN-3D	84.5
UFC crime dataset	[85]	CNN PT [Adithya et al.]	CNN PT	N	3D-CNN	N	99.57
	[72]	CNN PT + LSTM [Mugunga et al.]	CNN PT	LSTM	VGG-16 pre-trained with ImageNet	Bi-ConvLSTM	99.1
	[31]	CNN PT [Huszár et al.]	CNN PT	N	Transfer-Learned X3D-M pre-trained on ImageNet	N	84.2
BEHAVE	[72]	CNN PT + LSTM [Mugunga et al.]	CNN PT	LSTM	VGG-16 pre-trained with ImageNet	Bi-ConvLSTM	99.3
	[67]	CNN PT + LSTM [Asad et al.]	CNN PT	LSTM	VGG-16	Wide Dense Residual Blocks (WDRB) + LSTM	95.9
	[87]	CNN PT + CNN [Qu et al.]	CNN PT	CNN	C3D pre-trained with THUMOS2014 dataset	DC3D	73.8
Surveillance Camera	[80]	Transformer [Aktı et al.]	Transformer	N	ViT	N	84.6
	[39]	LSTM + LSTM [Fath et al.]	LSTM	LSTM	Conv-LSTM	GRU Network	75.9
	[32]	CNN + LSTM [Ullah et al.]	CNN	LSTM	Darknet+Residual Optical Flow CNN	M-LSTM	74
AIRTLab	[50]	CNN PT + LSTM [Sernani et al.]	CNN PT	LSTM	CNN-3D trained with Sports 1M	Conv-LSTM	97.32
	AIRTLab	[91]	CNN PT + LSTM [Contardo et al.]	CNN PT	LSTM	MobileNet-V2	ConvLSTM

Industrial Surveillance	[59]	CNN PT + LSTM [Mumtaz et al.]	CNN PT	LSTM	VGG-19	Bi-LSTM	81.22
	[39]	LSTM + LSTM [Fath et al.]	LSTM	LSTM	Conv-LSTM	GRU Network	80
AVA	[34]	CNN [Monteiro et al.]	CNN	N	SlowfastNetwork	SlowfastNetworw + X3D Network	84.5
XD-V	[31]	CNN PT [Huszár et al.]	CNN PT	N	Transfer-Learned X3D-M pre-trained on ImageNet	N	89.31

The accuracy values achieved by the selected articles are shown in Table 13 for the five most used datasets in order of highest to lowest number of citations; as shown in Figure 2: hockey fights, action movies, violent flow, RWF-2000 and RLVS. It can be seen how the best results are obtained with combinations of CNN and CNN + LSTM, as there is a large difference in accuracy between those combinations and other types of algorithms. Nevertheless, other articles such as of Hu et al. [33] and Mohtavipour et al. [38] are an exception as they achieve excellent results through the use of manual feature extraction methods in combination with CNN. Overall, it has been observed that the choice of the dataset affects the obtained results, as datasets with more varied scenes, poorer image quality, more people in the scene, etc., have poorer results. It has also been observed how the use of CNN and CNN + LSTM renders the highest accuracy. Nevertheless, good results have been obtained by applying manual feature extraction and CNN, or skeleton-based deep learning algorithms and transformers.

5.9. Trustworthiness in the Selected Articles

Section 1 highlighted the criticality of developing trustworthy artificial intelligence, so that its operation and results can be understood, as opposed to unintelligible, opaque, black-box algorithms. We referred to the European Union (EU) report [10] which defines trustworthy artificial intelligence in three pillars, namely, lawful, ethical and robust. Subsection 5.9.1 addresses the algorithms which had been used in the selected articles for violence detection, and which indirectly promote the explainability of the violence detection process, despite the fact that the concept of trustworthiness was not always explicitly mentioned.

In the selected articles, we searched for mentions of keywords that make up trustworthy AI according to the European Union report [10]: trustworthy, lawful, ethical, robust and explainable. Out of the total of 63 selected articles, 33 papers (52%) do not refer to or make use of trustworthy artificial intelligence; 30 papers (47%) mention trustworthy artificial intelligence to some degree, however it is not enough to achieve trustworthy models. Table 14 presents the mentions of these keywords in the selected articles, ordered from the highest to lowest number of mentions. The *Trustworthy Components* column contains the mentioned elements that make up the trustworthy AI. The *Article Location Mention* column states where in the text that word is found; The term *Mention* means that the the keyword is stated in the introduction, related work or conclusion but does not refer to the model proposed in the article. If the term is *Model* it means that the keyword refers to the model developed in the article. If it is *Dataset* it means that the keyword refers to the dataset developed in the paper. Finally, the *Cite Count* and *Cite* columns contain the number of citations and the citations of the selected articles that mention those terms.

Table 14 shows how the most cited term is *Robust*, as both a mention in the introductory/concluding sections and in reference to the proposed model. This is because, as discussed in Section 5.4, violence detection is principally binary problem, i.e., there may be violence or not. It was discussed how a widespread method for analyzing the results is the use of a confusion matrix, which is based on the analysis of true positives, true negatives, false positives, and false negatives. Analysing the elements of the confusion matrix allows

Table 14: Mention of trustworthy AI components in the selected articles.

Trustworthy Components	**Article Location Mention**	**Cite Count**	**Cite**
Robust	Mention	15	[32], [60], [63], [46], [50], [66], [94], [34], [82], [47], [48], [77], [81], [106], [105]
	Model	10	[61], [31], [64], [50], [66], [101], [41], [76], [80], [106]
Ethical	Dataset	1	[40]
Explainable	Model	1	[33]

us to assess how robust the algorithm is to other actions that appear in the video. The detection of violence is complex given that it may be confused with non-violent actions that may resemble physical aggression, such as a hug. Thus, robustness is a widely cited term because of both, the binary nature of the problem and its ambiguity with respect to other actions. The concept of *ethics* is only mentioned by Nadeem et al. [40] who addressed the ethical problems involved in labeling violent scenes for the generation of datasets facing the people involved in the videos. The authors therefore proposed the development of an automatic labeling algorithm with ethical guarantees. The mention of *explainability* is also much lower. Hu et al. [33] addressed the problem of "black-box" deep learning algorithms and pointed to uncertainty quantification (UQ) as a means of quantifying the level of uncertainty of these algorithms, with the application of Monte Carlo method and Hierarchical Cross-Validation. No direct mention of the term *trustworthy* or *lawful* has been found.

5.9.1. Algorithms in the Selected Articles for Increased Explainability

Explainability of Artificial Intelligence (XAI) focuses on the ability to understand and explain the decision-making process of an artificial intelligence model; where opacity in AI models can be a challenge to acceptance and adoption, especially in crucial applications where understanding the reasoning behind automated decisions is critical [10]. As shown in Table 14, there is a low number of references to trustworthy artificial intelligence and related concepts. In Section 5.9.1 we compile, from among the algorithms that had been employed in the different parts of the violence detection process, those that can boost the trustworthiness of the detection system. None of the algorithms and processes that are mentioned below had been used with the direct intention of achieving trustworthiness; nevertheless, given their characteristics, they can serve for this purpose.

In the process of **characteristic frame extraction** a division was made in Table 10 into different categories of algorithms according to their operation. The algorithms belonging to the category of *object detection* consist in the detection of the persons appearing in the video, *Yolo* is an examples of algorithm based on the detection of objects.

These algorithms can boost explainability as they help to understand the reason why the algorithm selected a given set of frames as potential acts of violence.

In the process of providing **input to the violence detection algorithms**, the different types of inputs were divided in Table 11 into categories according to whether they were picture-centered, motion-centered or audio-centered. Inputs from motion-focused algorithms can provide greater explainability, as in the case of the *separate motion energy picture*, which emphasizes areas with rapid pixel intensity changes across frames. These inputs can help to understand the results to be obtained as output.

Finally, in the process of **violence detection**, transformers, as well as CNNs and LSTMs that extract spatial and temporal information, make it difficult to understand their process and result. Machine Learning algorithms as well as manual feature extraction have simpler logic that may be somewhat more understandable. Although the simpler the algorithm the more comprehensible it may be, the ideal choice would be complex algorithms which are at the same time highly trustworthy. In this regard, skeleton-based algorithms are highly comprehensible, since their purpose is to detect the joints of the people present in the video and their evolution over time. Thus, one of the algorithms that in our opinion has a greater degree of explainability is the one developed by Naik et al. [94]. In that article, the authors used heat maps to highlight the people detected in the video in the pre-processing. The Deep Learning skeleton-based algorithm was used to mark the joints in green, yellow and red color code, which indicated non-violence, potential violence and definite violence, respectively.

Overall, the concern of the selected works for the development of a trustworthy artificial intelligence for the detection of physical aggression in video is practically null. It has been observed that the concept of robustness has been used to describe the good results of the developed algorithms, given the binary nature of the problem and the quantification of this by means of elements of a confusion matrix. In Section 5.9.1, algorithms were identified that can enhance the process's understandability, although the primary objective of the authors was not specifically to increase explainability but rather to make the explanation clearer for the reader.

6. Conclusion and Future Work

Physical aggression is a serious and widespread concern in our society, affecting individuals globally and impacting nearly every aspect of life. This influence extends not just to the immediate victims but also to their families, the broader community. AI-based violence detection is the last barrier to defend victims from its attacks and can perform large-scale surveillance constantly over time [24] [72]. Furthermore, in recent years, the focus on establishing trustworthy artificial intelligence has grown significantly. Prominent organizations, such as the European Union have released reports to define and set standards for these concepts [10]. This article presents a systematic mapping study of violence detection in video focusing on trustowrthy artificial intelligence. Given the lack of up-to-date reviews that provide a comprehensive analysis of video violence detection, this paper presents a systematic mapping study that also aims to highlight the limited use and mention of trustworthy artificial intelligence techniques in recent articles.

A rigorous methodology has been established, from defining objectives to database selection and filtering criteria, ensuring a meticulous approach in gathering and selecting articles. A total of 28 violence detection datasets have been collected and categorised, showing how many new datasets have been developed in recent years. A total of 13 evaluation parameters used in violence detection have been compiled, where "Accuracy" is the most used. 21 key feature extraction methods have been collected and categorised in the selected articles, which are essential for a lighter commentary on the process of video violence detection. The diversity of inputs for violence detection algorithms was analyzed, with RGB video emerging as the most common input focus for feature extraction. Different types of algorithms and classifiers used in violence detection were compared, showcasing the strengths and weaknesses of traditional versus deep learning methods. Analysis of accuracy in violence detection pointed to *CNN* and *CNN+LSTM* combinations as often yielding superior results, although other algorithm combinations also showed promise.

Of the selected papers, none make explicit use of trustworthy artificial intelligence techniques with the aim of making the developed model more reliable, although a total of 15 articles mention the robustness of their model, this being one of the three pillars of trustworthy artificial intelligence according to the European Union. On the other hand, several papers employ techniques or algorithms in the process of detecting violence in video that can facilitate the explainability of the model; for example, the use of object detection algorithms or Skeleton-based algorithms. These findings collectively underline the current landscape of violence detection research, emphasizing the need for comprehensive assessments, diverse datasets, and continued exploration of algorithmic approaches for more accurate and trustworthy outcomes.

In a future research we will develop a proposal in which the algorithms that have been highlighted in Section 5.9.1, are applied throughout in the process of violence detection. This will be done with the aim of increasing the explainability of the decisions and results taken by the algorithm. The algorithms with the best results, namely CNN and the combination of *CNN + LSTM*, are not explainable. Nevertheless, explainable artificial intelligence algorithms could be developed to enable the application of CNN or LSTM in an explainable manner. As future work, a review article will be produced expanding on each finding from this systematic mapping study to offer a deeper analysis and enhance understanding within this research area. Another further study is required regarding the ethical and legal aspects of the algorithms and datasets employed in the process of violence detection, as those would be critical elements in a trustworthy artificial intelligence system, according to the report developed by the European Union [10].

Acknowledgements

This research was partially supported by the project Preven-TIA "Smart Platform for the prevention of ambient and ergonomic labour risks in the post-COVID era based on Edge Computing and Trustworthy Artificial Intelligence", financed by Consejería de Industria, Comercio y Empleo de la Junta de Castilla y León (Ref. INVESTUN/22/SA/0003).

References

[1] D. Long, L. Liu, M. Xu, J. Feng, J. Chen, L. He, Ambient population and surveillance cameras: The guardianship role in street robbers' crime location choice, Cities 115 (2021) 103223.

[2] P. Negre, R. S. Alonso, J. Prieto, A. G. Arrieta, J. M. Corchado, Review of physical aggression detection techniques in video using explainable artificial intelligence, in: International Symposium on Ambient Intelligence, Springer, 2023, pp. 53–62.

[3] A. Muarifah, R. Mashar, I. H. M. Hashim, N. H. Rofiah, F. Oktaviani, Aggression in adolescents: The role of mother-child attachment and self-esteem, Behavioral Sciences 12 (5) (2022).

[4] S. Hillis, J. Mercy, A. Amobi, H. Kress, Global prevalence of past-year violence against children: a systematic review and minimum estimates, Pediatrics 137 (3) (2016).

[5] S. Wilkinson, Meet the heroic campaigners making cities safe for women, Global Street Harassment-Making Street Safer: Action Aid (2016).

[6] A. Enaifoghe, M. Dlelana, A. A. Durokifa, N. P. Dlamini, The prevalence of gender-based violence against women in south africa: A call for action, African Journal of Gender, Society & Development 10 (1) (2021) 117.

[7] F. Fekih-Romdhane, D. Malaeb, A. Sarray El Dine, S. Obeid, S. Hallit, The relationship between smartphone addiction and aggression among lebanese adolescents: the indirect effect of cognitive function, BMC pediatrics 22 (1) (2022) 735.

[8] F. Jing, L. Liu, S. Zhou, J. Song, L. Wang, H. Zhou, Y. Wang, R. Ma, Assessing the impact of street-view greenery on fear of neighborhood crime in guangzhou, china, International journal of environmental research and public health 18 (1) (2021) 311.

[9] H. Yue, H. Xie, L. Liu, J. Chen, Detecting people on the street and the streetscape physical environment from baidu street view images and their effects on community-level street crime in a chinese city, ISPRS International Journal of Geo-Information 11 (3) (2022) 151.

[10] Ethics guidelines for trustworthy ai, European Commission (2019).

[11] ISO/IEC, Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence, Technical Report ISO/IEC TR 24028:2020 (2020).

[12] K. Abhishek, D. Kamath, Attribution-based xai methods in computer vision: A review, arXiv preprint arXiv:2211.14736 (2022).

[13] T. Speith, A review of taxonomies of explainable artificial intelligence (xai) methods, in: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, 2022, pp. 2239–2250.

[14] S. Afra, R. Alhajj, Early warning system: From face recognition by surveillance cameras to social media analysis to detecting suspicious people, Physica A: Statistical Mechanics and its Applications 540 (2020) 123151.

[15] S. Vosta, K.-C. Yow, A cnn-rnn combined structure for real-world violence detection in surveillance cameras, Applied Sciences 12 (3) (2022).

[16] R. S. Alonso, I. Sittón-Candanedo, R. Casado-Vara, J. Prieto, J. M. Corchado, Deep reinforcement learning for the management of software-defined networks and network function virtualization in an edge-iot architecture, Sustainability 12 (14) (2020) 5706.

[17] Z. Ageed, S. Zeebaree, A comprehensive survey of big data mining approaches in cloud systems 1 (2021) 29–38.

[18] B. Sharma, D. Koundal, R. A. Ramadan, J. M. Corchado, Emerging sensor communication network-based ai/ml driven intelligent iot, Sensors 23 (18) (2023).

[19] O. Ali, A. Shrestha, J. Soar, S. F. Wamba, Cloud computing-enabled healthcare opportunities, issues, and applications: A systematic review, International Journal of Information Management 43 (2018) 146–158.

[20] D. Ding, Z. Ma, D. Chen, Q. Chen, Z. Liu, F. Zhu, Advances in video compression system using deep neural network: A review and case studies, Proceedings of the IEEE 109 (9) (2021) 1494–1520.

[21] A. Rivas, A. González-Briones, G. Hernández, J. Prieto, P. Chamoso, Artificial neural network analysis of the academic performance of students in virtual learning environments, Neurocomputing 423 (2021) 713–720.

[22] Y. Zhang, Z. Fang, J. Fan, Generalization analysis of deep cnns under maximum correntropy criterion, Neural Networks (2024) 106226doi:https://doi.org/10.1016/j.neunet.2024.106226.
URL https://www.sciencedirect.com/science/article/pii/S0893608024001503

[23] B. Omarov, S. Narynov, Z. Zhumanov, A. Gumar, M. Khassanova, State-of-the-art violence detection techniques in video surveillance security systems: a systematic review, PeerJ Computer Science 8 (2022) e920.

[24] H. Yao, X. Hu, A survey of video violence detection, Cyber-Physical Systems 9 (1) (2023) 1–24.

[25] M. Siddique, M. S. Islam, R. Sinthy, K. Mohima, M. Kabir, A. H. Jibon, M. Biswas, State-of-the-art violence detection techniques: A review, Asian Journal of Research in Computer Science (2022) 29–42.

[26] G. Kaur, S. Singh, Violence detection in videos using deep learning: A survey, Advances in Information Communication Technology and Computing: Proceedings of AICTC 2021 (2022) 165–173.

[27] M. S. M. Shubber, Z. T. M. Al-Ta'i, A review on video violence detection approaches, International Journal of Nonlinear Analysis and Applications 13 (2) (2022) 1117–1130.

[28] B. A. Kitchenham, D. Budgen, O. P. Brereton, Using mapping studies as the basis for further research–a participant-observer case study, Information and Software Technology 53 (6) (2011) 638–651.

[29] K. Petersen, S. Vakkalanka, L. Kuzniarz, Guidelines for conducting systematic mapping studies in software engineering: An update, Information and software technology 64 (2015) 1–18.

[30] K. Petersen, R. Feldt, S. Mujtaba, M. Mattsson, Systematic mapping studies in software engineering. ease'08 proceedings of the 12th international conference on evaluation and assessment in software engineering, 68–77 (2008).

[31] V. D. Huszár, V. K. Adhikarla, I. Négyesi, C. Krasznay, Toward fast and accurate violence detection for automated video surveillance applications, IEEE Access 11 (2023) 18772–18793.

[32] F. U. M. Ullah, M. S. Obaidat, K. Muhammad, A. Ullah, S. W. Baik, F. Cuzzolin, J. J. Rodrigues, V. H. C. de Albuquerque, An intelligent system for complex violence pattern analysis and detection, International Journal of Intelligent Systems 37 (12) (2022) 10400–10422.

[33] X. Hu, Z. Fan, L. Jiang, J. Xu, G. Li, W. Chen, X. Zeng, G. Yang, D. Zhang, Top-alcm: A novel video analysis method for violence detection in crowded scenes, Information Sciences 606 (2022) 313–327.

[34] C. Monteiro, D. Durães, Modelling a framework to obtain violence detection with spatial-temporal action localization, in: World Conference on Information Systems and Technologies, Springer, 2022, pp. 630–639.

[35] B. Lohithashva, V. M. Aradhya, Violent video event detection: a local optimal oriented pattern based approach, in: Applied Intelligence and Informatics: First International Conference, AII 2021, Nottingham, UK, July 30–31, 2021, Proceedings 1, Springer, 2021, pp. 268–280.

[36] Y. Chen, B. Zhang, Y. Liu, Estn: Exacter spatiotemporal networks for violent action recognition, in: 2021 IEEE 6th International Conference on Signal and Image Processing (ICSIP), IEEE, 2021, pp. 44–48.

[37] A. Srivastava, T. Badal, A. Garg, A. Vidyarthi, R. Singh, Recognizing human violent action using drone surveillance within real-time proximity, Journal of Real-Time Image Processing 18 (2021) 1851–1863.

[38] S. M. Mohtavipour, M. Saeidi, A. Arabsorkhi, A multi-stream cnn for deep violence detection in video sequences using handcrafted features, The Visual Computer (2022) 1–16.

[39] F. U. M. Ullah, K. Muhammad, I. U. Haq, N. Khan, A. A. Heidari, S. W. Baik, V. H. C. de Albuquerque, Ai-assisted edge vision for violence detection in iot-based industrial surveillance networks, IEEE Transactions on Industrial Informatics 18 (8) (2021) 5359–5370.

[40] M. S. Nadeem, F. Kurugollu, S. Saravi, H. F. Atlam, V. N. Franqueira, Deep labeller: automatic bounding box generation for synthetic violence detection datasets, Multimedia Tools and Applications (2023) 1–18.

[41] H. Kim, H. Jeon, D. Kim, J. Kim, Lightweight framework for the violence and falling-down event occurrence detection for surveillance videos, in: 2022 13th International Conference on Information and Communication Technology Convergence (ICTC), IEEE, 2022, pp. 1629–1634.

[42] R. Madhavan, Utkarsh, J. Vidhya, Violence detection from cctv footage using optical flow and deep learning in inconsistent weather and lighting conditions, in: Advances in Computing and Data Sciences: 5th International Conference, ICACDS 2021, Nashik, India, April 23–24, 2021, Revised Selected Papers, Part I 5, Springer, 2021, pp. 638–647.

[43] H. M. B. Jahlan, L. A. Elrefaei, Mobile neural architecture search network and convolutional long short-term memory-based deep features toward detecting violence from video, Arabian Journal for Science and Engineering 46 (9) (2021) 8549–8563.

[44] T. Z. Ehsan, M. Nahvi, S. M. Mohtavipour, An accurate violence detection framework using unsupervised spatial–temporal action translation network, The Visual Computer (2023) 1–21.

[45] T. Z. Ehsan, S. M. Mohtavipour, Vi-net: a deep violent flow network for violence detection in video sequences, in: 2020 11th International Conference on Information and Knowledge Technology (IKT), IEEE, 2020, pp. 88–92.

[46] A. Mumtaz, A. Bux Sargano, Z. Habib, Fast learning through deep multi-net cnn model for violence recognition in video surveillance, The Computer Journal 65 (3) (2022) 457–472.

[47] K. Aarthy, A. A. Nithya, Crowd violence detection in videos using deep learning architecture, in: 2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon), IEEE, 2022, pp. 1–6.

[48] K. Gkountakos, K. Ioannidis, T. Tsikrika, S. Vrochidis, I. Kompatsiaris, A crowd analysis framework for detecting violence scenes, in: Proceedings of the 2020 International Conference on Multimedia Retrieval, 2020, pp. 276–280.

[49] R. Vijeikis, V. Raudonis, G. Dervinis, Efficient violence detection in surveillance, Sensors 22 (6) (2022) 2216.

[50] P. Sernani, N. Falcionelli, S. Tomassini, P. Contardo, A. F. Dragoni, Deep learning for automatic violence detection: Tests on the airtlab dataset, IEEE Access 9 (2021) 160580–160595.

[51] E. Bermejo Nievas, O. Deniz Suarez, G. Bueno García, R. Sukthankar, Violence detection in video using computer vision techniques, in: Computer Analysis of Images and Patterns: 14th International Conference, CAIP 2011, Seville, Spain, August 29-31, 2011, Proceedings, Part II 14, Springer, 2011, pp. 332–339.

[52] T. Hassner, Y. Itcher, O. Kliper-Gross, Violent flows: Real-time detection of violent crowd behavior, in: 2012 IEEE computer society conference on computer vision and pattern recognition workshops, IEEE, 2012, pp. 1–6.

[53] M. Cheng, K. Cai, M. Li, Rwf-2000: An open large scale video database for violence detection, in: 2020 25th International Conference on Pattern Recognition (ICPR), 2021, pp. 4183–4190.

[54] M. M. Soliman, M. H. Kamal, M. A. El-Massih Nashed, Y. M. Mostafa, B. S. Chawky, D. Khattab, Violence recognition from videos using deep learning techniques, in: 2019 Ninth International Conference on Intelligent Computing and Information Systems (ICICIS), 2019, pp. 80–85.

[55] W. Sultani, C. Chen, M. Shah, Real-world anomaly detection in surveillance videos, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 6479–6488.

[56] P. Wu, J. Liu, Y. Shi, Y. Sun, F. Shao, Z. Wu, Z. Yang, Not only look, but also listen: Learning multimodal violence detection under weak supervision, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16, Springer, 2020, pp. 322–339.

[57] C. Schuldt, I. Laptev, B. Caputo, Recognizing human actions: a local svm approach, in: Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004., Vol. 3, IEEE, 2004, pp. 32–36.

[58] M. S. Islam, M. M. Hasan, S. Abdullah, J. U. M. Akbar, N. Arafat, S. A. Murad, A deep spatio-temporal network for vision-based sexual harassment detection, in: 2021 Emerging Technology in Computing, Communication and Electronics (ETCCE), IEEE, 2021, pp. 1–6.

[59] N. Mumtaz, N. Ejaz, S. Aladhadh, S. Habib, M. Y. Lee, Deep multi-scale features fusion for effective violence detection and control charts visualization, Sensors 22 (23) (2022) 9383.

[60] J. Mahmoodi, H. Nezamabadi-pour, D. Abbasi-Moghadam, Violence detection in videos using interest frame extraction and 3d convolutional neural network, Multimedia tools and applications 81 (15) (2022) 20945–20961.

[61] M. Magdy, M. W. Fakhr, F. A. Maghraby, Violence 4d: Violence detection in surveillance using 4d convolutional neural networks, IET Computer Vision (2023).

[62] Y. Bi, D. Li, Y. Luo, Combining keyframes and image classification for violent behavior recognition, Applied Sciences 12 (16) (2022) 8014.

[63] S. G. Jaiswal, S. W. Mohod, Classification of violent videos using ensemble boosting machine learning approach with low level features.

[64] S. Sharma, B. Sudharsan, S. Naraharisetti, V. Trehan, K. Jayavel, A fully integrated violence detection system using cnn and lstm., International Journal of Electrical & Computer Engineering (2088-8708) 11 (4) (2021).

[65] M. Ahmed, M. Ramzan, H. U. Khan, S. Iqbal, M. A. Khan, J.-I. Choi, Y. Nam, S. Kadry, Real-time violent action recognition using key frames extraction and deep learning (2021).

[66] D. Freire-Obregón, P. Barra, M. Castrillón-Santana, M. D. Marsico, Inflated 3d convnet context analysis for violence detection, Machine Vision and Applications 33 (2022) 1–13.

[67] M. Asad, J. Yang, J. He, P. Shamsolmoali, X. He, Multi-frame feature-fusion-based model for violence detection, The Visual Computer 37 (2021) 1415–1431.

[68] R. Halder, R. Chatterjee, Cnn-bilstm model for violence detection in smart surveillance, SN Computer science 1 (4) (2020) 201.

[69] N. Appavu, et al., Violence detection based on multisource deep cnn with handcraft features, in: 2023 IEEE International Conference on Advanced Systems and Emergent Technologies (IC_ASET), IEEE, 2023, pp. 1–6.

[70] A. Kumar, A. Shetty, A. Sagar, A. Charushree, P. Kanwal, Indoor violence detection using lightweight transformer model, in: 2023 4th International Conference for Emerging Technology (INCET), IEEE, 2023, pp. 1–6.

[71] Z. Islam, M. Rukonuzzaman, R. Ahmed, M. H. Kabir, M. Farazi, Efficient two-stream network for violence detection using separable convolutional lstm, in: 2021 International Joint Conference on Neural Networks (IJCNN), IEEE, 2021, pp. 1–8.

[72] I. Mugunga, J. Dong, E. Rigall, S. Guo, A. H. Madessa, H. S. Nawaz, A frame-based feature model for violence detection from surveillance cameras using convlstm network, in: 2021 6th International Conference on Image, Vision and Computing (ICIVC), IEEE, 2021, pp. 55–60.

[73] A. Wintarti, R. D. I. Puspitasari, E. M. Imah, Violent videos classification using wavelet and support vector machine, in: 2022 International Conference on ICT for Smart Society (ICISS), IEEE, 2022, pp. 01–05.

[74] N. Singh, O. Prasad, T. Sujithra, Deep learning-based violence detection from videos, in: Intelligent Data Engineering and Analytics: Proceedings of the 9th International Conference on Frontiers in Intelligent Computing: Theory and Applications (FICTA 2021), Springer, 2022, pp. 323–332.

[75] H. Gupta, S. T. Ali, Violence detection using deep learning techniques, in: 2022 International Conference on Emerging Techniques in Computational Intelligence (ICETCI), IEEE, 2022, pp. 121–124.

[76] A. Jain, D. K. Vishwakarma, Deep neuralnet for violence detection using motion features from dynamic images, in: 2020 third international conference on smart systems and inventive technology (ICSSIT), IEEE, 2020, pp. 826–831.

[77] A. Traoré, M. A. Akhloufi, 2d bidirectional gated recurrent unit convolutional neural networks for end-to-end violence detection in videos, in: International Conference on Image Analysis and Recognition, Springer, 2020, pp. 152–160.

[78] A. Traoré, M. A. Akhloufi, Violence detection in videos using deep recurrent and convolutional neural networks, in: 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC), IEEE, 2020, pp. 154–159.

[79] K. R. Talha, K. Bandapadya, M. M. Khan, Violence detection using computer vision approaches, in: 2022 IEEE World AI IoT Congress (AIIoT), IEEE, 2022, pp. 544–550.

[80] Ş. Aktı, F. Ofli, M. Imran, H. K. Ekenel, Fight detection from still images in the wild, in: Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2022, pp. 550–559.

[81] Y. Su, G. Lin, J. Zhu, Q. Wu, Human interaction learning on 3d skeleton point clouds for video violence recognition, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16, Springer, 2020, pp. 74–90.

[82] L. Zhou, End-to-end video violence detection with transformer, in: 2022 5th International Conference on Pattern Recognition and Artificial Intelligence (PRAI), IEEE, 2022, pp. 880–884.

[83] F. Santos, D. Durães, F. S. Marcondes, S. Lange, J. Machado, P. Novais, Efficient violence detection using transfer learning, in: International Conference on Practical Applications of Agents and Multi-Agent Systems, Springer, 2021, pp. 65–75.

[84] A. Jayasimhan, P. Pabitha, A hybrid model using 2d and 3d convolutional neural networks for violence detection in a video dataset, in: 2022 3rd International Conference on Communication, Computing and Industry 4.0 (C2I4), IEEE, 2022, pp. 1–5.

[85] H. Adithya, H. Lekhashree, S. Raghuram, Violence detection in drone surveillance videos, in: International Conference on Smart Computing and Communication, Springer, 2023, pp. 703–713.

[86] S. Blunsden, R. Fisher, The behave video dataset: ground truthed video for multi-person behavior classification, Annals of the BMVA 4 (1-12) (2010) 4.

[87] W. Qu, T. Zhu, J. Liu, J. Li, A time sequence location method of long video violence based on improved c3d network, The Journal of Supercomputing 78 (18) (2022) 19545–19565.

[88] Ş. Aktı, G. A. Tataroğlu, H. K. Ekenel, Vision-based fight detection from surveillance cameras, in: 2019 Ninth International Conference on Image Processing Theory, Tools and Applications (IPTA), IEEE, 2019, pp. 1–6.

[89] Y. Shang, X. Wu, R. Liu, Multimodal violent video recognition based on mutual distillation, in: Chinese Conference on Pattern Recognition and Computer Vision (PRCV), Springer, 2022, pp. 623–637.

[90] P. Sernani, N. Falcionelli, S. Tomassini, P. Contardo, A. F. Dragoni, Deep learning for automatic violence detection: Tests on the airtlab dataset, IEEE Access 9 (2021) 160580–160595.

[91] P. Contardo, S. Tomassini, N. Falcionelli, A. F. Dragoni, P. Sernani, Combining a mobile deep neural network and a recurrent layer for violence detection in videos (2023).

[92] Y. Ji, Y. Wang, J. Kato, K. Mori, Predicting violence rating based on pairwise comparison, IEICE TRANSACTIONS on Information and Systems 103 (12) (2020) 2578–2589.

[93] A. Srivastava, T. Badal, P. Saxena, A. Vidyarthi, R. Singh, Uav surveillance for violence detection and individual identification, Automated Software Engineering 29 (1) (2022) 28.

[94] A. J. Naik, M. Gopalakrishna, Deep-violence: individual person violent activity detection in video, Multimedia Tools and Applications 80 (12) (2021) 18365–18380.

[95] S.-T. Cheng, C.-W. Hsu, G.-J. Horng, C.-R. Jiang, Video reasoning for conflict events through feature extraction, The Journal of Supercomputing 77 (2021) 6435–6455.

[96] M. Sj¨oberg, Y. Baveye, H. Wang, V. L. Quang, B. Ionescu, E. Dellandréa, M. Schedl, C.-H. Demarty, L. Chen, The mediaeval 2015 affective impact of movies task., in: MediaEval, Vol. 1436, 2015.

[97] Z. Zheng, W. Zhong, L. Ye, L. Fang, Q. Zhang, Violent scene detection of film videos based on multi-task learning of temporal-spatial features, in: 2021 IEEE 4th International Conference on Multimedia Information Processing and Retrieval (MIPR), IEEE, 2021, pp. 360–365.

[98] A. Li, M. Thotakuri, D. A. Ross, J. Carreira, A. Vostrikov, A. Zisserman, The ava-kinetics localized human actions video dataset (2020). arXiv:2005.00214.

[99] Q. Liang, C. Cheng, Y. Li, K. Yang, B. Chen, Fusion and visualization design of violence detection and geographic video, in: Theoretical Computer Science: 39th National Conference of Theoretical Computer Science, NCTCS 2021, Yinchuan, China, July 23–25, 2021, Revised Selected Papers 39, Springer, 2021, pp. 33–46.

[100] C.-H. Demarty, C. Penet, M. Soleymani, G. Gravier, Vsd, a public dataset for the detection of violent scenes in movies: design, annotation, analysis and evaluation, Multimedia Tools and Applications 74 (2015) 7379–7404.

[101] M. D. Mahalle, D. V. Rojatkar, Audio based violent scene detection using extreme learning machine algorithm, in: 2021 6th international conference for convergence in technology (I2CT), IEEE, 2021, pp. 1–8.

[102] M. D. Mahalle, D. V. Rojatkar, Audio based violent scene detection using extreme learning machine algorithm, in: 2021 6th International Conference for Convergence in Technology (I2CT), 2021, pp. 1–8.

[103] S. Narynov, Z. Zhumanov, A. Gumar, M. Khassanova, B. Omarov, Detecting school violence using artificial intelligence to interpret surveillance video sequences, in: Advances in Computational Collective Intelligence: 13th International Conference, ICCCI 2021, Kallithea, Rhodes, Greece, September 29–October 1, 2021, Proceedings 13, Springer, 2021, pp. 401–412.

[104] U. Rachna, V. Guruprasad, S. D. Shindhe, S. Omkar, Real-time violence detection using deep neural networks and dtw, in: International Conference on Computer Vision and Image Processing, Springer, 2022, pp. 316–327.

[105] L.-P. Hung, C.-W. Yang, L.-H. Lee, C.-L. Chen, Constructing a violence recognition technique for elderly patients with lower limb disability, in: International Conference on Smart Grid and Internet of Things, Springer, 2021, pp. 24–37.

[106] Z. Zhang, D. Yuan, X. Li, S. Su, Violent target detection based on improved yolo network, in: International Conference on Artificial Intelligence and Security, Springer, 2022, pp. 480–492.