



Web-Based Deepfake Detection Using VERITAS: Integrating Vision-Based Excitation with Transformer-Driven Intelligence

Alam Rahmatulloh^{1,2}, Herman Dwi Surjono¹, Fatchul Arifin¹, Rohmat Gunawan², Randi Rizal²

^{1,2,3}Doctoral Program in Engineering Science at the Faculty of Engineering, Yogyakarta State University, Yogyakarta, Indonesia

^{1,4,5}Department of Informatics, Faculty of Engineering, Siliwangi University, Tasikmalaya, Indonesia

¹alam@unsil.ac.id, ²hermansurjono@uny.ac.id, ³fatchul@uny.ac.id, ⁴rohmatgunawan@unsil.ac.id, ⁵randirizal@unsil.ac.id

Abstract

This study proposes a web-based deepfake detection system that integrates Vision-Based Excitation technology and Transformer-based intelligence, called VERITAS (Vision-based Excitation and Robust Intelligence for Transformer-Assisted Deepfake Detection). The system is designed to automatically detect manipulated images and videos by leveraging the Vision Transformer (ViT) model architecture, equipped with the Grad-CAM mechanism for interpretability of detection results. The study conducted a series of tests to measure the system's performance in various scenarios and ensure its reliability in dealing with various types of input. Load testing results showed that up to 30 simultaneous users, the system can operate with good responsiveness (average response time of 130 ms) without experiencing errors. However, when the number of users reaches 40 or more, the system performance drops drastically with a very high error rate, reflecting limitations in handling server load. Real-world testing showed the system can detect deepfakes with an accuracy of 73.61%, with results varying depending on the quality of the tested images. Furthermore, unit functional testing and coverage analysis demonstrated an excellent test pass rate (85%), with all major functions running smoothly and error handling needed to be fixed in some code sections. Overall, the VERITAS system demonstrates strong potential for web-based deepfake detection, with high reliability under low load and adequate performance in functional testing. However, further optimization is needed to handle higher user loads.

Keywords: performance testing; unit testing; VERITAS; vision transformer; web-based deepfake detection

How to Cite: Alam Rahmatulloh, H. D. Surjono, F. Arifin, R. Gunawan, and R. Rizal, "Web-Based Deepfake Detection Using VERITAS: Integrating Vision-Based Excitation with Transformer-Driven Intelligence", *J. RESTI (Rekayasa Sist. Teknol. Inf.)*, vol. 10, no. 1, pp. 181 - 188, Feb. 2026.

Permalink/DOI: <https://doi.org/10.29207/resti.v10i1.7320>

Received: November 4, 2025

Accepted: February 23, 2026

Available Online: February 28, 2026

This is an open-access article under the [CC BY 4.0 License](https://creativecommons.org/licenses/by/4.0/)

Published by [Ikatan Ahli Informatika Indonesia](https://www.ikatanahliinformatika.org/)

1. Introduction

Rapid advances in artificial intelligence and machine learning have led to the proliferation of deepfake technology, which holds both impressive potential and concerns [1]. Deepfakes, as artificially generated media, are capable of convincingly impersonating real individuals or events, potentially spreading misinformation and disinformation on an unprecedented scale [2-4]. This phenomenon underscores the crucial need for sophisticated and reliable deepfake detection methods to maintain the integrity of digital content and public trust [3], [5]. These sophisticated video manipulation techniques enable the creation of highly realistic content, including face swapping, expression changes, and facial feature modifications, all of which pose serious threats such as identity theft and fraud [6], [7]. Therefore, the urgency

of developing effective deepfake detection systems is crucial to maintaining the authenticity and trust in digital media [3], [8]. This threat demands a more robust cyber defense strategy, especially considering that countries like Indonesia are still making slow progress in building comprehensive cyber defenses [9]. Furthermore, the legal framework in Indonesia still faces challenges in addressing the complexity of AI-driven crimes, including deepfakes, which requires a review of the TPKS Law and a revised Criminal Code [10].

While numerous studies have attempted to develop deepfake detection methods, many existing datasets still do not capture the latest deepfake videos generated by state-of-the-art methods circulating on various platforms, hampering the ability to keep pace with the rapid evolution of generative AI techniques used in real-world deepfake production [11]. As a result, detection

methods often fail to recognize multimodal deepfakes that combine video and textual manipulation, or that synchronize fake speech with lip movements, because the complexity of the artifacts produced by these techniques requires a more holistic analysis [12], [13]. This limitation highlights the need for more adaptive and comprehensive detection approaches, capable of identifying subtle anomalies produced by multimodal deepfake techniques and adapting to the rapid advancement of generative technologies [14]. The deepfake phenomenon has triggered impostor bias, where individuals doubt the authenticity of multimedia content due to the awareness of AI capabilities [15]. In this context, in-depth research into deepfake digital forensics is crucial for identifying effective detection methods to combat this media manipulation [16]. Therefore, deepfake detection has become a very important research area, focusing on identifying subtle inconsistencies and artifacts using machine learning techniques, especially Convolutional Neural Networks [15]. However, these approaches are often limited by semantic discrepancies and dataset-specific biases, which limit universal generalization, especially when dealing with complex multidimensional manipulations that are not fully explored within the current multimodal framework [12]. Therefore, advanced detection strategies capable of integrating vision-based excitation with transformer-based intelligence are needed to overcome these limitations and improve the model's generalization ability to unknown deepfakes [12], [17].

Despite the extensive body of research on deepfake detection, a critical gap remains unresolved. Most existing approaches focus either on improving detection accuracy under controlled benchmark datasets or on proposing new model architectures without sufficient consideration of explainability and real-world deployment constraints. There is still a lack of detection systems that can simultaneously generalize to rapidly evolving deepfake generation techniques, provide interpretable visual evidence for their predictions, and operate reliably in a web-based, multi-user environment. This gap becomes increasingly significant as modern deepfakes generated by transformer and diffusion-based models exhibit fewer low-level artifacts and require models capable of capturing both local and global inconsistencies while remaining transparent and practically deployable.

In this context, this study attempts to address these challenges by integrating VERITAS, a vision-based verification and explanation framework, with a transformer-based model to improve accuracy and transparency in deepfake detection [18]. The importance of this research is further reinforced by the continued advancement of deepfake technology, particularly through the integration of transformer and diffusion models, which yield superior generative quality and enhance content authenticity [19], [20]. While this technology offers significant creative potential in entertainment and education [21], its implications also include the spread of misinformation,

erosion of public trust, and violation of personal privacy, making deepfake detection a critical challenge [22]. Existing countermeasures often suffer from low performance, poor generalization, and limited resilience to various attacks due to the continuous evolution of forgery techniques [23]. Nevertheless, the integration of a vision-based deepfake detection framework with transformer-based intelligence offers a promising approach to improving accuracy and resilience against increasingly sophisticated deepfake threats [3].

This study aims to strengthen trust, authenticity, and security in online communication and information sharing by analyzing the impact of deepfakes and examining existing strategies [24]. Based on this motivation, the contribution of this study is threefold. First, this work proposes the VERITAS system, which integrates vision-based excitation mechanisms with transformer-driven intelligence to improve robustness against diverse and evolving deepfake manipulation techniques. Second, the system incorporates explainability through Grad-CAM-based visualizations, enabling users to interpret detection results by highlighting manipulated facial regions. Third, VERITAS is implemented and evaluated as a web-based deepfake detection prototype, with comprehensive testing covering functional correctness, system performance under user load, and real-world detection effectiveness. These contributions collectively position VERITAS as a practical and transparent deepfake detection framework that bridges the gap between algorithmic research and real-world application, especially in computationally limited environments such as IoT devices [25].

2. Methods

In this section, we outline the methodology adopted for developing the VERITAS deepfake detection system, which integrates Convolutional Neural Network (CNN) and an optimized transformer architecture with the addition of a Squeeze-and-Excitation (SE) Block [26]. The methodology consists of several key phases: system architecture design, data preparation, model development, and evaluation. These phases ensure the model's capability to accurately detect deepfake content in both controlled and real-world environments.

2.1 System Architecture Design

The architectural structure of the developed web-based deepfake detection application is presented in Figure 1. The development of the web-based deepfake detection application was carried out using the Django framework, utilizing Web Server Gateway Interface (WSGI) connectivity and server management via Unicorn. The VERITAS model consists of three main components with strong interconnections: spatial feature extraction using CNN, global context modeling using Vision Transformer (ViT), and a channel attention module implemented based on the Squeeze-and-Excitation (SE) Block concept. Furthermore, the model architecture is strengthened by two self-learning

branches in the transformer section, which aim to improve the generalization quality of visual representations following the FSFM framework approach [27].

Conceptually, the VERITAS system operates as a sequential deepfake detection pipeline that transforms user-provided visual input into an authenticity prediction through several interconnected processing stages. The process begins with image acquisition and preprocessing, followed by hierarchical feature extraction to capture local visual cues. These features are then refined through attention-based mechanisms to emphasize informative regions before being analyzed by a transformer-based model that captures global contextual dependencies. The final output is a binary classification indicating whether the input is authentic or manipulated, accompanied by visual explanation to support interpretability.

Figure 1 shows the web-based deepfake detection workflow. The process begins with an Input Image (1) uploaded by the user via the Client Browser (2). The image is sent from the client browser to the Web Server (Nginx/Apache) (3), which is responsible for receiving the request and forwarding it to the Gunicorn WSGI (Web Server Gateway Interface) (4). Gunicorn runs a Python application using the Django framework (5), which manages the application flow and is responsible for processing the image through views.py (6). In this section, the Predictor (7) is the result of the VERITAS model processing on the received image to detect whether it is a deepfake or authentic.

Next, the application displays the detection results via a web page in the form of veritas.html (8) using the Django framework. This is then forwarded back to the client (10) via WSGI and the Web Server (9), indicating whether the image is "Real" or "Fake" as the Output Image (11) displayed to the user.

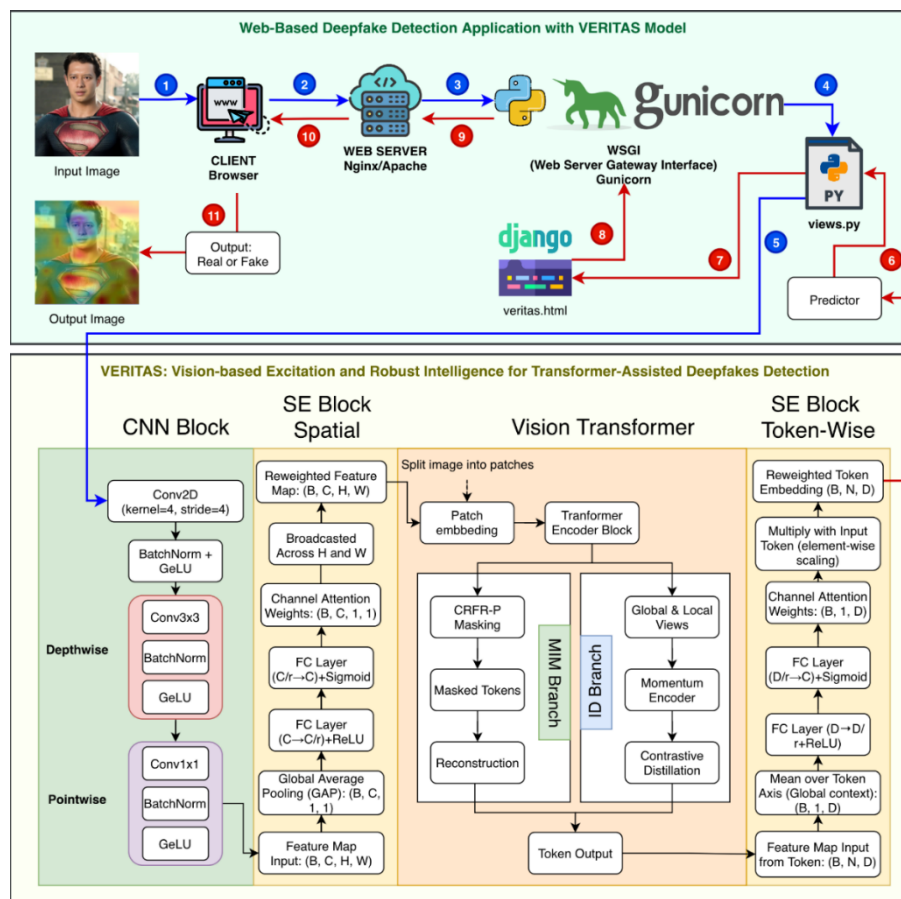


Figure 1. Proposed Web-Based Deepfake Detection Application with VERITAS Model

2.2 Data Preparation

The following description focuses on the internal model architecture and computational operations, rather than repeating the overall system workflow. In the VERITAS model, the image is first processed through a CNN Block that uses 2D convolution (4x4 kernel, stride 4) to extract basic image features (this process is indicated by the Conv2D step (kernel=4, stride=4)). The features are then processed with several convolutional

layers, including Conv3x3 and Conv1x1, as well as BatchNorm and GeLU techniques for normalization and non-linear activation. In the next stage, the extracted features are processed with Global Average Pooling (GAP) to reduce dimensionality and simplify further processing.

After that, the image is processed through two main blocks: the SE Block Spatial and the SE Block Token-Wise. In the SE Block Spatial, spatial attention is

applied to the image by calculating important channels using Channel Attention Weights. Next, the image is divided into small pieces (patches) and processed using the Patch Embedding and Transformer Encoder Blocks, part of the Vision Transformer.

In the Vision Transformer, these image pieces are further processed using the CRFFR-P Masking technique to filter out less relevant information, and the MIM Branch to reconstruct the image parts that are relevant. missing. All these processes aim to improve the accuracy of deepfake detection.

At the end of the process, the model generates an Output Token that will be used to determine whether an image is deepfake or real. This process ensures that the application can effectively detect image manipulation using a combination of CNN, SE Block, and Vision Transformer techniques.

2.3 System Development

Overall, this application combines the strengths of CNN Block, SE Block Spatial, and Vision Transformer to accurately detect deepfakes, leveraging deep image processing and a powerful transformer-based model.

The VERITAS model was tested through a training experiment using the FaceForensics++ dataset [28], with an image size of 224×224, a batch size of 64, and 10 epochs. Evaluation was carried out using two datasets, CelebDF-v2 (CDFV2) [29] and Wild Deepfake (WDF) [30]. Furthermore, an ablation study was conducted, revealing that the full configuration (CNN+SE+TF+SE) achieved the best performance, with AUC values of 83.79% and 87.15%. The precision reached 80.02%, and recall saw a significant improvement to 88.61%, leading to the highest F1-score of 84.10%.

With this architecture, VERITAS not only functions as a detection tool but also provides transparency in the analysis results. This makes the system more trustworthy and relevant for application in real-world contexts, particularly in supporting digital information security and reducing the risk of spreading fake content.

2.4 System Evaluation

The system evaluation was conducted through a series of tests to assess the performance and effectiveness of the web-based VERITAS deepfake detection system. The first test aimed to ensure that each system component functioned properly and had adequate code coverage. Next, end-to-end testing was conducted to examine the overall system's interactions in handling a wide variety of user inputs. Load testing was conducted to measure the system's ability to handle multiple concurrent users, while real-world testing used data from online platforms to assess system performance under more varied conditions. Finally, the results of these tests were compared to assess the advantages and limitations of the VERITAS system in detecting deepfakes.

3. Results and Discussions

VERITAS outperforms several leading methods, achieving AUC values of 83.79% at the frame level and 87.15% at the video level, surpassing the FSFM baseline which recorded AUC values of 76.39% and 82.31%, respectively. Additionally, VERITAS demonstrates strong detection performance with a precision of 80.02%, recall of 88.61%, and an F1-score of 84.10%, particularly excelling in identifying manipulated faces while maintaining a low false-positive rate [18].

A web-based deepfake detection system using the VERITAS (Vision-based Excitation and Robust Intelligence for Transformer-Assisted Deepfake Detection) model has been successfully implemented in the form of a web-based application prototype. The system is designed with a Django backend architecture that integrates the Vision Transformer model with the Squeeze-and-Excitation Attention mechanism, enabling it to process image and video input and generate predictions regarding the authenticity of the content. On the frontend, the system provides a simple and interactive web interface, allowing users to upload media files for automatic analysis.

The system's main functions include: (1) input upload and validation, which ensures only files with valid formats are processed and provides warnings if files are not; (2) deepfake detection, which displays classification results in the form of real or fake labels along with confidence scores in the form of probabilities; and (3) heatmap visualization, which provides visual interpretation by highlighting facial areas suspected of manipulation. Functional trials through end-to-end testing demonstrated that all scenarios performed as expected, from processing valid files, rejecting files with incorrect formats, to handling empty inputs. This demonstrates that the system is robust enough to anticipate varying input conditions.

Being web-based, the system is widely accessible to users without requiring special software installation. The integration of Vision Transformer and SE-Attention allows the model to be more sensitive to subtle manipulation details, while heatmap visualization enhances explainability, allowing users to receive not only detection results but also an understanding of suspected facial features. The implementation of the system in the form of a web prototype demonstrates that research is not limited to theory but is directed toward practical solutions ready for further testing in real-world contexts and with the potential for integration into broader digital security platforms.

A series of tests have been conducted on the VERITAS-based deepfake detection system, and the results are presented. A comprehensive analysis was conducted to evaluate system performance across various aspects, from functional stability through unit testing and coverage, system robustness to input variations through

end-to-end testing, user load handling through load and performance testing, to detection effectiveness under real-world conditions through testing of deepfake image and video samples.

The discussion focuses on the test results, reflecting the system's reliability, limitations, and opportunities for future improvement. In addition, each test result is analyzed by providing academic and practical justification, so that the findings of this study not only have theoretical value, but also applicable value.

Accordingly, the following subsections not only report experimental results but also examine the underlying factors influencing system behavior, limitations, and comparative performance with existing deepfake detection approaches.

3.1. Functional Analysis of the System through Unit Test and Coverage

Unit testing (Table 1) was conducted on 19 scenarios using the Django TestCase framework. The results showed that all tests ran without errors, with code coverage reaching 85%. The core module `detector.py` had 97% coverage, `views.py` 93%, `models_vit.py` 82%, and `grad_cam.py` 73%.

A high coverage level is evidence that the system has met basic stability requirements. Weaknesses in the `grad_cam.py` module indicate the need for additional

testing on error cases to improve robustness. A high coverage level is evidence that the system has met basic stability requirements. Weaknesses in the `grad_cam.py` module indicate the need for additional testing on error cases to improve robustness.

Table 1. Unit Test Results and Coverage

Modul	Coverage	Interpretation
<code>detector.py</code>	97%	Almost all core functionality has been tested and is stable without errors.
<code>Grad_cam.py</code>	73%	Most of the functionality works well, but error handling has not been fully tested.
<code>Models_vit.py</code>	82%	Input variations have been tested, but some code paths remain untested.
<code>Views.py</code>	93%	The web display functionality has been tested well, except for error handling.

3.2. End-to-End Testing Analysis

E2E testing (Table 2) was performed to validate the system's interaction with users. The results showed that all scenarios performed as expected: valid files were processed correctly, malformed files were rejected, and empty input generated a warning.

E2E testing proves a system is robust enough to handle input variations and user errors. This is essential to ensure system reliability when used in public environments.

Table 2. E2E Testing Result

Scenario	Expected Results	Actual Results	Status
Upload a valid image	Preview + detection results appear	accepted	✓
Upload an invalid file (.txt)	Rejected / warning	accepted	✓
Upload an empty file	Browser warning	accepted	✓

3.3. Load Analysis and Performance Testing

The performance tests in Table 3 used Apache JMeter with 5–50 simultaneous users. Results showed the system was stable for up to 30 users, with a response time of ± 100 –130 ms and 0% error. However, at 40 users, the response time spiked (1304 ms) with a 57.44% error rate. At 50 users, the error rate increased dramatically to 91.40%.

Table 3. Load Testing Result

Number of Users	Average Response Time	Error %	Interpretation
5–30	97–131 ms	0%	The system is stable and performing well. The server is starting to become overwhelmed, and performance is dropping.
40	1304 ms	57,44%	The system is failing to withstand the high load.
50	1967 ms	91,40%	

The system is still limited to small-to-medium scale. Server optimization, load balancing, or a container-based architecture are needed to support more simultaneous users.

3.4. Real World Testing Analysis

The 72 deepfake image/video samples used for real-world testing were sourced from a combination of publicly available datasets and possibly custom collections specifically created for this experiment. The samples are custom, their collection involved sourcing content from real-world platforms where deepfake videos are increasingly circulated, such as social media sites and online repositories known for distributing manipulated media. Additionally, videos created using deepfake tools or services like DeepFakes, FaceSwap, and GANs were included to ensure a diverse representation of manipulation methods.

To ensure the authenticity of the dataset, a thorough validation process was conducted. Real samples were manually verified, likely sourced from well-known public video repositories that feature genuine content, such as celebrity interviews or publicly available news footage. The fake samples were also verified by

confirming that they were generated using recognized deepfake creation techniques. This validation process was further strengthened by cross-referencing the samples with trusted sources to maintain a balanced mix of authentic and manipulated content.

The 72 samples were divided into a balanced set, with an equal distribution of real and fake content. The breakdown was as follows: 36 real samples and 36 fake samples. This balanced approach ensured that the real-world testing provided an accurate representation of how well the system could differentiate between genuine and manipulated content under diverse conditions. Real-world testing was conducted on 72 deepfake image/video samples, including those in Figures 2 and 3. The results showed an accuracy rate of 73.61%. The system was able to distinguish between genuine and fake content fairly well, but still made frequent errors in borderline cases (predictions approaching 50%).

The performance gap between benchmark evaluation results and real-world testing accuracy (73.61%) can be primarily attributed to several interacting factors. First, a domain shift is evident between curated benchmark datasets, such as FaceForensics++ and CelebDF-v2, and real-world data collected from social media and online platforms. Benchmark datasets are typically constructed under controlled conditions with relatively consistent lighting, resolution, and compression settings, whereas real-world content often undergoes aggressive compression, resizing, and re-encoding, which can obscure subtle manipulation artifacts. This phenomenon has been widely reported in prior deepfake detection studies as a major cause of degraded generalization performance.

Second, the diversity of manipulation techniques in real-world samples introduces additional complexity. Unlike benchmark datasets that focus on specific generation pipelines, real-world deepfakes may involve hybrid or post-processed manipulations, including manual editing, style transfer, and temporal smoothing. Such variations can reduce the discriminative power of models trained primarily on standard deepfake datasets, even when transformer-based architectures are employed.

Third, borderline cases observed in the experiments, where prediction confidence approaches 50%, suggest that some real-world samples exhibit visual characteristics that closely resemble authentic content. This is consistent with recent findings indicating that diffusion-based and transformer-driven generators tend to minimize low-level artifacts, thereby challenging detectors that rely on subtle visual inconsistencies. Consequently, although VERITAS demonstrates robustness under benchmark conditions, its performance in unconstrained environments reflects the inherent difficulty of real-world deepfake detection rather than implementation deficiencies.

These results (Table 4) demonstrate that the system is adaptive to real-world data beyond the benchmark dataset. However, accuracy remains below the target of $\geq 85\%$, requiring additional strategies, such as domain adaptation or data augmentation, to improve generalization.

Table 4. Real World Test Results

Total Sample	Accuracy	Major Weaknesses
72	73,61%	Misclassification in borderline cases

Figure 2 illustrates the raw image input that the VERITAS system processes for deepfake detection. It represents a sample image that is uploaded to the system via the web interface for analysis. This input image will undergo preprocessing, feature extraction, and manipulation analysis as part of the detection pipeline. It is essential for verifying how the system handles real-world image data and distinguishing it from manipulated media.

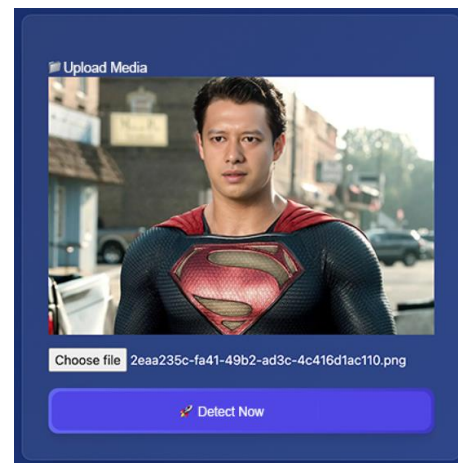


Figure 2. Test Input Image

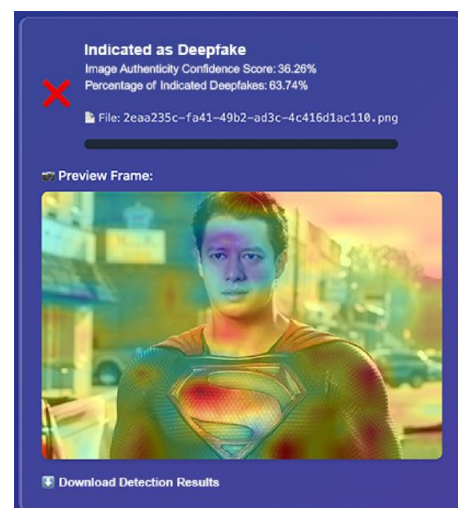


Figure 3. Test Result Image

Figure 3 shows the result after the VERITAS system has processed the test input image. This could include visualizations such as heatmaps, confidence scores, and a classification output (real or fake). The heatmap might highlight facial features or other regions that are

identified as manipulated, demonstrating the model's interpretability using the Grad-CAM mechanism. This figure provides valuable insights into the system's ability to visualize and explain detection results, an important factor for building trust and transparency in deepfake detection.

3.5. Comparison of Test Results

Figure 4 shows a comparison of the VERITAS system test results. E2E testing was 100% successful as expected, with 85% unit test coverage, and real-world accuracy at 73.61%. In load testing, the system was stable for up to 30 users (0% error) but failed for 50 users with an error rate of 91.4%. This confirms that the system is functional and stable on a small-to-medium scale, but still requires accuracy improvements and scalability optimization.

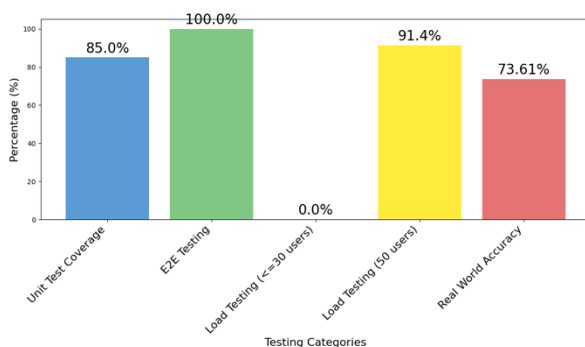


Figure 4. Comparison of VERITAS System Test Results

When contextualized within existing deepfake detection literature, the performance of VERITAS aligns with trends reported by recent transformer-based and hybrid detection approaches. Prior studies such as GenConViT and FSFM have demonstrated strong benchmark performance but similarly report noticeable performance degradation when applied to unseen datasets or real-world scenarios. The AUC improvements achieved by VERITAS over the FSFM baseline indicate that the integration of vision-based excitation mechanisms contributes positively to feature discrimination, particularly in capturing informative facial regions.

However, comparable to findings in other studies, the real-world accuracy achieved by VERITAS remains lower than benchmark results, reinforcing the widely acknowledged generalization challenge in deepfake detection. Unlike some high-complexity models that prioritize detection accuracy at the cost of interpretability, VERITAS explicitly incorporates Grad-CAM-based visualization, offering an advantage in transparency and forensic usability. This design choice positions VERITAS as a balanced solution that emphasizes explainability and deployability, even if absolute accuracy under unconstrained conditions has not yet reached the upper bounds reported in controlled experimental settings.

Furthermore, compared to lightweight or real-time detectors that sacrifice contextual modeling for speed, VERITAS leverages transformer-based global reasoning while maintaining acceptable responsiveness in a web-based environment. This trade-off highlights the system's relative strength in practical applicability rather than purely offline performance optimization. As such, the results of this study support existing conclusions in the literature that robust deepfake detection in real-world settings requires not only architectural innovation but also dataset diversity, domain adaptation strategies, and system-level considerations.

4. Conclusions

This research successfully implemented a VERITAS (Vision-based Excitation and Robust Intelligence for Transformer-Assisted Deepfake Detection) based deepfake detection system in the form of a stable, user-accessible web prototype. Functional and end-to-end testing indicate that the system meets internal stability requirements, with 85% code coverage and consistent handling of valid and invalid input scenarios.

In terms of technical performance, load testing results showed the system could serve up to 30 simultaneous users with a stable response time (100–130 ms) without errors. However, when the number of users exceeded 40, system performance decreased drastically, with error rates reaching over 50%. This indicates the need for infrastructure optimization, for example, with load balancing or container-based deployment, to ensure the system's ability to operate at scale.

In terms of detection effectiveness, real-world testing results on 72 image/video samples demonstrated a detection accuracy of 73.61%. While these results are quite good for a prototype, they are still below the target of $\geq 85\%$ commonly used in standard deepfake detection research. These findings indicate the need for enhanced model training strategies, such as domain adaptation or more varied data augmentation, to improve the system's generalizability to data beyond the benchmark dataset.

These findings suggest that while architectural integration and explainability can enhance practical usability, achieving robust performance in unconstrained environments remains a central challenge for deepfake detection systems.

Overall, this research contributes to the field of deepfake detection by presenting a web-based prototype that combines transformer-based detection with visual explanation mechanisms, as demonstrated through functional testing, performance evaluation, and real-world experiments. This advantage increases the transparency of detection results and strengthens user trust. Thus, VERITAS can be positioned as a web-based deepfake detection system with demonstrated functional reliability and interpretability, while further improvements in accuracy generalization and

scalability are required prior to large-scale real-world deployment.

Acknowledgements

This research was funded by the Institute for Research and Community Service, Siliwangi University (UNSIL) Indonesia in 2025.

References

- [1] G. Petmezas, V. Vanian, K. Konstantoudakis, E. E. I. Almaloglou, and D. Zarpalas, "Video deepfake detection using a hybrid CNN-LSTM-Transformer model for identity verification," *Multimedia Tools and Applications*, Mar. 2025. DOI: [10.1007/s11042-024-20548-6](https://doi.org/10.1007/s11042-024-20548-6)
- [2] M. R. Shoaib, Z. Wang, M. T. Ahvanooy, and J. Zhao, "Deepfakes, Misinformation, and Disinformation in the Era of Frontier AI, Generative AI, and Large AI Models," in *2023 International Conference on Computer and Applications (ICCA)*, 2023, pp. 1–7. DOI: [10.1109/ICCA59364.2023.10401723](https://doi.org/10.1109/ICCA59364.2023.10401723)
- [3] A. H. Soudy et al., "Deepfake detection using convolutional vision transformers and convolutional neural networks," *Neural Computing and Applications*, vol. 36, no. 31, pp. 19759–19775, Nov. 2024. DOI: [10.1007/s00521-024-10181-7](https://doi.org/10.1007/s00521-024-10181-7)
- [4] A. Rahmatulloh, H. D. Surjono, F. Arifin, and G. F. Nugraha, "Exploring Deepfake Models for Image Translation: A Systematic Review of Current Techniques and Future Directions," *Social Science Research Network (SSRN) Elsevier*, 03-Dec-2024. DOI: [10.2139/ssrn.5042547](https://doi.org/10.2139/ssrn.5042547)
- [5] A. A. Hasanaath, H. Luqman, R. Katib, and S. Anwar, "FSBI: Deepfakes Detection with Frequency Enhanced Self-Blended Images," 2024 [Online]. Available: <https://github.com/gufranSabri/>
- [6] D. Wodajo Deressa, H. Mareen, P. Lambert, S. Atnafu, Z. Akhtar, and G. Van Wallendael, "GenConViT: Deepfake Video Detection Using Generative Convolutional Vision Transformer," 2025 [Online]. Available: <https://github.com/erprogs/GenConViT>
- [7] R. Babaei, S. Cheng, R. Duan, and S. Zhao, "Generative Artificial Intelligence and the Evolving Challenge of Deepfake Detection: A Systematic Analysis," *Journal of Sensor and Actuator Networks*, vol. 14, no. 1, p. 17, Feb. 2025. DOI: [10.3390/jsan14010017](https://doi.org/10.3390/jsan14010017)
- [8] V. L. L. Thing, "Deepfake Detection with Deep Learning: Convolutional Neural Networks versus Transformers," in *2023 IEEE International Conference on Cyber Security and Resilience (CSR)*, 2023, pp. 246–253. DOI: [10.1109/CSR57506.2023.10225004](https://doi.org/10.1109/CSR57506.2023.10225004)
- [9] A. Mardamsyah, G. Winandika Saragih, A. Farras Zhafran, I. R. Simanjuntak, and M. F. Riadi, "Penerapan Komunikasi ESP-NOW dengan Topologi Jaringan Mesh Sebagai Bentuk Interoperability dalam Konsep Network Centric Warfare," *Journal of Education, Humaniora and Social Sciences (JEHSS)*, vol. 7, no. 2, pp. 707–727, Nov. 2024. DOI: [10.34007/jehss.v7i2.2374](https://doi.org/10.34007/jehss.v7i2.2374)
- [10] M. Hafiz, "The Era of Artificial Intelligence: Examining Indonesia's Adaptability and Legal Challenges," 25-Sep-2024. DOI: [10.31219/osf.io/vk7jq](https://doi.org/10.31219/osf.io/vk7jq)
- [11] B. Cho et al., "Towards Understanding of Deepfake Videos in the Wild," 2025 [Online]. Available: <https://arxiv.org/abs/2309.01919>
- [12] P. Liu, S. Member, Q. Tao, and J. Tianyi Zhou, "From Single-modal to Multi-modal Facial Deepfake Detection: Progress and Challenges," 2024 [Online]. Available: <https://arxiv.org/html/2406.06965v4>
- [13] Y. S. Krubha et al., "Robust AI-Generated Face Detection with Imbalanced Data," 2025 [Online]. Available: <https://github.com/Purdue-M2/SP>
- [14] Dr. Sheshang Degadwala and Vishal Manishbhai Patel, "Advancements in Deepfake Detection: A Review of Emerging Techniques and Technologies," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 10, no. 5, pp. 127–139, Nov. 2024. DOI: [10.32628/CSEIT24105811](https://doi.org/10.32628/CSEIT24105811)
- [15] I. Amerini et al., "Deepfake Media Forensics: State of the Art and Challenges Ahead," 2024 [Online]. Available: <https://arxiv.org/abs/2408.00388>
- [16] S. M. Qureshi, A. Saeed, S. H. Almotiri, F. Ahmad, and M. A. Al Ghamdi, "Deepfake forensics: a survey of digital forensic methods for multimodal deepfake identification on social media," *PeerJ Computer Science*, vol. 10, p. e2037, May 2024. DOI: [10.7717/peerj-cs.2037](https://doi.org/10.7717/peerj-cs.2037)
- [17] D. A. Coccomini, R. Caldelli, F. Falchi, and C. Gennaro, "On the Generalization of Deep Learning Models in Video Deepfake Detection," *Journal of Imaging*, vol. 9, no. 5, p. 89, Apr. 2023. DOI: [10.3390/jimaging9050089](https://doi.org/10.3390/jimaging9050089)
- [18] A. Rahmatulloh, H. D. Surjono, F. Arifin, I. Darmawan, and N. Ambarsari, "VERITAS: Vision-based Excitation and Robust Intelligence for Transformer-Assisted Deepfakes Detection," *International Journal of Intelligent Engineering and Systems*, vol. 18, no. 8, pp. 472–484, Sep. 2025. DOI: [10.22266/ijies2025.0930.29](https://doi.org/10.22266/ijies2025.0930.29)
- [19] H. Ranka et al., "Examining the Implications of Deepfakes for Election Integrity," 2024 [Online]. Available: www.aaai.org
- [20] R. Lanzino, F. Fontana, A. Diko, M. R. Marini, and L. Cinque, "Faster Than Lies: Real-time Deepfake Detection using Binary Neural Networks," 2024 [Online]. Available: <https://arxiv.org/abs/2406.04932>
- [21] R. Babaei, S. Cheng, R. Duan, and S. Zhao, "Generative Artificial Intelligence and the Evolving Challenge of Deepfake Detection: A Systematic Analysis," *Journal of Sensor and Actuator Networks*, vol. 14, no. 1, p. 17, Feb. 2025. DOI: [10.3390/jsan14010017](https://doi.org/10.3390/jsan14010017)
- [22] M. Abbasi, P. Váz, J. Silva, and P. Martins, "Comprehensive Evaluation of Deepfake Detection Models: Accuracy, Generalization, and Resilience to Adversarial Attacks," *Applied Sciences*, vol. 15, no. 3, p. 1225, Jan. 2025. DOI: [10.3390/app15031225](https://doi.org/10.3390/app15031225)
- [23] H. Wu et al., "Combating Deepfakes: An Entropy-Aware Framework for Detecting Forged Visual and Textual Information," Nov. 2024. DOI: [10.21203/rs.3.rs-5380506/v1](https://doi.org/10.21203/rs.3.rs-5380506/v1)
- [24] P. L. Kharvi, "Understanding the Impact of AI-Generated Deepfakes on Public Opinion, Political Discourse, and Personal Security in Social Media," *IEEE Security & Privacy*, vol. 22, no. 4, pp. 115–122, Jul. 2024. DOI: [10.1109/MSEC.2024.3405963](https://doi.org/10.1109/MSEC.2024.3405963)
- [25] S. Bethu, M. Trupthi, S. K. Mandala, S. Karimunnisa, and A. Banu, "AI-IoT Enabled Surveillance Security: DeepFake Detection and Person Re-Identification Strategies," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 7, 2024. DOI: [10.14569/IJACSA.2024.0150799](https://doi.org/10.14569/IJACSA.2024.0150799)
- [26] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141 [Online]. DOI: [10.1109/CVPR.2018.00745](https://doi.org/10.1109/CVPR.2018.00745)
- [27] G. Wang, F. Lin, T. Wu, Z. Liu, Z. Ba, and K. Ren, "FSFM: A Generalizable Face Security Foundation Model via Self-Supervised Facial Representation Learning," *arXiv preprint arXiv:2412.12032*, Dec. 2024 [Online]. Available: <https://arxiv.org/abs/2412.12032>
- [28] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, "FaceForensics++: Learning to Detect Manipulated Facial Images," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 1–11. DOI: [10.1109/ICCV.2019.00009](https://doi.org/10.1109/ICCV.2019.00009)
- [29] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3204–3213. DOI: [10.1109/CVPR42600.2020.00327](https://doi.org/10.1109/CVPR42600.2020.00327)
- [30] B. Zi, M. Chang, J. Chen, X. Ma, and Y. G. Jiang, "WildDeepfake: A Challenging Real-World Dataset for Deepfake Detection," in *MM 2020 - Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 2382–2390. DOI: [10.1145/3394171.3413769](https://doi.org/10.1145/3394171.3413769)