International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

F.A.S.T. VISION: A Multi-Modal Framework for Digital Media Tamper Detection

Dr. V. Ravi Kumar
Professor

K. Jyothi Sai
UG Student
Department of CSE,
ACE Engineering college
Hyderabad-501 301, India
Email:
kavitijyothisai@gmail.com

D. Sai Rohith
UG Student
Department of CSE,
ACE Engineering college
Hyderabad-501 301, India
Email:
dangetisairohith@gmail.com

M. Abhinav
UG Student
Department of CSE,
ACE Engineering college
Hyderabad-501 301,India
Email:
abhinavgoud403@gmail.com

***Abstract*:**

The increasing availability of digital editing tools and AI-generated content has led to a rise in manipulated media, including images, screenshots, and videos, creating challenges for authenticity verification. This paper presents F.A.S.T. VISION (Forensic AI System for Tamper Detection), a multi-modal framework designed to detect tampering across different media types. The system integrates classical forensic techniques such as Error Level Analysis (ELA), copy-move detection, splicing detection, metadata analysis, and OCR-based text verification, along with AI-based feature analysis and frame-level video processing. The framework follows a modular pipeline architecture and uses a hybrid scoring mechanism to generate the final decision. Experimental evaluation on a custom dataset demonstrates an accuracy of approximately 81.67% in detecting both manual and AI-based manipulations.

Keywords: Digital Forensics, Image Tampering Detection, Video Forensics, Screenshot Analysis, Error Level Analysis (ELA), Copy-Move Detection, Splicing Detection, Metadata Analysis, Optical Character Recognition (OCR), Multi-Modal Framework, Hybrid Detection, AI-Based Forensics.

I. INTRODUCTION:

The rapid growth of digital technologies has led to an increase in manipulated media, including forged images, altered screenshots, and tampered videos. The widespread availability of advanced editing tools and AI-based content generation techniques has made verifying the authenticity of digital media increasingly difficult. This poses significant challenges in fields such as cybersecurity, journalism, and legal investigations, where the integrity of digital evidence is critical.

Existing forensic methods, such as Error Level Analysis (ELA), metadata inspection, and copy-move detection, are often limited to specific media types or rely on manual inspection. In addition, these techniques are typically applied independently and lack integration across different media formats, reducing their effectiveness in practical scenarios.

To address these limitations, this paper presents F.A.S.T. VISION (Forensic AI System for Tamper Detection), a multi-modal digital forensics framework designed to analyze images, screenshots, and videos. The system adopts a modular pipeline architecture and integrates classical forensic techniques with AI-based analysis to provide an automated and interpretable tamper detection process.

The main contributions of this work are as follows:

1. A unified multi-modal forensic framework for detecting tampering across images, screenshots, and videos.

2. A hybrid decision mechanism that combines multiple forensic signals to generate a final tampering verdict.

3. An automated multi-stage analysis pipeline supporting efficient processing and structured forensic report generation.

II. LITERATURE SURVEY

Digital media forensics has evolved significantly with advancements in image processing and artificial intelligence (AI). Early methods primarily focused on detecting inconsistencies in image compression and structure, while more recent approaches leverage deep learning techniques to enhance detection accuracy across a wider range of manipulation types.

Error Level Analysis (ELA), introduced by Jessica Fridrich et al., is one of the earliest methods for image tamper detection. It identifies variations in JPEG compression levels to highlight potentially modified regions. Although ELA is simple and provides visual

evidence of tampering, it is limited to compressed images and may produce false positives in complex or high-quality images.

With the advent of deep learning, Belhassen Bayar and Matthew Stamm proposed a Convolutional Neural Network (CNN)-based approach for image forgery detection. Their method uses constrained convolution layers to capture manipulation-specific artifacts, enabling automatic feature extraction. While effective, this approach requires large training datasets and significant computational resources.

In video forensics, Darius Afchar et al. explored deep learning methods to detect manipulated videos, particularly deepfakes. Their approach analyzes both spatial and temporal inconsistencies across frames. Despite strong performance, these methods depend heavily on large labeled datasets and may struggle with novel manipulation types.

Chen Ming et al. introduced multi-modal forensic frameworks that combine visual, textual, and metadata features for enhanced detection. While these approaches improve performance by integrating multiple data sources, they often require complex architectures and large multimodal datasets.

The literature indicates that most existing methods focus on a single media type or rely heavily on large-scale training data. Traditional techniques offer interpretability but lack robustness, whereas deep learning methods enhance accuracy at the cost of increased complexity. This underscores the need for a unified and efficient system that integrates multiple forensic techniques. The proposed F.A.S.T. VISION framework addresses this gap by combining classical and AI-based methods within a modular pipeline for multi-modal tamper detection.

III. PROBLEM STATEMENT:

The rapid advancement of digital editing tools and artificial intelligence has significantly increased the generation of manipulated media, including images, screenshots, and videos. These manipulations, ranging from simple edits to AI-generated content such as deepfakes, pose significant challenges in verifying the authenticity of digital information.

Existing digital forensic methods are often limited to specific media types or rely on manual inspection, making them inefficient, less scalable, and prone to errors. Traditional techniques, such as Error Level Analysis (ELA) and metadata inspection, are typically applied in isolation, reducing their effectiveness in detecting complex manipulations. While recent AI-based approaches improve detection accuracy, they require large training datasets and high computational resources, limiting their practical applicability.

Moreover, most existing systems lack a unified framework capable of handling multiple media types within a single pipeline, which is essential for real-world forensic scenarios. Therefore, there is a clear need for an efficient, automated, and scalable system that integrates multiple forensic techniques to enhance detection reliability across diverse forms of digital content.

IV. SYSTEM DESIGN

The F.A.S.T. VISION system is a multi-modal digital forensics framework designed to analyze images, screenshots, and videos. Rather than using a unified interface, the system provides separate dashboards for each media type, allowing users to independently upload and analyze different forms of digital content.

A. Dashboard-Based Architecture

The system consists of three dedicated user interfaces:

- Image Analysis Dashboard
- Screenshot Analysis Dashboard
- Video Analysis Dashboard

Each dashboard is linked to its corresponding backend analysis pipeline. This modular design simplifies user interaction and ensures that each media type is processed using specialized forensic techniques.

B. Processing Workflow

The overall workflow of the system is as shown in Fig. 1:

1. The user selects the appropriate dashboard (image, screenshot, or video).
2. The media file is uploaded through the selected interface.
3. The request is sent to the backend via FastAPI endpoints.
4. The corresponding forensic pipeline is triggered.
5. Multiple analysis modules process the input.
6. Intermediate results and visual outputs are generated.

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

7. A rule-based decision mechanism produces the final verdict.
8. Results, visualizations, and reports are displayed to the user.

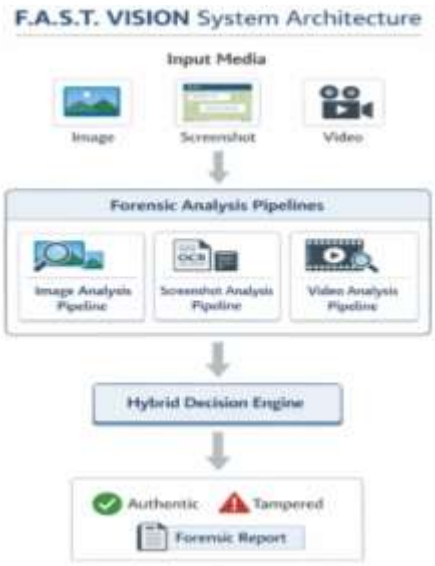

Fig. 1. System architecture of F.A.S.T. VISION

C. System Components

The F.A.S.T. VISION system comprises the following components:

- **Frontend Dashboards**: Independent interfaces for image, screenshot, and video analysis.
- **Backend Server (FastAPI)**: Handles API requests and routes them to the appropriate forensic pipelines.
- **Forensic Pipelines**:
  - Image Pipeline
  - Screenshot Pipeline
  - Video Pipeline
- **Decision Engine**: Combines outputs using rule-based logic to generate the final verdict.
- **Report Generator**: Produces structured PDF forensic reports.

D. Design Advantages

The proposed architecture offers several advantages:

- **Simplified User Interaction**: The separation of dashboards reduces interface complexity and improves usability.
- **Optimized Processing**: Each media type is analyzed using dedicated forensic techniques, improving accuracy and efficiency.
- **Improved Maintainability**: Independent pipelines facilitate easier debugging, updates, and system modifications.
- **Scalability**: The modular design enables seamless integration of future enhancements and additional features.

V. METHODOLOGY

The proposed **F.A.S.T. VISION** system adopts a modular and multi-modal approach for detecting tampering across images, screenshots, and videos as shown in Fig. 2. Each media type is processed using a dedicated forensic pipeline, where multiple analysis techniques are applied to extract relevant features. The outputs from these modules are combined using a rule-based hybrid decision mechanism to determine the authenticity of the input media.

A. Image Forensics Pipeline

The image forensics pipeline integrates multiple complementary techniques to detect various types of manipulation:

1. **Error Level Analysis (ELA)**

   ELA is used to identify compression inconsistencies within an image. The input image is recompressed, and the difference between the original and recompressed images is analyzed. Regions with higher error levels indicate potential tampering.

2. **AI-Based Content Analysis**

   A pre-trained deep learning model is used to estimate the probability of the image being AI-generated. This module outputs:
   - AI-generated probability score
   - Additional feature-based signals

   This helps identify synthetic or AI-generated images.

3. **Copy-Move Detection**

Copy-move detection identifies duplicated regions within the image. The system analyzes local features to detect repeated patterns and generates:

- Copy-move score
- Heatmap highlighting suspicious regions

4. **Splicing Detection**

   Splicing detection identifies inconsistencies in color distribution and noise patterns across different regions of the image. This technique helps detect composite images created from multiple sources.

5. **Metadata Analysis**

   Image metadata (EXIF data) is analyzed to detect anomalies such as:

   - Missing metadata fields
   - Inconsistent timestamps
   - Unusual device information

   These inconsistencies may indicate possible tampering.

6. **Decision Mechanism**

   The outputs from all image analysis modules are combined using a rule-based approach. Tampering is detected if:

   - Multiple forensic indicators are triggered, or
   - The AI-generated probability exceeds a predefined threshold

B. Screenshot Forensics Pipeline

The screenshot analysis pipeline combines text-based and visual forensic techniques:

1. **Text Extraction (OCR)**
   Optical Character Recognition (OCR) is used to extract textual content from the screenshot.
2. **Text Pattern Analysis**
   The extracted text is analyzed for inconsistencies in:
   - Alignment
   - Spacing
   - Structural layout

   These inconsistencies may indicate manipulation in text-based content such as chats or documents.
3. **Visual Inconsistency Detection**
   The system evaluates multiple visual features, including:
   - Noise inconsistency
   - Sharpness variation
   - Edge irregularities

   These features help detect local editing artifacts.
4. **ELA-Based Analysis**
   Error Level Analysis is applied to screenshots to detect compression-related anomalies.
5. **Tamper Scoring**
   All extracted features are combined to compute a tamper score. A threshold-based classification is used to label the screenshot as authentic or suspicious.

C. Video Forensics Pipeline

The video forensics pipeline focuses on frame-level and metadata-based analysis:

1. **Frame Extraction**

   The input video is divided into frames at fixed intervals for analysis.

2. **Frame-Level Analysis**

   Each frame is analyzed using feature-based and statistical techniques to detect:

   - Potential AI-generated frames
   - Visual inconsistencies

3. **Temporal Analysis**

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

Differences between consecutive frames are analyzed to identify:

- Sudden transitions
- Unnatural changes

4. **Metadata Analysis**

Video metadata is examined, including:

- Frame rate (FPS)
- Codec information
- Duration

Inconsistencies in these attributes may indicate tampering.

5. **Final Scoring**

All extracted features are aggregated to compute a suspicion score. A threshold-based decision is used to classify the video as authentic or tampered.

D. Hybrid Decision Framework

The final decision in the **F.A.S.T. VISION** system is obtained using a rule-based hybrid mechanism. Instead of relying on a single technique, the system combines outputs from multiple forensic modules to improve reliability. This reduces dependence on individual methods and enhances the overall robustness of the decision-making process.

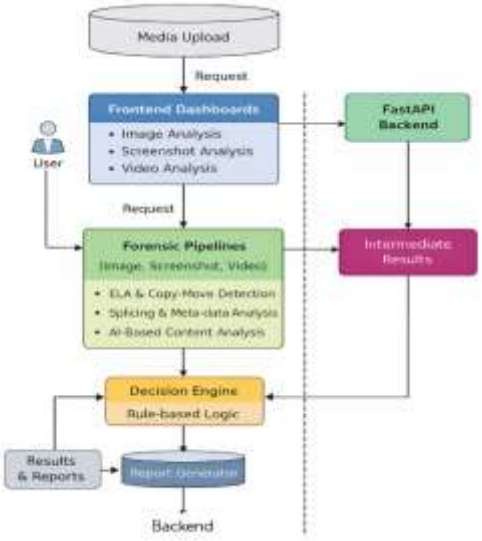

Fig. 2. Processing pipeline of F.A.S.T.VISION

VI. EXPERIMENTAL RESULTS AND ANALYSIS

The proposed **F.A.S.T. VISION** framework was evaluated using a public dataset of 300 images, comprising 150 authentic and 150 tampered samples. The tampered images include multiple manipulation types such as copy-move, splicing, and AI-generated content, ensuring a comprehensive evaluation under realistic conditions.

A. Evaluation Metrics

The system performance was evaluated using standard classification metrics:

- **Accuracy**: Overall correctness of classification
- **Precision**: Ratio of correctly identified tampered samples to total predicted tampered samples
- **Recall**: Ability to correctly identify actual tampered samples
- **F1-Score**: Harmonic mean of precision and recall

These metrics are essential for forensic systems, where precision reduces false positives and recall ensures effective detection of tampered content. The F1-score provides a balanced evaluation of both.

B. Experimental Results

The obtained performance metrics are:

- **Accuracy**: 81.67%
- **Precision**: 78.79%
- **Recall**: 86.67%
- **F1-Score**: 82.54%

The results indicate effective detection performance, with high recall demonstrating strong capability in identifying tampered media.

C. Confusion Matrix Analysis

The confusion matrix is given as shown in Fig. 3:

$$\begin{bmatrix} 115 & 35 \\ 20 & 130 \end{bmatrix}$$

Fig. 3. Confusion Matrix

where 115 and 130 denote true negatives and true positives, respectively, while 35 and 20 represent false positives and false negatives. The relatively low number of false negatives indicates

effective detection of manipulated samples, which is critical in forensic applications.

D. Discussion

The experimental results demonstrate that the proposed framework effectively detects various types of digital manipulations. The high recall (86.67%) ensures that most tampered samples are correctly identified, which is essential in forensic applications.

Although precision (78.79%) is slightly lower, this trade-off is acceptable, as detecting manipulated content is prioritized over minimizing false positives. The integration of classical forensic techniques with AI-based methods improves the overall robustness and reliability of the system.

Overall, these results demonstrate the suitability of the proposed framework for practical digital forensic applications.

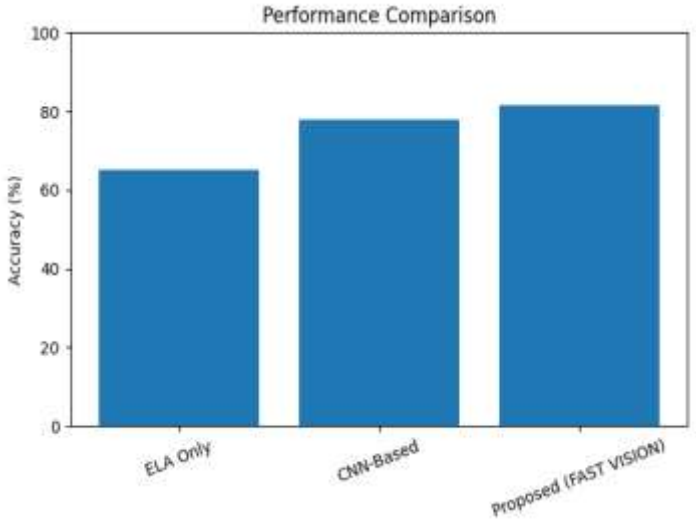

Fig. 4. Comparison Table

VII. CONCLUSION

This paper presented **F.A.S.T. VISION**, a multi-modal digital forensics framework for detecting tampering in images, screenshots, and videos. The system integrates multiple forensic techniques, including Error Level Analysis (ELA), copy-move detection, splicing detection, metadata analysis, and AI-based content evaluation, within a unified and modular architecture.

Unlike traditional approaches that rely on a single method or media type, the proposed framework adopts a hybrid strategy that combines classical forensic techniques with pre-trained AI models. This integration improves detection reliability while maintaining computational efficiency and interpretability.

Experimental results show that the system achieves an accuracy of 81.67% and a recall of 86.67%, demonstrating strong capability in detecting tampered content. These results validate the effectiveness of the hybrid decision mechanism in handling a wide range of manipulation types, including AI-generated content.

The modular design, combined with a dashboard-based interface, enhances scalability, usability, and adaptability for real-world forensic applications. In addition, the ability to generate structured forensic reports improves the system's practical applicability in investigative scenarios.

Future work will focus on improving precision, extending the framework to handle emerging manipulation techniques, and incorporating advanced deep learning models to further enhance detection performance on larger and more complex datasets.

VIII. REFERENCES

[1] J. Fridrich, M. Goljan, and R. Du, "Detecting copy-move forgery in digital images," in *Proc. Digital Forensic Research Workshop (DFRWS)*, 2003.
[2] B. Bayar and M. C. Stamm, "A deep learning approach to universal image manipulation detection," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, 2017, pp. 20–24.
[3] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A compact facial video forgery detection network," in *Proc. IEEE Int. Workshop Inf. Forensics Security (WIFS)*, 2018.
[4] T. Bianchi and A. Piva, "Image forgery localization via block-grained analysis of JPEG artifacts," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 3, pp. 1003–1017, Jun. 2012.
[5] H. Farid, "Image forgery detection," *IEEE Signal Process. Mag.*, vol. 26, no. 2, pp. 16–25, Mar. 2009.
[6] I. Goodfellow *et al.*, "Generative adversarial nets," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2014, pp. 2672–2680.
[7] M. Hussain, S. R. H. Rizvi, and A. Mahmood, "Image forgery detection using machine learning: A survey," *ACM Comput. Surveys*, vol. 51, no. 5, pp. 1–36, 2018.

[8] Y. Li *et al.*, "FaceForensics++: Learning to detect manipulated facial videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 1–11.
[9] A. Piva, "An overview on image forensics," *ISRN Signal Process.*, vol. 2013, pp. 1–22, 2013.
[10] Tesseract OCR, "Tesseract OCR Engine." [Online]. Available: https://tesseract-ocr.github.io/ (accessed Mar. 2026).