BullyBuster

LEVERAGING ARTIFICIAL INTELLIGENCE TO FIGHT (CYBER)BULLYING FOR HUMAN WELL-BEING

THE «BULLYBUSTER» PROJECT

Ital-IA
ITALIA INTELLIGENZA ARTIFICIALE
cini National Lab AIIS

PARTNERSHIP

- Università degli Studi di Napoli «Federico II»
  - Carlo Sansone, Stefano Marrone, Michela Gravina, Antonio Galli
- Università degli Studi di Bari «Aldo Moro»
  - Donato Impedovo, Vincenzo Gattulli, Lucia Sarcinella
- Università degli Studi di Foggia
  - Donatella Curtotti, Angela Procaccino, Grazia Terrone, Wanda Nocerino
- Università degli Studi di Cagliari
  - Gian Luca Marcialis, Giulia Orrù, Giovanni Puglisi, Sara Concas, Marco Micheletto, Gianpaolo Perelli

MAIN ISSUE: THE «BULLY» PROFILE

- The behavior is carried out voluntarily: the bully acts with the precise aim of dominating the other and damaging him.

- The attacks are the result of **cognitive planning**

- **Intention to harm and lack of compassion:**

▪ the "persecutor" takes pleasure in insulting, beating or trying to dominate the "victim";

▪ she/he continues even when it is evident that the victim is very ill and distressed

ACTIONS

Flaming

Denigration

Harassment

Cyberbashing

Cyberstalking

Exclusion **Exposure**

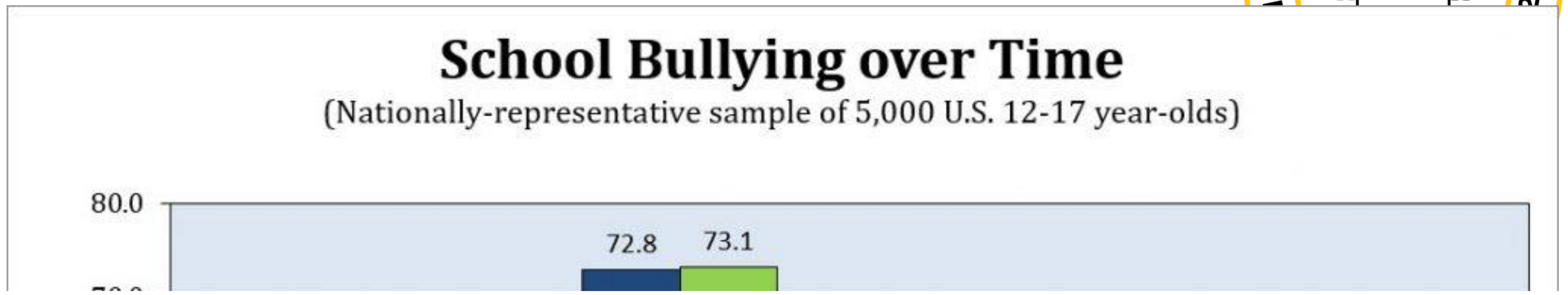
(CYBER)BULLYING AND WELL-BEING

- Change in sleep-wake rhythm
- Nightmares
- Changes in appetite
- Psychomotor agitation
- Tic
- Widespread fears
- Avoidance of group contexts

- Headache
- Gastrointestinal problems
- Abdominal pain
- Dermatitis

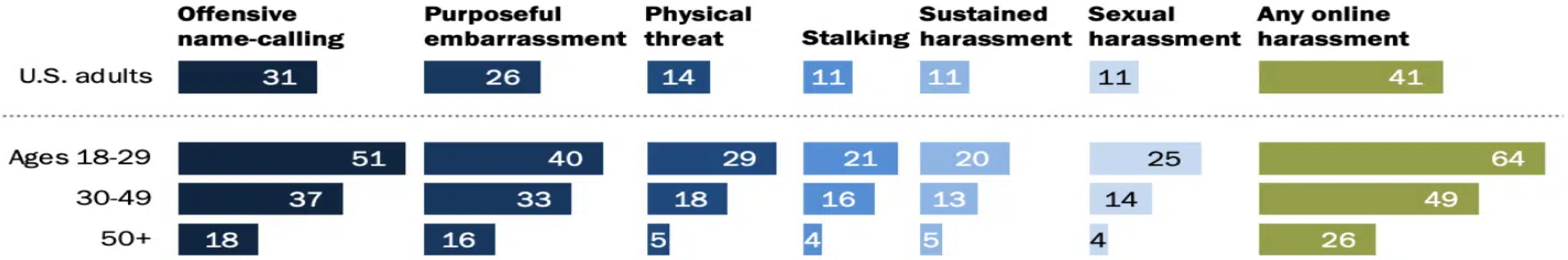
- Sadness
- Apathy and disinterest widespread
- Fatigue and asthenia
- Outbursts of unjustified anger
- Isolation

(CYBER)BULLYING TRENDS



Adults under 30 are more likely than any other age group to report experiencing any form of harassment online

% of U.S. adults who say they have personally experienced the following behaviors online

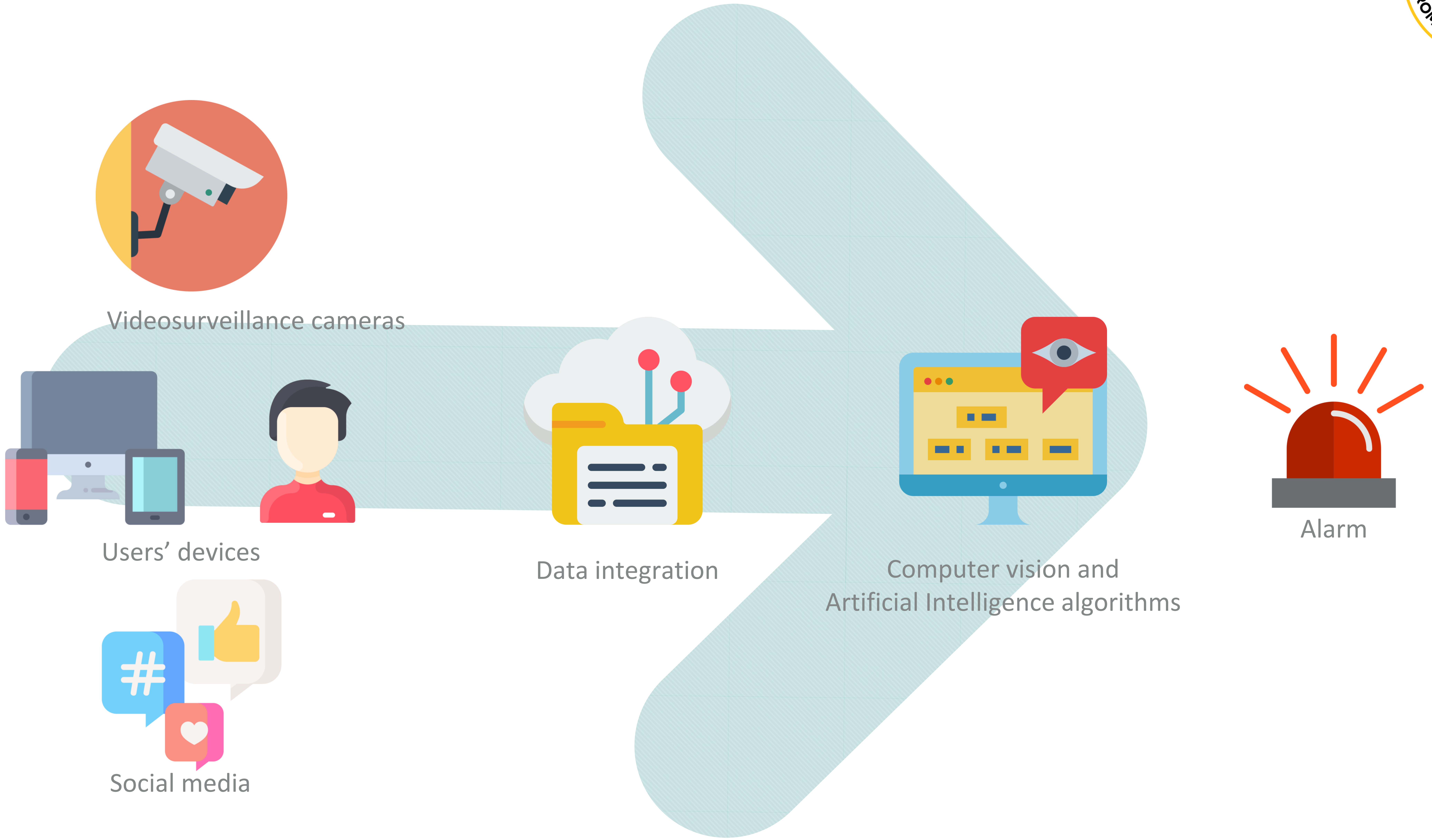


Note: Those who did not give an answer are not shown.
Source: Survey of U.S. adults conducted Sept. 8-13, 2020.
"The State of Online Harassment"

PEW RESEARCH CENTER

THE «BIG PICTURE»

IRCAI GLOBAL TOP 100 PROMISING PROJECT 2022



FIGHT AGAINST CYBERBULLYING IN ITALY: THE LEGAL PERSPECTIVE

Legge 18 giugno 2017, n. 71, recante «Disposizioni a tutela dei minori per la prevenzione ed il contrasto del cyberbullisimo»

Fatti penalmente rilevanti

- Molestie (art. 660 c.p.)
- Ingiuria (art. 594 c.p.) — **Depenalizzata, ex d.lgs. 7/2016**
- Diffamazione (art. 595 c.p.)
- Ricatto
  - Estorsione (art. 629 c.p.)
  - Minaccia di danno ingiusto (art. 612 c.p.)
- Furto d'identità
  - Sostituzione di persona (art. 494 c.p.)
  - Accesso abusivo ad un sistema informatico o telematico (art. 615 ter c.p.)
- Alterazione, acquisizione illecita e manipolazione di dati personali (art. 635 bis c.p.)
- Trattamento illecito di dati personali (art. 167, d.lgs. 196/2003)

Fatti non penalmente rilevanti

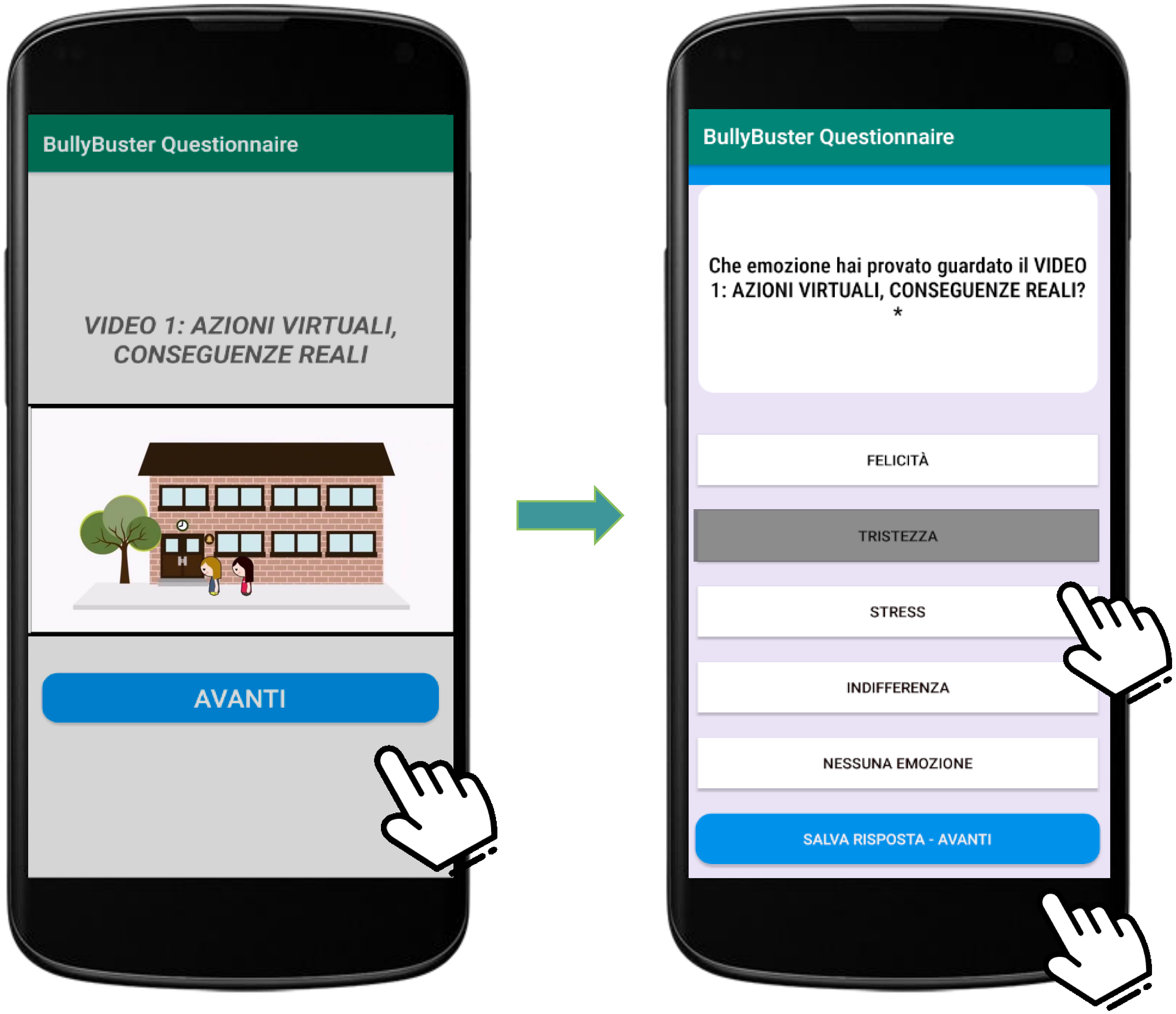
- Pressione
- Aggressione
- Denigrazione
- Diffusione di contenuti online

COLLECTING SIGNIFICANT DATA:THE BB-QUESTIONNAIRE

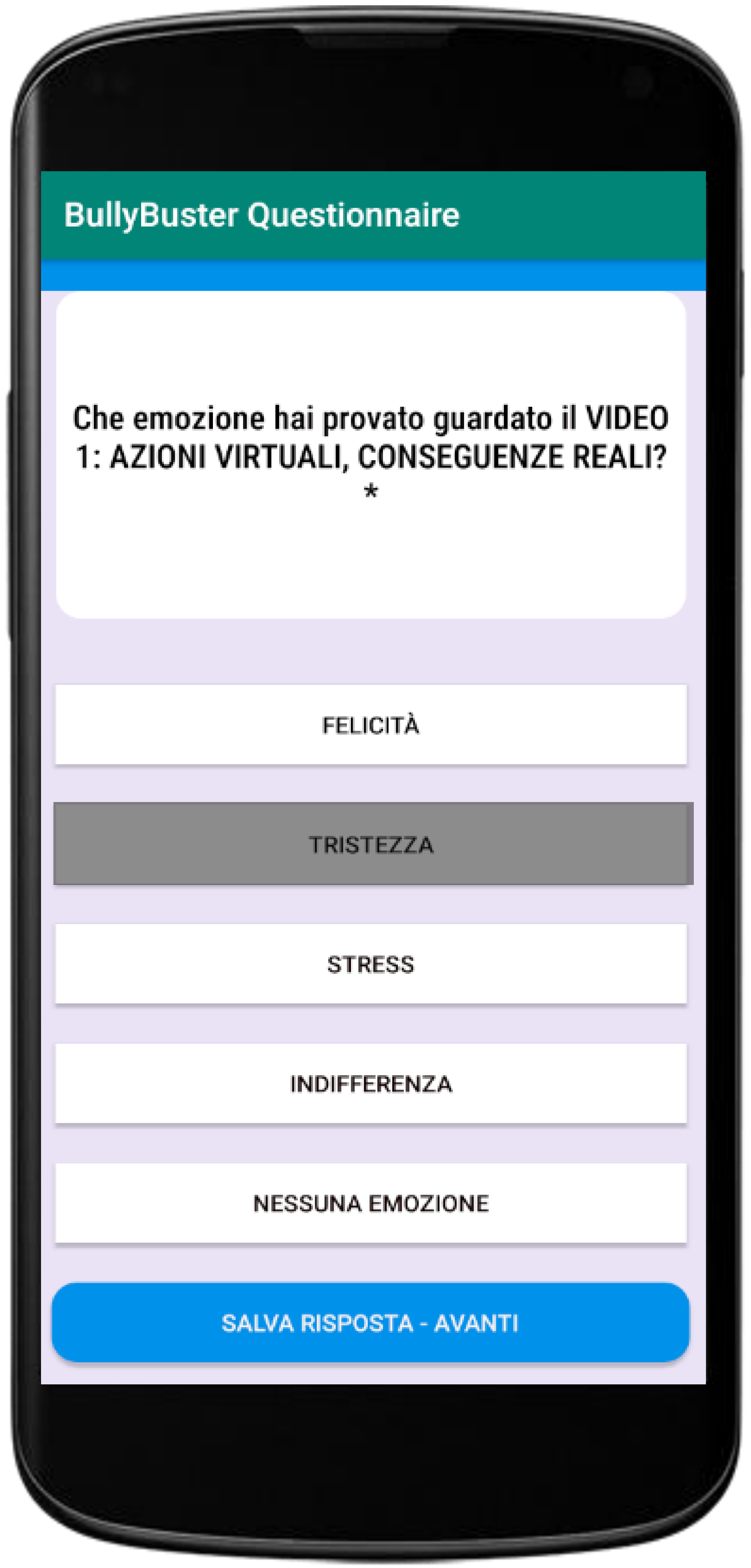
BullyBuster.pythonanywhere.com

android

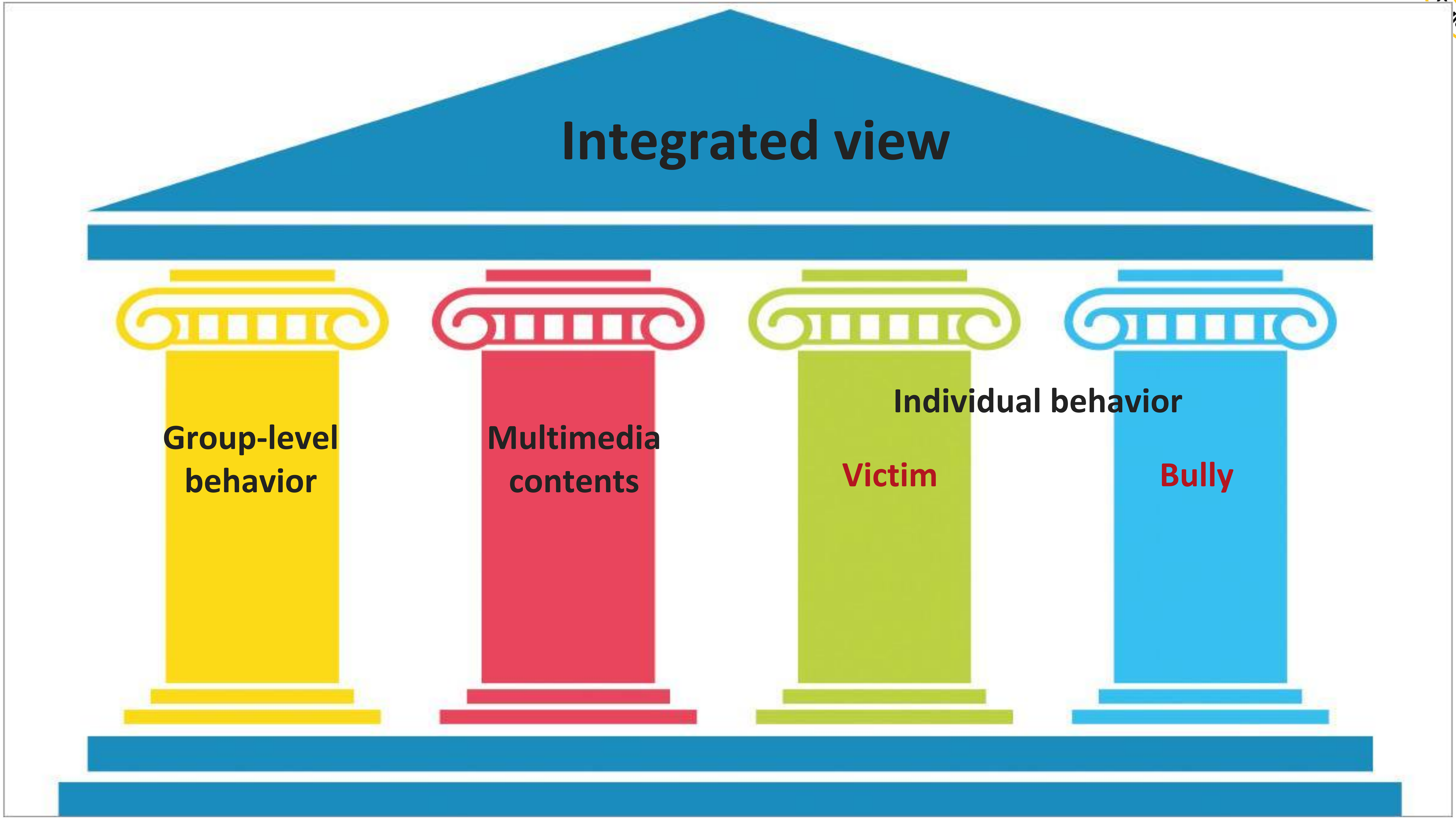
G. Terrone, A. Gori, E. Topino, A. Musetti, A. Scarinci, C. Guccione, V. Caretti, The Link between Attachment and Gambling/Internet Addiction in Adolescence: A Multiple Mediation Analysis with Developmental Perspective, Theory of Mind (Friend) and Adaptive Response, Journal Personalized Medicine, vol. 11, no. 3, 2021; https://doi.org/10.3390/jpm11030228.


BullyBuster Questionnaire
VIDEO 1: AZIONI VIRTUALI, CONSEGUENZE REALI
AVANTI
BullyBuster Questionnaire
Che emozione hai provato guardato il VIDEO 1: AZIONI VIRTUALI, CONSEGUENZE REALI? *
FELICITÀ
TRISTEZZA
STRESS
INDIFFERENZA
NESSUNA EMOZIONE
SALVA RISPOSTA - AVANTI


BullyBuster Questionnaire
VIDEO 1: AZIONI VIRTUALI, CONSEGUENZE REALI
AVANTI


BullyBuster Questionnaire
Che emozione hai provato guardato il VIDEO 1: AZIONI VIRTUALI, CONSEGUENZE REALI? *
FELICITÀ
TRISTEZZA
STRESS
INDIFFERENZA
NESSUNA EMOZIONE
SALVA RISPOSTA - AVANTI

THE BULLYBUSTER «PILLARS»



INDIVIDUAL BEHAVIOR

TEXT ANALYSIS: AGGRESSIVE CONTENTS

KEYSTROKE DYNAMICS: WELL-BEING EVALUATION

IDENTIFYING WELL-BEING STATES WITH KEYSTROKE DYNAMICS

- A person using the keyboard is unaware that their actions are being monitored resulting in an unbiased typing rhythm

- We introduced a time-windowing approach that allows analysing users' writing sessions in different batches, even when the considered writing window is relatively small

- This is very relevant in the field of social media, where the exchanged messages are usually very small and the typing rhythm is very fast

Marrone S. and Sansone C. (2022). Identifying Users' Emotional States through Keystroke Dynamics. In Proceedings of the 3rd International Conference on Deep Learning Theory and Applications - Volume 1: DeLTA, ISBN 978-989-758-584-5, pages 207-214. DOI: 10.5220/0011367300003277

TEXT ANALYSIS: FEATURE EXTRACTION AND RESULTS

- We leverage 20 high-level features based on the dwell time (i.e., the time elapsed between a key press and the same key release), on the flight time (i.e., the time elapsed between a key release and the next key press) and on the D2D-time (down to down, i.e., the time elapsed between a key press and the next key press)

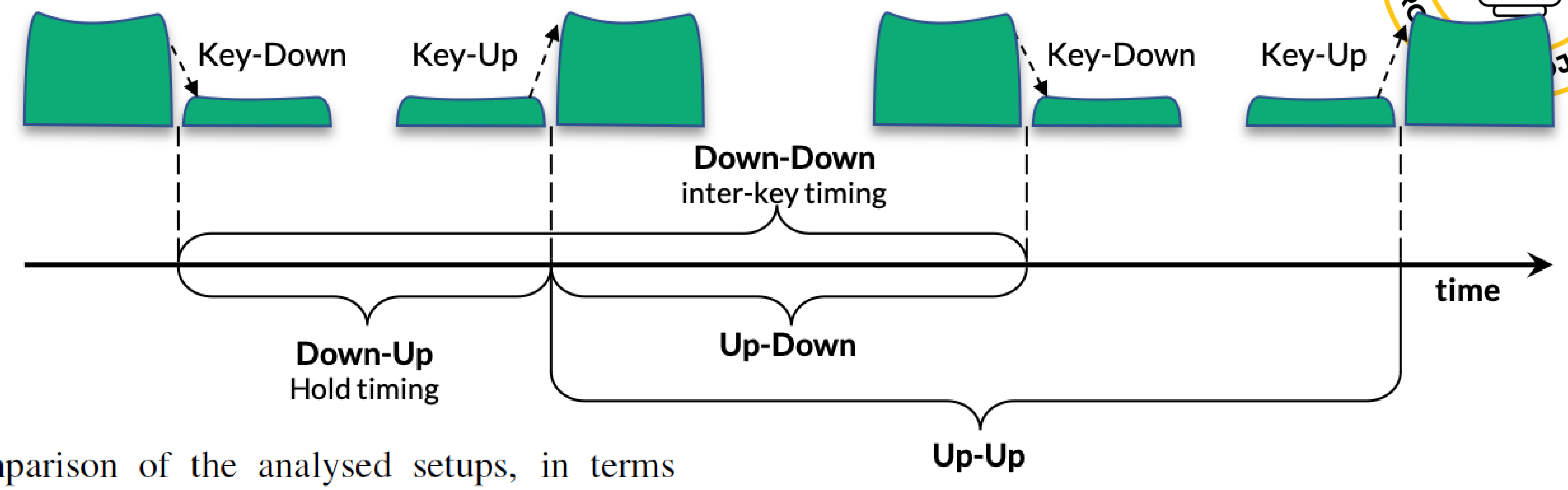
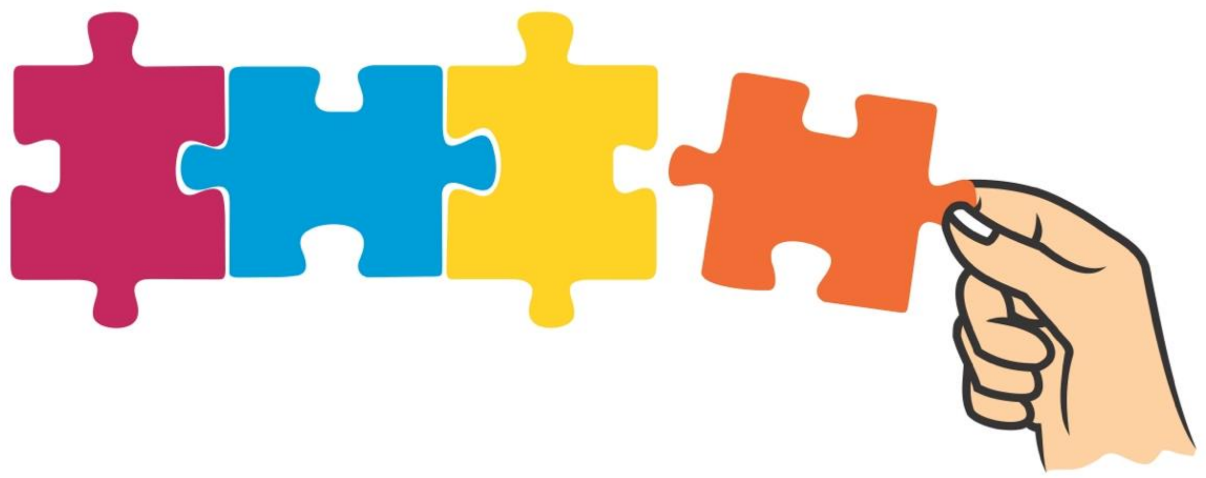


Table 4: Comparison of the analysed setups, in terms of classification accuracy (Acc), precision (Pre), recall (Rec) and F1-score (F1), varying the bag type (Fixed Bags - FB, Variable Bags - VB), the balancing technique (Class weights - CW, Undersampling - US, Oversampling - OS, Under-oversampling - UOS) and the voting approach (Highest probability voting - HPV, Most-frequent voting - MV). Best results are reported in bold.

Approach	Acc	Pre	Rec	F1
CNN CW-HPV	0.48	0.58	0.48	0.50
CNN CW-MV	0.44	0.56	0.43	0.43
CNN US-HPV	0.57	0.43	0.57	0.48
CNN US-MV	0.57	0.43	0.57	0.48
CNN OS-HPV	0.46	0.45	0.46	0.43
CNN OS-MV	0.41	0.43	0.41	0.40
CNN UOS-HPV	0.52	0.48	0.52	0.49
CNN UOS-MV	0.54	0.5	0.54	0.5
MIL-SVM VB	**0.76**	**0.80**	**0.69**	**0.74**
MIL-SVM FB-HPV	0.52	0.6	0.52	0.53
MIL-SVM FB-MV	0.48	0.52	0.48	0.47

VERBAL ABUSE DETECTION

- Design and implementation of a Machine Learning system that identifies cyberaggression in comments
- Creation of a vocabulary of Italian words considering four types of categories: *Bad Words, Second Person, Threats, and Bulling Terms*
- *Aggressive Italian Dataset*: Creating and labeling a balanced Italian (aggressive and non-aggressive comments)



FEATURE EXTRACTION

1. **Number of negative words** (Dictionary of 540 negative words)
2. **Number of "no/not“;**
3. **Uppercase**: Boolean value that indicates whether the comment is capitalized
4. **Positive/negative weight of the comment**: positive and negative weight of the comment within the range[0,1].
5. **Use of the second person** (24-word Dictionary);
6. **Presence of threats** (314-word dictionary);
7. **Presence of bullying terms** (359-word dictionary);
8. **Comment Length**.

RESULTS

	SVM			DT			RF			MLP		
Achille Lauro	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Not-aggressive	0.98	0.88	0.93	0.94	0.86	0.90	0.99	0.91	0.95	0.96	0.91	0.94
Aggressive	0.70	0.94	0.81	0.64	0.83	0.72	0.77	0.98	0.86	0.75	0.75	0.81
Accuracy	0.90			0.85			0.93			0.90		

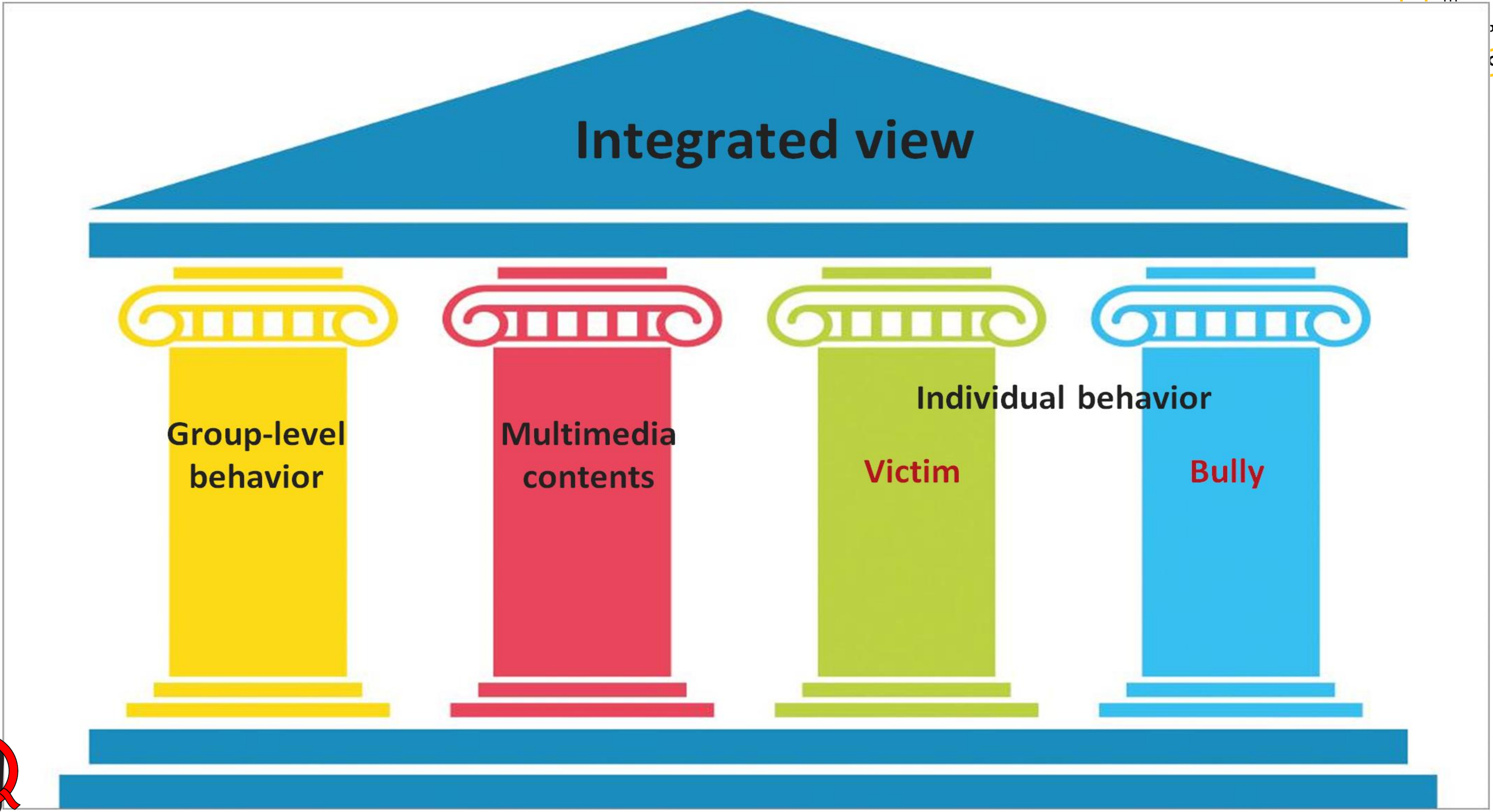
	SVM			DT			RF			MLP		
Fabio Rovazzi	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Not-aggressive	0.94	0.84	0.89	0.89	0.78	0.84	0.98	0.83	0.90	0.92	0.86	0.89
Aggressive	0.75	0.90	0.82	0.66	0.82	0.73	0.75	0.97	0.85	0.76	0.87	0.81
Accuracy	0.86			0.80			0.88			0.86		

	SVM			DT			RF			MLP		
Matteo Renzi	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Not-aggressive	0.98	0.95	0.97	0.98	0.95	0.96	0.99	0.98	0.98	0.98	0.96	0.97
Aggressive	0.74	0.89	0.81	0.71	0.84	0.77	0.85	0.95	0.90	0.76	0.88	0.82
Accuracy	0.94			0.94			0.97			0.95		

	SVM			DT			RF			MLP		
Giuseppe Conte	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Not-aggressive	0.93	0.84	0.89	0.89	0.77	0.83	0.96	0.85	0.90	0.90	0.85	0.88
Aggressive	0.73	0.87	0.79	0.64	0.81	0.71	0.75	0.92	0.82	0.82	0.84	0.78
Accuracy	0.85			0.80			0.87			0.84		

V. Gattulli, D. Impedovo, G. Pirlo, and L. Sarcinella, “Cyber aggressionand cyberbullying identification on social networks,” in ICPRAM.Scitepress, 2 2022, pp. 644–651.

THE BULLYBUSTER «PILLARS»



DEEPFAKES

«An image or recording that has been convincingly altered and manipulated to misrepresent someone as doing or saying something that was not actually done or said»

A deepfake is an image, or a video or audio recording, that has been edited using an algorithm to replace the person in the original with someone else (especially a public figure) in a way that makes it look authentic.

- The **fake** in deepfake is transparent: deepfakes are not real.
- The **deep** is less self-explanatory: this half of the term is specifically influenced by deep learning—that is, machine learning using artificial neural networks with multiple layers of algorithms.

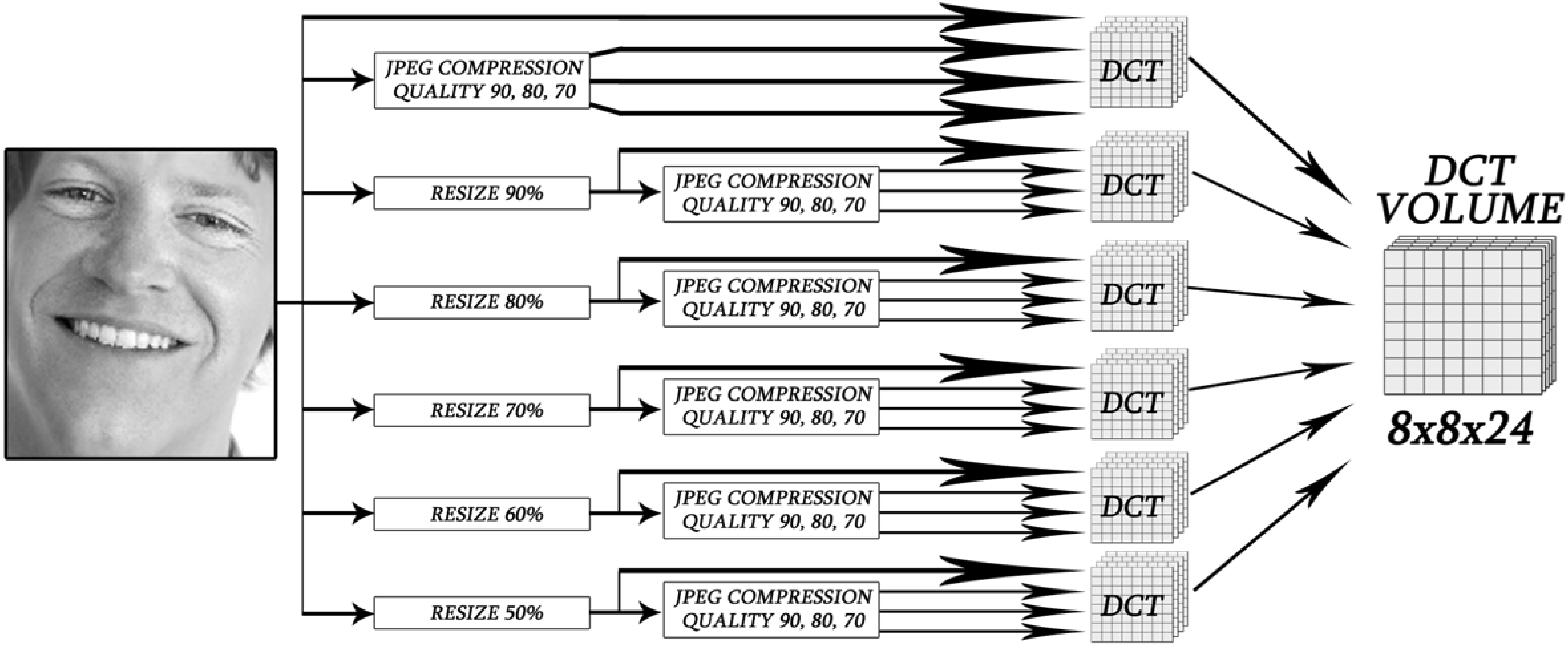
Merriam-Webster dictionary

DEEPFAKES AS A THREAT

https://www.bbc.com/news/technology-56404038

https://ars.electronica.art/center/en/obama-deep-fake/

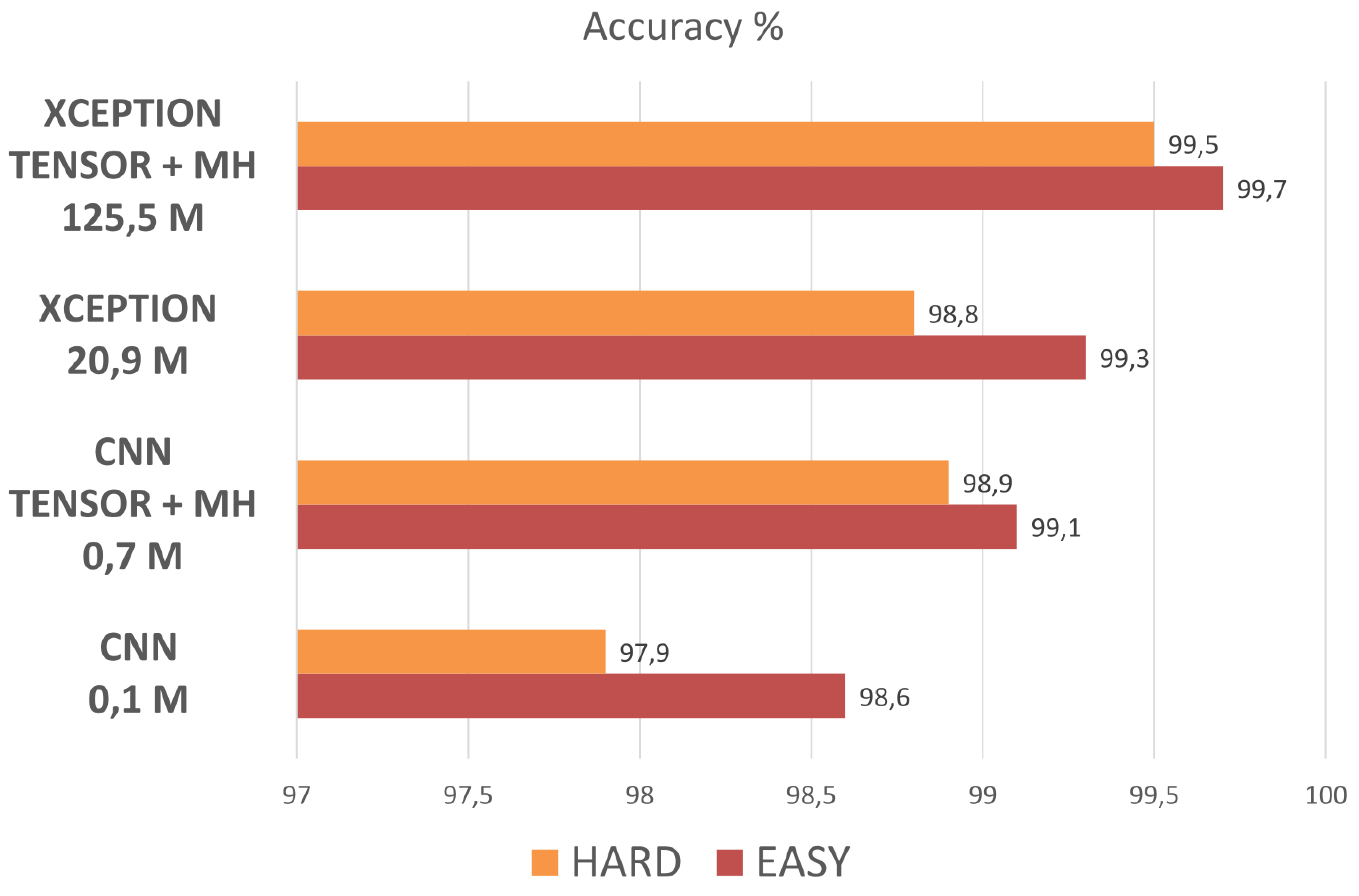
HANDLING SCALE AND COMPRESSION



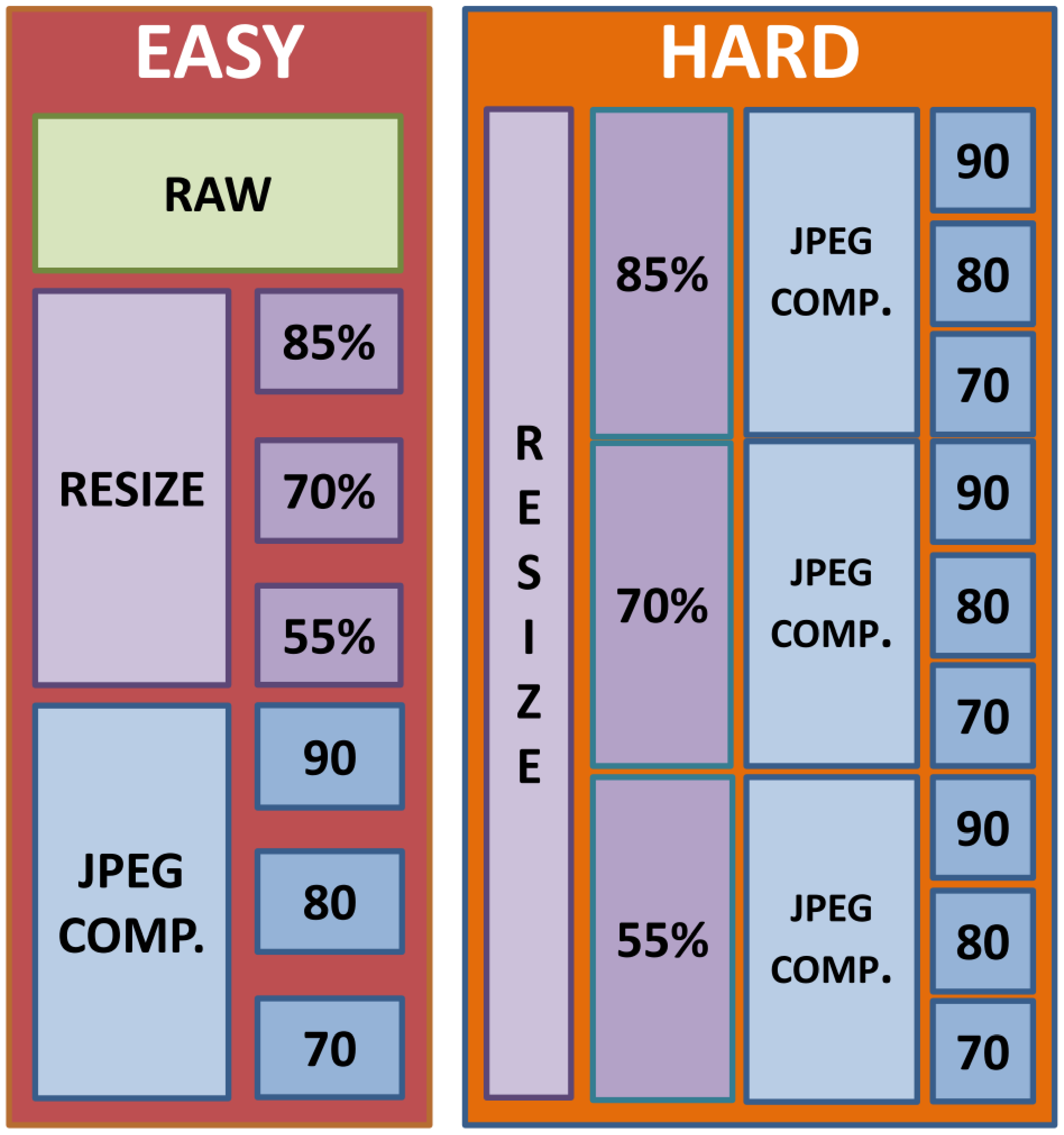
S. Concas, G. Perelli, G.L. Marcialis, G. Puglisi, Tensor-based deepfake detection in scaled and compressed images, 29th IEEE Int. Conf. on Image Processing (ICIP 2022), 16-19 October, 2022, Bordeaux (France), pp. 3121-3125, DOI: 10.1109/ICIP46576.2022.9897606

RESULTS

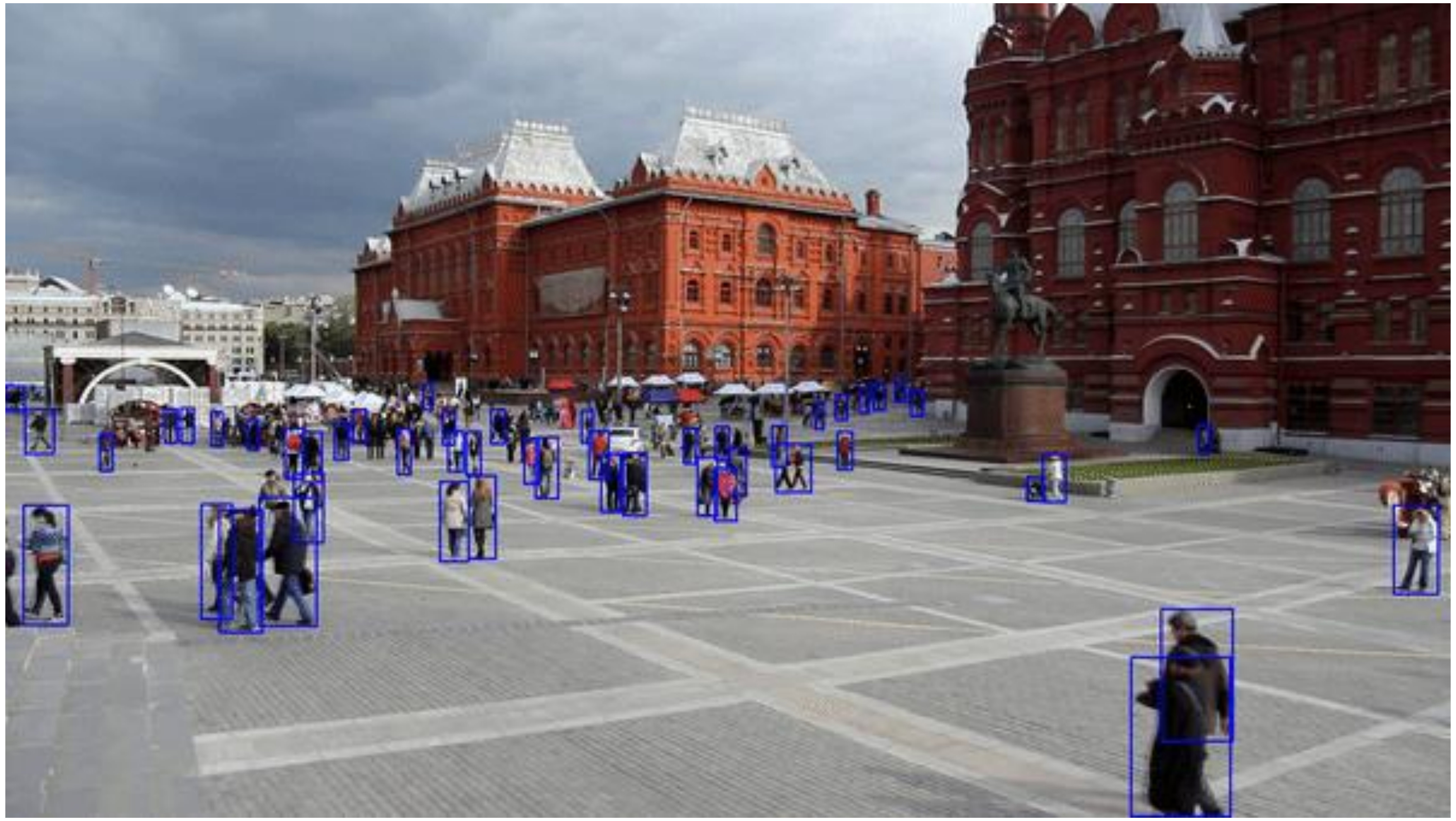
IRCAI GLOBAL TOP 100 PROMISING PROJECT 2022

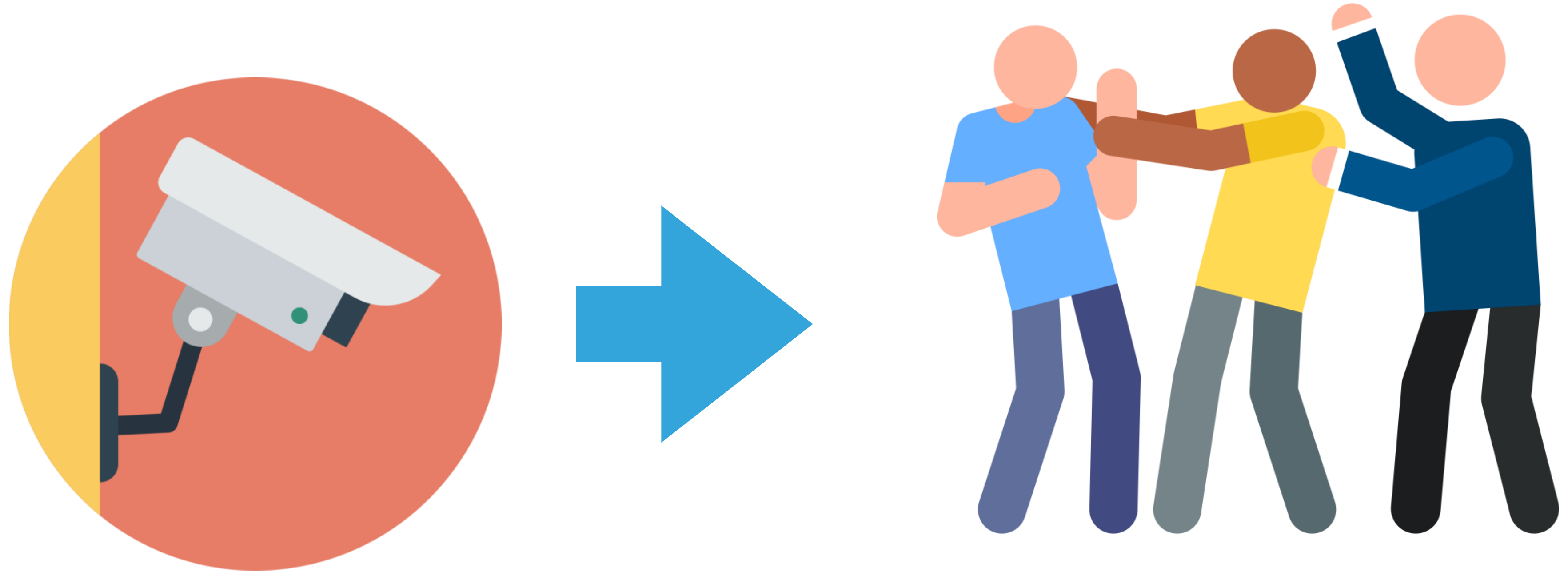


«Complexity» of the test set



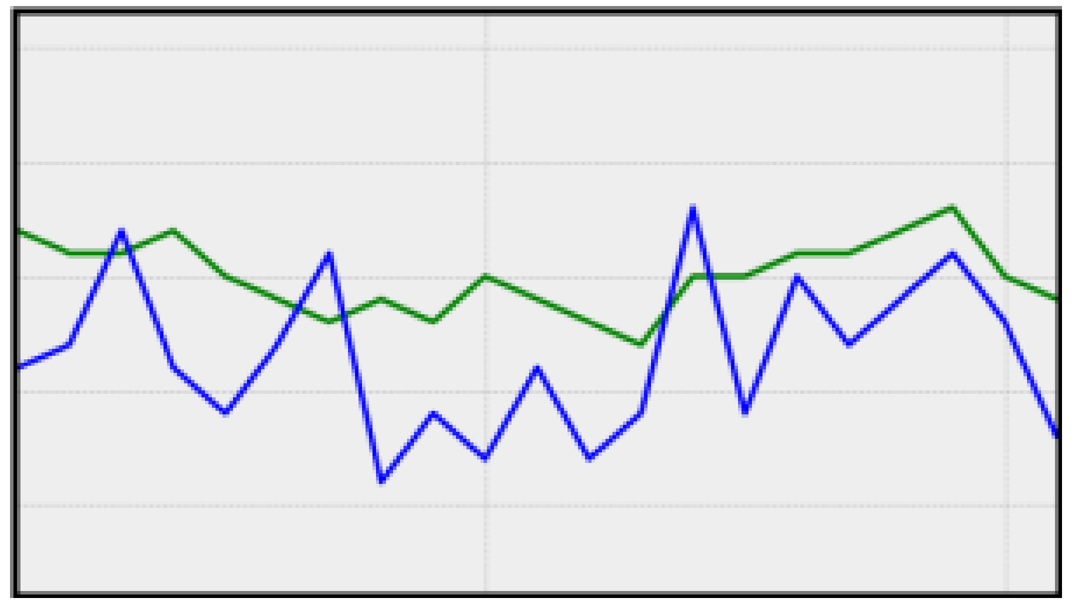
ANOMALOUS EVENTS DETECTION IN CROWDS

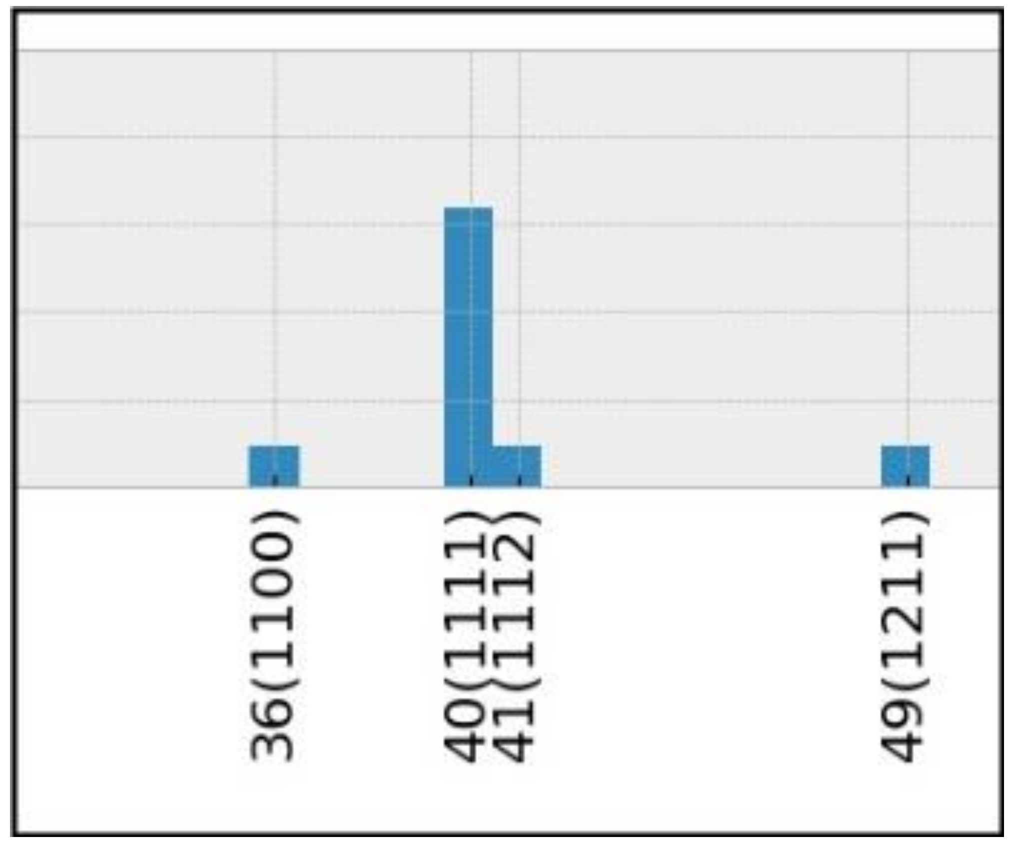


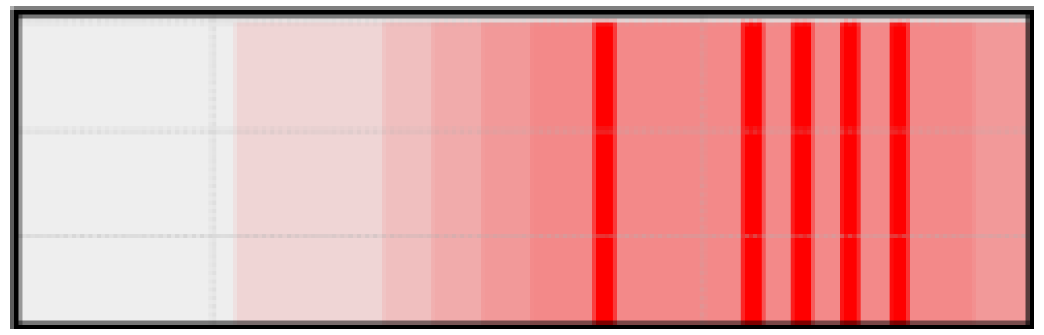


Violent behaviors

FEATURE EXTRACTION AND DESCRIPTION







FRAME SEGMENTATION – GROUP IDENTIFICATION → Group dynamics: aggregation and disgregation measurements → Descriptor of the group dynamics → Analysis of the system response: false alarms and true alarms

G. Orrù, D. Ghiani, M. Pintor, G.L. Marcialis, F. Roli, Detecting Anomalies from Video-Sequences: a Novel Descriptor, IEEE/IAPR 25th Int. Conf. on Pattern Recognition (ICPR 2021), Milano (Italy), 10-15th, Jan., 2021, https://arxiv.org/abs/2010.06407,
DOI: 10.1109/ICPR48806.2021.9412855

RESULTS

Motion-Emotion Data set

	Supervised			Leave-one-out		
All ME videos	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
MC	88.89%	94.12%	91.43%	79.31%	71.87%	75.41%
COF	71.11%	88.89%	79.01%	52.50%	60.00%	56.00%
CD	75.00%	91.67%	82.50%	73.17%	83.33%	77.92%
BD	70.45%	86.11%	77.50%	56.52%	74.29%	64.20%

MC – Manual counting

COF – Clustering of Optical Flow

CD – Cascade Detector

BD – Blob Detection

INTEGRATED VIEW

IRCAI GLOBAL TOP 100 PROMISING PROJECT 2022

BullyBuster : Tool chat di gruppo

RIschio contenuti multimediali manipolati

Rischio alto

Sono stati analizzati: 10 video inviati in chat nel periodo di riferimento
Di questi 7 video sono stati manipolati con tecniche deepfake
In media il 77% dei frame dei video presentava manipolazioni

Rischio violenza verbale

Medio-basso

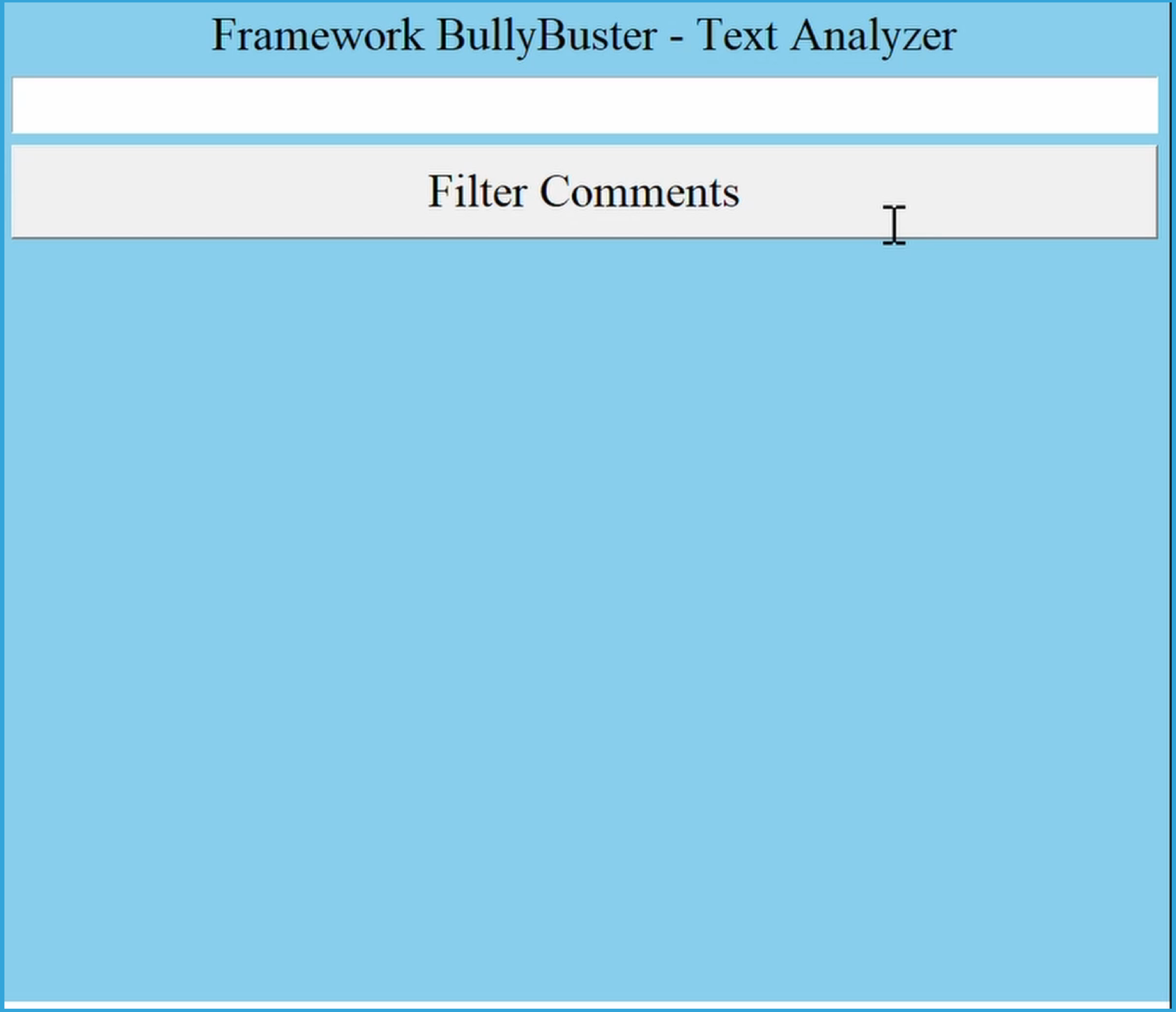
Nel periodi di riferimento sono state mandate 5 parole volgari o offensive (2% dei messaggi inviati)
Gli studenti coinvolti sono 3 su 20 attivi nella chat

Rischio di azioni di violenza fisica

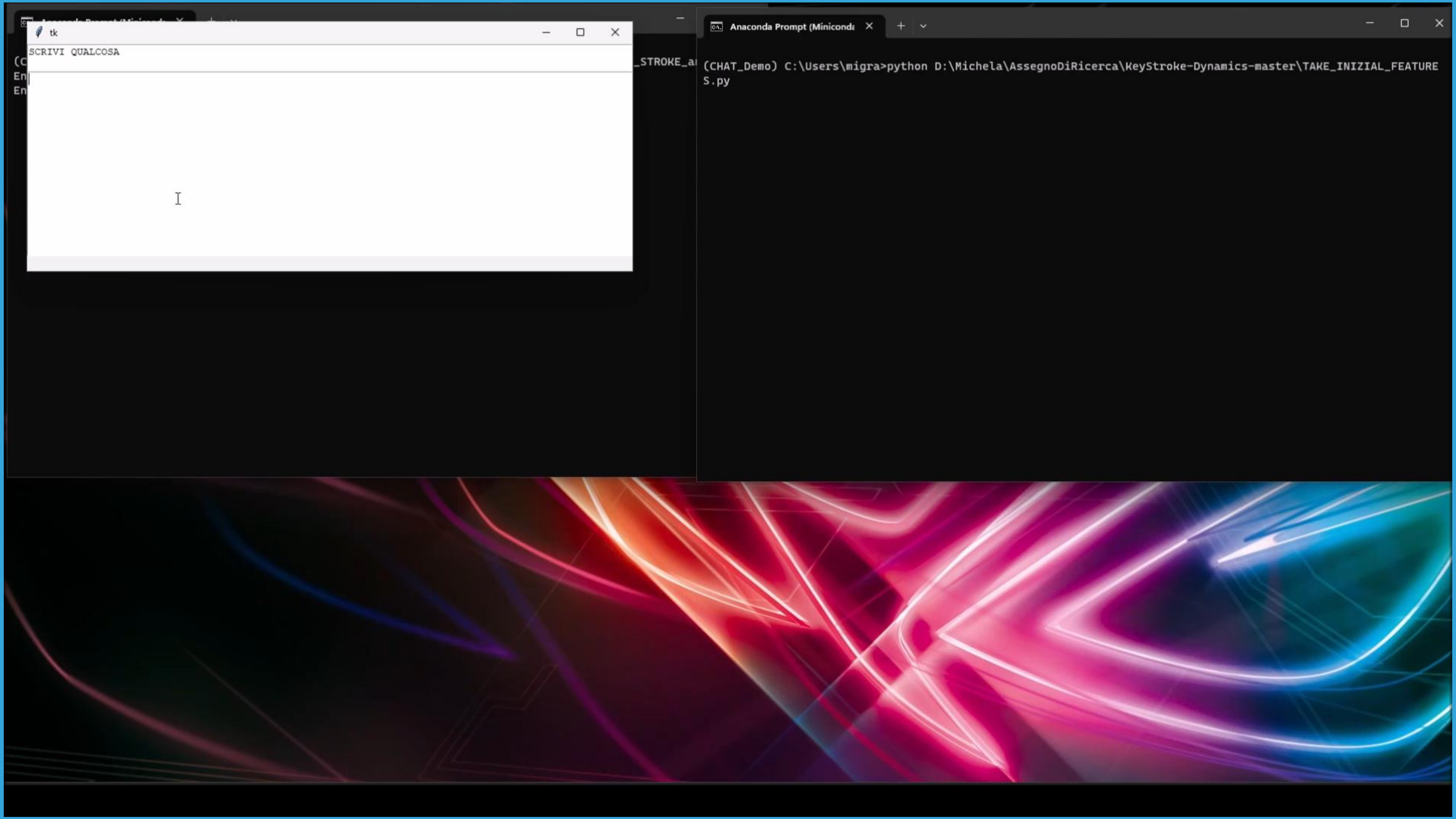
Basso

Nel periodi di riferimento sono stati rilevati 2 comportamenti anomali
I video contenenti le anomalie sono:
31_10_2022.mp4
6_12_2022.mp4
I comportamenti anomali sono durati in media 30 secondi

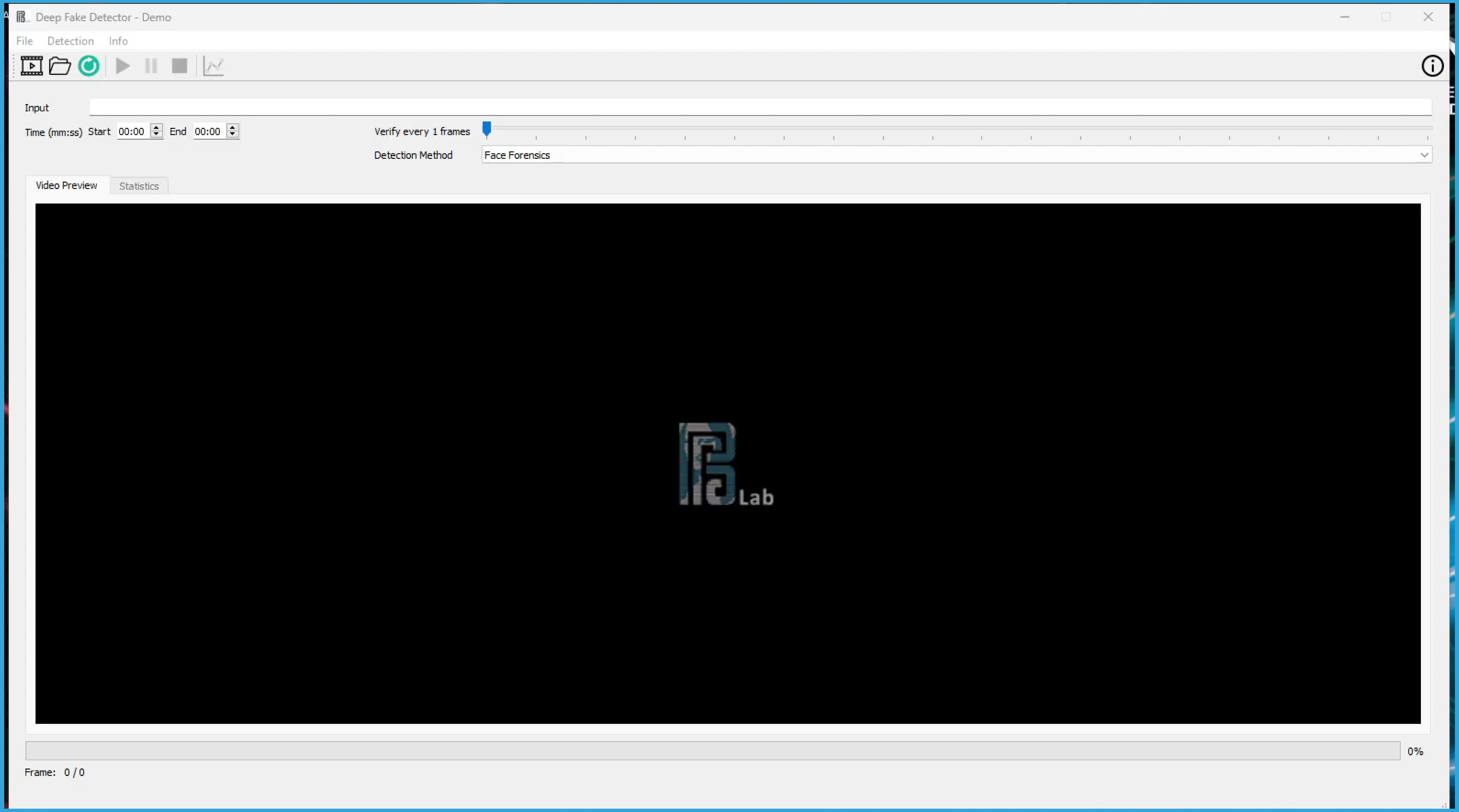
DEMO TIME



DEMO TIME

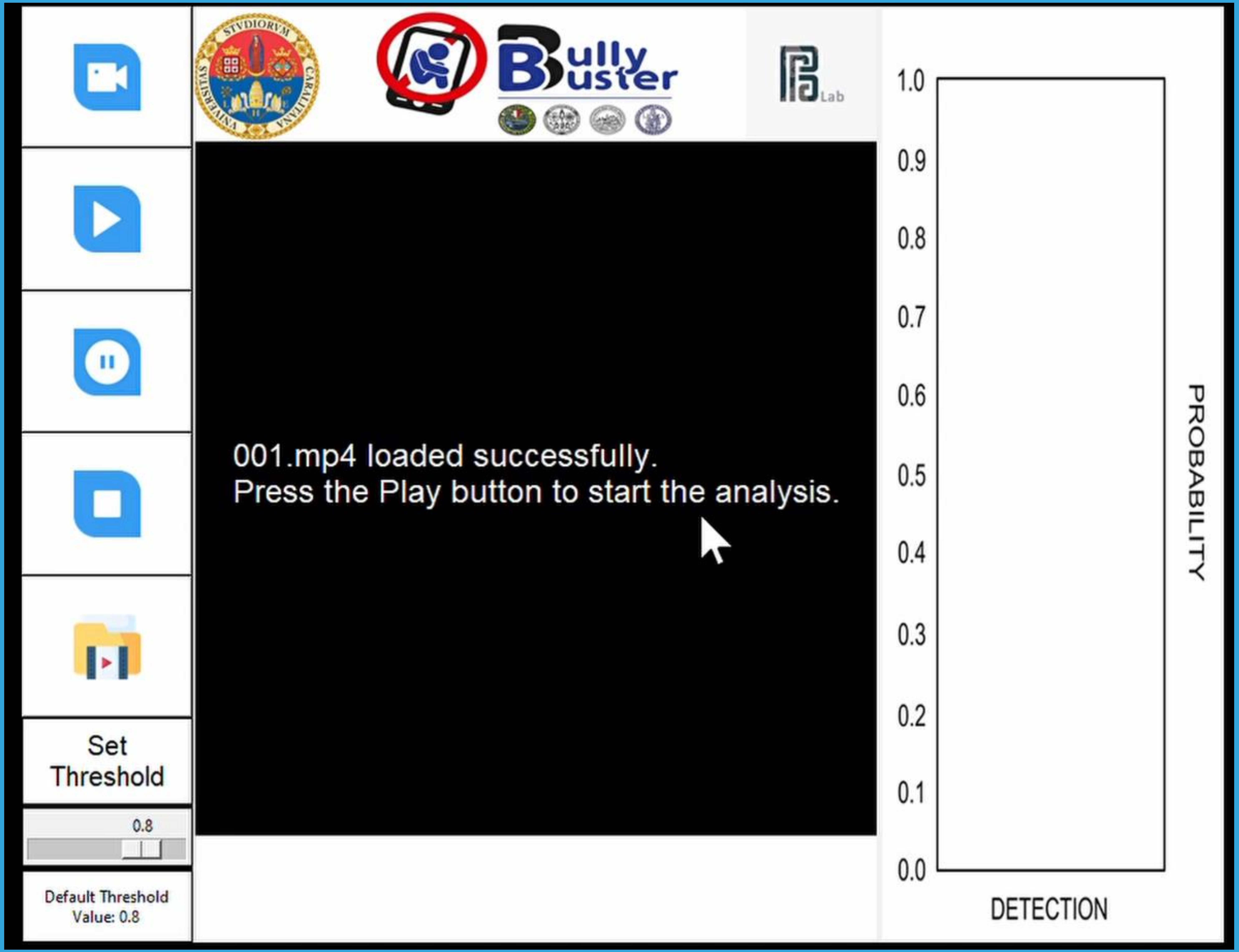

tk
SCRIVI QUALCOSA
_STROKE_a
Anaconda Prompt (Miniconda
(CHAT_Demo) C:\Users\migra>python D:\Michela\AssegnoDiRicerca\KeyStroke-Dynamics-master\TAKE_INIZIAL_FEATURES.py

DEMO TIME



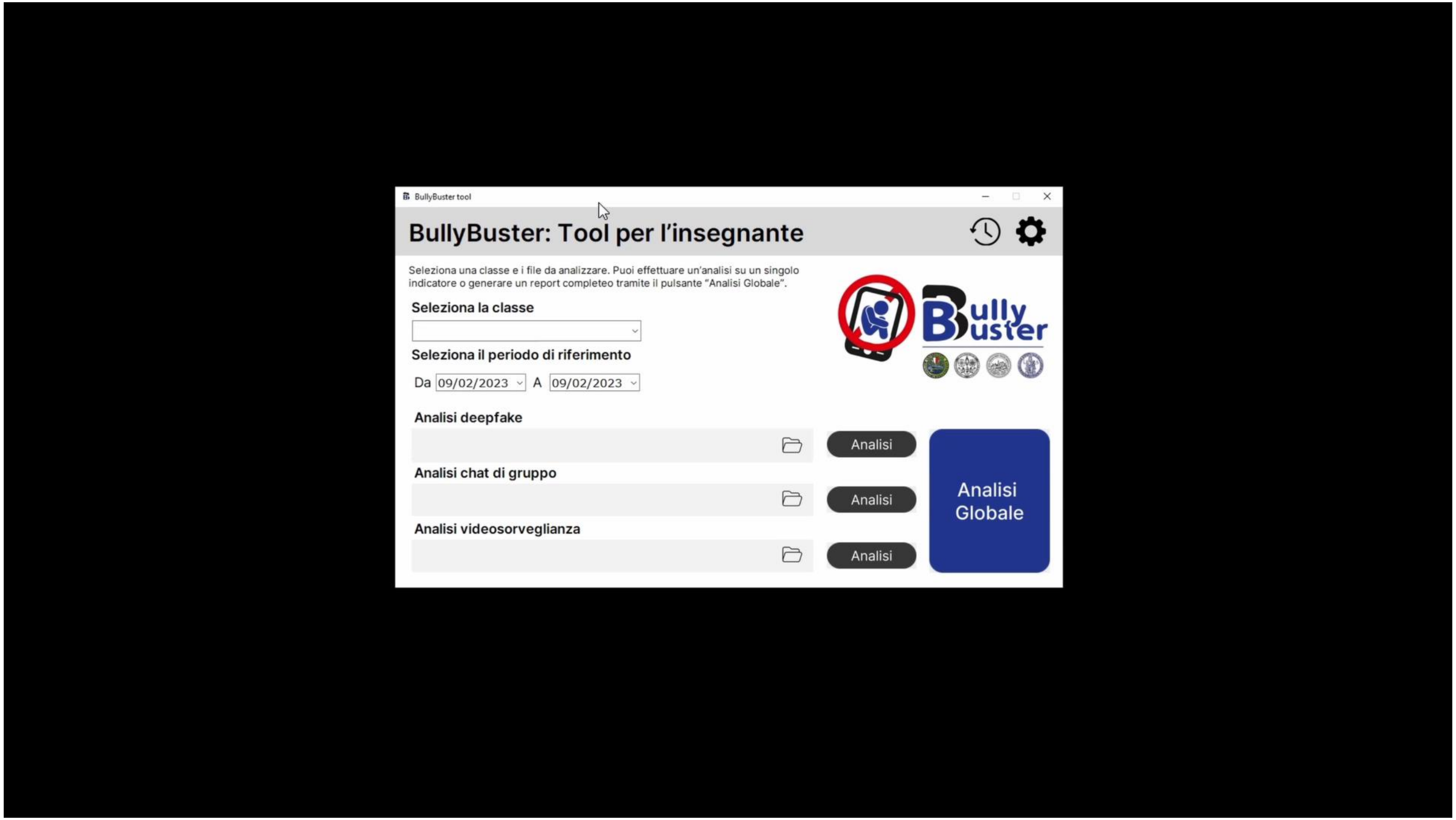
Deep Fake Detector - Demo
File
Detection
Info
Input
Time (mm:ss)
Start
00:00
End
00:00
Verify every 1 frames
Detection Method
Face Forensics
Video Preview
Statistics
Lab
0%
Frame: 0 / 0

DEMO TIME



Bully Buster
Lab
001.mp4 loaded successfully.
Press the Play button to start the analysis.
Set Threshold
0.8
Default Threshold Value: 0.8
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
PROBABILITY
DETECTION

THE BULLYBUSTER TOOL



THANK YOU

PI: DONATELLA CURTOTTI, DONATO IMPEDOVO, GIAN LUCA MARCIALIS, CARLO SANSONE

STAFF: SARA CONCAS, ANTONIO GALLI, VINCENZO GATTULLI, MICHELA GRAVINA, MARCO MICHELETTO, STEFANO MARRONE, GIULIA ORRÙ, WANDA NOCERINO, ANGELA PROCACCINO, GRAZIA TERRONE

WEB SITE: WWW.BULLYBUSTER.UNINA.IT