Explainable Multimodal Deepfake Detection Using Image and Audio Transformers with Attention Mechanisms

Charan G[1], Dhanush C[2], Dhanush S[3], Dheemanth M[4], and Dr. Pallavi GB[5]
[1–5]BMS College of Engineering, Department of Computer Science and Engineering, Bengaluru, India
Email: charan.cs22@bmsce.ac.in, {dhanushc.cs22, dhanushs.cs22, dheemanth.cs22m, pallavi.cse}@bmsce.ac.in

***Abstract*—The sophistication of deepfake generation technology has reached alarming levels, creating synthetic audio and video content that is nearly indistinguishable from real media. This proliferation across digital platforms raises serious concerns about misinformation campaigns, unauthorized identity manipulation, and the fundamental erosion of trust in digital content. Our survey examines contemporary deepfake detection methodologies with particular emphasis on multimodal approaches that analyze both visual and acoustic elements simultaneously. We investigate developments in CNN-based detection systems, transformer architectures designed for image and audio analysis, and attention mechanisms that facilitate enhanced feature discrimination. Beyond evaluating the capabilities and limitations of existing detection techniques, we analyze widely-used benchmark datasets, standard evaluation metrics, and persistent challenges including inadequate cross-domain performance, insufficient multimodal integration, and limited model transparency. Drawing from this comprehensive analysis, we present a conceptual multimodal detection architecture that combines modality-specific transformers with attention-driven explainability mechanisms to enable transparent, robust, and real-time deepfake identification.**

***Index Terms*—Deepfake Detection, Vision Transformers, Audio Transformers, Explainable AI, Attention Mechanisms, Multimodal Fusion**

I. Introduction

What started as an experimental curiosity in research laboratories has rapidly evolved into a significant threat affecting digital media authenticity worldwide. Deepfakes—synthetically generated audiovisual content created using advanced neural networks—now possess the capability to replicate human facial expressions, speech characteristics, and visual identities with disturbing accuracy [1]. The consequences extend far beyond technical novelty, manifesting as threats to political discourse integrity, personal security, financial systems, and biometric authentication mechanisms [3].

Initial detection strategies predominantly concentrated on examining single data modalities. Visual detection methods searched for telltale signs like textural anomalies, edge artifacts, and unnatural facial motion patterns [2], [4]. These unimodal approaches performed adequately when confronting early-generation deepfakes that exhibited obvious imperfections. However, they demonstrate significantly reduced effectiveness against modern forgeries produced through advanced generative adversarial networks and sophisticated neural face manipulation techniques [8]. The audio domain presents comparable difficulties. Contemporary voice synthesis and cloning technologies can accurately reproduce speaker-specific vocal characteristics including tone, prosody, and speaking patterns, facilitating voice-based fraud and impersonation schemes [9], [10]. Current audio detection systems, despite their utility, continue to struggle with cross-linguistic generalization and interpretability issues [11], [12].

Recognizing the inherent limitations of analyzing individual modalities in isolation, researchers have increasingly turned toward multimodal detection frameworks that exploit both visual and acoustic information streams [7]. The underlying rationale suggests that integrating complementary signal types should yield more resilient detection capabilities, particularly when dealing with sophisticated real-world deepfakes. Unfortunately, existing multimodal implementations encounter significant obstacles: suboptimal fusion methodologies, temporal synchronization problems between modalities, and diminished robustness when encountering previously unseen manipulation techniques [13].

The interpretability problem deserves equal consideration. A substantial portion of contemporary deepfake detectors—particularly those leveraging deep neural architectures—function essentially as black boxes. While these systems may achieve high accuracy rates, they typically provide minimal insight into their decision-making processes, creating difficulties for security analysts, forensic investigators, and end-users who need to understand the reasoning behind classifications. This opacity has fueled growing interest in Explainable AI methodologies, with techniques like LIME, Grad-CAM, and attention visualization mechanisms offering pathways toward understanding model behavior and identifying decision-critical features [5].

Transformer architectures have emerged as promising candidates for both visual analysis (Vision Transformers) and audio processing (spectral transformers), primarily due to their capacity for modeling long-range dependencies and capturing complex contextual relationships [6]. Additionally, novel research directions including quantum-inspired learning paradigms and computationally efficient real-time architectures underscore the continued evolution occurring within this research domain [14], [17].

However, substantial gaps persist. The absence of scalable, generalizable, and interpretable multimodal systems represents a critical barrier to practical deployment. This survey delivers a thorough synthesis of existing model architectures, available datasets, evaluation methodologies, and unresolved research challenges. We subsequently present a conceptual multimodal transformer framework specifically engineered to enhance both detection robustness and system transparency.

II. Detection Model Architectures

Deepfake detection research has undergone substantial transformation driven by advances in transformer-based architectures and explainability techniques. Contemporary research trends clearly indicate movement toward integrating attention mechanisms within both image and audio processing pipelines, enabling detection of manipulation patterns that conventional CNNs and manually engineered features frequently miss.

A. Visual Deepfake Detection Approaches

Convolutional neural networks formed the foundation of early visual deepfake detection systems. These architectures exploited spatial feature hierarchies to identify subtle irregularities introduced during manipulation processes. CNNs showed promise under controlled experimental conditions but their restricted interpretability and limited capacity for capturing long-range spatial relationships motivated exploration of alternative architectural designs [2].

Hybrid architectures combining CNNs with attention mechanisms quickly gained momentum. Attention layers proved valuable for highlighting manipulated regions and emphasizing artifact-related features [8]. The introduction of Vision Transformers (ViTs) represented a more fundamental architectural shift. By processing images as sequences of patch embeddings and employing self-attention mechanisms, ViTs excel at capturing global spatial relationships—a capability particularly beneficial for detecting inconsistencies across manipulated facial regions and textural patterns [6]. Specialized variants including Class-Attention Vision Transformers have demonstrated exceptional performance in identifying subtle manipulation traces [9].

Interpretability considerations have progressively become central to detector design philosophy. Ensemble methodologies now routinely combine CNNs, transformers, and gradient-based visualization techniques to deliver interpretable evidence supporting classification decisions [4], [5]. Parallel research efforts focus on GAN-aware detection models that leverage semi-supervised learning to recognize synthesis patterns across diverse manipulation method families [3]. Nevertheless, modality-specific interpretability mechanisms remain relatively underdeveloped, suggesting opportunities for further investigation.

Several challenges continue to limit practical applicability: insufficient cross-dataset generalization, difficulties processing heavily compressed video content, and inadequate temporal modeling capabilities all restrict real-world deployment potential.

Key Finding: Research literature consistently demonstrates that hybrid architectural approaches—integrating convolutional feature extraction, transformer-based modeling, and explainability components—deliver superior detection performance. Specific architectures including CVT, CaiT, M2TR, and ensemble configurations incorporating Grad-CAM or LIME visualization achieve enhanced accuracy while maintaining decision transparency [4], [6], [9]. These hybrid systems substantially outperform traditional standalone CNN architectures by simultaneously capturing localized image artifacts and modeling long-range spatial dependencies.

B. Audio Deepfake Detection Methodologies

Audio deepfake detection presents distinct challenges compared to visual analysis, requiring identification of anomalies within temporal and spectral signal characteristics. Early detection approaches relied predominantly on manually engineered acoustic features including mel-frequency cepstral coefficients (MFCCs), spectral centroid measurements, and pitch contour analysis. As neural vocoder technology advanced, these handcrafted feature representations proved increasingly inadequate for distinguishing authentic speech from synthetically generated audio [12], [15].

Contemporary detection methodologies typically employ deep learning architectures including convolutional neural networks, deep neural networks, and long short-term memory networks to model complex speech dynamics. Some approaches explicitly analyze transitions between voiced and unvoiced speech regions, exploiting the observation that synthesis artifacts frequently manifest during these transitions [14]. Transformer-based audio models process spectrogram representations using multi-head self-attention mechanisms, enabling detection of subtle anomalies in rhythmic patterns, prosodic characteristics, and spectral structure [11], [13]. Beyond their detection capabilities, attention mechanism outputs serve as interpretability tools by highlighting temporally suspicious audio segments.

Novel research directions include quantum-inspired learning models that embed acoustic signals within high-dimensional Hilbert spaces, offering alternative perspectives particularly relevant for low-data scenarios [13]. Multimodal lip-synchronization detection frameworks evaluate consistency between lip motion and corresponding speech content, demonstrating effectiveness for detecting audio-visual manipulation [16].

Audio detection systems face several persistent obstacles: limited generalization to emerging speech synthesis techniques, insufficient cross-lingual robustness, vulnerability to adversarial perturbations, and the notable absence of standardized explainability frameworks specifically designed for auditory models.

Key Finding: Across the audio detection domain, models integrating deep feature learning with attention-based mechanisms or metric learning strategies exhibit the strongest generalization capabilities. The AASIST architecture, employing graph attention networks for analyzing voiced and

unvoiced speech regions, demonstrates particularly robust performance [14]. Siamese LSTM configurations utilizing triplet loss functions prove effective for one-shot learning and resource-constrained detection scenarios [18]. Lightweight spectrogram-based CNN architectures remain appropriate for real-time applications with strict latency requirements.

III. Benchmark Datasets

Dataset quality fundamentally shapes research progress in deepfake detection. As manipulation techniques have grown more sophisticated, benchmark datasets have correspondingly expanded in scale, diversity, realism, and multimodal coverage. This section examines key datasets employed for image-based, audio-based, and multimodal deepfake detection while highlighting their strengths and continuing limitations that influence model evaluation.

A. Visual Deepfake Datasets

Visual deepfake research depends heavily upon several publicly accessible datasets including FaceForensics++ (FF++), Celeb-DF, and the DeepFake Detection Challenge (DFDC) dataset. FF++ has established itself as one of the most extensively utilized benchmarks, providing multiple compression quality levels and diverse manipulation techniques that facilitate model training and cross-method comparison [1]. Celeb-DF distinguishes itself through high-quality synthetic video content that presents significantly greater detection difficulty and more closely approximates real-world deepfake quality [2].

More recent dataset releases include CelebAMask-HQ, which provides detailed semantic segmentation masks for facial components. These pixel-level annotations prove especially valuable for explainability research by enabling evaluation of whether model attention maps correlate with semantically meaningful facial regions [2]. The Kaggle 140k Real and Fake Faces dataset introduces balanced distributions of authentic photographs and StyleGAN-generated synthetic images, facilitating learning of generalizable GAN artifact representations.

Competition-driven datasets like DFWild-Cup incorporate deepfakes collected from diverse real-world platforms. While offering valuable demographic and stylistic diversity, this dataset's 5:1 ratio of fake to real samples introduces challenges for balanced model training and unbiased evaluation [6]. DFDC maintains its position as the most comprehensive and practically relevant visual benchmark, encompassing over 100,000 manipulated videos exhibiting variation in ethnicity, illumination conditions, and manipulation methodologies [10]. CelebDF-v2 sees frequent use in cross-dataset performance assessment due to its realistic manipulation quality and challenging evaluation conditions.

Despite these advances, several limitations characterize current image-based datasets: demographic imbalances, absence of temporal annotations, dependence on manipulation techniques potentially rendered obsolete by rapid generator improvements, and critically, minimal availability of synchronized audio-visual samples that would support multimodal detection research.

Assessment: Considering dataset diversity, manipulation realism, and explainability support, DFDC and Celeb-DF provide the strongest foundations for evaluating contemporary visual deepfake detectors. DFDC delivers large-scale samples with substantial variation while Celeb-DF and CelebDF-v2 facilitate rigorous generalization assessment under challenging conditions.

B. Audio Deepfake Datasets

Audio deepfake detection has progressed alongside specialized datasets targeting voice spoofing and synthetic speech. The ASVspoof challenge series—particularly ASVspoof2019-LA and ASVspoof2021-DF—has established itself as the primary benchmark for assessing speaker verification system security against spoofing attacks [18]. These datasets contain thousands of genuine and spoofed audio samples generated using diverse speech synthesis and voice conversion algorithms.

The Fake-or-Real (FoR) dataset, especially its FoR-2sec variant, has gained popularity through its emphasis on short-duration audio clips appropriate for rapid and real-time inference applications [15]. The MLAAD dataset adopts a more analytical perspective by segmenting audio into voiced and unvoiced components, reflecting empirical evidence that unvoiced segments frequently expose synthesis artifacts more readily [14]. Additional resources including DFAD and augmented LibriSpeech-based datasets facilitate evaluation of real-time classification systems intended for deployment on resource-constrained or mobile platforms.

Multimodal datasets including FakeAVCeleb and TMC deliver synchronized audio-video recordings labeled across four explicit categories: authentic audio with authentic video, synthetic audio with authentic video, authentic audio with synthetic video, and synthetic audio with synthetic video [10]. These datasets prove particularly valuable for audio-visual consistency analysis including lip-synchronization deepfake detection [16].

While many audio datasets encompass diverse synthesis techniques, several challenges persist: restricted multilingual representation, scarcity of emotionally expressive or conversational speech samples, and inconsistent recording conditions that collectively limit model generalization capabilities.

Assessment: For audio-exclusive detection tasks, ASVspoof2019-LA and FoR-2sec provide the most comprehensive and widely adopted foundations. For multimodal detection applications, FakeAVCeleb offers particular value through its synchronized audio-video structure that enables cross-modal alignment analysis while supporting explainability investigations.

IV. Evaluation Metrics and Performance Measurement

Effective evaluation of deepfake detection models requires metrics that capture not only classification quality but also

cross-dataset robustness and, where feasible, prediction interpretability. Both image-based and audio-based detection systems utilize standard machine learning metrics while incorporating domain-specific measurements tailored to manipulation detection requirements.

A. Metrics for Visual Detection Systems

Within the visual detection domain, researchers commonly employ accuracy, precision, recall, F1-score, and Area Under the ROC Curve (AUC) as primary evaluation metrics [1], [3], [4]. Among these, AUC has gained particular prominence due to its emphasis on class discrimination capability independent of classification threshold selection, making it especially appropriate for datasets exhibiting class imbalance.

Confusion matrices find frequent use for visualizing Type I and Type II error distributions and assessing potential model bias toward either real or fake sample detection [3]. For models incorporating explainability mechanisms, Intersection over Union (IoU) metrics evaluate alignment between predicted manipulation regions—derived from saliency maps or Grad-CAM visualizations—and manually annotated ground-truth masks when such annotations are available [2].

Cross-dataset evaluation protocols have become essential components of contemporary research methodology. Models undergo training on one dataset (such as FF++) before evaluation on an entirely distinct dataset (such as Celeb-DF) to assess real-world generalization capabilities. Compression-based robustness testing is widely adopted to simulate practical deployment scenarios where video content experiences heavy recompression through social media platforms [6].

Assessment: The combination of AUC, F1-score, and IoU delivers comprehensive understanding of model behavior. AUC quantifies discriminative capability, F1-score balances precision-recall tradeoffs, and IoU measures interpretability alignment—collectively forming a well-suited metric suite for explainable deepfake detection research.

B. Metrics for Audio Detection Systems

Audio deepfake detection models similarly rely upon accuracy, precision, recall, and F1-score for baseline performance characterization [12], [15], [17]. However, this domain places particular emphasis on Equal Error Rate (EER), a metric with extensive usage history in biometrics and speaker verification systems [18]. EER identifies the operating point where false acceptance rate equals false rejection rate, providing a unified quantitative measure of security-usability balance.

Temporal alignment metrics including Normalized Levenshtein Distance have been adopted specifically for lip-synchronization deepfake detection, where the objective involves comparing textual transcriptions or phoneme sequences extracted from audio against predicted lip movements derived from video [16]. Latency measurements and inference speed evaluations are increasingly recognized as essential factors, particularly for systems designed for real-time deployment on mobile or edge computing devices [17].

Assessment: EER maintains its position as the most domain-relevant evaluation metric for audio-based detection, complemented by F1-score and latency measurements for real-time deployment contexts. This metric combination effectively characterizes both detection performance and practical deployment viability.

V. Research Gaps and Open Challenges

Despite considerable research progress, several critical gaps continue to impede development of reliable, scalable, and explainable deepfake detection systems deployable in real-world scenarios.

A dominant limitation across current literature involves the prevalence of unimodal detection approaches. While visual-only and audio-only systems achieve reasonable effectiveness within their respective domains, they fundamentally fail to capture cross-modal inconsistencies—often the most revealing indicators in sophisticated deepfake scenarios [10]. Even existing multimodal implementations frequently employ simplistic late fusion strategies that neglect higher-level cross-modal relationship modeling.

Explainability constitutes another major unresolved challenge. The majority of deepfake detectors, particularly transformer-based architectures, generate classification decisions without providing transparent justification. Although saliency visualization methods and attention mechanism analysis offer partial insight, these approaches lack standardization and exhibit significant reliability variation across different model architectures [5], [7]. Remarkably few research efforts evaluate how interpretability mechanisms impact user trust, system usability, or forensic investigation value.

Generalization capability represents an equally critical concern. Models trained on specific datasets frequently exhibit substantial performance degradation when evaluated on alternative datasets, particularly when confronting previously unseen manipulation styles, varying compression levels, or diverse demographic distributions [1], [6]. Robust adversarial training methodologies and domain adaptation techniques remain comparatively underexplored within the deepfake detection research landscape.

Real-time detection requirements introduce additional complexity. High-performing models, particularly transformer-based architectures, impose considerable computational demands that preclude straightforward deployment on edge computing platforms or mobile devices [17]. Furthermore, latency constraints, energy consumption profiles, and memory requirements receive inadequate attention in existing literature.

Finally, fairness considerations and forensic attribution capabilities are largely absent from current research efforts. Many benchmark datasets exhibit demographic imbalances, and very few detection models attempt to identify the specific manipulation method employed in generating detected deepfakes. Additionally, inconsistent multimodal evaluation standards complicate meaningful cross-study performance comparisons.

These accumulated gaps clearly motivate the need for detection frameworks that simultaneously achieve multimodal integration, interpretable decision-making, cross-domain generalization, and computational efficiency suitable for practical deployment scenarios.

VI. Proposed Multimodal Detection Framework

To systematically address the limitations identified through our comprehensive literature analysis, we outline a unified multimodal detection framework designed for joint analysis of visual and auditory signals while maintaining clear interpretability. Our proposed system architecture incorporates two dedicated transformer-based processing branches—one handling image frame sequences extracted from video content and another analyzing corresponding speech signals. Figure 1 presents the conceptual system design.

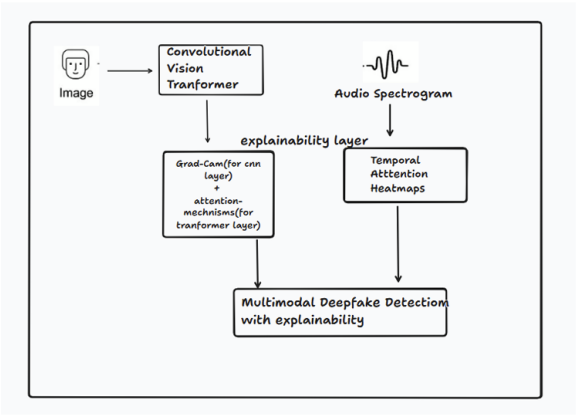

Fig. 1. Conceptual workflow of the proposed explainable multimodal deepfake detection system combining CVT and AASIST.

The visual processing pipeline builds upon a Vision Transformer architecture augmented with convolutional feature extraction modules. This hybrid structural design enables simultaneous capture of localized facial artifacts and long-range contextual patterns that frequently reveal subtle manipulation indicators in deepfake video content. To enhance interpretability, we apply gradient-weighted class activation mapping (Grad-CAM) techniques to convolutional layers while leveraging self-attention maps from transformer stages to provide insight into global dependency relationships across facial regions.

The audio processing branch employs a spectrogram-based Audio Spectral Transformer for examining temporal and spectral characteristics of speech signals. Transformer self-attention mechanisms enable the model to highlight irregularities in rhythmic patterns, prosodic features, and harmonic structure—acoustic properties frequently distorted by synthetic voice generation systems. Attention heatmap visualizations identify specific speech segments contributing most significantly to classification decisions, rendering audio predictions more transparent and interpretable.

A cross-modal attention fusion layer forms the architectural core. This module integrates high-level feature embeddings from both modalities and explicitly models relationships between facial movements and corresponding speech characteristics. By directly modeling audio-visual coherence, the framework facilitates identification of inconsistencies that unimodal systems routinely overlook, including asynchronous lip motion or acoustically unnatural voice patterns.

To ensure practical deployment viability, lightweight architectural variants such as MobileViT and Lite-AST can be incorporated to substantially reduce computational overhead. We additionally recommend implementing domain adaptation strategies to enhance cross-dataset robustness, particularly critical for deployment environments encountering diverse manipulation techniques or variable recording conditions.

The overall design simultaneously addresses three fundamental challenges: (1) bridging modality gaps through explicit cross-modal fusion, (2) improving generalization through transformer-based representation learning, and (3) enabling decision transparency through layered explainability mechanisms. The resulting system establishes a foundation for developing trustworthy and scalable multimodal deepfake detection platforms appropriate for deployment in security applications, media verification systems, and biometric authentication contexts.

VII. Conclusion and Future Directions

The continued advancement of deepfake generation technologies has substantially amplified risks associated with misinformation propagation, identity fraud, and digital content manipulation. This survey has examined a comprehensive spectrum of deepfake detection approaches encompassing CNN-based systems, transformer-driven architectures, audio-specific detection models, and multimodal frameworks. The progressive adoption of attention mechanisms and interpretability tools reflects the research community's ongoing effort toward developing more transparent and trustworthy detection solutions.

Despite these advances, several persistent gaps remain unresolved: inadequate cross-dataset generalization, computational efficiency limitations, suboptimal multimodal fusion quality, and insufficient model interpretability. These shortcomings underscore the fundamental requirement for detection systems that achieve not merely high accuracy but also explainability, scalability, and adaptability to emerging manipulation techniques.

The multimodal transformer-based framework proposed in this work specifically targets these challenges through integration of visual and audio analysis with explicit attention-driven interpretability mechanisms. By simultaneously capturing localized artifacts and modeling global cross-modal relationships, the system provides a robust foundation for future multimodal deepfake detection research. Promising extensions include multilingual audio modeling, fairness-aware evaluation protocols, adversarial robustness enhancement, and real-time

deployment optimizations to facilitate broad practical adoption across diverse application domains.

REFERENCES

[1] M. S. Rana, M. N. Nobi, B. Murali, and A. H. Sung, "Deepfake Detection: A Systematic Literature Review," *IEEE Access*, vol. 10, pp. 25494–25513, 2022.

[2] B. M. G. *et al.*, "Detecting AI-generated Images with CNN and Interpretation using Explainable AI," in *InC4*, 2024.

[3] J. John and B. V. Sherif, "Comparative Analysis on DeepFake Detection Methods and Semi-Supervised GAN Architecture," in *I-SMAC*, 2022.

[4] A. A. Kumar *et al.*, "XAI–Empowered Ensemble Deep Learning for Deepfake Detection," in *ICCCNT*, 2024.

[5] N. Mansoor and A. I. Iliev, "Explainable AI for DeepFake Detection," *Applied Sciences*, 2025.

[6] "A Timely Survey on Vision Transformer for Deepfake Detection," arXiv:2405.08463, 2024.

[7] J. Jheelan and S. Pudaruth, "Deepfake Detection with Explainable AI," *Computers*, 2025.

[8] "Hybrid Deepfake Image Detection with CNN and Attention," arXiv:2502.10682, 2025.

[9] A. Pandey and B. Rudra, "Enhancing Deepfake Detection through Quantum Transfer Learning and Class-Attention ViT," *Applied Sciences*, 2025.

[10] S. Muppalla, S. Jia, and S. Lyu, "Integrating Audio-Visual Features for Multimodal Deepfake Detection," in *MIT URTC*, 2023.

[11] B. Sarada *et al.*, "Audio Deepfake Detection and Classification," in *APCIT*, 2024.

[12] H. H. Kilinc and F. Kaledibi, "Audio Deepfake Detection Using Machine and Deep Learning," in *WINCOM*, 2023.

[13] A. Pandey and B. Rudra, "Deepfake Audio Detection Using Quantum Learning Models," in *MECOM*, 2024.

[14] G. Sivaraman, H. Tak, and E. Khoury, "Investigating Voiced and Unvoiced Speech Regions for Audio Deepfake Detection," in *ICASSP*, 2025.

[15] V. Sundaram *et al.*, "Leveraging Acoustic Features and Deep Neural Architectures for Audio Deepfake Detection," in *ICACC*, 2024.

[16] M. Bohacek and H. Farid, "Lip-sync Deepfake Detection from Audio-Video Mismatch," in *CVPRW*, 2024.

[17] P. Chiddarwar, "Real-Time Detection of AI-Generated Audio," in *ICT-BIG*, 2024.

[18] A. Khan and K. M. Malik, "One-Shot Learning for Audio Deepfake Detection," in *WIFS*, 2023.