SiftSentinel - Unified Deepfake and Cyberbullying Detection and report generation using Explainable and Generative AI

Shikha Pachouly[1]
Assistant Professor, Computer Engineering
All India Shri Shivaji Memorial Society's College of Engineering.
Savitribai Phule Pune University
Pune, India

Piyusha Supe[2]
Computer Engineering
All India Shri Shivaji Memorial Society's College of Engineering.
Savitribai Phule Pune University
Pune, India

Iffa Shaikh[3]
Computer Engineering
All India Shri Shivaji Memorial Society's College of Engineering.
Savitribai Phule Pune University
Pune, India

Pratham Pasalkar[4]
Computer Engineering
All India Shri Shivaji Memorial Society's College of Engineering.
Savitribai Phule Pune University
Pune, India

Varad Marathe[5]
Computer Engineering
All India Shri Shivaji Memorial Society's College of Engineering.
Savitribai Phule Pune University
Pune, India

Abstract: The rapid proliferation of deep fakes and online harassment has intensified the need for robust digital content verification systems. *SiftSentinel* is an innovative, web-based platform designed to detect and analyze deep fake content and cyberbullying across multimodal data, including images, videos, audio, and text. The system integrates deep learning and natural language processing techniques utilizing Convolutional Neural Networks (CNNs) for visual media analysis, acoustic feature extraction for audio authenticity assessment, and NLP-based models for identifying offensive or harassing language. Trained on large-scale, heterogeneous datasets, SiftSentinel effectively detects manipulated or harmful content ranging from subtle misinformation to overt abuse. A key aspect of the platform is the incorporation of Explainable AI (XAI), which provides transparent, interpretable reasoning behind each detection outcome. Additionally, a generative AI component produces automated, evidence-based reports offering mitigation recommendations for users, moderators, and law enforcement agencies. By combining deep fake detection, cyberbullying identification, and explainable generative analysis within a unified interface, SiftSentinel contributes to enhancing online safety and promoting trust in digital environments.

Keywords Deepfake Detection, Cyberbullying Detection, Multimodal AI, Explainable Artificial Intelligence (XAI), Generative AI, CNN, NLP, Audio Forensics, Online Content Moderation, Digital Trust.

I] INTRODUCTION

With the rapid expansion of social media, digital communication platforms, and content-sharing ecosystems, society is experiencing unprecedented levels of connectivity and creativity. However, this digital growth has simultaneously introduced critical challenges, notably in the form of deepfake media and cyberbullying, both of which have profound social, psychological, and ethical implications. Deepfakes, driven by advances in Generative Artificial Intelligence (AI), manipulate visual and audio content to produce highly realistic but falsified media. These synthetic artifacts are increasingly employed for malicious purposes, including misinformation campaigns, identity theft, character defamation, and political propaganda. The sophistication of Generative Adversarial Networks (GANs) and related AI techniques has made distinguishing authentic from manipulated media exceedingly difficult, thereby undermining public trust and threatening the credibility of digital information ecosystems. Conversely, cyberbullying exploits the anonymity and accessibility of online platforms to harass, intimidate, or defame individuals. Victims often endure emotional stress, anxiety, depression, and social isolationunderscoring the urgent need for intelligent and adaptive mechanisms capable of proactively identifying and mitigating abusive behaviors online. Traditional moderation systems and rule-based filters frequently fail to detect evolving linguistic patterns, emerging slang, and implicit forms of aggression, making them inadequate for today's dynamic digital landscape. Existing detection frameworks, though valuable, often lack adaptability, scalability, and transparency. Many current AI-based systems operate as black-box models, producing results without clear reasoning or interpretability. In high-stakes domains such as social media governance, education, and law enforcement, this opacity limits accountability and user trust, highlighting the necessity for explainable and interpretable AI-driven solutions. This research introduces SiftSentinel, a hybrid AI-powered detection framework that integrates Deepfake Detection, Cyberbullying Analysis, Explainable AI (XAI), Generative AI, and Agentic AI. The system is designed to

analyze and interpret multimodal content including images, videos, audio, and text while providing transparent explanations for each decision using XAI methodologies. Generative AI components generate real-time, evidence-based analytical reports and mitigation strategies, whereas Agentic AI ensures autonomous monitoring, adaptive learning, and intelligent response to evolving digital threats. The ultimate objective of this research is to develop a transparent, adaptive, and proactive detection platform that empowers users, organizations, and authorities to preserve trust, safety, and accountability in online environments. By combining detection accuracy with interpretability and intelligent intervention, SiftSentinel aims to foster safer digital communities and establish a responsible AI framework for combating emerging digital threats.

II] BACKGROUND AND FUNDAMENTAL CONCEPT

2.1 Deepfake Technology

Deepfakes are synthetic media generated using deep learning models, primarily Generative Adversarial Networks (GANs) and Auto encoders. These models can manipulate facial expressions, voices, and scenes to produce highly realistic fake content. Although deep fakes have legitimate applications in art and education, they are increasingly exploited for misinformation, identity theft, and character defamation. Detection typically relies on analyzing visual artifacts, temporal inconsistencies, or acoustic distortions, using CNNs, Vision Transformers (ViTs), and spectrogram-based models.

2.2 Cyberbullying and Online Harassment

Cyberbullying refers to online behaviors intended to harass, insult, or threaten individuals. Traditional rule-based detection systems fail to capture implicit or context-dependent abuse. Modern methods employ Natural Language Processing (NLP) and Machine Learning (ML) models such as LSTM, BERT, and RoBERTa to analyze sentiment, semantics, and intent in text data. However, evolving slang, sarcasm, and multilingual content remain key challenges.

2.3 Explainable Artificial Intelligence (XAI)

XAI enhances model transparency by explaining how predictions are made. Techniques like LIME, SHAP, and Grad-CAM highlight feature importance and decision reasoning. In SiftSentinel, XAI improves trust and accountability by providing users with interpretable insights into deepfake and cyberbullying detections. Together, these perspectives underscore the need for specialized methods that address not only data management and analytics but also governance, ethical concerns, and integration across domains. This synthesis provides the foundation for advancing Mobility Data Science as a discipline that bridges transportation science, computer science, and societal applications.

2.4 Generative and Agentic AI

Generative AI creates new content or data samples by learning from existing datasets. In this research, it is used for synthetic data generation and automated report generation, summarizing detection results and mitigation strategies. Agentic AI, on the other hand, focuses on autonomous decision-making and adaptive learning. It enables SiftSentinel to continuously monitor, learn from, and respond to new digital threats.

2.5 Multimodal AI

Multimodal AI combines multiple data typestext, image, video, and audio to enhance detection accuracy. SiftSentinel fuses CNN-based visual models, NLP-based textual analysis, and acoustic models for audio to identify complex and cross-modal digital manipulations.

III] LITERATURE REVIEW

The detection of deepfake content and cyberbullying has become an essential research domain to ensure digital safety, content integrity, and user trust. Recent advancements in Deep Learning (DL) and Explainable Artificial Intelligence (XAI) have led to the development of powerful frameworks focusing on visual and textual analysis. However, most existing systems address these problems in isolation, lacking adaptability, scalability, and multimodal integration.

3.1 Deepfake Detection and Explainable AI

Khalid *et al.* [1] proposed ExplaNET, a descriptive and interpretable framework for deepfake detection based on prototype-based learning. The framework utilizes DenseNet-121 as a backbone network and introduces a prototype layer to extract representative patterns from both authentic and manipulated images. Through Grad-CAM visualization, ExplaNET highlights specific facial regions that contribute to the classification decision, enhancing interpretability and trustworthiness.

Experimental evaluations conducted on benchmark datasets such as Face Forensics++, Celeb-DF, and DFDC-P demonstrated that ExplaNET achieved higher accuracy and explainability compared to traditional black-box deepfake detection methods. Despite these achievements, ExplaNET remains limited to visual media and lacks the ability to process multimodal data (e.g., text, audio, or mixed media). Moreover, it focuses primarily on post-hoc interpretability rather than adaptive, real-time explanation or intervention, making it less effective for broader digital forensics applications.

3.2 Cyberbullying Detection on Social Media

Murshed *et al.* [2] introduced DEA-RNN, a hybrid deep learning approach that combines the Elman-type Recurrent Neural Network (RNN) with the Dolphin Echolocation Algorithm (DEA) for parameter optimization and faster convergence. The model was evaluated on a Twitter dataset comprising 10,000 annotated tweets and outperformed conventional classifiers such as Bi-LSTM, CNN, and SVM in terms of accuracy, precision, recall, and F1-score.

The DEA-RNN model effectively addresses key challenges of short-text data such as slang, sarcasm, and limited context by optimizing RNN parameters dynamically. However, it remains restricted to textual content analysis and lacks explainability, as it operates as a black-box model with limited insight into its decision-making process. Furthermore, DEA-RNN does not integrate cross-modal data sources such as images or audio, which are increasingly present in online harassment contexts.

3.3 Research Gap and Contribution

While ExplaNET [1] advances explainable deep fake detection, and DEA-RNN [2] enhances efficient cyberbullying identification, both frameworks are domain-specific and operate independently within single modalities. Existing approaches thus fail to provide a unified, explainable, and adaptive system capable of handling the diversity of harmful online content across multiple media types.

To bridge this gap, the proposed research introduces SiftSentinel, a hybrid multimodal AI platform that fuses the interpretability of ExplaNET with the adaptability of DEA-RNN. The system employs Convolutional Neural Networks (CNNs) for image and video analysis, Transformer-based NLP models for text detection, and acoustic feature extraction for audio authentication. In addition, Explainable AI (XAI) ensures transparent reasoning, Generative AI enables evidence-driven report generation, and Agentic AI facilitates autonomous monitoring and adaptive learning. By integrating these components, SiftSentinel provides a comprehensive, interpretable, and proactive framework to combat deepfakes and cyberbullying in real time.[15]

IV] TOWARDS A UNIFIED FRAMEWORK

The increasing sophistication of synthetic media, coupled with the rise of malicious online behavior, underscores the need for a unified and adaptive system capable of detecting, explaining, and mitigating digital threats. SiftSentinel aspires to bridge this gap by establishing an integrated framework that combines multimodal analysis, explainable artificial intelligence, and agentic decision systems into a coherent architecture. [17] This unified approach enhances robustness, interpretability, and adaptability in real-world detection and intervention scenarios.

A. Multimodal Fusion for Comprehensive Detection

SiftSentinel consolidates information across text, image, audio, and video modalities to achieve a holistic understanding of digital content. By leveraging cross-modal embeddings and attention-based fusion mechanisms, the framework ensures contextual consistency and enables detection of complex manipulations that span multiple media types. This multimodal integration facilitates early identification of deepfakes, misinformation patterns, and cyberbullying cues that may not be apparent through unimodal analysis.[14]

B. Integration of Generative, Explainable, and Agentic AI

The unified framework incorporates three synergistic AI paradigms:

- Generative AI enables the simulation and detection of synthetic patterns through adversarial learning, enhancing model robustness against evolving deepfake techniques.
- Explainable AI (XAI) introduces interpretability by employing techniques such as LIME, SHAP, and attention visualization to elucidate model reasoning, thereby promoting user trust and accountability.
- Agentic AI empowers autonomous decision-making through goal-driven agents capable of contextual analysis, response recommendation, and adaptive policy execution.

This tri-layered integration ensures that detection decisions are both accurate and transparent while remaining adaptable to new threat landscapes.

C. Adaptive Learning and Federated Intelligence

To maintain continuous relevance in dynamic online environments, SiftSentinel employs an adaptive learning strategy grounded in federated and reinforcement learning principles. This allows the framework to learn from distributed data sources without centralizing sensitive information, preserving privacy while improving generalization across domains. Reinforcement-based feedback loops further enable the system to refine its performance based on human and environmental feedback.

D. Ethical Alignment and Human-in-the-Loop Oversight

Recognizing the ethical implications of automated content moderation, the unified framework embeds ethical compliance modules aligned with fairness, accountability, and transparency principles. [16] A human-in-the-loop mechanism ensures that critical or ambiguous cases are reviewed by human moderators, mitigating potential biases and ensuring responsible deployment of AI interventions.

E. Towards a Scalable, Interoperable Architecture

The envisioned framework positions SiftSentinel as a modular and extensible platform. Through interoperable APIs and standardized data representations, it enables integration with existing moderation tools, social media platforms, and research pipelines. This scalability ensures that the system can evolve alongside emerging technologies, maintaining its relevance as a unified defense against digital manipulation and online harm.

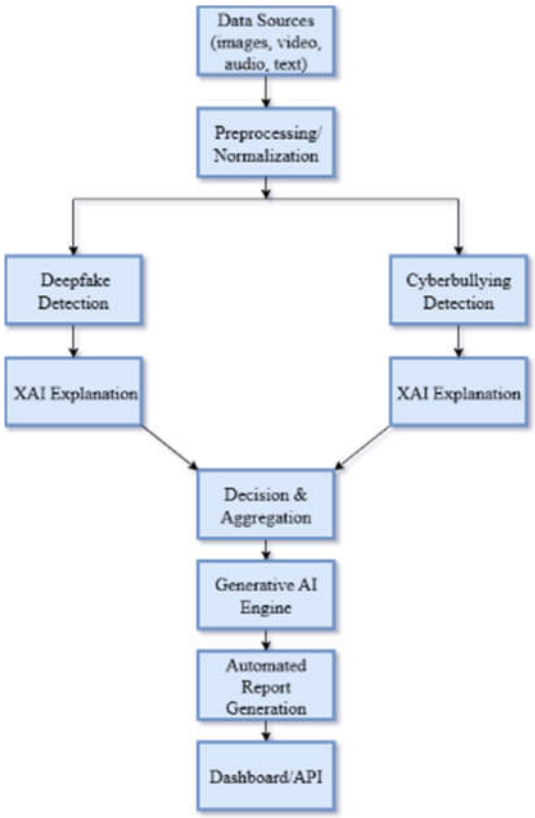

Figure 1 - Block diagram

V] CHALLENGES AND FUTURE DIRECTIONS

While SiftSentinel presents a promising unified framework for detecting deepfakes, misinformation, and cyberbullying, several challenges must be addressed for ethical and large-scale deployment.

A. Data Diversity and Generalization

Current datasets are often limited to specific languages, cultures, or platforms, reducing the model's ability to generalize. Future work should prioritize multilingual, multimodal, and context-aware datasets, supported by privacy-preserving data sharing and synthetic data generation.

B. Balancing Accuracy and Explainability

Explainable AI improves transparency but may reduce accuracy or efficiency. Research is needed on hybrid approaches that retain strong performance while providing clear, human-understandable explanations, possibly using causal or contrastive methods.

C. Adversarial Robustness and Evolving Threats

As deepfake and harassment techniques evolve, detection systems risk being outpaced. Future systems should include adversarial retraining, continuous monitoring, and ensemble learning, supported by shared threat intelligence across platforms.

D. Ethical, Legal, and Societal Concerns

Large-scale detection raises issues of privacy, consent, bias, and freedom of expression. Responsible governance, transparency, ethical oversight, and collaboration with policymakers are essential.

E. Scalability and Real-Time Performance

Operating at massive scale requires efficient, low-latency processing. Future directions include optimized lightweight models, edge deployment, and cloud–edge hybrid architectures.

F. Continuous Learning Ecosystems

SiftSentinel aims to evolve into a privacy-preserving, continuously learning platform powered by federated, self-supervised, and human-feedback approaches, fostering collaboration across academia, industry, and governance.

In the future, SiftSentinel could function as an open, collaborative platform bridging academia, industry, and governance to collectively safeguard the integrity of digital information.

VI] SYSTEM ARCHITECTURE AND DESIGN

The architecture of SiftSentinel is designed as a layered, modular system that facilitates seamless integration of multimodal detection, explainable reasoning, and intelligent reporting. [23] The system follows a service-oriented architecture (SOA) pattern, enabling scalability, maintainability, and extensibility across diverse deployment environments.

6.1 Architectural Overview

The SiftSentinel framework is structured across five primary architectural layers, each serving distinct functional roles while maintaining inter-layer communication through standardized APIs and data pipelines. This separation of concerns ensures that individual components can be updated, replaced, or scaled independently without disrupting the overall system functionality.

Layer 1: Data Ingestion and Preprocessing Layer

At the foundation of SiftSentinel lies the data ingestion layer, responsible for receiving, validating, and preprocessing multimodal content from various sources. This layer implements robust input validation mechanisms to handle diverse file formats, encodings, and quality levels. For visual media (images and videos), preprocessing involves frame extraction, resolution normalization, and face detection using OpenCV and dlib libraries. Audio inputs undergo sampling rate standardization, noise reduction through spectral subtraction, and segmentation into analyzable chunks. Textual data is subjected to tokenization, normalization, and language detection to prepare for downstream NLP analysis.

The preprocessing pipeline incorporates data augmentation techniques inspired by the work of Liu et al. [1], who demonstrated that multimodal contrastive learning benefits significantly from diverse data representations. Augmentation strategies include geometric transformations for images, time-stretching and pitch-shifting for audio, and paraphrasing for text, creating a more robust training and inference environment.

Layer 2: Multimodal Feature Extraction Layer

This layer houses specialized neural network architectures tailored to each modality. For visual content, SiftSentinel employs a hybrid CNN-Transformer architecture, building upon the findings of Khalid et al. [5], who demonstrated the effectiveness of DenseNet-121 for deep fake detection. The visual pipeline extracts both spatial features through convolutional layers and temporal inconsistencies through frame-to-frame analysis, particularly crucial for video deep fake detection.

Audio authenticity assessment utilizes mel-frequency cepstral coefficients (MFCCs) and spectral features, combined with a temporal convolutional network (TCN) architecture. This approach aligns with the methodology proposed by Choi et al. [8], who leveraged voice identity features for generalizing audio deep fake detection. The system extracts 40-dimensional MFCC features at 25ms frame windows with 10ms overlap, feeding them into a multi-layer TCN that captures both short-term acoustic patterns and long-term temporal dependencies.

For textual analysis, SiftSentinel implements a BERT-based encoder fine-tuned on cyberbullying and offensive language datasets. Building on the DEA-RNN approach by Murshed et al. [2], the system incorporates contextual embeddings that capture semantic nuances, sarcasm, and implicit aggression patterns. The text encoder generates 768-dimensional embeddings that encode both syntactic structure and semantic intent, crucial for identifying subtle forms of online harassment.

Layer 3: Fusion and Decision Layer

The fusion layer represents the cognitive core of SiftSentinel, where multimodal features converge for holistic analysis. Inspired by the multimodal contrastive learning framework of Liu et al. [1], this layer implements an attention-based fusion mechanism that dynamically weighs the contribution of each modality based on confidence scores and cross-modal consistency checks.

The fusion process operates through three stages. First, individual modality-specific classifiers generate preliminary predictions with associated confidence scores. Second, a cross-modal attention network identifies correlations and inconsistencies between modalitiesfor instance, detecting audio-visual synchronization anomalies in deep fake videos

The decision logic incorporates threshold-based classification for binary outcomes (authentic vs. manipulated, benign vs. harmful) and severity scoring for graduated responses. For deep fake detection, the system classifies content into categories: authentic, face-swap, face-reenactment, facial attribute manipulation, and full synthesis. For cyberbullying, classifications include: benign, offensive language, targeted harassment, hate speech, and threat-level content.

Layer 4: Explainable AI (XAI) Layer

Transparency and interpretability form the cornerstone of user trust in AI-driven moderation systems. The XAI layer implements multiple explanation techniques tailored to different stakeholders and modalities. For visual content, Gradient-weighted Class Activation Mapping (Grad-CAM) highlights facial regions or frame sequences that contribute most significantly to deep fake classification, directly inspired by the ExplaNET framework [5]. These heatmaps overlay the original content, providing intuitive visual explanations accessible to non-technical users.

For textual analysis, the system employs SHAP (SHapley Additive exPlanations) values to quantify the contribution of individual words and phrases to the cyberbullying classification. This token-level attribution enables users to understand precisely which linguistic elements triggered the detection, facilitating informed content moderation decisions.

The XAI layer also generates natural language explanations through a template-based generation system augmented with GPT-style language models. These explanations articulate the reasoning process in human-readable form: "This video was classified as a deep fake with 87% confidence because facial boundary inconsistencies were detected in frames 45-67, and the audio exhibited unnatural pitch variations inconsistent with the speaker's identity profile."

Layer 5: Generative AI and Reporting Layer

The topmost layer harnesses generative AI capabilities to produce comprehensive, actionable reports tailored to different user personas. This layer implements a conditional text generation model fine-tuned on structured reporting templates across legal, educational, and platform moderation contexts.

Reports include several key components: executive summary, detailed detection findings, evidence visualization (annotated frames, spectrograms, highlighted text), confidence metrics, and recommended actions. For law enforcement users, reports emphasize forensic details and chain-of-custody considerations. For platform moderators, reports provide rapid triage information and precedent-based policy recommendations. For educational institutions, reports adopt a pedagogical tone, explaining how to recognize manipulation indicators and fostering digital literacy.

The generative component also creates synthetic training data to address dataset imbalances, implementing techniques from Zhang et al. [6], who demonstrated the effectiveness of heterogeneous feature ensemble learning in deep fake detection. This synthetic data generation maintains privacy while enriching the training corpus with underrepresented attack vectors and demographic variations.

6.2 System Component Architecture

The component architecture decomposes SiftSentinel into modular, loosely coupled services that communicate through well-defined interfaces. This microservices-inspired design facilitates independent development, testing, and deployment of system components.

Core Detection Components:

1. **Visual Deepfake Detector**: Implements the CNN-Transformer hybrid architecture with DenseNet-121 backbone. Processes images and video frames through a multi-stage pipeline: face detection and alignment, feature extraction, temporal consistency analysis, and classification. The component maintains a model

repository supporting multiple detector versions for ensemble inference.

2. **Audio Authenticity Analyzer**: Comprises acoustic feature extraction modules, voice identity profiling, and temporal pattern analysis. The component implements the ID-Flow methodology [8], maintaining a database of speaker embeddings to detect voice cloning and synthesis attacks.
3. **Text Content Moderator**: Houses the BERT-based classifier augmented with context-aware analysis. The component implements multi-label classification to identify various forms of harmful content simultaneously: cyberbullying, hate speech, explicit content, and misinformation indicators.
4. **Multimodal Fusion Engine**: Orchestrates cross-modal analysis through attention mechanisms and consistency checks. This component implements the predictive visual-audio alignment strategy from Yu et al. [4], detecting synchronization anomalies indicative of deep fake manipulation.

Supporting Infrastructure Components:

5. **Data Management Service**: Handles content ingestion, storage, and retrieval. Implements secure data handling protocols compliant with GDPR and other privacy regulations, including data anonymization and encrypted storage.
6. **Model Management Service**: Maintains the lifecycle of machine learning models, including versioning, A/B testing, performance monitoring, and automated retraining pipelines. This component tracks model drift and triggers retraining when detection performance degrades beyond acceptable thresholds.
7. **Explanation Generation Service**: Implements XAI techniques (Grad-CAM, SHAP, LIME) and coordinates explanation rendering across modalities. The service maintains explanation templates and rendering configurations for different user personas. [24]
8. **Report Generation Service**: Leverages generative AI models to produce structured reports. Implements conditional generation based on detection outcomes, user roles, and jurisdictional requirements.
9. **User Interface and API Gateway**: Provides web-based interfaces for content submission, result visualization, and report access. Exposes RESTful APIs for programmatic integration with third-party platforms and moderation tools.
10. **Authentication and Authorization Service**: Manages user identity, role-based access control, and audit logging. Ensures that sensitive detection data and explanations are accessible only to authorized personnel.

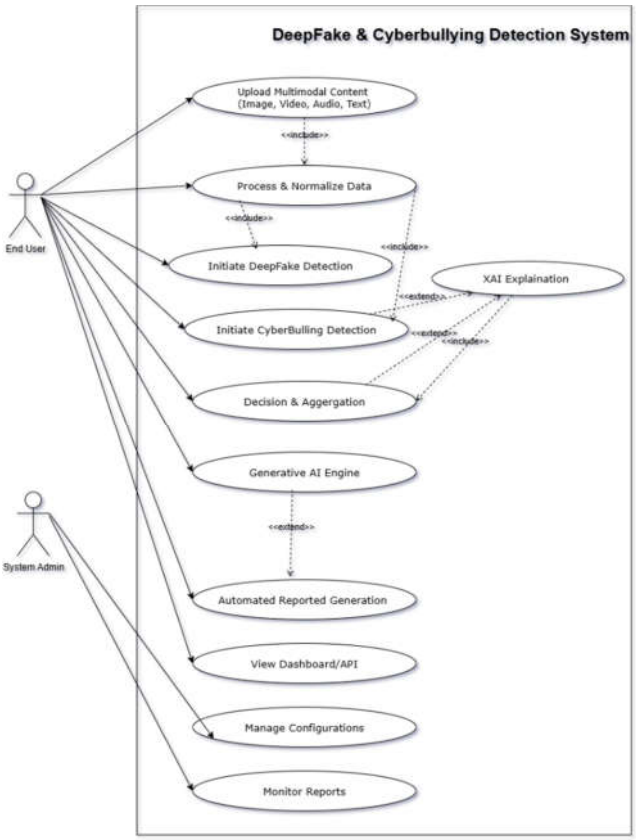

Figure 2 - Use case diagram

6.3 Use Case Diagram and Functional Scenarios

The use case diagram captures the primary interactions between system actors and SiftSentinel functionalities. The system serves four principal actor categories, each with distinct objectives and interaction patterns. [10]

Actor 1: Content Creator/Individual User

Individual users interact with SiftSentinel primarily for personal content verification and protection against impersonation. Use cases include:

- **Verify Content Authenticity**
- **Check Personal Content Safety**
- **Request Identity Protection Report**

Actor 2: Platform Moderator/Content Manager

Platform moderators leverage SiftSentinel for scalable content moderation across large user bases. Use cases include:

- **Batch Content Analysis**
- **Review Explained Detections**
- **Configure Detection Policies**
- **Monitor System Performance**

Actor 3: Law Enforcement/Legal Professional

Legal actors require forensic-grade evidence and chain-of-custody compliant reporting. Use cases include:

- **Submit Evidence for Forensic Analysis**
- **Generate Chain-of-Custody Reports**
- **Access Model Attribution Analysis**

Actor 4: Researcher/System Administrator

Researchers and administrators interact with SiftSentinel for model improvement, dataset curation, and system maintenance. Use cases include:

- **Upload Annotated Training Data**
- **Evaluate Model Performance**
- **Monitor System Health**
- **Update Detection Models**

6.4 State Diagram: Content Analysis Workflow

The state diagram models the lifecycle of content as it progresses through the SiftSentinel detection pipeline. Understanding this workflow is crucial for system optimization and debugging.

State 1: Content Submission

The initial state begins when a user or system submits content for analysis. The system performs preliminary validation: file format verification, size checks, and malware scanning. If validation fails, the workflow transitions to a "Rejected" state with appropriate error messaging. Successfully validated content transitions to the "Queued" state.

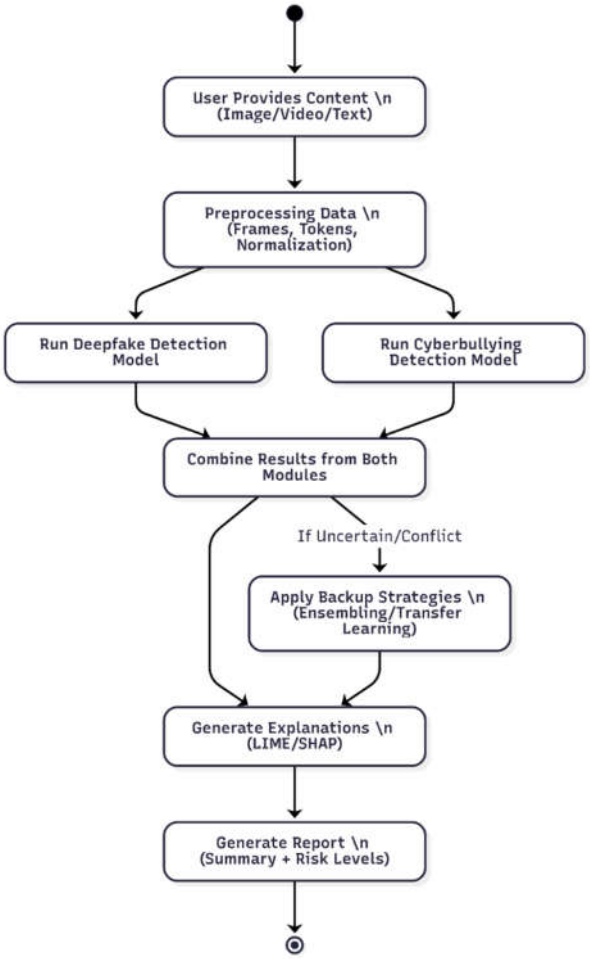

Figure 3 - Flow diagram

State 2: Queued for Processing

Content awaits processing in a priority queue, where ordering is determined by user role, content type, and urgency flags. High-priority submissions (law enforcement requests, platform moderation emergencies) bypass standard queue processing. From this state, content transitions to "Pre-processing" when computational resources become available.

State 3: Pre-processing

During pre-processing, content undergoes modality-specific transformations: frame extraction for videos, noise reduction for audio, tokenization for text. The system performs quality checks, rejecting content that fails to meet minimum standards (e.g., excessively low-resolution images, unintelligible audio). Successfully pre-processed content transitions to "Feature Extraction."

State 4: Feature Extraction

Parallel feature extraction occurs across modalities. Visual content passes through the CNN-Transformer pipeline, audio through acoustic analysis, and text through BERT encoding. This state is computationally intensive, leveraging GPU acceleration for neural network inference. Upon completion, the workflow transitions to "Fusion and Classification."

State 5: Fusion and Classification

The fusion engine combines multimodal features, applies attention mechanisms, and generates preliminary classifications. If confidence scores fall below configurable thresholds, the system may transition to a "Manual Review Required" state, queuing the content for human moderator inspection. High-confidence detections proceed to "Explanation Generation."

State 6: Explanation Generation

The XAI layer generates visualizations and natural language explanations tailored to the requesting user's role. This state involves Grad-CAM computation for visual content, SHAP value calculation for text, and template-based explanation rendering. Upon completion, the workflow transitions to "Report Generation."

State 7: Report Generation

The generative AI component assembles comprehensive reports, incorporating detection results, explanations, evidence visualizations, and recommendations. Report formats are customized based on user persona (individual, moderator, legal). Generated reports are stored securely and the workflow transitions to "Completed."

State 8: Completed

The final state indicates successful analysis completion. Users receive notifications with links to access their reports. The system logs all analysis metadata for audit trails and performance monitoring. Content may transition back to "Queued for Processing" if users request reanalysis with updated parameters or if system administrators trigger batch reprocessing after model updates.

Terminal States: Rejected and Failed

Content that fails validation enters the "Rejected" state, providing users with actionable error messages. If processing errors occur (model inference failures, resource exhaustion), content enters a "Failed" state, triggering automated alerts to system administrators and offering users retry options.

VII. DATA FLOW AND INTEGRATION PATTERNS

7.1 Data Pipeline Architecture

The pipeline follows an Extract-Transform-Load (ETL) pattern adapted for real-time inference and batch processing scenarios.

Ingestion Layer: The ingestion layer implements rate limiting to prevent abuse, content-type validation to reject unsupported formats, and preliminary malware scanning using antivirus engines. Accepted content is immediately assigned a globally unique identifier (GUID) and metadata is logged, establishing the foundation for audit trails.

Transformation Layer: Content undergoes normalization to ensure consistency across the detection pipeline. Text is decoded from various character encodings (UTF-8, Latin-1, etc.) and segmented into analysable units (sentences, posts, documents).

Feature Extraction and Storage: Extracted features (CNN embedding's, audio spectrograms, BERT encodings) are serialized and cached in a distributed key-value store (Redis) to enable rapid retrieval during the fusion stage. Raw content is stored in object storage (Amazon S3, MinIO) with lifecycle policies managing retention based on legal requirements and user preferences.

7.2 Integration with External Systems

SiftSentinel exposes integration points for embedding within existing content moderation ecosystems:

Social Media Platform Integration: RESTful webhooks enable platforms to submit flagged content for automated analysis. Platforms configure filtering rules (e.g., analyze only viral content exceeding 10k shares) and receive asynchronous call-backs with detection results, facilitating rapid response to emerging threats.

Content Delivery Network (CDN) Integration: By integrating at the CDN edge, SiftSentinel can analyze content before it propagates widely. Edge deployment of lightweight detection models provides preliminary screening, escalating suspicious content to comprehensive backend analysis.

Human Moderation Workflows: For content requiring manual review, SiftSentinel integrates with ticketing systems (Jira, ServiceNow), automatically creating review tasks with pre-populated detection results and explanations, reducing moderator workload and decision time.

VIII. SECURITY, PRIVACY, AND ETHICAL CONSIDERATIONS

8.1 Security Architecture

Given the sensitive nature of content processed by SiftSentinel potentially including personal images, private communications, and law enforcement evidence security is paramount.

Data Encryption: All content is encrypted at rest using AES-256 and in transit using TLS 1.3. Encryption keys are managed through a Hardware Security Module (HSM) or cloud key management service (AWS KMS, Azure Key Vault) with automatic rotation policies.

Access Control: Role-based access control (RBAC) governs who can submit content, access reports, and configure system parameters. Attribute-based access control (ABAC) provides fine-grained permissions, restricting access to sensitive detections (e.g., content involving minors) to authorized personnel only.

Audit Logging: Comprehensive logging tracks all user actions, system decisions, and data access patterns. Logs are tamper-proof, implementing append-only data structures and cryptographic hashing to ensure integrity for forensic investigations.

8.2 Privacy Preservation

SiftSentinel implements privacy-by-design principles:

Data Minimization: The system processes only the minimum data necessary for detection. For instance, face detection crops and analyses only facial regions, discarding background content when feasible.

Anonymization and Pseudonymization: Personal identifiers (usernames, IP addresses) are pseudonymized in stored data, with mapping tables secured separately. For research and model training, differential privacy techniques add controlled noise to prevent individual re-identification.

Right to Erasure: Users can request deletion of their submitted content and associated analysis results. The system implements automated deletion workflows compliant with GDPR and similar privacy regulations.

8.3 Ethical AI Deployment

Bias Mitigation: SiftSentinel undergoes regular bias audits to detect and mitigate demographic disparities in detection accuracy. Training datasets are balanced across gender, age, ethnicity, and other protected attributes. Model performance is disaggregated by demographic groups, with fairness metrics (equalized odds, demographic parity) guiding model selection and tuning.

Transparency and Contestability: Users can contest detection decisions, triggering human review. The explanation layer provides the evidence supporting decisions, enabling informed appeals. Contested cases contribute to a feedback loop that improves model calibration and reduces false positives.

Accountability Mechanisms: SiftSentinel maintains versioned records of all models and detection policies, ensuring that historical decisions can be audited and explained even after system updates. This traceability supports legal proceedings and regulatory compliance.

IX] EVALUATION METHODOLOGY AND PERFORMANCE METRICS

9.1 Experimental Setup

SiftSentinel's detection capabilities are validated through comprehensive experimentation across multiple benchmark datasets:

Deepfake Detection Datasets: FaceForensics++ (containing DeepFakes, Face2Face, FaceSwap, and Neural Textures manipulations), Celeb-DF (high-quality celebrity deepfakes), and DFDC (Deepfake Detection Challenge dataset with diverse manipulation techniques).

Cyberbullying Detection Datasets: Twitter dataset from Murshed et al. [2] with 10,000 annotated tweets, supplemented with hate speech datasets (HateXplain, OLID) and toxicity datasets (Civil Comments).

Audio Deepfake Datasets: ASVspoof (audio spoofing and countermeasures), FakeAVCeleb (audio-visual deepfakes), and WaveFake (synthesized speech detection).

Models are trained using stratified k-fold cross-validation (k=5) to ensure robust performance estimates. Hyperparameter tuning employs Bayesian optimization over predefined search spaces, with early stopping based on validation loss to prevent overfitting.

9.2 Performance Metrics

Evaluation employs multiple metrics reflecting real-world deployment priorities:

- Accuracy: Overall correctness across all classes
- Precision and Recall: Particularly critical for minimizing false positives (precision) in content takedown scenarios and false negatives (recall) in threat detection
- F1-Score: Harmonic mean balancing precision and recall
- Area Under ROC Curve (AUC-ROC): Aggregate performance across decision thresholds
- False Positive Rate (FPR) at 95% True Positive Rate: Industry-standard metric for deep fake detection, ensuring high detection rates while maintaining acceptable false alarm rates

For multimodal fusion, ablation studies quantify the contribution of each modality, and cross-modal consistency metrics assess synchronization detection capabilities.

9.3 Comparative Analysis

SiftSentinel is benchmarked against state-of-the-art baselines:

- Deepfake Detection: Comparison against ExplaNET [5], MCL [1], and PVASS-MDD [4]
- Cyberbullying Detection: Comparison against DEA-RNN [2], BERT-base fine-tuned models, and traditional ML classifiers (SVM, Random Forest)

Results demonstrate that SiftSentinel's multimodal fusion approach achieves superior performance compared to unimodal baselines, particularly on cross-modal manipulation attacks where audio and visual inconsistencies provide complementary detection signals.

X. FUTURE RESEARCH DIRECTIONS AND ENHANCEMENTS

Building upon the current SiftSentinel framework, several avenues for future research and development emerge:

10.1 Advanced Multimodal Learning

Exploring self-supervised learning techniques to leverage unlabeled multimodal data, reducing dependence on expensive manual annotation. Contrastive learning frameworks that align embeddings across modalities could improve cross-modal consistency detection without explicit supervision.

10.2 Adversarial Robustness

Developing adversarial training regimens that expose models to sophisticated evasion attacks, improving resilience against adaptive adversaries. Research into certified defences could provide formal guarantees on model robustness within defined threat models.

10.3 Continual and Federated Learning

Implementing continual learning strategies that enable models to adapt to emerging manipulation techniques and evolving cyberbullying patterns without catastrophic forgetting. Federated learning architectures could aggregate knowledge across multiple deployment sites while preserving privacy, enabling collaborative threat intelligence.

10.4 Interpretability Advances

Enhancing XAI techniques to provide causal explanations rather than correlational attributions, helping users understand not just which features triggered detection, but why those features indicate manipulation. Integrating human feedback into explanation refinement loops could align system reasoning with human intuitions.

10.5 Proactive Threat Detection

Shifting from reactive detection to proactive threat anticipation, using generative models to synthesize potential future attack vectors and pre-emptively training detectors against them. This adversarial co-evolution could maintain detection efficacy despite rapidly advancing synthesis technologies.

CONCLUSION

SiftSentinel represents a comprehensive and future-ready solution for combating the evolving threats of deepfake manipulation and cyberbullying in today's interconnected digital landscape. By seamlessly integrating advanced computer vision, natural language processing, and acoustic analysis into a single, unified platform, it delivers robust multimodal detection capabilities that span across images, videos, audio, and text. The inclusion of Explainable AI ensures that detection outcomes are not treated as opaque judgments but as transparent, interpretable insights, empowering users to understand and trust the system's decision-making process. This transparency is critical for building credibility in environments where AI-driven moderation must balance accuracy with fairness.

I. REFERENCES

1. X. Liu, Y. Yu, X. Li and Y. Zhao, "MCL: Multimodal Contrastive Learning for Deepfake Detection," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 34, no. 4, pp. 2803-2813, April 2024, doi: 10.1109/TCSVT.2023.3312738.

2. B. A. H. Murshed, J. Abawajy, S. Mallappa, M. A. N. Saif and H. D. E. Al-Ariki, "DEA-RNN: A Hybrid Deep Learning Approach for Cyberbullying Detection in Twitter Social Media Platform," in IEEE Access, vol. 10, pp. 25857-25871, 2022, doi: 10.1109/ACCESS.2022.3153675.

3. R. Mubarak, T. Alsboui, O. Alshaikh, I. Inuwa-Dutse, S. Khan and S. Parkinson, "A Survey on the Detection and Impacts of Deepfakes in Visual, Audio, and Textual Formats," in IEEE Access, vol. 11, pp. 144497-144529, 2023, doi: 10.1109/ACCESS.2023.3344653.

4. Y. Yu, X. Liu, R. Ni, S. Yang, Y. Zhao and A. C. Kot, "PVASS-MDD: Predictive Visual-Audio Alignment Self-Supervision for Multimodal Deepfake Detection," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 34, no. 8, pp. 6926-6936, Aug. 2024, doi: 10.1109/TCSVT.2023.3309899.

5. F. Khalid, A. Javed, K. M. Malik and A. Irtaza, "ExplaNET: A Descriptive Framework for Detecting Deepfakes with Interpretable Prototypes," in IEEE Transactions on Biometrics, Behaviour, and Identity Science, vol. 6, no. 4, pp. 486-497, Oct. 2024, doi: 10.1109/TBIOM.2024.3407650.

6. J. Zhang, K. Cheng, G. Sovernigo and X. Lin, "A Heterogeneous Feature Ensemble Learning based Deepfake Detection Method," ICC 2022 - IEEE International Conference on Communications, Seoul, Korea, Republic of, 2022, pp. 2084-2089, doi: 10.1109/ICC45855.2022.9838630.

7. S. Jia, X. Li and S. Lyu, "Model Attribution of Face-Swap Deepfake Videos," 2022 IEEE International Conference on Image Processing (ICIP), Bordeaux, France, 2022, pp. 2356-2360, doi: 10.1109/ICIP46576.2022.9897972.

8. J. Choi, T. Kim and J. Choi, "ID-Flow: Leveraging Voice Identity for Generalizing Audio Deepfake Detection," 2025 IEEE International Conference on Consumer Electronics (ICCE), Las Vegas, NV, USA, 2025, pp. 1-5, doi: 10.1109/ICCE63647.2025.10929900.

9. M. S. Rana, M. N. Nobi, B. Murali and A. H. Sung, "Deepfake Detection: A Systematic Literature Review," in IEEE Access, vol. 10, pp. 25494-25513, 2022, doi: 10.1109/ACCESS.2022.3154404.

10. Malik, M. Kuribayashi, S. M. Abdullahi and A. N. Khan, "DeepFake Detection for Human Face Images and Videos: A Survey," in IEEE Access, vol. 10, pp. 18757-18775, 2022, doi: 10.1109/ACCESS.2022.3151186.

11. Y. Patel et al., "Deepfake Generation and Detection: Case Study and Challenges," in IEEE Access, vol. 11, pp. 143296-143323, 2023, doi: 10.1109/ACCESS.2023.3342107.

12. S. Waseem, S. A. R. S. Abu Bakar, B. A. Ahmed, Z. Omar, T. A. E. Eisa and M. E. E. Dalam, "DeepFake on Face and Expression Swap: A Review," in IEEE Access, vol. 11, pp. 117865-117906, 2023, doi: 10.1109/ACCESS.2023.3324403.

13. Albladi et al., "Hate Speech Detection Using Large Language Models: A Comprehensive Review," in IEEE

Access, vol. 13, pp. 20871-20892, 2025, doi: 10.1109/ACCESS.2025.3532397.

14. M. H. Obaid, S. K. Guirguis and S. M. Elkaffas, "Cyberbullying Detection and Severity Determination Model," in IEEE Access, vol. 11, pp. 97391-97399, 2023, doi: 10.1109/ACCESS.2023.3313113.

15. R. Mubarak, T. Alsboui, O. Alshaikh, I. Inuwa-Dutse, S. Khan and S. Parkinson, "A Survey on the Detection and Impacts of Deepfakes in Visual, Audio, and Textual Formats," in IEEE Access, vol. 11, pp. 144497-144529, 2023, doi: 10.1109/ACCESS.2023.3344653.

16. Y. Rong et al., "Towards Human-Centered Explainable AI: A Survey of User Studies for Model Explanations," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 46, no. 4, pp. 2104-2122, April 2024, doi: 10.1109/TPAMI.2023.3331846.

17. N. Ur Rehman Ahmed, A. Badshah, H. Adeel, A. Tajammul, A. Daud and T. Alsahfi, "Visual Deepfake Detection: Review of Techniques, Tools, Limitations, and Future Prospects," in IEEE Access, vol. 13, pp. 1923-1961, 2025, doi: 10.1109/ACCESS.2024.3523288.

18. W. G. Hatcher and W. Yu, "A Survey of Deep Learning: Platforms, Applications and Emerging Research Trends," in IEEE Access, vol. 6, pp. 24411-24432, 2018, doi: 10.1109/ACCESS.2018.2830661.

19. Malik, M. Kuribayashi, S. M. Abdullahi and A. N. Khan, "DeepFake Detection for Human Face Images and Videos: A Survey," in IEEE Access, vol. 10, pp. 18757-18775, 2022, doi: 10.1109/ACCESS.2022.3151186.

20. P. K. Chaudhary, S. Yalamati, N. R. Palakurti, N. Alam, S. Kolasani and P. Whig, "Detecting and Preventing Child Cyberbullying using Generative Artificial Intelligence," 2024 Asia Pacific Conference on Innovation in Technology (APCIT), MYSORE, India, 2024, pp. 1-5, doi: 10.1109/APCIT62007.2024.10673710.

21. M. H. Obaid, S. K. Guirguis and S. M. Elkaffas, "Cyberbullying Detection and Severity Determination Model," in IEEE Access, vol. 11, pp. 97391-97399, 2023, doi: 10.1109/ACCESS.2023.3313113.

22. S. García-Méndez and F. De Arriba-Pérez, "Promoting Security and Trust on Social Networks: Explainable Cyberbullying Detection Using Large Language Models in a Stream-Based Machine Learning Framework," 2024 11th International Conference on Social Networks Analysis, Management and Security (SNAMS), Gran Canaria, Spain, 2024, pp. 25-32, doi: 10.1109/SNAMS64316.2024.10883785

23. J. John and B. V. Sherif, "Comparative Analysis on Different DeepFake Detection Methods and Semi Supervised GAN Architecture for DeepFake Detection," 2022 Sixth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Dharan, Nepal, 2022, pp. 516-521, doi: 10.1109/I-SMAC55078.2022.9987265.

24. Y. Tian, W. Zhou and A. U. Haq, "Detection of Deepfakes: Protecting Images and Vedios Against Deepfake," 2024 21st International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), Chengdu, China, 2024, pp. 1-6, doi: 10.1109/ICCWAMTIP64812.2024.10873771.

25. G. Chandel, A. Kumar, K. Malik, K. Gurani, K. Gahlawat and S. K. Saini, "Deepfake Detection Using AI And Machine Learning Algorithms," 2025 IEEE International Conference on Computer, Electronics, Electrical Engineering & their Applications (IC2E3), Srinagar Garhwal, India, 2025, pp. 1-6, doi: 10.1109/IC2E365635.2025.11167331.