Leveraging Artificial Intelligence to Fight (Cyber)Bullying for Human Well-being: The BullyBuster Project

Giulia Orrù[1,*], Antonio Galli[3], Vincenzo Gattulli[2], Michela Gravina[3], Stefano Marrone[3], Marco Micheletto[1], Angela Procaccino[4], Wanda Nocerino[4], Grazia Terrone[5], Donatella Curtotti[4], Donato Impedovo[2], Gian Luca Marcialis[1] and Carlo Sansone[3]

[1] *University of Cagliari, piazza d'Armi, 09123, Cagliari, Italy*
[2] *University of Bari, Via Edorado Orabona 4, 70121, Bari, Italy*
[3] *University of Naples Federico II, Via Claudio 21, 80125 Naples, Italy*
[4] *University of Foggia, Via Antonio Gramsci 89, 71122 Foggia, Italy*
[5] *Tor Vergata University, Via Columbia 1, 00133 Roma, Italy*

Abstract

Bullying and cyberbullying are phenomena which, due to their growing diffusion, have become a real social emergency. In this context, artificial intelligence can be a powerful weapon to identify episodes of violence and fight bullying both in the virtual and in the real world. Through machine learning, it is possible to detect the language patterns used by bullies and their victims and develop rules to detect cyberbullying content automatically. The BullyBuster project merges the know-how of four interdisciplinary research groups to develop a framework useful for maintaining psycho-physical well-being in educational contexts.

Keywords

bullying, detection, crowd, text analysis, keystroke dynamics, deepfake

1. Introduction

Bullying and cyberbullying are pervasive social issues that adversely affect the well-being and mental health of millions of children adolescents and adults worldwide. These harmful behaviors impact the immediate victims and contribute to a toxic environment in educational institutions, workplaces, and online spaces. With the growing reliance on digital platforms for communication and socialization, the need to develop innovative solutions to identify, prevent, and address bullying and cyberbullying has become increasingly critical.

In this paper, we present the "BullyBuster - A framework for bullying and cyberbullying action detection by computer vision and artificial intelligence methods and algorithms" project (BB), funded under the tender relating to Projects of Relevant National Interest (PRIN) in 2017, which involves four multidisciplinary research groups belonging to four universities in Southern Italy (University of Bari Aldo Moro, University of Cagliari, University of Foggia, University of Naples Federico II). This interdisciplinary project combines artificial intelligence, technology, law, and psychology expertise to develop a comprehensive framework for detecting and addressing bullying and cyberbullying. The BB approach integrates cutting-edge AI techniques with psychological models to create tools that can effectively assess the risk of perpetuating, assisting, or suffering from bullying and violence in both physical and digital environments. The BB research group has gained significant experience during the project, and their preliminary results have earned them recognition, including selection for inclusion in the "Maker Faire European Edition 10th Anniversary Book" and selection as a promising project by the "Research Centre on Artificial Intelligence under the auspices of UNESCO" (IRCAI) in the Global Top 100 list of AI projects addressing the 17 United Nations Strategic Development Goals[1].

2. The BullyBuster framework

In this section, we will go through the core modules of the BullyBuster architectural framework, as depicted in Figure 2. The BullyBuster project's (Figure 1) purpose is to integrate behavioural biometrics, content analysis and crowd analysis into computer vision systems in order to detect all of the behaviours and indications that might define a bullying occurrence. The system, in particular, may be separated into four modules: (1) a module for crowd analysis for identifying through video-surveillance

Ital-IA 2023: 3rd National Conference on Artificial Intelligence, organized by CINI, May 29–31, 2023, Pisa, Italy

*Corresponding author.

giulia.orru@unica.it (G. Orrù)

[1] https://ircai.org/top100/entry/bullybuster-a-framework-for-bullying-and-cyberbullying-action-detection-by-computer-vision-and-artificial-intelligence-methods-and-algorithms/

Figure 1: BullyBuster project logo.

camera potential bullying incidents; (2) a text analysis module, that analyze textual content for the detection of signs of bullying or cyberbullying, (3) a third module of keystroke dynamics analysis to assess a potential victim's emotional state; and (4) a module of deepfake detection, to prevent the spread of malicious multimedia content on social networks and chatlines. These modules have been designed based on the psychological modeling of the behavior of bullies, victims, and spectators. The proposed solutions have been studied from a legal and juridical point of view to resolve the crucial aspects of privacy and data protection.

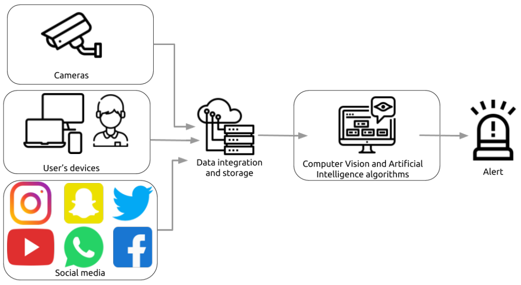

Cameras
User's devices
Social media
Data integration and storage
Computer Vision and Artificial Intelligence algorithms
Alert

Figure 2: Schematization of the BullyBuster system: the system analyses the video flow of surveillance cameras, a text-based analysis of the data present in the social accounts to detect harassment, oppression and stalking, keystroke dynamics, and deepfake detection. Collected data is analyzed to build effective computer vision and artificial intelligence-based algorithms to detect bullying and cyberbullying actions.

2.1. Behavioural modeling of the phenomenon

Bullying is a repeated, aggressive and anti-social behaviour of one or more individuals against a specific victim. It can have severe and long-term consequences for the victim's mental, emotional, and physical wellbeing [1]. Anxiety, sadness, low self-esteem, social isolation, and even suicidal thoughts or behaviours may be experienced by victims [2]. To build safe and inclusive environments for everyone, it is critical to identify and prevent bullying in all manifestations. Screening for (cyber)bullying behavioural indicators, both for perpetrators and victims, is critical for early detection and intervention. Schools, parents, and communities can take appropriate action to support victims and prevent bullying behaviours by recognizing these signals. In Italy, various strategies are being employed to combat these harmful behaviours, including education and awareness programs, online safety education and anti-bullying policies that discourage negative behaviour and spread a mental and cultural attitude emphasising diversity and tolerance [3].

2.2. Legal and juridical aspects

In Italy, prevention is carried on by promoting a cultural and social environment that discourages such behaviour and diffuses a mental and cultural attitude that emphasizes diversity and tolerance as a mean of mutual enrichment thanks to the efforts of the Ministry of Education, University and Research. The cyberbullying detection is in charge of Ministry of Justice as it was recently acknowledged as a crime by the Italian Law no. 71 from June, 18th, 2017[2]. This law officially described the criminal action of "cyberbullying" and the way (rules, investigative tools, burden of proof) to prevent and counter it. Moreover, it has been promoted the National Cyberbullying Observatory (http://www.cyberbullismo.com/), by which it is possible to report these criminal acts. In particular, the smartphone application named "Youpol" was implemented thanks to the efforts of Ministry of Internal Affairs and released for public use. This app allows reporting several criminal actions, including bullying and cyberbullying. The legal basis of the research is the art. 9, lett. j) GDPR:

> 1. [...] processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, shall be prohibited. 2. Paragraph 1 shall not apply if one of the following applies: [...]
>
> j. processing is necessary for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes in accordance with Article 89(1) based on Union or Member State law which shall be proportionate to the aim pursued, respect the essence of the right to data protection and provide for suitable and specific measures to safeguard the fundamental rights and the interests of the data subject.

[2] http://www.gazzettaufficiale.it/eli/id/2017/06/3/17G00085/sg

From the privacy point of view, the Garante per la protezione dei dati personali[3] , the italian authority responsible that people's privacy is not violated, recently allowed the installation of video surveillance cameras in the Istituto Galileo Ferraris in Verona. Furthermore, in Italy, with Law No. 71, May, 29th, 2017, the use of "special investigative tools", without identifying people, is admitted. It represents a considerable motivation to elaborate by technicians new forms of electronic evidence to meet as soon as possible the legislative needs.

2.3. Physical violence detection

Understanding human behaviours and actions through images and videos can be highly beneficial in the fight against bullying. Physical violence, isolation, or other physical patterns such as encirclement are among the behavioral "indicators" that can signify the presence of a problem, both as a victim and as a bully. Based on these psychological models, designing a computer vision system to identify and report anomalous occurrences attributed to bullying is possible. With this purpose, we designed a novel descriptor for crowd behavior analysis and anomaly detection [4]. It enables the observation of groups of persons that are not individually identifiable but can provide enough information to detect violence or panic. The created technique, inspired by the concept of the one-dimensional Local Binary Pattern algorithm (1D-LBP) [5], aims to evaluate the speed of creation and disintegration of crowd groups using appropriate patterns. The number of groups observed in a time interval can discriminate an ordinary scene from an abnormal one. We hypothesize that abrupt changes in the number of groups are caused by an anomalous occurrence, which may be recognized by translating these variations onto a temporal sequence of strings, which will be considerably different from those corresponding to a condition without anomalies (3). Through these algorithms, we have developed a detector that analyzes a video sequence and returns the probability of anomaly in the scene. The system alerts the human operator when this anomaly exceeds a certain threshold, as shown in Figure 4.

2.4. Verbal abuse detection

Cyberbullying is a widespread phenomenon that generally includes discrimination, aggression, and harassment. This phenomenon has become more widespread due to new communication channels allowing people to send and receive messages worldwide. The definition of cyberbullying has, among its consequences, cyber aggression. Cyber aggression is defined as aggressive behavior exercised on the Internet that uses digital media content (text, images, videos, etc.) to cause harm to another person, referred to as a victim. The following definition is very close to cyberbullying, which is *"an intentional aggressive act by an individual or group of individuals, using electronic forms of contact, repeated over time against a victim who cannot easily defend himself or herself"* [6]. Unlike standard forms of bullying, Cyberbullying is very dangerous because it can be perpetrated anywhere, anytime, and victims cannot often stop or reduce the spread of these activities [7].

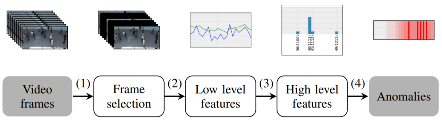


Figure 3: Anomaly detection pipeline. (1) A subset of the total frames is selected from the whole sequence of frames; (2) Low-level features are extracted to obtain the number of groups in each scene; (3) High-level features compute statistics of dynamics patterns; (4) Anomalies are obtained through thresholding a specific pattern.

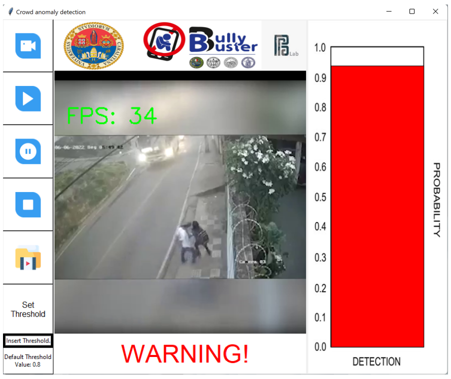


Figure 4: Demonstrator of the crowd analyzer built into the BullyBuster framework. The system detects suspicious behaviour and reports it to a human operator.

This project branch focuses on cyber aggression (textual or verbal), starting from textual comments of various post-Italians on Twitter through innovative steps of Feature Engineering, including slang and modern and close-to-aggressive speaking methods. According to the previously reported definition, Cyberbullying is a repeated cyber-attack by one or more users against one or more specific victims, repeated in multiple posts over time. An example of victims of these attacks are "celebrities" who are constantly attacked on their public profiles. In addition, actual conflicts arise from these aggressive comments between users with different thoughts. The goal was to search for recurring patterns in the wording of sentences or types of attacks. First, we focused on the type of language used by the attackers, noting that in

[3] http://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/1651744

almost all cases, it turns out to be approximate, full of expressions and words belonging to a vulgar jargon that leaves little room for misunderstanding; this led to thinking about the possibility of working with a list of these words, to recognize them in a comment. Next, it was noted how to measure a word's weight in a sentence, both negative and positive. Again, it was noticed that some negative comments were written in capitals as if to simulate a higher tone of voice. This led to thinking of a way to keep track of this peculiarity. Finally, another feature highlighted was the presence of the word "no/not" ("not" in English) in the aggressive comments; in many cases, it was noted that the aggression sessions began with this word to contradict the victim. All these observations were the basis for determining the right features to extract in the preprocessing phase. This study was conducted through Twitter comments in Italian, a language little used in the current state-of-the-art studies.

The Feature Engineering created is considered because they have been studied and analyzed. In contrast, others were considered because they have been exploited in some state-of-the-art studies, namely: *Number Negative words (BW), Number of "not/not" (NN), Uppercase (U), Positive/Negative weight of the comment (PW/NW), Use of the second person (SP), Presence of threats (TR), Presence of bullying terms (KW), Comment length (L).*

More specifically, based on the feature, it was deemed appropriate to create a dictionary as well:

- *Number of negative words (BW).* Through a "BadWord" vocabulary containing 540 extremely vulgar negative words used for aggressive purposes, insults, and humiliation [8].
- *Number of "not/not" (NN).* Using "no/not" within a sentence completely changes the sentence's meaning from positive to negative or vice versa.
- *Uppercase (U).* This is a Boolean value that indicates whether the comment is capitalized. It can be interpreted as an attack on someone [9].
- *Positive/negative comment weight (PW/NW).* This feature includes two values: the positive and negative weight of the comment within the range [0,1]. For this purpose, each word's synset and relative weight were extracted using WordNet and SentiWordNet [10].
- *Use of the second person (SP).* This Boolean value indicates the presence or absence of a second singular or plural form in the comment. This feature was extracted through a specially created dictionary containing 24 words [11].
- *Presence of threats (TR).* incitement to violence, or suicide. These expressions were identified using a specifically dedicated vocabulary containing 314 violent or inciting words [12].
- *Presence of bullying terms (KW).* A Boolean value indicates cyberbullying keywords (e.g., idiot, stupid, jerk, clown, whale, trash, ...). In this case, a vocabulary containing 359 terms identified as insults and possible insults was created.
- *Comment length (L).* This feature represents the length of the comment in terms of words. It has been noted that most negative comments consist of only a few words, usually no more than three.

The innovation related to this branch is Feature Engineering, dictionary creation, and a Dataset of aggressive and non-aggressive comments. The models used and tested are basic Shallow learning approaches to testing features[13].

2.5. Stress detection

In the context of (cyber)bullying detection and prevention, there is a growing need to identify and prevent negative emotional states of users while they engage in online communication. Affective computing, the ability to recognize users' emotional states, has been an ambition in this field for some time. However, current technologies are either expensive or obtrusive, making them impractical for widespread use. An alternative solution is Keystroke Dynamics (KD), which uses behavioural biometrics to identify or verify a person's identity by analyzing habitual rhythm patterns they use while typing on a keyboard.

On this line, we focused on the application of KD to continuously anticipate users' emotional states during message-composing sessions [14]. In particular, in the study we introduced a time-windowing approach that allows for the analysis of users' writing sessions in various batches, even when the writing window under consideration is quite brief. This approach is particularly relevant in the world of social media, where communications are frequently shared quickly and sparingly. The results of the study suggest that even extremely brief writing windows (on the order of 30 seconds) are sufficient to recognize the subject's emotional state with the same level of accuracy as systems based on the analysis of longer writing sessions. This finding is significant in the context of cyberbullying detection and prevention, as it enables real-time identification of negative emotional states, such as anger or sadness, during online communication. By using KD to detect and prevent cyberbullying, it may be possible to reduce the emotional harm caused by these negative interactions and promote healthier online communication.

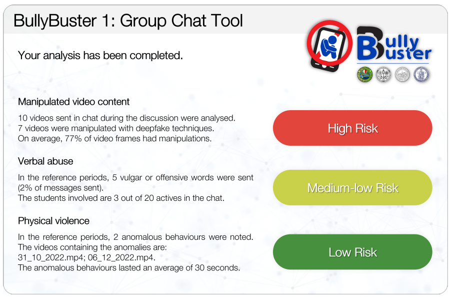
2.6. Manipulated video content detection

The issue of face-manipulated videos has gotten a lot of attention in the last two years, especially after the introduction of deep fake technology, which uses deep learning technologies to change the identity of people in images or videos. In the context of cyberbullying, the seriousness of the problem is indicated by the fact that even a single person's act can spread extensively and be repeated by others, resulting in repetition and an imbalance of power. The problem is worsened by the ability of current mobile devices to generate counterfeit content with a simple application. For these reasons, developing trustworthy algorithms for appropriately classifying videos as real or fake, i.e., deepfake detectors, is critical. Although many deepfake detectors have achieved excellent accuracy levels, generalization remains a significant challenge. Deepfake detectors, in particular, can identify just the manipulations on which they have been trained. For the BullyBuster deepfake detector, we exploited the complementarity of different individual classifiers with appropriate fusion rules to increase the generalization capacity of modern deepfake detection systems [15, 16]. Moreover, we designed a novel deepfake detection approach based on the Discrete Cosine Transform (DCT) representation of manipulated and original images at different scaling and compressing levels [17].

2.7. Prototype and use cases

The BullyBuster framework takes into account three different use cases:

- The BB questionnaire aims to collect data from specializing the BullyBuster automatic tools. The student watches animated videos illustrating bullying and cyberbullying before completing the questionnaire. We chose questions based on psychological analyses to assess how much a person's actions in real life and on the internet put them at risk of perpetuating, helping, or suffering bullying and violence. The development of the questionnaire and data collection was carried out in compliance with the aspects of privacy and ethics.
- The "Teacher Tool" enables the teacher to transfer data obtained from the class chat and videos retrieved from the video surveillance system to a desktop application and evaluate them using a deepfake detector, a text analyzer, and a crowd analyzer. The system generates a report that includes the risk percentages of bullying/cyberbullying behaviors for each module and the overall risk percentage for the class.
- The "Guided discussion tool" requires students to utilize the Bullybuster chat software on the desktop machines in the school's computer room to discuss an assigned topic (environment, politics, current events, etc.). The system analyzes chat data and determines the presence of deepfakes, violent comments, and stress levels. The outcome is a report that the instructor can consult; it provides the percentage risk of cyberbullying actions for the specific modules and the overall risk for the class.


BullyBuster 1: Group Chat Tool

Your analysis has been completed.

Manipulated video content

10 videos sent in chat during the discussion were analysed.
7 videos were manipulated with deepfake techniques.
On average, 77% of video frames had manipulations.

Verbal abuse

In the reference periods, 5 vulgar or offensive words were sent (2% of messages sent).
The students involved are 3 out of 20 actives in the chat.

Physical violence

In the reference periods, 2 anomalous behaviours were noted.
The videos containing the anomalies are:
31_10_2022.mp4; 06_12_2022.mp4.
The anomalous behaviours lasted an average of 30 seconds.

Figure 5: Final dashboard of the BullyBuster framework that allows to analyze the behaviour of a group of individuals to assess the risk of actions attributable to the problem of (cyber) bullying.

3. Conclusions

Bullying and cyberbullying are social issues that affect individuals and communities on various levels. They have broader implications for the overall well-being and safety of the social environment because they affect the individual's mental health and disrupt social relationships. The BullyBuster project presents a comprehensive, multidisciplinary approach to addressing the pervasive of these issues. The framework addresses many aspects of these phenomena in both physical and digital environments by combining cutting-edge artificial intelligence algorithms, computer vision, and psychological models. The partnership of four Southern Italian universities resulted in a comprehensive system that includes crowd analysis, text analysis, keyboard dynamics, and deepfake detection modules. The BullyBuster framework has demonstrated effectiveness in recognizing and preventing bullying and cyberbullying events. Its modular design enables ongoing improvement and adaptation to new challenges and technical advances. Furthermore, the legal and juridical component of the project assures that the developed solutions comply with privacy and data protection rules, making the BullyBuster framework a feasible and ethical solution for educational institutions, workplaces, and online platforms.

4. Acknowledgment

This work is supported by the Italian Ministry of Education, University and Research (MIUR) within the PRIN2017 - BullyBuster - A framework for bullying and cyberbullying action detection by computer vision and artificial intelligence methods and algorithms (CUP: F74I19000370001). The project has been included in the Global Top 100 list of AI projects addressing the 17 UNSDGs (United Nations Strategic Development Goals) by the International Research Center for Artificial Intelligence under the auspices of UNESCO.

References

[1] E. Menesini, M. Modena, F. Tani, Bullying and victimization in adolescence: Concurrent and stable roles and psychological health symptoms, The Journal of genetic psychology 170 (2009) 115–33. doi:10.3200/GNTP.170.2.115-134.

[2] P. K. Smith, J. Mahdavi, M. Carvalho, S. Fisher, S. Russell, N. Tippett, Cyberbullying: its nature and impact in secondary school pupils, Journal of Child Psychology and Psychiatry 49 (2008) 376–385. doi:https://doi.org/10.1111/j.1469-7610.2007.01846.x.

[3] E. Menesini, A. Nocentini, B. E. Palladino, Empowering students against bullying and cyberbullying: Evaluation of an italian peer-led model, International Journal of Conflict and Violence 6 (2012) 313–320.

[4] G. Orrù, D. Ghiani, M. Pintor, G. L. Marcialis, F. Roli, Detecting anomalies from video-sequences: a novel descriptor, in: 2020 25th International Conference on Pattern Recognition (ICPR), IEEE, 2021, pp. 4642–4649.

[5] N. Chatlani, J. J. Soraghan, Local binary patterns for 1-d signal processing, in: 2010 18th European Signal Processing Conference, IEEE, 2010, pp. 95–99.

[6] R. Dredge, J. Gleeson, X. de la Piedad Garcia, Presentation on facebook and risk of cyberbullying victimisation, Computers in Human Behavior 40 (2014) 16–22. URL: https://www.sciencedirect.com/science/article/pii/S0747563214004099. doi:https://doi.org/10.1016/j.chb.2014.07.035.

[7] R. Slonje, P. K. Smith, A. Frisén, The nature of cyberbullying, and strategies for prevention, Computers in Human Behavior 29 (2013) 26–32. URL: https://www.sciencedirect.com/science/article/pii/S0747563212002154. doi:https://doi.org/10.1016/j.chb.2012.05.024, including Special Section Youth, Internet, and Wellbeing.

[8] H. M. A. Ishara Amali, S. Jayalal, Classification of cyberbullying sinhala language comments on social media (2020) 266–271. doi:10.1109/MERCon50084.2020.9185209.

[9] D. Chatzakou, N. Kourtellis, J. Blackburn, E. D. Cristofaro, G. Stringhini, A. Vakali, Mean birds: Detecting aggression and bullying on twitter (2017). arXiv:1702.06877.

[10] M. Raghavan, M. K. Poongavanam, S. R. Ramachandran, R. Sridhar, Emotion and sarcasm identification of posts from facebook data using a hybrid approach, ICTACT Journal on Soft Computing 07 (2017) 1427–1435. doi:10.21917/ijsc.2017.0197.

[11] S. Shtovba, M. Petrychko, O. Shtovba, Detection of social network toxic comments with usage of syntactic dependencies in the sentences, Conference the Second International Workshop on Computer Modeling and Intelligent Systems, CEUR Workshop 2353 (2019).

[12] M. Raza, M. Memon, S. Bhatti, R. Bux, Detecting Cyberbullying in Social Commentary Using Supervised Machine Learning, 2020, pp. 621–630. doi:10.1007/978-3-030-39442-4_45.

[13] V. Gattulli, D. Impedovo, G. Pirlo, L. Sarcinella, Cyber aggression and cyberbullying identification on social networks, 2022, pp. 644–651. doi:10.5220/0010877600003122.

[14] S. Marrone, C. Sansone, Identifying users' emotional states through keystroke dynamics, in: Proceedings of the 3rd International Conference on Deep Learning Theory and Applications, SciTePress, Lisbon, Portugal, 2022, pp. 207–214. doi:10.5220/0011367300003277.

[15] S. Concas, S. M. La Cava, G. Orrù, C. Cuccu, J. Gao, X. Feng, G. L. Marcialis, F. Roli, Analysis of score-level fusion rules for deepfake detection, Applied Sciences 12 (2022). URL: https://www.mdpi.com/2076-3417/12/15/7365. doi:10.3390/app12157365.

[16] S. Concas, J. Gao, C. Cuccu, G. Orrù, X. Feng, G. L. Marcialis, G. Puglisi, F. Roli, Experimental results on multi-modal deepfake detection, in: S. Sclaroff, C. Distante, M. Leo, G. M. Farinella, F. Tombari (Eds.), Image Analysis and Processing – ICIAP 2022, Springer International Publishing, Cham, 2022, pp. 164–175.

[17] S. Concas, G. Perelli, G. L. Marcialis, G. Puglisi, Tensor-based deepfake detection in scaled and compressed images, in: 2022 IEEE International Conference on Image Processing (ICIP), IEEE, 2022, pp. 3121–3125.