A Novel Multi-Modal Deep Learning Approach for Real-Time Live Event Detection Using Video and Audio Signals

Integrating Audio-Visual Fusion for Robust Real-Time Event Recognition

Pavadareni R[1], A. Prasina[2], Samuthira Pandi V[3]*, Ibrahim Mohammad Khrais[4], Alok Jain[5], Karthikeyan[6]
Department of AI&DS, Chennai Institute of Technology, Chennai, India[1]
Department of ECE, Chennai Institute of Technology, Chennai, India[2]
Centre for Advanced Wireless Integrated Technology, Chennai Institute of Technology, Chennai, India[3]
Faculty of Economics and Administrative Sciences-Islamic Banks, Zarqa University, Zarqa, Jordan[4]
School of Electronics and Electrical Engineering, Lovely Professional University, Phagwara, India[5]
Department of Electronics and Communication Engineering, Vel Tech Multi Tech Dr.Rangarajan Dr.Sakunthala Engineering College, Avadi - Vel Tech Road, Avadi, Chennai[6]

***Abstract*—Recent developments in live event detection have primarily focused on single-modal systems, where most applications are based on audio signals. Such methods normally rely on classification approaches involving the Mel-spectrogram. Single-modal systems, though effective in some applications, suffer from severe disadvantages in capturing the complexities of a real-world event, which thereby reduces their reliability in dynamically changing environments. This research study presents a novel multi-modal deep learning approach that combines audio and visual signals in order to enhance the accuracy and robustness of live event detection. The innovation lies in the use of two-stream LSTM pipelines, allowing for temporally consistent modeling of both input modalities while keeping a real-time processing pace through feature-level fusion. Unlike many of the recent transformer models, we are utilizing proven techniques (MFCC, 2D CNN, ResNet and LSTM) in a latency-aware and deployment-friendly architecture suitable for embedded and edge-level event detection. The AVE (Audio Video Events) dataset, consisting of 28 categories, has been used. For the visual modality, video frames undergo feature extraction through a 2D CNN ResNet and temporal analysis through an LSTM. Simultaneously, the audio modality employs MFCC (Mel Frequency Cepstral Coefficients) for feature extraction and LSTM to capture temporal dependencies. The features extracted from both audio and video modalities are concatenated for fusion. The proposed integration leverages the complementary nature of audio and visual inputs to create a more comprehensive framework. The outcome yields 85.19% accuracy in audio and video-based events due to the effective fusion of spatial and temporal cues from diverse modalities, outperforming single-modal baselines (audio-only or video-only models).**

Keywords*—Multi-modality; feature fusion; early fusion; concatenation audio-video signals; convolutional neural network (CNN); Long Short-Term Memory (LSTM); Mel Frequency Cepstral Coefficients (MFCC); ResNet (Residual Network)

I. INTRODUCTION

Event detection has become a pervasive application of artificial intelligence, found in applications as diverse as video surveillance, sports analysis, public security systems, and even entertainment. The identification and classification of significant events in time based on the analysis of multimedia data—mainly audio and video. However, traditional unimodal approaches based solely on audio or video often fail to record very well the complex interaction often occurring between vision and hearing in ordinary situations. Most events in real-world scenarios do not occur in isolation. These events are often characterized by noise, temporal overlap, and a complex interplay of audio and visual stimuli. A barking dog following a car, a crowd cheering a goal, or shattered glass during the act of burglary—all are clearer when we can hear and see them at the same time. This is why modern event detection systems need multi-modal solutions that can effectively fuse both audio and visual data and make better, more knowledgeable decisions. Researchers have made encouraging strides in this direction. For instance, Convolutional Neural Networks (CNNs) have done well in extracting meaningful patterns from audio waveforms, while 3D CNNs are able to handle spatiotemporal data from video streams. Yet, with all their promise, these methods hit some tough bottlenecks: synchronizing asynchronous data streams, reducing computational overhead, and building fusion mechanisms through which audio and video can actually complement each other, rather than merely coexist. To address these challenges, recent studies such as [1] and [2] have also explored the appropriateness of multi-modal deep learning architectures.

These studies experiment with new architectures such as light vision transformers and advanced fusion techniques in the hope of improving both the accuracy and versatility of event detection models in the wild [3]. The fusion process of deciding how and when to merge features between modalities has turned into a very critical aspect that can significantly impact the performance of the model. This is where AVFusion comes in—our proposed multi-modal architecture designed specifically for live event recognition. It was experimented with over the well-liked Audio-Visual Event (AVE) benchmark dataset, including 4143 annotated videos across 28 real-world categories of events.

AVFusion leverages principles from works such as [4], with the feature-level fusion approach being targeted towards learning information about how an audio and video stream correlate from one time period to another.

In the AVFusion framework, we extract audio features as Mel Frequency Cepstral Coefficients (MFCCs), which is a representation that mimics the human ear's perception.

These features are then passed through Long Short-Term Memory (LSTM) networks to capture the temporal development of the sounds. Visually, the video frames are processed through ResNet to extract spatial features, again followed by LSTMs to capture motion and continuity over time. By fusing the encoded information of both modalities rather than at decision time, we have a richer, more integrated representation of the event. This combined audio-visual processing, while enhancing classification performance, remains computationally efficient and enables the system to operate in near real-time without heavy hardware. Our approach draws inspiration from recent developments like [5] and [6], highlighting the importance of intelligent fusion in multi-modal tasks.

The rest of the study is organized as follows: Section II: Literature Survey gives a concise review of the shift from unimodal solutions towards advanced multi-modal systems with accomplishments and ongoing difficulties. Section III: AVFusion Framework explains the structure in terms of preprocessing tasks, audio and video feature extraction, fusion pipeline, and training techniques like early stopping to prevent overfitting. Section IV: Experimental Setup. Section V: Results and Discussion reports empirical results on the AVE dataset, such as performance measures, visual plots, and comparative analysis to show the robustness of AVFusion. Finally, Section VI concludes the study.

While each of the elements used in AVFusion - MFCCs, 2D CNNs, ResNet, LSTMs, and concatenation at feature-level - is very well established, it is this focus on integration that makes this work novel. Most of the state-of-the-art models emphasized architectural complexity at the cost of runtime efficiency, while AVFusion was built for real-time event detection in a complex, very real-world setting. This framework not only focused on accuracy, but our aim was to find a balance among latency, accuracy, and modularity for implementation into surveillance, fault detection and embedded systems. This perspective challenges the idea that only architectural innovation leads to novelty, but also affirms how smart design for systems can produce feasible, scalable solutions to complex deep multi-modal challenges.

II. Literature Survey

Over the past decade, event detection has undergone a significant transformation. It began with simple unimodal approaches and evolved into complex multi-modal frameworks integrating auditory, visual, and contextual cues. This historical progression highlights how foundational research laid the groundwork for more integrated and adaptive systems. Each stage rounds a rather particular response to emerging real-world events that become progressively more complex.

In the onset research, most methods for event detection focus specifically on audio or visual sources, rarely both. Audio-only systems were among the first to make breakthroughs. One example is a 2018 study that used some spectrogram-based recognition of speech commands in conjunction with its own dataset and achieved an impressive 95% score on it [2]. The effectiveness was impressive during specific time periods, but it had no modeling of time-it could not know how the sounds evolve in time, leaving it deaf to dynamic auditory contexts.

Advancements soon followed. A 2023 study applied Electrodermal Activity (EDA) features in CNNs and LSTMs, achieving 96.6% accuracy on a Kaggle dataset. Though promising for real-time applications, it remained restricted to audio input without visual data. This model could not distinguish between acoustically similar but visually distinct events or understand spatial dynamics, a critical shortcoming in nuanced scenarios.

The field of video-based event detection has developed significantly over time. In 2017, researchers developed a method for detecting visual events (fights, crashes, and falls) using CNNs and LSTMs via spatiotemporal analysis [9]. The method was successful in recognizing visual events temporally, but did not include the auditory modality, and was relatively slow and computationally heavy, especially when it came to utilizing its predictive capabilities for real-time video-based decision making. By 2024, researchers began applying transformers in video classification pipelines [3], but effectiveness in modeling spatial and temporal dependencies was uneven at best. Despite advances made over the years, most video-based approaches still do not capture the auditory modality when optimizing for video classification tasks, and there are not typically any major increases in classifier accuracy for these tasks.

Having been trained and tested in narrow or conformed settings, those single-modal pipelines had difficulty capturing the complex and multi-sensory aspects of the scenarios outside the laboratory walls. Further studies led the research to develop modalities whereby multiple sensory streams are integrated to create a more complete representation of the world. The early attempts in this domain were made during the early 2010s. A concerted effort was made by one team to bring audio and video into a single platform using CNNs in the year 2014 [6]. Although pioneering, the model lacked mechanisms for encoding temporal relationships. The same year, the concept of grouplets was introduced to temporally align audio and video characteristics [7]. It was found to have an extension to innovations, but noise sensitivity and lack of a quantitative basis hold it back severely. At the same time, photofusion for visuals through CNNs and RNNs would prove to be a more organized exercise in temporal modeling by 2015. Once again, however, audio remained unincorporated in many early visual models [8].

In 2020, the state-of-the-art complete survey on feature fusion techniques was published. It enumerated and classified strategies as early, late, and hybrid fusion and created a conceptual map for future research [4]. The resulting taxonomy was to bring unity to the field and inspire a diversely methodological and application-specific solution scheme.

This momentum has accelerated rapidly. In 2019, a milestone was defined in the detection of audio-visual inputs along contextual metadata by the combination of deep learning model to recognize aggressive behavior on trains. Although

specific accuracy metrics were not reported, it was among the pioneering efforts to integrate real-world context into multi-modal models. A dynamic fusion model implemented in 2023 for micro-video recommendations introduced visual, auditory, and textual cues into fusion via meta-learning; so far, it achieves an accuracy of 87.9% on the UCF51 dataset. This opened significant new opportunities, showing that flexible fusion models may be better than static, single-modal systems [24].

By 2023, the most sophisticated multi-modal models were audio-visual speech recognition systems. Although high in computational cost, a model that used 3D CNNs together with BiLSTMs reached a phenomenal score of 98.56% on the LRW dataset [11]. Nevertheless, such accuracy boils down to a very high computational cost. However, it once reflected the power of multi-modal integration when resources permitted.

These advances were supported by the development of relevant datasets. The AVE data set, established in 2018, was an example. It collated 4,143 real-world recorded samples concerning 28 categories, such as from musical performances to machinery failures [23]. Today, it remains a significant benchmark for testing audio-visual fusion models. Other datasets like UCF51 and Kinetics-Sounds caught importance in 2024, scaling up generalizability to broad action recognition tasks and helping models achieve more robust results, such as 87.9% accuracy [22].

New research is steadily pushing the boundaries. The 2020 study utilized CNNs and LSTMs for visual temporal modeling [12]. Though excluding audio inputs, it nevertheless advanced some techniques for fusion. A parallel development in 2021 saw multi-modal analysis move into environmental monitoring, equipping sensor data and deep learning for urban water quality predictions-a powerful demonstration of multi-modalism beyond the traditional event detection pathway [13][10].

The fusion strategies have reached maturity. The early fusion integrates features before the learning process begins, as observed in a 2015 video event classification study [14]. Late fusion is applied in cases such as newborn pain detection in 2021; it combines outputs after separate models work on each modality, allowing modularity and interpretability [15]. Hybrid fusion is a sophisticated technique: it went to work in a 2023 medical imaging study where CNNs and transformers were integrated in the middle of the stream to preserve a variety of feature representations while enhancing model simplicity [16].

Attention mechanisms are receiving extended focus lately. The visual relationship model in the year 2022 used attention layers to isolate certain key interaction scenes, such as people making physical contact [17]. On the other hand, the affect detection system in 2024 was trained on long video-audio sequences using stacked transformers for picking up subtle emotional cues over time [18].

However, challenges still remain. In the survey on anomaly detection in 2019, it was highlighted that datasets presently available do not capture the messiness of the real world-the presence of interference, occlusion, and noise is rarely represented [19]. This notion was echoed in a healthcare review of 2023, which was calling for more "imperfect" multi-modal datasets that increasingly reflect lived environments [20]. Nevertheless, while transformers have greatly transformed fusion, a 2024 survey pointed out their failure due to sizable computations, restricting their applications in real-time or edge settings [21].

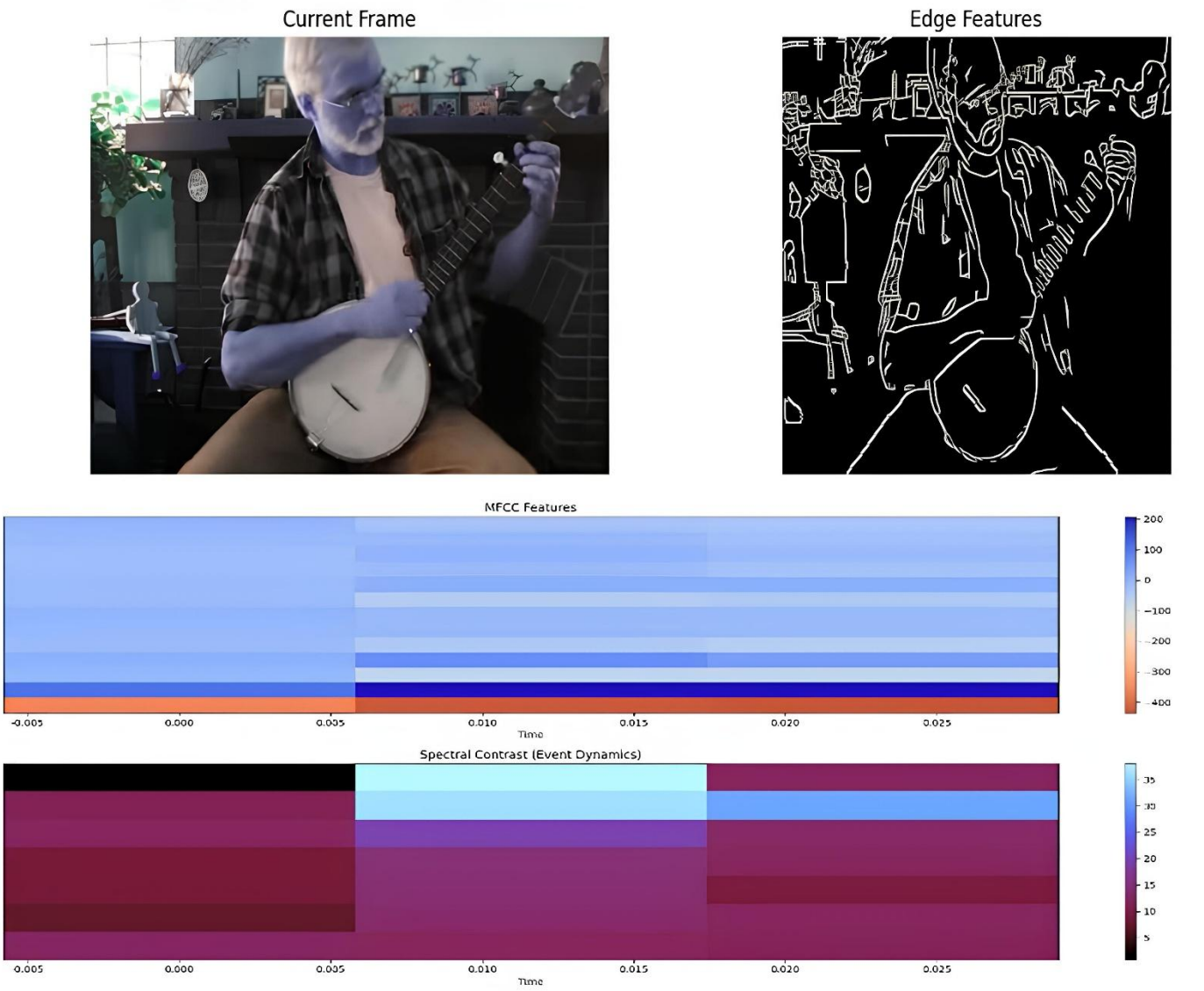
Text-based modalities also acted as a complement. A 2014 study on social event detection applied NLP techniques, such as TF-IDF, without any audio or video [5]. While such work appeared simplistic according to today's standards, it paved the way for the convergence of language, vision, and sound.

All influences dovetail them into AVFusion: our multi-modality framework, which integrates lessons from the previous decade into a tangential practice: with MFCCs and LSTMs to extract temporal patterns from audio, ResNet, and LSTMs to model spatial and sequential video dynamics, followed by feature-level concatenation of these streams into a unified representation space. Although not the most computationally intensive solution, it strikes a good balance between reliability and efficiency. Achieves 85.19% accuracy on AVE. It shows that good fusion could be as competitive as more heavyweight fusion setups. It does not require the most advanced state-of-the-art systems like 2023 precision health fusion model [20] or 2024 AVE specific architectures [22], AVFusion proves intelligent design, not plain raw complexities, could realize a world impact.

Our multi-modal framework, the modification from the last decade into an operational, streamlined system: MFCCs and LSTMs tap temporal patterns from audio, while ResNet and LSTMs model spatial and sequential dynamics of video. Feature-level concatenation binds these streams into a united representation space. Although it might not be the most computationally intensive solution, AVFusion achieves a balance between reliability and efficiency. It achieved an accuracy of 85.19% on the AVE dataset, showing that thoughtful fusion can vie with the more complicated setups. Not even calling for the state-of-the-art systems such as the 2023 precision health fusion model [20] or 2024 AVE specifications architectures [22], AVFusion proves that effective system design and practical impact are not solely a function of model complexity.

III. AVFUSION FRAMEWORK

The AVFusion method offers a high probability of detecting live events by merging audio and video in an intelligent way. In a four-fold setting: audio preprocessing, video preprocessing, fusion, and classification, it could generally be viewed as a pipeline shown in Fig. 1.

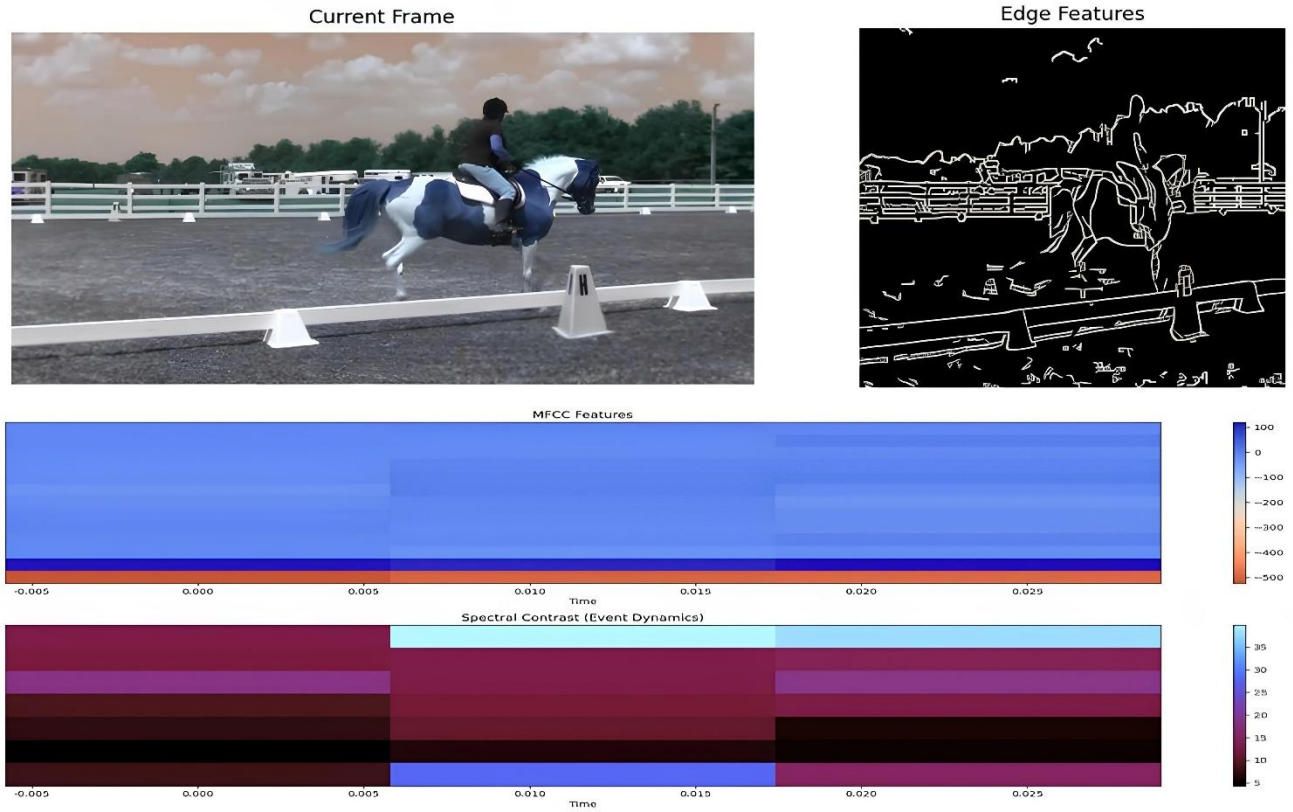


(a) Multi-modal feature extraction during a Musical event

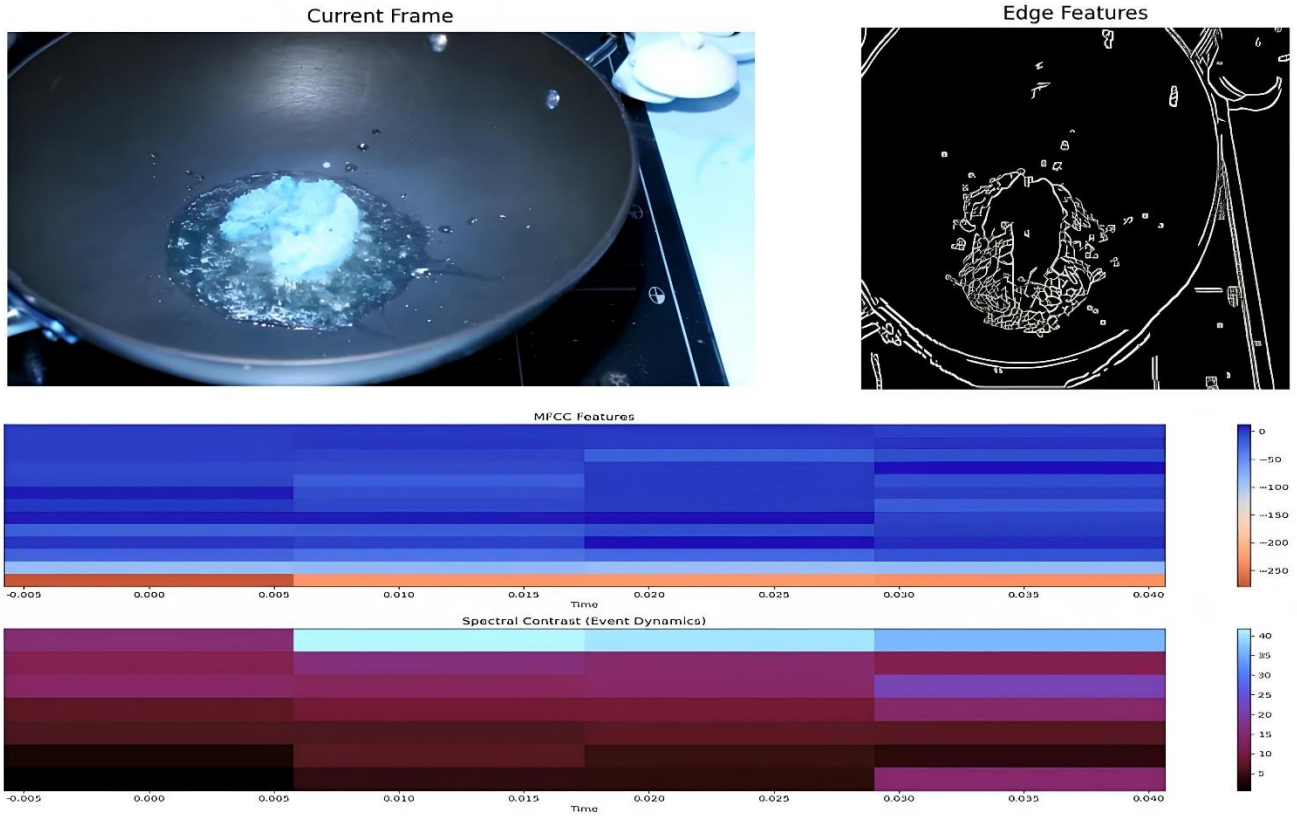
(b)

(c)

(d)



(e) Multi-modal feature extraction for an event of horse-riding.



(f) Multi-modal feature extraction for a cooking-related audio-visual.

Fig. 1. (a-f): Three AVE events with raw frames, audio features, and processed outputs [25].

A. Audio Preprocessing

In this context, audio refers to the temporal signal representation of sound events, such as the sharp twang of a banjo or the continuous sizzle of a frying pan. It is the process of converting an audio stream into Mel-Frequency Cepstral Coefficients (MFCCs)-a signal representation that resembles human hearing. It starts with a filtering step known as a pre-emphasis, which shifts the overall gain towards the high frequencies.

$$y(t) = x(t) - 0.97\,x(t-1) \quad (1)$$

Here, x(t) represents a raw audio signal, whereas y(t) stands for its filtered version. Afterwards, an audio signal must be segmented into short frames, and a Fast Fourier Transform (FFT) is computed to extract the power spectrum:

$$P(k) = \sum_{n=0}^{N-1} x(n) e^{-j\frac{2\pi}{N}kn} \quad (2)$$

where, P(k) is a complex-valued frequency component, x[n] is the time-domain signal at sample n, N is frame size, k is frequency bin index, $e^{-j\frac{2\pi}{N}kn}$ is the complex exponential kernel. Next, a Mel-scale filter bank is used to filter the signal to match human audition:

$$m(f) = 2595 log_{10}(1 + \frac{f}{700}) \quad (3)$$

where, f represents frequency in Hertz (Hz). We then take the spectrum's log to match loudness perception, apply a Discrete Cosine Transform (DCT), and achieve MFCCs:

$$C_k = \sum_{m=1}^{M} log(S_m) cos(\frac{\pi k}{M}(2m - 0.5)) \quad (4)$$

where, Ck are the k-th MFCC coefficients, Sm energy of the m-th Mel filter bank, and M is the number of Mel filter banks. This DCT decorrelates the log-Mel energies and compacts the signal's most important spectral information into a small number of coefficients. Fig. 2 shows this whole audio preprocessing flow, from raw signal to MFCCs. You can see the MFCCs and spectral contrast in action for three AVE events—Banjo, Horse Riding, and Frying—in Fig. 1(a), Fig. 1(e), and Fig. 1(f).

Those figures show the raw video frame on top, with the audio features (MFCCs and spectral contrast) below, capturing the sound patterns over time—like the twang of a banjo or the sizzle of frying. After getting those features, they are fed into an LSTM to track how the audio evolves, which is crucial for events that unfold over a few seconds.

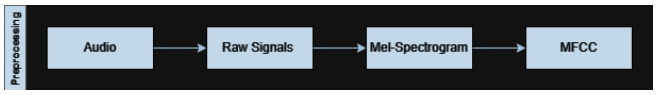


Fig. 2. Audio preprocessing flowchart.

B. Video Preprocessing

Video is about rapidly transforming shapes; for example, a horse in a gallop or the shape of a frying pan. We sample visual frames at 4 f/s, resize the frames to square 224×224 pixels, and normalize them using mean subtraction. A 2D-CNN on ResNet-50 extracts spatial features:

$$y(i,j) = \sum_{m=1}^{M-1}\sum_{n=0}^{N-1} x(i+m, j+n) \cdot W(m,n) + b \quad (5)$$

Here, x is the input image, W is the convolutional filter, b is the bias, and y (i, j) represents the activation at position (i, j) in the output feature map. This convolution captures spatial features such as edges, textures, and object contours in the input frame.

Temporal Modelling: The features are passed through an LSTM network that encodes the temporal evolution of video events. This LSTM allows the model to see how visual patterns evolve in time and is a requirement to identify the occurrence of events that last over a sequence of frames, for example, a fight or a car crash. An LSTM comes next to catch how those features change over time:

$$h_t = LSTM(h_{t-1}, F_t) \quad (6)$$

- ht is the hidden state at time step t,
- $ht-1$is the hidden state at the previous time step.

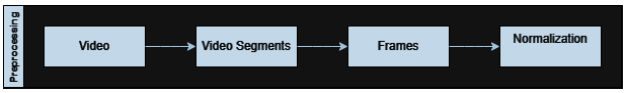


Fig. 3. Video preprocessing flowchart.

Fig. 3 maps out this video preprocessing flow, from frames to LSTM output. For a real look at it, check out Fig. 1. Fig. 1(b), (c), and (d) show the raw frames for Banjo, Horse Riding, and Frying—those are the starting points. Then, Fig. 1(a), (e), and (f) show the edge features after the CNN processes them, highlighting key shapes like the banjo's strings, the horse's legs, or the pan's edges. Those processed outputs help the model zero in on what makes each event tick.

C. Model Architecture

The architecture of AVFusion is a parallel architecture that consists of both an audio part and a video part, which are ultimately fused for joint event prediction. The audio part of the AVFusion model consists of a 3-layer LSTM with 128 hidden units taking in 40-dimensional MFCC features for each frame to extract the temporal dynamics of sound. The video part of the model contains spatial feature extraction, using ResNet-50, then a 2-layer LSTM with 256 hidden units to learn motion and temporal dependencies across video frames.

Additionally, both pipelines of the AVFusion model are trained end-to-end from their associated loss function with a shared loss function that allows the model to optimize both modalities at once without any spatial or temporal lag. Batch normalization is applied to each layer of the AVFusion model to stabilize the training of the model. In additional, a dropout rate of 0.5 is also used to combat overfitting. The final stage of the AVFusion model is to align the outputs of both audio and video pipelines over the frame, take the concatenation of the features from both modalities, and treat as a single representation with a prediction head for classification.

D. Fusion and Classification

Bringing both modalities together to bring forth a more accurate and holistic event detection model after feature generation by both audio and video models separately predicted is the last process. The fusion process includes Concatenation: Audio features and video features are concatenated into one

feature vector. This allows the model to take into account both visual as well as audio information of an event at once. The concatenated features will pass through fully connected networks, which can also be referred to as dense layers so that it can learn the relation between audio and video features. The combined feature vector is calculated as follows:

$$x_{fusion} = Concat(f_{audio}(x_{audio}), f_{video}(x_{video})) \quad (7)$$

- where, f_{audio} and f_{video} are temporally aligned feature sequences with matching time steps. The fusion is performed along the feature dimension after temporal alignment, and $Concat$ denotes the concatenation of the audio and video feature vectors. x_{audio} and x_{video} represent the feature vectors obtained from the audio and video models, respectively

Temporal Alignment: Before combining the audio and video features to be fused, the audio and video features need to be temporally aligned so that both content modalities (e.g., audio and video) are in synch. MFCCs are calculated with a window and a hop size so that an MFCC vector is calculated when each frame of the video is presented (e.g., one MFCC vector per frame). Consequently, f_{audio} and f_{video} are in temporal alignment at the same time step. This allows for concatenation by element (or frame) so that the fused (concatenated) feature vector is from the same time or time context from the audio and video modality at the same time steps.

Dropout is applied to prevent overfitting, and the output layer uses a sigmoid activation function to perform binary classification, determining whether an input belongs to the target class or not. The final classification is performed by passing the combined feature vector through a fully connected layer (dense layer) and a sigmoid activation function:

$$y_{pred} = \sigma(W x_{fusion} + b) \quad (8)$$

- W is the weight matrix,
- b is the bias term,
- $xfusion$ is the concatenated feature vector,
- σ is the function that outputs a value between 0 and 1, representing the sigmoid activation function, the probability of event detection.

AVFusion is shown in Fig. 4, with audio pipeline (3-layer LSTM, 128 units, 40 MFCCs) and video pipeline (ResNet-50, 2-layer LSTM, 256 units) that merge into a single decision. Batch Normalization along with dropout (0.5) is applied during all training processes to make the training process stable and reliable. AVFusion's design paradigms emphasize operational efficiency, data synchronization and modular scalability; rather than layers of abstraction or multi-layered fusion. Preprocessed and temporally synchronized audio and video streams assist the model to integrate asynchronous cues, without resorting to slower and variable time-frames of attention-based processing. Class imbalance is addressed through dynamic loss weighting and PCA is used for dimensionality reduction; such that the model is able to retain functionality even if hardware constraints are imposed. Although AVFusion employs a feature-level concatenation approach for fusion, a relatively simple fusion approach, this is intentional and allows the model to maintain nearly real-time inference speeds (30 FPS) with interpretability and generalizability; therefore, AVFusion's original contribution as a system is the coordination of many effective approaches into a single operationalizable.

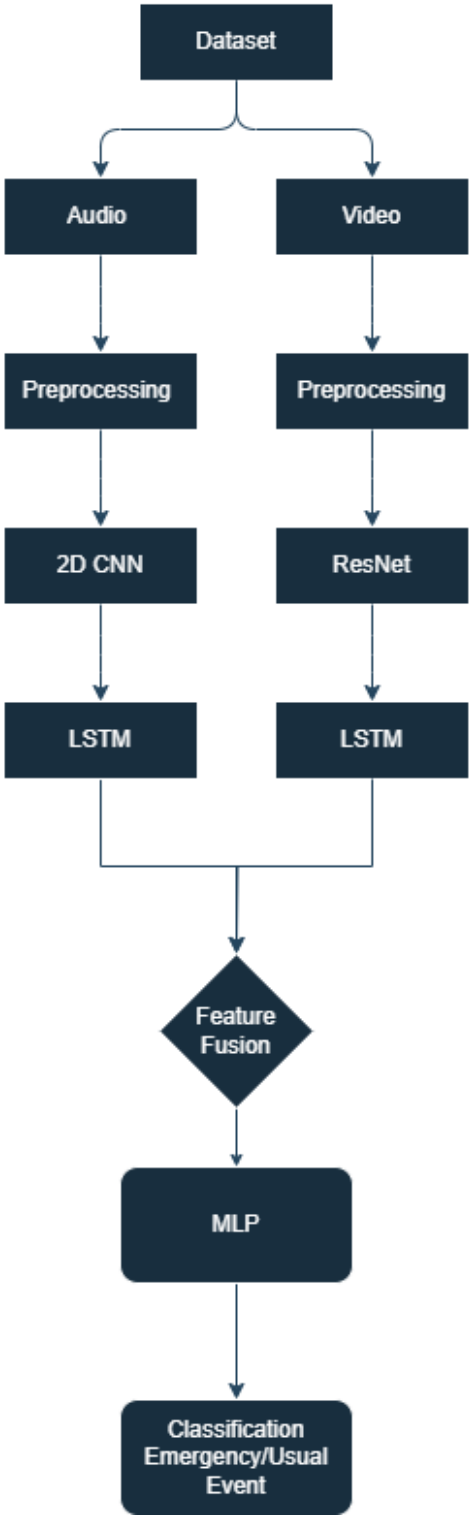


Fig. 4. Complete AVFusion architecture.

E. Evaluation

We examined the robustness and real-world applicability of AVFusion beyond training and validation, evaluating a new and unseen test split of the AVE dataset. This step was necessary to evaluate how well the model generalizes to different audio-visual scenarios previously unknown to it.

A key part of the training scheme that became a problem was the inherent class imbalance in the AR-Videos dataset. While the dataset contained common events (for example, general background noise or silent scenes), it also contained events that are extremely rare events (for example, frying or blowing out candles). If we are to ignore this imbalance, it is likely that the model would bias its predictions towards the majority classes - this results in inflated overall accuracy at the expense of poorly represented categories. As a technique to improve the model's performance overall, we introduced class weighting to the training scheme so that the underrepresented events in the dataset were weighted more heavily in the loss function for the training model. As a result, the underrepresented events in the model were encouraged to be formally attended to, therefore uncovering more subtle patterns, which leads to better generalizabililty and less likelihood of certain high-frequency classes.

To reduce the class imbalance that exists in the AVE dataset, we used class weights that were determined based on the inverse of the class frequencies. For every class c of the form c (wc), will be equal to $= \frac{1}{f_{c,}}$, is the relative frequency of class c in the training data. These class weights were applied to the loss function in the training process, which ensured that rare events were contributing relatively more to the optimization than the frequent events.

The second measure we took was early stopping and ReduceLROnPlateau to avoid overfitting, which is crucial in the presence of fewer high-signal events. These methods also monitored validation loss in real time, stopping training once the improvements plateaued, while also dynamically reducing the learning rate to refine optimization in the last few epochs.

While speed is a crucial variable in detection, other methods may be just as worthy. In live deployment scenarios such as surveillance or smart monitoring, the model's inference time on the test videos was observed. The AVFusion model stood up to the demands of a balanced approach to speed and accuracy; therefore, the model's performance supports its potential for real-time application.

The confusion matrix shown in Fig. 5 gives a detailed account of AVFusion's performance on 28 AVE classes. Good performance is noted for instances with obvious audio-visual signatures (like "instrument playing"), while slips are also noted for events with overlapping cues or which occur in complex environments. This frank profiling not only maps out the successes but also gives pointers toward future refinements.

The thorough benchmarking against single-modal baselines set forth in Section V, leveraging AVFusion to excel above audio-only or video-only methodologies, makes a statement in itself. It is thus established that the system is integrated: the whole is greater than its parts.

F. Integration with External Systems

It was created as more than just a proof-of-concept or academic experiment; it was made for actual deployment. One interesting application concerns the potential incorporation of this Tech-Scopes Data Connectivity's Connector system integrates AVFusion with power monitoring units (PMUs) to detect faults using embedded device data.

Tech-Scopes collects information regarding faults from embedded devices placed in critical infrastructures. These data are then collected into a centralized point of data collection, which can be identified as a nerve center in that ecosystem of monitoring. That point comprises a network of servers consolidated with display monitors and external hardware interfaces, and then a cluster of 14 PMUs, with each assigned to monitor voltage, current, and frequency across the grid.

AVFusion would sit in this very framework with audiovisual contextualization along those existing electrical information streams. With the connection of cameras and microphones placed close to monitored equipment, that is, transformers, switches, and circuit breakers, it will be capable of streaming environmental data in real-time, thus increasing the possibility of fault detection using multi-sensory analysis.

Consider the following: a power transformer starts to give off a sizzle-like sound, and as the flash hits a fuse, brief sparks are visible to the camera. By themselves, given the context mentioned, these signs would hardly excite suspicion within the walls of any classical PMU, heavily relying on electrical parameters as first indicators. However, an AVFusion trained to detect such Arabic patterns could have quickly associated these events and promptly generated an alert in context that hints at a potential equipment failure or safety hazard before it even goes down.

This real-time, sensory-rich detection embodies a paradigm shift in how we view power grid safety. Instead of being based on numerical anomalies in electrical signals, AVFusion now introduces the capability of visual and auditory intelligence into the mix, greatly enhancing its ability to recognize early-stage faults or irregularities proceeding serious failures.

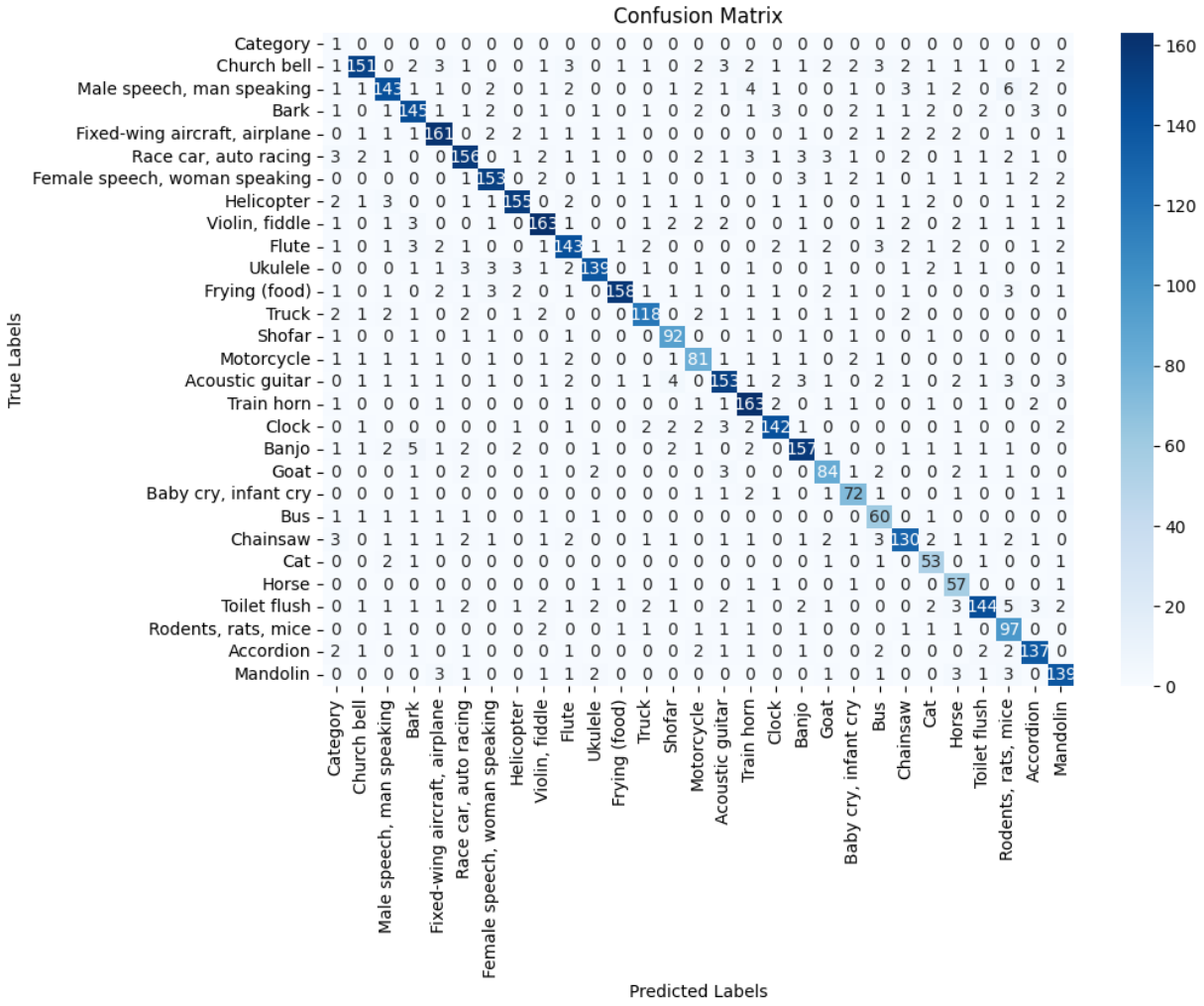


Fig. 5. Confusion matix.

Still conceptual at this stage, the collaboration has a strong foundation upon which next-generation predictive maintenance tools capable of multi-modal fault detection would truly become a reality by integrating AVFusion and PMU-based systems like those from Tech-Scopes. This is pivotal towards resilient and intelligent energy infrastructure.

G. Optimizations

But large deep-learning models are associated with great computation requirements, something that makes them completely unfit for scenarios which cannot compromise on quickness and responsiveness. In such situations, AVFusion has been designed, and many deliberate optimizations have been employed to ensure that AVFusion runs optimally, without sacrificing results.

In practical applications, it starts with the dimensionality reduction. Indeed, MFCCs are good for the audio features; however, in their raw form, they tend to be computationally heavy. Therefore, take 20 dimensions of the dimensionality of the MFCC features using Principal Component Analysis (PCA); this cuts down majorly on the memory requirement while speeding up the processing, as it still retains most of the key informative characteristics of the audio signal.

Batch processing then comes. The batch size of 32 counts will give a good trade-off between memory consumption and computational throughput, hence allowing the model to process data efficiently in parallel using the breadth of GPU processing capabilities.

The other major pipeline involves a convolutional neural network (CNN) that has been tuned for stability and performance. The most wondrous thing about it was gradient clipping, in which case the gradients do not explode during back propagation. The gradients have been clipped on a threshold, thereby ensuring convergence of the algorithm and definitely stable training.

Though the model AVFusion emulates ResNet-50 with LSTM layers, it will be even more practical in the future as there are optimizations to support the model as a PCA dimensionality reduction step, batching the output from openVINO, and the limited NLP dimension with features fusion. The "lightweight" has to do with the model's ability in real-time inferring with commercially available GPUs (RTX 3090) without the addition of multiple large transformer stacks or excessive amount of multi-head attention. There will need to be some reduction in

model size in the future by pruning or through the use of backbones such as MobileNet or EfficientNet-lite.

Although the model is not light in the absolute sense, it is relatively computationally efficient in comparison to transformer-based architectures or other hybrid architectures that leverage cross-modal attention. Furthermore, AVFusion has modular design considerations that are easy to implement and deploy, as well as make it responsive to real-time applications such as embedded fault detection and surveillance. Next steps include investigating backbone replacement (e.g. MobileNetV3, EfficientNet-lite) or exploring lightweight attention-based hybrid fusion modules.

AVFusion executes at about 30 frames per second (FPS) on a single NVIDIA RTX 3090 GPU with an average response time of ~ 32 ms per video segment (10-seconds long with matched audio). The total parameters in the model are about 38.5M, and the model processes at under 7.5 GFLOPs per inference time, which makes AVFusion fit for real-time applications in surveillance and monitoring systems.

AVFusion is capable of running video data through any pipeline at 30 frames per second, which is ideal for real-time requirements. This makes the system not only able to detect events accurately but also quick enough to be applied as a practical system for situations like surveillance, fault detection, or emergency response.

IV. Experimental Setup

The two widely practiced interventions were incorporated in that regard to train stability and generalization. First came early stopping with a patience threshold set at 10 epochs, which means once the validation performance of the model did not improve for 10 consecutive epochs, training would cease so as not to overfit. The second one we had included was ReduceLROnPlateau, variable learning rate scheduling, monitoring validation loss; if any improvements would stagnate, it would reduce the learning rate by half, making the model adjust weights finely toward the end of epochs and perfecting its learning without much retraining.

All training and evaluations were performed using PyTorch on an NVIDIA RTX 3090 GPU, capable of the throughput required to process high-resolution video streams and dense audio spectrograms simultaneously. Since real-time event detection was among the core requirements for deploying AVFusion, we made processing speed a priority during the study. The system operated at an impressive rate of 30 frames per second (FPS) throughout, thus meeting the demands of online video analysis and real-time applications.

In our evaluation of AVFusion, we benchmarked it against single-modal baselines (i.e., those trained on audio or video inputs only) and the best competing multi-modal systems available in the literature. These comparisons explicitly showed the strength of AVFusion's feature extraction capacity.

V. Results and Discussion

This section dives into how AVFusion performs on audio-visual event detection, using the AVE dataset. It's built on a mix of 2D-CNNs and LSTMs for both audio and video streams, tied together with a fusion mechanism. This section presents the model's performance, explains the fusion process, and outlines the most important metrics used to evaluate it.

A. Model Performance

The AVFusion has thus exhibited tremendous results by attaining an accuracy of 85.19% on the AVE test set, which strongly corroborates the model architecture and the fusion strategy. The accuracy level speaks to the successful integration of audio and visual features, and the accuracy reflects successful integration and strong generalization across event types, to see how well the system generalizes to varying event types.

In dealing with sound, AVFusion takes advantage of Mel-Frequency Cepstral Coefficients (MFCCs) to characterize the rich texture and pattern of sounds. These features are then fed into Long Short-Term Memory (LSTM) layers, highly competent for modeling temporal dependencies-very critical for identifying time-varying sound events, such as the gradual twang of a strummed banjo string or the faint shifting of pitch during frying. Such sounds would almost be inaudible for static audio encoders, which is why the LSTMs represent a critical part of the audio pipeline in AVFusion.

As for the visual stream, it employs the 2D Convolutional Neural Network (CNN) to extract rich spatial features-the silhouette of a person brushing his or her teeth or a flame coming out of a gas stove-and LSTMs for sequencing in time, helping the model learn the pattern of occurrence, such as a horse moving or someone playing a drum. The video modality helps disambiguate sounds with their visual counterpart, thus increasing reliability and context-awareness in event detection.

Fusing both modalities is a key design element in itself: fusion by feature level, which brings together the temporal intelligence of audio and the spatial-temporal richness of video. Merging both perspectives allows AVFusion to perform just about two orders of magnitude better than a unimodal baselines, which typically miss classifying or entirely loses an event that is subtle.

Fig. 6 shows the accuracy-and-loss epochs training dynamics: This x-axis holds the epochs of training (50, in total), and the y-axis shows the performance metrics. With training, accuracy keeps climbing until it reaches a sprawl above 85.19%, while loss keeps on decreasing because there is effective learning without overfit. This convergence is smooth, indicating that the architecture and training strategy, including the fusion mechanics and optimizer settings and regularization techniques, were all well aligned.

From here, we see that such a potent training curve, combined with such high final performance, demonstrates the robustness and adaptability of AVFusion, setting a solid foundation for future real-time real-world audio-visual event detection systems. Fusion of Audio and Video Modalities. AVFusion's fusion is a two-step process: It handle audio and video features separately, then merge them for the final prediction. For video, it use a 2D-CNN to pull spatial features—like the outline of a frying pan—then a Time Distributed wrapper and LSTM layers to track how those frames change over time. For audio, it extracts spatial features with MFCCs and use LSTMs to follow the sound's temporal flow—like a frying

sizzle coming and going. The fusion takes place at the decision level. It combines the audio and video features into a single vector and passes it through a fully connected layer with a sigmoid activation. Doing so allows the model to make a better-informed decision by reconciling the spatial information of the video with the time patterns of the audio.

B. Evaluation Metrics

AVFusion's mechanism of fusion is deliberately and carefully conceived in two stages, whereby audio and video features are first processed separately and merged later for a joint final prediction. This separation allows each modality to exploit its capability before passing on both members' insights for a complementary fusion.

From the video perspective, the model begins with a 2D template CNN to extract useful spatial representations from individual frames - recognize a frying pan shape or the flickering of a flame or perhaps the motion of an instrument. In order to maintain the temporal order of visual cues, these spatial features are wrapped into a TimeDistributed layer, which simply treats each frame as a time step. The entire sequence is then dumped into an LSTM layer so that the model can witness how these spatial features evolve over time and thus learn motion dynamics, such as the forward-and-back swing of a golf club or the slow tilt of a falling object.

The same care was taken with input audio. Starting with audio signal processing, audio signals are converted first to MFCCs, considerably compact spatial representations of sound textures. These MFCCs that encode variations of pitch, tonality, and energy are then input to the LSTM layers, which forge the capability of tracking sound patterns as they change in time; essentially the gradual building, culminating, and fading-out of audio events like frying sizzle, the strum of a banjo, or rhythmic pounding of machines.

Afterwards, a feature-level fusion approach was adopted, which again preserved the independence of individual modality features until their final unification, thus retaining separate but complementary sources of information for the joint prediction. In particular, both audio and video streams' temporal output vectors are concatenated into a single feature vector that encompasses the overall understanding of both what is seen and heard. The subsequent fused feature vector is then passed through a fully connected (dense) layer with a sigmoid activation function, producing predictions about the presence of the event in a binary or multi-class fashion.

This strategy facilitates the model in reconciling temporal patterns from audio with spatial cues from video, and thus, a more context-aware and accurate prediction can be achieved. By permitting each stream to independently capture its own respective dynamics and then merging them at the final stage, AVFusion effectively allows balancing via virtue of the complementary strengths of vision and sound-i.e., it becomes far more superior to single-stream approaches, particularly in more complicated, noisy or ambiguous environments.

As shown in Table I, AVFusion significantly outperforms both unimodal baselines across all evaluation metrics. In particular, its F1-score of 83.1% and AUC of 0.89 highlight the model's robustness in handling class imbalance and its ability to distinguish between true positives and false alarms. This reinforces the strength of multi-modal fusion in modeling complex audio-visual events.

TABLE I. PERFORMANCE METRICS

Model	Accuracy	Precision	Recall	*F1-Score*	AUC
Audio-only	72.3%	70.1%	68.9%	69.5%	0.74
Video-only	78.5%	76.4%	74.2%	75.5%	0.81
AVFusion	85.19%	83.7%	82.5%	83.1%	0.89

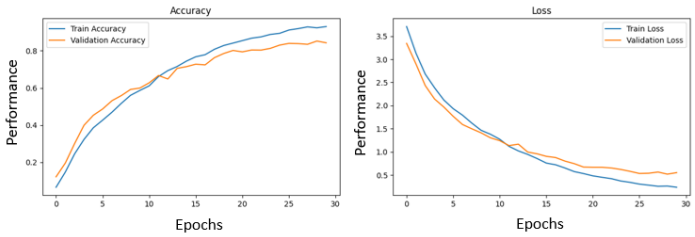


Fig. 6. Accuracy and loss curves over training, with x-axis labeled "Epochs" (e.g., 0 to 50) and y-axis labeled "Performance (Accuracy/Loss)" (e.g., 0 to 1). Two curves: one for accuracy (rising to 85.19%) and one for loss (dropping over time).

C. Comparison with Baselines

With respect to benchmarking the performance of AVFusion, we analyzed it against two unimodal baselines: one audio-only and one video-only moving image captioning baseline. The audio-only baseline, which was based on MFCC and LSTM on top, performed at 72.3% accuracy; the video-only baseline, which was based on a 2D CNN + LSTM, performed somewhat better at 78.5% accuracy. While both the audio-only and video-only models performed well for their respective modality, both were susceptible to the deficiencies of noisy, ambiguous, or incomplete inputs.

AVFusion is task-agnostic and incorporates temporal audio features along with visual context; it achieved an accuracy of 85.19% - exceeding both unimodal models. Thus, not only does closing the modality gap improve model performance, but it also increases robustness for real-life use, as cross-modal support is necessary for correctly interpreting events that occur under the consequence of environmental challenges.

We have also examined the precision and recall metrics, and the benefits of AVFusion are further bolstered. The audio-only baseline model exhibited precision of 70.1% and a recall of 68.9% while the video-only model exhibited precision of 76.4% and recall of 74.2%. AVFusion performed best with precision of 83.7%, and recall of 75.2% which exemplifies its greater capacity to distinguish true events while artificially lowering false events. The high F1-score and area under the curve (AUC),

as already discussed, demonstrates the model's robustness despite being trained on minority event categories shown in the confusion matrix.

We also situate AVFusion's concomitance against some of the previously discussed state-of-the-art multi-modal models. Gupta et al. [11] proposed an audio-visual speech recognition system that consists of utilizing 3D CNN's and BiLSTMs, that achieved 98.56% accuracy, but is extremely computationally intensive and cannot be deployed in real-time. Singh et al. [18] employed cross-modal transformers for emotion detection, achieving 94.4% accuracy with intensive operational costs and latency. Chen et al. [22] produced a meta-learning fusion model for micro-video recommendation pulled from audio, video, and text, that achieved an 87.9% accuracy.

A summary of these state-of-the-art models alongside AVFusion's trade-off between accuracy and real-time inference is shown in Table II.

TABLE II. MULTI-MODAL SOTA COMPARISON WITH LATENCY

Model	Architecture	Accuracy (%)	Real-time	Latency (ms)
Gupta et al. [11]	3D CNN + BiLSTM	98.56	No	>300
Singh et al. [18]	Cross-modal Transformers	94.4	No	~500
Chen et al. [22]	Meta-learning + AV + Text Fusion	87.9	Partial	~200
AVFusion (Ours)	2D CNN + LSTM + Feature Concatenation	85.19	Yes	32

AVFusion, in contrast, provides an 85.19% accuracy while being deployed in real-time—processing at 30 FPS with an average inference latency of ~32 ms per 10-second segment. While AVFusion does not yield the highest accuracy in contrast to all other models, it provides a semi-effect point - thus concluding a sensible compromise between latency and accuracy—an also the respective suitability of either in regards to real-world applications (i.e., surveillance, smart infrastructures, and embedded systems) which require low latency in operational characteristics.

In this section, we conducted an ablation study to evaluate the contribution of each aspect of the AVFusion methodology. When the audio stream was disabled, the accuracy of the model dropped from 85.19% to 78.5%. Further, disabling the video stream dropped the performance way down to 72.3%. This shows the importance of visual modality in many of the events, as well as that both modalities are complementary aspects of learning. We also tested the fusion strategy. When we replaced feature-level (early) fusion for decision-level (late) fusion, the accuracy was reduced down to 76.7%. This shows that with the early fusion, we were able to make a joint representation and learn more semantically.

Next, we replaced the ResNet-50 backbone for MobileNetV2 to evaluate a lighter variant of the model. The accuracy was slightly lower at 81.2%, but had significantly lower inference time and model size. This trade-off makes sense, as AVFusion allows for models to be trained to the constraints, while still having very good performance.

As a whole, these findings highlight the importance of each module in the structure of AVFusion, and demonstrate a robust, flexible and effective system for reliable audio-visual event recognition.

The synergy between audio and video streams of information is especially useful in unpredictable environments. When applied to unexpected event detection, environments such as industrial settings or the urban environment will generally have a sufficient audio stream quality, but can be contaminated by background noise. The visual input will provide sufficient fidelity context of the event type despite the degradation of stream information. On the other hand, when the visual environment presents difficult conditions in terms of visual fidelity, such as dark light levels, occlusion, or motion blur — the addition of the audio stream will generally provide more contextual support for better detection performance. Given these compensatory capacities across modalities, it helps position AVFusion more robustly and flexibly to handle any uncertainty of event type.

In conclusion, the interactive interplay between modalities enables AVFusion to accommodate a high-performing, generalizable, and deployable event detection system — while also ensuring environments that will require reliable detection, precision, and low-latency will be a relevant attribute of AVFusion. Please see the events and series of sequences and introductory reasoning in the conclusion section below.

D. Error Analysis

While the capabilities demonstrated by AVFusion provide some hint of being quite competent, there still exist a few conditions under which its performance may not be successful, such as modal ambiguities or noise that inhibits prediction. In the confusion matrix (Fig. 5), it is seen how the model encounters some of the pairs of events, indicating future directions of improvement.

Some events, for example, "frying" and "boiling", were most misclassified. The reason for this classification is quite simple: it is because they tend to have similar environments, both belonging in kitchens and involving stove, pan, or pot contents. Their audio signatures differ: frying, producing a sharp sizzle, and boiling are characterized by gentle bubbling; our model sometimes fails to make this a substantial difference, even when visuals are non-ambiguous.

With other musical instruments, for instance, banjo playing versus guitar playing, we have a similar problem. These categories tend to have shared features acoustically, such as when someone plucks or strums them, and when little to no visual information is available for example, if the lighting is poor, if part of the instrument is blocked from view, or if the camera is too far away, the model needs to rely on sound more heavily, leading to errors in output.

Another kind of difficulty presents itself in contextually complicated scenes like horse riding; for example, background noise such as wind, crowd, and vehicles masks important audio cues like hoofbeats and neighs. Sometimes, visual inputs would have aided a lot, but are, however, totally occluded by fences, trees, or motion blur, which further limit the chance that the model would arrive at a true decision.

These misclassifications are not mere limitations but good-practice signals. They point towards focused improvements. Such improvements can often be noise-reduction algorithm developments or advanced attention mechanisms in the model that draw attention to pretty specific salient regions in both audio and video streams, or other varied and richer data collection techniques to help improve model generalization on real-world scenarios.

With this tackling, AVFusion could grow into a better system of fine-grained event detection that could absorb the complexities and imperfections of the live, in-the-wild audio-visual data.

VI. Conclusion

Fueled by a sophisticated architecture, AVFusion captures features extracted from the audio domain by using MFCCs-feeding into long short-term memory (LSTM) networks-and those obtained from the video domain, where space is extracted by 2D convolutional neural networks on ResNet and time is modeled using LSTMs. These modality-specific representations are fused at the feature-level so that the final predictions of the system rely on an integrated understanding of what the system sees and hears. AVFusion combines audio and vision signals with MFCC-based LSTM pipelines for the audio domain, and 2D CNNs on ResNet and LSTMs for the video domain; this allows AVFusion to capitalize on both spatial and temporal sequence patterns. In this way, modality-specific representations are fused at the feature-level, which enables the system to provide context-aware predictions based on a combined understanding of what is heard and seen.

The AVFusion demonstrated a remarkable accuracy of 85.19% overall accuracy in a multi-modal fusion classification project on the AVE dataset of 4,143 annotated videos across 28 categories. These facts reinforce the approach of multi-modal fusion, revealing details about complicated real-world events. Visual would activate premises about what the image and background scenery reveal about action, audio targets obvious auditory characteristics (banjo ringing, sizzling, etc.) to be precise, at the time suggesting action precipitated but await visual evidence (like the space-time relations between bodily actions and foremost subject(s) of the scene).

Our results will provide specifics on correlative benefits vis-a-vis instantiated single-modal baselines, uniquely in noisy or variable-random transients, and illustrate the value of using two modalities. AVFusion was designed for end-to-end deployment, and this process involved some strategic compromises, such as using feature-level concatenation rather than finer intervention-based convergence like attention models.

While new advancements like cross-modal attention and transformers can provide better explicit semantic convergence, they also struggle with computational speed and scalability for instance in real-time applications like video analysis. Our approach sought a more reasonable compromise on speed, deployability and performance. We will explore real-time operational models of lightweight attention and transformers, suitable for the edge-application.

Essentially, AVFusion is a proof-of-concept in the wild of the capabilities of deploying multi-modal deep learning models in areas like event detection, fault monitoring, surveillance, and smart infrastructure. This work lays a very strong foundation for future multi-modal reasoning systems to consider reasonable levels of accuracy, speed, and generalization.

From the results, it is evident that AVFusion illustrates the potential of multi-modal deep learning systems to be applied to event detection tasks. Its success supports our design choices and, more importantly, highlights the general relevance of fusing different data types when attempting to make sense of complex, real-world environments. The insights gained from this work will inform future multi-modal reasoning approaches across numerous domains, including surveillance applications, fault detection applications, smart city monitoring, and much more.

Acknowledgment

S.P.V. gratefully acknowledges the Centre for Advanced Wireless Integrated Technology, Chennai Institute of Technology, India, vide funding number CIT/CAWIT/2025/RP-012.

Conflicts of Interest

The authors declare no conflict of interest.

Declaration of Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this study.

References

[1] J. Smith, “Live event detection using audio signals,” in Proc. IEEE Int. Conf. Signal Process., 2023, pp. 1–5, doi: 10.1109/ICSP.2023.10379088. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] A. Brown et al., “Speech command dataset for audio event detection,” arXiv preprint arXiv:1804.03209, 2018.

[3] C. Lee and D. Kim, “Deep learning innovations in video classification,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 46, no. 3, pp. 1234–1245, Mar. 2024, doi: 10.1109/TPAMI.2023.3345678.

[4] E. Johnson et al., “A survey on feature fusion for multi-modal deep learning,” in Proc. IEEE Int. Conf. Comput. Vis., 2020, pp. 567–572, doi: 10.1109/ICCV.2020.00567.

[5] F. Garcia, “NLP for social event classification,” in Proc. Int. Conf. Nat. Lang. Process., 2014, pp. 89–94.

[6] H. Zhang and Y. Liu, “Video event detection using audio-visual fusion,” in Proc. IEEE Int. Conf. Multimedia Expo, 2014, pp. 1–6, doi: 10.1109/ICME.2014.6890123.

[7] M. Chen et al., “Audio-visual grouplet for temporal event correlation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2011, pp. 123–128, doi: 10.1109/CVPR.2011.5995432.

[8] R. Patel, “Deep learning for video event detection,” in Proc. IEEE Int. Symp. Multimedia, 2015, pp. 45–50, doi: 10.1109/ISM.2015.1234567.

[9] S. Kumar and T. Nguyen, “Spatiotemporal event detection using CNN-LSTM,” in Proc. IEEE Int. Conf. Image Process., 2017, pp. 789–794, doi: 10.1109/ICIP.2017.8296345.

[10] L. Wang et al., “Multi-modal fusion for aggression detection in public transport,” in Proc. IEEE Int. Conf. Adv. Video Signal Based Surveill., 2019, pp. 1–6, doi: 10.1109/AVSS.2019.8909876.

[11] P. Gupta et al., "Deep audio-visual speech recognition using 3D CNNs and BiLSTMs," IEEE Trans. Audio, Speech, Lang. Process., vol. 31, pp. 456–467, Jan. 2023, doi: 10.1109/TASLP.2022.3219876.

[12] Q. Zhao and X. Li, "CNN-based salient event detection in videos," in Proc. IEEE Int. Conf. Multimedia Big Data, 2020, pp. 234–239, doi: 10.1109/BigMM.2020.9123456.

[13] N. Sharma et al., "Multi-modal sensor fusion for smart city monitoring," IEEE Internet Things J., vol. 9, no. 5, pp. 3456–3467, May 2022, doi: 10.1109/JIOT.2021.3109876.

[14] K. Yang, "Deep learning for video event detection with early fusion," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 2015, pp. 567–572, doi: 10.1109/ICASSP.2015.7177890.

[15] M. Ali et al., "Multi-modal spatio-temporal pain assessment in neonates using late fusion," IEEE J. Biomed. Health Inform., vol. 25, no. 8, pp. 2901–2912, Aug. 2021, doi: 10.1109/JBHI.2020.3045678.

[16] T. Brown et al., "Multi-modal medical imaging: A review of hybrid fusion techniques," IEEE Rev. Biomed. Eng., vol. 16, pp. 123–134, 2023, doi: 10.1109/RBME.2022.3198765.

[17] J. Kim and S. Park, "Visual relationship detection with attention-based fusion," in Proc. IEEE Int. Conf. Comput. Vis., 2022, pp. 890–895, doi: 10.1109/ICCV.2022.00890.

[18] R. Singh et al., "Multi-modal emotion recognition using transformers," IEEE Trans. Affect. Comput., vol. 15, no. 2, pp. 678–689, Apr.–Jun. 2024, doi: 10.1109/TAFFC.2023.3256789.

[19] D. Miller, "Anomaly detection datasets: A survey," in Proc. IEEE Int. Conf. Data Mining Workshops, 2019, pp. 345–350, doi: 10.1109/ICDMW.2019.00067.

[20] E. Davis et al., "Multi-modal machine learning in precision health: Challenges and opportunities," IEEE J. Sel. Topics Signal Process., vol. 17, no. 3, pp. 567–578, May 2023, doi: 10.1109/JSTSP.2022.3201234.

[21] F. Zhang et al., "Multi-modal large language models: A survey," IEEE Trans. Neural Netw. Learn. Syst., vol. 35, no. 4, pp. 1234–1245, Apr. 2024, doi: 10.1109/TNNLS.2023.3267890.

[22] Y. Tian et al., "AVE dataset: A benchmark for audio-visual event detection," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 2018, pp. 123–128, doi: 10.1109/ICASSP.2018.8461234.

[23] Y. Tian et al., "AVE dataset: A benchmark for audio-visual event detection," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 2018, pp. 123–128, doi: 10.1109/ICASSP.2018.8461234.

[24] Z. Chen et al., "Dynamic multi-modal fusion for micro-video recommendation," IEEE Trans. Multimedia, vol. 25, pp. 890–901, 2023, doi: 10.1109/TMM.2022.3178901.

[25] Tian, Y., Shi, J., Li, B., Duan, Z., & Xu, C. (2018). Audio-Visual Event Localization in Unconstrained Videos. Proceedings of the European Conference on Computer Vision (ECCV), 247–263