# Ph.D. Research Proposal

# Multimodal Fusion and Explainable AI(XAI) for Multimedia Content Verification

## Data and computer sciences

## spanning theoretical computer science, data analytics, artificial intelligence, and machine learning,

Alidor Mbayandjambe Masheke, M.Sc
alidor.mbayandjambe@unikin.ac.cd

June 27, 2025

**Abstract**

The proliferation of manipulated multimedia content, ranging from deepfakes to synthetic audio and contextually misleading videos, poses a significant challenge to public trust, democratic processes, and social cohesion. This Ph.D. research project aims to develop an advanced system for multimedia content verification by integrating multimodal fusion techniques and explainable artificial intelligence (XAI). The proposed approach leverages deep learning and multimodal learning to analyze cross-modal dependencies and inconsistencies, combining visual-linguistic transformers with temporal modeling for spatial and semantic alignment. Explainable AI techniques such as SHAP (SHapley Additive exPlanations), attention visualization, LIME, and counterfactual analysis will be integrated to ensure transparency and reliability. The research will address key use cases, including deepfake detection, verification of video-speech consistency, and classification of misleading social media posts. Expected outcomes include a state-of-the-art verification framework, scientific contributions to multimedia forensics, publications in high-impact venues, and a prototype tool for practical use in journalism and public institutions. By advancing trustworthy AI and media forensics, this research will contribute to global efforts in combating disinformation while fostering local capacity building and interdisciplinary collaboration.Data privacy and ethical considerations will be strictly adhered to, ensuring responsible use of sensitive information.

## 1 Overview and Background

The proliferation of manipulated multimedia content via social media and news outlets has heightened the need for robust verification systems. While unimodal detection methods—such as analyzing visual or audio anomalies exist, they often fail to detect complex cross-modal manipulations [N+21]. Recent deep learning models like ViLT and CLIP show promise in capturing such inconsistencies across modalities [L+22], yet their lack of explainability hinders adoption in sensitive areas such as journalism and law enforcement.

This project addresses this gap by combining multimodal fusion and Explainable AI (XAI) to enable both the detection and transparent interpretation of manipulated content. Multimodal fusion, integrating data such as image capture or video transcript pairs, improves robustness and generalizability [BAM18, NMN+25a]. At the same time, XAI techniques provide critical insights into model

decisions, which is essential for building trust in high-stakes applications [ADRDS+20]. In an age of fake news and deepfakes threatening public trust and democratic processes [CC19, MNN+25], this research proposes a unified, interpretable framework for verifying multimedia content.

# 2 Motivation

By using multimodal fusion and explainable AI, we can analyze complex data from different modalities (image, video, audio, text) to detect inconsistencies that signal potential manipulations. This approach not only allows for more accurate identification of falsified content but also explains why and how this content was classified as authentic or manipulated. Existing tools often operate as "black boxes," undermining trust in AI-driven verification. By developing a transparent, multimodal system, this project aims to empower journalists, policymakers, and civil society with actionable insights while advancing scholarly understanding of cross-modal forensics.The integration of explainability mechanisms is therefore crucial to ensure responsible deployment and strengthen end-user trust, whether they are journalists, content moderators, or ordinary citizens [GA19, AB18]. The motivations behind this research are multifold:

- **The rise of synthetic and manipulated media** (e.g., deepfakes, AI-generated text) increases the difficulty of manual content verification.
- **Most current models focus on a single modality**, ignoring complementary information available across modalities (e.g., inconsistencies between video and audio or image and text).
- **Lack of explainability** in AI systems hinders their adoption by non-technical stakeholders such as journalists, investigators, and policy-makers.
- **Need for early detection tools** to stop the spread of misinformation before it becomes viral.

# 3 Objectives

## Objective 1: Design a Multimodal Fusion Framework

Develop a hybrid architecture (early/late fusion) leveraging transformer-based models (e.g., VideoBERT, CLIP) to detect inconsistencies across video, audio, and text modalities. The framework will employ contrastive learning to distinguish authentic from manipulated content.

## Objective 2: Integrate Explainable AI Techniques

Incorporate XAI methods (SHAP, Grad-CAM, LIME) to generate human-interpretable explanations, highlighting manipulated regions and semantic contradictions. Evaluate explanation fidelity using user studies and quantitative metrics.

## Objective 3: Validate on Real-World Datasets

Test the system on benchmarks (FakeAVCeleb, DFDC) and DRC-specific datasets to assess generalizability. Metrics include AUC, precision-recall, and cross-modal alignment scores.

## Objective 4: Deploy a Prototype Tool

Build a user-friendly interface for content moderators and journalists, enabling semi-automatic verification with explainable outputs.

### Suggested Journal Titles and Potential Article Titles from the PhD Work

- X-MediaCheck: An Explainable AI Framework for Multimedia Content Verification: - Suitable for journals like IEEE Transactions on Multimedia or Pattern Recognition.

- Enhancing Multimedia Verification Capabilities: Integrating Multimodal Fusion Architectures and Explainability Techniques for Manipulation Detection: - Could be submitted to Multimedia Tools and Applications or Computer Vision and Image Understanding.
- Detecting Cross-Modal Inconsistencies in Media Content Using Explainable Deep Learning: - Potential submission to Information Fusion or Expert Systems with Applications.
- Towards AI-Assisted Fact-Checking Systems: An Explainable Approach for Automatic Multimedia Content Verification: - Target journals like Digital Journalism or New Media Society.

# 4 Research Questions

Q1 How can multimodal fusion architectures be optimized to effectively detect inconsistencies between different modalities (image, audio, text) in manipulated content?

Q2 Which explainable AI methods are most appropriate for making multimedia verification systems' decisions transparent and interpretable?

Q3 How can multimodal verification systems be adapted to withstand new manipulation techniques and maintain their effectiveness in the face of evolving deepfakes and other synthetic content?

Q4 To what extent do the explanations generated by XAI systems improve the trust and acceptance of end users (journalists, fact-checkers) in the context of content verification?

# 5 Targeted Issues

This research focuses on key multimedia manipulations that pose significant risks to digital information integrity. The main challenges addressed include:

1. **Video Deepfakes**: These highly realistic fake videos depict individuals saying or doing things they never did [ML21]. Deepfakes have surged by 900% between 2019 and 2021, threatening reputations, public trust, and democratic systems [APCC19]. Their realism makes manual detection unreliable, requiring advanced automated techniques.
2. **Synthetic Audio**: Voice cloning technologies can replicate a person's voice using minimal data [WTY20]. Recent fraud cases using synthetic audio have led to serious financial and reputational harm. Detection is difficult due to subtle acoustic cues that often escape conventional analysis.
3. **Decontextualization**: Authentic content is misrepresented by being placed in false contexts [WD18]. This includes outdated videos presented as current or real images paired with misleading captions. Such manipulations are common on social media and distort public perception.
4. **Hybrid Multimodal Manipulations**: These involve combinations of manipulations across modalities, e.g., a deepfake video with synthetic audio and misleading text [ZWZ20]. They present the greatest challenge, demanding coordinated multimodal analysis.

Tackling these challenges through multimodal fusion and explainable AI can improve detection accuracy and enhance transparency [NMN+25b], fostering greater trust in content verification systems.

# 6 Methodology

## Phase 1: Literature Review and Problem Formalization

- Analyze state-of-the-art techniques in multimodal learning, content verification, and explainable AI.
- Identify open challenges and define evaluation criteria.
- Discussion with experts in the relevant fields (AI specialists, fact-checkers, journalists)

### Phase 2: Dataset Construction and Preprocessing

- Acquisition of data from multiple sources such as reference datasets (FakeAVCeleb, DFDC, multimodal Twitter datasets)
- Preprocessing of data to handle missing values, remove duplicates, and normalize data to ensure consistency
- Identification of relevant features that may signal manipulations across different modalities
- Annotation of additional data for specific use cases not covered by existing datasets

### Phase 3: Model Design and Training

- Design multimodal deep learning architectures (e.g., transformers, late fusion models, contrastive learning).
- Integrate explainability modules such as Grad-CAM, LIME, SHAP, and cross-modal attention visualization.
- Validation of proposed architectures by comparing their performance with reference unimodal models

### Phase 4: Evaluation and Validation

- Perform experiments using accuracy, precision, recall, F1-score, AUC-ROC and human interpretability metrics.
- Compare against unimodal and non-explainable baselines.

### Phase 5: Prototype Development and Dissemination

- Design of a user interface allowing interaction with the generated explanations
- Share code, data, and results through open-access platforms and scientific publications.
- Conducting user studies with journalists and fact-checkers to evaluate the practical utility of the system
- Iterative refinement of the system based on user feedback

## 7 Expected Outcomes

At the end of our research, we expect the following outcomes:

- Development of new theories, models, or technologies in the field of multimodal content verification
- Publications in high-impact journals, like *IEEE Transactions on Multimedia* and *ACM FAT** and presentations at international conferences
- Strengthening of research skills and knowledge in advanced multimodal fusion and explainable AI methodologies
- Contribution to improving information verification practices through more accurate and transparent tools
- A prototype system demonstrating the practical applicability of the developed approaches

# 8 Ethical Considerations

- Anonymize sensitive data (e.g., political speech videos).
- Ensure XAI outputs do not inadvertently expose private information.
- Collaborate with local stakeholders to mitigate unintended societal impacts.

# 9 Research Timeline

This section outlines the planned schedule for the research activities over a period of three years, starting in February 2026 and ending in January 2029, as detailed in Table 1.

Table 1: Research Timeline.

| **Phase** | **Duration** | **Activities** |
|---|---|---|
| Literature Review | February to May 2026 | Collection of relevant literature on multimodal fusion and explainable AI techniques. Writing a systematic review and submitting to a WoS-indexed journal. |
| Data Collection and Preprocessing | June to August 2026 | Full data collection and preprocessing. Annotation of additional data if necessary. |
| Development of Multimodal Fusion Architectures | September to December 2026 | Design and implementation of transformer-based architectures. Exploration of multimodal fusion strategies. |
| Integration of Explainable AI Methods | January to March 2027 | Implementation of interpretability techniques and development of attention visualizations. |
| Experimentation and Analysis | April to July 2027 | Execution of extensive experiments and performance evaluation. |
| Prototype Development and User Evaluation | August to October 2027 | Implementation of a prototype system and conduction of user studies. |
| Thesis Writing and Dissemination of Results | November 2027 to Jun 2028 | Finalization of the dissertation, submission of papers to high-impact journals, and presentation of results at international conferences. |

# 10 Conclusion

This doctoral research project represents a strategic step in advancing my academic trajectory and contributing meaningfully to the field of multimedia content verification. By leveraging multimodal fusion and explainable AI, the project seeks to develop robust and transparent methods for detecting and understanding content manipulation. The expected outcomes aim to support the scientific community and institutional stakeholders engaged in the fight against disinformation. Conducted

within a stimulating and interdisciplinary research environment, this Ph.D. will foster methodological innovation and strengthen the scientific impact of ongoing efforts in AI and multimedia analysis.

# References


[AB18] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, 6:52138–52160, 2018.

[ADRDS+20] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Arturo Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020.

[APCC19] Henry Ajder, Giorgio Patrini, Francesco Cavalli, and Laurence Cullen. The state of deepfakes: Landscape, threats, and impact. https://www.sensity.ai/reports/the-state-of-deepfakes-2019, 2019. Sensity AI Report.

[BAM18] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, 2018.

[CC19] Robert Chesney and Danielle Keats Citron. Deepfakes and the new disinformation war: The coming age of post-truth geopolitics. *Foreign Affairs*, 98(1):147–155, 2019.

[GA19] David Gunning and David Aha. Darpa's explainable artificial intelligence (xai) program. *AI Magazine*, 40(2):44–58, 2019.

[L+22] Jiasen Lu et al. Vilt: Vision-and-language transformer without convolution or region supervision. *International Conference on Machine Learning (ICML)*, pages 1–15, 2022.

[ML21] Yisroel Mirsky and Wenke Lee. The creation and detection of deepfakes: A survey. *ACM Computing Surveys (CSUR)*, 54(1):1–41, 2021.

[MNN+25] Alidor M. Mbayandjambe, Grevi B. Nkwimi, Darren Kevin T. Nguemdjom, Fiston Oshasha, Célestin Muluba, and Xavier F. Kutuka. Ai-generated sneaker detection: Leveraging gans and convolutional neural networks for image classification. *International Journal of Innovative Science and Research Technology*, 10(4):3413–3422, 2025.

[N+21] Thanh Thi Nguyen et al. Deepfake detection: Current challenges and next steps. *IEEE International Conference on Multimedia & Expo (ICME)*, pages 1–6, 2021.

[NMN+25a] Darren Kevin T. Nguemdjom, Alidor M. Mbayandjambe, Grevi B. Nkwimi, Fiston Oshasha, Célestin Muluba, Héritier I. Mbengandji, and Ibsen G. Bazie. Explainable ai (xai) for obesity prediction: An optimized mlp approach with shap interpretability on lifestyle and behavioral data. *International Journal of Innovative Science and Research Technology*, 10(4):3192–3200, 2025.

[NMN+25b] Darren Kevin T. Nguemdjom, Alidor M. Mbayandjambe, Grevi B. Nkwimi, Fiston Oshasha, Célestin Muluba, Héritier I. Mbengandji, Ibsen G. Bazie, Raphael Kpoghomou, and Alain M. Kuyunsa. Enhancing the robustness of computer vision models to adversarial perturbations using multi-scale attention mechanisms. *International Journal of Innovative Science and Research Technology*, 10(4):3565–3578, 2025.

[WD18] Claire Wardle and Hossein Derakhshan. Information disorder: Toward an interdisciplinary framework for research and policymaking. *Council of Europe Report DGI*, 9:1–107, 2018.

[WTY20] Xin Wang, Shinji Takaki, and Junichi Yamagishi. Defence against adversarial attacks in speaker recognition using multi-channel audio. In *ICASSP 2020 - IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6469–6473. IEEE, 2020.

[ZWZ20] Xinyi Zhou, Ziqi Wu, and Reza Zafarani. Joint multimodal fake news detection with news content and social engagements. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 347–356. ACM, 2020.