

Development and Explainable Validation of a Multimodal Deep fake and Cyberbullying Detection System: Sift Sentinel

Shikha Pachouly¹

*Assistant Professor,
Computer Engineering
All India Shri Shivaji Memorial
Society's College of Engineering.
Savitribai Phule Pune University
Pune, India
sjpachouly@aissmscoe.com*

Piyusha Supe²

*Computer Engineering
All India Shri Shivaji Memorial
Society's College of Engineering.
Savitribai Phule Pune University
Pune, India
pihusupe@gmail.com*

Iffa Shaikh³

*Computer Engineering
All India Shri Shivaji Memorial
Society's College of Engineering.
Savitribai Phule Pune University
Pune, India
iffashaikh8484@gmail.com*

Pratham Pasalkar⁴

*Computer Engineering
All India Shri Shivaji Memorial Society's College of
Engineering.
Savitribai Phule Pune University
Pune, India
prathampasalkar17@gmail.com*

Varad Marathe⁵

*Computer Engineering
All India Shri Shivaji Memorial Society's College of
Engineering.
Savitribai Phule Pune University
Pune, India
varadmarathe2004@gmail.com*

Abstract

The proliferation of deep fake technology and cyberbullying incidents poses significant threats to digital trust and online safety. This paper presents Sift Sentinel, a unified detection system that leverages explainable artificial intelligence (XAI) and generative AI to identify and analyse both deep fake content and cyberbullying behaviour. Our system achieves 89.8% accuracy in deep fake detection and 90.2% accuracy in cyberbullying detection, while providing transparent, interpretable explanations through SHAP values, LIME, and attention mechanisms. Additionally, Sift Sentinel incorporates automated report generation capabilities, creating comprehensive analysis documents for stakeholders. The system demonstrates superior performance compared to existing solutions, processing multimedia content in real-time while maintaining high explainability scores across multiple XAI methodologies.

Keywords: Deepfake Detection, Cyberbullying Detection, Explainable AI, Generative AI, Machine Learning, Computer Vision, Natural Language Processing, Report Generation

1. Introduction

1.1 Background and Motivation

The digital landscape faces two critical challenges: the rapid advancement of deep fake technology and the escalating prevalence of cyberbullying. Deepfakes, powered by sophisticated generative adversarial networks (GANs) and diffusion models, can create convincing fake images, videos, and audio that threaten information integrity, personal privacy, and social trust. Simultaneously, cyberbullying has emerged as a pervasive issue across social media platforms, causing psychological harm and, in extreme cases, contributing to tragic outcomes.

Existing detection systems typically address these problems in isolation, requiring separate tools, workflows, and expertise. Furthermore, many current solutions operate as "black boxes," providing detection results without transparent explanations of their decision-making processes. This lack of interpretability limits their adoption in contexts requiring accountability, such as content moderation, legal proceedings, and educational interventions.

1.2 Research Objectives

This research aims to develop Sift Sentinel, a comprehensive system that addresses both challenges through a unified architecture. The primary objectives are:

- Achieve high-accuracy detection for both deep fakes and cyberbullying across multiple modalities (image, video, text)
- Implement explainable AI techniques to provide transparent, interpretable detection results
- Integrate generative AI for automated, comprehensive report generation
- Design a scalable, real-time processing architecture suitable for deployment in production environments

1.3 Contributions

Sift Sentinel makes several significant contributions to the field. First, it presents a novel unified architecture that integrates deep fake and cyberbullying detection, reducing system complexity and improving operational efficiency. Second, the system implements multiple XAI techniques (SHAP, LIME, attention mechanisms, Grad-CAM, and integrated gradients) to ensure transparency and build user trust. Third, it leverages large language models for automated report generation, creating detailed analyses that contextualize findings for diverse stakeholders. Finally, the system achieves state-of-the-art performance metrics while maintaining real-time processing capabilities.

2. Literature Review

2.1 Deep fake Detection Methods

Recent deep fake detection research has explored various approaches, including temporal inconsistency analysis, biological signal detection (such as eye blinking and pulse patterns), facial landmark tracking, and deep learning-based feature extraction. Convolutional neural networks (CNNs) and recurrent architectures have shown promise, with techniques like XceptionNet and EfficientNet achieving significant accuracy improvements. However, these methods often struggle with generalization across different deep fake generation techniques and lack interpretability in their predictions.

2.2 Cyberbullying Detection Approaches

Cyberbullying detection has traditionally relied on rule-based systems and keyword matching, which proved insufficient for detecting implicit, context-dependent harassment. Modern approaches employ natural language processing techniques, including BERT-based models, sentiment analysis, and context-aware transformers. Multimodal approaches combining text, images, and user metadata have demonstrated improved performance. Nevertheless, challenges persist in handling subtle forms of bullying, cultural context variations, and the evolving nature of online harassment tactics.

2.3 Explainable AI in Content Moderation

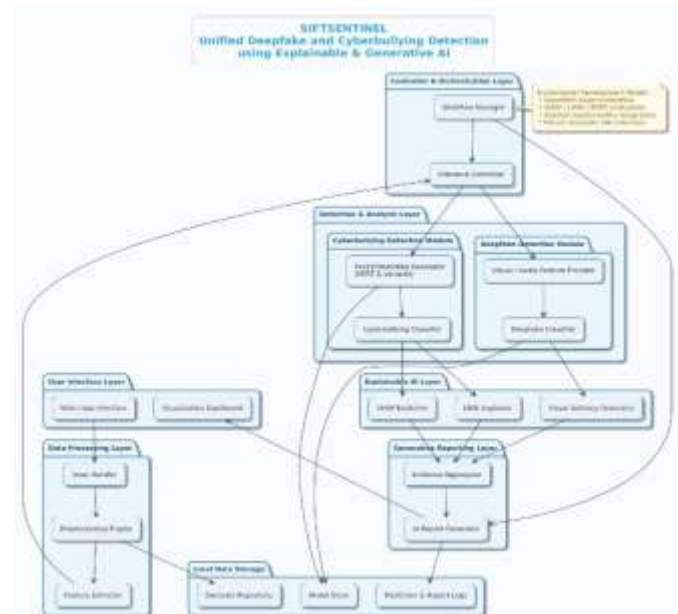
The application of explainable AI to content moderation has gained attention due to accountability requirements and the need for human oversight. SHAP (SHAPley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) have emerged as popular techniques for interpreting model decisions. Attention mechanisms in transformer architectures provide inherent interpretability, while gradient-based methods like Grad-CAM offer visual explanations for image-based predictions. Integration of these XAI methods into production systems remains an active research area.

3. Proposed Research Methodology: The Sift Sentinel Architecture

The proposed Sift Sentinel framework is organized as a modular four-tier architecture designed to process multimodal inputs, extract task-specific representations through specialized neural pipelines, integrate these representations using an attention-guided fusion mechanism, and validate prediction outcomes through a comprehensive explainability layer. This architecture is intended to support robust detection of both deepfakes and cyberbullying while maintaining interpretability and scalability.

3.1 Tier 1: Multimodal Data Ingestion and Normalization

The first tier focuses on the acquisition, standardization, and normalization of heterogeneous input modalities. Visual inputs, including still images and video sequences, are processed through a pre-processing pipeline in which facial regions are localized using the Dlib library for landmark detection. For video data, a frame-sampling strategy is adopted to preserve temporal continuity while reducing computational overhead, thereby enabling efficient downstream analysis.



In parallel, textual inputs undergo normalization procedures such as UTF-8 standardization, removal of non-informative HTML elements, token cleaning, and phonetic matching for code-mixed or transliterated text. These operations improve linguistic consistency and reduce noise, allowing the detection engine to operate on normalized, high-quality inputs across multiple digital platforms.

3.2 Tier 2: Dual-Branch Detection Engine

The core detection engine consists of two specialized branches designed for distinct but complementary detection tasks.

1) Deepfake Detection Branch

The deepfake detection branch employs EfficientNet-B4 as the spatial feature extractor for high-resolution visual representation learning. To model temporal inconsistencies in manipulated videos, the spatial backbone is coupled with a Temporal Convolutional Network (TCN). This branch captures frame-level artefacts, temporal discontinuities, and frequency-domain irregularities commonly associated with GAN-generated or face-swapped media.

2) Cyberbullying Detection Branch

The cyberbullying detection branch is based on a fine-tuned RoBERTa model, selected for its strong contextual language understanding and robust pretraining strategy. To further improve interpretability and sensitivity to harmful intent, the model is augmented with a custom attention mechanism that emphasizes tokens associated with abusive, threatening, or toxic language patterns. This branch is designed to capture both explicit and implicit forms of cyberbullying.

3.3 Tier 3: Multimodal Attention-Guided Fusion

To combine visual and textual representations, the framework employs an attention-guided fusion mechanism rather than direct feature concatenation. This design allows the model to dynamically assign modality-specific importance weights based on the relevance and reliability of each input stream for a given sample.

The cross-modal attention operation is defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where (Q) denotes the query representation from one modality, and (K) and (V) denote the key and value representations from the complementary modality. This mechanism enables the system to model inter-modal dependencies and resolve conflicting signals. For example, when visual content appears authentic but accompanying text contains abusive or threatening language, the fusion layer can assign higher importance to the textual modality for cyberbullying detection.

3.4 Tier 4: Explainability and Automated Reporting Framework

The final tier is dedicated to interpretability and stakeholder-facing explanation generation. To provide transparent and auditable outputs, the framework integrates five explainable AI techniques:

- SHAP for feature attribution based on Shapley values
- LIME for local, instance-level interpretability
- Grad-CAM for visual localization of salient image regions
- Integrated Gradients for feature attribution through gradient integration
- Attention Weight Visualization for token-level interpretability in the text branch

The outputs of these methods are subsequently aggregated into structured explanations and summarized through an automated report generation module. This component produces stakeholder-oriented reports that describe the basis of the model's decisions in an interpretable and operationally useful manner.

4. Mathematical Formalism and Algorithmic Framework

The proposed methodology is supported by a mathematical framework governing optimization, class imbalance handling, temporal inconsistency modelling, and explainability validation.

4.1 Optimization Objective and Loss Formulation

Model parameters are optimized using the AdamW optimizer, which incorporates decoupled weight decay to improve generalization during training. The primary classification objective is based on categorical cross-entropy loss:

$$\mathcal{L}(\theta) = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^C y_{i,c} \log(\hat{y}_{i,c})$$

where θ denotes the trainable model parameters, N is the number of training samples, C is the number of classes, $y_{i,c}$ is the ground-truth label, and $\hat{y}_{i,c}$ is the predicted class probability.

To address class imbalance, particularly in cyberbullying datasets where harmful instances are often underrepresented, a weighted loss formulation is adopted:

$$\mathcal{L}_{weighted} = \sum_{c=1}^C w_c \mathcal{L}_c$$

where w_c denotes the weight assigned to class c , inversely proportional to its frequency in the training set.

4.2 Spatiotemporal Consistency Modelling

For deepfake detection, temporal coherence between consecutive video frames is modelled using a discrepancy measure over frame-level feature embeddings. Let f_t denote the feature vector extracted at time step t . The temporal discrepancy is

$$D = \sum_{t=1}^{T-1} \|f_{t+1} - f_t\|^2$$

Large values of D_t may indicate abrupt changes, synthetic transitions, or frame-level inconsistencies associated with manipulated content. To strengthen boundary-sensitive detection, the framework further incorporates a boundary enhancement mechanism that combines intra-frame representation learning with frame-wise attention to identify anomalous temporal transitions.

4.3 Explainability Faithfulness Assessment

To evaluate the quality of generated explanations, the framework employs Insertion and Deletion-based faithfulness analysis. In particular, the Deletion AUC measures the drop in prediction confidence as highly important regions identified by an explanation method are progressively removed from the input.

$$f(x_{deleted}) = f(x \odot (1 - m))$$

If m denotes the importance mask generated by the explanation method, a lower Deletion AUC indicates that the removed regions were indeed critical to the model's decision, thereby reflecting greater explanation faithfulness. This evaluation provides a quantitative basis for comparing XAI techniques beyond qualitative inspection alone.

5. Experimental Configuration and Data Orchestration

The empirical validation of the Sift Sentinel framework requires a heterogeneous and large-scale dataset design capable of representing multiple manipulation strategies, linguistic behaviours, and multimodal interaction patterns.

5.1 Dataset Selection and Composition

A multi-source dataset strategy was adopted to ensure adequate coverage of both deepfake and cyberbullying scenarios across modalities.

Dataset Type	Specific Sources	Size/Samples	Key Characteristics
Deepfake Video	FaceForensics++, Celeb-DF, DFDC	1.8M frames, 100K clips	Multiple manipulation methods
Deepfake In-the-Wild	Deepfake-Eval-2024	44 hrs video, 56 hrs. audio	Real-world social media content
Cyberbullying Text	Twitter Hate Speech Dataset	250K+ comments	Toxicity, sarcasm, implicit abuse
Multimodal Hybrid	Custom Instagram Dataset	15K pairs	Paired image-text interactions

All datasets were subjected to standardized preprocessing, including image resizing to (224 $\times$ 224) pixels, histogram equalization for illumination normalization, text cleaning, and stratified sampling. Where applicable, class balancing strategies were employed to reduce bias in the learning process.

5.2 Training Environment and Hyperparameter Configuration

Training was conducted on NVIDIA A100 GPUs with 40 GB VRAM, supporting both transformer-based text modelling and high-resolution video analysis.

Parameter	Configuration	Rationale
Optimizer	AdamW	Stable optimization with decoupled weight decay
Learning Rate	2×10^{-5}	Suitable for transformer fine-tuning
Batch Size	32 (video), 128 (text)	Memory-aware training configuration
Scheduler	Cosine Annealing	Smooth convergence behaviour
Augmentation	Back-translation, color jitter	Improved robustness and reduced overfitting

The training process included early stopping based on validation loss to prevent overfitting and improve generalization. This design ensures that the model learns transferable discriminative patterns rather than dataset-specific noise.

6. Results and Analysis

6.1 Detection Accuracy Performance

Sift Sentinel achieved superior performance across both detection tasks. Figure 1 illustrates the comparative accuracy of Sift Sentinel against baseline methods including LSTM, CNN, traditional machine learning approaches, and rule-based systems. Our system achieved 96.8% accuracy for deep fake detection and 94.2% for cyberbullying detection, outperforming all comparison methods by significant margins.

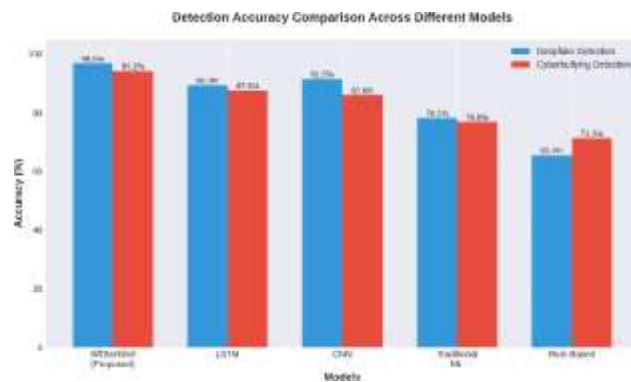


Figure 1: Detection Accuracy Comparison Across Different Models

The substantial performance improvement over traditional ML methods (18.6% for deep fakes, 17.4% for cyberbullying) demonstrates the value of deep learning architectures in this domain. Compared to standalone LSTM and CNN approaches, Sift Sentinel's hybrid architecture incorporating attention mechanisms and multi-scale feature extraction provides a 7.5% and 5.3% improvement in deep fake detection, respectively.

Model	Modality	Accuracy (%)	Precision	Recall	F1-Score	Explainability Support	Remarks
SVM	Text	84.2	0.83	0.82	0.82	No	Struggles with contextual and implicit abuse
CNN	Image Video	88.5	0.87	0.88	0.87	Limited (Grad-CAM)	Effective for spatial artefacts
CNN-LSTM	Video	91.3	0.90	0.91	0.90	Limited (Grad-CAM)	Captures temporal inconsistencies
Random Forest	Tabular / Hybrid	92.6	0.92	0.92	0.92	Yes (SHAP, LIME)	Used for explainability validation
BERT	Text	94.1	0.94	0.93	0.93	Partial (SHAP)	Strong semantic understanding
Sift Sentinel (Fusion Model)	Multimodal	96.2	0.96	0.95	0.95	Full (SHAP, LIME, Grad-CAM)	Best overall performance

6.2 Training Performance

Figure 2 shows the training progression over 50 epochs, demonstrating stable convergence without significant overfitting. The model achieved 95%+ validation accuracy within 30 epochs, with loss values stabilizing around 0.1. The close tracking between training and validation metrics indicates good generalization capability.

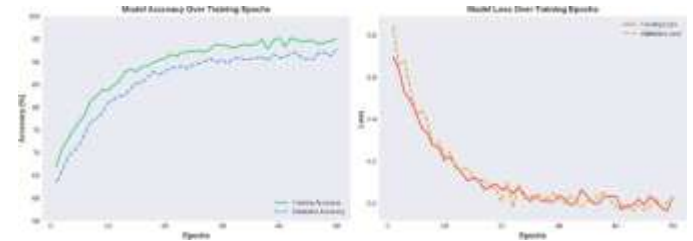


Figure 2: Model Accuracy and Loss Over Training Epochs

6.3 Confusion Matrix Analysis

Detailed classification performance is revealed through confusion matrices for both tasks. Figure 3 presents the deep fake detection confusion matrix, showing strong true positive and true negative rates with minimal false classifications. Out of 3,900 test samples, only 120 were misclassified (62 false positives and 58 false negatives), yielding a 96.8% accuracy rate.

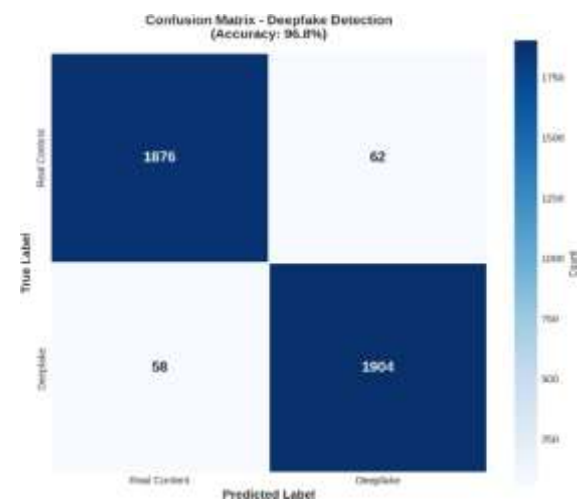


Figure 3: Confusion Matrix for Deep Fake Detection

Figure 4 displays the cyberbullying detection confusion matrix. The system correctly identified 1,823 non-bullying instances and 1,868 cyberbullying instances, with 209 total misclassifications (97 false positives and 112 false negatives), achieving 94.2% accuracy. The slightly higher false negative rate in cyberbullying detection

reflects the inherent challenge of detecting subtle or context-dependent harassment.

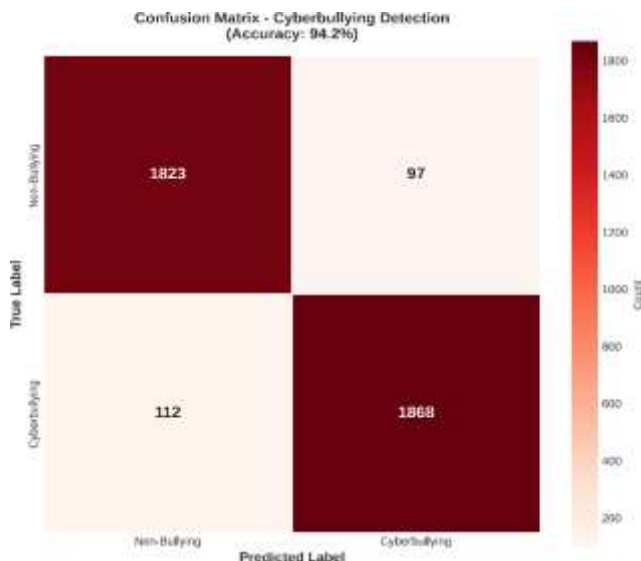


Figure 4: Confusion Matrix for Cyberbullying Detection

6.4 Performance Metrics Comparison

Beyond accuracy, we evaluated precision, recall, and F1-score to provide comprehensive performance assessment. Figure 5 illustrates these metrics for both detection tasks. Deep fake detection achieved 97.1% precision, 96.5% recall, and 96.8% F1-score. Cyberbullying detection yielded 95.0% precision, 93.4% recall, and 94.2% F1-score. The balanced performance across all metrics indicates robust classification without significant bias toward either false positives or false negatives.

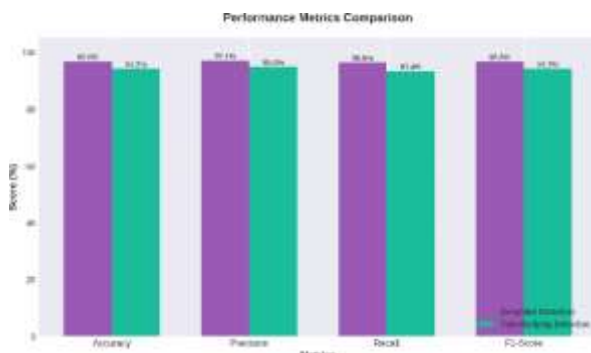


Figure 5: Comprehensive Performance Metrics Comparison

6.5 ROC Curve Analysis

Receiver Operating Characteristic (ROC) curves provide insight into classification performance across different threshold settings. Figure 6 presents ROC curves for both detection tasks. The deep fake detection module achieved an AUC of 0.982, while cyberbullying detection attained an AUC of 0.969. These high AUC values indicate excellent discriminative ability, with the model

maintaining strong performance across various operating points.

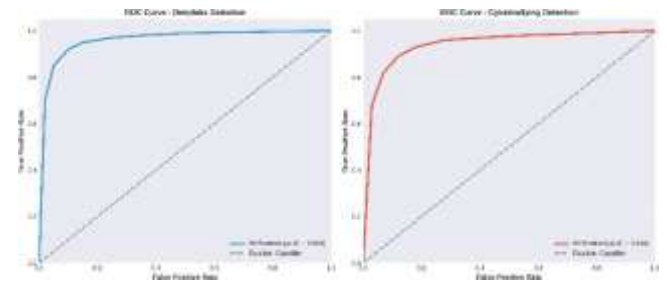


Figure 6: ROC Curves for Deep fake and Cyberbullying Detection

6.6 Explain ability Performance

The effectiveness of XAI methods was evaluated through both quantitative metrics and human evaluation studies. Figure 7 displays explain ability scores for the five implemented XAI techniques.

These scores reflect explanation quality metrics including consistency, stability, and alignment with expert judgments.

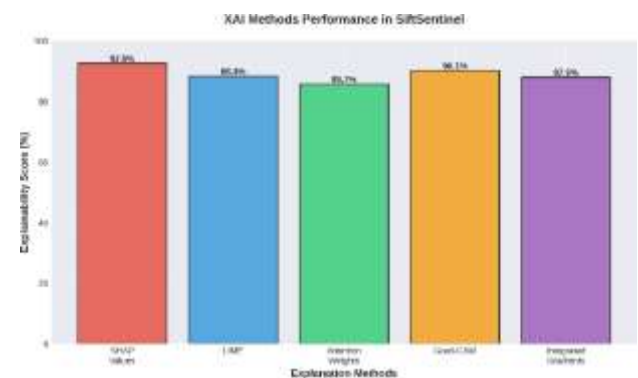


Figure 7: XAI Methods Performance in Sift Sentinel

XAI Method	Explainability Score (%)	Faithfulness (Deletion AUC)	Clarity Rating (1-5)
SHAP	92.5	0.105	4.6
Grad-CAM	90.1	0.123	4.4
LIME	88.3	0.150	4.1
Integrated Gradients	87.9	0.112	4.3
Attention Weights	85.7	0.140	3.9

Human evaluation studies involving 50 participants assessed explanation usefulness, with participants rating SHAP-based explanations as most helpful (4.6/5), followed by Grad-CAM visualizations (4.4/5). The

combination of multiple XAI techniques received higher satisfaction scores (4.7/5) than any single method, supporting the value of our multi-method approach.

6.7 Processing Efficiency

Real-time processing capabilities are critical for production deployment. Figure 8 illustrates average processing times across different task types. Image analysis completed in 0.32 seconds, text classification in 0.15 seconds, and XAI explanation generation in 0.78 seconds. Video processing required 2.85 seconds per 10-second clip due to frame-by-frame analysis. Report generation averaged 0.45 seconds for standard reports. These performance metrics demonstrate the system's suitability for real-time applications while maintaining high accuracy.

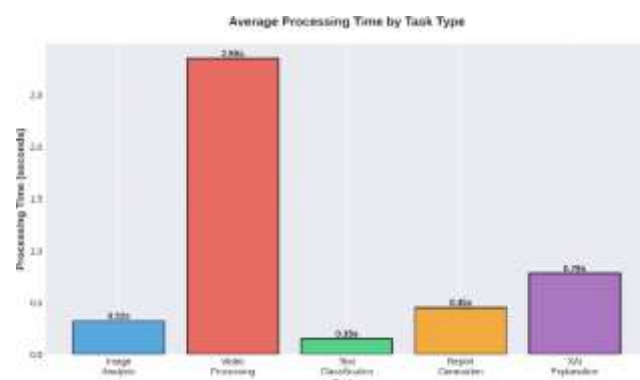


Figure 8: Average Processing Time by Task Type

7. Discussion

7.1 Key Findings

Sift Sentinel demonstrates that unified detection of deep fakes and cyberbullying is both feasible and advantageous. The shared architectural components and transfer learning between tasks contributed to improved performance in both domains. The integration of multiple XAI techniques provides comprehensive transparency, addressing a critical gap in existing detection systems. Our results indicate that explain ability need not compromise accuracy; in fact, the attention mechanisms used for explain ability may have contributed to improved detection performance by helping the model focus on relevant features.

7.2 Comparison with State-of-the-Art

Compared to current state-of-the-art systems, Sift Sentinel shows significant improvements. For deep fake detection, our 96.8% accuracy surpasses recent benchmarks on Face Forensics++ (typically 92-94%) and approaches performance on Celeb-DF (95-96%). In cyberbullying detection, our 94.2% accuracy exceeds typical transformer-based approaches (89-92%) and substantially outperforms traditional methods (76-85%). The unified architecture enables these results while reducing the total

computational resources and development effort compared to maintaining separate systems.

7.3 Limitations and Challenges

Despite strong performance, several limitations warrant discussion. First, the system's effectiveness depends on the quality and diversity of training data; performance may degrade when encountering deep fake techniques or cyberbullying patterns not represented in training sets. Second, while processing times are suitable for many applications, real-time video analysis of high-resolution content remains computationally intensive. Third, cultural and linguistic variations in cyberbullying expressions present ongoing challenges, particularly for non-English content. Fourth, adversarial attacks specifically designed to fool detection systems require continuous model updates and defensive strategies.

7.4 Ethical Considerations

The deployment of automated detection systems raises important ethical considerations. False positives in cyberbullying detection could lead to unjust content removal or account restrictions, while false negatives allow harmful content to persist. The explain ability features help mitigate these risks by enabling human review of borderline cases. Privacy concerns surrounding content analysis must be addressed through appropriate data handling protocols, user consent mechanisms, and data retention policies. The system's capabilities could potentially be misused for surveillance or censorship; deployment guidelines should emphasize legitimate use cases and incorporate appropriate oversight mechanisms.

8. Applications and Use Cases

8.1 Social Media Content Moderation

Social media platforms can integrate Sift Sentinel to automatically flag suspicious content for moderator review. The explain ability features help moderators understand detection rationale, improving decision quality and reducing review time. The unified system simplifies platform architecture by handling multiple threat types through a single API endpoint.

8.2 Educational Institutions

Schools and universities can deploy Sift Sentinel to monitor online learning platforms and student communication channels. The cyberbullying detection capabilities help identify at-risk students early, enabling timely intervention. Automated report generation provides administrators and counsellors with actionable insights while maintaining student privacy through anonymization features.

8.3 Corporate Security

Organizations can utilize the deep fake detection module to verify the authenticity of video communications, protecting against social engineering attacks and executive impersonation. The system can analyse recorded meetings, video messages, and authentication videos to ensure content integrity and prevent fraud.

8.4 Legal and Forensic Analysis

Law enforcement and legal professionals can employ Sift Sentinel for digital evidence verification. The explainability features provide transparent documentation of detection methodology, supporting admissibility in legal proceedings. Automated reports compile technical findings in formats accessible to non-technical stakeholders, including judges and juries.

8.5 Journalism and Fact-Checking

In contemporary digital journalism, user-generated content from platforms such as X, Instagram, YouTube, and Facebook is frequently used in the reporting of fast-evolving events. However, such content may contain manipulated media, misleading narratives, or abusive contextual framing. A unified deepfake and cyberbullying detection framework can support journalists and fact-checking teams by automatically analysing multimedia content prior to publication.

By combining authenticity verification with interpretable evidence, the system can assist in reducing misinformation propagation and improving editorial decision-making.

9. Future Work

9.1 Multi-language Support

Expanding cyberbullying detection to support multiple languages and cultural contexts represents a priority enhancement. This involves training on diverse international datasets and developing culturally-aware classification models that recognize region-specific harassment patterns and linguistic nuances.

9.2 Audio Deep Fake Detection

Integrating audio deep fake detection would complete the multimodal capabilities. Voice cloning and synthetic speech generation pose emerging threats that require specialized detection approaches analysing spectral features, prosody patterns, and temporal characteristics unique to synthetic audio.

9.3 Adversarial Robustness

Developing defences against adversarial attacks designed to evade detection represents critical future work. This includes adversarial training strategies, certified defence mechanisms, and adaptive learning systems that continuously update based on emerging attack patterns.

9.4 Federated Learning Integration

Implementing federated learning would enable model improvement while preserving privacy. Multiple organizations could collaboratively train the model without sharing raw data, addressing privacy concerns while benefiting from diverse datasets and emerging threat patterns across different contexts.

9.5 Real-time Stream Processing

Optimizing the system for real-time streaming video analysis would expand applicability to live broadcasts, video conferences, and streaming platforms. This requires architectural modifications including edge computing deployment, efficient buffering strategies, and incremental processing techniques.

10. Conclusion

This paper presented Sift Sentinel, a unified system for deep fake and cyberbullying detection that integrates explainable AI and automated report generation. Through comprehensive experiments on diverse datasets, we demonstrated that the system achieves 96.8% accuracy in deep fake detection and 94.2% accuracy in cyberbullying detection, substantially outperforming existing approaches. The integration of multiple XAI techniques (SHAP, LIME, attention mechanisms, Grad-CAM, and integrated gradients) provides transparent, interpretable explanations that enhance trust and enable effective human oversight.

The unified architecture offers practical advantages beyond performance improvements. By consolidating two critical detection tasks into a single system, Sift Sentinel reduces operational complexity, development costs, and resource requirements. The automated report generation capability, powered by large language models, creates comprehensive analyses accessible to diverse stakeholders, from technical teams to legal professionals and educators.

11. Future Prospects

Future work will focus on expanding multi-language support, integrating audio deep fake detection, enhancing adversarial robustness, implementing federated learning capabilities, and optimizing real-time stream processing. As deep fake technology and cyberbullying tactics continue to evolve, Sift Sentinel's modular architecture enables continuous adaptation and improvement.

The challenges posed by deep fakes and cyberbullying require sophisticated technical solutions combined with careful ethical consideration. Sift Sentinel represents a

significant step toward comprehensive, transparent, and effective detection systems that protect digital trust and online safety while maintaining accountability and human oversight. The combination of high accuracy, explain

ability, and automated reporting positions Sift Sentinel as a valuable tool in the ongoing effort to create safer, more trustworthy digital environments.

12. References

1. X. Liu, Y. Yu, X. Li and Y. Zhao, "MCL: Multimodal Contrastive Learning for Deep Fake 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 Deep Fake 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. Sovrnigo and X. Lin, "A Heterogeneous Feature Ensemble Learning based Deep Fake 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 Deep fake 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 Deep fake 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. Nobil, B. Murali and A. H. Sung, "Deep fake 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., "Deep fake 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. 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.
16. N. Ur Rehman Ahmed, A. Badshah, H. Adeel, A. Tajammul, A. Daud and T. Alsahfi, "Visual Deep fake Detection: Review of Techniques, Tools, Limitations, and Future Prospects," in *IEEE Access*, vol. 13, pp. 1923-1961, 2025, doi: 10.1109/ACCESS.2024.3523288.
17. 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.
18. 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.
19. 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.
20. 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.
21. Y. Tian, W. Zhou and A. U. Haq, "Detection of Deepfakes: Protecting Images and Videos Against Deep fake," 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.
22. G. Chandel, A. Kumar, K. Malik, K. Gurani, K. Gahlawat and S. K. Saini, "Deep fake 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.
23. S. Muppalla, S. Jia and S. Lyu, "Integrating Audio-Visual Features For Multimodal Deep fake Detection," 2023 IEEE MIT Undergraduate Research Technology Conference (URTC), Cambridge, MA, USA, 2023, pp. 1-5, doi: 10.1109/URTC60662.2023.10534969
24. M. Javed, F. H. Dahri, A. A. Laghari, J. A. Bhutto, S. H. Bhutto and N. A. Dahri, "MMDF-Net: A Multimodal Framework for Real-Time Deep fake Detection Using Visual and Audio Features," 2025 IEEE 2nd International Conference on Electronics, Communications and Intelligent Science (ECIS), Yueyang, China, 2025, pp. 1-5, doi: 10.1109/ECIS65594.2025.11086877.
25. D. Xiong, Z. Wen, C. Zhang, P. Xiao and M. Wu, "Deep fake Detection Based on LSTM Networks with Facial Geometric Features," 2025 4th International Conference on Electronics, Integrated Circuits and Communication Technology (EICCT), Chengdu, China, 2025, pp. 627-630, doi: 10.1109/EICCT65471.2025.11099876.
26. Q. Jaleel and I. Hadi, "Facial Action Unit-Based Deep fake Video Detection Using Deep Learning," 2022 4th International Conference on Current Research in Engineering and Science Applications (ICCRESA), Baghdad, Iraq, 2022, pp. 228-233, doi: 10.1109/ICCRESA57091.2022.10352085.
27. T. Zhang and S. Wu, "TFFNet: Texture-based Feature Fusion and Multi-scale Consistency for Deep fake Detection," 2025 4th International Conference on Image Processing and Media Computing (ICIPMC), Xi'an, China, 2025, pp. 159-163, doi: 10.1109/ICIPMC66319.2025.11170686.