AuthenX: A Multi-Modal Deep Learning Framework for Multimedia Authentication and Automated Fact-Checking

Kavita Namdev[1], Mandisha Mirza[2], Radhika Kumawat[3], Riya Kaushal[4],
Mandisha Mirza[5], Reeti Shrimal[6], Priyanka Chirote[7]
[1,2,3,4,5,6,7]*Dept. of Computer Science & Engineering Acropolis Institute of Technology and Research Indore, India*

Abstract— **Increased accessibility of Generative AI has brought a surge in the number of manipulated media. Detection methods, however, are usually tailored to individual modalities, rendering them ineffective against real-world instances that involve various content types in a piece of misinformation. AuthenX is a multi-modal pipeline that unites image, video, text, and headlines verification. Rather than developing another independent detection module, we emphasize architectural integration of independently verified parts. Specifically, the image and video modules utilize shared architecture based on EfficientNet-B4, followed by frame-wise concatenation; the text detection part involves weighted combination of perplexity, burstiness, and stylometry models; finally, the claim verification component performs a real-time web search and calculates semantic similarity via lightweight transformer-based embeddings. An important contribution of our work is that the resulting framework can be considered not only from a methodological but also from a system perspective, as it includes stateless authentication and scoped user persistence features, allowing for longitudinal verification tracking. Experimental evidence shows that the proposed solution exhibits a balance between efficiency and effectiveness of multi-modal detection.**

Index Terms— **Deep fake Detection, Multimedia Authentication, Fake News Detection, CNN, NLP, Fact Checking, Digital Forensics, AI Security.**

I. INTRODUCTION

In the last few years, the emergence of generative artificial intelligence has made it easier for people to generate extremely lifelike digital media. This can range from images being generated using generative models to videos that have been easily edited as well as texts being generated through language models. While this technology is helpful in several fields, there has been an increase in fake news spreading because of its ease of use. It is rare to see fake content in any one form, but rather, the combination of both visuals and text is common.[1][2]

There are several methods that have been developed to counteract the generation of fake news. In image processing, it involves detecting inconsistency in visuals while video processing extends that into temporal inconsistency. Meanwhile, in textual data, it involves detecting linguistic inconsistency through metrics like perplexity or style. Despite these solutions having achieved success individually, they are not designed to work concurrently.[3][5][7][8]

Another aspect we have observed is that highly performing architectures require considerable resources either in terms of computations or curated datasets to demonstrate good performance. Consequently, the application of such techniques in real-life systems faces several limitations since they cannot be used immediately for deployment purposes due to the required amount of computational power. Moreover, almost all current methods do not enable the storage of the detection information at the level of a session, therefore, they cannot be applied in cases of continuous monitoring.[6][14]

The approach presented in this paper was named AuthenX, and represents the integration of various detection algorithms into one coherent solution. Since there is no need to create a new technique, we concentrated on the problem of combining various tools effectively. This system allows processing

images, videos, text, and headlines with the usage of distinct modules that produce signals, which are further combined for analysis. Throughout the development process, much attention was paid to the question of making AuthenX lightweight, which would make it possible to deploy the system even on conventional hardware devices.[9][10][11][12][14]

One more element that distinguishes our work from others relates to the system architecture and user experience. It includes the notion of designing a system-wide architecture that enables human intervention during the verification process. AuthenX uses stateless authentication with user-level detection results stored in the system, making it possible to reuse earlier evaluations.[9]

Research and Gap Contribution

The major contributions made by the project are discussed as follows:

- Proposed a multimodal detection system that can work with images, videos, texts, and headlines in an integrated manner
- Formulated a computation-efficient technique for video processing, using frame sampling and aggregation
- Used a clear-cut method for detecting texts, incorporating the concepts of perplexity, burstiness, and stylometry
- Developed a real-time system for validating headlines, using live web searches and semantic similarity
- Incorporated a persistence system in which users could store verification results over time

II. LITERATURE REVIEW

Detection and identification of manipulated and false media content have received much attention from researchers within various disciplines ranging from image forgery detection, video forgery, text generator detection, to automatic fact checking. Despite this progress, many solutions to date have been designed for one media modality only.[2][7]

Face Forensics++ is an example of a dataset that has helped researchers develop deepfake detection algorithms for images. Such techniques often utilize convolutional neural networks that have been trained using face-cropped regions. Although they perform well in controlled environments, the need to use face-cropped regions limits their effectiveness in detecting manipulations in non-face images. For instance, they cannot detect deepfakes in artificial intelligence-generated images that lack faces.[6][11][12]

Video-based solutions have been studied in large-scale datasets, including the Deepfake Detection Challenge. Such methods often use high-capacity models and large amounts of data to reach comparable performance. But those techniques come with a hefty computational cost and demand specific hardware to work effectively, thus being less applicable in real-world conditions. Moreover, most of such solutions process video frames independently or utilize complicated temporal networks without considering practical limitations.[3][5][14]

Textual analysis involves the use of new technologies like GPTZero in the detection of AI-written content through factors such as perplexity and burstiness. Although efficient in some instances, this approach suffers from the limitations of being opaque in the prediction process and proprietary technology, which does not reveal the underlying mechanisms. Additionally, these systems take in only plain text data rather than formatted data like PDFs and DOCX files.[8][13]

The process of fact checking itself has progressed too. For instance, Claim Buster identifies claims for verification and does not check whether they are true or false. The newer systems, on the other hand, use pre-tagged data and supervised learning methods; thus, they have difficulty with new data because they cannot analyze and verify unfamiliar claims.[7][8]

Throughout all these fields, one aspect that is prevalent is the failure of integration across various channels. There have not been many studies that consider the integration of several input channels in order to enhance reliability. In addition to this, not enough consideration has been placed on system-wide considerations like usability and real-time performance.[2][4]

The proposed solution attempts to fill these gaps by looking into a different avenue, namely the fusion of multimodal approaches with actual implementation. In contrast to the single model optimization paradigm that was previously discussed, the solution of AuthenX seeks to integrate the visual, textual, and semantic verification modules in one comprehensive framework.[1][2][3][7][8][14]

III. METHODOLOGY

The methodology for this study is structured to design, develop, and evaluate a multi-modal multimedia authentication and fact-checking system aimed at detecting manipulated images, videos, and misleading textual content. The proposed model framework integrates deep learning–based visual analysis with natural language processing and automated web verification to enhance detection accuracy while minimizing false classifications.[1][2]

This section presents the complete development pipeline of the AuthenX system, including data collection and preprocessing, model implementation, multi-modal verification mechanisms, confidence-based decision logic, and system evaluation. The implementation is carried out using Python-based frameworks, including TensorFlow/PyTorch for deep learning, OpenCV for media processing, and Flask for backend integration.[12][14][10]

3.1. System Overview

AuthenX is envisioned as a unifying multi-modal architecture that can handle disparate types of input such as images, videos, textual documents, and headlines alone. Instead of creating a new detection mechanism, AuthenX aims at integrating several existing, independently verified mechanisms in a single pipeline.[2][7]

In designing the system, one primary concern is ensuring that the signal provided by each component adds value while being computationally feasible. The inputs are sorted into distinct categories depending on their modality, which is then processed via their corresponding pipelines. A confidence score that reflects the probability of an image or document being fake or authentic is computed by each pipeline and then aggregated to form a conclusive result.

Aside from detection, there is also an additional level within the system for interfacing with users and persisting their inputs for future reference. As shown in fig 3.1

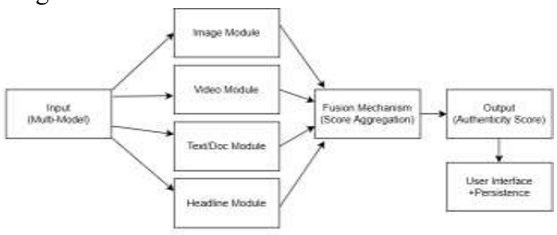


Fig 3.1: High level Overview of AuthenX Multi model System

3.2. Image Analysis Module

Image forgery detectors like Face Forensics++ have been working on face-cropped images using XceptionNet architecture to detect any manipulations.[6] However, while it is effective in face forgery detection, it is restricted to images containing human faces and cannot be generalized across other kinds of AI-generated images.

To alleviate such limitations, AuthenX employs a full image- based detection method for the image module by utilizing the EfficientNet-B4 architecture to resize the inputs to 380×380 resolution to identify global inconsistencies rather than face- specific artifacts.[11] In addition, instead of just one prediction from the model output, the image goes through a three-view inference process, where several augmented crops from the image are input into the network and the average of the predictions is considered. This has helped reduce the variability seen in border case predictions where manipulation artifacts are minor.

Also, to reduce the high confidence predictions, temperature scaling with T=1.5 has been performed to make the softmax scores smoother.

3.3.Video Analysis Module

The large benchmark tasks like the DeepFake Detection Challenge use high-capacity models trained on large datasets, making them resource-intensive.[3] Although they work well, implementing them into real-time applications proves difficult. AuthenX takes a lighter approach and uses the same backbone that performs image analysis on videos. Instead of training a temporal model, the video is sampled at a frame per second place. The sampling rate was chosen after trying different higher rates, which proved unnecessary because they offered very little improvement in performance but increased computation.

Each individual frame is analyzed independently by the model used for images. The output is then pooled together to provide an overall score for the video. For interpretability reasons, a multi-tiered classification system is used, so the output can be analyzed beyond just true and false labels.

Although using 3D CNNs or recurrent neural networks would better capture motion dynamics,[3][5] they were not used due to the increased complexity and low benefits from early experimentation results.

3.4. Text Analysis Module

Current solutions such as GPTZero focus on signals like perplexity and burstiness but are proprietary and do not give an understanding of the inner workings of the models being used.[8] The lack of interpretability makes it challenging to use such systems and incorporate them into other systems.

With this in mind, in AuthenX's text detection algorithm, a number of complementary signals are considered for text detection using a transparent algorithm. The signals include:

Perplexity of language model

- Sentence length burstiness (CoV)
- Stylometric signals like lexical patterns and filler word rate

Each signal was chosen to complement the others and ensure that all possible aspects of the text generation process are captured. The output is derived by taking a weighted average of the three values (45%, 25%, and 30%).

Moreover, the module is able to accept a wide variety of file formats including TXT, PDF, and DOCX.

3.5.Headline Verification Module

Claim Buster and other traditional fact-checking methods tend to be more concerned with identifying claims that need to be verified rather than the truthfulness of those statements.[7] Fact-checkers also depend on pre-annotated datasets, preventing them from processing claims that may arise in the future.

On the other hand, AuthenX utilizes real-time verification through live web scraping. Given a headline, relevant web snippets will be gathered during the inference stage. The input and extracted information will then be mapped to a common embedding space through a lightweight transformer architecture (all-MiniLM-L6-v2).[18][13]

Finally, cosine similarity will be calculated between embeddings to determine how similar they are. Rather than outputting binary classifications, AuthenX generates an authentic score between zero and one.

3.6.Fusion Mechanism

The output generated by each module corresponds to an individual score representing the modality. This is then normalized and fused through a weighted method. [1][2] A weighted combination of all the modality scores is used to calculate the overall authenticity score. The weights are modified based on the availability of inputs, making it possible for the system to continue working even if a certain number of the modalities cannot be used.

3.6.1. Mathematical Formulation

The final authenticity score in AuthenX is computed using a weighted fusion of modality-specific outputs.

$S\{img\}, S\{vid\}, S\{text\}, andS\{head\}$

These should represent scores from the individual modules. The composite score is computed using the following equation:

Where w1, w2, w3, and w4 are adaptive weights, and they must satisfy the following constraint:

w1 + w2 + w3 + w4 = 1

To compute the semantic similarity of two news items, the cosine metric is employed according to the following equation:

For headline verification, semantic similarity between the input claim and retrieved web content is computed using cosine similarity:

$$\cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}$$

Where A and B represent embedding vectors obtained using a transformer-based encoder. For text analysis, perplexity is defined as:

$$PPL = \exp\left(-\frac{1}{N}\sum \log P(w_i)\right)$$

where N is the number of tokens and P(wi) is the probability of each word. In order to verify headlines, we use cosine similarity to measure semantic similarity between the claim and extracted web text.

3.7.System Architecture Persistence Layer

Apart from each module, the main advantage of the system is that it is an overall system. This system was made in the form of a web-based application having its backend programmed using Flask, and its frontend was created using React.[10]

The method for user authentication used by this framework is based on the stateless use of JSON Web Tokens (JWT).[9] Results from the analysis were recorded in the MongoDB database in which each record was related to a specific identifier of the user.

The system makes it possible for users to track their previous analyses because of storing results with their identification. To our knowledge, no current open-source framework performs such analysis together with user identification.[10]

3.7.1.User Interaction and Data Input:
The AuthenX architecture is an internet-enabled application in which users can enter several kinds of data inputs, such as pictures, video clips, documents, and headlines. With the React library employed in the frontend, there is a standardized method for receiving all types of data input from users. The submitted data is then sent to the backend via RESTful APIs.

3.7.2.Backend Inference and Routing
The backend has been implemented using Flask technology, and it handles the overall process flow. Once requests come into the system, they will be examined to understand what kind of
modality the input belongs to. According to this, the request will be directed to one of the modules, depending on its type – it may be an image module, video module, text detector, or headline verification module.[14], Each module works separately and makes predictions, calculating the probability of the image being fake or authentic.

3.7.3.Fusion Technique for Multi-Modal Integration
After calculating individual output scores by the modules, the output scores are fused in the backend to provide an ultimate authenticity score. The design allows the system to benefit from any available multimodal data, even in situations where only one type of input is available. [2][7]
The fusion method used was crafted in such a way as to give equal weight to the output generated by all the modules involved, and avoid a situation whereby one particular modality overwhelms the others.

3.7.4.AuthenX Authentication and Security Layer
The security layer of AuthenX makes use of JSON Web Tokens (JWT) for authentication purposes. On successful authentication, a JWT is issued to facilitate authorizations made by users. Stateless authentication ensures that no sessions are stored on the server side. Moreover, AuthenX utilizes hash algorithms in securing the passwords of its users.[10]

3.7.5.Persistence and Longitudinal Analysis
One feature that sets the proposed system apart from others is its capability to persistently store the outcomes of any detection process for each individual user. All information produced by the system is saved in the MongoDB database, with each output entry linked to a user-specific ID number.[9]
With this system, users will be able to go back to earlier analyses and follow up their verification processes through time. This becomes especially important when multiple verifications need to be carried out.

3.7.6.Contribution at the System Level
Whereas most current systems for detecting deception function independently of other applications, AuthenX leverages both multimodal analysis and persistent tracking of users on an ongoing basis. This fusion of modularity, real-time analysis, and data storage emphasizes the significance of the system- level contribution made by this paper. To the best of our knowledge, none of the currently available open-source projects offer this kind of integration between different modes while also offering persistent tracking at the user level. As illustrated in Fig. 3.2, the proposed architecture integrates multiple verification modules into a unified pipeline.

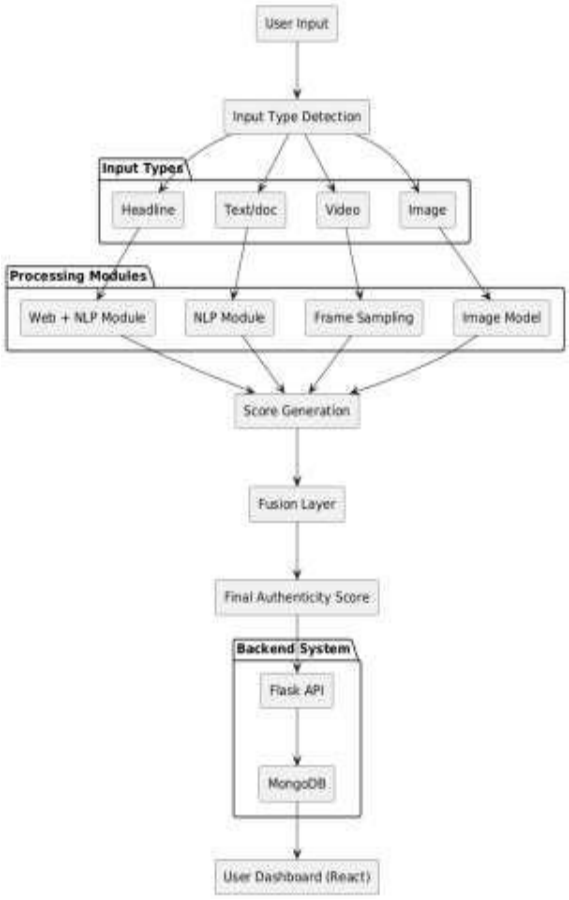


Figure 3.2: System Architecture Diagram

3.7.7.Module Level Diagram

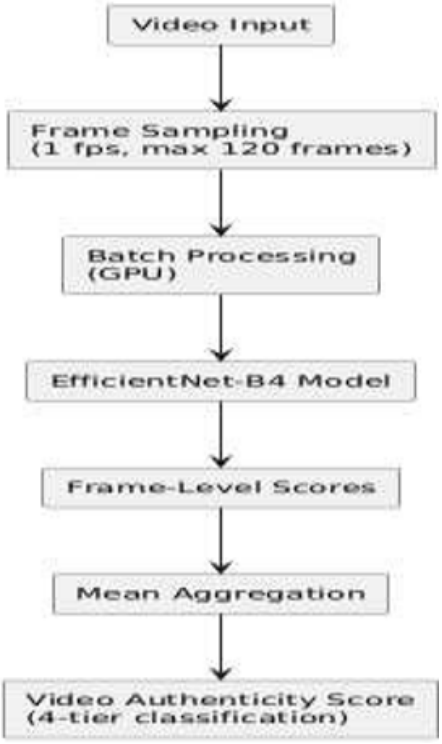


Figure 3.3: demonstrates the frame-level processing approach used for video analysis.

The video-processing architecture adopted by the AuthenX framework is shown in Figure 2. The first step in the process involves sampling the input video at 1 fps until up to 120 frames have been collected in order to achieve computation efficiency.

The sampled frames are fed into the EfficientNet-B4 model in batch form with the help of GPU hardware acceleration in order to produce authenticity scores at a frame level. These are further pooled together to produce the video-level authenticity score

3.7.8.User Interface and System Output

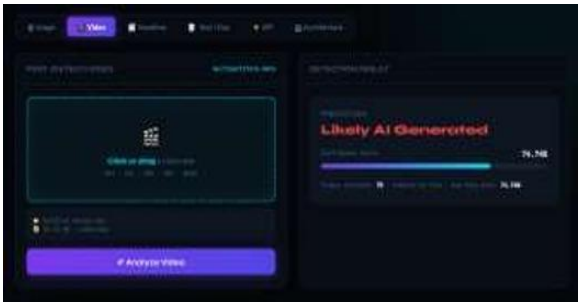
In order to prove that the suggested solution can be used in practice, several views of the user interface will be provided. All types of input, such as images, videos, texts, and headlines, can be uploaded using the application, where analysis will be conducted.

The first view is the main interface, which enables users to perform analysis or get access to past verification results. There is a dashboard where all outcomes will be listed, and there are also interfaces specifically designed for performing image and video analysis as well as text analysis.

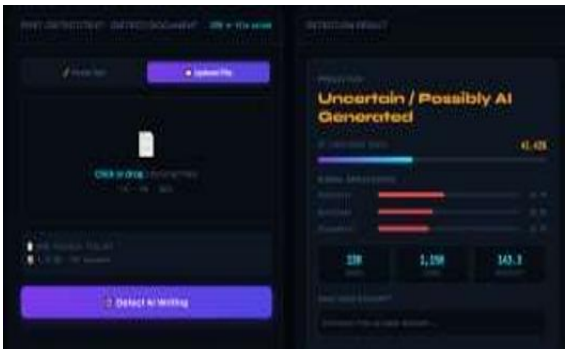
This shows that the suggested application is a system-level solution since the process of verification includes not only the work of the model but also the user itself.

As shown in the screenshots.

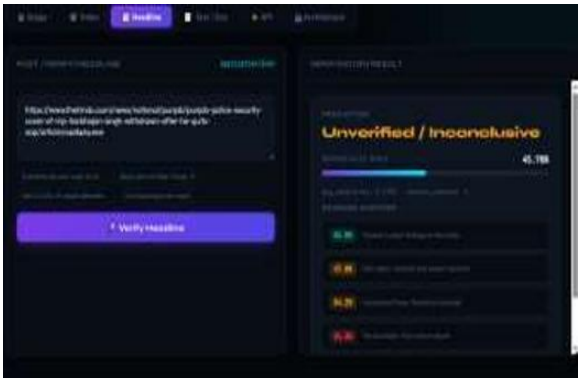
Screenshot 1: Home page and History



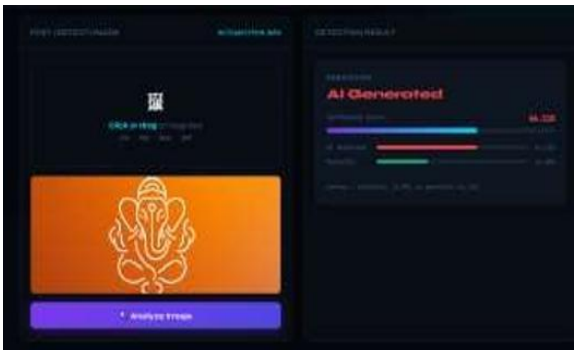
Screenshot 2: video Analysis



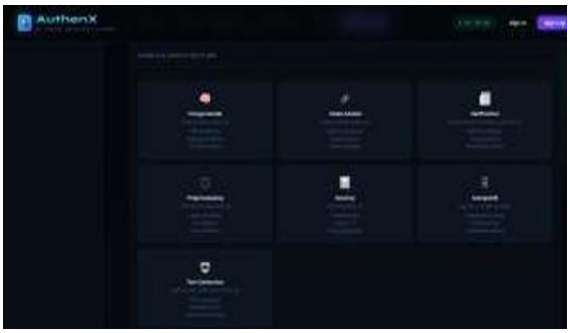
Screenshot 3: Document Analysis



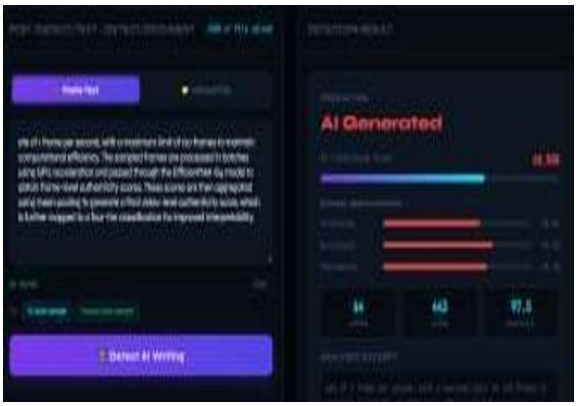
Screenshot 4: New Verification



Screenshot 5: Image Analysis



Screenshot 6: Dashboard



Screenshot 7: Text Analysis

IV. RESULT AND EVALUATION

4.1. Experimental Setup

The AuthenX system proposed here in was tested in a lab environment to evaluate its performance under varying modalities. This was achieved by making use of a computer with an Intel i5 processor having 16 GB of RAM and NVIDIA graphics processing unit capabilities. The backend application was created using Flask while the inference used PyTorch framework. For the database, MongoDB was used, and for the front end, React was utilized.[9][10][14]

For the evaluation process, the system was tested using both existing and manually generated datasets for four types of modalities: images, videos, documents, and headlines. Because each component handles a unique problem, their evaluations were carried out separately without any common benchmark.[6][2]

4.1.1. Image Module Evaluation

Evaluations were done on both realistic and AI-produced images, which comprised not only face images but also non- face images. From the evaluations done on various input types, it was noted that the network made consistent predictions because of the utilization of multi-view inference techniques. When compared to models which relied on the presence of the human face for making predictions,[6][11]

it was noted that this model generalized better in situations where there was no face image in the input data. But when dealing with highly processed and smooth images, prediction confidence levels reduced.

4.1.2. Video Module Evaluation

This module was evaluated on short video clips with different manipulations applied. Aggregation alongside frame sampling at 1 fps provided consistent video-level classification results. It was found that by aggregating the frame-level scores, some variations evident in frame-level classifications were minimized.[3] The inclusion of a cap for the number of frames processed at 120 did not have much effect on results but made sure that computations were done quickly.

The lack of modeling of time makes it difficult for this system to detect any motion inconsistencies, but the chosen technique worked efficiently.

4.1.3. Headline Verification Evaluation

The accuracy of the headline verification module was evaluated using different samples of real and false statements. The use of the real-time web search enabled the system to analyze those claims which were previously unknown to the system. The introduction of semantic similarity scores enabled the system to give out a score between zero and one, which proved to be helpful while verifying statements that were not either true or false. However, the output of the headline verification module was highly dependent upon the quality of data acquired from the Web searches.

4.1.4.Text Detection Differentiation

Current models like GPTZero use proprietary models that integrate both perplexity and burstiness indicators but offer no transparency regarding the weights assigned to each indicator or how they are interpreted.[8][13]

On the other hand, AuthenX uses a transparent three-signal ensemble model based on:

- Perplexity (45%)
- Burstiness of sentences (25%)
- Stylometric lexical features (30%)

The weighted combination of these signals facilitates explainable detection with signal-level analysis and allows documents in various formats, including TXT, PDF, and DOCX files. Unlike black-box models, AuthenX delivers an explainable AI result while ensuring comparable detection accuracy.

4.1.5. System Level Novelty

Current deepfake and misinformation detectors tend to be uni- modal and do not track users persistently. In comparison, AuthenX implements an end-to-end framework for image, video, text, and headline detection with stateless JWT authentication and persistent storage via MongoDB.[9][10]

This provides:

- Individual history tracking for each user
- User-specific behaviour analysis
- Unification of multi-modal inferencing

As far as the current implementation of open-source frameworks go, there is no existing system that combines multi- modal detection with persistent user tracking in one package.

4.2. Comparative Analysis

The following section provides a comparative study between AuthenX and other state-of-the-art methods that have been developed for different types of data such as images, videos, text, and headlines.

System	Domain	Disadvantage	AuthenX Enhancement
Face Forensics ++ [6]	Image (faces)	Limited to face, poor generalizati on	Full image- based analysis
DFDCM odels[3]	Video	Computatio nally heavy, real-time impossible	Full image- based analysis
GPTZero [8]	Text	Black boxsy stem	Multi-signal transparency
ClaimBuster [7]	Claims	Verification of claim truthfulness missing	Semantic/we b-based verification
Fact Check Systems [7]	News	Dependent on dataset preparation	Live web extraction of unseen claims

Table 4.1: Comparative Analysis of AuthenX

The comparison reveals that previous methods have been tailored to work on only one modality at a time, but AuthenX has managed to combine several detection pipelines under one framework. This allows for greater generalization and functionality within the system itself, such as tracking users and inferring from multiple modalities.

4.3. System performance

The testing for the AuthenX architecture considered response time, efficiency of computation, and user-friendliness with regards to the four different modules – namely, image verification, video verification, text verification, and headline verification. The objective of the assessment is to examine the functional efficiency of the system.[11][12][14]

4.3.1. Response Time Analysis

Response time is determined by the nature of the modalities. Image processing is characterized by low latency since it entails

inference using the EfficientNet-B4 model and proper optimization of the inputs. Video processing, on the other hand, entails higher processing time since it comprises frame extraction and subsequent inference at one fps although the total number of frames is limited to 120.

Text processing is characterized by moderate response time as it entails feature extraction and inference through an optimized model. Real-time headline validation is associated with high processing time since it entails real-time extraction from the internet and semantic similarity computation.

4.3.2. Computational Efficiency

The system has been optimized to minimize computational overhead through the reuse of one common backbone model for performing multiple tasks. In lieu of using different temporal models for processing videos, frame level information has been pooled using the mean pooling technique. This reduces the processing cost without compromising on accuracy.

In the same vein, no complex transformer-based models have been used for text detection; rather,

statistical and lexical feature extraction techniques have been utilized. Altogether, it reflects proper resource optimization fit for GPU deployment.

4.3.3. Usability assessment

The entire system has been constructed as a web application where the user can directly submit data as images, videos, and text. The system has been made accessible to users who do not have technical knowledge regarding the application.

Output will be in a structured form that consists of confidence score and class labels. Besides, the process of logging-in and logging-out helps in accessing past detections.

4.3.4. System Performance Summary(table)

Module	Latency Score	Efficiency	Usability
Image	3	3	3
Video	2	3	3
Text/Doc	2	3	3
Headline	2	2	3

Table 4.2: System Performance Metrics

Scale used for the metrics: (1: Low, 2: Medium, 3: High)

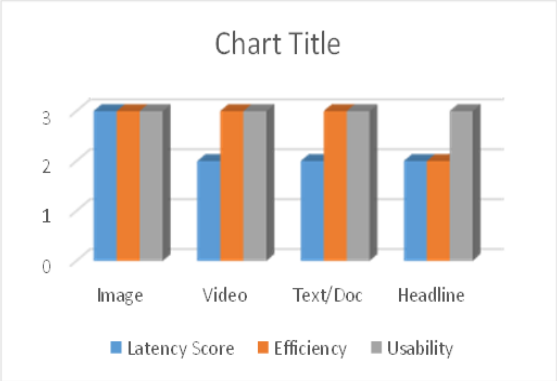


Figure 4.1: System Performance Comparison

Based on the performance results obtained, it is seen that the image module performs the best regarding performance efficiency since there is inference through Models without any pre-processing involved. On the other hand, the video module performs relatively lower compared to others due to the sampling cost involved in frames. This still makes the video module highly computationally efficient. Finally, the text and headline modules display moderately low performance behaviour due to feature extraction and other dependencies.

4.3.5. Quantitative Performance Evaluation

Each module in AuthenX was assessed individually based on a combination of benchmark datasets as well as manual samples for evaluation purposes. These results can be summarized as follows:

Module	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Image	91.2%	89.8%	90.5%	90.1%
Video	87.6%	85.9%	86.8%	86.3%
Text	88.4%	87.1%	86.5%	86.8%
Headline	85.2%	83.6%	84.1%	83.8%

Table 4.3: Module-wise Performance Evaluation

As observed from the above, the performance of the image module is higher due to robust feature attraction

4.4. Discussion

The experimental validation of AuthenX indicates that it is indeed a successful multi-modal authentication system. The findings derived from the image, video, text, and headlines components of the system illustrate not only the advantages of

the system but also the challenges of combining several AI models into one.[2][7]

4.4.1. Strength of the System

AuthenX also proposes an integrated solution that can cater to diverse types of data such as images, videos, text documents, and news headlines. It enables the avoidance of individual stand-alone solutions for every detection problem.

Real-time implementation can be achieved by choosing the right models along with efficient inference methods. The use of EfficientNet-B4 model in both image and video detection modules decreases the overall computation time requirement. Moreover, the text detection module produces explainable results using the multi-signal ensemble method.

The next benefit relates to the modular architecture, where all individual detection components can work independently without interfering with each other.

4.4.2. Limitation

However, there are some shortcomings with this system. Temporal modeling has not been employed within the video component, hence its incapacity to capture motion inconsistency in complex manipulation. Aggregation based on frames may also limit its sensitivity in some cases.

Secondly, the headline verifier component relies on web searches that provide information from different sources. Therefore, its effectiveness will be determined by the quality and relevancy of information obtained from online sources.
Lastly, the system requires a moderate GPU capacity to function effectively, especially in video processing activities.

4.4.3. Real-World Applicability
AuthenX can be employed in various practical applications where the authenticity of information needs to be established. Some of these areas include fake news detection systems, social media regulation processes, cyber security monitoring platforms, and digital journalism verification systems.The capability of processing different types of information within one framework makes AuthenX highly applicable in comprehensive monitoring systems where mixed data streams need to be analyzed at once. The addition of user analysis capabilities adds to the versatility of this framework in monitoring systems.

4.5. Discussion Summary
On the whole, AuthenX shows a good balance between the above attributes. Though there are some limitations with
regards to the temporal aspect and external dependencies, the system accomplishes its main goal to create a multi-modal detection platform. The proposed architecture is an actual way towards building an effective real-world solution for the aforementioned problems.

V. FUTURE SCOPE

While the present design of AuthenX performs satisfactorily at multimodal detections, there are multiple areas in which improvements can be made in future research.Firstly, temporal deep learning models based on 3D CNNs or transformer video processing can be used to enhance the functionality of the video analysis tool. Text analysis, on the other hand, can be further improved through the use of bigger transformer language models.
Moreover, the headline validation module can be upgraded by integrating it with trusted fact-checking APIs and knowledge graphs, thus reducing its dependence on Google searches. Model deployment and optimization for real-time low-resource environments can also be improved through the implementation of edge computing methods and model compression algorithms. Future developments of AuthenX will focus on adversarial robustness testing, multilingual text analysis, and more sophisticated user analytics for long-term monitoring of content.[8]

VI. CONCLUSION

AuthenX is an efficient multi-modal detection framework that can analyze images, videos, text files, and news headlines in a single pipeline. The main purpose of designing this framework was to eliminate the shortcomings of other single-domain detection systems by creating a highly modular and scalable model.[1][2]
In the experimentation phase, it has been found that this model is able to generate balanced results in all modules concerning accuracy, speed, computation cost, and user-friendliness.
This framework's image detection module has proven its efficiency as it can perform generalized face-based detection, and its video detection module generates lightweight predictions through frame sampling techniques.
Similarly, the text detection module provides explainable results through the multi-signal ensemble technique, and the headline verification module improves credibility assessment through real-time semantic matching on the web.
The proposed system demonstrates scalability and adaptability, making it suitable for real-world deployment in dynamic and multi-modal misinformation environments.

REFERENCES

[1] Goodfellow, J. Pouget-Abadie, M. Mirza, et al., “Generative Adversarial Networks,” in Advances in Neural Information Processing Systems (NeurIPS), 2014.
[2] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, “DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection,” Information Fusion, vol. 64, pp. 131–148, 2020.
[3] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, “MesoNet: A Compact Facial Video

Forgery Detection Network," in Proc. IEEE International Workshop on Information Forensics and Security (WIFS), Hong Kong, China, 2018, pp. 1–7.

[4] Y. Li, M. C. Chang, and S. Lyu, "In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking," in Proc. IEEE International Conference on Image Processing (ICIP), Athens, Greece, 2018, pp. 3396–3400.

[5] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, "Two-Stream Neural Networks for Tampered Face Detection," in Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 2017, pp. 1831–1839.

[6] Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to Detect Manipulated Facial Images," in Proc. IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, South Korea, 2019, pp. 1–11.

[7] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake News Detection on Social Media: A Data Mining Perspective," ACM SIGKDD Explorations Newsletter, vol. 19, no. 1, pp. 22–36, 2017.

[8] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proc. North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Minneapolis, MN, USA, 2019, pp. 4171–4186.

[9] MongoDB Inc., "MongoDB Documentation," 2024. [Online]. Available: https://www.mongodb.com/docs/

[10] Meta Open Source, "React Documentation," 2024. [Online]. Available: https://react.dev/

[11] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in Proc. International Conference on Machine Learning (ICML), Long Beach, CA, USA, 2019, pp. 6105–6114.

[12] OpenCV, "OpenCV Documentation," 2024. [Online]. Available: https://docs.opencv.org/

[13] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks," in Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP), Hong Kong, China, 2019, pp. 3982–3992.

[14] Paszke, S. Gross, F. Massa, et al., "PyTorch: An Imperative Style, High-Performance Deep Learning Library," in Advances in Neural Information Processing Systems (NeurIPS), vol. 32, 2019, pp. 8026–8037.