Team Report 1.2 Project Solution Design and Business Applicability

DeepFakeChain

Due date	17/10/2022
Submission date	18/10/2022
Version	1.0
Authors	Nikos Giatsoglou (CERTH) Symeon Papadopoulos (CERTH) Dora Kallipolitou (Zelus) Stella Markopoulou (Zelus)

Grant Agreement No.: 957228
Call: H2020-ICT-2018-2020
Topic: ICT-54-2020
Type of action: RIA

Document Revision History

Version	Date	Description of change	List of contributor(s)
v0.1	09/09/2022	Template circulated by the TruBlo team.	TruBlo team
v0.2	20/9/2022	Structure created	CERTH
v0.3	23/9/2022	Completed Project Drivers	CERTH
v0.4	27/9/2022	Completed Summary of Existing Functions	CERTH & Zelus
v0.5	30/9/2022	Completed Solution & Requirements	CERTH
v0.6	3/10/2022	Completed Risks and Mitigation	CERTH
v0.7	14/10/2022	Revisions	CERTH & Zelus
v1.0	18/10/2022	Version ready for final submission.	

Disclaimer

The information, documentation and figures available in this deliverable are written by the DeepFakeChain team and do not necessarily reflect the views of the TRUBLO consortium or of the European Commission.

The TRUBLO consortium and the European Commission is not liable for any use that may be made of the information contained herein.

Project co-funded by the European Commission in the H2020 Programme	
Nature of the deliverable:	R: Document, report
Dissemination Level:	CO: Confidential to TruBlo project and Commission Services

Funded by the Horizon 2020 Framework Programme of the European Union

EXECUTIVE SUMMARY

This document is the second deliverable of the research project DeepFakeChain, selected by the cascade-funding project NGI TruBlo (GA 957228), which aims to develop a scientific framework and testbed for collaborative deepfake media detection. The second deliverable focuses on the analysis of the state-of-the-art and the existing solutions for collaborative media annotation and deepfake detection, which will help design the DeepFakeChain testbed. Towards this goal, the deliverable specifies functional and non-functional requirements and identifies potential risks to the success of the DeepFakeChain project.

TABLE OF CONTENTS

Funded by the Horizon 2020 Framework Programme of the European Union

Funded by the Horizon 2020
Framework Programme of the European Union

LIST OF FIGURES

Funded by the Horizon 2020
Framework Programme of the European Union

LIST OF TABLES

Funded by the Horizon 2020
Framework Programme of the European Union

ABBREVIATIONS

Abbreviations	Definitions
AI	Artificial Intelligence
API	Application Programming Interface
CGI	Computer Generated Imagery
CNN	Convolutional Neural Network
DDOS	Distributed Denial-of-Service
DFD	DeepFake Detection (dataset by Google)
DFDC	DeepFake Detection Challenge
DL	Deep Learning
EU	European Union
GAN	Generative Adversarial Network
GPU	Graphics Processing Unit
GRU	Gated Recurrent Unit
IPFS	InterPlanetary File System
LSTM	Long Short Term Memory
ML	Machine Learning
MPEG	Moving Picture Experts Group
MQTT	MQ Telemetry Transport
NFS	Network File System
OCR	Optical Character Recognition
PRNU	Photo-Response Non-Uniformity

Funded by the Horizon 2020
Framework Programme of the European Union

REST	REpresentational State Transfer
RNN	Recurrent Neural Network
SIFT	Scale-Invariant Feature Transform
SOTA	State-Of-The-Art
SURF	Speeded Up Robust Features
URL	Uniform Resource Locator
US	United States
VAE	Variational AutoEncoder
VGG	Visual Geometry Group
XAI	Explainable Artificial Intelligence

Funded by the Horizon 2020
Framework Programme of the European Union

1 PROJECT DESCRIPTION

Project name	DeepFakeChain
Link to project on TruBlo website	https://www.trublo.eu/deepfakechain/
Primary contact	Dr Symeon Papadopoulos, papadop@iti.gr
Project members	Mr Nikolaos Giatsoglou, ngiatsog@iti.gr Dr George Kordopatis, georgekordopatis@iti.gr Ms Stella Markopoulou, s.markopoulou@zelus.gr
Organisation(s)	CERTH, Zelus
Organisation's website	https://www.certh.gr, https://www.zelus.gr
Short project summary	
What is the focus of your project?	The development of a scientific testbed to research innovative deepfake detection algorithms and combinations with human judgement. The project targets the media sector and can also be extended to the content moderation use case by extending to the detection of more general harmful content. Blockchain technology will be used to notarise the data that is uploaded on our testbed and enhance the trustworthiness of the media evaluations.
Why is a new/better solution needed?	The generation of deepfake synthetic media that are exceedingly difficult to detect is a worrying trend with potentially devastating impact on society. Currently, no solution exists for perfect automatic detection, nor one is expected in the near future due to the ongoing refinement of generated deepfakes. To address this issue, DeepFakeChain combines algorithmic solutions with human expert opinion, which is better in discerning context and valuable for detection. Our solution could also cover content moderation, which is an emerging market with few established solutions.
How will your solution be better?	By integrating expert human judgement with automatic evaluations, capitalising on the best of both worlds. By offering access and explainability features of cutting-edge AI algorithms to non-technical professionals. By increasing the trustworthiness of our

Funded by the Horizon 2020 Framework Programme of the European Union

	platform’s data and decisions through blockchain.
Extra: How does this project contribute to “trustable content on future blockchains”	By storing proofs of authenticity of our platform’s data (media, annotations, user profiles, decisions) in a distributed blockchain network, making them available to the platform’s end users and third parties for validation and auditing.
Type of project	(X) Scientific/research () Commercial, potential startup () Open source, non-commercial () Other, pls add 1-4 words if selected
Technologies used	AI-based deepfake detection algorithms, ensemble learning, consensus and truth-discovery algorithms, permissioned blockchain network, reverse media search
Use of Alastria resources	Yes

Funded by the Horizon 2020
Framework Programme of the European Union

2 PROJECT SCOPE

The DeepFakeChain project seeks to address the threat of fake media and harmful content by investigating new approaches for **collaborative media annotation and deepfake detection**. This platform will allow human users to collaborate on the verification of deepfakes and access innovative algorithms for automatic detection. We strongly believe that neither human nor machine intelligence alone can successfully mitigate deepfakes, hence our solution will capitalise on the individual strengths of both, i.e., the contextual intelligence of humans and the raw processing power of machines. Furthermore, blockchain technology will be key in ensuring the trustworthiness of the platform's uploaded data and decisions.

DeepFakeChain targets primarily the media sector, aiming to equip journalists and fact-checkers with advanced digital tools to address deepfakes. Media annotation in particular is an emerging market with a few established solutions, which are described and analysed in terms of features in Section 4 of this deliverable. In the bigger picture and in the long run, DeepFakeChain aspires to address more general types of harmful content, being mindful of the high demand of large social media and news companies for **automatic content moderation** solutions.

DeepFakeChain is a project with a **strong research focus**. Due to this, we intend to develop a **scientific framework and testbed** with minimal features compared with the ones offered by the solutions of Section 4 in order to experiment with innovative ways and models for deepfake detection. Specifically, the research agenda of DeepFakeChain, elaborated in the scientific drivers of Section 3 and summarised here for convenience, is:

- ➔ New algorithms that generalise better than the state-of-the-art deepfake media seen "in the wild".
- ➔ New ways of integrating human and machine evaluations of deepfakes.
- ➔ New ways to explain the decisions of deepfake detection algorithms.

In the following, we explicitly state the in-scope and out-of-scope features of the platform.

2.1 IN SCOPE

The high level features of DeepFakeChain where our research and development effort will focus include the following:

- A scientific framework and testbed to develop innovative collaborative approaches for the detection of deepfakes by humans and machines.
- New algorithmic services for the automatic evaluation of multimedia content.
- New approaches to combine machine and human judgements.
- New explainability features to increase the trustworthiness of deepfake detection algorithms.
- A repository of training data for research.
- Persistent distributed storage of proofs of authenticity of the platform's data.
- An independent mechanism to validate the platform's data against tampering**.**

In terms of implementation, we intend to capitalise and expand on existing functionality by CERTH and Zelus, described in section 5.4. As proof of concept of the above features, we plan to integrate them in a web-based demonstrator application with the following features:

- Uploading, storage and collaborative annotation of multimedia content, offering access to algorithmic services for deepfake detection.
- A simple authentication system for controlled access to the application.
- A channel-based scheme to organise media and workflows on specific topics.
- A reverse search mechanism to find similar or near-duplicate videos.
- Visualisation panels that highlight the explainability features of our algorithms.

2.2 OUT OF SCOPE

The following features are out-of-scope for the DeepFakeChain project although they may become relevant in a future evolution of our solution.

- A ready-to-enter the market solution for collaborative media annotation.
- A platform for educating citizens on deepfakes and false information.
- A service for producing legal evidence of media tampering that can be used in the courtroom.
- An open platform offering public users unrestricted access to its services.

Funded by the Horizon 2020
Framework Programme of the European Union

3 PROJECT DRIVERS

The main drivers of DeepFakeChain are *innovation*, *research*, *social impact*, and *demand* drivers. These concern the final output, after Phase 2 of the TruBlo project, and are elaborated in the following.

3.1 INNOVATION DRIVERS

In the previous years, a series of key events have dramatically changed our attitude towards the Internet and digital technology. The rise of fake news in the 2016 and 2020 US elections and the COVID pandemic have alarmed us over the impact of false information and urged us towards the detection and filtering of harmful content detection. In addition, the advancements of AI and ML have fuelled our dreams of widespread automation but also raised the issues of bias, lack of transparency, and opaque decision making that are inherent to these technologies. These issues have created fertile ground for innovation, where DeepFakeChain intends to grow.

Innovative algorithms for automatic deepfake detection: Deepfakes are a type of false information, potentially harmful. Compared with traditional fake media, the photorealism of deepfakes has been made possible only recently, thanks to deep learning advances. Similarly, it is commonly thought that new technology will be required to mitigate deepfakes.

→ DeepFakeChain will develop innovative ML algorithms for deepfake detection.

Human-machine collaboration for deepfake detection: Currently, the issue of maintaining humans-in-the-loop in automated decision making is hotly debated yet no definite solution has emerged. Deepfake detection is an example of a task that could greatly benefit from the contextual knowledge of humans, which is difficult for machines to mimic at this point.

→ DeepFakeChain will work towards an approach for human-machine collaboration on the task of deepfake detection.

Explainable decisions for deepfake detection: Along with the demand for AI services, concerns are growing over the reliability and predictability of AI. Furthermore, new regulatory frameworks like the European Digital Services Act place great emphasis on the transparency and explainability of automated decision making. Deepfake detection is a typical task that requires explanations to convince users why a media file is classified as deepfake.

→ DeepFakeChain will aim to provide accessible and comprehensive explanations for the judgement of media content as deepfake.

Distributed secure storage for the transparency of centralised platforms: While decentralised technologies are becoming once again popular, with such notable examples as the IPFS network and distributed ledgers, centralised platforms will continue playing a key role to the Internet due to their improved user experience, and ease of deployment, management, and efficiency. Transparency remains a thorny issue of centralised platforms that needs to be addressed. Blockchain can be a solution by notarising the operations of the platform in an independent network, available for all users to check. While a wealth of such applications have been proposed for blockchain, few have been implemented and gained traction.

→ DeepFakeChain will build on blockchain to secure the provenance and authenticity of human and algorithmic annotations on uploaded media files and metadata.

Funded by the Horizon 2020
Framework Programme of the European Union

3.2 RESEARCH DRIVERS

In the past years, deepfake detection has sparked intense research, raising intriguing questions over what constitutes a genuine media product and whether AI algorithms can discern it. The focus of these efforts has been primarily ***human face manipulation***, due to the potentially negative impact of impersonations and identity-based fraud. Below we present key challenges of deepfake detection that form the research drivers of DeepFakeChain, as well as its research agenda. Our focus is on deepfake detection although other types of manipulations may be considered in the duration of the project.

Generalisation of deepfake detection algorithms: Since the appearance of deepfakes in mid 2010s, many ML techniques have been proposed for their detection, for example, based on traditional forensics (signatures, spectral methods), handcrafted features (image texture, blending artefacts, physiological signals), and deep-learning methods that learn the optimal features for the task. Currently, deep learning methods are the most effective techniques, frequently achieving >90% accuracy in standard datasets [1]. This performance, however, does not generalise to deepfake media "in the wild", which exhibit less ideal characteristics, e.g., multiple faces, sub-optimal lighting, reprocessing (compression). Other approaches require the knowledge of the deepfake generation algorithm, which is not generally available. There has been initial work to improve the generalisation of deepfake detection, notably based on ensemble and transfer learning, but it is still at an early stage.

→ DeepFakeChain will address the poor generalisation of deepfake detection algorithms by thoroughly researching promising techniques including ensemble and transfer learning.

Integration of human input: Currently, humans can still discern low-quality deepfakes from genuine media based on apparent artefacts or due to contextual knowledge (e.g., Nicolas Cage's face appearing in wildly implausible contexts or simply an uncredited movie). As deepfake generation becomes more competent and leaves less trails for forensics, we believe that human reasoning will be a necessary ingredient of deepfake media mitigation, alongside detection algorithms, although their combination is not clear at the moment. For a human-oriented Internet, we also require human supervision and control over automatic decisions.

→ DeepFakeChain will research effective ways to integrate human input with detection algorithms in order to achieve high levels of trust in the classification of media as deepfakes. This will ensure humans stay in the loop of decisions and can override machine decisions if they are found to be unreliable.

Explainability of AI algorithms: Despite impressive achievements of modern AI algorithms, they can be unpredictable and introduce bias. This causes unease with regard to their usage in decision making. To address these concerns, researchers have recently investigated techniques to explain the outcomes of AI algorithms, and thus make them more dependable. There are many explainable AI (XAI) approaches. Some AI techniques, notably decision trees, are already reasonably understandable even to human non-experts. The biggest challenge however lies in artificial neural networks, which are notoriously opaque and have driven the current advancements of AI. Approaches to explain neural networks include saliency maps (or heatmaps), attribution methods, and distillation to simpler explainable models. Interestingly, the drive towards explainability has recently revived the debate between symbolic and sub-symbolic (or connectionist) AI [2]. Symbolic AI, chiefly represented by expert systems and knowledge bases, were popular during the 70s and later fell in popularity to connectionist systems such as neural networks. Symbolic methods however are inherently explainable and some approaches seek to combine them with neural networks.

→ DeepFakeChain will research effective techniques to reason about the operation of deepfake detection algorithms so that they can be relied upon by media professionals.

Funded by the Horizon 2020 Framework Programme of the European Union

3.3 SOCIAL IMPACT DRIVERS

The potential negative social impact of deepfakes have been extensively documented[3, 4]. Below, we summarise the most important threats.

Politics: Deepfakes can be used to impersonate politicians, state officials, and other public figures with the goal to manipulate public opinion, sway voters, and create diplomatic incidents. Threats include smearing campaigns, election interference, fake news and disinformation, fabrication of fake terrorist content, compromise of national security and destabilisation. Perpetrators can be corrupted governments, political parties, intelligence offices, and other shady companies.

Business: Deepfakes can be used for fraud and other types of malicious activities. Threats include bypassing biometric authentication systems, brand sabotage, defamation of corporate management, or even stock manipulation, considering e.g., the effect of tweets from prominent figures such as Elon Musk.

Judicial system: The main threat is evidence tampering. The effort to verify the presented multimedia in the courtroom will strain trials, which are already lengthy and expensive.

Pornography: Creating false identities in pornography is a sensitive issue with significant legal and ethical ramifications. Threats include celebrity, child, and revenge pornography.

Intimidation: Deepfakes can be used as part of a variety of intimidation attacks to individuals, complementary to what has already been described. Examples include blackmail, doxing, bullying, etc.

Post-truth society: On a more general note, perhaps the greatest concern over deepfakes is their capacity to sow distrust over all types of visual information, ushering us in a post-truth world where no multimedia can be trusted. This is by no means a new concern; history and art have already shown us that not even genuine footage can be relied upon to reveal the truth, with such poignant examples as the Zapruder film, which captured the assassination of John F. Kennedy, and Antonioni’s movie BlowUp, in which a highly magnified blow up of a photograph may or may not provide evidence of a crime. Tampered media also have a long history but deepfakes can reach new levels of unreality due to the photorealism of very improbable depictions.

→ As deepfakes become more prevalent and harder to detect, they will put great pressure on professionals such as journalists, lawyers, and police authorities. DeepFakeChain targets media professionals but its services could be made available to any concerned citizen. Some limitations are foreseen, for example requiring registration and approval for human evaluation, considering the limited deployment resources. In addition, DeepFakeChain’s decisions are not envisioned to have legal standing, at least in the first design phase.

3.4 DEMAND DRIVERS

As the threat of deepfakes and other digital harmful content grows, there is increasing demand for technological solutions that integrate the best approaches from research and allow non-technical professionals to use them. Below we present the main demand drivers of DeepFakeChain.

Connecting cutting-edge research with users: As the pace of advancements in science and technology has been steadily increasing, there is increased pressure in making use of the new knowledge as quickly as possible. This is due to both market pressures and the emergence of new threats, which exploit the lack of understanding over new technologies. In the case of

Funded by the Horizon 2020
Framework Programme of the European Union

deepfakes, their potential harm, as described in the Societal Impact Drivers subsection, is grave enough to justify rapidly developing mitigation measures and making them available to the public.

→ DeepFakeChain will connect cutting-edge research on deepfake detection with the end users.

Trustworthy online content: As the Internet becomes infested by false and harmful content, there is a growing need for services that guarantee the trustworthiness of online media. These should offer transparent and explainable decisions over which content is fake, as opaque decisions can either create a false sense of security or make people doubtful about them.

→ DeepFakeChain will provide explanations that justify the evaluations of media content as deepfakes and demarcate their limitations so that they can be trusted.

Flexible cross-organisational collaboration: Organisations spend a lot of time and effort in organising their internal workflows, with frequently suboptimal results. This problem is exacerbated in the case of cross-organisational collaboration, which are nevertheless crucial in the modern connected world and its demands for extraversion and top-quality output. This is certainly true for the media sector where collaborative and citizen journalism have been highly successful. Cross-organisational collaboration should be easy to deploy, flexible to give freedom to the collaborators, yet retain editorial control to individual companies over the information that they want to share.

→ DeepFakeChain intends to build a basis for seamless collaboration on content annotation and moderation.

Funded by the Horizon 2020
Framework Programme of the European Union

4 SUMMARY OF EXISTING FUNCTIONALITY OR SOLUTION

The services of DeepFakeChain will ultimately target journalists and media professionals. In TR1.1, we overviewed existing solutions that are close to the vision of our project and fall into the categories of *collaborative media verification* and *deepfake detection*. In the following, we present representative examples from each category, analyse their features, and conclude by presenting relevant state-of-the-art on deepfake detection.

4.1 COLLABORATIVE VERIFICATION PLATFORMS

From the overview of TR1.1, Check and TrulyMedia stand out as the most complete and updated solutions for media verification. The other platforms either refer to more general forms of annotation, as in the case of Hypothes.is and Birdwatch, or are unmaintained, or specifically related to initiatives at a given context, e.g., elections.

4.1.1 CHECK

Check, formerly known as CheckDesk, is a platform by the not-for-profit company Meedan, allowing journalists and fact-checkers to collaborate for the verification of online content. Interestingly, Check also supports the use case of citizen journalism, by providing "tiplines" for social media users to upload and request verification of seen media. Tiplines are implemented through integration with popular chat applications and the help of bots. The platform has powered various verification campaigns, most notably, Propublica's Electionland, reporting on voting issues during the 2016 US elections, and FirstDraft's CrossCheck, addressing misinformation in the 2017 French Presidential Election. Free access to the platform is provided through registration and the source code is publicly available[1], although not intended for production-grade deployment.

Check's workflow is organised around *workspaces* and *items*. Workspaces are collaborative channels in which users can join and verify content on a specific topic such as a diplomatic incident, elections, or a war. Currently, Workspaces are *private* and only members of the workspace can contribute to them, where membership is attained through invitation. Items include content to be verified, such as text, audio, images, and videos, and are published to a specific workspace. Older versions of the platform differentiated between text-only content (called *claim*) and multimedia content (called *media item*) but the current interface does not allow uploading text-only claims. This indicates a shift of emphasis of the platform towards multimedia.

In more detail, items represent the smallest unit of information requiring verification. Each item has a *status*, which can be *unstarted*, *inconclusive*, *in progress*, *false, verified*, while custom values can be defined as well. Users can evaluate the veracity of the items in two ways: *notes*, which are free-form comments of the users, and *annotations*, which are structured questions about the items like what is the time and place of the item. They can further assign the item to workspace members, lock and unlock it for editing, flag it as spam or graphic, and even embed it in external sites, attracting the attention of external collaborators. Finally, through collective effort, a terminal status is reached which can be either *verified* or *false.* It is worth stressing that the consensus for the judgement is entirely manual as any user can change the status of the item.

Regarding the user authentication and roles, Check offers light authentication as only an email and a pseudonym is required to join the platform. Trust is nevertheless a requirement, as users can join workspaces only through extended invitations and join links, the difference being that

[1] https://github.com/meedan/check

join links require approval. Other features provide further security such as two-factor authentication, security alerts, cookies encrypted to be invisible to third parties, protection against distributed denial-of-service (DDoS) attacks through Cloudfare, etc. User roles fall into three categories with fixed privileges:

- **Admins** have full rights over the workspaces such as managing new members, the workspace's folders and rules, sources, etc.
- **Editors** have similar rights to admins except for some specific privileges like removing / duplicating workspaces, adding bot services, changing languages and creating new admins.
- **Collaborators** represent the basic users of Check, lacking rights for user management, editing existing workspaces, and adding new sources.

Interestingly, Check is quite liberal with managing items, as all user roles can change their status, lock and unlock them, and even assign tasks to each other. Some access control can be customised at the level of folders, which organise collections of items inside a workspace.

Some algorithmic features help during the annotation process. These are:

- optical character recognition (OCR) to automatically extract text from images
- audio and video transcription, supporting 13 languages
- reverse image search, provided by Google
- similarity search of new media with past media once they are uploaded. Interestingly, these are offered as suggestions to users, who can accept or decline them, and the decisions help train the algorithm and increase its accuracy.
- automatic detection of graphic content and classification as Adult, Medical, or Violence through Google's Vision API[2]. Admins can configure rules to automatically filter the incoming content or display a warning screen over it.

Another interesting feature of Check is *bots*, i.e., non-human users that can automate non-intelligent processing on incoming content. This processing includes importing fact-checks from external databases, crawling for tweets, and integrating with external platforms such as Slack. Most importantly, bots integrate with external chat platforms such as WhatsApp, Viber, Telegram, Facebook, Twitter, and LINE, in order to create *tiplines.* These are services that can be used by any Internet user to upload media and request verification from Check's fact-chekers. After processing the requests, the latter can reply back to the user reports. This tipline feature allows Check to support citizen journalism.

Finally, thanks to Check being open-source, its architecture is available through the Github site and can be seen in Figure 1. Check is structured based on the **micro-service pattern**, where distinct components are decoupled via REST APIs. We see that the data layer contains various storage solutions allowing different functionalities on the uploaded content, notably, indexes for enabling similarity searches. The services layer then decouples the core system from other processing functions like parsing content from URL links (e.g., videos from YouTube), analysing text and media (e.g., for similarity search), taking screenshots from web pages, etc. The client layer then provisions for different types of interfaces like web, mobile, and browser extension interfaces. Finally, the bot layer contains the bots offered by the platform.

[2] https://cloud.google.com/blog/products/ai-machine-learning/filtering-inappropriate-content-with-the-cloud-vision-api

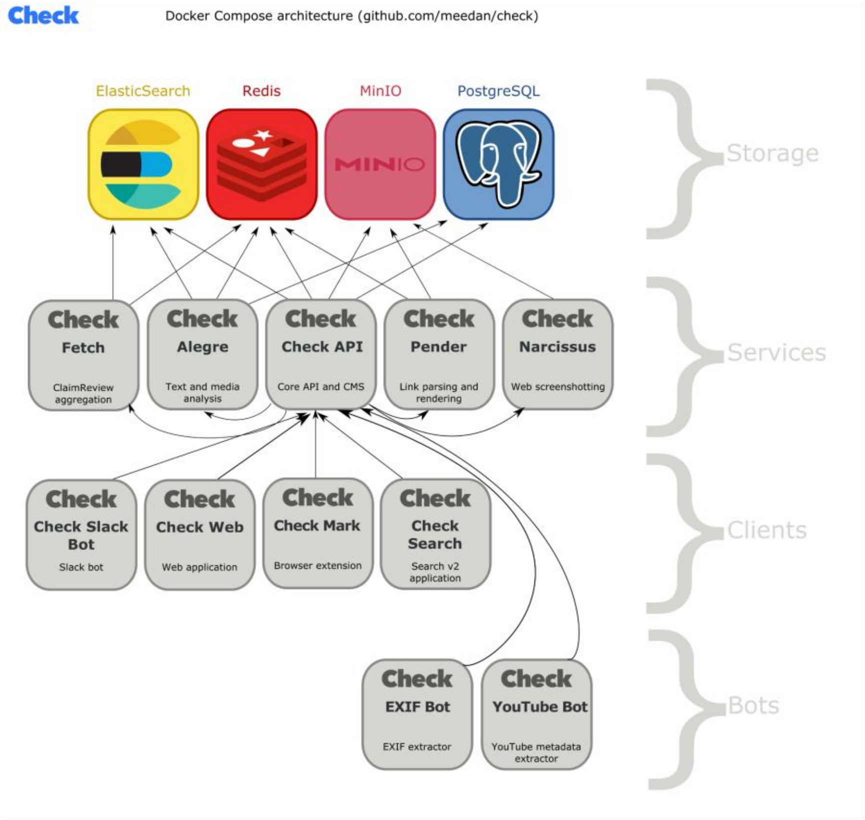


FIGURE 1: THE ARCHITECTURE OF CHECK (TAKEN FROM [3])

4.1.2 TRULYMEDIA

TrulyMedia is a collaboration platform for the verification of online digital content, especially coming from social networks. It has been developed jointly by the German broadcaster Deutsche Welle and the Greek software company ATC, and has received funding from the European Commission and Google's Digital News Initiative. It targets journalists, as well as human rights investigators. It is a popular platform; after being tested in the 2017 German Elections, it has been used by various organisations such as the Amnesty International[4] and the German public broadcaster ZDF[5], as well as in other initiatives[6].

In terms of features, the platform bears some similarities to Check. It allows users to organise media content in collections, annotate and tag them, save key frames in videos, and identify issues to verify. Based on these issues, the content can then be marked as *raw*, *pending*, *unclear*, *verified*, or *fake*. In addition, users can mark content as being graphic, having licence issues, or containing profanity. Sources can also be stored to document their reliability and authoritativeness. The platform connects with various external services such as Google Maps, TinEye, WolframAlpha, Google Reverse Image Search, Yandex, Snopes, Pipl in order to provide useful shortcuts to commonly used verification tools. Journalists collaborate by creating profiles, joining teams, and receiving live updates of their activities. They can further assign tasks with priorities to each other and monitor their progress. TrulyMedia finally supports chatting for direct communication, through personal or group messages, which is not supported by Check.

[3] https://github.com/meedan/check
[4] https://www.truly.media/amnesty-uses-truly-media/
[5] https://www.truly.media/zdf-trusts-truly-media-to-tackle-disinformation/
[6] https://www.truly.media/our-verification-network-gets-bigger-new-pilots-in-myanmar-and-georgia/

Funded by the Horizon 2020 Framework Programme of the European Union

The biggest distinction of TrulyMedia with Check concerns data acquisition. TrulyMedia integrates very closely with social media like Twitter, Facebook, YouTube, displaying content streams directly in its page and enabling search with advanced filters. These include searching by time, media type, hashtags, authors, mentioned entities, and sentiment, among others. Users can then drag-and-drop content to their collection in addition to uploading media manually. The platform also displays detailed social media analytics (called *insights*) such as popular hashtags, media, most mentioned entities, top influencers, and various trends.

Currently, TrulyMedia is actively developed and receives funding to implement new features. For example, in the context of the EU project AI4Media, TrulyMedia has reported developing AI features for i) the analysis and monitoring of social media items (clustering, community detection, similarity detection, topic detection, etc.), ii) the analysis and monitoring of Twitter accounts (analysis of users' history, hate speech detection, user classification, etc.), iii) the search across audiovisual verification repositories (audio/video comparison, reverse audio/video search, duplicate detection, etc.)[7]. Finally, the Digger project[8], where Fraunhofer IDMT was partner, investigated detecting deepfakes by using state-of-the-art audio forensics technologies.

4.2 DEEPFAKE DETECTION SERVICES

From the overview of TR1.1, we present the DeepWare Scanner and the DeepFake-o-Meter services. Both focus exclusively on deepfake detection, are open to the public, and open-source. In contrast, other tools are proprietary, offer limited access without a paid licence, do not reveal the underlying technology, and/or use deepfake detection as an internal feature. For example, RealityDefender is a software that sprang from Microsoft's research and targets large enterprises and organisations such as governments. DuckDuckGoose is more commercial-oriented but does not provide adequate information about its implemented algorithm. Sensity and BioID focus on cybersecurity, targeting identity verification systems. In the following, we present DeepWare Scanner and DeepFake-o-meter, extracting their most important features.

4.2.1 DEEPWARE SCANNER

Deepware Scanner is a tool for deepfake detection developed by the Bosnian cybersecurity company Deepware. The tool is accessible online for free[9] and, despite being offered by a private company, the code for its command line version is publicly available[10]. According to the company claims, with this move, they want to boost collaboration on the deepfake detection task and commit to sharing further advancements with the community. Scanner focuses on human face manipulations, as the majority of existing deepfake detection tools and algorithms.

Feature-wise, Deepware Scanner is a simple interface with a search bar for users to fill in video links or upload them from their local storage. Only YouTube, Facebook, or Twitter sources are currently supported. Once users initiate a scan, the video is first uploaded, with a delay that depends on the duration and size of the video, and then processed. The outcome of this processing is a classification (deepfake detected / undetected) paired with the prediction scores of a few algorithms. In addition, the interface allows replaying the video file with the identified faces marked inside it and presents the video and audio metadata of the media file. Two interesting features are that users can request expert review of the media, offered by the company, and request the takedown of the media from the original platform.

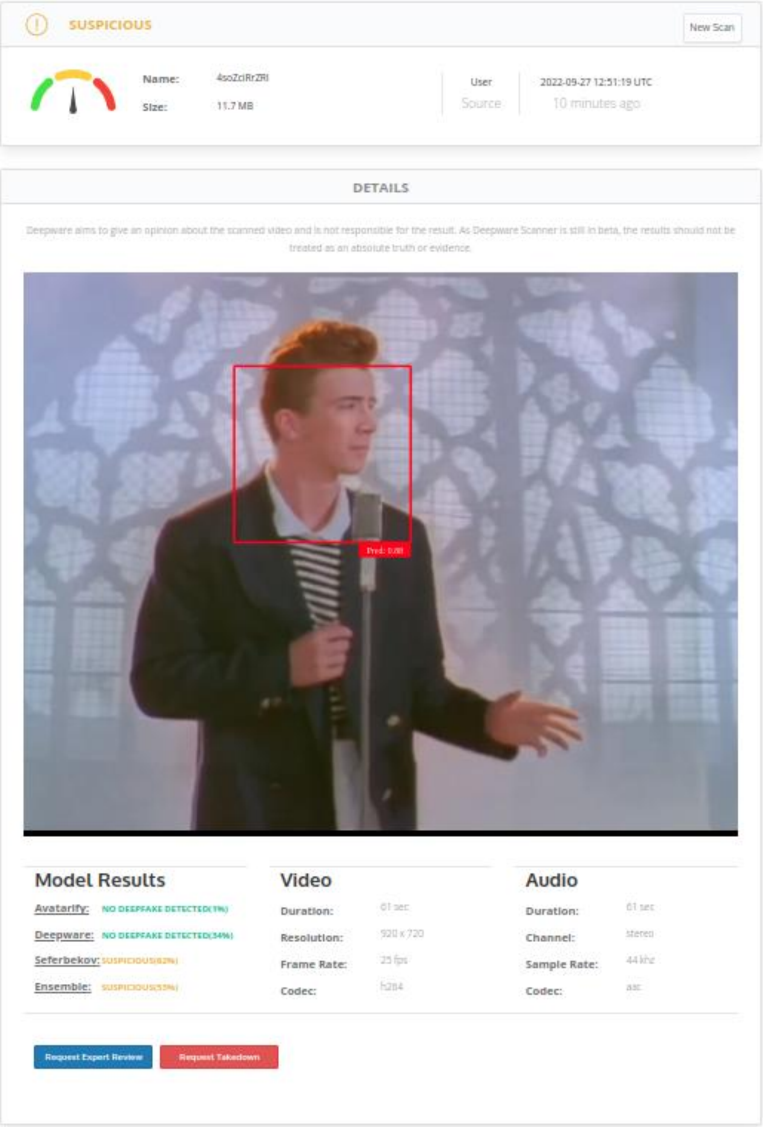
[7] https://www.truly.media/truly-media-now-exploited-in-ai4media-an-h2020-project-advancing-ai-solutions-for-the-media-industry/

[8] https://digger-project.com/

[9] https://scanner.deepware.ai/

[10] https://github.com/deepware/deepfake-scanner

From a technical perspective, the processing is taking place on a frame-by-frame basis, i.e., ignoring temporal and audio information and includes the following steps: i) detect and extract faces from each frame, ii) classify faces with the AI algorithm, iii) cluster faces to identify persistent faces and ignore outliers, iv) calculate a single score per identity, and v) calculate a single score per video. With respect to the implemented detection algorithms, a pure deep learning approach is used with a classification head on top of the EfficientNet B7 convolutional network. This approach is similar to the winning solution of the DFDC challenge by Seferbekov. In fact, Scanner provides the results for both Deepware's in-house solution and the best model by Seferbekov's ensemble, as well as their combination.



Name: 4soZciRrZRI
Size: 11.7 MB
User
Source
2022-09-27 12:51:19 UTC
10 minutes ago

DETAILS

Deepware aims to give an opinion about the scanned video and is not responsible for the result. As Deepware Scanner is still in beta, the results should not be treated as an absolute truth or evidence.

Model Results

Avatarify: NO DEEPFAKE DETECTED(1%)
Deepware: NO DEEPFAKE DETECTED(34%)
Seferbekov: SUSPICIOUS(82%)
Ensemble: SUSPICIOUS(53%)

Video

Duration: 61 sec
Resolution: 920 x 720
Frame Rate: 25 fps
Codec: h264

Audio

Duration: 61 sec
Channel: stereo
Sample Rate: 44 khz
Codec: aac

FIGURE 2: THE RESULT SCREEN OF DEEPWARE SCANNER (TAKEN FROM [11])

[11] https://scanner.deepware.ai/

4.2.2 DEEPFAKE-O-METER

DeepFake-o-meter[12] is a deepfake detection service developed and hosted by the Media Forensics Lab of the University of Buffalo. The service is free to use and open source[13], and offers access to 12 detection algorithms selected from the literature of deepfake detection. It is clearly the output of academic work as more emphasis was placed on the diversity of the algorithms compared instead of the user interface design. While the service was designed with extensibility in mind according to the creators' words, at the time of writing its Github code has not been updated since October 2020 and the link to instructions on how to integrate a new algorithm is broken.

Regarding its features, DeepFake-o-meter has a very simple interface where users can upload videos from a URL link or local storage and select the algorithms to be run. These steps do not require registration but to receive the processing results, the user must provide a valid email address. The results are then returned in a report form. Interestingly, DeepFake-o-meter returns not only a single classification of a video being a deepfake but a comprehensive graph of the detection probability for each identified face for each frame. This level of detail is clearly intended for researchers and can also be seen as an explainability feature.

From a technical perspective, the architecture of DeepFake-o-Meter is shown in Figure 3. The DeepFake-o-meter developers designed a uniform pipeline which could accommodate the pre-processing steps and particularities of each algorithm and could in theory remain extensible for future additions. Each algorithm is implemented as a separate Docker container. There is clearly an emphasis on variety: the majority of algorithms are still based on deep learning methods but some of them stray beyond the dominant convolutional neural network (CNN) paradigm, e.g., using capsule networks, and others target handcrafted features such as visual artefacts at the face's features and the frame's spectrum. These algorithms are not combined in a unique score.

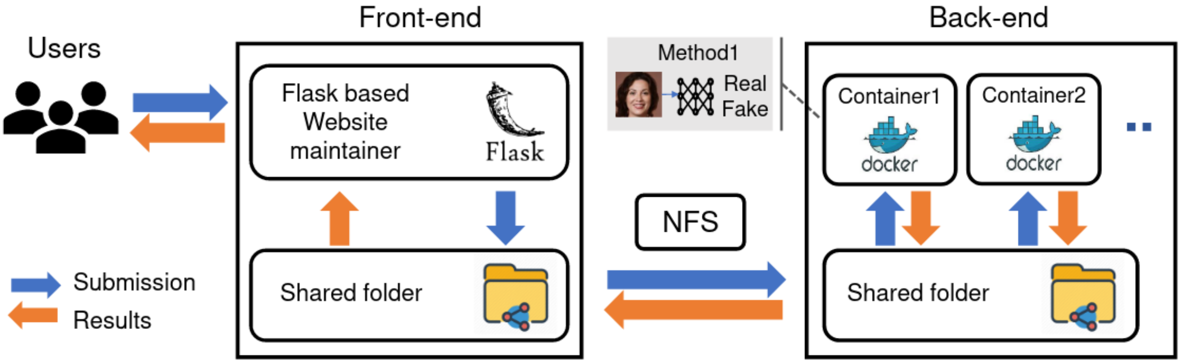


FIGURE 3: THE ARCHITECTURE OF DEEPFAKE-O-METER[14](TAKEN FROM [5])

[12] https://zinc.cse.buffalo.edu/ubmdfl/deep-o-meter/
[13] https://github.com/yuezunli/deepfake-o-meter
[14] NFS refers to Network File System protocol.

Funded by the Horizon 2020 Framework Programme of the European Union

4.3 SCIENTIFIC STATE-OF-THE-ART

In this section, we overview the most frequently used datasets in the deepfake literature, the main detection techniques, as well as approaches towards generalisation and explainability.

4.3.1 DEEPFAKE DATASETS

Currently, deepfakes are easy to find online in video platforms like YouTube or generate with publicly available software like FaceApp[15] and Reface[16]. To have a common reference for testing, the research community has created a number of standard datasets that are frequently used in the literature. These datasets contain real and fake videos, and can be easily converted to image datasets by sampling frames.

According to [6], deepfake datasets can be categorised to three generations, as shown in Table 1. Each generation improves on the previous by increasing the number of frames and videos by an order of magnitude, while the 3rd generation also increased the number of participating humans beyond 100, in an effort to avoid overfitting to a few human identities. Additionally, in the 3rd generation the human participants granted consent and received payment for their participation in the video datasets, which did not occur in the previous generations (with the exception of Google's DFD). Ref. [6] provides further information about these datasets like the number of actors and perturbation methods.

Recently, datasets have emerged that try to increase the diversity of fake content and better represent deepfakes "in the wild". These are OpenForensics [7], WildDeepFake [8], and ForgeryNet [9]. OpenForensics introduces multiple faces to address multi-face forgery detection. WildDeepfake contains a small number of deepfake videos found online. Finally, ForgeryNet is currently the largest dataset of deepfakes both in terms of numbers and perturbation methods.

TABLE 1: DEEPFAKE DATASETS[17]

1st generation	2nd generation	3rd generation	Other
UADFV (98 / 49)	Google DFD (3,000 / 3,000)	DeeperForensics1.0 (60,000 / 1,000)	OpenForensics (115,325 / 70,325)
DF-TIMIT (960 / 640)	DFDC Preview (5,244 / 5,244)	DFDC (128,154 / 104,500)	WildDeepfake (707 / 707)
FF++ DF (5,000 / 4,000)	Celeb-DF (6,229 / 5,639)		ForgeryNet (221,247 / 121,617)

4.3.2 DEEPFAKE DETECTION

What does a deepfake detection algorithm detect? The answer is not as straightforward compared with standard object detection tasks, which detect patterns of pixels resembling the object of interest. In contrast, a deepfake image can represent any scene and is an assortment of pixels seemingly like any other image. In fact, the image is fake due to its generation method,

[15] https://play.google.com/store/apps/details?id=io.faceapp

[16] https://play.google.com/store/apps/details?id=video.reface.app&hl=en&gl=US

[17] The numbers in the parentheses denote the total number of videos and the number of fake videos, in this order.

Funded by the Horizon 2020 Framework Programme of the European Union

which is what the detection algorithm tries to detect. This is done through tell-tale artefacts left in the image's pixels by the generation algorithm, which would not appear if the image had been produced with a camera. This implies two things. First, different types of fake media like traditionally manipulated media and CGI graphics require different algorithms [10]. Second, generation algorithms can conceivably be so competent in the future so as to leave no detectable artefacts. Interestingly, instead of detecting fake images, the opposite approach is also productive and has been explored in the literature: detect real images by artefacts left by cameras, specifically, the photo response non-uniformity (PRNU) [11].

The next question is what kind of features can be used to detect the said artefacts. The features are mathematical descriptions of images that can be used as inputs to an ML algorithm. Low-level features are based on pixels and their surrounding areas and can capture artefacts that may not be visible to the human observer. These features however are easily degraded with simple manipulations such as compression and adversarial noise. High-level features in contrast identify larger structures, which may make intuitive sense to humans (e.g., mismatched eye colours) and be more robust to perturbations at the pixel level. These features however may not be discriminative enough to detect deepfakes, especially with improved generation algorithms.

4.3.3.1 Handcrafted features

Initial approaches in the literature use handcrafted features, borrowing successful techniques from computer vision. Upon extraction, these features are typically used with traditional ML algorithms such as *logistic regression*, *support vector machines* (*SVMs*), *fully-connected neural networks*, and *random forests*. Examples are [12]:

- Local features, captured by descriptors like SIFT and SURF vectors.
- Image spectra, identifying artefacts in the frequency domain.
- Visual artefacts on facial landmarks such as mismatched head poses and eye colours, distorted teeth, nose tips, and face borders.
- Biological signals such as irregular eye blinking and pulse rate (for videos).
- Media stream descriptors from the MPEG standard (for videos).

Handcrafted features are based on our human insight of what may go wrong with fake content. They thus target specific artefacts, usually high-level, that are explainable to humans and computationally lightweight. On the other hand, they are not optimal in terms of their discriminative ability and are easily fooled by improved generation methods.

4.3.3.2 Deep learning features

Subsequent approaches have turned towards neural networks that operate on pixels and learn optimal features given appropriately labelled datasets (*representation learning*). With respect to images, the most popular approaches use CNNs to extract feature vectors, followed by a classification head for the final decision. Models moved from shallow architectures to deeper ones like *XceptionNet*, *ResNet*, and *VGG,* which showed better accuracy [1]. Deep networks combine the best of low and high level features: they operate on pixels but subsequent layers find higher level details. On the other hand, they are prone to overfitting: while they do display > 90% accuracy on train-test splits of given datasets, their performance is degraded on unseen datasets. The standard CNN pipeline has recently been extended to *capsule networks* [13], which organise CNN layers in multiple capsules and route the input dynamically through them, and to *multi-task learning* [14], which learns additional tasks to detection, e.g., finding the manipulated bits inside the image.

With respect to videos, the detection is more involved due to the richer information that they contain. This makes detection more expensive, explaining why most detection algorithms work

Funded by the Horizon 2020
Framework Programme of the European Union

on short video segments, e.g. 10 minutes. On the positive side, videos are harder to forge due to their complexity, and offer more opportunities for inconsistencies to detect than images. In general, there are three approaches to detect deepfake videos:

Frame-based detection: processes a subset of frames from the video, sampled randomly or with a more sophisticated approach. The intuition is that faces change little between consecutive frames, hence a representative selection should suffice. Frame-based detection is also attractive because it is lightweight. For the processing, the faces from different frames must first be connected, typically with a face clustering algorithm which distinguishes different faces and false positives from the face detection algorithm [15]. The individual faces are then evaluated with an image detection algorithm, typically CNN-based, and the final result is derived from the aggregation of all frames and faces.

Spatiotemporal detection: processes an entire succession of frames. The intuition is that consecutive frames can reveal inconsistencies in the temporal domain that can increase the accuracy of detection, at an added computational cost. For the processing, faces are first detected within consecutive frames and possibly aligned [16]. The most popular approaches then encode each frame separately and then pass the encodings through a recurrent neural network (RNN) network like the *gated recurrent unit* (*GRU)* and the *long-short term memory* (*LSTM)* networks [1]. Other approaches include *3D CNNs* [17], which operate directly on the video sequence, *optical-flow models* [18], which operate on the displacement field of pixels, and *transformer-based models* [19], which picks frames through attention.

Audiovisual detection: processes the audio stream along the video. As with spatiotemporal detection, the intuition is that additional information will reveal more inconsistencies, this time in the audio domain, and lead to higher detection accuracy. This approach is the least explored in the literature due to its complexity and the limited number of audiovisual generation. For the processing, the audio and the image streams can be processed i) jointly, ii) separately in the beginning and fused mid-point (*early fusion*), iii) separately and fused at the end (*late fusion*) [20]. Joint processing can achieve the highest accuracy but at increased computational cost.

4.3.3 GENERALISATION OF DETECTION

While deepfake detection algorithms have good accuracy on the datasets on which they have been trained, their accuracy drops considerably on different generation algorithms and unseen data, as well as data which have undergone simple processing like compression [21, 22]. Such conditions are typical for media encountered online, therefore impressive detection results of published algorithms should be interpreted cautiously. For algorithms that use deep features, this is due to *overfitting* on the training set. A standard way to increase robustness to simple processing is by perturbing the images in the training set with various transformations such as compression, cropping, adding noise, and rotations. This indirectly helps generalisation as the algorithm is forced to learn high-level discriminative features instead of relying on the easily corrupted pixels. Some high-level features have also been found to generalise well, such as inconsistent lip movements [23] and blending artefacts in face-swaps due to the blending of a target face to a source image [24]. Nevertheless, there is no guarantee that these methods will not be superseded in the future.

Detection algorithms can be evaluated for their generalisation abilities by testing them in the datasets of Section 4.3.1. In general, *zero-shot performance* (evaluating on an unseen dataset) is bad, hence *retraining* is needed to generalise to the new datasets. Simply retraining however on a new dataset does not work as intended due to *catastrophic forgetting,* i.e., the network forgets the correct operation on previous datasets. Catastrophic forgetting is a classic problem of *incremental learning* (also called *continuous* or *lifelong learning*) [25], which can be addressed in several ways [26]. At one extreme, the network can be retrained from scratch using all the datasets; this may yield the best performance but is computationally expensive and the older datasets may not be available in practice. Rehearsal approaches relax this issue by maintaining a few representative examples of the previous datasets and including them in the retraining routine. To avoid storing old examples, the model can be retrained on only new data but penalising deviations from the responses of the old model to the new data. This approach is popular but may fail to optimise effectively due to the penalty constraint and is prone to forgetting after many iterations of learning [27].

Ideally, retraining should be lightweight so that the model can adapt to new tasks or datasets with only a few samples (*few-shot learning*). This is achieved by *transfer learning*, which tries to transfer knowledge from past to new tasks with light fine-tuning [28]. A common approach is to use a pre-trained network which has learnt general features of the input dataset, and fine-tune it with a classification head. The approaches of section 4.3.3.2 that use deep neural networks fall into this category. Freezing the lower layers of the model during retraining also retains the low level features of model and reduces the computational load at a possible reduction in accuracy. Finally, autoencoders can be used to learn representations that can both solve a task and recreate the original data. Autoencoders have been applied successfully in deepfake detection, showing good generalisation ability [21] and inspiring further work [14].

A different approach to generalisation is *ensemble learning*, which combines multiple algorithms so that the ensemble is more accurate and robust than each individual component [29]. An ensemble architecture comprises the *base learners* (also called *individual*, *component,* or *level-0 learners*) and a combination layer or *meta-learner* (also called *level-1 learner*) to combine their outputs. There is an important distinction between an ensemble that is trained end-to-end, training both the individual learners and the meta-learner, and an ensemble for which the individual learners are provided as-is and only the meta-learner is trained. In the first case, which is closer to the original idea of ensemble learning, the learners are part of the model design; they are typically simple homogeneous algorithms whose combination intends to correct the inaccuracies of each component. This case is well represented by *random forests* and contains the techniques of *bagging* and *boosting*. In the second case, individual learners are heterogeneous, focusing on different regions of the feature space or even different tasks and the learner's objective is to fuse their outcomes. This case features the *stacking* technique, which offers large freedom on how to design the ensemble. If the individual learners are highly specialised and accurate in their intended feature area, selecting the correct algorithm may be more productive than simple averaging. Stacking ensembles were the dominant approach in the solutions of the DFDC challenge and are considered a promising way to increase the generalisation of deepfake detection [30].

4.3.4 EXPLAINABILITY OF DETECTION

XAI has become a hot research topic recently due to the explosion of AI applications, especially in high-stake use cases like medicine, law, and defence [31]. The key concern of AI is its inherent unpredictability and bias, which is carried from the data on which the AI algorithm is trained. To increase the transparency of the developed models, *model cards* have been proposed, which report the datasets used for training, their characteristics, details about training, benchmark evaluations, and other related information, highlighting the strengths and limitations of the models [32]. While undoubtedly this is a step in the right direction, explainability is a more ambitious goal that seeks to reason about the outcome of an AI algorithm in a way that is understandable by humans. As explainability offers insights for the correct operation of the algorithm, it can also be a valuable tool for the development of AI algorithms.

Funded by the Horizon 2020
Framework Programme of the European Union

Some AI algorithms are inherently explainable. The symbolic algorithms of the 80s, e.g., expert systems, used symbols for inference that were recognisable to humans, e.g., words, but fell out of favour due to their poor scaling. In addition, classical ML models like linear and logistic regression, K-nearest neighbours, Bayesian methods, and, especially, decision trees can be interpreted by technically-minded users, as long as they don't grow in complexity. These algorithms have been characterized *interpretable* / *transparent / white-box,* contrasting with *opaque* / *black-box* models, chiefly represented by neural networks, which defy simple explanation. Another terminology which is gaining traction is ***post-hoc*** and ***ante-hoc explainable*** models which refers to models that require and don't require explanation respectively[18]. It is worth noting here that even large ante-hoc models may become unwieldy for humans and require explanation.

For post-hoc models, explainability can be classified at a high level as:

- **model-specific**, relating to a specific model type.
- **model-agnostic**, applying to any type of model, focusing on the relations and correlations of input features and output.

In terms of scale, it can also be:

- **global**, explaining the whole model.
- **local**, explaining a single prediction.

Many further taxonomies have appeared in the literature. From a result perspective, i.e., what explainability ultimately offers, there are three main approaches [33]:

Feature importance, which tries to highlight the most important features that contribute to an output, and it is the most popular approach. For a traditional ML task with handcrafted features, e.g., health indicators, this can be a contribution percentage for each feature. For an image classification task, this may be a *saliency map* or *heatmap* indicating which pixels contributed the most to the activation of the network's neurons. Feature importance is typically evaluated by examining the algorithm's behaviour at small deviations of the current features. This can be done with *perturbations* or, more conveniently for neural networks, *gradients.*

Surrogate models, which tries to simplify a whole model or parts of it with a simpler, preferably interpretable, model. This is similar to the distillation technique, where the output of a complex teacher model is used to train a simpler student model. Simplification can also entail *architecture modification*, e.g., exchanging convolutional layers with max pooling layers. Sparsity can also be achieved through pruning which may make features and neural connections more interpretable. There have also been approaches to combine neural networks with symbolic AI, resulting in neuro-symbolic models.

Examples, which tries to find representative examples of the model's behaviour. In the simplest case, the examples are inputs which maximise the confidence of the neural network's output. Additionally, *counterfactual* and *adversarial* examples have been proposed which find examples close to a given input which produce a different result. These negative examples can be very helpful during training to make the algorithm more robust.

[18] It is worth noting here that the terminology of XAI is diverse and inconsistent due to the rapid progress of the field. The terms transparent and white-box models can also be misleading as a neural network with known weights *is* transparent in the common sense but by no means interpretable. Hence, [33] advocates in favor of *ante-hoc* and *post-hoc* terminology, which we will also follow in the text.

Funded by the Horizon 2020
Framework Programme of the European Union

Focusing on the deepfake detection task, there are two prominent approaches for explainability [34]:

Source attribution, which tries to detect the source of media. In forensics, the cameras, used to produce a genuine image, have been found to insert a tell-tale PRNU sequence, which can be uncovered in the frequency domain and can even help identify the camera's model. Interestingly, GAN-generated images have also been found to contain similar traces, mainly due to the upsampling operations in the convolutional layers. Source attribution can increase the credibility of a deepfake evaluation but its application to videos is much harder due to the extensive pre-processing and post-processing.

Artefact localization, which tries to localise anomalous areas inside an image that are suspected of tampering. This is also useful to identify the forgery type, e.g., a face-swap or an attribution manipulation or a completely synthetic media. In general, localisations based on high-level features (e.g., mismatched eye colour) are more understandable to humans provided that they exist. If they don't exist however, deep learning techniques cannot be explained easily. Heatmaps and saliency maps have been used to reveal areas which have activated the neurons [35] but they are not always informative as shown in the classic example of Figure 4: Saliency maps for object detection where similar pixels are activated in the same image for a correct and an incorrect classification.

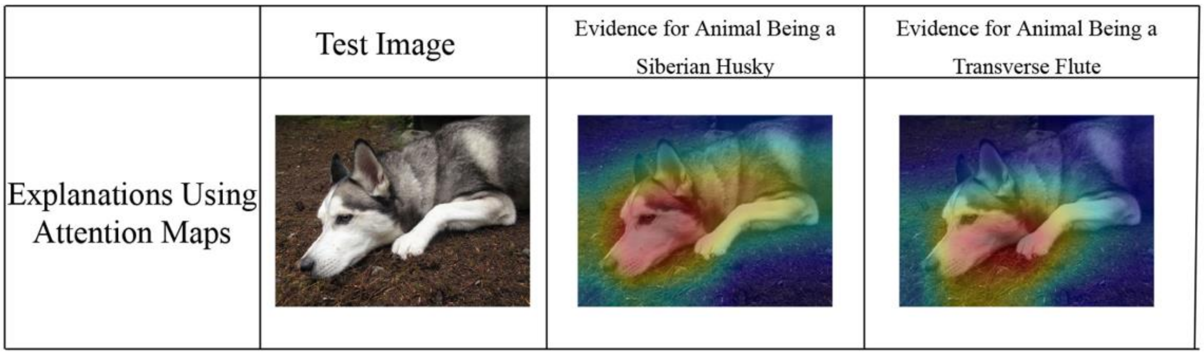

Test Image
Evidence for Animal Being a Siberian Husky
Evidence for Animal Being a Transverse Flute
Explanations Using Attention Maps

FIGURE 4: SALIENCY MAPS FOR OBJECT DETECTION (TAKEN FROM [36])

Funded by the Horizon 2020
Framework Programme of the European Union

5 SOLUTION DESIGN

In this section, we describe the design of the DeepFakeChain platform from the *user perspective*. In Section 5.1 of TR1.1, we described different user roles and privileges of DeepFakeChain, as well as different channel types for the organisation of content. However, in order to focus on the scientific aspects of our project, we have simplified our design considerably to include a single type of user and channel. Note that the described design is meant as a proof-of-concept demonstration application with the purpose of easily connecting end users (journalists, fact-checkers) with the results of the deepfake detection algorithms such as the ones that have been described in the previous state-of-the-art review section.

5.1 CONCEPTS

The design of DeepFakeChain contains the following concepts:

- **User profiles,** containing details about registered users such as name, description, affiliations, contact information, etc. All users receive a user profile upon registration.
- **Media,** uploaded from online sources or local storage. They are stored to the platform in conventional storage, to be available to the platform in case of future take-downs. In addition, proofs of authenticity are stored in the blockchain network.
- **Channels,** workspaces where users from potentially different organisations can collaborate on a particular topic, e.g., the Ukrainian war. Users can search for channels, join and associate uploaded media with them.
- **Annotations**, notes by users referring to a media's authenticity. They can pinpoint to a specific spatiotemporal window and also contain a judgement, i.e., a concrete evaluation of the media being a deepfake.
- **Comments**, general notes by users about the media. They are a form of asynchronous communication among DeepFakeChain's users.

5.2 WORKFLOWS

Based on the above concepts, we foresee the following workflows:

- **Register workflow:** users authenticate and create profiles in the DeepFakeChain platform.
- **Collaborate workflow:** users create, join, and manage channels.
- **Organise workflow:** users assign media to channels.
- **Search workflow**: users search channels and media inside the DeepFakeChain platform. DeepFakeChain will provide *reverse media search* to identify similar and near-duplicate media, based on an available solution by CERTH.
- **Verify workflow**: users annotate media and evaluate them as genuine or deepfakes, as well as request verification by the algorithms of DeepFakeChain.

Funded by the Horizon 2020
Framework Programme of the European Union

5.3 ARCHITECTURE

The architecture of DeepFakeChain has already been described in Section 5.2 of TR1.1 and is repeated in Figure 5 for convenience. Briefly, DeepFakeChain communicates with its users through a front-end server (User Layer), which offers access to AI services (AI Layer) and the platform's data (user profiles, media files, and annotations) stored in conventional storage (Data and Content Layer). The AI layer in particular will be extensible so that new services can be easily incorporated to the platform. The platform will also communicate with an external blockchain network (blockchain layer) in order to store proofs of authenticity of its data. The users will be able to verify these proofs through an independent interface, which will ensure the trustworthiness of our platform. Smart contracts may also be used to implement the trust and reputation scheme of the platform.

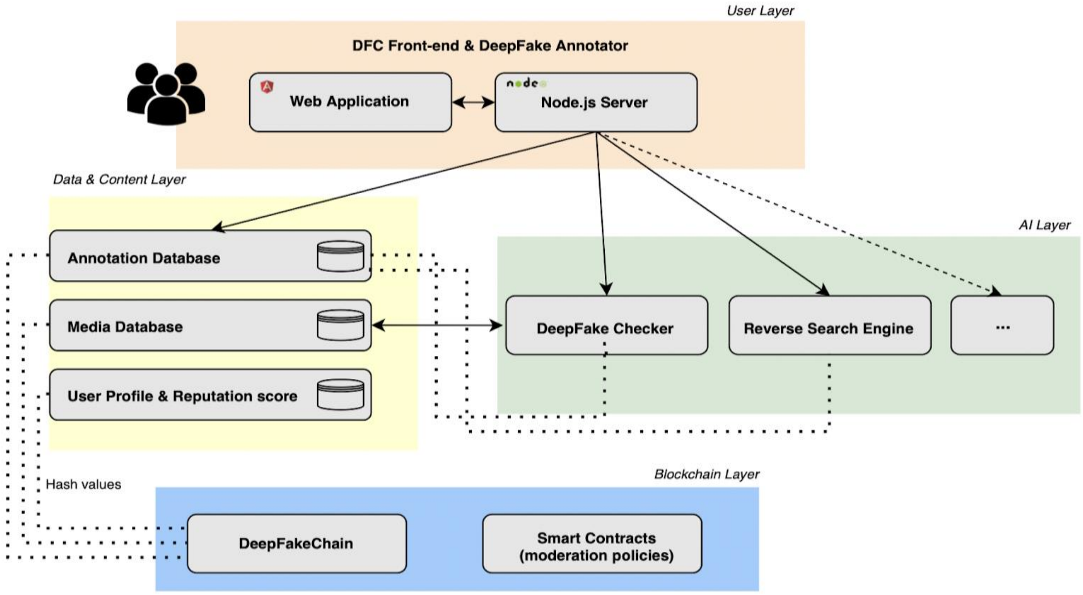


FIGURE 5: THE ARCHITECTURE OF DEEPFAKECHAIN

5.4 AVAILABLE IMPLEMENTATIONS

CERTH and Zelus have developed components and tools that are relevant to DeepFakeChain and can form an initial technical basis for the proof-of-concept web-based application. Specifically, CERTH maintains a prototype asset management system for media uploading from online sources (e.g., YouTube) and annotation, a reverse image search service, as well as a deepfake detection service based on an ensemble of trained models for deepfake detection [38]. Zelus maintains a SmartViz toolkit for creating sophisticated and informative dashboards for the DeepFakeChain UI. The UI design is particularly important in this project, considering the highly technical nature of the implemented algorithms and the need for explainability features to increase user trust in the algorithmic outputs.

CERTH's platform can form a blueprint for the development of uploading media, downloading them from external sources, organising them in channels, and annotating them. The already implemented deepfake detection service can form a base for the development of more state-of-the-art algorithms with the desired features (accuracy, generalisation, explainability). The existing reverse search engine will also be integrated in the platform. The SmartViz toolkit by Zelus will help implement effective panels and visualisations. Finally, the Alastria network – access to which is offered by Trublo - will be used for implementing the blockchain layer.

Funded by the Horizon 2020 Framework Programme of the European Union

5.4.1 CERTH'S PROTOTYPE FOR MEDIA VERIFICATION

CERTH has implemented a prototype application, called DFDLab, for media uploading, annotation, browsing/search and automatic processing. Users authenticate in the platform via a simple email-password registration and can upload videos through URLs or local storage, which is standard among similar tools. Supported videos include YouTube, Facebook, Twitter, TikTok and direct URLs. Once uploaded, the videos belong to the user's personal library and can be marked as public or private, depending on whether they are searchable by other users. For each video, users can provide a title, a description, tags, comments, and spatiotemporal annotations (tied to a specific area in the video for a specific duration).

The application provides access to two AI-based services, which will form the starting point for DeepFakeChain research activities. First, videos can be reverse searched to find similar media in the DFDLab index. This service is based on work by CERTH on content-based video indexing and near duplicate retrieval [37]. Second, the uploaded videos can be processed with a deepfake detection algorithm to estimate their probability of being a deepfake. This algorithm is also based on previous work by CERTH and supports an ensemble of five deep learning models [38].

5.4.2 ZELUS' SMARTVIZ TOOLKIT

Zelus' SmartViz is a data visualisation toolkit which facilitates collecting data from multiple sources and presenting them in attractive visual interfaces, tailored to the end user needs. These interfaces can be predefined and/or user-defined, built from a large selection of widgets, and supporting a wide variety of visualisations, ranging from simple charts and graphs to complex timeline representations and geospatial depictions. Other features include advanced filtering mechanisms, interconnected visualisations, and data comparison panes to enable multiple data presentations and exploratory analysis.

Internally, SmartViz contains three components: the data intake adapters, a middleware, and a frontend, as shown in Figure 6. The data adapters are responsible for connecting with different data sources, with the current implementation supporting RESTful APIs and real-time data feeds like Kafka topics and MQTT. Other adapters can easily be added to support additional data sources. The middleware receives information from the adapters and transforms it into internal representation which are useful for visualisation. It also hosts configuration options for user authentication and authorisation, dashboard sharing options, interface layout options, and other parameters. The frontend finally serves as a web application that is directly accessible by the end users.

In DeepFakeChain, the SmartViz toolkit can be used in the design of the user interface, displaying the uploaded media and the related metadata (annotations, comments, tags) in an attractive way. Its interactive features may also enable panels for exploratory analysis, for general trends of the uploaded deepfakes, as well as the progress of their evaluation. Most importantly, innovative visualisations can be developed that explain the consensus of human and machine judgement over a media evaluation, based on the methods proposed by the DeepFakeChain project.

Funded by the Horizon 2020
Framework Programme of the European Union

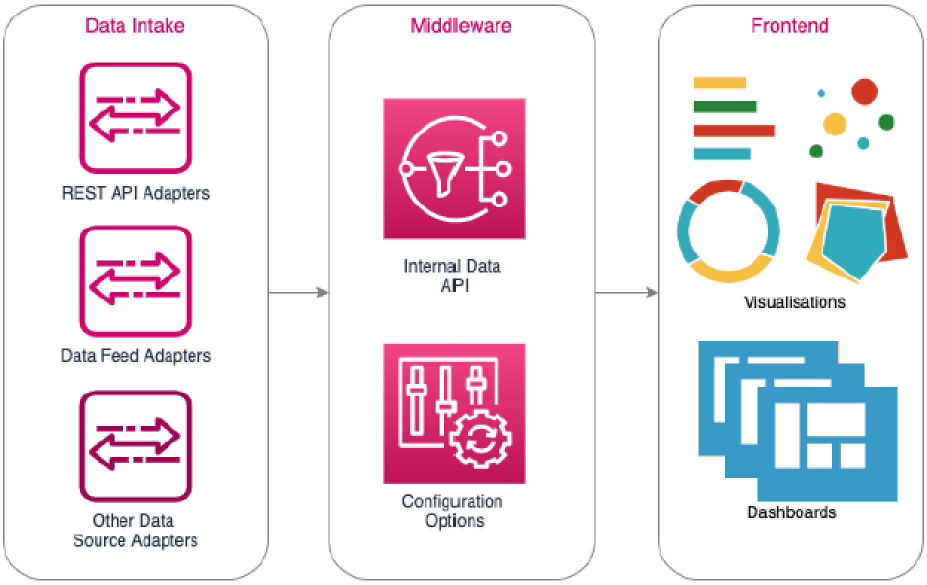


FIGURE 6: THE ARCHITECTURE OF SMARTVIZ

Funded by the Horizon 2020
Framework Programme of the European Union

6 FUNCTIONAL REQUIREMENTS

Based on the solution design of Section 5, we describe the functional requirements of the DeepFakeChain platform. Note that these are *platform-application requirements*, describing what the platform can do based on the user actions of Section 5. The requirements contain features, which are not important for the DeepFakeChain project, but have value for a future evolution of DeepFakeChain towards a marketable product. The requirements are divided in terms of priority in the categories of Table 2: Functional Requirement prioritiesTable 2.

TABLE 2: FUNCTIONAL REQUIREMENT PRIORITIES

Value	Rating	Description
1	Critical	This requirement is critical to the success of the project. The project will not be possible without this requirement.
2	High	This requirement is high priority, but the project can be implemented at a bare minimum without this requirement.
3	Medium	This requirement is somewhat important, as it provides some value but the project can proceed without it.
4	Low	This is a low priority requirement, or a “nice to have” feature, if time and cost allow it.
5	Future	This requirement is out of scope for this project, and has been included here for a possible future release.

In the following tables, we organise the functional requirements according to the workflows defined in Section 5.2, including a table at the end for our algorithmic layer.

TABLE 3: FUNCTIONAL REQUIREMENTS FOR THE REGISTER WORKFLOW

ID	Requirement	Priority
REG 1	The platform shall allow users to create user profiles upon registration.	1
REG 2	The platform shall allow users to edit their user profiles.	1
REG 3	The platform shall allow users to delete their user profiles by unregistering from the platform.	1
REG 4	The platform shall allow users to configure the visibility of their user profiles.	4
REG 5	The platform shall allow users to verify their identity.	4

Funded by the Horizon 2020
Framework Programme of the European Union

TABLE 4: FUNCTIONAL REQUIREMENTS FOR THE COLLABORATE WORKFLOW

ID	Requirement	Priority
COL 1	The platform shall allow users to create channels for collaborating on specific topics.	2
COL 2	The platform shall allow users to edit the descriptions of channels.	2
COL 3	The platform shall allow users to join channels.	2
COL 4	The platform shall allow users to invite other users to channels.	4
COL 5	The platform shall allow users to delete channels.	2
COL 6	The platform shall allow users to configure the visibility of channels.	4
COL 7	The platform shall allow users to communicate directly with each other via direct messages.	5
COL 8	The platform shall allow users to communicate in groups via group messages.	5

TABLE 5: FUNCTIONAL REQUIREMENTS FOR THE ORGANISE WORKFLOW

ID	Requirement	Priority
ORG 1	The platform shall allow users to upload media files.	1
ORG 2	The platform shall allow users to delete media that they have uploaded from the platform.	1
ORG 3	The platform shall allow users to organise media inside collections and folders	3
ORG 4	The platform shall allow users to add their uploaded media to channels that they have joined.	2
ORG 5	The platform shall allow users to remove their uploaded media from channels that they have joined.	2
ORG 6	The platform shall allow users to configure the visibility of their uploaded media.	4

Funded by the Horizon 2020
Framework Programme of the European Union

TABLE 6: FUNCTIONAL REQUIREMENTS FOR THE SEARCH WORKFLOW

ID	Requirement	Priority
SEA 1	The platform shall allow users to search the uploaded media, subject to their visibility constraints.	1
SEA 2	The platform shall allow users to search for similar and near-duplicate media subject to their visibility constraints.	2
SEA 3	The platform shall allow users to search user profiles, subject to their visibility constraints.	4
SEA 4	The platform shall allow users to search annotations and comments subject to their visibility constraints.	4

TABLE 7: FUNCTIONAL REQUIREMENTS FOR THE VERIFY WORKFLOW

ID	Requirement	Priority
VER 1	The platform shall allow users to annotate and comment on media.	1
VER 2	The platform shall allow users to edit their annotations and comments.	2
VER 3	The platform shall allow users to delete their annotations and comments.	2
VER 4	The platform shall allow users to configure the visibility of their annotations and comments.	4
VER 5	The platform shall allow users to evaluate media as deepfakes.	1
VER 6	The platform shall allow users to verify that the uploaded data is genuine.	2

Funded by the Horizon 2020
Framework Programme of the European Union

TABLE 8: FUNCTIONAL REQUIREMENTS FOR THE AI LAYER

ID	Requirement	Priority
ALG 1	The platform shall allow users to view the available algorithms for deepfake detection.	1
ALG 2	The platform shall allow users to check media with selected deepfake detection algorithms.	1
ALG 3	The platform shall allow users to request a consensus opingion of the existing algorithmic and human judgements.	1

Funded by the Horizon 2020
Framework Programme of the European Union

7 NON-FUNCTIONAL REQUIREMENTS

In this section, we present the non-functional requirements of the AI layer, corresponding to the functional requirements of Table 8, which are the most relevant to the scientific focus of our project.

TABLE 9: NON-FUNCTIONAL REQUIREMENTS FOR THE AI LAYER

ID	**Requirement**
NFR 1	The AI layer must achieve very high accuracy (>95%) on datasets seen during training.
NFR 2	The AI layer must achieve high generalisation, i.e. sufficiently high accuracy (>70%) on datasets not seen during training.
NFR 3	The AI layer must be able to propose optimised combinations of algorithms for a given video.
NFR 4	The AI layer must present to the users the capabilities and limitations of each algorithm using a model card.
NFR 5	The AI layer must provide a panel to user with accessible and thorough explanations on the taken decision.
NFR 6	The deepfake detection algorithms of the AI layer should be able to run on consumer-level GPUs.

Funded by the Horizon 2020
Framework Programme of the European Union

8 RISKS AND MITIGATION

We conclude the deliverable by summarising the risks faced by the DeepFakeChain project and our planned mitigation strategies. The risks are separated in the *scientific*, *business*, and *resource* categories. We stress that due to the scientific focus of the project, the scientific risks are the most relevant, at least during Phase 1.

8.1 SCIENTIFIC RISKS

TABLE 10: SCIENTIFIC RISK OF GENERALISATION

Risk	Poor generalisability of algorithms.
Description	DeepFakeChain may not succeed in producing deepfake detection algorithms that work well in unseen real-life videos.
Reasons	The inherent difficulty of the deepfake detection task. The constant appearance of new deepfakes. The lack of diverse training resources. The weakness of the employed methods.
Mitigation	CERTH has expertise on deepfake detection and already implemented an ensemble of deepfake detection models. This provides a good start for experimentation. The literature of deepfake detection will be thoroughly reviewed to identify the most promising approaches to experiment with.

TABLE 11: SCIENTIFIC RISK OF EXPLAINABILITY

Risk	Poor explainability of the platform’s decisions.
Description	DeepFakeChain may not succeed in producing good explanations for the decisions reached by both the human users and its algorithms.
Reasons	The inherent lack of interpretability of artificial neural networks and deep learning architectures, which are typically used in deepfake detection algorithms. The weakness of the employed methods.
Mitigation	At a minimum, DeepFakeChain will implement established methods of explainability such as heatmaps, to offer basic insight on the algorithmic decisions. Other promising approaches will be investigated based on a thorough literature review. The human judgement will also offer opportunities for explainability which are not possible with algorithmic detection alone.

Funded by the Horizon 2020
Framework Programme of the European Union

TABLE 12: SCIENTIFIC RISK OF MAN-MACHINE COLLABORATION

Risk	Poor combination of human and machine judgements.
Description	DeepFakeChain may not succeed in combining the judgements of algorithms and humans effectively.
Reasons	The difficulty and "black magic" aspect of designing stacking ensembles. The asynchronous speeds of human and machine operations. The contextual nature of human annotations compared with the binary output of classification algorithms.
Mitigation	DeepFakeChain will review automatic approaches for ensemble design that scale better than manual tuning. It will provide a baseline judgement based on human consensus while ensuring that the automatic judgement will be at least as accurate as the most accurate model in the ensemble.

8.2 BUSINESS RISKS

TABLE 13: BUSINESS RISK OF DEMAND

Risk	Lack of demand for a collaborative deepfake detection platform.
Description	DeepFakeChain may not succeed in securing demand and curving a market share for collaborative deepfake detection.
Reasons	The existence of established services like TrulyMedia and Check. The narrow focus of deepfake detection inside verification. The inhibition of media companies to collaborate and invest in such software, and the poor financial state of the news sector. Ineffective promotion of the platform.
Mitigation	DeepFakeChain will aim at developing and offering highly specialised services that are not available from other platforms and promote them effectively. This will open additional business avenues of merging with existing platforms. In addition, owing to the extensibility and good design of the platform, it could be repurposed easily towards more general collaborative tasks that have a much higher demand. One such example is collaborative moderation of content in large social media platforms.

TABLE 14: BUSINESS RISK OF CAPITAL

Risk	Lack of capital for the evolution of the platform.
Description	DeepFakeChain may not succeed in securing funds for its successful transition to the market beyond the end of the TruBlo project.
Reasons	The scientific focus of DeepFakeChain. The potential lack of funding for follow-up projects.
Mitigation	DeepFakeChain will search for additional funding from public and private sources. For the latter, it will focus on developing highly innovative algorithms that will attract the interest of investors. During Phase 2, a comprehensive business plan will be developed that will explore opportunities for the economic sustainability of the platform.

Funded by the Horizon 2020
Framework Programme of the European Union

8.3 RESOURCE RISKS

TABLE 15: RESOURCE RISK OF COMPUTATIONAL RESOURCES

Risk	Lack of computation resources for the development of deepfake detection
Description	DeepFakeChain may not have sufficient computational resources to develop innovative AI algorithms.
Reasons	The high cost of AI competent hardware (e.g., GPUs). The complexity of DL algorithms. The computational burden of video processing.
Mitigation	CERTH already has competent hardware that has enabled the development of a state-of-the-art ensemble of deepfake detection models. The project will focus on developing lightweight and efficient algorithms, which is also currently a very pertinent sustainability demand.

TABLE 16: RESOURCE RISK OF TRAINING DATA

Risk	Lack of diverse training sets for deepfake detection
Description	DeepFakeChain may not achieve good algorithmic performance due to the lack of sufficiently large and diverse datasets.
Reasons	The cost of creating large and diverse datasets, especially on video content that require permissions from the actors. The focus of the deepfake literature on face manipulation.
Mitigation	DeepFakeChain will demonstrate the competence of its algorithms on the most updated and established datasets. Due to its operation, it will also become a source of real-life videos to test and tweak these algorithms. In the long-term, the DeepFakeChain platform will become a good source of training data as it naturally crowdsources human annotations from expert users based on its everyday usage.

TABLE 17: RESOURCE RISK OF STORAGE

Risk	Lack of resources for distributed storage
Description	DeepFakeChain may not find sufficient resources for distributed storage.
Reasons	Distributed storage requires the deployment of a network. The potential limitations of specific designs of distributed storage.
Mitigation	DeepFakeChain will make use of Alastria's permissioned network during the project's duration, access to which is offered by TruBlo. The premissioned nature of Alastria ensures that performance will not be sacrificed and a minimum level of trust exist among Alastria's nodes. After the end of the project, DeepFakeChain will evaluate the optimal solution for its blockchain storage taking into account its economic sustainability and business plan. TruBlo's mentorship will be pivotal to evaluate potential solutions.

Funded by the Horizon 2020
Framework Programme of the European Union

9 REFERENCES

[1] M. S. Rana, M. N. Nobi, B. Murali, and A. H. Sung, “Deepfake detection: a systematic literature review,” *IEEE Access*, 2022.

[2] E. Ilkou and M. Koutraki, “Symbolic vs sub-symbolic ai methods: Friends or enemies?” in *CIKM (Workshops)*, 2020.

[3] M. Sharma and M. Kaur, “A review of deepfake technology: an emerging ai threat,” *Soft Computing for Security Applications*, pp. 605–619, 2022.

[4] S. A. Buo, “The emerging threats of deepfake attacks and countermeasures,” *arXiv preprint arXiv:2012.07989*, 2020.

[5] Y. Li, C. Zhang, P. Sun, L. Ke, Y. Ju, H. Qi, and S. Lyu, “Deepfake-o-meter: an open platform for deepfake detection,” in *2021 IEEE Security and Privacy Workshops (SPW)*. IEEE, 2021, pp. 277–281.

[6] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. C. Ferrer, “The deepfake detection challenge (dfdc) dataset,” *arXiv preprint arXiv:2006.07397*, 2020.

[7] T.-N. Le, H. H. Nguyen, J. Yamagishi, and I. Echizen, “Openforensics: Large-scale challenging dataset for multi-face forgery detection and segmentation in-the-wild,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10117–10127.

[8] B. Zi, M. Chang, J. Chen, X. Ma, and Y.-G. Jiang, “Wilddeepfake: A challenging real-world dataset for deepfake detection,” in *Proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 2382–2390.

[9] Y. He, B. Gan, S. Chen, Y. Zhou, G. Yin, L. Song, L. Sheng, J. Shao, and Z. Liu, “Forgerynet: A versatile benchmark for comprehensive forgery analysis,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 4360–4369.

[10] I. Castillo Camacho and K. Wang, “A comprehensive review of deep-learning-based methods for image forensics,” *Journal of Imaging*, vol. 7, no. 4, p. 69, 2021.

[11] O. Gouda, A. Bouridane, M. A. Talib, and Q. Nasir, “Machine learning-based methods in source camera identification: A systematic review,” in *2022 International Conference on Business Analytics for Technology and Security (ICBATS)*. IEEE, 2022, pp. 1–7.

[12] M. Masood, M. Nawaz, K. M. Malik, A. Javed, A. Irtaza, and H. Malik, “Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward,” *Applied Intelligence*, pp. 1–53, 2022.

[13] H. H. Nguyen, J. Yamagishi, and I. Echizen, “Capsule-forensics networks for deepfake detection,” in *Handbook of Digital Face Manipulation and Detection*. Springer, Cham, 2022, pp. 275–301.

[14] H. H. Nguyen, F. Fang, J. Yamagishi, and I. Echizen, “Multi-task learning for detecting and segmenting manipulated facial images and videos,” in *2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS)*. IEEE, 2019, pp. 1–8.

Funded by the Horizon 2020
Framework Programme of the European Union

[15] P. Charitidis, G. Kordopatis-Zilos, S. Papadopoulos, and I. Kompatsiaris, “Investigating the impact of pre-processing and prediction aggregation on the deepfake detection task,” *arXiv preprint arXiv:2006.07084*, 2020.

[16] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan, “Recurrent convolutional strategies for face manipulation detection in videos,” *Interfaces (GUI)*, vol. 3, no. 1, pp. 80–87, 2019.

[17] O. de Lima, S. Franklin, S. Basu, B. Karwoski, and A. George, “Deepfake detection using spatiotemporal convolutional networks,” *arXiv preprint arXiv:2006.14749*, 2020.

[18] I. Amerini, L. Galteri, R. Caldelli, and A. Del Bimbo, “Deepfake video detection through optical flow based cnn,” in *Proceedings of the IEEE/CVF international conference on computer vision workshops*, 2019, pp. 0–0.

[19] S. A. Khan and H. Dai, “Video transformer for deepfake detection with incremental learning,” in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 1821–1828.

[20] Y. Zhou and S.-N. Lim, “Joint audio-visual deepfake detection,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 14800–14809.

[21] D. Cozzolino, J. Thies, A. Rössler, C. Riess, M. Nießner, and L. Verdoliva, “Forensictransfer: Weakly-supervised domain adaptation for forgery detection,” *arXiv preprint arXiv:1812.02510*, 2018.

[22] J. Sabel and F. Johansson, “On the robustness and generalizability of face synthesis detection methods,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 962–971.

[23] A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic, “Lips don’t lie: A generalisable and robust approach to face forgery detection,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 5039–5049.

[24] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, and B. Guo, “Face x-ray for more general face forgery detection,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5001–5010.

[25] R. Ratcliff, “Connectionist models of recognition memory: constraints imposed by learning and forgetting functions.” *Psychological review*, vol. 97, no. 2, p. 285, 1990.

[26] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska *et al.*, “Overcoming catastrophic forgetting in neural networks,” *Proceedings of the national academy of sciences*, vol. 114, no. 13, pp. 3521–3526, 2017.

[27] Z. Li and D. Hoiem, “Learning without forgetting,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 12, pp. 2935–2947, 2017.

[28] K. Weiss, T. M. Khoshgoftaar, and D. Wang, “A survey of transfer learning,” *Journal of Big data*, vol. 3, no. 1, pp. 1–40, 2016.

[29] O. Sagi and L. Rokach, “Ensemble learning: A survey,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, p. e1249, 2018.

Funded by the Horizon 2020
Framework Programme of the European Union

[30] M. S. Rana and A. H. Sung, “Deepfakestack: A deep ensemble-based learning technique for deepfake detection,” in *2020 7th IEEE international conference on cyber security and cloud computing (CSCloud)/2020 6th IEEE international conference on edge computing and scalable cloud (EdgeCom)*. IEEE, 2020, pp. 70–75.

[31] A. B. Arrieta, N. Dáz-Rodrguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcá, S. Gil-López, D. Molina, R. Benjamins *et al.*, “Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai,” *Information fusion*, vol. 58, pp. 82–115, 2020.

[32] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru, “Model cards for model reporting,” in *Proceedings of the conference on fairness, accountability, and transparency*, 2019, pp. 220–229.

[33] T. Speith, “A review of taxonomies of explainable artificial intelligence (xai) methods,” in *2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 2239–2250.

[34] B. Khoo, R. C.-W. Phan, and C.-H. Lim, “Deepfake attribution: On the source identification of artificially generated images,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, p. e1438, 2022.

[35] B. Malolan, A. Parekh, and F. Kazi, “Explainable deep-fake detection using visual interpretability methods,” in *2020 3rd International Conference on Information and Computer Technologies (ICICT)*. IEEE, 2020, pp. 289–293.

[36] C. Rudin, “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead,” *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[37] G. Kordopatis-Zilos, C. Tzelepis, S. Papadopoulos, I. Kompatsiaris, and I. Patras, “Dns: Distill-and-select for efficient and accurate video indexing and retrieval,” *arXiv preprint arXiv:2106.13266*, 2021.

[38] S. Baxevanakis, G. Kordopatis-Zilos, P. Galopoulos, L. Apostolidis, K. Levacher, I. Baris Schlicht, D. Teyssou, I. Kompatsiaris, and S. Papadopoulos, “The mever deepfake detection service: Lessons learnt from developing and deploying in the wild,” in *Proceedings of the 1st International Workshop on Multimedia AI against Disinformation*, 2022, pp. 59–68.

Funded by the Horizon 2020
Framework Programme of the European Union