



iJRASET

International Journal For Research in
Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** X **Month of publication:** October 2026

DOI: <https://doi.org/10.22214/ijraset.2026.85260>

www.ijraset.com

Call: ☎ 08813907089

E-mail ID: ijraset@gmail.com

DeepFake Shield: A Multi-Modal Generative AI Fraud Detection Pipeline with Explicit Uncertainty Handling

Mrs. Reena Dewangan¹, Ms. Yogini Verma², Ms. Sanskruti Chaple³, Ms. Lingam Sharon⁴, Ms. Kritika Tandulkar⁵, Mr. Om Margamwar⁶, Mr. Vansh Bangre⁷

¹Assistant Professor, Department of Computer Science and Engineering, G H Rasoni Skill Tech University, Nagpur

^{2, 3, 4, 5, 6, 7}Research Scholar, Department of Computer Science and Engineering, G H Rasoni Amravati University, Nagpur

Abstract: Synthetic media is now cheap to make and hard to judge by eye. A cloned voice can approve a payment, a generated face can get past a casual identity check, and a fluent paragraph written by a language model can pass for a statement from a real person. Most detectors cover only one of these cases, and almost all return a confident verdict even for input far outside their training data. This paper presents DeepFake Shield, a Flask-based web application that screens images, text, audio and video through one interface. Each modality is served by a pre-trained deep learning classifier, and a threshold layer turns the raw model score into one of three outcomes: authentic, synthetic or inconclusive. For video, the visual and audio streams are analysed separately and any disagreement between them is reported as a conflict. The paper gives a detailed justification for the topic, describes the architecture and decision logic, and sets out an evaluation protocol based on precision, recall, F1, ROC-AUC and Brier score with thresholds tuned on validation data only. We report the engineering checks completed so far and state clearly that they are not accuracy measurements.

Keyword: Deepfake detection, Multi-modal fraud detection, Generative AI, Synthetic speech, AI-generated text detection, Selective classification, Abstention, Uncertainty handling, Flask web application, Transformers, Digital forensics, Misinformation.

I. INTRODUCTION

Generative models now sit inside ordinary tools. Anyone with a laptop can swap a face in a video, clone a few seconds of speech, or ask a language model to draft a convincing e-mail. The harm that follows is not hypothetical. Identity fraud, impersonation of company executives, fabricated evidence and large-scale misinformation all rest on the same capability, which is producing believable content at almost no cost.

The defensive side has not kept up in usability. A journalist who receives a suspicious video may need one tool for the frames, another for the soundtrack and a third for the caption that came with it. A content moderator faces the same split. Worse, many tools give back a single number and leave the user to guess what it means, even though the raw scores of neural networks are known to be poorly calibrated [8].

We built DeepFake Shield to close two practical gaps. The first is fragmentation: one application should accept an image, a block of text, an audio clip or a video and send it to the right detector. The second is overconfidence: the system should be allowed to say that it does not know. The main contributions of this work are the following.

- A unified screening pipeline for four modalities built on open pre-trained models.
- A three-way decision rule with an explicit inconclusive class that is stored and counted separately.
- A video procedure that analyses frames and audio independently and flags disagreement between them.
- An evaluation protocol that keeps engineering smoke tests apart from accuracy claims, together with a plain statement of limitations.

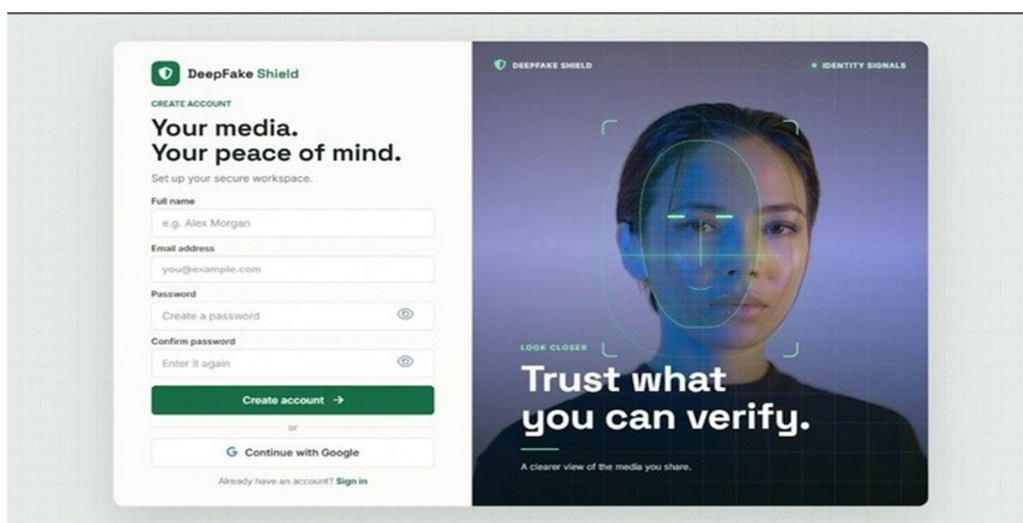
II. RELATED WORK

Mirsky and Lee reviewed how deepfakes are created and detected, and organised the field around reenactment and replacement attacks [1]. Tolosana et al. looked at face manipulation more specifically and separated entire-face synthesis, identity swap, attribute manipulation and expression swap, with detection methods grouped under each [2].

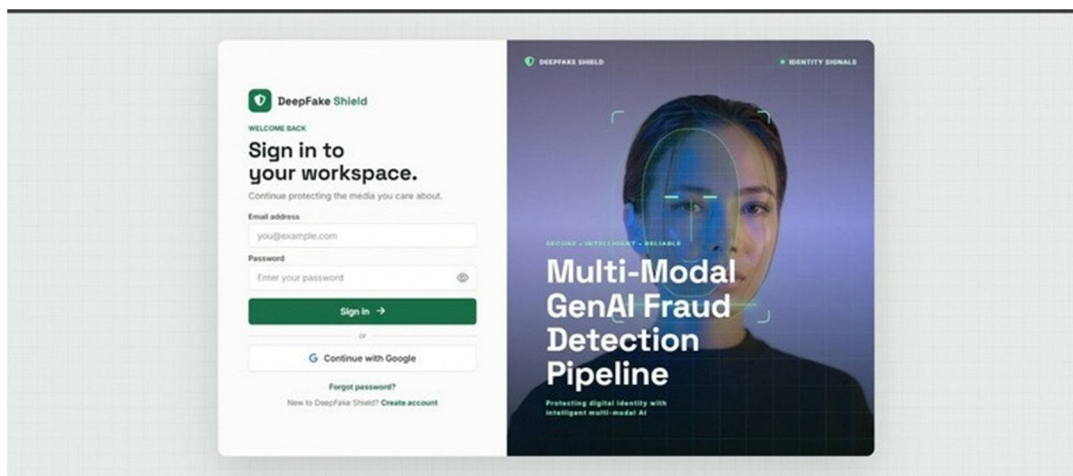
Benchmarks made comparison possible. Rössler et al. released FaceForensics++, a collection of manipulated face videos produced with several methods at different compression levels, and showed that detection becomes harder as compression increases [3]. Dolhansky et al. published the DeepFake Detection Challenge dataset with more than one hundred thousand clips recorded with paid actors, and the challenge results showed that detectors struggle to generalise to unseen manipulations [4]. Li et al. introduced Celeb-DF, a set of higher-quality fake videos on which earlier detectors performed noticeably worse [5].

For speech, the ASVspoof 2019 challenge covered both logical access attacks (speech synthesis and voice conversion) and physical access attacks (replay) [6]. For text, Guo et al. built the HC3 corpus, which pairs answers written by human experts with answers from ChatGPT, and used it to evaluate detectors [7]. RoBERTa [9] is the pre-trained encoder most commonly fine-tuned for this task.

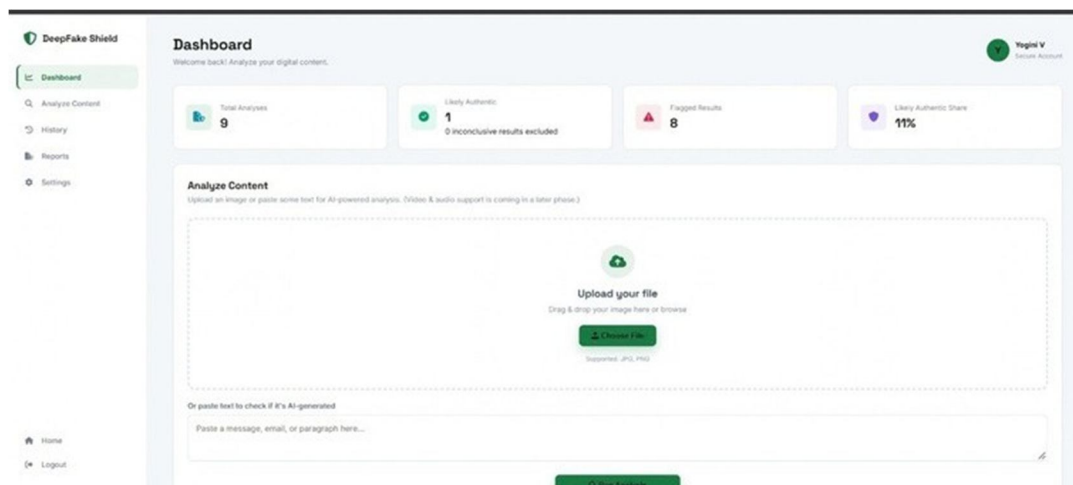
Our treatment of uncertainty draws on older work on classification with a reject option. Chow described the trade-off between error rate and rejection rate [10], and Geifman and El-Yaniv brought the idea to deep networks as selective classification [11]. Guo et al. showed that modern networks tend to be over-confident and that simple post-hoc scaling can improve calibration [8]. Together these papers support a simple design principle: when the model is unsure, it is better to abstain and hand over to a person than to guess.



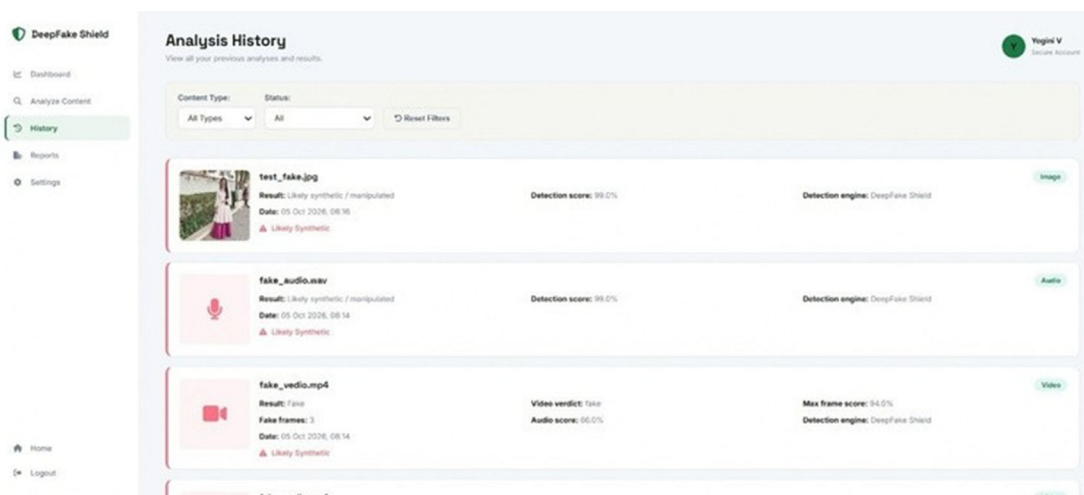
The login interface maintains visual continuity with the signup process, utilizing the same dual-panel configuration optimized for returning users. The input area transitions to the "Sign in to your workspace" configuration, prompting users for their registered Email Address and Password to continue protecting their media assets. A secure "Sign in" navigation trigger processes the credentials, supported by text links for account recovery and new user registration. The right-hand graphic updates its text messaging to introduce the platform's architectural capability: a "Multi-Modal GenAI Fraud Detection Pipeline" engineered to safeguard digital identities



DeepFake Shield features a modern split-screen authentication interface designed for user onboarding. The left panel houses the registration workspace under the title "Your media. Your peace of mind," providing intuitive text fields for Full Name, Email Address, Password, and Confirm Password verification. Users can register directly via a prominent green "Create account" call-toaction or utilize a secondary single-sign-on option via "Continue with Google." The right panel serves as a visual anchor, showcasing a high-tech graphic of a person's face overlaid with a biometric scanning mesh alongside the brand's core ethos: "Trust what you can verify"



The central command workspace provides an optimized environment for digital content analysis, anchored by a persistent left-hand navigation sidebar for rapid system transit. The top header features user profile indicators alongside a high-level metrics layout that breaks down system usage into distinct data cards: Total Analyses (9), Likely Authentic (1), Flagged Results (8), and a Likely Authentic Share (11%) tracking index. The primary screen real estate is dedicated to the "Analyze Content" workspace, which features a central file ingestion drop-zone supporting direct JPG and PNG uploads, complemented by a lower secondary text input area for evaluating copy-pasted written text



The Analysis History log displays a comprehensive ledger of all past media validation requests, complete with structural dropdown filters to sort data by Content Type and Status. The historical records dynamically categorize media types, explicitly showcasing processing rows for images, audio tracks, and full video file formats. Each individual ledger entry outputs critical diagnostic variables generated by the system, including the file name, precise timestamp, categorical result verdict, an engine attribution tag, and highly accurate metric evaluations such as 99.0% Detection scores for manipulated assets

III. PROBLEM STATEMENT

This section explains in detail why the topic is worth studying and why the design choices of DeepFake Shield follow from the problem and not from convenience.

A. *The threat is real and already costly*

Deepfake fraud has moved from demonstration to practice. In 2019 a UK energy company was reported to have transferred money after a caller imitated the voice of its parent company's chief executive [12]. In early 2024 a finance employee at a multinational firm in Hong Kong was reported to have paid out about 25 million US dollars after a video call in which colleagues, including the chief financial officer, appeared to be present but were in fact synthetic [13]. In both cases the victim did what most people would do: they trusted a familiar face or voice. The lesson is that a check made by human eye and ear is no longer enough, and that a tool must support the person making the decision.

B. *Why more than one modality must be covered*

Real fraud rarely uses one medium. A phishing message may be machine-written text, followed by a cloned voice call, followed by a video meeting with swapped faces. A detector that covers only images leaves the other three doors open. Mixed-media forgeries are even harder: a genuine video with a replaced soundtrack passes a purely visual check, and a genuine recording under generated frames passes a purely audio check. Screening all four modalities in one place, and comparing the visual and audio verdicts for video, is therefore a requirement of the problem and not a feature added for show.

C. *Why existing tools fall short for ordinary users*

Research detectors are usually released as code for one modality, tested on one benchmark, and run from a command line. Commercial tools are often closed, paid, and give little explanation. The people who most need help, such as journalists, moderators, teachers and small businesses, are rarely machine-learning engineers. They need an interface that accepts a file, returns a clear answer and keeps a record. A web application with accounts, a private history and a CSV export meets this need directly.

D. *Why an inconclusive class is necessary*

Detectors are known to degrade on generators they have not seen [3, 4, 5], and their scores are not calibrated probabilities [8]. A system that always chooses between real and fake is therefore guaranteed to be confidently wrong on some inputs. In a fraud setting the two kinds of error carry different costs. A false positive may wrongly accuse a person or discredit an authentic recording. A false negative may clear a forgery. By adding an explicit inconclusive outcome, with thresholds set on validation data, the system trades some coverage for reliability and passes the hardest cases to a human reviewer. This follows the established theory of classification with a reject option [10, 11].

E. *Why open pre-trained models and local inference*

Training a detector for each modality from scratch needs large labelled datasets and substantial compute, and the result would age quickly as generators change. Using publicly available checkpoints makes the system reproducible, lets other researchers swap in better models, and keeps the cost low. Running inference locally means that sensitive uploads, such as a private voice recording or an identity photograph, do not have to be sent to a third-party service, which addresses an obvious privacy concern.

F. *Who benefits*

- Journalists and fact-checkers who must decide whether to publish a received video or audio clip.
- Platform moderators who need a first-pass filter and a clean way to route doubtful items to manual review.
- Organisations that want to protect staff identities and keep an audit trail of what was checked and when.
- Researchers who need a harness for measuring precision, recall and calibration of detectors on labelled data.
- Ordinary users and content creators who want to check an image or message before trusting or posting it.

G. *Research questions*

The work is guided by three questions. First, can four detectors for different media be combined behind one interface without hiding their limits? Second, does a three-way decision rule give a more honest and more useful output than a binary one? Third, how should such a system be evaluated so that engineering checks are not mistaken for accuracy evidence? The rest of the paper addresses these questions in turn.

IV. METHODOLOGY

A. Models and Resources

No model was trained from scratch in this work. Each modality uses a publicly available pre-trained checkpoint from the Hugging Face hub. Table I summarises the inputs and the main limit of each detector. Labelled data are needed only for evaluation, and we plan to use public benchmarks such as FaceForensics++, Celeb-DF, DFDC, ASVspoof 2019 and HC3 for that purpose [3, 4, 5, 6, 7].

Modality	Model	Accepted input	Main limitation
Image	CNN classifier (dima806 checkpoint)	JPEG, PNG, GIF	Unseen generators; drift from training data
Text	RoBERTa detector (Hello-SimpleAI checkpoint)	English, at least 80 words, up to 500,000 characters	English only; trained on human vs. early ChatGPT text
Audio	Voice classifier (MelodyMachine checkpoint)	Speech, up to 10 minutes	Not validated for music or general audio
Video	Frame sampling plus audio analysis	Common video containers	32 frames can miss brief edits; no lip-sync check

Table I. Modalities handled by the pipeline.

B. Data Pre-processing

Before inference, each input is converted into the form its model expects. For images, files are opened with Pillow, converted to a standard colour format and resized to the model input size. For text, the input is checked against the minimum of 80 words and the maximum of 500,000 characters, and short or non-English text is flagged with a warning because the detector is not reliable there. For audio, speech is loaded with librosa and soundfile, resampled, and split into segments when the clip is long; silent audio is treated as an edge case and reported as such. For video, OpenCV extracts 32 sampled frames and the audio track is separated for the audio detector.

C. Modeling

1) Image detector

A convolutional classifier scores each image between 0 and 1. It is the same detector that is applied to every sampled video frame. It is trained on specific generators, so it cannot be expected to recognise images from new ones, and this is stated in the interface.

2) Text detector

A RoBERTa-based classifier [9] scores a passage as human-written or machine-written. It was trained on human answers against early ChatGPT answers [7], so newer models, paraphrased text and heavily edited text are likely to be handled worse.

3) Audio detector

A transformer-based classifier scores speech for signs of synthesis. Long recordings are segmented and the segment scores are combined. It is limited to speech and has not been validated on music.

4) Video analysis

Video is treated as two signals. The visual stream is sampled to 32 frames, each frame is scored by the image detector, and the frame scores are aggregated into one visual verdict. The audio track goes through the audio detector. When the two verdicts disagree, for example convincing frames over a synthetic voice, the system reports a conflict instead of averaging them. A dubbed or voice-replaced clip is exactly the case where an average would hide the problem.

D. Decision Logic

Each detector returns a score s between 0 and 1, which we read as the model's confidence that the input is synthetic. Two thresholds divide the range, as shown in Fig. 1. If s is at or above the upper threshold the result is labelled synthetic. If s is at or below the lower threshold it is labelled authentic. Anything in between is labelled inconclusive and is marked for manual review. One pair of thresholds is kept for each modality and is chosen on validation data only. The score is an uncalibrated model output and is not presented as the probability that the content is fak.

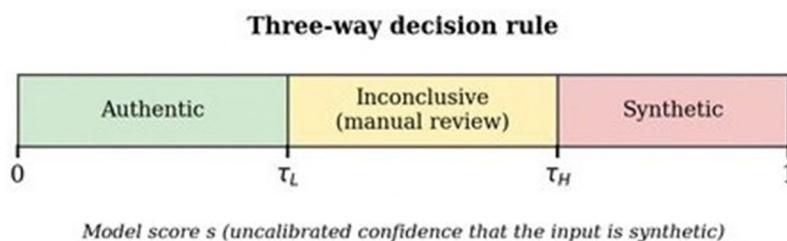


Fig 1. Three-way decision rule with an explicit inconclusive region.

E. System Architecture

The system has three layers, shown in Fig. 2. A browser front end provides the landing page, login and signup, dashboard, upload and analysis page, history, reports and account settings. A Flask back end handles sessions, routing, file upload and persistence. A detection engine selects the modality, loads and caches the model, runs inference and applies the threshold logic. Results are stored through SQLAlchemy, using SQLite during development and PostgreSQL in production. Each analysis record holds the content type, source name, optional image path, label, model score, a threat flag, a list of warnings and a timestamp. Passwords are stored only as hashes created with Werkzeug, and sessions are managed with Flask-Login. History can be filtered by content type and by result, and the records can be exported to CSV.

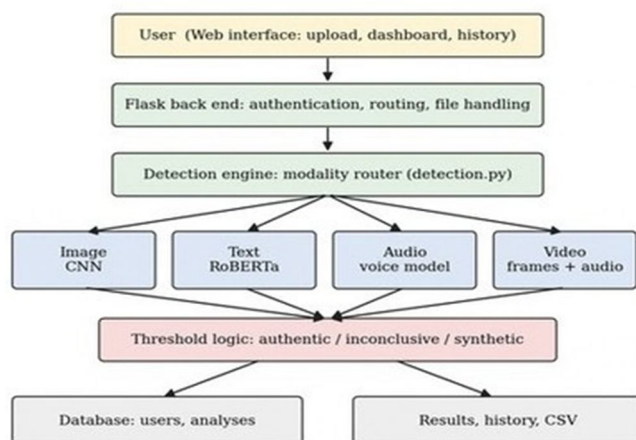


Fig 2. System architecture of DeepFake Shield.

The implementation uses Python 3.11 or later, PyTorch and the Transformers library [14, 15], Pillow, librosa, OpenCV and pytest. For staging the application is containerised with Docker. Production is planned behind Gunicorn or uWSGI and an nginx reverse proxy with TLS and rate limiting.

V. RESULTS AND DISCUSSION

A. Verification completed so far

We separate three kinds of checking because they answer different questions and are easy to confuse. Table II lists what each one tells us.

Check	What it covers	What it does not show
Unit tests	34 pytest cases for label validation, abstention logic, sampling, audio segmentation, video frame extraction and threshold binding. All 34 pass.	Whether the detectors are correct
Smoke tests	All four models load and run inference on generated fixtures.	Any accuracy figure
Labelled evaluation	Precision, recall, F1, ROC-AUC and Brier score on held-out labelled data, with coverage reported for abstention.	Not yet completed for this version

Table II. Verification activities and their meaning.

B. Evaluation protocol

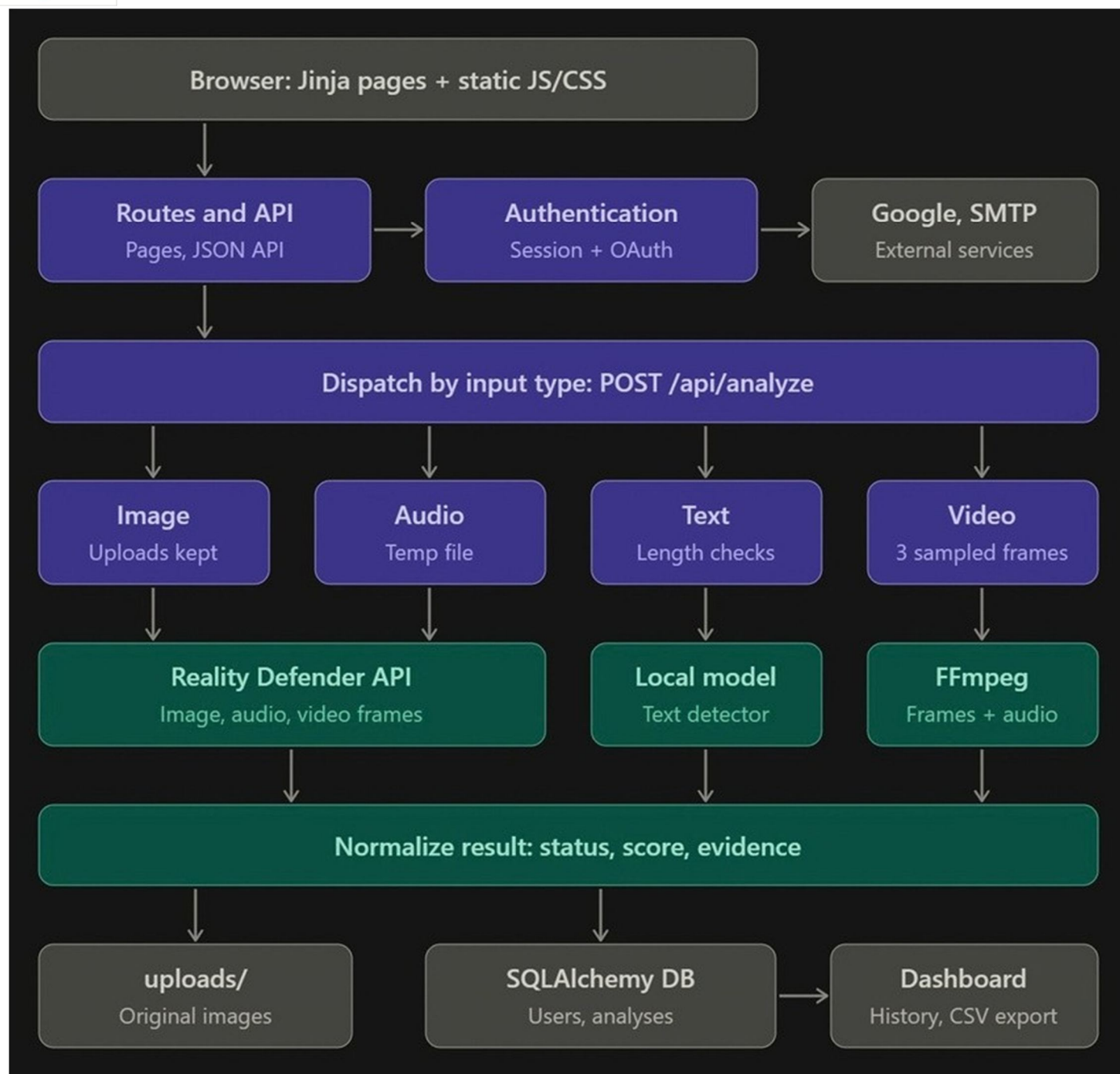
The labelled evaluation is run with a separate script on test sets that do not overlap with the validation data used to set thresholds. Precision, recall and F1 describe the quality of the decisions, ROC-AUC measures how well the scores rank real above fake, and the Brier score [16] exposes calibration problems that ranking metrics hide. Because the system can abstain, we also report coverage, the share of inputs that receive a decisive label, and covered accuracy, the accuracy on that share. Without both, a high accuracy figure could be reached by quietly abstaining on every hard case.

C. Discussion

We have deliberately left accuracy numbers out of this version. Publishing figures from pre-trained checkpoints before the labelled runs are finished would contradict the principle on which the system is built. Perfect scores of 1.0 on every metric, which are sometimes reported in small studies, usually point to a test set that is too easy or too close to the training data. We would treat such a result with suspicion in our own work. The design also sets acceptance targets for response time: under 2 seconds for an image, 3 for text, 5 for audio and 10 for video. These are targets, and measured timings on fixed hardware will be reported with the labelled results.

D. Limitations

The image model cannot be expected to detect generators that were absent from its training data. The text model supports English only, and a false positive on text can harm a real person; writing by non-native English speakers is a known source of misclassification. The audio model is limited to speech. The video procedure samples only 32 frames and does no lip-sync analysis. The tool does not check facts: an authentic label says nothing about whether the content is true. Adversarial evaluation has not yet been done. For these reasons the output should be treated as one input to human judgement and never as a final verdict.



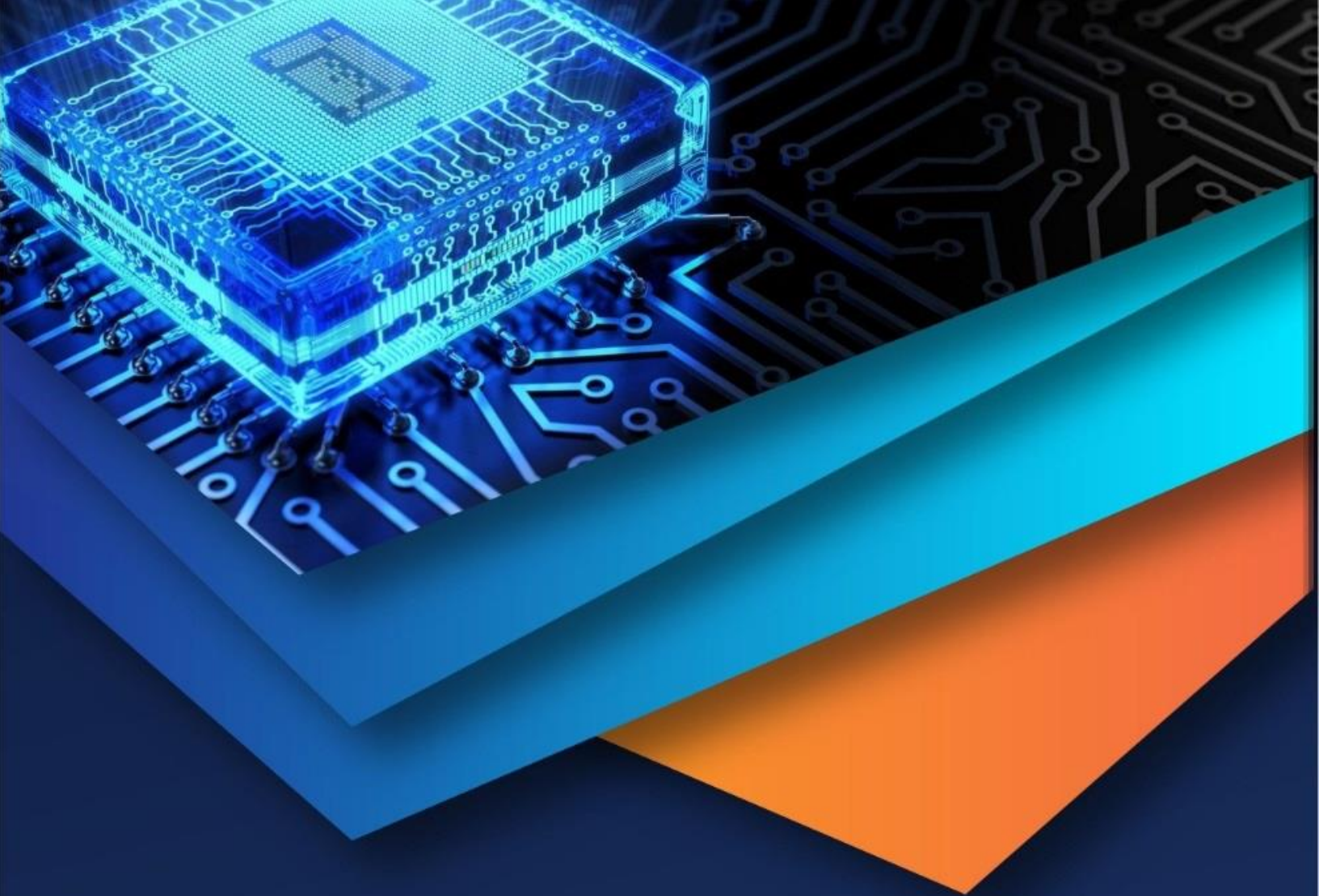
VI. CONCLUSION AND FUTURE SCOPE

DeepFake Shield shows that one accessible tool can screen images, text, audio and video, and that doing this honestly means admitting uncertainty. Returning to the research questions, four detectors can be placed behind one interface with their limits stated openly, a three-way decision rule gives a more honest output for ambiguous content, and a strict separation between testing and accuracy claims keeps the evaluation credible. The system is meant to support human reviewers and not to replace them. Future work will first complete the labelled evaluation and report the results with coverage and calibration. After that we plan API endpoints with batch processing, simple ensembles of detectors, adversarial evaluation, automatic threshold tuning on validation data and support for languages other than English. Longer-term directions include stream processing, a mobile client and federated learning, each of which raises its own privacy and robustness questions.

REFERENCES

- [1] Y. Mirsky and W. Lee, "The Creation and Detection of Deepfakes: A Survey", ACM Computing Surveys, vol. 54, no. 1, 2021.
- [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, 2020.
- [3] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies and M. Nießner, "FaceForensics++: Learning to Detect Manipulated Facial Images", IEEE/CVF International Conference on Computer Vision (ICCV), 2019.
- [4] B. Dolhansky et al., "The DeepFake Detection Challenge (DFDC) Dataset", arXiv:2006.07397, 2020.
- [5] Y. Li, X. Yang, P. Sun, H. Qi and S. Lyu, "Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics", IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.
- [6] M. Todisco et al., "ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection", Interspeech, 2019.
- [7] B. Guo et al., "How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection", arXiv:2301.07597, 2023.
- [8] C. Guo, G. Pleiss, Y. Sun and K. Q. Weinberger, "On Calibration of Modern Neural Networks", International Conference on Machine Learning (ICML), 2017.
- [9] Y. Liu et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach", arXiv:1907.11692, 2019.
- [10] C. K. Chow, "On Optimum Recognition Error and Reject Tradeoff", IEEE Transactions on Information Theory, vol. 16, no. 1, 1970.
- [11] Y. Geifman and R. El-Yaniv, "Selective Classification for Deep Neural Networks", Advances in Neural Information Processing Systems (NeurIPS), 2017.
- [12] C. Stupp, "Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case", The Wall Street Journal, August 2019.
- [13] H. Chen and K. Magramo, "Finance worker pays out \$25 million after video call with deepfake 'chief financial officer'", CNN, February 2024.
- [14] A. Paszke et al., "PyTorch: An Imperative Style, High-Performance Deep Learning Library", Advances in Neural Information Processing Systems (NeurIPS), 2019.
- [15] T. Wolf et al., "Transformers: State-of-the-Art Natural Language Processing", Proceedings of EMNLP: System Demonstrations, 2020.
- [16] G. W. Brier, "Verification of Forecasts Expressed in Terms of Probability", Monthly Weather Review, vol. 78, no. 1, 1950.

©IJRASET: All Rights are Reserved5



10.22214/IJRASET



45.98



IMPACT FACTOR:
7.129



IMPACT FACTOR:
7.429



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089  (24*7 Support on Whatsapp)