



Machine Learning Based Audio-Only football Goal Detection Under Clean and Degraded Acoustic Conditions

Iqra Andleeb¹, Muhammad Shakeel Tariq², Rana Muhammad Saleem¹, Irsa Shahid³, Sania Qayyum⁴ and Arbish Munawar¹

¹Department of Computer Science, University of agriculture Faisalabad

²School Education Department (SED), Government of Punjab, DG Khan

³Department of Computer Science Arid Agriculture Burewala

⁴Department of Computer Science University of Gloucestershire United Kingdom

ARTICLE INFO

Article History:

Received: July 30, 2026
 Revised: August 18, 2026
 Accepted: September 27, 2026
 Available Online: September 30, 2026

Keywords:

Goal Detection, Audio Only Event Detection, Mel-Spectrogram, Convolutional Neural Network (CNN), Audio Analysis, Audio Classification, Degraded Audio, Sports Analytics, Live Streaming

Classification Codes:

Funding:

This research received no specific grant from any funding agency in the public or not-for-profit sector.

ABSTRACT

Automatic football goal detection is impotent for real time events monitoring sports analysis system and intelligent broad casting system. And this detection works based on videos of live streaming, audio of commentary and crowd cheering, and many other technologies. As existing work relies on the visual information mainly or combined audio video modalities. However, during live streaming the video stream may become unavailable, delayed or corrupted, while audio remains accessible. This creates a need for alternative information source for reliable goal detection using audio only. This research works on football goal detection using only audio, particularly under imperfect conditions of audio. This research proposed CNN + Mel-spectrogram based model for football goal detection using only audio, with particular focus on detection under degraded acoustic conditions. The proposed approach uses audios of match, convert them into Mel- spectrogram, which further use as input to Convolutional neural network (CNN) to classify goal or no goal event. This study also investigates the acoustic patterns related to change in audio quality to support reliable goal detection. For examining the real-world streaming conditions, the system evaluated under both clean and artificially degraded audios by including different levels of noise. The no goal situation where strong crowd reaction arises is also evaluated as false goal detection. Performance is evaluated using accuracy, precision, recall and F1- score. The aim of research is to investigate the use of only audio for football goal detection and to determine the feasibility of CNN based models for acoustic representation while visual information is temporarily unavailable and audio quality also compromised.



© 2026 The authors published by JCIS. This is an Open Access Article under the Creative Common Attribution Non-Commercial 4.0

Corresponding Author's Email: rana.m.saleem@uaf.edu.pk

1. Introduction

In the era of artificial intelligence and technology, sports data analysis highly intermingled technological advancements. Technological advancements changed the way of decision making, analysis of sports events, and also effect the sports information and their impacts on peoples [1]. In the modern broadcasting system for sports large amount of audio and visual data is generated which is further used for automatic event detection, match summarization, highlight generation, and real time sports analysis. Among different sports, football has received great attention in this area because a football match contains sequence of meaningful events, as like goals, shots, fouls, penalties, and other actions that can be automatically identified. In football, goals are the particularly important event of the match, so goal detection is most significant and the rapid growth of digital streaming and live

sports broadcasting created and increasing demand for automatic systems, for the detection of important events in sports matches such as goal in football.

Traditional approaches for sports event detection heavily rely on visual information [2]. In the video-based systems a variety of cues are used such as player movements, replay sequences, ball trajectories, scoreboard changes, camera transitions, and changes in the visual characteristics of broadcast. Now with the emergence of deep learning and computer vision, video-based methods have achieved considerable success in the recognizing and localizing of sports actions. However, these approaches have complete dependency on the continuous availability and sufficient quality of the video streaming. But in practical broadcasting, the situation may change. As live streaming may affect by several technical factors, such as network instability, bandwidth limitations, transmission errors, packet loss, buffering, encoding problems, and temporary interruptions. So, the video of a part of a live broadcast may become unavailable, corrupted, delayed, or may be unusable for a short period of time while the audio components remain accessible. Figure.1 clearly represent the football event detection scenario.

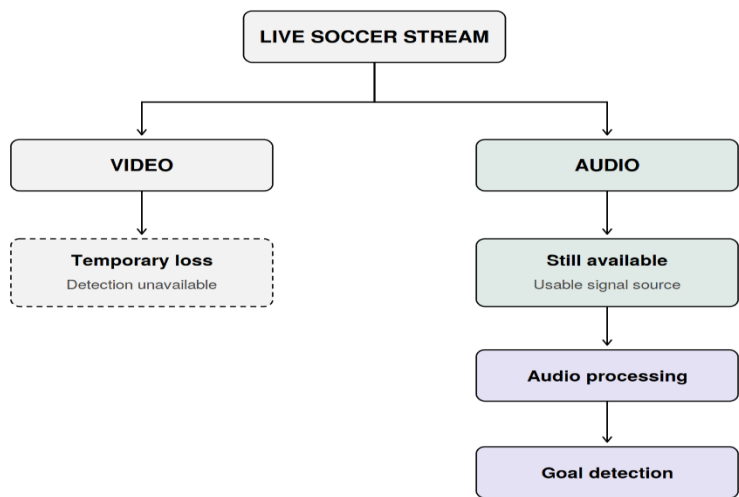


Figure. 1. The audio event detection scenarios.

Temporary visual missing creates a need for alternative information of events. Audio can be such source because it is generated continuously throughout the match, also can contain information about both match environment and audience response. The audio of football broadcasts may include commentator speech, referee whistles, player sounds, and most importantly crowd reactions. The event of goal may produce a sudden and significant change in acoustic environment as all the spectators react simultaneously. So, as a result, a rapid and noticeable increase in acoustic energy appears. These changes provide information for differentiating the situation of goal event or normal match. So, only audio-based systems can continue goal detection even when corresponding video information is temporarily unavailable.

Use of audio for event identification in sports is not a new research direction. However, goal detection from audio is more challenging than a simpler change detection in crowd activity. A football match can contain many movements that can generate a strong emotional response in crowds, like a shot that narrowly misses a goal, a penalty appeal etc. If a detection system associates high acoustic energy and large volume of crowd cheering with goal event, then it may cause false prediction. Thus, a reliable goal detection system should learn acoustic patterns that can sufficiently discriminate between an actual goal and other high intensity nongoal events.

Quality of available audio is another important challenge. In the ideal experimental environment clear audio recording signals are used with limited interference, but in real-world live broad casting environment audio signal may can contain background noise, crowd ambience, commentator speech, compression artifacts, transmission distortions, and variations in recording quality. Therefore, detection in clear audio environment may not be adequate for systems performance in practical scenario. For more realistic detection system performance should be considered by analyzing the system's behavior when audio quality is degraded. So, audio quality degraded by applying different methods including background noise can mask goal-related acoustic patterns, Reduced sampling rates can affect the representation of higher-frequency components, and variations in recording conditions can

change the acoustic characteristics of the same event. Evaluation of system under controlled degradation provides better understanding of system's behavior while quality of input signals change.

Recent advancement in deep learning provides effective methods for the automatic analysis of audio signals. In this research, for the representation of football audio as time frequency signal Mel-Spectrogram is used (shown in Figure.2), which further allows a Convolutional Neural Network (CNN) to learn acoustic patterns associated with goal and non-goal events. This mel-spectrogram captures the changes in acoustic energy and frequency with time, which are useful for the analysis of crowd reaction and other sounds during football matches.

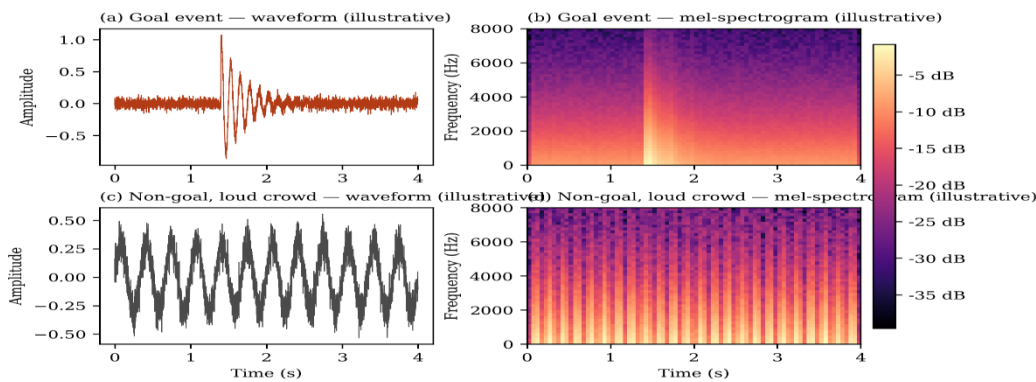


Figure. 2. Mel-spectrogram used in audio analysis.

The proposed framework uses Mel-spectrogram with Convolutional Neural Network (CNN) to perform binary classification as goal or no goal. For the investigation of the practical robustness, system tested under degraded acoustic conditions by introducing controlled noise and reduced audio quality. Performance of the system under both clean and degraded conditions is compared to determining reliability of the model that can detect goals when both visual information and high-quality audio may not be available.

2. Related Work

Automatic analysis of sports broadcasts an active area of research for more than two decades, with the work that research investigating how audio-visual information detects important events and is useful in summarization. In football match events often produce observable changes in both the visual scene and the acoustic environment, so it is particularly suitable for this type of analysis. Different visual changes and indicators are used for match events detection as like goal. A proposed approach with real time goal detection systems completely depends on visuals by using high frame cameras around goal area. Focus on the system on ball and on checking whether it crosses the goal line. Although the system demonstrates the feasibility of the real time event detection system, it completely depends on the visual data capture by cameras for decision making. so the system not able to detect any event if the video stream miss or its quality degraded[2]. Support vector machine is used for visual classification for event detection in football matches, where images-based data is used as input. It demonstrates how machine learning-based classification is useful for football event detection. But it not address the situation where the video component of the stream temporarily unavailable [3]. Spatial and temporal information is used with Convolutional neural networks and resulting representation combined through auto encoder. Their results demonstrated the potential of deep representations for distinguishing goal events from non-goal sequences [4]. A machine learning-based soccer video summarization framework which shot segmentation, replay detection, scoreboard detection, and excitement-event detection. Use different machine learning algorithms, like SVMs and neural networks., while k-means clustering used for excitement detection which highlights the importance of combining multiple indicators of excitement and match events [5]. As Goals detection from audio concerned, previous research proves that it can be detect by different perspectives using audios only including changes in commentator speech, crowd intensity, and other acoustic characteristics. Therefore, researchers have explored audio as an independent or important source of information for automatic football event detection. Such as study [6] explored that crowd response patterns can exploited for sports event detection. Similarly, another study demonstrate that variation in acoustic energy and repetition characteristics can help in the identification of important events in football match [7]. These studies provide foundational work for considering crowd response as important acoustic cue for goal detection.

Traditional audio systems often relied on handcrafted features such as MFCCs, energy, spectral descriptors, and other acoustic statistics. Recent developments in deep learning have significantly changed the way of audio classification. Now, modern approaches learn discriminative representations automatically from either waveforms or time-frequency representations. [8] proposed an approach based on Convolutional Recurrent Neural Network (CRNN) combining Convolutional layers with recurrent layers, for rare event detection based on sound. An end-to-end 1D CNN proposed system instead of relying on handcrafted frequency-domain features, learned filters directly from the raw waveform [9]. A researcher trained a deep sound representation using two million unlabeled videos and exploited the natural synchronization between visual and audio information. Support the idea that deep neural networks can learn useful acoustic representations without depending exclusively on manually engineered features [10]. Large-Scale Pretrained Audio Neural Networks based work investigated CNN-based architectures use to train on the large-scale dataset named AudioSet, and showed that pretrained audio networks can provide useful representations for multiple audio-related tasks [11]. [12] Audio Spectrogram Transformer (AST) treat audio spectrogram as input to transformer architecture, showed that spectrogram representations can be effectively used for modern deep-learning-based audio classification. Availability of labeled dataset for a reliable audio classification system is another major factor. [13] introduced UrbanSound8K and discussed the scarcity of large-scale, carefully annotated real-world audio datasets. Temporal action spotting with crowd reaction and commentator response is important for sufficient detection of an event, context aware loss function work is an important example for this phenomena [14]. Further enhance its functionality by using camera pointing for localizing player positioning for soccer action spotting [15]. Use of both audio and video stream for football action spotting and classification analyzed. Which is demonstrated by using experiments the SoccerNet benchmark that audio can provide useful complementary information to visual features [16].

Thus, previous research clearly indicates that audio visual information is a combination and separately used for event classification in football. Different machine learning algorithms applied to achieve desired result including SVM, CNNs etc. but a few studies work on only audio-based systems, and they also use simple traditional methods for data extraction from audios.

Research gap:

The review literature showed that automatic football event detection is an established research area. As discussed, earlier studies use different methods like SVM, use both audio and video, action spotting methods etc. some modern research use CNNs, RNNs and spectrogram-based architecture for audio classification. Despite all this progress, several main aspects remain significantly unaddressed. First, much of the recent literature is primarily video-based or multimodal. Second, earlier audio-based studies established the usefulness of crowd and acoustic information but generally relied on handcrafted acoustic characteristics or traditional machine-learning approaches. Third, the identification of binary goal versus non-goal classification using audio alone remains different from general excitement. Finally, the effect of degraded audio quality on audio-only football goal detection has not been sufficiently integrated into the problem formulation represented by the reviewed studies.

3. Methodology

Overview of proposed approach:

This research proposed an audio-only based machine learning framework (Figure.3) that is used for FIFA goal detection by using match audio. The main motivation is a situation in which the video components of match stream temporarily become unavailable, or damage while audio remains available but may degrade in quality. In such cases goal detection is potentially dependent on the available acoustic information, such as changes in acoustic patterns of crowd reaction and overall stadium noise. The proposed method consists of four major stages: 1) preparation of audio dataset 2) conversion of audio signal to Mel-spectrogram 3) CNN based binary classification of goal and no goal events 4) evaluation under controlled degraded audio. Complete workflow illustrated in fig. Evaluation of the model use 5-fold cross validation for out of fold performance estimation. Moreover, a degraded audio experiment performed by introducing controlled Gaussian noise at different signals to noise ratio (SNR).

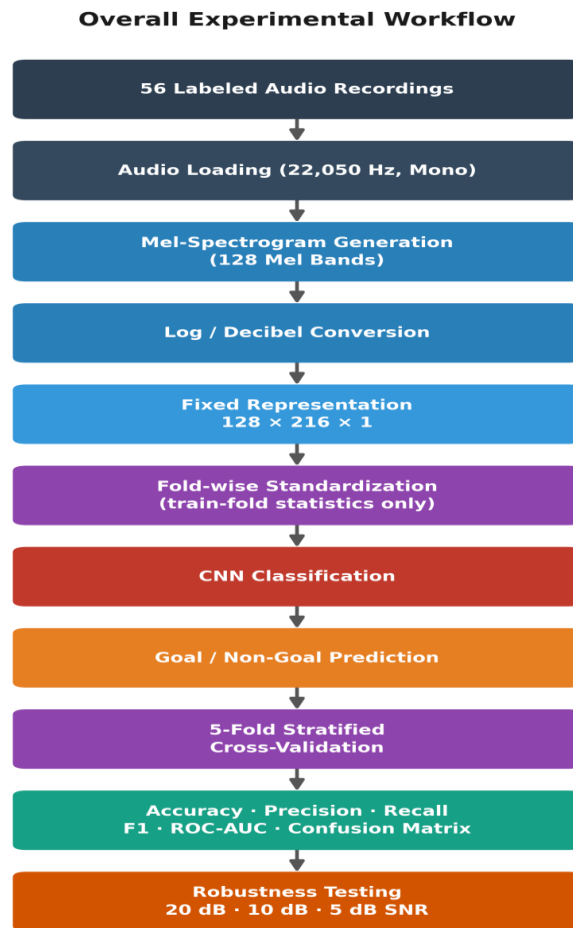


Figure. 3. Overview of framework.

Dataset preparation:

Dataset used in this research available on kaggle publically and used in competition “FIFA Goal & Crowd Pulse Detection” contains 56 audio files for training and 24 traditional testing audio files. 56 audios of training data use for model development and cross validation and 24 test audio samples used as unseen samples for predictions. The training data is divided into two target class goals and no goal. For binary classification label target variable is goal is used and a value of 1 represents goal event and value of 0 represents no goal event. For training balanced set is used 28 are goal and 28 are no goal recordings. Metadata of dataset contains an Id field corresponding to each file and is goal field representing target class.

Preprocessing:

The audio recordings are loaded by using the Librosa library. Audio was converted to mono during loading and set sampling rate used for the CNN input representation 22,050 Hz. Each audio recording transformed from a one-dimensional waveform to a two-dimensional time-frequency representation using a Mel-spectrogram. For audio classification Mel-spectrograms are useful because they represent the distribution of acoustic energy across frequency and time. For each audio recording, a Mel-spectrogram with 128 Mel frequency bands generated. The resulting power spectrogram was subsequently converted into the decibel scale using logarithmic power conversion. The resulting representations were arranged into a fixed input size of 128×216 . A single channel dimension was then added, producing the final CNN input shape: $128 \times 216 \times 1$

Thus, each audio recording was represented as an image-like time-frequency matrix rather than being provided directly to the neural network as a raw waveform.

Mel-Spectrogram Representation:

For input audio signals spectrogram can be represented as

$$M(m, t) = \sum_k |X(k, t)|^2 H_m(k)$$

Where $X(k, t)$ represents the short Fourier transform of the audio signals, and $H_m(k)$ is the m -th Mel filter, m use for the representation of the Mel-frequency band, t represents the time frame, k use for the frequency bin. By implementation of this with audio sampling rate of 22,050 Hz 128 Mel -frequency bands generated. The power values were converted to the decibel scale using:

$$M_{db}(m, t) = 10 \log_{10} \left(\frac{M(m, t)}{M_{max}} \right)$$

Where $M(m, t)$ is the Mel-spectrogram power at Mel band m and time frame t , while $M(max)$ represents the maximum reference power. Fig.4 represents the Mel-spectrogram of an audio file

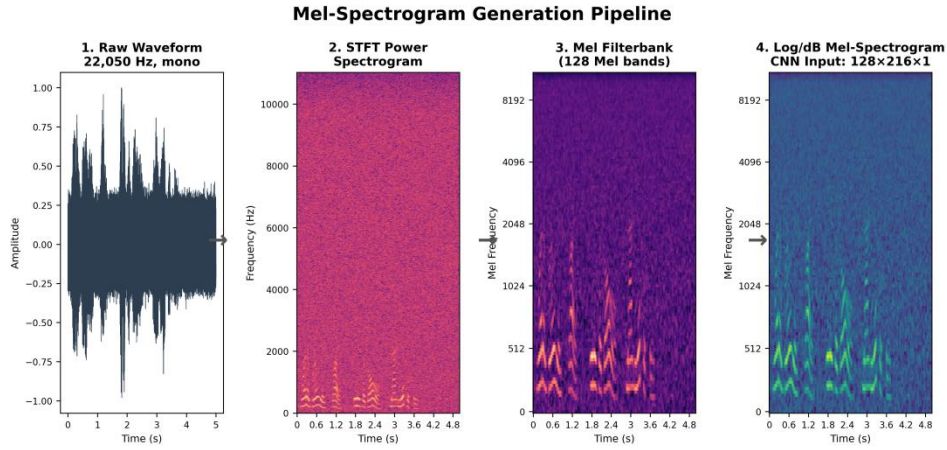


Figure. 4. Mel- spectrogram representation of audio file.

Resulting Mel-spectrogram stored as Numpy array and shapes of training samples and target are following:

$$X \in \mathbb{R}^{56 \times 128 \times T}$$

And

$$y \in \{0,1\}^{56}$$

Mel-spectrogram of a sample shown in fig.

Data Standardization:

Standardization was applied during cross-validation, to improve the numerical stability of CNN training. The main thing is that only training portion of that fold is used to calculate the mean and standard deviation of that fold. Each training and validation input was transformed using:

$$X' = \frac{X - \mu_{train}}{\sigma_{train} - \epsilon}$$

Here μ_{train} and σ_{train} denote the meaning and standard deviation calculated from the training fold, respectively. And ϵ is used to prevent division by zero.

For training:

$$X'_{train} = \frac{X_{train} - \mu_{train}}{\sigma_{train} - \epsilon}$$

For validation:

$$X'_{val} = \frac{X_{val} - \mu_{train}}{\sigma_{train} - \epsilon}$$

CNN Architecture:

To classify Mel-spectrogram into goal and no goal classes a Convolutional Neural Network (CNN) is used. Convolutional layer can learn local time-frequency related patterns associated with different acoustic events from image like two-dimensional structure of Mel-spectrogram. CNN architecture (shown in Figure.5) consists of two convolutional blocks followed by global feature

aggregation and fully connected classification layers. The first convolutional layer with ReLU activation, contains 16 filters of size 3×3 and followed by a 2×2 max-pooling layer. The second convolutional layer contains 32 filters of size 3×3 , which further followed by a 2×2 max-pooling layer. Then Global Average Pooling used to reduce the spatial feature maps into a compact feature representation. Which is followed by a fully connected layer containing 32 neurons with ReLU activation and a dropout rate of 0.5 also applied. At the end, a single neuron with a sigmoid activation function produces the probability of the input belonging to the goal class, represented in table 1.

Table. 1. CNN architecture layers.

Layer	Configuration
Input	$128 \times 216 \times 1$
Convolutional Layer 1	16 filters, 3×3 , ReLU
Max Pooling 1	2×2
Convolutional Layer 2	32 filters, 3×3 , ReLU
Max Pooling 2	2×2
Global Average Pooling	Feature aggregation
Dense Layer	32 neurons, ReLU
Dropout	0.5
Output Layer	1 neuron, Sigmoid

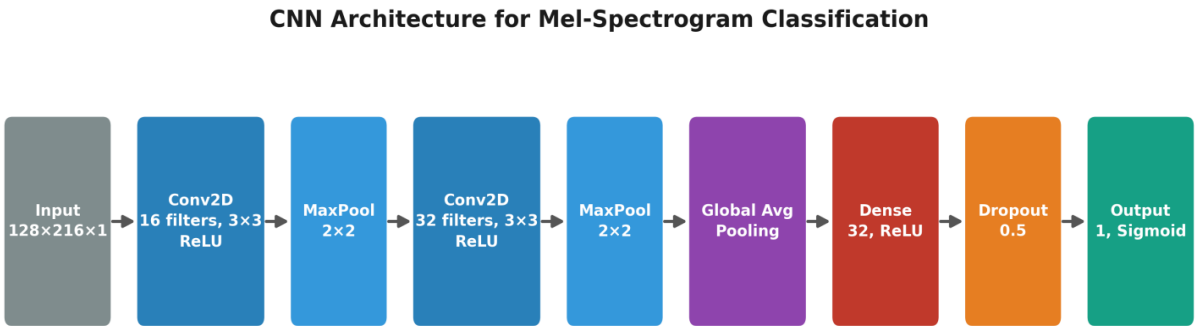


Figure. 5. CNN architecture.

Model Training:

As the task is binary classification, Adam optimizer with Binary-cross-entropy is used to train the CNN model. The binary cross-entropy loss is defined as:

$$L_{BCE} = -[y\log(p) + (1 - y)\log(1 - p)]$$

where y denoted actual label and p is the predicted probability of goal class.

With maximum 100 training epochs and a batch size of 8 for each is used. To reduce unnecessary training when validation loss was stopped improving, early stopping employment. For binary prediction, a probability threshold of 0.50.5 was used:

$$y^{\wedge} = \begin{cases} 1, p \geq 0.5 \\ 0, p < 0.5 \end{cases}$$

Where 1 is goal and 0 is no goal state.

Five-Fold Stratified Cross-Validation:

Five-fold stratified cross-validation was used for model evaluation, as the data set is relatively small. Stratification preserves the proportion of goal and non-goal samples across the folds. The dataset was divided into five folds; in each iteration four folds were used for training and one for validation. Process repeat until every sample validated.

In each fold, new CNN model initialized and trained independently. Standardization parameters calculated from training portions only and applied to both the training and validation portions. The predictions from all folders combined to get out of fold validation and it is further used to get overall performance and confusion matrix. Figure.6 explain each fold.

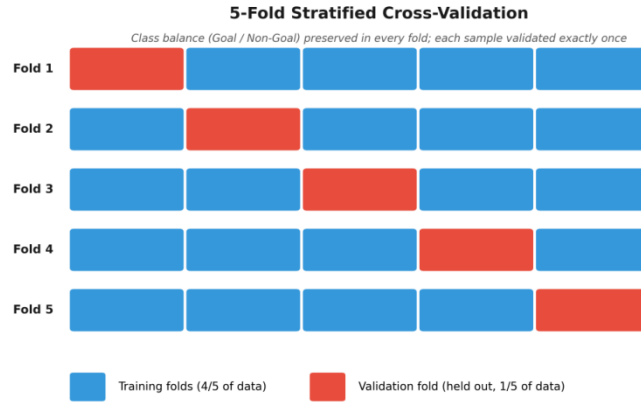


Figure. 6. Five-fold cross validation

Evaluation Metrics:

The performance of the classifier was evaluated using Accuracy, Precision, Recall, F1-score, and ROC-AUC.

Accuracy measures the proportion of all samples that were correctly classified:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision measures the proportion of predicted goal events that were goals:

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall measures the proportion of actual goal events that were correctly detected:

$$\text{Recall} = \frac{TP}{TP + FN}$$

F1-score is the harmonic means of precision and recall:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

It can also be expressed directly in terms of the confusion-matrix components such as:

$$F1 = \frac{2TP}{2TP + FP + FN}$$

Where TP represents True Positives, TN for True Negatives, FP for False Positives and FN for False Negatives.

ROC-AUC was additionally used to evaluate the model's ability to distinguish between goal and non-goal samples across different classification thresholds. The ROC curve represents the relationship between the True Positive Rate (TPR) and False Positive Rate (FPR):

$$\text{TPR} = \frac{TP}{TP + FN}$$

$$\text{FPR} = \frac{FP}{FP + TN}$$

The Area Under the ROC Curve (ROC-AUC) summarizes the discriminative performance of the classifier across all possible classification thresholds.

Confusion matrices were also generated to examine the distribution of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

Final CNN Model:

Finally, CNN model was trained using the complete set of 56 labeled training samples and epochs, batch size, and loss function same as cross validation. The final model achieved the accuracy of 98%. This results from training dataset, which distinguishes it from the cross-validation results. The trained model further uses to make predictions on the 24 unseen data samples and predict this data successfully as goal and no goal classes.

Degraded-Audio Robustness Experiment:

Then another main experiment was performed to test the hypothesis that model can predict the goal event even in the situation when audio quality is degraded. controlled Gaussian noise was added to the original audio waveform, instead of adding direct noise. Which is more realistic simulation of acoustic degradation as the signal is corrupted before the generation of time frequency representation. Three signals to noise ratio (SNR) conditions were considered, 20 dB SNR for relatively mild degradation, 10 dB SNR for moderate degradation and 5 dB SNR for strong degradation samples shown in Figure.7.

Degraded-Audio Robustness Experiment: Gaussian Noise Injection at Multiple SNR Levels

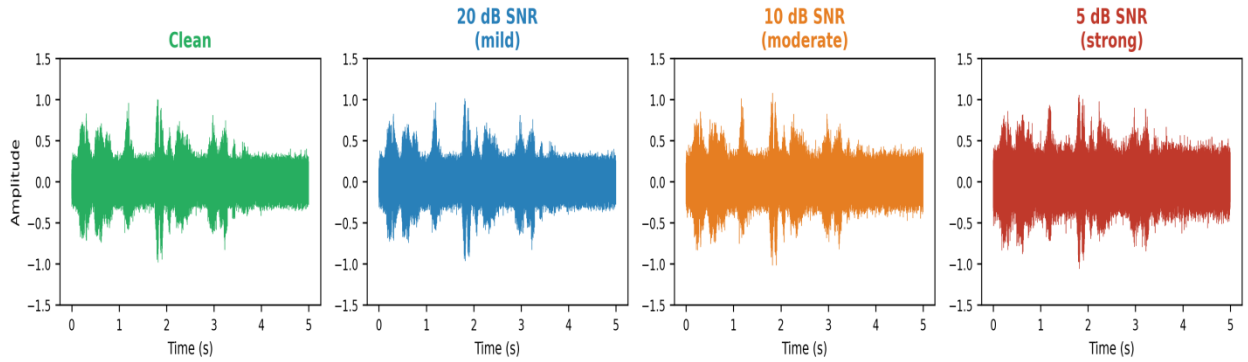


Figure. 7. Audio degradation.

For a given audio signal x , the signal power was calculated as:

$$P_s = \frac{1}{N} \sum_{i=1}^n x_i^2$$

Where N is total number of audio samples and x is the amplitude of audio.

For a target signal-to-noise ratio, the required noise power was calculated as:

$$P_n = \frac{P_s}{10^{SNR \text{ db}/10}}$$

The standard deviation of the Gaussian noise obtained from the noise power:

$$\sigma_n = \sqrt{P_n}$$

And Gaussian noise signal generated as:

$$n_i \sim N(0, \sigma_n^2)$$

and added to the original waveform:

$$x_{\text{noisy}} = x + n$$

where x_{noisy} used to represent the degraded audio waveform and n represents the generated Gaussian noise.

This waveform converts into Mel-spectrogram using same procedure and resulting into 128*126*1 format CNN input. The degraded dataset contains same 56 labeled training audio files and CNN were independently evaluated using fivefold cross validation for each of the three SNR levels. Accuracy, Precision, Recall, F1-score, ROC-AUC, and confusion matrices were calculated for each degradation condition also.

This experiment was implemented to check the condition where audio for goal detection is degraded in quality due to external noise.

4. Results and Discussion

Experimental Evaluation:

The proposed approach of CNN based audio classification framework evaluated training by using fivefold cross validation. For evaluation accuracy, Precision, Recall, F1-score, and ROC-AUC are used. With the analysis of clean audio, the robustness of model evaluated under three controlled noise conditions.

The experiment on clean audios performs consistently high across all five folds. The mean and standard deviation of all samples are 1.0000 ± 0.0000 . results of out of fold confusion matrix show that all samples classified well.

Clean-Audio Classification Performance:

Five-fold cross-validation performance obtained using the clean audio represented in table.2

Table. 2. Clean-Audio Classification Performance.

Metric	Mean $\pm$ Std
Accuracy	1.0000 ± 0.0000
Precision	1.0000 ± 0.0000
Recall	1.0000 ± 0.0000
F1-Score	1.0000 ± 0.0000
ROC-AUC	1.0000 ± 0.0000

The out-of-fold confusion matrix contains 28 true negatives, 28 true positives, and no false positives or false negatives shown in table.3

Table. 3. Clean-Audio Classification Performance Confusion Matrix.

Actually / Predicted	Non-Goal	Goal
Non-Goal	28	0
Goal	0	28

These results indicate that CNN model can clearly classify two separate classes on the available training data. However, that dataset contains 56 training data recordings so it can be interpreted in context of the limited sample size.

Final Model Training:

Following cross validation, CNN model was trained using labeled data and model achieve 98% accuracy at the end of training. This value is a training-set result and that's why it is different from the five-fold out-of-fold evaluation mentioned above. This difference is important because the final model was trained using all available labeled samples. Confusion matrix shown in Figure.8.

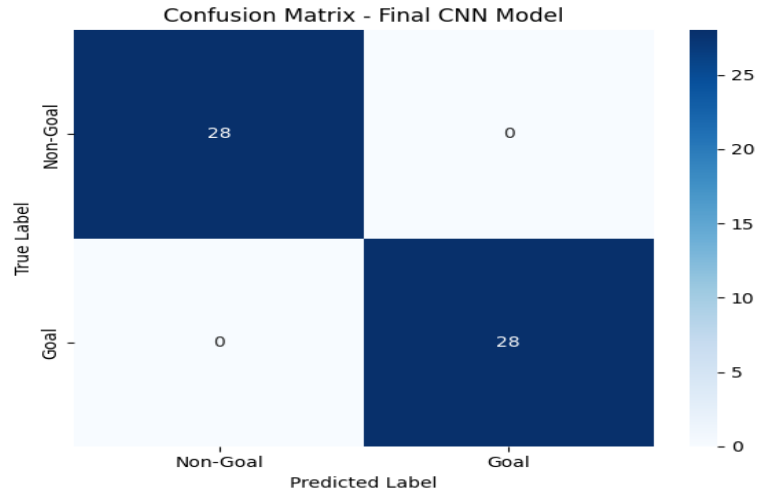


Figure. 8. Final model training confusion matrix.

Model also test on 24 samples of unseen data, and it predict 12 as goal and 12 as no goal events, these results indicate that model successfully learn the training data pattern.

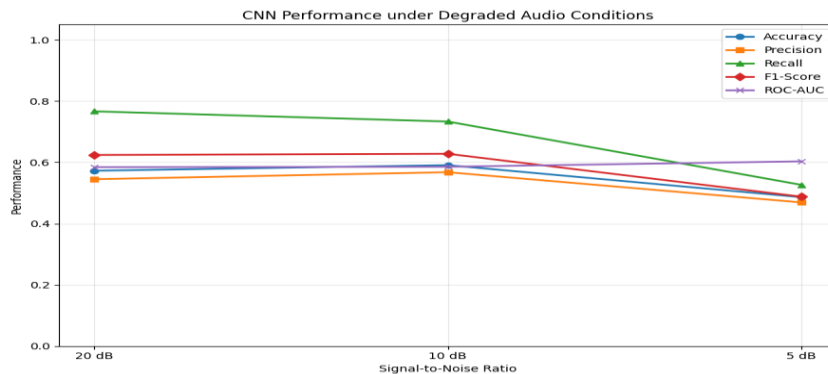
Robustness to Degraded Audio and comparison with clean audios:

The proposed approach is evaluated by adding controlled Gaussian noise to the original audio waveform 20 dB, 10 dB, and 5 dB SNR. The same 56 labeled training dataset as used in clean audio experiment degraded and used. At 20 dB SNR, model achieved a mean accuracy of 0.5727 ± 0.0610 with a mean F1-score of 0.6242 ± 0.1267 . And at 10 dB SNR the mean accuracy was 0.5909 ± 0.0909 and the mean F1-score was 0.6280 ± 0.1317 . Finally, at 5 dB SNR the performance decreased more, with a mean accuracy of 0.4864 ± 0.1469 and a mean F1-score of 0.4879 ± 0.1936 . The results of 20db and 10 db are close to the 10 db noise addition produce slighter higher mean accuracy and F1 score. The overall trend indicates clear degradation in accuracy with degraded audio datasets. These results show that acoustic noise has strong influence on CNN classifiers and acoustic patterns learned are sensitive to change in input signals. Table .4 summarizes this result of comparison of all the conditions with clean audios.

Table. 4. comparison of all the conditions with clean audios.

Audio Condition	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Clean	1.0000	1.0000	1.0000	1.0000	1.0000
20 dB SNR	0.5727	0.5450	0.7667	0.6242	0.5844
10 dB SNR	0.5909	0.5679	0.7333	0.6280	0.5856
5 dB SNR	0.4864	0.4692	0.5267	0.4879	0.6033

The results clearly show a considerable difference between the clean and degraded acoustic conditions. On the clean-audio cross-validation produced perfect scores on the available dataset, while the introduction of acoustic noise considerably reduced



the classification performance of same model. Figure.9 And Figure.10 show the noise added audios accuracy and ROC-AUC in following under different SNR conditions.

Figure. 9. Evaluation matrix with SNR audios.

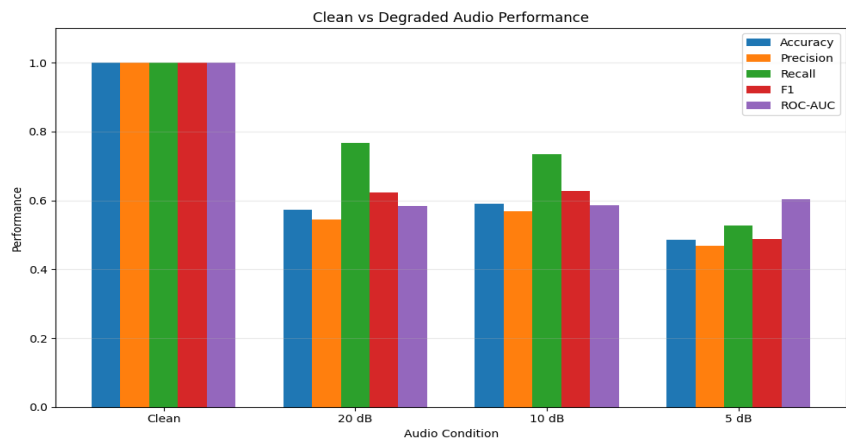


Figure. 20. Evaluation matrix representation with SNR audios.

Confusion-Matrix Analysis under Noise:

The out-of-fold confusion matrices provide additional insight into the effect of noise. After adding noise of three different level classification performance of model shown in table

For the 20 dB SNR condition, the confusion matrix shown in Figure.11.:

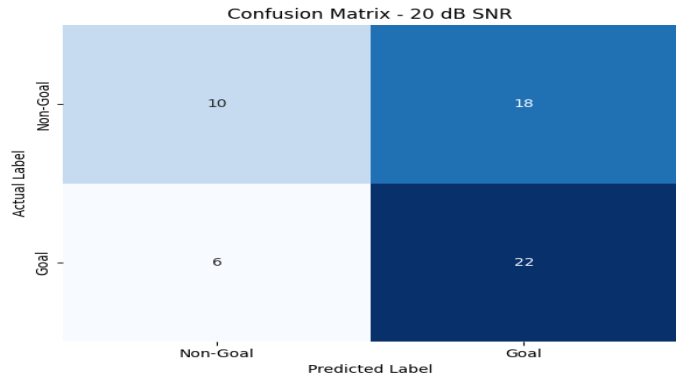


Figure. 31. 20 dB SNR condition.

The model correctly identified 22 of the 28 goal recordings but incorrectly classified 18 non-goal recordings as goals.

For the 10 dB SNR condition results shown in Figure.12: At this level of noise, 21 goal events detected accurately, and also for 12 no goal event model give accurate result.

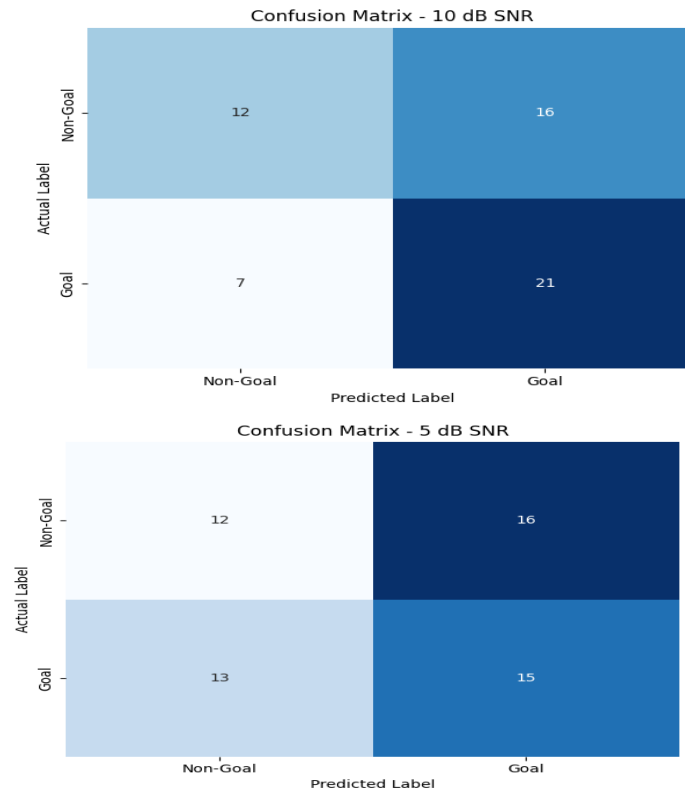


Figure. 42. 10 dB SNR condition.

For the 5 dB SNR condition results in confusion matrix Figure.13:

Figure. 53. 5 dB SNR condition.

The 5 dB condition predict 15 is goal and 12 no goal accurately, so produces a more balanced distribution of false positives and false negatives, indicate that more noise makes it more difficult for the model to differentiate goal-related acoustic patterns from non-goal audio.

5. Conclusion

This study investigated the feasibility of detection of football goal events from commentary and crowd audio alone, using a CNN model working with Mel-spectrogram representations, with the aim of providing a detection pathway for scenarios where the video stream is delayed, corrupted, or temporarily unavailable.

Under five-fold stratified cross-validation on clean audio, the proposed CNN achieved perfect classification (Accuracy = 1.00, F1 = 1.00) on the 56-clip labeled dataset. When the same models trained only on clean audio were evaluated on artificially degraded versions of their held-out clips, performance remained perfect down to 20 dB SNR, degraded to 91% accuracy at 10 dB SNR, and collapsed to 50% accuracy at 5 dB SNR and below. Critically, this collapse was systematic rather than random like at high noise levels the model predicted "goal" for every clip, indicating that its decision boundary is substantially driven by overall acoustic energy, and that injected noise raises the perceived loudness of every clip toward the "goal-like" region of that boundary.

These results support a qualified answer to the research question: a CNN operating on Mel-spectrograms is a feasible approach for audio-only goal detection under clean-to-moderately-degraded conditions (≥ 10 dB SNR), but its reliability cannot currently be assumed under heavier acoustic degradation, where it fails in a specific, energy-driven way rather than degrading gracefully. However, the have some limitations, which should be consider in future extension of this work.

6. Future Work

After this research several directions explore to improve football event detection based on audio data only. First is the main problem of limitation of dataset. Current work is based on only 56 training dataset recordings limiting the ability to evaluate generalization across different matches and recording conditions. So future datasets should include large number of events and multiple match recordings and a greater number of files.

Second, more advanced audio features should be investigated. Mel-spectrograms, features such as MFCCs, chroma features, spectral contrast, zero-crossing rate should be used for audio analysis, it gives more realistic scenario of match like crowd reaction intensity etc.

Third, more advanced deep learning architecture should be investigated. Fourth, real world audio degradation should be studied like crowd noise, commentary interference, compression effect etc. Fifth, instead of offline segments classification, real time event detection should be performed.

7. Acknowledgements

Author is grateful to the UAF Sub Campus Burewala.

References

1. Vidal F, Nicolas C, Bahaeddine E, Fakir E, Billingham J. Automatic event detection in football using tracking data. *Sports Engineering*. 2022;25(1):1–15. doi:10.1007/s12283-022-00381-6
2. Orazio TD, Leo M, Spagnolo P, Nitti M, Mosca N, Distante A. A visual system for real time detection of goal events during soccer matches. Vol. 113. 2009;113:622–32. doi:10.1016/j.cviu.2008.01.010
3. Ancona N, Cicirelli G, Branca A, Distante A, Elaborazione I, Amendola V. Goal detection in football by using Support Vector Machines for classification 1. 2001.
4. Tsagkatakis G, Jaber M, Tsakalides P. Goal !! Event detection in sports video.
5. Zawbaa HM, El-bendary N. Machine Learning-Based Soccer Video Summarization System. 19–20.
6. Workshop PR, July J, Baillie M, Jose JM. An Audio-based Sports Video Segmentation and Event Detection Algorithm. 2004;(July).
7. Duxans H, Anguera X, Conejero D. Audio based Soccer Game Summarization.
8. Scenes A. CONVOLUTIONAL RECURRENT NEURAL NETWORKS FOR RARE SOUND EVENT DETECTION Emre C , akir and Tuomas Virtanen Tampere University of Technology Finland. 2017;(November):1–5.
9. Abdoli S, Cardinal P, Koerich AL. End-to-End Environmental Sound Classification using a 1D Convolutional Neural Network. 2019;1–24.
10. Torralba A. SoundNet : Learning Sound Representations from Unlabeled Video arXiv : 1610 . 09001v1 [cs . CV] 27 Oct 2016. 2016;(Nips).
11. Kong Q, Member S, Cao Y, Iqbal T. PANNs : Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition. (1).
12. Gong Y, Chung Y an, Glass J. AST : Audio Spectrogram Transformer.
13. Salamon J, Jacoby C, Bello JP. A Dataset and Taxonomy for Urban Sound Research. (1):1041–4.
14. Cioppa A, Deli A, Giancola S, Ghanem B, Droogenbroeck M Van, Moeslund TB. A Context-Aware Loss Function for Action Spotting in Soccer Videos.
15. Cioppa A, Deli A, Giancola S, Barnich O, Ghanem B, Droogenbroeck M Van. Camera Calibration and Player Localization in SoccerNet-v2 and Investigation of their Representations for Action Spotting.
16. Vanderplaetse B. Improved Soccer Action Spotting using both Audio and Video Streams. (iii).