

## Article

# A Prototype Web Application to Support Human-Centered Audiovisual Content Authentication and Crowdsourcing

Nikolaos Vryzas <sup>\*</sup>, Anastasia Katsaounidou, Lazaros Vrysis , Rigas Kotsakis and Charalampos Dimoulas 

Multidisciplinary Media & Mediated Communication Research Group (M3C), Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece; akatsaounidou@gmail.com (A.K.); lvrysis@auth.gr (L.V.); rkotsakis@auth.gr (R.K.); babis@auth.gr (C.D.)

\* Correspondence: nvryzas@auth.gr

**Abstract:** Media authentication relies on the detection of inconsistencies that may indicate malicious editing in audio and video files. Traditionally, authentication processes are performed by forensics professionals using dedicated tools. There is rich research on the automation of this procedure, but the results do not yet guarantee the feasibility of providing automated tools. In the current approach, a computer-supported toolbox is presented, providing online functionality for assisting technically inexperienced users (journalists or the public) to investigate visually the consistency of audio streams. Several algorithms based on previous research have been incorporated on the backend of the proposed system, including a novel CNN model that performs a Signal-to-Reverberation-Ratio (SRR) estimation with a mean square error of 2.9%. The user can access the web application online through a web browser. After providing an audio/video file or a YouTube link, the application returns as output a set of interactive visualizations that can allow the user to investigate the authenticity of the file. The visualizations are generated based on the outcomes of Digital Signal Processing and Machine Learning models. The files are stored in a database, along with their analysis results and annotation. Following a crowdsourcing methodology, users are allowed to contribute by annotating files from the dataset concerning their authenticity. The evaluation version of the web application is publicly available online.

**Keywords:** tampering; authentication; misinformation; web application; news; machine learning; deep learning; crowdsourcing



**Citation:** Vryzas, N.; Katsaounidou, A.; Vrysis, L.; Kotsakis, R.; Dimoulas, C. A Prototype Web Application to Support Human-Centered Audiovisual Content Authentication and Crowdsourcing. *Future Internet* **2022**, *14*, 75. <https://doi.org/10.3390/fi14030075>

Academic Editor: Carlos Filipe Da Silva Portela

Received: 7 January 2022

Accepted: 24 February 2022

Published: 27 February 2022

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

News authentication is considered a vital task for reliable informational services. The COVID-19 pandemic situation that we currently experience showcased the importance of fact-checking in fighting disinformation to protect our societies and democracies. The role of audiovisual recording is considered crucial in documenting news articles, thus convincing audiences about the truth of the underlying events [1–3]. With the advancement of Information and Communication Technologies and the availability of easy-to-use editing and processing tools, one unwanted side-effect is the falsification of multimedia assets (i.e., images, audio, video) to alter the presented stories, making them more appealing (or intentionally doctored). In this context, unimodal solutions have been implemented to inspect each of the individual media entities, while multimodal forensic services are also deployed through online collaborative environments, plug-ins, serious games, and gamification components [1,2,4,5].

While the detection of manipulated photos/images and the evaluation of the associated forgery attacks remain critical [6], audio and video content have become even more popular nowadays. In this context, audio offers some unique features, such as less demanding processing needs and the inherent time continuity, making tampering inconsistencies easier to reveal [7,8]. Semantic processing and machine learning technologies empower

today's digital forensics tools. However, these new capabilities can also be exploited for counter-/anti-forensic means, requiring constant and continuous effort.

### 1.1. Related Work

Content verification has always been a very important part of journalistic workflows and a crucial factor of journalistic ethics and deontology. In the context of Journalism 3.0, where new business models of low or no pay journalism, combined with news aggregation and republishing and the reuse of amateur user-generated content (UGC) [9], disinformation has become a major problem for journalistic practice. As a result, several fact-checking organizations have appeared in the past decade, intending to find and debunk false claims that are spread throughout the Web and social media services [10]. Recent research of academics and organizations has been directed towards highlighting the best practices for content verification, through international cooperation networks [11].

In the modern media ecosystem, data variety is a very important parameter of big data volumes [12]. This means that fact-checkers need to manage content in many different modalities (e.g., text, audio, image, video). Different approaches and methodologies have to be defined for each case [13]. In disinformation, media assets may be used in a misleading context to support a false claim, or may be manipulated themselves. In the first case, an image/audio/video file is followed by an untrue description or conclusion, while in the latter, the media file has been maliciously edited. Such manipulations may include actions, such as copying and moving parts of the file to a different place and splicing in segments of a different file, aiming at affecting the semantic meaning of the file [14]. Common cases can be found in all file types, whether image, audio, or video. In the case of image tampering detection, spatial techniques can be used to locate suspicious regions and discontinuities within an image file. Media Verification Assistant is a project that allows users to upload images and applies several algorithms to provide forensics analysis [14,15]. In contrast to static images, audio and video files introduce the dimension of time. In video files, besides the spatial analysis of single image frames, the detection of temporal discontinuities can be crucial for the spatiotemporal location of malicious tampering [16]. Such techniques are expected to be computationally heavy. Audio is a very important modality present in the majority of video files. In this sense, audio can be used autonomously for the authentication of both audio and video assets. Audio information retrieval techniques are much less computationally complex. Audio forensics tools are not, however, as well-explored as those applied to visual information. Two important toolboxes on the market are the ARGO-FAAS [17] and the EdiTracker plugin [1]. They are, however, paid services, and not publicly available.

Audio forensics techniques address the processes of audio enhancement, restoration, and authentication of an audio asset so that it can be considered as evidence in court [18,19]. Authentication techniques aim at detecting artifacts within an audio file that can indicate malicious editing. Traceable edits can be found in the file container information or in the audio content [20,21]. Techniques that inspect file container inconsistencies investigate the metadata, descriptors, or the encoding structure. When the audio content is investigated, the aim is to use dedicated software to detect certain artifacts that may be inaudible by human subjects. Several different approaches can be found in the literature.

Electronic Network Frequency (ENF) techniques make use of the phenomenon of the unintentional recording of an ENF through interference. Electronic networks provide alternating current with a nominal frequency of 50 or 60 Hz, depending on the region. However, the real frequency of the current fluctuates around this value. The electronic equipment that is used for recordings captures this frequency fluctuation, which can act as a timestamp of the recording. It is possible to isolate and track the ENF in recordings to check whether there is phase inconsistency in the fluctuation, or even to find the exact time of the recording from the log files of the electronic networks [22–25].

Other approaches investigate the acoustic environment of the recording, such as the Signal-to-Reverberation ratio of a room [26,27]. The specifications of a recording device

have also been proven to be traceable in research, providing an indicator of whether parts of an audio file were recorded with a different device [20,28–30]. Dynamic Acoustic Environment Identification (AEI) may rely on statistical techniques that rely on reverberation and background noise variance in a recording [31]. Machine learning techniques are proven to be very useful for acoustic environment identification. Machine learning models do not rely on the definition of a set of rules for decision-making, but require a dataset of pre-annotated samples to train a classification model that can identify different classes, in this case, acoustic environments [31–33].

Another methodology for audio tampering detection investigates file encoding and compression characteristics. A huge number of highly configurable audio encodings are available that differ in terms of compression ratio, bitrate, use of low-pass filters, and more. A file that comes from audio splicing is very likely to contain segments encoded with different configurations, which can be traceable [34–36]. Most encoding schemes depend on psychoacoustic models that apply algorithms to discard redundant, inaudible frequencies. The Modified Discrete Cosine Transform (MDCT) coefficients can be investigated using statistical or machine learning methods to detect outliers in specific segments of the file [37]. Even when the file is reencoded in another format, there are often traces of the effect of previous compression algorithms [38–41].

Media authentication can be supported during content production using container and watermarking techniques, such as hash-code generation and encryption, and MAC timestamp embedding. Recovery of the inserted hash code that was generated by algorithms, such as SHA-512, enables the detection of tampered points within an audio stream [42]. Similarly, embedding timestamp information in files can allow the identification of an audio excerpt with a different MAC timestamp that has been maliciously inserted [20].

Whether the aim is training machine learning models or evaluating proposed analysis methods, one crucial part of every audio authentication project is the formation of a dataset. This is a very complex procedure due to the task's peculiarities, and it often acts as a bottleneck for the robustness of such techniques. Not many datasets are available for experimentation. In [43], a dataset was recorded featuring different speakers, acoustic rooms, and recording devices. In [44] a dataset with different encodings was created through an automated process. In [45], existing recordings were edited to create a dataset. In [7], an automated process was proposed for the creation of a multi-purpose dataset using an initial set of source files provided by the user.

### *1.2. Project Motivation and Research Objectives*

It has been made clear that machine learning solutions for audio tampering detection require a dataset for the training of models. Since datasets with real cases of tampered files are not available, most works require the formulation of artificial datasets for model evaluation. Such datasets are often difficult to handcraft, so they follow automated procedures for dataset creation, simulating real-world scenarios. As a result, the implemented models are case- and dataset-specific. There is no evidence for the generalization of the models in multiple scenarios and tampering techniques. For this reason, it is not yet feasible to integrate automated audio authentication into professional workflows without supervision, as they cannot be considered reliable for production and real-world applications. Furthermore, models that are pre-trained in known datasets and conditions may be more vulnerable to adversarial attacks [46].

On the other hand, traditional audio forensics techniques require expertise and fluency with audio analysis tools. In such an approach, human intelligence and experience play a crucial role in the process of authentication. While this is the most reliable solution and the preferable option in courtrooms, it cannot provide a viable alternative with massive appeal. There is an urgent need for tools that can help in the fight against disinformation. Such tools should be accessible to a broad audience of journalists, content creators, and simple users, to improve the overall quality of news reporting. Average users do not have the expertise to apply audio analysis techniques in the same way as professionals of audio forensics.

The motivation for the current research emerges from the hypothesis that it is feasible to strengthen a user's ability to recognize tampered multimedia content using a toolbox of supervisory tools provided online through an easy-to-use interface. State-of-the-art approaches for audio analysis and tampering detection were integrated into a web application. The application is available publicly through a web browser. The results of the algorithms do not provide an automated decision-making scheme, but rather a set of visualizations that can assist the user in a semi-automated approach. This means that the framework does not include a model that performs binary classification of files as tampered, or not-tampered, but the final decision is the responsibility of the user, taking advantage of their perception and experience as well as the context of the media asset. Through the use of the application, crowdsourcing is promoted for the creation of a dataset with real-world tampered files for future use.

The remaining of the paper is structured as follows. In Section 2, the proposed web framework is presented, in terms of the functionality, aims, and technical specifications. The integrated algorithms and their operating principles are listed without emphasizing technical details. In Section 3, the evaluation results from the reverberation estimation models and the initial implementation of the prototype web application are presented. In Section 4, the research results are summarized and discussed, and the future research goals of the project are defined. In Section 5, some of the limitations of the presented research are analyzed.

## 2. Materials and Methods

As stated in the problem definition section, the proposed approach consists of a framework for the assistance of professional journalists and the public in detecting tampered audiovisual content. The core of the framework is a web application with a graphic user interface provided to the public for the submission and analysis of content. The application incorporates an ensemble of algorithms that provide the user with supervisory tools for semi-automatic decision-making. The analysis strategy is audio-driven, as it makes use of the audio channel. The integrated algorithms do not classify files as tampered or not, but rather support the users in decision-making. The application offers the necessary crowdsourcing functionality for dataset creation and user cooperation. The framework was designed and implemented as a component of the Media Authentication Education (MAthE) project, which aims at providing educational and gamification tools to battle misinformation [4].

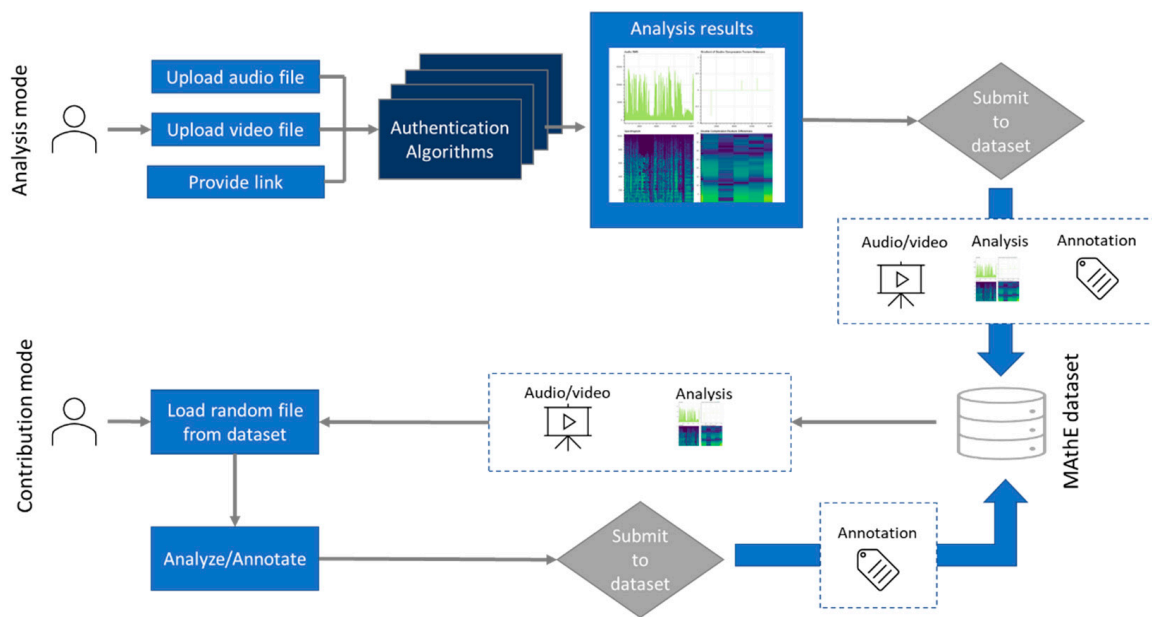
### 2.1. A Web Application for Audio Tampering Detection and Crowdsourcing

The main goal of the web application is to combine the effectiveness of state-of-the-art signal processing, machine learning advances and human perception for computer-assisted audio authentication. The application:

1. Implements state-of-the-art analysis options. An ensemble of algorithms is incorporated, addressing multiple audio tampering strategies. Such strategies may include encoding detection, recording conditions, background noise clustering, and others.
2. Follows a modular approach. The algorithms that are provided in the initial implementation are available as individual modules. This allows the existing algorithms to be upgraded in the future, as well as the extension of the initially provided toolbox.
3. Supports human-centered decision-making. As was explained, it is within the rationale of the MaThe solutions to promote computer-assisted decision making. The algorithmic implementations provide intuitive visualizations aiming at assisting the user in content authentication, taking also into consideration the user's personal experience and perception, as well as the context of the asset under investigation.
4. Is publicly available. As was explained, the web framework aims to address a wide public. An important prerequisite for this is that it is freely available for anyone to use and contribute.

5. Requires no audio or technical expertise. The design principles prioritize ease-of-use, following a typical workflow. A more experienced user with a technical and signal processing background, may get better insight and understanding of the produced visualizations. However, the detection of outliers or suspicious points in a file timeline is self-explanatory and does not require a deep understanding of the algorithms and mechanisms.
6. Promotes crowdsourcing. Users and teams can become involved and contribute to the project in several ways to further advance the field of audio tampering detection. They can submit files, annotated as tampered or not tampered, with a brief justification. Users can also randomly browse files from the dataset, analyze them, and mark them as tampered or not tampered. Finally, as this is an open-source project following a modular architecture, researchers and teams are encouraged to contribute with code and extensions.

The main functionality of the MAtHE AudioVisual Authentication framework is shown in Figure 1. Users can submit files for analysis and investigation, or contribute by annotating existing files concerning their authenticity. Once a file is submitted, the application returns analysis results, and the user can decide if they want to submit the file to the database along with an annotation (tampered or not tampered), submit the file to the database without annotation, or not submit anything to the database. Contributing users can access submitted files, annotated or not, examine the analysis results, and provide annotation (tampered or not tampered), following a crowdsourcing methodology.



**Figure 1.** The MAtHE AudioVisual Authentication framework and functionality.

### 2.2. The Computer-Supported Human-Centered Approach

The main concept of our approach depends on the idea that actors with no expertise in signal processing, machine learning, and computational methods can benefit from the visualization output of such techniques with little or no training. Computational methods in media authentication usually try to detect anomalies within the file under investigation. Such anomalies can be visually depicted (e.g., with a change in color). A non-expert user can perceive such depictions and interpret them accordingly, even without understanding or knowledge of the technical details that led to this visualization. After locating the suspicious points within the file, the user can base their reaction based on contextual information and their own critical thought. For example, an object within an image that looks like an anomaly in the visualizations and also dramatically alters the



image's semantic meaning probably indicates tampering. This is the main idea of the ReVeal project [14], which deals with tampered images and was a major inspiration of the present work. There is evidence from experiments that users with no technical knowledge were able to detect tampering of images with the support of such visualizations [6]. In this approach, a gamification approach was also tested that allowed users to ask for the help of such a visualization toolbox [5] in order to detect fake news and proceed in the game [4]. While such techniques are not robust in the automated detection of media content tampering, they can push the limits of human intellect and support users to make better decisions on fake content recognition.

In this direction, in the present work, several visualizations based on anomalies that are detected by audio processing are proposed. Since the audio channel is commonly part of video files, this toolbox aims at supporting users with no technical expertise to make decisions on the authenticity of audio and video files.

### *2.3. An Ensemble of Methods for Audio Tampering Detection*

In the related work section, several approaches for tampering detection are presented, which may fall into specific categories. Such categories include relevant audible or inaudible artifacts that are produced during the malicious editing of audiovisual files. As a result, depending on the type of forgery and the technical flaws of such an action, one technique may be more or less suitable. Hence, the motivation of the project derives from the hypothesis that it is irrelevant to try to evaluate different approaches to choose the most efficient, since this cannot be applied universally to every case [8].

The MATHe AudioVisual Authentication approach proposes a superposition of methods in a modular architecture that includes a dynamic group of algorithmic elements. Such techniques are either outcomes of previous research work within the project [7,8,47] or were found in the literature. The modular architecture allows for the modification of existing functions in the future, as well as its extension with new modules that come from new research, literature review, or contribution within the academic society.

Another hypothesis that has played an important role in the MATHe architecture design is that the lack of real-world datasets, as well as the diversity of the characteristics of tampered files, sets a bottleneck to the maturity of automated decision-making schemes. Most models are trained with artificially created datasets that address a specific type of tampering (recording device, room acoustics, encoding, etc.). Moreover, disinformation is only relevant at certain time points of a file, where the editing alters the semantic meaning of the recording. This is something that a human subject may easily understand. For this reason, the proposed design incorporates signal processing tools and machine learning models in a semi-automated approach [48]. It is not within the project's expectations to provide automated massive authentication of archive files, but rather to assist humans in analyzing and authenticating a specific file under investigation (FUI). The outcomes of the system require a human-in-the-loop [49] strategy. This is considered an effective combination of machine processing capabilities and human intelligence.

The initial toolbox of signal processing algorithms that was included in the prototype version of the MATHe AudioVisual Authentication application is presented below. It is noted that the technical presentation and validation of every approach is not within the scope of the current paper. Instead, a short description of the main functional principles of every category of techniques is given, along with references to publications with the technical details of different algorithms. The toolbox is dynamic, and it will be supported by incorporating state-of-the-art feature-based [50,51] and deep [52] (machine) learning approaches for audiovisual semantic analysis. It can also be deployed as a mobile application [53]. It is expected to further grow and evolve through the use of the application and the continuous dissemination of the MATHe project.

### 2.3.1. Common Audio Representations

This family of tools includes typical audio representations, such as waveform, energy values, and spectrograms. Such tools are available in most typical audio editing applications. Sound waveforms are the depiction of the amplitude of sound pressure of every audio sample, expressing the variation in the audio signal in time. For an audio signal with a common sampling frequency of 44,100 samples per second, waveforms may include a huge number of samples to be depicted, which can be computationally heavy to represent in an interactive graph running on a web browser. For this reason, in the proposed toolbox, time integration is performed, showing the root mean square (RMS) value for successive time windows as a bar diagram. This is used as a tradeoff to avoid the excess information redundancy of the waveform. Mel scale spectrograms provide a spatiotemporal representation of audio signals, depicting the evolution of the spectral characteristics through a time interval [54]. Spectral information is given for specific frequency bands that apply to the Mel scale, which is inspired by the psychoacoustic characteristics of human auditory perception. They are included in the toolbox because they can be useful, and they enhance the MAtHE framework's all-in-one solution so that users do not have to make use of more than one piece of software for analysis and decision-making.

### 2.3.2. Different Encoding Recognition

This family of techniques investigates the existence of small audio segments in the FUI that have different compression levels or encoding characteristics. This indicates that they may be segments of another file that were inserted in the original file. One common naïve approach that can be very effective in some cases is the calculation of the bandwidth, because most compression algorithms apply low-pass filtering to eliminate the higher frequencies that are of minor importance to the human auditory perception.

Feature vectors are descriptors of several attributes of a signal. Different encoding and compression levels, even if they are often proven to be inaudible in listening tests with human subjects, can affect the features that describe an audio signal. In the Double Compression technique for audio tampering detection that was proposed in [8], the FUI was heavily compressed. Features are extracted from the FUI and the compressed signal. For every time frame, the feature vector difference is calculated between the two signals. Parts of the FUI that have different encoding are expected to have different feature vector distances. Moreover, the gradient of differences is calculated. This measure is expected to reach peak values when there is an alteration in the compression levels, indicating suspicious points.

The double compression algorithm is summarized as follows [8]:

1. Heavy compression to the audio file under investigation (FUI), thus creating a double-compressed file (DCF).
2. A feature vector is extracted from the FUI and the DCF, creating the  $(T \times F)$  matrices  $F_i(t)$ , where  $i = 1, 2$ ,  $T$  is the number of time frames and  $F$  is the length of the feature vector.
3. For every time frame, the Euclidean distance  $D(t)$  of the two matrices is calculated
4.  $D'(t) = D(t) - D(t - 1)$  is calculated to show the differentiation between successive time frames.
5.  $D'(t)$  is expected to present local extrema in time frames that include a transition between audio segments of different compression, indicating possible tampering points.

For the feature selection, an audio feature vector was evaluated in [7]. Using a dedicated dataset creation script, a set of audio files were created, containing audio segments of different compression formats and bitrates. Specifically, segments of mp3-compressed audio in different bitrates were inserted randomly within an uncompressed file containing speech. Subjective evaluation experiments with three experts in the field of media production indicated that human listeners failed completely to detect the inserted segments for mp3 bitrates above 96 kbps, while they recognized approximately 10% of the inserted segments for mp3s of 64 kbps [7]. The dataset that was created in [7], along with the script for customized dataset generation, are documented and provided pub-

licly at <http://m3c.web.auth.gr/research/datasets/audio-tampering-dataset/> (accessed on 26 January 2022).

The selected feature set includes several frequency domain attributes, namely spectral brightness, with predefined threshold frequencies 500 Hz, 1000 Hz, 1500 Hz, 2000 Hz, 3000 Hz, 4000 Hz, and 8000 Hz, as well as rolloff frequencies, which are the upper boundary frequencies that contain energy ratios of 0.3, 0.5, 0.7, or 0.9 to the total signal energy, and spectral statistics (Spectral Centroid, Spread, Skewness, Kurtosis, Spectral Flatness, 13 Mel Frequency Cepstral Coefficients, Zero Crossing Rate, and RMS energy). The technical details of the aforementioned feature vectors are outside the scope of the current paper, but the methodology and feature evaluation process are presented thoroughly in [7].

### 2.3.3. Reverberation Level Estimation

Another indicator that several segments of a FUJ may have been inserted from a different file is the effect of acoustic conditions on the recording. Every space has different reverberation levels that affect the recording. Especially since most newsworthy events are not recorded in ideal conditions of professional recording studios, a regression model based on a Convolutional Neural Network architecture [47] was trained using a big dataset of simulated reverberation to provide a numerical estimations of the Signal-to-Reverberation ratio for every audio segment. Segments with outlier values are possibly related to malicious audio splicing.

Convolutional Neural Networks (CNNs) are a type of deep learning architecture that have gained popularity in audio recognition and event detection tasks [55,56]. One main reason for their recent widespread is that there is no need for a handcrafted feature vector. Instead, a visual representation of the audio information is fed to the networks as an image, and the input layers extract hierarchical features in an unsupervised manner during training. Different kinds of input have been evaluated for deep learning techniques, with spectrograms being the dominant approach [54].

Signal-to-Reverberation-Ratio (SRR) can be a useful attribute that can indicate audio slicing. SRR expresses the ratio of the energy of the direct acoustic field to the reverberation acoustic field. It is determined by the acoustic characteristics of the space of the recording, the positioning of the sound source and the recording device. The distance where the levels of the direct and the reverberation sound are equal ( $SRR = 1$ ) is called the critical distance. At distances closer than the critical distance, we can assume  $SRR > 1$ , and at distances that are farther than the critical distance,  $SRR < 1$ . The critical distance itself depends on the room acoustic attributes.

For recordings that take place under different conditions, the SRR is expected to differ. When segments from different recordings are pieced together, it is possible to detect the inconsistency in their SRR, even if it is not audible by human listeners. Calculating the SRR for different time windows can provide another criterion for audio tampering detection.

In the proposed approach, a deep learning regression model is used for a data-driven estimation of the SRR, based on simulation data. A 3600-second-long audio file containing pink noise was created, using the Adobe Audition generator. Using the same software, reverberation was added to the file with different SRRs. Ten different SRRs were chosen, resulting in 11 audio files (including the original), producing a 39600-second-long dataset. The same source audio file was used for all SRRs, so that the model is trained to recognize the reverberation and not information related to the content of different audio streams. The selected SRRs are shown in Table 1.

**Table 1.** The different Signal-to-Reverberation Ratios that were used for the model training.

Signal (%)	100	90	80	70	60	50	40	30	20	10	0
Reverberation (%)	0	10	20	30	40	50	60	70	80	90	100



The dataset was used for the training of a CNN regression model. The output of the model is a continuous value from 0 (no reverberation) to 1 (only reverberation). The model architecture is provided in Table 2.

**Table 2.** The architecture and hyper-parameters of the CNN model for reverberation level estimation.

Layer	Type	Configuration
1	Convolutional 2D Layer	16 filters Kernel size = (3,3) Strides = (1,1)
2	Max Pooling 2D Layer	Pool size = (2,2)
3	Dropout	Rate = 0.25
4	Convolutional 2D Layer	32 filters Kernel size = (3,3) Strides = (1,1)
5	Max Pooling 2D Layer	Pool size = (2,2)
6	Dropout	Rate = 0.25
7	Convolutional 2D Layer	64 filters Kernel size = (3,3) Strides = (1,1)
8	Dropout	Rate = 0.25
9	Convolutional 2D Layer	128 filters Kernel size = (3,3) Strides = (1,1)
10	Convolutional 2D Layer	256 filters Kernel size = (3,3) Strides = (1,1)
11	Flatten Layer	
12	Dense Neural Network	Output weights = 64 Activation = ReLU L2 regularizer
13	Dense Neural Network	Output weights = 64 Activation = ReLU
14	Dropout	Rate = 0.25
15	Dense Neural Network	Output weights = 24 Activation = Linear

#### 2.3.4. Silent Period Clustering

Besides room acoustics, the background noise also characterizes a recording. The environmental noise that is recorded in speech recordings is often inaudible, since it is mixed with a speech signal of a much higher level. However, in the small periods of silence that occur between words and syllables, the background noise signal is dominant. In the case of combining two or more recordings to create a tampered audio file, different background noise patterns may be distinguishable. Initial investigation has shown that by exporting a feature vector from small segments of silence (~25 ms) and providing them to a clustering algorithm, it is feasible to separate the different environmental audio classes that are present in an unsupervised way [8].

#### 2.4. Crowdsourcing for Dataset Creation, Validation, and User Cooperation

Crowdsourcing is a methodology for distributed problem-solving that happens online by the collective intelligence of a community in a specific predefined direction set by an organization [49]. It has gained interest thanks to its efficiency and applicability in multiple domains and tasks [57–61]. In machine learning and automation, it has become very popular for collaborative problem solving and dataset formulation and validation [57,59]. Users are expected to participate according to intrinsic (fun, personal growth, etc.) and extrinsic (payment, rewards, etc.) motives [59,60].

Within the MAtHE approach, crowdsourcing is promoted in several ways. First of all, crowdsourcing was promoted for the collaboration on the formulation of a database

of tampered audiovisual files. The importance of a real-world dataset for the training and evaluation of different machine learning and computer-assisted tampering detection approaches was highlighted in the previous sections. When a user provides a file for analysis in the web application, the file is stored temporarily to perform the analysis. After the results are provided, the user is asked for permission to save the file in our dataset. The file can be stored with or without accompanying metadata concerning the user's final decision. In case of a positive response, the file is stored.

Besides submitting files, users can also provide validation of existing files in the database. Such files may have been uploaded by other users but have no evaluation concerning their integrity, or may be files that are already annotated. It is common practice to include several annotators to strengthen the reliability and overall quality of the dataset.

In the case of crowdsourcing, the validation of files that have been uploaded by other users without an evaluation, collective intelligence, and collaboration are integrated into the framework. Users are encouraged to submit their files even if the analysis did not help them determine the authenticity of the file, to get help from other users. In case of a response, the uploader is informed about the other users' suggestions. This facilitates collaborative decision making, and is also a more efficient dataset creation strategy because files are not submitted only when the uploader can decide with confidence.

### 2.5. Common Use Case Scenarios

For a more efficient presentation of the functionality offered by the interface, the three most common use case scenarios are presented. In the results section, the implementation of the functionality in terms of the user experience (UX) choices, and the technical details are presented.

Scenario 1: A user submits a file and uses the toolbox to determine its authenticity.

In this common scenario, a user submits a file by uploading an audio or video file or by providing a YouTube link. After the analysis takes place on the server side, the visualizations are provided to the user. The user locates the points in time where inconsistencies are observed, listens to the audio, and makes a determination concerning the authenticity of the file, as was explained in Section 2.2 concerning the user-centered computer supported design. The user is then asked whether they are willing to contribute the file and their decision to the database. If they decide not to contribute, the file is deleted.

Scenario 2: A user is unable to decide and asks for help from the community.

The user submits the files, and, after they investigate the visualizations provided by the toolbox, they are unable to make a decision on the authenticity of the file. The user then decides to upload the file to the dataset unlabeled, so that it can be accessed by other users.

Scenario 3: A user browses the database to annotate files.

A user wants to contribute by annotating files that are already in the dataset. The interface provides randomly selected files from the database, along with the visualizations that come as outputs of the analysis. The user investigates the visualizations and makes a decision concerning the authenticity of the file, then chooses to submit their decision. Their decision is saved in the database, containing a label of the file (tampered/not tampered), and, optionally, the point in time where forging was detected and a short justification. The media file will still be available to other users for investigation after the annotation process. This means that a file may have multiple labels from different users, which is a common practice in crowdsourcing projects, since the input of one user cannot be considered totally reliable on its own.

## 3. Results

### 3.1. Convolutional Neural Network Regression Model for Signal-to-Reverberation-Ratio Estimation

As described in Section 2.2, among the integrated visualizations based on previous work, a CNN model was trained for the estimation of SRR from audio. The loss function

that was used for model training was mean square error (mse), which is a common choice for regression tasks, and Adamax was the optimizer. The goal of a regression training task is to minimize the difference between the estimated and the real values. The architecture and hyperparameters of the network were selected based on the existing literature and the trial-and-error-based micro-tuning during training. For the implementation, the Keras Python library was used [62]. Mel spectrograms were extracted to be used as input to the network, with 128 Mel scale coefficients. The spectrograms were extracted using the librosa library in Python [63]. They were extracted from overlapping windows of 1 s with a 50% overlap, leading to a final dataset of approximately 79,200 audio samples and a sampling frequency of 44,100 samples/second. For the output layer of the network, a fully connected network with a linear activation function was used to provide the predicted continuous value from 0 to 1. The mean square error was used as a metric for the evaluation of the network performance. A test set was used, that equals 20% of the entire dataset. The resulting mse was 0.029 (2.9%), a value that is considered acceptable for the task described.

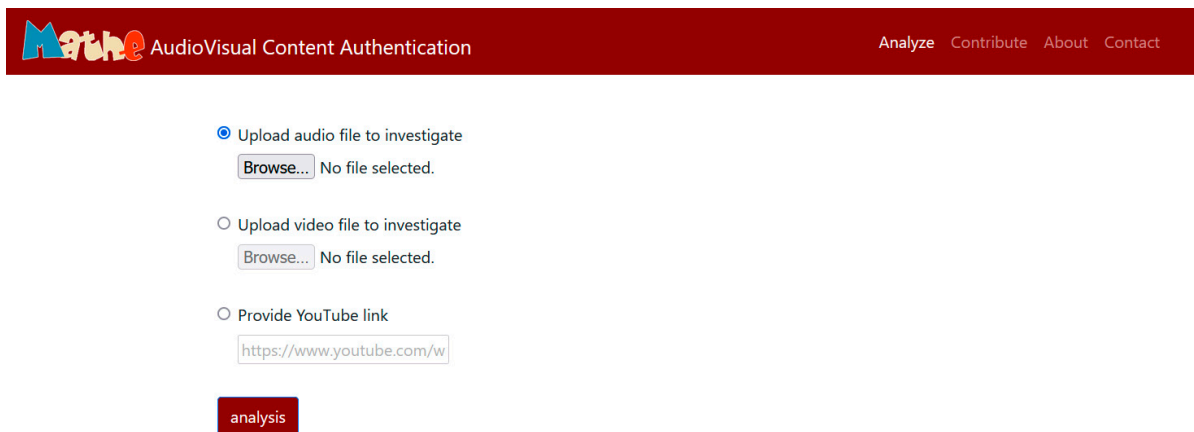
### *3.2. Implementation and Deployment of the Prototype for Human-Centered Audio Authentication Support and Crowdsourcing*

The web application was designed using the Flask web framework. This was a rational choice, taking into consideration the popularity of the Python programming language for data analysis and the successful deployment of similar web applications by our team [61]. The selection of a popular programming environment for such tasks may make the extensibility of the framework by contributors more appealing and viable. It was deployed on a dedicated Ubuntu virtual machine and was run on a Waitress production-quality pure-Python WSGI server. The application provides a back-end, where the algorithmic procedures take place, and a front-end graphical user interface.

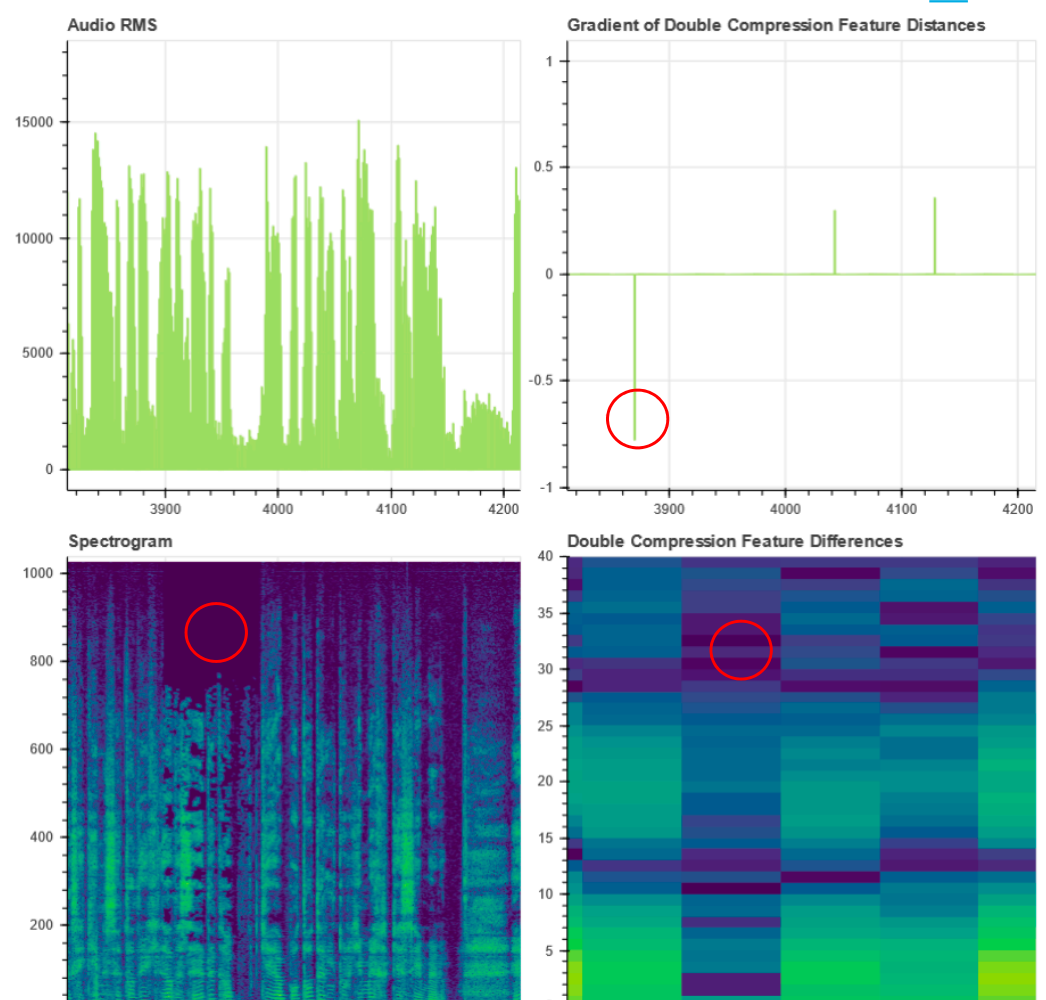
For audio analysis, namely audio file read, write, and audio feature and spectrogram extractions, the librosa Python library was used. The PyTube Python library was used for YouTube video downloading. The AudioSegment Python library was used for mp3 transcoding, in order to implement the double compression algorithm. The CNN model for SRR estimation was saved as a TensorFlow HDF5 [64] model, and it was loaded during server launch in order to perform regression on the back end for the provided files.

The interface offers two main functionalities: Analyze and Contribute.

In Analyze mode, the user can provide an audiovisual object for investigation (Figure 2). The interface gives the choice of uploading a media file or providing a YouTube link. The analysis takes place on the server and returns a set of interactive visualizations based on the algorithmic procedures. As explained in Section 2.2, the framework follows a modular approach, allowing extension with more visualizations in future versions. The diagrams are generated using the Bokeh library. The main idea is the use of a linked  $x$ -axis for all figures, which is the time axis. This means that, by zooming in on a specific time value of one of the available diagrams, the user zooms automatically in on the same time value on all diagrams. This enables a simultaneous combined investigation of the results of all available algorithmic procedures for the detection of suspicious points within the file. For example, in Figure 3, by zooming in on a peak of the gradient of the double compression feature distances (upper right), which indicates a suspicious point, it is clear that the other diagrams also indicate a possible tampering point. The red circles noting the suspicious behavior in the three visualizations were added for presentation reasons in this paper and were not part of the original results by the interface. For further information concerning the principles of the aforementioned algorithmic procedures, refer to [8]. Moreover, a media player is provided, where the user can play the audio/video file at a certain point in time to assist with their decision-making.



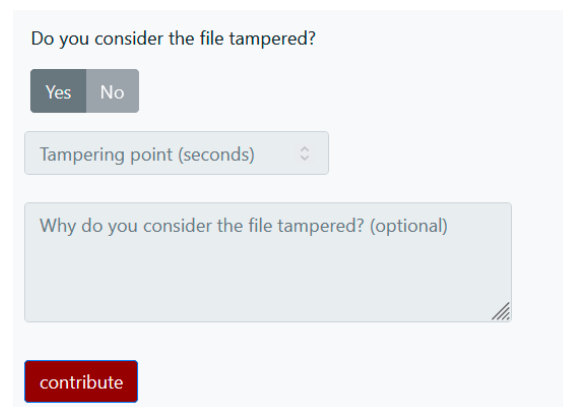
**Figure 2.** The interface where users can provide audio/video files for analysis.



**Figure 3.** An example of combined analysis. The user has zoomed in at a suspicious point in time, and three of the visualizations indicate forgery.

The media file is uploaded in a temporary folder for the needs of analysis, to be deleted later. However, the application provides the user with the option to submit the file to the dataset, along with the analysis results and, optionally, a personal opinion on the authenticity of the file, as described in Section 2.3. The files, the analysis results, and the user annotations are stored in the database.

In the Contribute mode, users can annotate existing files of the database, in a crowd-sourcing approach (Figure 4). A file is selected randomly. The same environment as in the Analyze mode is provided. The user has access to a media player, as well as the resulting set of interactive visualizations for the detection of possible tampering points. It is noted that the analysis results have already been saved to the database and are loaded directly. The analysis is a computationally complex procedure that requires time, which makes the annotation process much more time-consuming. After seeing the analysis, the user can determine whether and at which point in time the file is tampered, and can also optionally justify this decision with a short description. It is also possible to skip a certain file and select another one randomly. The user's opinion is stored in the database for the extension of the annotated dataset.



The image shows a web form for contributing to a dataset. At the top, it asks "Do you consider the file tampered?". Below this question are two buttons: "Yes" and "No". Underneath the buttons is a dropdown menu labeled "Tampering point (seconds)". Below the dropdown is a text input field with the placeholder text "Why do you consider the file tampered? (optional)". At the bottom of the form is a red button labeled "contribute".

**Figure 4.** In Contribute mode, users can browse files from the database along with their analysis visualizations, and annotate them concerning the detection of audio tampering.

There is also an About section, for users who wish to get more information concerning the project and the specifications of the algorithmic implementations, and a Contact section for anyone who wishes to ask questions or contribute to the project. The current version of the web application is uploaded to the domain [m3capps.jour.auth.gr](http://m3capps.jour.auth.gr) (accessed on 26 January 2022) in testing mode, for evaluation.

It has been explained that contribution to the project is encouraged and sought after. Contribution cannot only be achieved through the use of the interface by users who want to submit files to the database or to annotate existing entries. It can also refer to providing new models or algorithms and improving the ones we have already incorporated, following the modular architecture that has been described in Section 3. To address such needs and also to strengthen the transparency of the proposed procedure, the code of the interface and the backend functionality has been uploaded to GitHub and can be retrieved at <https://github.com/AuthJourM3C/MATHE-authentication> (accessed on 26 January 2022) under a GNU General Public License v3.0.

#### 4. Discussion

An AudioVisual Authentication application is presented, part of the MAtHE project on computer-aided support of journalists and simple users against misinformation. It specializes in the authentication of audiovisual content in an audio-driven technical approach. It has the form of a web application and implements the functionality of a framework that promotes machine-assisted, human-centered decision making, collective intelligence, and collaboration in the battle against the malicious tampering of audiovisual content. The functionality of the application is provided to the end-users through a very simple and intuitive interface where the user is asked to provide the FUI. The toolbox features several signal processing modules that are applied to the FUI, providing an interactive graph that contains several visualizations. These are based on different technical principles and algorithms that aim to assist the user who makes the final decision. Through crowd-



sourcing, a dataset of real-world tampered files is expected to be created and validated for the first time.

Along with a set of algorithms that are based on previous research, a CNN regression model for SRR estimation was presented and evaluated. The model performs SRR estimation with an MSE of 0.029, which is an acceptable resolution for the detection of different acoustic environments. Of course, like all the proposed techniques, it has several limitations, especially regarding studio recordings, using files with similar room acoustics for tampering, or simulating the reverberation environment to match the excerpts used for copy-move forgery.

The main contributions of the research are summarized as follows:

1. A novel approach is proposed for audio tampering detection, where decision making is held by the human-in-the-loop in a computer assisted environment. This approach makes use of technical advances, surpasses their limitations and unreliability, and proposes a solution that can be immediately applied in journalistic practice.
2. The solution is provided openly as a service, allowing its use by journalists and the audience, without any limitations on their equipment or platform.
3. The application follows a modular approach. This means that the modules that are integrated in the prototype can be updated easily, and more modules can be added in the near future.
4. A CNN model for data-driven SRR estimation to be used in the direction of audio authentication was presented and evaluated.
5. A crowdsourcing approach was introduced for both user collaboration in media authentication and dataset creation and annotation. Users contribute with their effort to the extension of the dataset of tampered media files and also assist other users who request support in the authentication of specific files.

Since this is the initial launch of the application, future research goals include a thorough evaluation of the interface and the provided tools. This can be done in focus groups containing professionals and also publicly, to evaluate how a broader audience receives the application. Such workshops can also provide publicity for the project. Crowdsourcing is a core aspect of MATHe, so it is crucial to approach potential users and engage them in using the application to collect initial results. The major outcome is expected to be the dataset, which will be publicly available. After the formulation of the initial dataset, more intense experimentation on the applicability of different machine learning architectures can take place. Moreover, the involvement of more contributors from the engineering world (researchers, students, coders, etc.) can aid the improvement and extension of the provided toolbox. For this reason, all the necessary material will be also publicly accessible through the application's website.

## 5. Limitations

One major limitation concerning the current research is that the interface, at the time of publishing this paper, was an evaluation prototype, as was indicated in the title and clarified throughout the paper. As a result, it will be used for the dissemination and evaluation of the framework, so several things are expected to change, be added, or be modified in future versions. Moreover, as explained in Section 3, the implementation depends on several third-party components, such as pyTube for YouTube content downloading, and the YouTube API itself. Since such components can be modified without warning, the application will have to be maintained to follow the latest functionality, updates and syntax of every component that is used. From the prospective of UX design, an error page has been integrated into the application to handle such exceptions, and to provide contact information for troubleshooting and bug reporting. Since the applied procedures are computationally heavy and require significant resources to guarantee a fast response time, for the evaluation version there is a restriction on the allowed file size and YouTube video length. A restriction in the allowed file types has been also set. These restrictions are expected to be lifted when the application is in production.

**Author Contributions:** Conceptualization, N.V., A.K., R.K. and C.D.; methodology, N.V., L.V. and R.K.; software, N.V., L.V. and R.K.; validation, A.K., R.K. and C.D.; formal analysis, N.V. and R.K.; investigation, N.V., R.K., A.K. and C.D.; resources, N.V., R.K. and A.K.; writing—original draft preparation, N.V., A.K., L.V., R.K. and C.D.; visualization, N.V. and L.V.; supervision, R.K. and C.D.; All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Data Availability Statement:** Not applicable.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Katsaounidou, A.N.; Dimoulas, C.A. Integrating Content Authentication Support in Media Services. In *Encyclopedia of Information Science and Technology*, 4th ed.; IGI Global: Hershey, PA, USA, 2018; pp. 2908–2919. [[CrossRef](#)]
2. Katsaounidou, A.; Dimoulas, C. The Role of media educator on the age of misinformation Crisis. In Proceedings of the EJTA Teachers' Conference on Crisis Reporting, Thessaloniki, Greece, 18–19 October 2018.
3. Katsaounidou, A.; Dimoulas, C.; Veglis, A. *Cross-Media Authentication and Verification: Emerging Research and Opportunities*; IGI Global: Hershey, PA, USA, 2019. [[CrossRef](#)]
4. Katsaounidou, A.; Vrysis, L.; Kotsakis, R.; Dimoulas, C.; Veglis, A. MATHe the game: A serious game for education and training in news verification. *Educ. Sci.* **2019**, *9*, 155. [[CrossRef](#)]
5. Katsaounidou, A.; Vryzas, N.; Kotsakis, R.; Dimoulas, C. Multimodal News authentication as a service: The “True News” Extension. *J. Educ. Innov. Commun.* **2019**, 11–26. [[CrossRef](#)]
6. Katsaounidou, A.; Gardikiotis, A.; Tsipas, N.; Dimoulas, C. News authentication and tampered images: Evaluating the photo-truth impact through image verification algorithms. *Heliyon* **2020**, *6*, e05808. [[CrossRef](#)] [[PubMed](#)]
7. Vryzas, N.; Katsaounidou, A.; Kotsakis, R.; Dimoulas, C.A.; Kalliris, G. Investigation of audio tampering in broadcast content. In Proceedings of the Audio Engineering Society Convention 144, Milan, Italy, 23–26 May 2018.
8. Vryzas, N.; Katsaounidou, A.; Kotsakis, R.; Dimoulas, C.A.; Kalliris, G. Audio-driven multimedia content authentication as a service. In Proceedings of the Audio Engineering Society Convention 146, Dublin, Ireland, 20–23 March 2019.
9. Bakker, P. New journalism 3.0—Aggregation, content farms, and Huffinization: The rise of low-pay and no-pay journalism. In Proceedings of the Future of Journalism Conference, Cardiff, UK, 8–9 September 2011.
10. Graves, L.; Cherubini, F. *The Rise of Fact-Checking Sites in Europe*; Reuters Institute for the Study of Journalism: Oxford, UK, 2016.
11. Bakir, V.; McStay, A. Fake news and the economy of emotions: Problems, causes, solutions. *Digit. Journal.* **2018**, *6*, 154–175. [[CrossRef](#)]
12. Verma, J.P.; Agrawal, S.; Patel, B.; Patel, A. Big data analytics: Challenges and applications for text, audio, video, and social media data”. *Int. J. Soft Comput. Artif. Intell. Appl.* **2016**, *5*, 41–51. [[CrossRef](#)]
13. Vlachos, A.; Riedel, S. Fact checking: Task definition and dataset construction. In Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science, Baltimore, MD, USA, 26 June 2014; pp. 18–22.
14. Zampoglou, M.; Papadopoulos, S.; Kompatsiaris, Y. Large-scale evaluation of splicing localization algorithms for web images. *Multimed. Tools Appl.* **2017**, *76*, 4801–4834. [[CrossRef](#)]
15. Zampoglou, M.; Papadopoulos, S.; Kompatsiaris, Y. Detecting image splicing in the wild (web). In Proceedings of the 2015 IEEE International Conference on Multimedia & Expo Workshops, Turin, Italy, 29 June–3 July 2015; pp. 1–6.
16. Sitara, K.; Mehtre, B.M. Digital video tampering detection: An overview of passive techniques. *Digit. Investig.* **2016**, *18*, 8–22. [[CrossRef](#)]
17. Grigoras, C.; Smith, J.M. Audio Enhancement and Authentication. In *Encyclopedia of Forensic Sciences*; Elsevier: Amsterdam, The Netherlands, 2013.
18. Maher, R.C. Audio forensic examination. *IEEE Signal Process. Mag.* **2009**, *26*, 84–94. [[CrossRef](#)]
19. Koenig, B.E. Authentication of forensic audio recordings. *J. Audio Eng. Soc.* **1990**, *38*, 3–33.
20. Zakariah, M.; Khan, M.K.; Malik, H. Digital multimedia audio forensics: Past, present and future. *Multimed. Tools Appl.* **2018**, *77*, 1009–1040. [[CrossRef](#)]
21. Gupta, S.; Cho, S.; Kuo, C.C.J. Current developments and future trends in audio authentication. *IEEE Multimed.* **2011**, *19*, 50–59. [[CrossRef](#)]
22. Rodríguez, D.P.N.; Apolinário, J.A.; Biscainho, L.W.P. Audio authenticity: Detecting ENF discontinuity with high precision phase analysis. *IEEE Trans. Inf. Forensics Secur.* **2010**, *5*, 534–543. [[CrossRef](#)]
23. Grigoras, C. Applications of ENF analysis in forensic authentication of digital audio and video recordings. *J. Audio Eng. Soc.* **2009**, *57*, 643–661.
24. Brixen, E.B. Techniques for the authentication of digital audio recordings. In Proceedings of the Audio Engineering Society Convention 122, Vienna, Austria, 5–8 May 2007.
25. Hua, G.; Zhang, Y.; Goh, J.; Thing, V.L. Audio authentication by exploring the absolute-error-map of ENF signals. *IEEE Trans. Inf. Forensics Secur.* **2016**, *11*, 1003–1016. [[CrossRef](#)]

26. Malik, H.; Farid, H. Audio forensics from acoustic reverberation. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Dallas, TX, USA, 14–19 March 2010; pp. 1710–1713.
27. Zhao, H.; Malik, H. Audio recording location identification using acoustic environment signature. *IEEE Trans. Inf. Forensics Secur.* **2013**, *8*, 1746–1759. [[CrossRef](#)]
28. Buchholz, R.; Kraetzer, C.; Dittmann, J. Microphone classification using Fourier coefficients. In Proceedings of the International Workshop on Information Hiding, Darmstadt, Germany, 8–10 June 2009; Springer: Berlin/Heidelberg, Germany, 2009; pp. 235–246.
29. Garcia-Romero, D.; Espy-Wilson, C.Y. Automatic acquisition device identification from speech recordings. In Proceedings of the 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, Dallas, TX, USA, 14–19 March 2010; pp. 1806–1809.
30. Hafeez, A.; Malik, H.; Mahmood, K. Performance of blind microphone recognition algorithms in the presence of anti-forensic attacks. In Proceedings of the 2017 AES International Conference on Audio Forensics, Arlington, VA, USA, 15–17 June 2017.
31. Malik, H. Acoustic environment identification and its applications to audio forensics. *IEEE Trans. Inf. Forensics Secur.* **2013**, *8*, 1827–1837. [[CrossRef](#)]
32. Narkhede, M.; Patole, R. Acoustic scene identification for audio authentication. In *Soft Computing and Signal Processing*; Springer: Singapore, 2019; pp. 593–602.
33. Patole, R.K.; Rege, P.P.; Suryawanshi, P. Acoustic environment identification using blind de-reverberation. In Proceedings of the 2016 International Conference on Computing, Analytics and Security Trends (CAST), Pune, India, 19–21 December 2016; pp. 495–500.
34. Qiao, M.; Sung, A.H.; Liu, Q. MP3 audio steganalysis. *Inf. Sci.* **2013**, *231*, 123–134. [[CrossRef](#)]
35. Yang, R.; Shi, Y.Q.; Huang, J. Detecting double compression of audio signal. In *Media Forensics and Security II, Proceedings of the IS&T/SPIE Electronic Imaging, San Jose, CA, USA, 17–21 January 2010*; SPIE: Bellingham, WA, USA, 2010; Volume 7541, p. 75410K.
36. Liu, Q.; Sung, A.H.; Qiao, M. Detection of double MP3 compression. *Cogn. Comput.* **2010**, *2*, 291–296. [[CrossRef](#)]
37. Seichter, D.; Cuccovillo, L.; Aichroth, P. AAC encoding detection and bitrate estimation using a convolutional neural network. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 2069–2073.
38. Lacroix, J.; Prime, Y.; Remy, A.; Derrien, O. Lossless audio checker: A software for the detection of upscaling, upsampling, and transcoding in lossless musical tracks. In Proceedings of the Audio Engineering Society Convention 139, New York, NY, USA, 29 October–1 November 2015.
39. Gärtner, D.; Dittmar, C.; Aichroth, P.; Cuccovillo, L.; Mann, S.; Schuller, G. Efficient cross-codec framing grid analysis for audio tampering detection. In Proceedings of the Audio Engineering Society Convention 136, Berlin, Germany, 26–29 April 2014.
40. Hennequin, R.; Royo-Letelier, J.; Moussallam, M. Codec independent lossy audio compression detection. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 726–730.
41. Luo, D.; Yang, R.; Huang, J. Identification of AMR decompressed audio. *Digit. Signal Process.* **2015**, *37*, 85–91. [[CrossRef](#)]
42. Maung, A.P.M.; Tew, Y.; Wong, K. Authentication of mp4 file by perceptual hash and data hiding. *Malays. J. Comput. Sci.* **2019**, *32*, 304–314. [[CrossRef](#)]
43. Khan, M.K.; Zakariah, M.; Malik, H.; Choo, K.K.R. A novel audio forensic data-set for digital multimedia forensics. *Aust. J. Forensic Sci.* **2018**, *50*, 525–542. [[CrossRef](#)]
44. Gärtner, D.; Cuccovillo, L.; Mann, S.; Aichroth, P. A multi-codec audio dataset for codec analysis and tampering detection. In Proceedings of the Audio Engineering Society Conference: 54th International Conference: Audio Forensics: Techniques, Technologies and Practice, London, UK, 12–14 June 2014.
45. Imran, M.; Ali, Z.; Bakhsh, S.T.; Akram, S. Blind detection of copy-move forgery in digital audio forensics. *IEEE Access* **2017**, *5*, 12843–12855. [[CrossRef](#)]
46. Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv* **2017**, arXiv:1706.06083.
47. Vryzas, N. Audiovisual Stream Analysis and Management Automation in Digital Media and Mediated Communication. Ph.D. Dissertation, Aristotle University of Thessaloniki, Thessaloniki, Greece, 2020.
48. Vrysis, L.; Tsipas, N.; Thoidis, I.; Dimoulas, C. 1D/2D Deep CNNs vs. Temporal Feature Integration for General Audio Classification. *J. Audio Eng. Soc.* **2020**, *68*, 66–77. [[CrossRef](#)]
49. Brabham, D.C. *Crowdsourcing*; MIT Press: Cambridge, MA, USA, 2013.
50. Vrysis, L.; Hadjileontiadis, L.; Thoidis, I.; Dimoulas, C.; Papanikolaou, G. Enhanced Temporal Feature Integration in Audio Semantics via Alpha-Stable Modeling. *J. Audio Eng. Soc.* **2021**, *69*, 227–237. [[CrossRef](#)]
51. Bountourakis, V.; Vrysis, L.; Konstantoudakis, K.; Vryzas, N. An enhanced temporal feature integration method for environmental sound recognition. *Acoustics* **2019**, *1*, 410–422. [[CrossRef](#)]
52. Vrysis, L.; Thoidis, I.; Dimoulas, C.; Papanikolaou, G. Experimenting with 1D CNN Architectures for Generic Audio Classification. In Proceedings of the Audio Engineering Society Convention 148, Vienna, Austria, 2–5 June 2020.
53. Vrysis, L.; Vryzas, N.; Sidiropoulos, E.; Avraam, E.; Dimoulas, C.A. jReporter: A Smart Voice-Recording Mobile Application. In Proceedings of the Audio Engineering Society Convention 146, Dublin, Ireland, 20–23 March 2019.

54. Korvel, G.; Treigys, P.; Tamulevicius, G.; Bernataviciene, J.; Kostek, B. Analysis of 2d feature spaces for deep learning-based speech recognition. *J. Audio Eng. Soc.* **2018**, *66*, 1072–1081. [[CrossRef](#)]
55. Ciaburro, G. Sound event detection in underground parking garage using convolutional neural network. *Big Data Cogn. Comput.* **2020**, *4*, 20. [[CrossRef](#)]
56. Ciaburro, G.; Iannace, G. Improving smart cities safety using sound events detection based on deep neural network algorithms. *Informatics* **2020**, *7*, 23. [[CrossRef](#)]
57. Estellés-Arolas, E.; González-Ladrón-de-Guevara, F. Towards an integrated crowdsourcing definition. *J. Inf. Sci.* **2012**, *38*, 189–200. [[CrossRef](#)]
58. Vrysis, L.; Tsiapas, N.; Dimoulas, C.; Papanikolaou, G. Crowdsourcing audio semantics by means of hybrid bimodal segmentation with hierarchical classification. *J. Audio Eng. Soc.* **2016**, *64*, 1042–1054. [[CrossRef](#)]
59. Vrysis, L.; Tsiapas, N.; Dimoulas, C.; Papanikolaou, G. Mobile audio intelligence: From real time segmentation to crowd sourced semantics. In Proceedings of the Audio Mostly 2015 on Interaction with Sound, Thessaloniki, Greece, 7–9 October 2015; pp. 1–6.
60. Cartwright, M.; Dove, G.; Méndez Méndez, A.E.; Bello, J.P.; Nov, O. Crowdsourcing multi-label audio annotation tasks with citizen scientists. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Glasgow, UK, 4–9 May 2019; pp. 1–11.
61. Vrysis, L.; Vryzas, N.; Kotsakis, R.; Saridou, T.; Matsiola, M.; Veglis, A.; Arcila-Calderón, C.; Dimoulas, C. A Web Interface for Analyzing Hate Speech. *Future Internet* **2021**, *13*, 80. [[CrossRef](#)]
62. Chollet, F.; Eldeeb, A.; Bursztein, E.; Jin, H.; Watson, M.; Zhu, Q.S. *Keras*; (v.2.4.3); GitHub: San Francisco, CA, USA, 2015; Available online: <https://github.com/fchollet/keras> (accessed on 26 January 2022).
63. McFee, B.; Raffel, C.; Liang, D.; Ellis, D.P.; McVicar, M.; Battenberg, E.; Nieto, O. librosa: Audio and music signal analysis in Python. In Proceedings of the 14th Python in Science Conference, Austin, TX, USA, 6–12 July 2015; Volume 8, pp. 18–25.
64. Collette, A. *Python and HDF5: Unlocking Scientific Data*; O'Reilly Media, Inc.: Newton, MA, USA, 2013.