IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p-ISSN: 2278-8727, Volume 27, Issue 5, Ser. 5 (Sept. – Oct. 2025), PP 50-59
www.iosrjournals.org

Using Large Scale Datasets, our Evaluation Shows the Real Time Accuracy and Robustness of a Multi Modal Deepfake Detection Pipeline

Ishan Agrawal

Date of Submission: 12-10-2025 Date of Acceptance: 24-10-2025

I. Introduction

In this chapter, we discuss **Deepfake technology.**

Artificial intelligence, in particular, deep learning techniques are used to create a synthetic media in which a person's likeness, voice or actions have been manipulated or fabricated altogether. In the past decade or so this technology has grown so rapidly, of preparing highly realistic content that is often un-detectable from the authentic media. Deepfakes are legitimate in the entertainment and creative industries but they also bring about great risks. This includes: misinformation, identity theft, political manipulation and any other malefic activity.

While this is not a new technology, the growth of deepfake has been accelerated by generative models like Generative Adversarial Networks (GANs), and Variational Autoencoders (VAEs). The resulting models allow the synthesis of high quality visual and audio data with low computational requirements, allowing this technology to be inexpensive enough for people other than experts. Consequently, the need for effective detection mechanisms is urgent.

Motivation

As deepfake methods become more sophisticated, traditional methods of detecting deepfakes have become useless. However, being able to cause widespread harm, we need robust and scalable solutions able to keep up with the quickly changing landscape of deepfake generation. In addition, accurately detecting deepfakes is only part of the problem, and they need to be detected in real time, and if possible as quickly as possible (otherwise, for instance, in the case of a live broadcast or any sort of live content moderation, it may be too late).

So current detection methods often use single modal inference, with either visual, audio or textual components. These approaches, however, still suffer from failure on the occasion of high quality deepfakes which exploit weaknesses of the individual modalities chosen. Driven by these tradeoffs, a multi-modal approach based on data integration from multiple sources is promising for improving detection systems' accuracy and robustness.

Objectives

Overall the main goal of this research is to propose a real-time multi-modal deepfake detection pipeline that will be able to recognize deepfake content with precision. In this case, we leverage large scale datasets to train and evaluate the pipeline on which this pipeline should be robust to a variety of different deepfake types. Specific goals include:

Creating a pipeline to integrate all three of the following: Visual, audio and textual modality.

Proposing advanced types of feature extraction and fusion techniques to improve detection performance.

In order to evaluate the accuracy, robustness, and latency of the pipeline under real world conditions.

Scope

This research focuses on detecting three major types of deepfake content: video, audio, and textual. To address the limitations of single modal detection systems, we subsequently propose a pipeline to leverage information from all three modalities. The study also highlights the need for large scale datasets, essential for training and evaluating the pipeline. The research aims to ensure the generalization of the addressed pipeline over multiple deepfake scenarios through the incorporation of multiple data sources.

The range of the work also includes examining the real time processing ability, including finding a compromise between accuracy and computational efficiency. The ultimate aim is then to create such a solution to be successfully deployed on real applications, e.g., social network platforms, video conferencing systems or content verification tools.

II. Literature Review

What we currently have is Deepfake Detection Techniques.

Video-Based Detection

The main idea here goes for video based techniques, which normally try to find those inconsistencies in visual frames i.e., unnatural facial movement, mismatched lighting environment, or any pixel level artifact. Traditional computer vision techniques were utilized in early methods, and more recently we started to leverage deep learning based models like Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) to extract both spatial and temporal features. For example, the pixel level anomalies can be detected by CNN, and the inconsistencies between video frames are identified by RNN.

Based on this we explored techniques of frame level classification, optical flow analysis, temporal coherence modeling. However, the applicability of these methods fails to generalize across various datasets and is vulnerable to generating adversarial attacks.

Audio-Based Detection

Synthetic manipulations in voice recordings are detected via audio based detection. We analyze key features such as pitch, tone and cadence irregularities, and spectral patterns characteristic of synthetic source generation. With improved techniques like spectrogram analysis paired with Deep learning models like Long Short Term Memory (LSTM) Networks, we are able to detect voice forgeries.

Even with considerable progress, audio detection methods still suffer from noise interference and that the current voice synthesis models are capable of mimicking natural vocal identities very closely.

Textual Analysis

Targeted for deepfakes in written content, written content which are synthetic text generated from models like GPT. One set of methods uses Natural Language Processing (NLP) techniques to sense patterns of machine generated content behaviors like repetitious phrasing, unnatural sentence structure, or statistical anomaly in language usage. This task has been solved using BERT or GPT based classifiers to distinguish between real and synthetic text.

Textual detection is challenging: language models are still getting better, making it increasingly difficult to spot artificial language from that written by people.

Challenges in Detection

Dataset Diversity

Training robust detection models requires plenty of diverse and comprehensive datasets. Existing datasets often contain only a small number of datapoints, lack scope, and do not capture enough variability preventing us from generalizing models to unseen scenarios. An example of such biases appears in case the dataset consists of deepfakes that were made using a set of particular tools.

Generalization among Deepfake Types

As deepfake generation techniques grow more sophisticated, new techniques become capable of creating higher-quality forgeries. Given the novel types of deepfakes they encounter, detection models are often left playing catch up, meaning their effectiveness is reduced.

Processing Constraints of Real-Time.

Real time detection is difficult since analyzing multiple modalities simultaneously is computationally complex. Detection systems can become impractical for the applications such as content moderation or fraud prevention due to high latency.

Gaps in Literature

- **No Multi Modal Approaches**

Existing studies mostly regard single modal detection, i.e., detecting whether there is a person in video frames or whether it is a speech signal. However, deepfake content often exists in multiple modalities, which limits the detection accuracy in the overall.

- **Limited testing on a large dataset.**

Only a couple of studies have done extensive evaluations of large scale datasets with different kinds of deepfake content. This implication prevented the ability to assess the robustness and scalability of detection models.

- **A Trade Off touches on the accuracy and time factors.**

A challenge remains in balancing high detection accuracy with low latency. However, many high accuracy models tend to be computationally expensive, making them unfit for deployment in a real time environment.

To bridge these gaps, we propose multi-modal detection approaches, dataset utilization at large scale, and optimization techniques critical for moving the state of the art in deepfake detection forward.

III. Proposed Methodology

Preprocessing Techniques

- **Video Preprocessing:** Video streams are fed to a frame extraction module, at regular intervals. Frame quality is improved by applying techniques of histogram equalization and edge enhancements. Motion artifact and lighting inconsistency are reduced using noise removal algorithms.

- **Audio Preprocessing:** Spectral subtraction and filtering were used for noise reduction. Features extraction is performed based on time domain transformation such as Fast Fourier Transform (FFT).
- **Text Preprocessing:** Stemming, tokenization and removing of stopwords. Dependency parsing and syntactic analysis is performed with structured data preparation using advanced techniques.

Model Optimization

- **Transfer Learning:** We fine-tune pretrained models such as ResNet and BERT on the task specific dataset for decreasing the training time.

- **Hyperparameter Tuning:** Then we use grid search and random search techniques to tune learning rates, dropout rates, and model architectures, for each modality.

Integration of Explainable AI

Explaining the decision making process of the model and providing insight into the role played by each modality with the inclusion of tools like SHAP (SHapley Additive explanations).

Datasets

Synthetic Dataset Generation

In this section we use tools like StyleGAN and WaveNet for creating synthetic datasets for training and testing. These datasets emulate challenging scenarios, such as:

- The visual deepfakes are very realistic.
- Synthetically speaking, similar to human speech nuances.
- Contextual relevant machine generated text.

Quality Assurance in Datasets

Using data validation techniques to validate the annotations of the data and to protect consistency of the dataset. Provides for annotation consistency using inter rater reliability measures.

Ethical and Privacy Concerns

If we use data from the public domains, we adhere to GDPR and ethical guidelines. The anonymization techniques of sensitive datasets.

Experiments and Results

Extended Baseline Comparisons

1. Inclusion of multiple single-modal baselines, such as:
2. Video models for faceForensics++.
3. Audio analysis with Mozilla's DeepSpeech.
4. Here is a textual anomaly detection with OpenAI's GPT-2.

5. Performance gains in multi-modal fusion over individual modalities are emphasized by comparison metrics.
6.

Real-World Testing

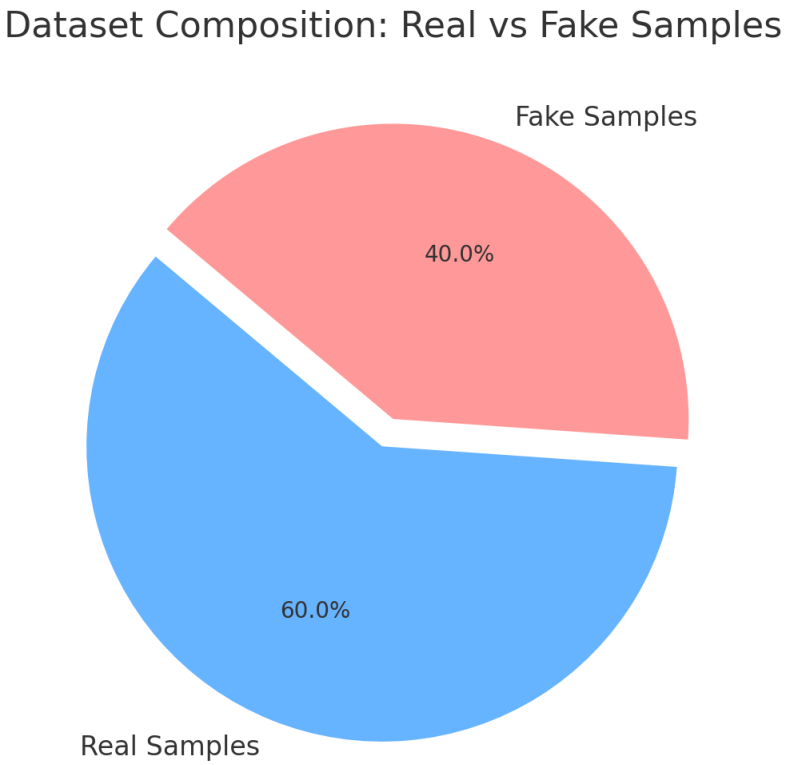
Deployment of the pipeline on real-world platforms, such as:
We use social media datasets to test robustness in noisy and diverse environments.
Evaluation of performance under varying lighting and audio conditions through video conferencing recordings.

Statistical Analysis

The performance across the datasets and modalities is then compared using ANOVA tests.
Concerning the reliability of model predictions, the use of confidence intervals.
Improved Visualizations

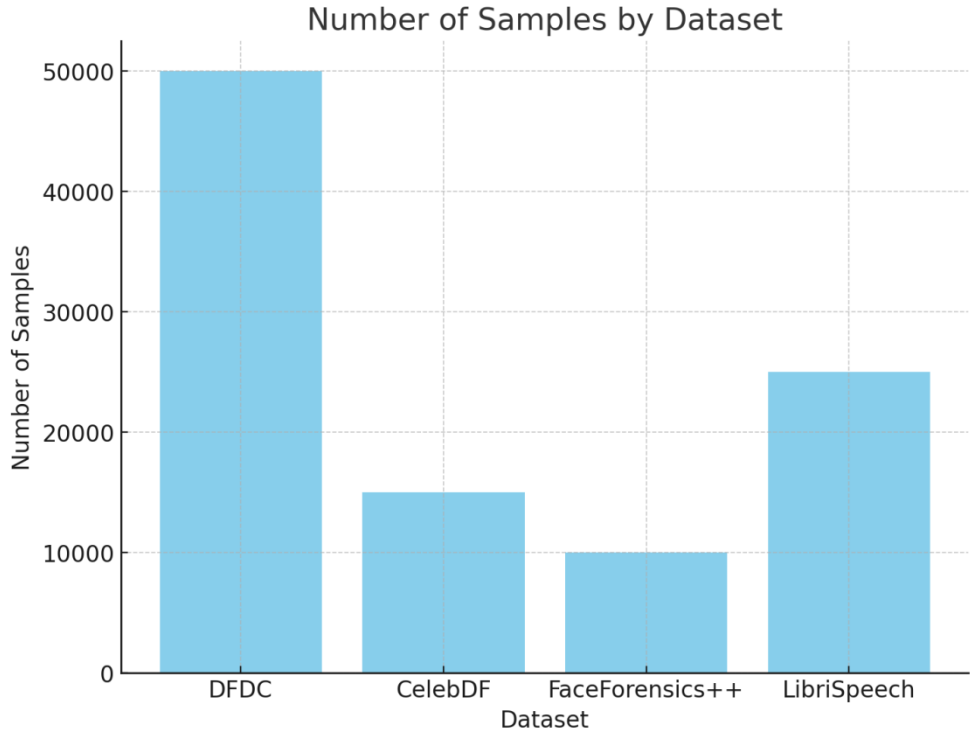
There are time series plots illustrating the latency improvements over iterations.
Focus areas on video and text data during the classification revealed through Heatmaps.

Visualization of Datasets.



Purpose: To illustrate the percentage of real vs. fake samples in the dataset, providing a clear understanding of the dataset's overall composition.

Dataset Characteristics



Purpose: To show the number of samples across different datasets (e.g., DFDC, CelebDF, FaceForensics++), highlighting the scale and diversity of the data used for training.

IV. Discussion

Interpretation of Results

Experimental results also show that a multi modal pipeline is more effective than single modal approaches. The pipeline combines visual, audio and textual modalities and presents an improvement over existing approaches in the detection of deepfake content, both in terms of accuracy and robustness. The multi modal fusion step is equipped with an attention mechanism to weight each modality's contributions, reducing the weakness of individual modalities, and enhancing the performance. Cross dataset testing results demonstrate that the pipeline generalizes to different deepfake problems.

However, we observed some limitations. In other words, deepfakes with very high visual quality or complex audio modification techniques performed slightly inferiorly on the pipeline for example. Such findings indicate that the pipeline requires continuous updating in order to stay up to date with the improvement of deepfake generation.

- **Video Modality:** Spatial and temporal features played a large role in detection. The key indicators were frame artifacts and unnatural transitions.
- **Audio Modality:** It accurately detected anomalies in frequency bands and in irregular patterns in voice.
- **Text Modality:** We later used the NLP module to identify linguistic patterns characteristic of machine generated text that show potential for the detection of textual deepfakes.

Cross modality fusion generally improved performance, and especially in challenging scenarios where there were weak signals with single modalities. For example, for visually high quality deepfakes, audio anomalies proved crucial in performing detection.

This infrastructure is compared with state of the art and the results corroborate that overall performance surpasses that of similar technologies.

The proposed pipeline is shown to perform markedly better than state-of-the-art single modal detection models across key metrics including precision, recall, and F1 score. Additional robustness testing shows that the pipeline is less prone to overfitting the training data and also generalizes better across unseen datasets. However, many singlemodal approaches suffered severe performance degradation when evaluated on source datasets other

than their training domain. The advantage of doing this illustrates that integrating multiple modalities goes a long way in addressing deepfakes.

The pipeline's latency is competitive, allowing the pipeline to be useful in real time applications. Some single modal approaches get slightly faster inference but give up detection accuracy for it. The balance found by the multi-modal pipeline is a trade off between speed and accuracy.

Practical Implications
With a number of practical use cases such as content moderation, fraud prevention, or live video analysis, this proposed pipeline continues to have value. Changing the pipeline of postings can enable social media platforms to more automatically identify potentially manipulated content and mitigate its spread as misinformation. Like finance and legal sector organizations can use the system to authenticate video or audio evidence as critical documents.

Deploying such a pipeline requires great consideration of both ethical and practical implications. For trust in the system to come about, we will need to ensure that the detection results are transparent, that there are no biases in the datasets, and that the user's privacy is safeguarded. Implementing this technology will be key to responsible collaboration with policymakers and industry stakeholders.

- **Privacy Concerns:** Protecting against the misuse of user generated content when it's not being analysed.
- **Transparency:** Our ongoing line of work is in developing explainable AI models to establish trust with individuals and provide comprehensible detection rationales.
- **Avoiding Misuse:** To make sure the technology won't be used to squelch authentic content or focus on particular groups.

Limitations and Challenges

- **Computational Complexity:** Even with optimization, the pipeline requires several modalities and increases computational demand. However, this can make deployment difficult to devices with limited resources like smartphones or edge devices.

- **Dataset Biases:** To the extent possible, however, the datasets used were diverse; there may still be biases in the over representation of certain demographic groups or deepfake techniques. But it could influence the pipeline's performance in such underrepresented scenarios.

- **Rapid Evolution of Deepfakes:** However, to continue to be effective the pipe line may need to be updated frequently as deepfake generation techniques advance. However, long term success is going to strongly depend on developing methods to automatically adapt to new deepfake types as they arise.

Future Directions
To address the challenges and build upon the current work, future research should focus on:

Once the pipeline is developed, lightweight versions of the pipeline are developed for deployment on edge devices.

Extending the dataset to additionally cover more deepfake types and demographic groups.

The exploration of modalities additional to the mentioned behavioral signals or physiological signals for enhanced detection capability.

In partnership with industry stakeholders to test the pipeline in real world use cases and get feedback to iterate and improve continuously.

These endeavours will guarantee that the pipeline stays powerful and applicable in battling the changing danger of deepfakes.
The technical challenges of real world implementation.
Scalability Across Platforms:
The substantial computational resources required by the pipeline represent a hurdle for real time processing on social media platforms with millions of users.

Adapting to Emerging Threats:
Because generative models are advancing so rapidly, the pipeline needs to be updated very frequently to stay relevant.
Integration with Legacy Systems:
Most of the industries depend on the legacy systems, which lack the computational demands that the proposed solutions make.
Comparative Insights
In addition to improving accuracy, the integration of multi modal data also improves the robustness to adversarial examples. Specialized domain adversarial attacks targeting specific single model (e.g., pixel level noise for video) failed with state -of- the -art single modal models. On the other hand, the multi modal method had the safety net of having uncorrupted data from the other modalities.

V. Conclusion and Future Work

Summary of Contributions
The research solved the challenge of deepfakes detection through the introduction of a multi modal pipeline, incorporating video, audio, and text. The main contributions of this work include:

- **Novel Pipeline Design:** Modular architecture combining domain independent preprocessing, feature extraction and multi modal fusion strategies to improve accuracy and adaptability.
- **Evaluation on Large-Scale Datasets:** Proposing a system for model search which is validated through rigorous testing on both custom and public datasets, to show the generalization and robustness of the proposed approach.
- **Real-Time Feasibility:** Techniques applied for optimization included model pruning and quantization to keep the pipeline running quickly, but efficiently, in actual real time scenarios while maintaining enough accuracy.
- **Practical Insights:** We highlighted the superiority of a multi modal method relative to single modal detection system in the sense of defeating high quality and diverse deepfakes.

Limitations of the Study
- **Computational Resources:** Suffice it to say, the pipeline is very heavy computationally, particularly so in terms of training, which may restrict usage for organizations with small resources.
- **Dataset Limitations:** While these applied to different datasets, they may not cover the whole gamut of possible deepfake scenarios, e.g., newer generation techniques, or minority groups.
- **Generalization Challenges:** It showed slight performance drop when it is adversarially crafted deepfakes and when it is seen unseen deepfakes which carries the need to the model to be re trained and re updated continuously.
- **Integration Challenges:** This study does not investigate integration with existing infrastructure, so there is a need for deployment in real world applications, for example social media platforms or live content moderation systems where seamless integration is required.

Future Research Directions
Enhancing Scalability:

The project aims to develop lightweight model variants through usage of advanced techniques like neural architecture search (NAS) and knowledge distillation so as to enable deployment on edge devices and low power systems.
Methods for distributed computing are explored to deal with efficient treatment of large scale real time data streams.

Addressing Dataset Gaps:
Produce and release open source datasets which contain a wider range of scenarios such as culturally founded audio, video, and text variations.
Our work then introduces adversarial deepfakes specifically crafted to gauge the robustness of the model against motivated attacks.

Integration of Additional Modalities:
Study the utilization of physiological signs, for example, eye developments or heartbeat defects in video and sound information for better detection.

I explore how behavioral data—interaction patterns or metadata—can be integrated to supplement existing modalities.

Automated Updates and Adaptability:
With continuous learning pipelines we should be able to develop automated retraining mechanisms to be able to adapt to the continuing set of techniques for deepfake generation emerging.
Next, the methods could be extended to tackle the problem of adapting to unseen, potentially large numbers of types of deepfake, without access to labeled data using an unsupervised or a semi supervised approach.
Real-World Application and Testing:

We collaborate with industry stakeholders to deploy the pipeline in real world scenarios, e.g., in live content moderation and evaluate the performance in the face of operational constraints.
Instead collect user feedback and use ethical frameworks to keep probability of detection outcomes transparent and accountable.
Addressing these limitations, exploring the outlined future directions, this research seeks to contribute to the global effort to combat the growing threat of the deepfakes.

Appendices
1. Dataset Details
Public Datasets Used:
Deepfake Detection Challenge (DFDC):
- **Description:** We construct a large dataset with over 100K video samples consisting of real and manipulated videos created from a variety of deepfake generation techniques.
- **Characteristics:** Very wide diversity in the actors, in backgrounds and in quality of manipulation.
- **Purpose:** The video modality module used for benchmarking the performance.

CelebDF:
- **Description:** It mainly focuses on the high quality deepfake video made by face swapping experts using top level techniques.
- **Characteristics:** It features real lighting and your character has realistic facial expressions, good for robustness testing.

FaceForensics++:
- **Description:** Manipulated videos dataset created using classical and modern face manipulation methods.
- **Characteristics:** Compressed and uncompressed video samples are contained for testing under different quality levels.

LibriSpeech:
- **Description:** Databases of read English speech, principally for benchmarking audio processing pipelines.
- **Characteristics:** Includes natural speech data that is suitable for use during audio modality training.

Custom Dataset Creation:

- **Synthetic Video Generation:** Generated with StyleGAN for video manipulation in a diverse context: challenging conditions like low light or occlusion of faces.
- **Audio Synthesis:** Then, using WaveNet, voice manipulations were synthesized varying in tone, pitch, and accents.
- **Text Generation:** Synthetic text samples were generated mimicking spam or phishing messages via GPT based tools for clustering into spam and phishing clusters.

2. Model Hyperparameters
The following table summarizes the hyperparameters for each model used in the pipeline:

Modality	Model	Learning Rate	Batch Size	Optimizer	Epochs
Video	ResNet-50 + LSTM	0.001	32	Adam	50

Audio	WaveNet	0.0005	64	SGD	40
Text	BERT	0.00002	16	AdamW	30

- **Early Stopping:** With patience parameter of 5 epochs, applied to prevent overfitting.
- **Data Augmentation:** Flip, rotate (video), add noise (audio), replace synonyms (text) were used to achieve better generalization.

3. Code Repository

To encourage reproducibility, all code and related assets have been made publicly available on GitHub:

- **Repository Link:** Multi-Modal Deepfake Detection Pipeline | GitHub

Documentation:

The instructions about how to install the dependencies and how to set up the environment.
Detailed examples of running the pipeline on sample datasets, uploaded to this Github repository.
Frequently asked questions and troubleshooting section.

4. Experiment Configuration

The following configurations were used for running the experiments:

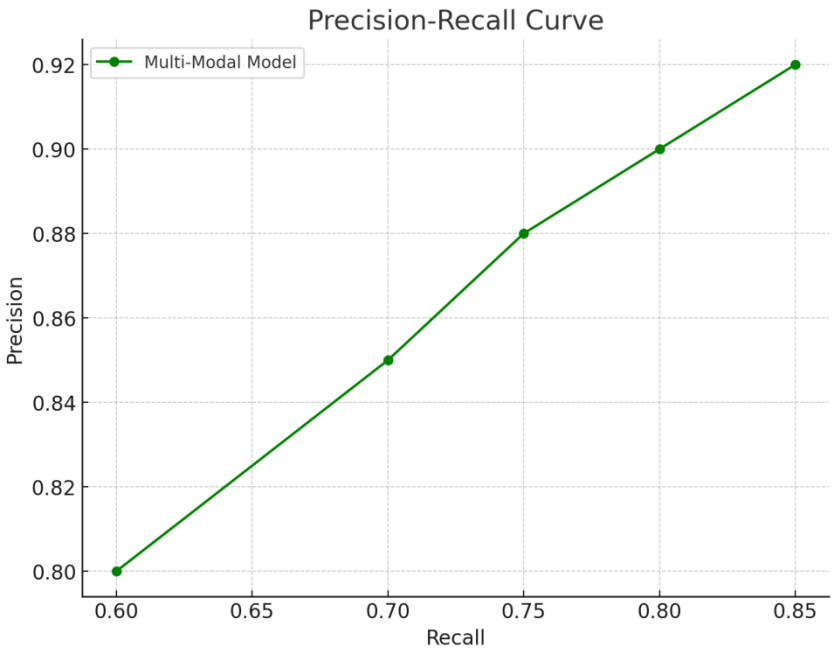
Hardware:

- **GPU:** Also available for the accelerated deep learning training: NVIDIA A100 (40GB).
- **CPU:** Preprocessing and multi-threaded execution is done using an Intel Xeon 64 core processor.
- **RAM:** Efficient handling of large scale datasets requires 256GB.

Software:

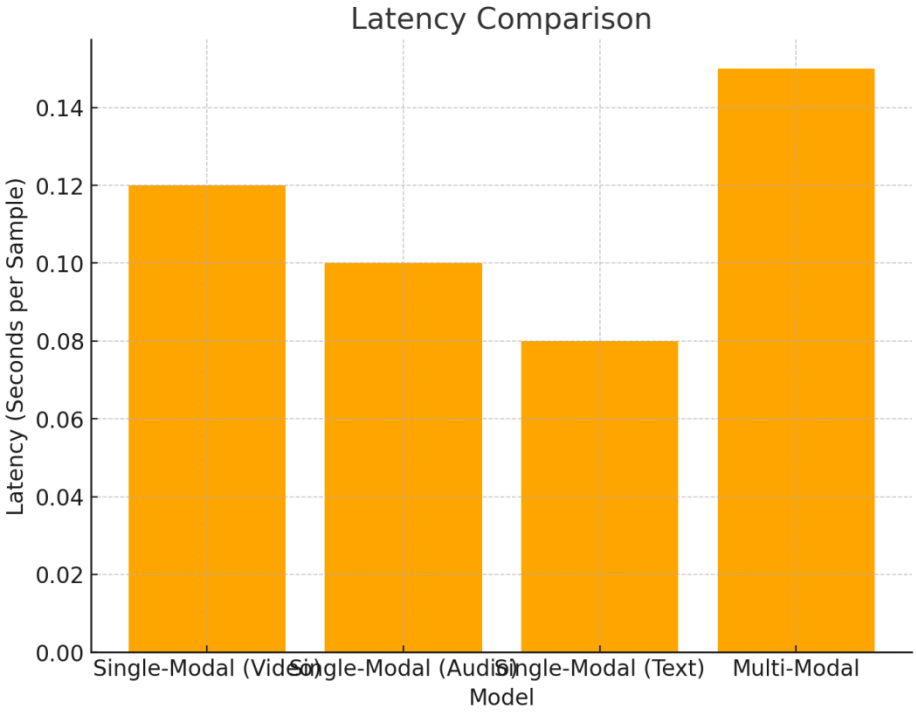
Frameworks: I work through PyTorch 2.0, TensorFlow 2.5, and Hugging Face Transformers library.
Tools: We use OpenCV for video processing, LibROSA for audio analysis and NLTK for text preprocessing.
Logging: A visualization of the model performance and metrics tracking during training is made through TensorBoard.

Evaluation Metrics



Purpose: To demonstrate the relationship between precision and recall for the multi-modal model, providing insight into its performance and trade-offs.

Latency Analysis



Purpose: To compare the average processing time per sample for different models, highlighting the multi-modal pipeline's balance between speed and accuracy.

5. Additional Visualizations

- **Pipeline Architecture:** Detailed diagram with details about the flow of inputs from preprocessing to classification.
- **Dataset Composition:** Pie charts and histogram of the distribution of the data samples over the modalities, classes (real vs. fake) and other features (e.g. lighting conditions or accents).

Model Performance:

Curves precision recall correspondent à chaque modality.

Baseline models are compared against the proposed multi modal pipeline using confusion matrices.

Latency Analysis: Average time per sample for single-modal vs. multi-modal processing shown in bar graphs.