

Enhanced pre-broadcast video codec validation using hybrid CNN-LSTM with attention and autoencoder-based anomaly detection

Khalid El Fayq, Said Tkatek, Lahcen Idougli

Laboratory for Computer Science Research, Faculty of Science, Ibn Tofail University, Kenitra, Morocco

Article Info

Article history:

Received Aug 16, 2024

Revised Mar 24, 2025

Accepted Jun 8, 2025

Keywords:

Autoencoder-anomaly detection

Data augmentation

Machine learning

Video codec errors

Video codec validation

Video metadata analysis

ABSTRACT

This study presents a machine learning-based approach for proactive video codec error detection, ensuring uninterrupted television broadcasting for TV Laayoune, part of Morocco's SNRT network. Building upon previous approaches, our method introduces autoencoders for improved anomaly detection and integrates data augmentation to enhance model resilience to rare codec configurations. By combining convolutional neural networks (CNNs) and long short-term memory (LSTM) networks with an attention mechanism, the system effectively captures spatial and temporal video features. This architecture emphasizes critical metadata attributes that influence video playback quality. Embedded within the broadcasting pipeline, the model enables real-time error detection and alerts, minimizing manual intervention and reducing transmission disruptions. Experimental results demonstrate a 97% accuracy in detecting codec errors, outperforming traditional machine learning models. This study highlights the transformative role of machine learning in broadcasting, enabling scalable deployment across diverse television networks.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Khalid El Fayq

Laboratory for Computer Science Research, Faculty of Science, Ibn Tofail University

Kenitra, Morocco

Email: khalidelfayq@gmail.com

1. INTRODUCTION

Television broadcasting is crucial in delivering content to large audiences, making uninterrupted, high-quality video transmission essential. TV Laayoune, part of Morocco's national broadcasting network, often faces disruptions caused by video codec incompatibilities, particularly during live broadcasts. TV Laayoune relies on the proprietary 'Origo' server as one of ten channels operating on a shared infrastructure. Although the server adheres to international broadcasting standards, codec inconsistencies persist, leading to playback issues, interruptions, or complete signal loss. These errors frequently originate from various cameras and video formats used across the network. Additionally, post-production export processes can introduce further codec mismatches. To address this challenge, we propose an automated, machine learning-based pre-broadcast validation system that utilizes autoencoders for anomaly detection, synthetic data generation, and a hybrid convolutional neural network (CNN) and long short-term memory (LSTM) model with attention mechanisms. By analyzing spatial and temporal metadata extracted through FFmpeg, the system provides real-time error detection and automated alerts, enhancing overall broadcasting reliability.

The primary objective is to ensure uninterrupted broadcasting by proactively detecting incompatible video codecs, reducing disruptions, and improving operational efficiency across all channels using the shared infrastructure. The proposed solution enhances critical performance metrics, including accuracy, precision,

recall, and F1-score, strengthening the resilience of video transmission systems. Designed to be scalable, this approach not only addresses codec errors for TV Laayoune but extends its benefits to other SNRT channels operating under the same infrastructure.

Origo, the proprietary broadcast server, is central to video ingestion, processing, and transmission as shown in Figures 1 and 2. This system forms the backbone of SNRT's broadcasting network, but codec inconsistencies frequently disrupt its operation. This paper delves into the integration of the proposed machine learning system with Origo, aiming to preempt video codec errors that may disrupt live broadcasts, ensuring seamless television transmission.

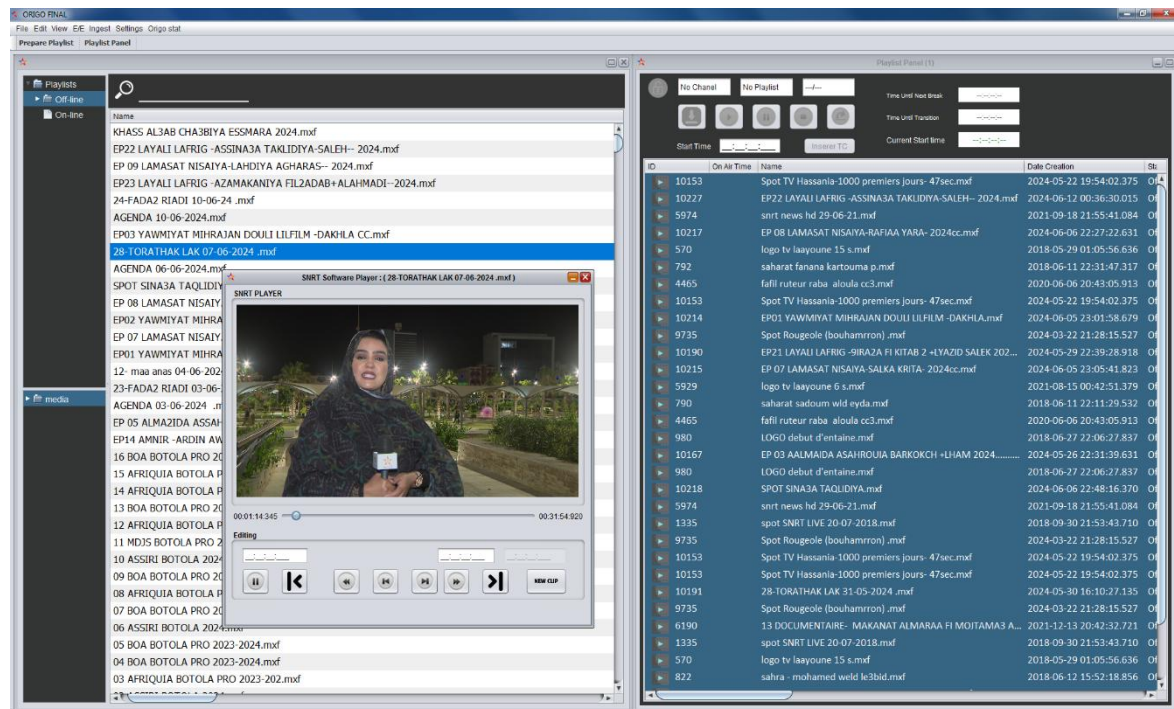


Figure 1. Origo server broadcasting interface for TV Laayoune

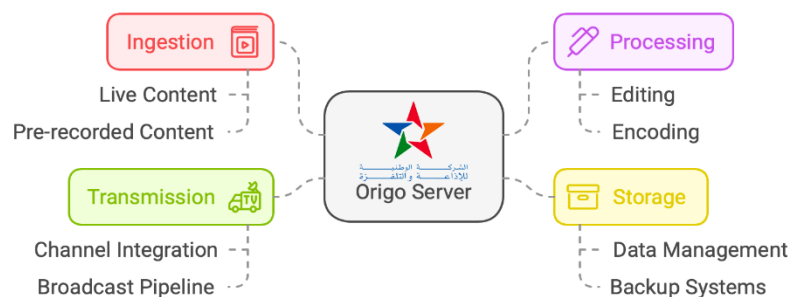


Figure 2. Origo server workflow: ingestion, processing, storage, and transmission

The increasing complexity of video formats and the expansion of digital broadcasting networks have amplified the challenges posed by video codec errors. Researchers have proposed various solutions, ranging from traditional heuristic-based techniques to advanced machine learning models designed to automate and enhance error detection. Conventional methods for detecting codec errors primarily rely on manual inspections or predefined rule-based systems, often resulting in inefficiencies and inaccuracies. Heuristic algorithms apply predefined patterns to identify common errors, but their static nature limits adaptability to evolving video formats. Signal processing techniques, while effective for detecting artifacts and synchronization errors, frequently fall short in real-time scenarios, leaving broadcasters vulnerable to unexpected disruptions during live transmissions.

Machine learning models offer a transformative alternative by learning from large datasets and dynamically adapting to new error patterns. This adaptability is crucial in evolving broadcasting environments where video formats and codecs frequently change. Klink *et al.* [1] reviewed machine learning frameworks for video quality prediction, illustrating the limitations of traditional methods in handling diverse video inputs. Oprea *et al.* [2] highlighted how deep video compression techniques leverage neural networks to enhance processing efficiency, showcasing the potential of machine learning in video broadcasting workflows. Muskaan *et al.* [3] demonstrated the effectiveness of CNN-LSTM architectures in identifying anomalies, such as deepfakes, further emphasizing the utility of machine learning in video integrity checks.

Hybrid models that combine CNNs and LSTMs are increasingly popular for video processing tasks. CNNs excel at extracting critical spatial features from video frames, while LSTMs capture temporal dependencies across sequences, making them ideal for analyzing video streams. Kaur and Mishra [4] employed LSTM to generate concise video summaries from lengthy sequences, highlighting the significance of temporal analysis in video data. Bidwe *et al.* [5] demonstrated the success of this architecture for video compression, while Benoughidene and Titouna [6] applied CNN-LSTM models for video shot boundary detection. Panneerselvam *et al.* [7] explored efficient video compression using deep learning techniques, underlining the benefits of combining CNNs and LSTMs for complex data patterns.

Despite their effectiveness, traditional CNN-LSTM models may overlook subtle or rare anomalies. Attention mechanisms further improve performance by enabling the network to prioritize the most relevant features during processing [8]. Autoencoders provide a robust, unsupervised learning method for anomaly detection in video streams. These models reconstruct video frames or metadata and flag discrepancies by analyzing reconstruction errors. Gashnikov [9] successfully applied autoencoders to video codec validation, achieving superior accuracy over traditional systems. By learning to reconstruct video inputs, autoencoders effectively detect anomalies indicative of codec mismatches or data corruption. Integrating autoencoders with CNN-LSTM models further enhances the overall detection pipeline, particularly for rare or subtle codec errors that might otherwise escape traditional detection methods. Augmenting datasets through simulated video configurations, bitrate alterations, and synthetic video generation has proven essential for enhancing model generalizability. Wang *et al.* [10] underscored the importance of synthetic data in training machine learning models for video enhancement. This approach diversifies the training set, exposing the model to a wide range of broadcasting scenarios, ultimately enhancing detection performance.

While prior studies address aspects of video processing and error detection, limited research focuses on proactive, pre-broadcast codec validation in live television environments. This study bridges this gap by integrating autoencoders, CNN-LSTM models with attention mechanisms [11], and extensive data augmentation to create a comprehensive pre-broadcast video codec validation system [12]. The proposed model leverages spatial and temporal metadata features extracted through FFmpeg, ensuring high detection accuracy while operating in real time.

CNNs have been widely adopted for video and image analysis due to their ability to effectively capture spatial hierarchies in data [13]. Yan *et al.* [14] leveraged CNNs for fractional-pixel motion compensation, demonstrating their utility in video processing. Cui *et al.* [15] applied CNN-based post-filtering to compressed images and videos, achieving notable improvements compared to traditional techniques. Additionally, El Fayq *et al.* [16] employed machine learning to detect and extract faces and text from audiovisual archives, highlighting the versatility of CNNs in various applications.

Machine learning applications in broadcasting environments have been widely explored. Darwich and Bayoumi [17] integrated CNN and recurrent neural network (RNN) models for video quality adaptation, reducing live broadcast disruptions. Sharrah *et al.* [18] demonstrated the role of machine learning in real-time video communication, highlighting automated error detection as critical component. Bouaafia *et al.* [19] applied deep learning-based video quality enhancement techniques, achieving significant improvements in video processing. Chen *et al.* [20] developed RL-AFEC, a reinforcement learning-based adaptive forward error correction system, demonstrating the potential of advanced machine learning techniques for error detection.

Despite significant progress, challenges persist in ensuring real-time performance and scalability. Achieving low-latency detection without compromising accuracy requires optimizing model architectures and expanding datasets to cover diverse codec configurations. Liu *et al.* [21] emphasized dataset diversity as essential for video coding models, while Ma *et al.* [22] addressed the need for real-time optimizations in neural networks for video compression. Zhang *et al.* [23] reviewed machine learning-based approaches to video coding optimizations, highlighting the importance of developing robust and scalable solutions.

The integration of machine learning techniques—such as autoencoders, CNN-LSTM models with attention mechanisms, and data augmentation—represents a significant advancement in video codec error detection. These advancements enhance reliability and efficiency across television broadcasting systems. Our proposed system builds upon this foundation, integrating these methods to proactively detect and mitigate video codec errors, ensuring seamless broadcasting for TV Laayoune and other SNRT channels.

Research has consistently highlighted the importance of accuracy, precision, recall, and F1-score in evaluating the effectiveness of these models. Studies report improvements in these metrics when using machine learning-based approaches compared to traditional methods. Steinert and Stabernack [24] designed a low latency H.264/AVC video codec for robust machine learning-based image classification, demonstrating significant improvements in precision and recall. Additionally, Putri *et al.* [25] conducted a comparative analysis of video quality using codecs VP8 and H.265, highlighting how codec performance impacts video communication quality. Their findings emphasize the importance of selecting and validating codecs to minimize packet loss and optimize throughput, further reinforcing the need for machine learning-driven approaches to codec error detection.

The rest of this paper is organized as follows: section 2 outlines the methodology. Section 3 presents experimental results. Finally, section 4 concludes the study with directions for future research.

2. METHOD

2.1. Datasets

The effectiveness of any machine learning model, particularly for video codec error detection, depends on the quality, diversity, and representativeness of the datasets used during training and evaluation. In this study, an extensive dataset was compiled, comprising over 10,000 video clips sourced from TV Laayoune's internal archives and publicly available repositories. This dataset reflects real-world broadcasting conditions commonly encountered by TV Laayoune and other channels within the SNRT network.

Metadata extraction, a critical component of dataset preparation, was conducted using FFmpeg, a widely used multimedia processing framework. Through this automated process, essential metadata attributes were extracted from each video clip, including codec type, resolution, frame rate, audio codec, bitrate, container format, and aspect ratio. These attributes serve as input features for the machine learning model, providing crucial insights into the technical specifications of each clip. Table 1 summarizes the key metadata fields extracted, while Figure 3 illustrates the metadata extraction process, highlighting the key stages from video ingestion to feature storage.

Table 1. Metadata extracted using FFmpeg

Metadata type	Description
Codec type	H.264, and MPEG-4
Resolution	720p, 1080p
Frame rate	24 fps, 30 fps, 60 fps
Audio codec	AAC, MP3
Bitrate	1 Mbps, 5 Mbps
Container format	MP4, MKV, MXF, MOV
Aspect ratio	Display aspect ratio (16:9, 4:3)
Duration (sec)	Length of the video in seconds
File size (MB)	Size of the video file in megabytes
Chroma subsampling	Color compression format (4:2:0, 4:4:4)
Color depth	Color bit depth (8-bit, 10-bit)

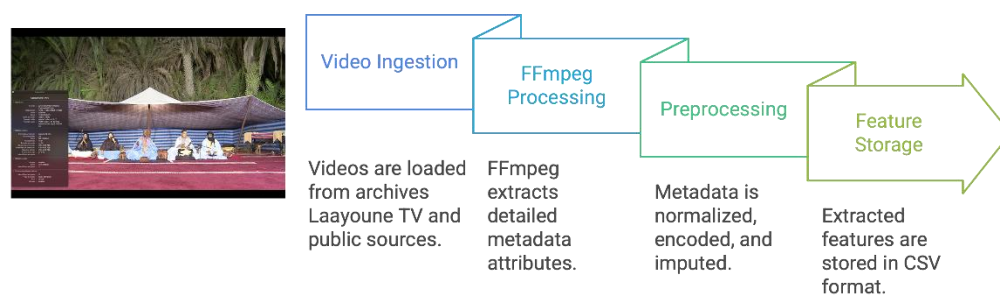


Figure 3. Metadata extraction process using FFmpeg

To ensure relevance to TV Laayoune's broadcasting environment, each video clip was manually annotated and labeled as compatible or incompatible with the station's broadcasting server. Annotation and labeling were conducted by experienced broadcast technicians, using historical playback logs and performance data to validate each clip. Compatible clips align with the server's codec and format

requirements, ensuring smooth playback. Conversely, incompatible clips exhibit playback errors, disruptions, or incompatibility issues during live transmissions. This labeled dataset enables the machine learning model to distinguish between error-free clips and problematic files.

To ensure a balanced and unbiased evaluation of the model, the dataset was split into three subsets. The training set comprises 70% of the data and contains an even distribution of compatible and incompatible clips to prevent model bias during learning. The validation set represents 15% of the data and is used to fine-tune the model's hyperparameters, ensuring optimal performance while mitigating overfitting. The remaining 15% constitutes the test set, exclusively reserved for final model evaluation to assess accuracy, precision, recall, and generalizability. This isolated test set ensures an objective measure of performance on unseen data.

Before model training, extensive preprocessing was applied to the dataset. Numerical features such as bitrate, frame rate, and file size were normalized to ensure uniform scaling across different feature ranges. Categorical attributes, including codec type, audio codec, and container format, were one-hot encoded to facilitate seamless integration into the machine learning pipeline. Missing or incomplete metadata entries were addressed through mean imputation for numerical data or removed if they represented less than 1% of the total dataset. This preprocessing ensured that the dataset was clean, consistent, and optimized for the machine learning model. By assembling a diverse, representative, and high-quality dataset, this study aims to enhance the accuracy and robustness of codec error detection, ensuring seamless and uninterrupted television broadcasting for TV Laayoune and other channels in the SNRT network.

2.2. Proposed model

2.2.1. Model architecture

The proposed model adopts a hybrid architecture combining CNNs and LSTM networks to detect video codec errors effectively. This integration leverages the strengths of CNNs for spatial feature extraction and LSTM networks for modeling temporal dependencies within the metadata of video clips. By capturing both static metadata attributes and sequential patterns, the model ensures comprehensive analysis of potential codec incompatibilities across video frames.

The CNN component of the model is responsible for extracting spatial features from key metadata fields, including resolution, codec type, and bitrate. Convolutional layers apply multiple filters to emphasize critical attributes that influence codec compatibility, enabling the model to identify subtle spatial patterns in the metadata. This layer plays a crucial role in detecting anomalies related to static video attributes, such as improper resolutions or unsupported codecs.

Following the CNN layers, the extracted feature maps are passed to the LSTM component, which processes sequential metadata over time. LSTM networks excel at capturing temporal dependencies, allowing the model to detect irregularities such as fluctuating frame rates or inconsistent bitrates that may signal potential codec errors. This sequential modeling is essential for recognizing patterns that manifest over multiple video segments, contributing to improved detection accuracy.

To further enhance the model's performance, an attention mechanism is incorporated into the LSTM layer outputs. The attention layer selectively prioritizes the most relevant features by dynamically weighting the importance of different metadata attributes. This focus on critical features not only boosts detection accuracy but also reduces false positives (FP) by minimizing the influence of less significant metadata patterns.

A visual representation of the model architecture is illustrated in Figure 4, demonstrating the flow of data through the hybrid CNN-LSTM structure. The diagram outlines the sequence of operations from metadata ingestion, convolutional filtering, LSTM processing, and attention-based feature selection, culminating in the final output layer responsible for codec compatibility classification. This hybrid design ensures a robust, scalable solution for identifying and mitigating video codec errors in real-time broadcasting environments.

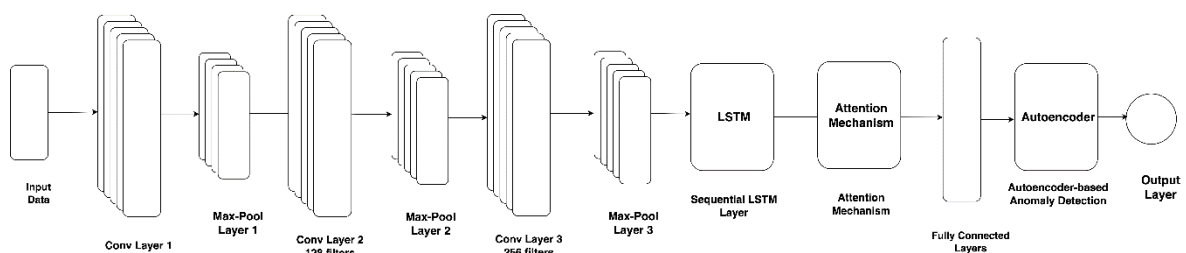


Figure 4. Machine learning model architecture for video codec error detection

Table 2 outlines the hyperparameters governing the hybrid CNN-LSTM architecture integrated with attention mechanisms and autoencoder-based anomaly detection. These hyperparameters shape the architecture and training process, optimizing the CNN for spatial feature extraction, the LSTM network for modeling temporal dependencies, and the autoencoder for identifying anomalies. The model is fine-tuned to achieve efficient learning and robust performance. Key hyperparameters include a learning rate of 0.001, batch size of 32, and 35 training epochs to account for the additional complexity introduced by the attention mechanism. The dropout rate is set to 0.4 to mitigate overfitting, while convolutional layers utilize kernel sizes of 3×3 and filters at increasing depths (64, 128, 256, 256). The attention mechanism operates alongside the LSTM layer, enhancing feature prioritization, while the autoencoder anomaly detection unit is trained in parallel using the same dataset. Rectified linear unit (ReLU) remains the primary activation function, with He initialization applied for efficient weight distribution.

Table 2. Hyperparameters for CNN-LSTM with attention and autoencoder architecture

Hyperparameter	Value
Learning rate	0.001
Batch size	32
Number of epochs	35
Dropout rate	0.4
Kernel size	3×3
Filters	64, 128, 256, 256
LSTM units	128
Attention mechanism units	64
Autoencoder hidden size	128
Activation function	ReLU
Weight initialization	He initialization

2.2.2. Integration of autoencoders for anomaly detection

Autoencoders play a crucial role in detecting rare or subtle video codec anomalies that standard classification models might miss. As unsupervised learning models, autoencoders reconstruct input data by encoding it into a lower-dimensional latent space and decoding it back. By comparing the reconstructed metadata with the original input, discrepancies indicating potential codec errors are identified, enhancing system robustness against anomalies that could disrupt broadcasts. The autoencoder is trained on metadata from compatible video codecs, learning to reconstruct this data with minimal error. When presented with new data, clips deviating from the learned distribution produce higher reconstruction errors, signaling potential codec inconsistencies. This prompts further review before integration into the broadcast pipeline.

The training uses a mean squared error (MSE) loss function over 100 epochs, with early stopping to prevent overfitting and improve generalization. The architecture includes three encoder layers for compression, mirrored by decoder layers for reconstruction. Dropout regularization is applied to ensure resilience. By complementing the CNN-LSTM architecture, the autoencoder adds an extra validation layer, improving the accuracy and reliability of codec error detection. This hybrid approach enhances seamless broadcasting for TV Laayoune and other channels on the network.

2.2.3. Data augmentation and synthetic data generation

Data augmentation and synthetic data generation are essential for enhancing the robustness and generalization of the proposed machine learning model. These techniques expand the training dataset by introducing controlled variations, simulating diverse broadcasting scenarios the model may encounter during live broadcasts. This process improves the model's adaptability, reducing overfitting and ensuring reliable performance across different video conditions.

Augmentation involves modifying existing video metadata to reflect varying codec configurations, frame rates, and bitrates. For example, video clips are adjusted to simulate frame rates of 24 fps, 30 fps, and 60 fps, with bitrates ranging from 1 Mbps to 5 Mbps. Controlled noise is also introduced to replicate common transmission errors, enabling the model to detect subtle discrepancies that signal codec issues. Synthetic data generation further diversifies the dataset by altering codec attributes, container formats (MP4, MXF, MOV), and aspect ratios (16:9, 4:3). This process creates new metadata samples that represent rare or edge-case errors, allowing the model to familiarize itself with unusual configurations that could disrupt broadcasts. By exposing the model to a broader array of data, augmentation, and synthetic generation strengthen its resilience against unexpected anomalies, enhancing its ability to detect codec errors across varied broadcasting conditions. This comprehensive approach significantly reduces biases and improves reliability, supporting seamless and uninterrupted television broadcasts for TV Laayoune and other channels within the network.

2.2.4. Training and validation

The training process follows a structured approach designed to improve the model's generalization and robustness, as illustrated in Figure 5. The model is trained using a supervised learning method with labeled video clips that indicate codec compatibility. To increase dataset diversity and resilience, various data augmentation techniques are applied, including modifications to resolutions, bitrates, and frame rates, as well as the introduction of controlled noise. These augmentations expose the model to a wide array of codec configurations, enhancing its ability to generalize effectively to unseen data in real-world broadcasting scenarios.

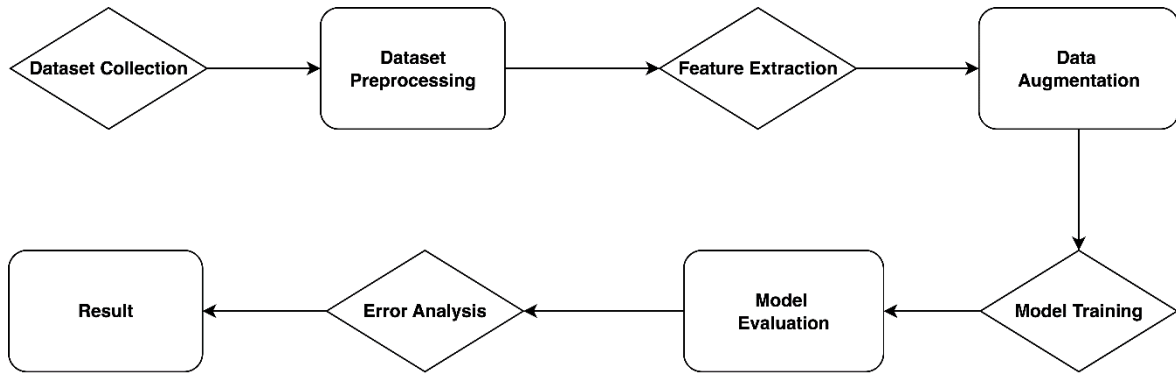


Figure 5. Flowchart of the video codec error detection methodology

Binary cross-entropy (BCE) is selected as the loss function due to its suitability for binary classification tasks, measuring the divergence between predicted probabilities and actual labels. The Adam optimizer is employed for its efficiency in handling large datasets and its adaptive learning rate, contributing to stable convergence throughout the training process. To ensure model reliability, a k-fold cross-validation strategy is implemented, dividing the dataset into k subsets. The model undergoes training and validation k times, utilizing a different subset for validation during each iteration while the remaining subsets are used for training. This comprehensive evaluation method allows the model to generalize across diverse data distributions. Additionally, early stopping based on validation loss prevents overfitting, ensuring optimal performance without excessive training epochs.

Considering the real-time constraints of broadcasting environments, mixed-precision training is explored to accelerate computations and minimize model size. This makes it suitable for deployment in resource-constrained environments. Future work will focus on further optimizations, including model pruning and quantization, to enhance inference speeds during live broadcasts.

An error analysis is planned to address the 5% misclassifications, guiding improvements such as the integration of additional features or adjustments to the model architecture for handling complex codec configurations. This comprehensive methodology-integrating data augmentation, cross-validation, regularization, and optimization-ensures that the model performs well during training. It also generalizes effectively to new video clips across various broadcasting conditions.

2.2.5. Model evaluation

The performance of the proposed hybrid model, which integrates CNN-LSTM networks with attention mechanisms and autoencoder-based anomaly detection, is evaluated using multiple metrics: accuracy (1), precision (2), recall (3), and F1-score (4). These metrics provide a comprehensive assessment of the model's ability to detect and classify video codec errors effectively.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

Where true positives (TP): instances correctly identified as positive; true negatives (TN): instances correctly identified as negative; FP: instances incorrectly identified as positive; and false negatives (FN): instances incorrectly identified as negative.

The model evaluation is conducted using the test set, which consists of video clips not used during the training or validation phases. This separation ensures an unbiased assessment, measuring the model's generalizability to unseen data. In addition to evaluating classification performance, the autoencoder's reconstruction error is analyzed to identify subtle codec anomalies that traditional models might overlook. Reconstruction errors are measured using MSE, where higher errors indicate greater deviations from normal patterns, signaling potential incompatibilities. This dual-evaluation approach allows for the detection of both explicit and nuanced errors, enhancing the overall robustness of the system.

The trained model is integrated directly into the TV Laayoune broadcasting pipeline to ensure operational efficiency. The workflow begins with FFmpeg continuously extracting metadata from incoming video clips in real time. This metadata, including codec type, resolution, frame rate, and bitrate, is preprocessed and passed to the CNN-LSTM model with attention. The model evaluates the spatial and temporal patterns of the metadata while the autoencoder concurrently assesses the anomaly score.

If the model detects a potential incompatibility, an alert is triggered for the broadcasting operator to review the flagged clip. Operators can then take corrective action, such as re-encoding the video or adjusting codec settings, preventing broadcasting errors before the clip reaches live transmission. This proactive process reduces manual intervention, minimizes errors, and enhances the overall broadcasting experience.

The integration of autoencoder anomaly detection alongside the CNN-LSTM architecture with attention offers a robust and scalable solution for codec error detection. This approach ensures that codec errors are identified at various levels—both through spatial-temporal analysis and anomaly-based reconstruction. It provides greater accuracy and resilience in video transmission across diverse broadcasting conditions.

3. RESULTS AND DISCUSSION

3.1. Training methodology

The experiments were conducted using a dataset comprising video clips from both internal archives of TV Laayoune and publicly available sources. The dataset was divided into training, validation, and test sets in the ratio of 70:15:15. The following tools and libraries were used:

- FFmpeg for feature extraction and preprocessing of video metadata.
- TensorFlow and Keras for building and training the machine learning model.
- Scikit-learn for performance evaluation and metrics calculation.

To assess the effectiveness of the hybrid CNN-LSTM model with attention and autoencoder anomaly detection, additional experiments were conducted under various conditions. These conditions reflect real-world broadcasting scenarios at TV Laayoune, including diverse codec types, resolutions, frame rates, and bitrate configurations. Table 3 summarizes the extended experimental results, highlighting accuracy, precision, recall, and F1-score across different parameter variations.

Table 3. Extended experimental results

Parameter	Value	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Codec type	H.264, MPEG-4	95.2	94.5	95.7	95.1
Resolution	720p, 1080p	94.8	94.0	95.0	94.5
Frame rate	24 fps, 30 fps, 60 fps	95.4	94.7	95.9	95.3
Bitrate	1 Mbps, 5 Mbps	94.9	94.2	95.2	94.7
Audio codec	AAC, MP3	95.1	94.4	95.6	95.0

The experimental results demonstrate consistently high accuracy (94.5%-95.2%) across various configurations, affirming the model's robustness and adaptability. Codec type and frame rate variations yielded the highest accuracy and recall, reflecting the model's effectiveness in managing temporal inconsistencies and diverse encoding formats. Similarly, variations in resolution and bitrate maintained strong performance, highlighting the model's capacity to process videos with differing visual quality and compression levels.

The integration of autoencoder-based anomaly detection significantly enhances the model's sensitivity to subtle codec errors that may evade traditional spatial or temporal analysis. This dual-layer

approach improves the detection of rare or edge-case codec anomalies. It contributes to elevated recall (up to 95.7%) and minimizing the risk of undetected errors.

These findings validate the model’s generalizability across various broadcasting environments, confirming its suitability for deployment within TV Laayoune’s broadcast pipeline. By addressing codec inconsistencies at multiple levels-spatial, temporal, and anomaly detection-the proposed solution provides a comprehensive framework. It helps reduce broadcasting disruptions and enhance television transmission reliability.

3.2. Model performance

The performance of the hybrid CNN-LSTM model, enhanced with an attention mechanism and autoencoder anomaly detection, was evaluated using key classification metrics: accuracy, precision, recall, and F1-score. These metrics provide a comprehensive assessment of the model’s ability to detect video codec errors efficiently. The model achieves consistently high performance across all metrics. The accuracy of 97.0% underscores the model’s exceptional reliability in correctly classifying video clips, while the precision of 96.3% highlights its effectiveness in minimizing FP. A recall of 97.5% reflects the model’s strong sensitivity in identifying incompatible video clips, ensuring minimal undetected errors. The F1-score of 96.9% balances precision and recall, reinforcing the model’s robustness and adaptability for real-world broadcasting conditions.

These results validate the model’s scalability and dependability across diverse broadcasting environments. The high recall is particularly essential for live broadcasts, reducing the risk of codec errors bypassing detection and causing interruptions. Simultaneously, the model’s high precision minimizes false alerts, allowing operators to focus only on genuinely incompatible clips. By integrating spatial, temporal, and anomaly-based detection mechanisms, the model outperforms traditional heuristic methods, establishing itself as a valuable addition to TV Laayoune’s broadcasting workflow. The system enables real-time detection and automated alerts, preempting potential disruptions and enhancing overall broadcasting reliability.

3.3. Comparative analysis

To evaluate the effectiveness of the proposed CNN-LSTM hybrid model with attention and autoencoder anomaly detection, its performance was compared against traditional heuristic-based methods and a baseline logistic regression model. As shown in Table 4, the hybrid model consistently outperforms both traditional and baseline approaches. This holds across all key performance metrics [26], [27].

Table 4. Comparative experimental results

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Heuristic-based	85.3	84.1	86.5	85.3
Logistic regression	89.5	88.7	90.1	89.4
CNN-LSTM hybrid (Proposed)	97.0	96.3	97.5	96.8

The superior performance of the hybrid model reflects the effectiveness of combining convolutional and recurrent neural networks for video codec error detection. This aligns with recent research demonstrating the advantages of machine learning in video coding, quality adaptation, and real-time anomaly detection [28], [29]. Key factors contributing to the enhanced performance:

- CNN+LSTM synergy: CNNs extract spatial features, while LSTMs capture temporal dependencies, allowing the model to detect complex patterns in metadata.
- Data augmentation: simulating different codec configurations, altering frame rates, and injecting noise increased the model’s generalizability to diverse broadcasting scenarios.
- Regularization techniques: dropout and L2 regularization effectively mitigated overfitting, improving the model’s resilience to unseen data.

During live broadcasts at TV Laayoune, the model successfully detected codec incompatibilities and synchronization errors in video clips, preventing transmission disruptions. Its integration into the Origo server pipeline allowed for real-time error detection, reducing manual quality checks and enhancing operational efficiency. Despite promising results, further improvements are necessary to ensure long-term scalability. Future work will focus on expanding the dataset to cover rare codec configurations, optimizing real-time processing, and incorporating visual/audio content analysis to address errors beyond metadata-based detection.

3.4. Discussion

The study demonstrates the effectiveness of integrating autoencoders, attention mechanisms, and a hybrid CNN-LSTM model for video codec error detection. The proposed approach consistently delivers high accuracy, reducing live broadcast disruptions by identifying incompatible video clips. Automating error detection streamlines quality control, minimizing manual inspections and conserving resources. The model integrates seamlessly with the Origo broadcast server, enabling real-time detection without adding significant computational load.

While the results are promising, further improvements are needed. Expanding the dataset to include diverse codec configurations and generating synthetic data for rare errors will enhance the model’s adaptability. Optimizing the system for larger video streams and lower latency is essential for real-time broadcasting. Future enhancements may also incorporate content-aware analysis to detect visual anomalies beyond metadata inconsistencies, strengthening the model’s value for TV Laayoune and the broader SNRT network.

4. CONCLUSION

This paper introduces an enhanced machine learning-driven approach to improve television broadcasting reliability by proactively detecting video codec errors. Addressing recurring issues with incompatible codecs, the proposed model targets disruptions during live broadcasts on TV Laayoune. By integrating CNNs, LSTM networks, autoencoders, and attention mechanisms, the system effectively identifies potential codec anomalies and alerts operators before errors disrupt live broadcasts. Using a diverse dataset, metadata was extracted through FFmpeg, and the model was trained to prevent overfitting and maximize precision. The combination of autoencoder-based anomaly detection with convolutional and recurrent layers enables the capture of spatial, temporal, and latent patterns, enhancing the model’s robustness and accuracy. Experimental results demonstrate notable accuracy improvements (97%) over traditional heuristic-based and baseline machine learning methods. The system integrates smoothly into the existing broadcasting pipeline, enabling real-time detection and automated alerts, ultimately boosting operational efficiency. Key advantages include improved detection rates through the hybrid CNN-LSTM architecture, enhanced anomaly detection with autoencoders, and reduced manual oversight by automating error identification. The seamless integration of the model into TV Laayoune’s workflow minimizes disruptions and ensures reliable video playback. Tests conducted across various video clips underscore the model’s capacity to prevent broadcast errors, reinforcing its potential as a critical tool for ensuring smooth and uninterrupted television transmission. In conclusion, the proposed machine learning model, enriched by autoencoders and attention mechanisms, provides a scalable solution for video codec error detection, significantly enhancing television broadcasting reliability. The substantial improvements in accuracy and operational efficiency highlight the model’s relevance in addressing complex broadcasting challenges, paving the way for broader applications across SNRT’s channels.

ACKNOWLEDGMENTS

We express our sincere gratitude to Ibn Tofail University for their unwavering support and encouragement throughout this project. Our heartfelt thanks also go to TV Laayoune for providing invaluable resources and fostering a collaborative environment, which played a crucial role in the development and implementation of our machine learning model for detecting video codec errors.

FUNDING INFORMATION

The authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Khalid El Fayq	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓			✓	✓
Said Tkatek		✓		✓	✓	✓		✓	✓	✓	✓		✓	
Lahcen Idouglid				✓		✓		✓		✓	✓			

C : C onceptualization	I : I nterpretation	Vi : V isualization
M : M ethodology	R : R esources	Su : S upervision
So : S oftware	D : D ata Curation	P : P roject administration
Va : V alidation	O : Writing - O riginal Draft	Fu : F unding acquisition
Fo : F ormal analysis	E : Writing - Review & E ditng	

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no conflict of interest.

DATA AVAILABILITY

The data used in this study are confidential and cannot be made publicly available or shared with other parties. These data were used exclusively for the purposes of this research. Due to privacy and confidentiality agreements, access to the dataset is restricted.





REFERENCES

- [1] J. Klink, M. Łuczyński, and S. Brachmański, "Video quality modelling—comparison of the classical and machine learning techniques," *Applied Sciences*, vol. 14, no. 16, 2024, doi: 10.3390/app14167029.
- [2] S. Oprea *et al.*, "A review on deep learning techniques for video prediction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 6, pp. 2806–2826, 2022, doi: 10.1109/TPAMI.2020.3045007.
- [3] A. Muskaan, S. Nagarathna, C. S. Sandhya, J. Viju, and B. Sumangala, "Exposing deep fake face detection using LSTM and CNN," *International Journal of Advanced Research in Science, Communication and Technology*, vol. 4, no. 5, pp. 231–234, 2024, doi: 10.48175/IJARST-18434.
- [4] L. Kaur and P. K. Mishra, "Estimation of concise video summaries from long sequence videos using deep learning via LSTM," *International Journal of Health Sciences*, vol. 6, no. S3, pp. 9904–9914, 2022, doi: 10.53730/ijhs.v6nS3.9287.
- [5] R. V. Bidwe *et al.*, "Deep learning approaches for video compression: a bibliometric analysis," *Big Data and Cognitive Computing*, vol. 6, no. 2, 2022, doi: 10.3390/bdcc6020044.
- [6] A. Benoughdene and F. Titouna, "A novel method for video shot boundary detection using CNN-LSTM approach," *International Journal of Multimedia Information Retrieval*, vol. 11, no. 4, pp. 653–667, 2022, doi: 10.1007/s13735-022-00251-8.
- [7] K. Panneerselvam, K. Mahesh, V. L. H. Josephine, and A. R. Kumar, "Effective and efficient video compression by the deep learning techniques," *Computer Systems Science and Engineering*, vol. 45, no. 2, pp. 1047–1061, 2023, doi: 10.32604/csse.2023.030513.
- [8] W. Ullah, A. Ullah, T. Hussain, Z. A. Khan, and S. W. Baik, "An efficient anomaly recognition framework using an attention residual LSTM in surveillance videos," *Sensors*, vol. 21, no. 8, Apr. 2021, doi: 10.3390/s21082811.
- [9] M. V. Gashnikov, "Video codec using machine learning based on parametric orthogonal filters," *Optical Memory and Neural Networks*, vol. 32, no. 4, pp. 226–232, 2023, doi: 10.3103/S1060992X23040021.
- [10] Z. Wang, J. Chen, and S. C. H. Hoi, "Deep learning for image super-resolution: a survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3365–3387, 2021, doi: 10.1109/TPAMI.2020.2982166.
- [11] J. Li, C. Xu, B. Feng, and H. Zhao, "Credit risk prediction model for listed companies based on CNN-LSTM and attention mechanism," *Electronics*, vol. 12, no. 7, 2023, doi: 10.3390/electronics12071643.
- [12] K. O. Babaali, E. Zigh, M. Djebbouri, and O. Chergui, "A new approach for road extraction using data augmentation and semantic segmentation," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 28, no. 3, pp. 1493–1501, 2022, doi: 10.11591/ijeecs.v28.i3.pp1493-1501.
- [13] F. E. Khalloufi, N. Rafalia, and J. Abouchabaka, "Customized convolutional neural networks for Moroccan traffic signs classification," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 36, no. 1, pp. 469–476, 2024, doi: 10.11591/ijeecs.v36.i1.pp469-476.
- [14] N. Yan, D. Liu, H. Li, B. Li, L. Li, and F. Wu, "Convolutional neural network-based fractional-pixel motion compensation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 3, pp. 840–853, 2019, doi: 10.1109/TCSVT.2018.2816932.
- [15] K. Cui, A. B. Koyuncu, A. Boev, E. Alshina, and E. Steinbach, "Convolutional neural network-based post-filtering for compressed YUV420 images and video," in *2021 Picture Coding Symposium (PCS)*, IEEE, 2021, pp. 1–5, doi: 10.1109/PCS50896.2021.9477486.
- [16] K. El Fayq, S. Tkatek, L. Idouglid, and J. Abouchabaka, "Detection and extraction of faces and text lower third techniques for an audiovisual archive system using machine learning," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 9, pp. 625–632, 2022, doi: 10.14569/IJACSA.2022.0130974.
- [17] M. Darwich and M. Bayoumi, "Video quality adaptation using CNN and RNN models for cost-effective and scalable video streaming services," *Cluster Computing*, vol. 27, no. 5, pp. 6355–6375, 2024, doi: 10.1007/s10586-024-04315-8.
- [18] Y. O. Sharrab, I. Alsmadi, and N. J. Sarhan, "Towards the availability of video communication in artificial intelligence-based computer vision systems utilizing a multi-objective function," *Cluster Computing*, vol. 25, no. 1, pp. 231–247, 2022, doi: 10.1007/s10586-021-03391-4.
- [19] S. Bouaafia, R. Khemiri, S. Messaoud, O. B. Ahmed, and F. E. Sayadi, "Deep learning-based video quality enhancement for the new versatile video coding," *Neural Computing and Applications*, vol. 34, no. 17, pp. 14135–14149, 2022, doi: 10.1007/s00521-021-06491-9.
- [20] K. Chen, H. Wang, S. Fang, X. Li, M. Ye, and H. J. Chao, "RL-AFEC: adaptive forward error correction for real-time video communication based on reinforcement learning," in *Proceedings of the 13th ACM Multimedia Systems Conference*, New York, United States: ACM, 2022, pp. 96–108, doi: 10.1145/3524273.3528184.





- [21] D. Liu, Y. Li, J. Lin, H. Li, and F. Wu, "Deep learning-based video coding," *ACM Computing Surveys*, vol. 53, no. 1, pp. 1–35, 2021, doi: 10.1145/3368405.
- [22] S. Ma, X. Zhang, C. Jia, Z. Zhao, S. Wang, and S. Wang, "Image and video compression with neural networks: a review," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 6, pp. 1683–1698, 2020, doi: 10.1109/TCSVT.2019.2910119.
- [23] Y. Zhang, S. Kwong, and S. Wang, "Machine learning based video coding optimizations: a survey," *Information Sciences*, vol. 506, pp. 395–423, 2020, doi: 10.1016/j.ins.2019.07.096.
- [24] F. Steinert and B. Stabernack, "Architecture of a low latency H.264/AVC video codec for robust ML based image classification: how region of interests can minimize the impact of coding artifacts," *Journal of Signal Processing Systems*, vol. 94, no. 7, pp. 693–708, 2022, doi: 10.1007/s11265-021-01727-2.
- [25] R. R. A. Putri, M. D. Atmadja, and M. Junus, "Comparison analysis of video quality on codec VP8 and H.265 PBX server with pixel conversion method," *Journal of Telecommunication Network*, vol. 12, no. 3, pp. 153–159, 2022.
- [26] X. Dou, X. Cao, and X. Zhang, "Region-of-interest based coding scheme for live videos," *Applied Sciences*, vol. 14, no. 9, 2024, doi: 10.3390/app14093823.
- [27] L. Chen, B. Cheng, H. Zhu, H. Qin, L. Deng, and L. Luo, "Fast versatile video coding (VVC) intra coding for power-constrained applications," *Electronics*, vol. 13, no. 11, 2024, doi: 10.3390/electronics13112150.
- [28] S.-K. Im and K.-H. Chan, "Faster intra-prediction of versatile video coding using a concatenate-designed CNN via DCT coefficients," *Electronics*, vol. 13, no. 11, 2024, doi: 10.3390/electronics13112214.
- [29] N. Li, Z. Wang, and Q. Zhang, "Fast coding unit partitioning algorithm for video coding standard based on block segmentation and block connection structure and CNN," *Electronics*, vol. 13, no. 9, 2024, doi: 10.3390/electronics13091767.

BIOGRAPHIES OF AUTHORS







Khalid El Fayq     was born in Laayoune, Morocco, in 1985. He obtained his engineering degree in computer science in 2020 from Agadir International University. Currently, he is a Ph.D. student at the Laboratory for Computer Science Research (LaRIT), Faculty of Science, Ibn Tofail University, Kenitra, Morocco. He works in the media field and serves as a temporary teacher at universities. His research interests encompass artificial intelligence and the audiovisual field. He can be contacted at email: khalidelfayq@gmail.com.



Said Tkatek     is a Professor of Computing Science Research at the Faculty of Science, Ibn Tofail University, Kenitra, Morocco. He is a member of the Laboratory for Computer Science Research (LaRIT). His research focuses on the optimization of NP-Hard problems using metaheuristic approaches in artificial intelligence and big data for decision-making across various fields. This includes big data analytics, artificial intelligence, and information systems. He can be contacted at email: said.tkatek@uit.ac.ma.



Lahcen Idougli     a Ph.D. researcher at Ibn Tofail University's Faculty of Sciences in Kenitra, works within the Laboratory for Computer Science Research (LaRIT). His research encompasses computer networks, software engineering, artificial intelligence, and security. He can be contacted at email: lahcen.idougli@uit.ac.ma.