Penerbit UTHM

JOURNAL OF SOFT COMPUTING AND DATA MINING
e-ISSN: 2716-621X

JSCDM

Vol. 6 No. 3 (2025) 20-32
https://publisher.uthm.edu.my/ojs/index.php/jscdm

Detection and Classification of Emergency Vehicles from Audio and Video Inputs using Deep Learning Techniques

Ephzibah E. P[1], Mareeswari V[1]*, Ashraf Osman Ibrahim[2]*, Nishanth Samson[1], Nasikethan R[1], Rozaida Ghazali[3]

[1] *School of Computer Science Engineering and Information Systems (SCORE),*
Vellore Institute of Technology (VIT), Vellore, INDIA
[2] *Department of Computing,*
Universiti Teknologi PETRONAS, Seri Iskandar, MALAYSIA
[3] *Faculty of Computer Science and Information Technology,*
Universiti Tun Hussein Onn Malaysia, Parit Raja, MALAYSIA

*Corresponding Author: ashrafosman2@gmail.com

DOI: https://doi.org/10.30880/jscdm.2025.06.03.002

Article Info

Received: 6 March 2025
Accepted: 11 July 2025
Available online: 30 December 2025

Keywords

Emergency vehicle detection, image processing, audio classification, YOLO, object detection, LSTM, explainable AI

Abstract

With the rapid advancement of autonomous vehicles, ensuring road safety is one of the biggest concerns of the automotive industry. One critical aspect of safety is the accurate and timely detection of emergency vehicles such as ambulances, and fire trucks and promptly switching lanes to ensure smooth passage. This paper proposed an efficient and straightforward method to locate and label emergency vehicles with the help of the most updated deep learning algorithms known as YOLOv8 and long short-term memory (LSTM). The accuracy and efficiency of emergency vehicle detection in terms of perfecting the models is the focus. Data augmentation methods are carried out to enhance day-night and low visibility performance of the model on the dataset. The system is capable of identifying and classifying emergency vehicles on the basis of audio and video data using several signal/image processing methods and accomplished with the means of explainable artificial intelligence (XAI) mechanisms providing detailed information. The given system can be used with self-driving and human-driven vehicles which can be fitted to advanced driver assistance systems (ADAS). It is found that the accuracy and performance of emergency vehicle detection have improved significantly with the 96.6% accuracy rate that will ameliorate the interaction safety of autonomous vehicles and emergency vehicles.

1. Introduction

As the automotive industry moves towards developing self-driving vehicles, one major area of concern is road safety. A major problem lies in the safe coexistence of ambulatory and autonomous vehicles and human-driven cars. Their detection and timely response is essential for safety. This problem is critical in the case of emergency vehicles such as ambulances and fire trucks where detection and immediate response capabilities for human-driven vehicles determines response efficiency. This research is critical as it proposes a new framework for detection and classification of emergency vehicles using powerful deep learning algorithms.

Autonomous vehicles also aid in improving the efficiency and economics of the transportation system. With the benefits identified, there also exist distinct hurdles such as the need for integrating unmanned vehicles into the paths of ambulatory vehicles. These vehicles need unobstructed pathways to their destinations and minimal hindrance from other vehicles in the queue. This implies other vehicles need to make way by swiftly and optimally in the presence of emergency vehicles on road. This paper focuses on enhancing the coexistence of autonomous and emergency vehicles by developing a robust, real-time solution for emergency vehicle detection and classification, which is a fundamental aspect of achieving harmonious interactions between these two vehicle types.

In this paper, we propose an effective yet straightforward solution for recognizing and categorizing emergency vehicles. Some of the focal points of this system include the deep learning algorithms YOLOv8 and LSTM networks. Advanced techniques of data augmentation and LSTM model fine-tuning will be done with the hope of achieving greater efficiency and accuracy in the emergency vehicle detection. The system utilizes data obtained from the vehicle's visual (camera) and auditory (microphone array) sensors to perform real-time, proactive scanning and detection of emergency vehicles. The analysis of data streams in the form of sound and sight is done. This results in accurate detection in complex scenarios. For instance, a camera-only system would fail to detect an emergency vehicle if the vehicle is obscured from the camera's view due to other traffic. Audio-only systems, on the other hand, are overly sensitive and prone to false detections. These shortcomings are resolved with the approach described in this paper.

The other objective on which this paper also focus is on enhancing model accuracy in complex processes such as day-night transitions and low visibility. The input systems of video and audio, along with other image and signal processing instruments make it possible to help recognize and classify emergency vehicles fast and correctly. The use of XAI brings more insight to the emergency vehicle recognition systems of EVAI among others. XAI explains how the decisions of the model were reached enabling other stakeholders to understand and trust the classification that has been made by the system, hence making working with the system safer.

Innovation through the creation of driverless cars in the automobile industry provides a radical way to safer and more viable transportation systems. Although, a major concern during the development of this is proper recognition followed by response to emergency vehicles. In spite of the great artificial intelligence and sensor technology achievements, self-driving cars can not accurately identify emergency vehicles, especially during low-light and foul weather conditions. It has caused several frightening incidents. These gaps in the identification of the distinctive communication of emergency vehicles (including blinking light pulses and the sirens) have been demonstrated through a numerically large number of accidents, and close calls by the current Advanced Driver Assistance Systems (ADAS) which have been addressed by a large number of collisions and close calls.

Hazardous situations on roads around the world have resulted from the inability to quickly identify these important indicators. Autonomous vehicles have frequently failed to yield to emergency vehicles, endangering lives and delaying prompt emergency responses, whether due to poor visibility, erratic traffic patterns, or intrinsic difficulties with sensor integration, faltered in yielding to emergency vehicles, thus putting lives in danger, and impeding swift emergency responses [1].

In 2022, 224 people died in emergency vehicle crashes, with 50% of deaths being non-emergency vehicle occupants. The majority of these deaths occurred in multi-vehicle crashes, with police vehicles, ambulances, and fire trucks being the most fatalities [23] Also, accident involving autonomous vehicles crashing into emergency vehicles due to the failure of ADAS systems onboard the autonomous vehicles. Most of these incidents involved cars crashing into emergency vehicles parked on the side of the road due to the inability of the system to detect flashing lights or lens flares at night. Addressing this gap in autonomous vehicle technology is essential for fostering a future prioritizing road safety and coexistence between autonomous and emergency vehicles.

Recent studies highlight advancements in emergency vehicle (EV) detection and navigation systems, emphasizing AI, IoT, and advanced computing. Mahmud et al. [2] and Rosayyan et al. [3] proposed IoT-based EVP systems leveraging edge computing and minimal sensors to improve EV transit and reduce congestion. Patel et al. [5] and Dodia et al. [10] integrated neural network siren classifiers and GPS-based traffic light systems for emergency lane creation, demonstrating significant reductions in waiting times. Cantarini et al. [9], Lisov et al. [18], and Choudhury & Nandi [19] explored acoustic EV detection using CNNs and LSTM models, addressing noise challenges and achieving high accuracy. Wang et al. [13] and Carvalho Barbosa et al. [14] introduced lightweight AI models for vehicle detection, showcasing efficiency on resource-constrained platforms. XAI's role in transparency and decision-making was emphasized by Nwakanma et al. [6], Ghaffarian et al. [7], and Omeiza et al. [8] for autonomous and connected vehicles. Kolekar et al. [15] and Nayak [16] developed vision-based frameworks for unstructured traffic, integrating explainability and precision. Agnew et al. [12] and Uma et al. [20] highlighted real-time navigation and traffic adaptation with advanced RCNNs and ECU-based systems. Collectively, these innovations address traffic management, road safety, and emergency response, enhancing EV detection, navigation, and overall urban traffic efficiency.

The system [2] lacks robustness in handling complex traffic scenarios where nearby vehicles can block the sensors leading to inconsistent detections. Further exploration is needed to enhance the system's adaptability to diverse traffic conditions and emergencies. Although the integration of edge computing and IoT sensors shows promise [3], there is a need for further research to address potential scalability issues and to explore the impact on system performance under various real-world conditions. While advancements in automotive technology are discussed [12], there is a gap in exploring the practical implementations of Electronic Control Units for real-time analysis of camera images and their effectiveness in detecting emergency vehicles in various road and weather conditions. Even the research works [16] discuss enhancing autonomous vehicle responses during emergencies; there is a gap in exploring the scalability and real-time responsiveness of the proposed model under various emergency scenarios.

The rest of this paper is organized as follows: Section 2 provides the materials and methods, Section 3 shows the results achieved beside the discussion, and Section 4 concludes the paper and outlines future work.

2. Material and Methods

The Emergency Vehicle Detection System stands as a cutting-edge solution to enhance public safety by providing real-time detection and classification of emergency vehicles. By seamlessly integrating visual and auditory information, the system significantly improves situational awareness for emergency services and traffic management. At the heart of this architecture lie two critical components: the YOLOv8 model for visual analysis and the LSTM model for audio analysis.

The system architecture for emergency vehicle detection is engineered to harness the capabilities of YOLOv8 and LSTM architectures within a highly efficient multi-threaded framework, thereby ensuring rapid and precise real-time predictions. A multi-threaded strategy better utilizes the available resources and increases the processing speed because the calculation problem can be allocated among threads that are executed concurrently. Besides accelerating the computations, this methodology allows adding separate modules to the design more easily. The performance of the system is further enhanced because of the individual functioning of each individual module in the allotted thread where performance and accuracy are the topmost priorities. This well-thought multithreading paradigm provides the architecture with the perfect compromise between computation speed and prediction accuracy, therefore being particularly desirable to be used in a time-sensitive setting such as an emergency. The interaction of these modules within the overall system architecture is shown in Fig.1, highlighting the proposed system architecture by using a multithreaded approach.

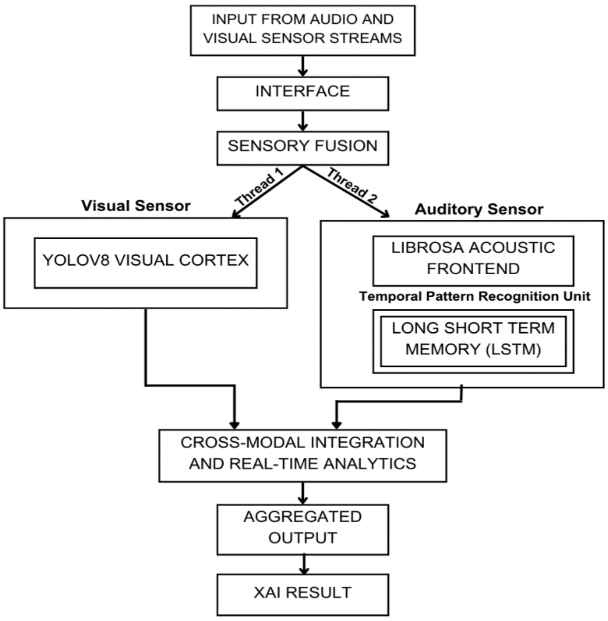
INPUT FROM AUDIO AND VISUAL SENSOR STREAMS
INTERFACE
SENSORY FUSION
Thread 1
Thread 2
Visual Sensor
Auditory Sensor
YOLOV8 VISUAL CORTEX
LIBROSA ACOUSTIC FRONTEND
Temporal Pattern Recognition Unit
LONG SHORT TERM MEMORY (LSTM)
CROSS-MODAL INTEGRATION AND REAL-TIME ANALYTICS
AGGREGATED OUTPUT
XAI RESULT

Fig. 1 *Proposed system architecture*

Video formats supported by the system include.avi,.mp4, and any other compatible format. The input is then intelligently split into two separate sensory streams through a process called "Sensory Fusion," which takes advantage of multi-modal sensory processing. To complete its task of detecting emergency vehicles, the visual sensor, which is a video component, takes two steps. It is first fed into the YOLOv8 model, which is a state-of-the-art object detection model for real-time object recognition. YOLOv8, which was specifically trained on a large dataset of emergency vehicles, demonstrates exceptional ability to recognize objects of interest in picture and video frames.

The Visual Cortex is dominated by the object detection model, YOLOv8. The Visual Cortex creates bounding boxes that contain the exact locations of identified emergency vehicles by sequentially processing each frame in the video stream and dynamically updating its "Visual Focus." Fundamentally, YOLOv8 uses deep Convolutional Neural Networks (DCNN) to function in a variety of difficult and complex situations, such as traffic situations, different lighting conditions, and weather patterns. YOLOv8 creates bounding boxes that contain the locations of the identified emergency vehicles by sequentially processing each frame in the video stream. The presence and exact location of emergency vehicles within the video feed are indicated visually by these bounding boxes.

At the same time, the audio clip is thoroughly examined. The librosa module, a powerful Python library for music and audio analysis, is used to extract relevant audio features that could suggest the presence of emergency vehicles. The Auditory Cortex is accessed through the Librosa Acoustic Frontend. In order to capture the subtleties of audio signals, such as frequency, intensity, and temporal characteristics, it performs sophisticated audio feature extraction. The "Temporal Pattern Recognition Unit" receives the processed audio features and analyzes them.

The video formats compatible with the system include.avi, and.mp4 among any other format. It is then smartly divided into two streams of sensory information via a phenomenon referred to as Sensory Fusion that exploits multi-modal sensory processing. In order to accomplish its mission of identification of emergency vehicles, the visual sensor which is a component of the video system carries out two steps. First, it is processed on the object detection state of the art real-time object recognition model, the YOLOv8. YOLOv8, specially trained on a large dataset of emergency vehicles, shows the outstanding object detection skill on picture and video frames with the objects in the frames of interest.

The object detection model, YOLOv8 prevails in the Visual Cortex. Visual Cortex uses their locations along with the date of identified emergency vehicles by accessing information frame by frame, in the video stream and dynamically changing its Visual Focus. In principle, with Deep Convolutional Neural Networks (DCNN), the YOLOv8 can operate in numerous hard and challenging conditions, e.g., traffic conditions, various lighting regards, and weather. YOLOv8 produces bounding boxes, which may enclose the position of the found emergency vehicles, one at a time, processing each frame in the video stream. These bounding boxes visually denote the presence of the emergency vehicles in the video feed as well as its specific location.

Otherwise, the audio clip is scanned in details. The relevant audio features will be extracted using the librosa module which is a powerful Python library package that performs music and audio analysis operations. The Librosa Acoustic Frontend has the Auditory Cortex. To capture such properties of audio signals like frequency, intensity and temporal properties, it carries out complex audio feature extraction. The audio features which have been processed are provided into the Temporal Pattern Recognition Unit that then analyses them.

These extracted audio features are then fed to the Temporal Pattern Recognition Unit that consists of the LSTM network. As it is recurrent neural network architecture, LSTM is quite excellent working with sequential analysis and time-series data. LSTM model is a requisite aspect of this system since it is an audio feature classifier. It has been tediously taught to do multi-class classification, that is, to classify the audio signatures into various classes of emergency vehicles. Due to the fact that LSTM is recurrent, it is capable of classifying the emergency vehicle sounds with reasonable precision by learning the sophisticated rhythms and time-dependent tendencies in the audio set. In order to demonstrate the functionality of the system, simple, but easy to use interface has been meticulously designed. The users will be able to deposit their sample files into this interface through which they can initiate the entire processing channel. According to the input data, the system indicates real-time status of the emergency vehicles that it has detected with similar accompaniment of audio classification of results. Once the execution has been completed, users can easily visualize the results that has been achieved through the visual and auditory sensor models.

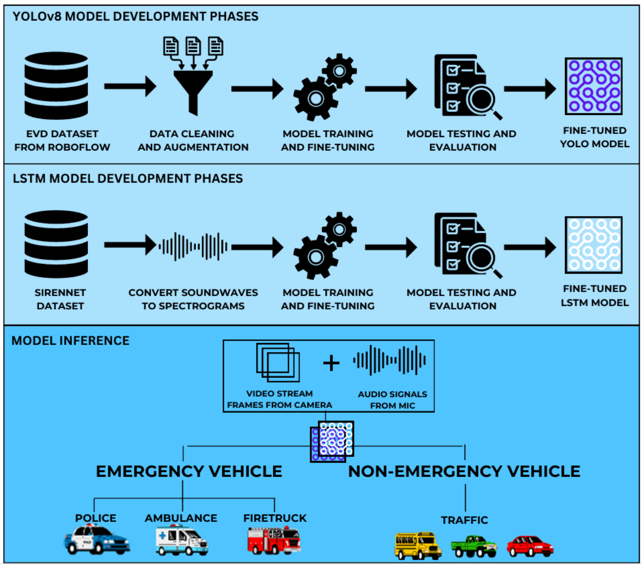
YOLOv8 MODEL DEVELOPMENT PHASES
EVD DATASET FROM ROBOFLOW
DATA CLEANING AND AUGMENTATION
MODEL TRAINING AND FINE-TUNING
MODEL TESTING AND EVALUATION
FINE-TUNED YOLO MODEL
LSTM MODEL DEVELOPMENT PHASES
SIRENNET DATASET
CONVERT SOUNDWAVES TO SPECTROGRAMS
MODEL TRAINING AND FINE-TUNING
MODEL TESTING AND EVALUATION
FINE-TUNED LSTM MODEL
MODEL INFERENCE
+
VIDEO STREAM FRAMES FROM CAMERA
AUDIO SIGNALS FROM MIC
EMERGENCY VEHICLE
NON-EMERGENCY VEHICLE
POLICE
AMBULANCE
FIRETRUCK
TRAFFIC

Fig. 2 *System workflow diagram*

Fig. 2 provides a concise explanation of how the system works and includes the data collection, preprocessing, model selection, training, and evaluation as well as inference (prediction) stages. A non-emergency or an emergency course should be expected. Transparency in system decision-making is enhanced, and it is also interpretable when XAI techniques are applied. XAI enables stakeholders to build an understanding and trust due to how the system comes up with its classifications. The stakeholders will be able to effectively interpret and validate the outputs of the system due to the handy context that the visualization of model predictions, feature importance, and decision justifications offer.

2.1 Data Collection and Preprocessing

A large subset of custom labeled data is initially mined and augmented and then is utilized to train the YOLOv8 model. In particular, the LSTM network is applied to the vehicle classification and audio feature extraction. The next steps are model selection, fine-tuning, post-processing and thorough evaluation.

The dataset used for training the YOLOv8 model contains images collected from the EVD dataset from Roboflow [24]. The data split up is mentioned in Table 1 and Table 2. Data augmentation is then performed on the dataset to generate new images to improve the diversity of data. The dataset captures various low-light and low-visibility conditions.

Table 1 *Data split up (YOLO)*

	Ambulance	Firetruck	Traffic
Train	360	360	381
Test	40	40	40
Total	400	400	421

Table 2 *SireNNet data split up (LSTM)*

	Ambulance	Firetruck	Traffic
Train	845	845	840
Test	211	211	210
Total	1056	1056	1050

The sireNNet dataset [21] was used to extract audio features and train the LSTM model. It consists of 1203 sounds from emergency vehicles and regular traffic. These sounds are classified and represent each type of emergency vehicle. Easch audio clip has a 3-second duration and is of .wav format. The data split up is mentioned in TABLE 2.

The sireNNet dataset encompasses a diverse range of sound profiles characteristic of various emergency vehicles and ambient traffic noise. Notably, these audio samples are meticulously categorized into distinct classes, according to each emergency vehicle type, and the sound waves have been transformed into spectrograms. These spectrograms are plotted to visually represent the frequency content of the audio signals over time.

The SireNNet and the Emergency Vehicle Detection Dataset from Roboflow datasets can limit and generalize the proposed system. We collected these two datasets and test them to the proposed framework. Moreover, the class imbalance issue may affect the emergency vehicles or sirens look less commonly than non-emergency. The overfitting is another issue, and it may affect the system by reducing the accuracy results.

Fig. 3 indicates the waveform of a fire truck siren sound, an ambulance siren, and the waveform of traffic noise. These figures represent the nature of soundwaves corresponding to different types of non-emergency and emergency vehicles on the road.

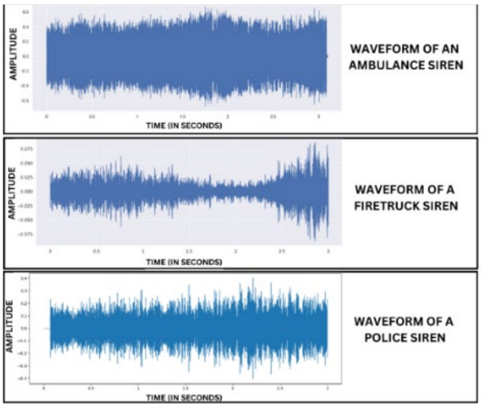

Fig. 3 *Waveform of emergency vehicles and traffic*

2.2 Model Selection

The effective detection of emergency vehicles in urban environments demands an integrative approach that combines audio and video modalities. This paper elucidates the meticulous model selection process, concentrating on two state-of-the-art neural network architectures: YOLOv8 for video-based object detection and LSTM for audio sequence analysis.

The YOLOv8 model is the latest release of the YOLO algorithm as of October 2023. The YOLOv8 architecture builds upon previous YOLO versions and comprises two main components: the backbone and the head. The backbone is based on a modified CSPDarknet53 architecture with 53 convolutional layers, utilizing cross-stage partial connections to enhance information flow. This can be mathematically represented as:

$$f(x) = W \cdot x + b \tag{1}$$

Here W represents the weight matrix, x denotes the input vector, and b signifies the bias vector. The head consists of multiple convolutional and fully connected layers, responsible for predicting bounding boxes, abjectness scores, and class probabilities. An essential feature is the self-attention mechanism in the head, enabling the model to focus on relevant image features. This can be expressed as:

$$A(Q,K,V) = softmax\left(\frac{Qk^T}{\sqrt{d_k}}\right)v \tag{2}$$

Where Q, K, and V denote the queries, keys, and values, respectively, and d_k represents the dimensionality of the key vectors.

Another critical feature is YOLOv8's multi-scaled object detection, facilitated by a feature pyramid network that detects objects of varying sizes within an image. This network includes multiple layers for detecting objects at different scales. Key metrics for evaluating YOLOv8 include mAP (mean Average Precision), which measures accuracy across object classes, FLOPs (floating-point operations) for estimating computational requirements, and speed (frames per second) for real-time performance, and "params," referring to trainable model parameters. "Size (pixels)" in YOLO typically denotes the input image size for training. YOLO model produces promising results in identifying the humans at the disaster location [22].

Table 3 *Model comparison*

	Results			
Model	Precision	Recall	F1-Score	Accuracy
CNN	0.310	0.450	0.450	0.454
LSTM	0.970	0.970	0.970	0.966

An extensive evaluation of two prominent deep learning architectures: CNN and LSTM networks was conducted to identify the most suitable model for audio detection. The results of the analysis were collected and evaluated as shown in TABLE 3.

These gates regulate whether data is retained or excluded. They are made up of a sigmoid neural net layer and a point wise multiplication operation. An LSTM has three gates that control and govern the cell state. The "forget gate layer," which makes use of the current input and the previous hidden state, determines which information from the cell state should be forgotten in the first step.

$$f_t = (W_f.[h_{t-1}, x_t] + b_f) \tag{3}$$

An input gate layer and a tanh layer that produces a candidate value for the new cell state are involved in the next step, which is determining what data should be stored.

$$i_t = (.[h_{t-1}, x_t] + b_i) \tag{4}$$

Next to that, the old cell state is updated to a new one by adding the newly selected, appropriately scaled candidate values and multiplying the old state by the output of the forget gate. Finally, the network uses the filtered cell state to determine what to output. To do this, a sigmoid layer is used to identify the pertinent cell state components, which are then passed through a tanh function and multiplied by the sigmoid gate's output.

$$C_t = \tanh\,(W_C.[h_{t-1},] + b_C\ ) \tag{5}$$

The LSTM uses the filtered version of the cell state to determine what to output after the cell state update. In order to do this, the output gate activation values (O_t) are calculated using the current input (x_t) and the previous hidden state (h_{t-1}). They are then multiplied by the sigmoid gate's output after the pertinent cell state components have been identified and run through a tanh function.

$$O_t \ = \ \sigma(W_o \ \cdot \ [h_{t-1}, \ x_t] + b_o) \tag{6}$$

$$h_t = o_t \cdot tanh\ tanh\ (c_t) \tag{7}$$

Each gate's distinct functionality contributes to the LSTM's capability to learn and process temporal patterns, making it a powerful tool for various sequential data tasks, including emergency vehicle detection.

2.3 Model Fine-Tuning

The YOLOv8 Model is trained on a custom EV dataset with 15 epochs. The ideal value was determined through trial and error by continuously monitoring the result metrics, such as loss and mean Average Precision (mAP). Cases of overfitting and underfitting were identified and eliminated after comparing the outcomes of several trial runs. The outcomes of multiple trial runs are thoroughly examined, with a focus on striking a balance between model complexity and generalization ability.

Additionally, the YOLOv8 Model's hyperparameters are changed during the fine-tuning process to improve its generalization and discriminative power. Through methodical experimentation and ongoing validation and DFL loss monitoring, parameters like learning rate, batch size, and augmentation techniques are adjusted to maximize the model's capacity to capture intricate patterns and features related to emergency vehicles while reducing overfitting to the training data.

In order to evaluate the model's robustness and generalization performance, the fine-tuning process also involves thorough validation and testing of the model on separate datasets. To find cases of overfitting, the results of several trial runs are carefully examined. The YOLOv8 Model is precisely calibrated to identify and categorize emergency vehicles in a variety of real-world scenarios thanks to this iterative approach to model training and optimization, strengthening the system's overall robustness and dependability.

A thorough fine-tuning procedure that included preprocessing audio data and fine-tuning hyperparameters was carried out in order to maximize the performance of the LSTM component within the emergency vehicle detection system. The extraction of audio features more especially, Mel-frequency cepstral coefficients is the basis of the model. A user-defined function that has been tailored for this use case contains the feature extraction procedure. After loading the audio data and computing the MFCCs for each audio file, feature normalization is performed. For model training, the scaled MFCC features that are produced are then combined into a data frame. The system's central component is the LSTM model, which is made to handle the classification task and accept the pre-processed MFCC features. Fig. 4 shows the model architecture.

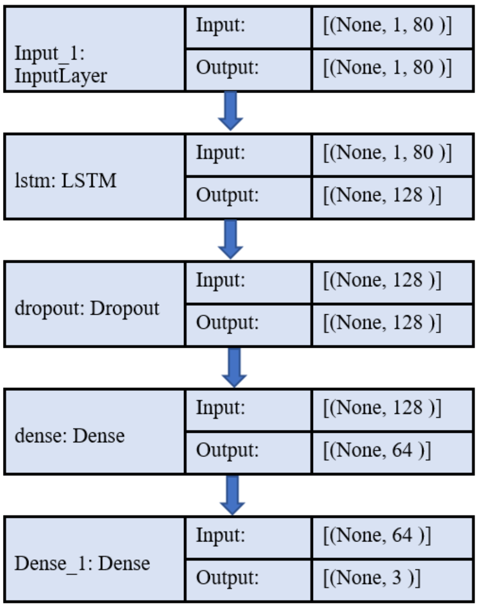

Fig. 4 *Visualization of LSTM model configurations*

A single LSTM layer consisting of 128 units is applied subsequent to a dropout layer to deal with the problem of overfitting. This process of classification considers a classification of two densely interconnected layers, and the final layer makes use of a soft-max activation with the aim of achieving probability scores of some classes. Some of the hyperparameters that were varied to check the optimizations included the number of LSTM units, rate of dropout, and learning rate. The primary measure of evaluation was accuracy, and the Adam optimizer was chosen in terms of model training. The dropout layer within the architecture played a significant role in minimizing the overfitting problems and the necessity to estimate the probabilities of the classes defined the fact that soft-max activation function should be chosen to provide the last dense layer.

The first part of the neural network is the Input Layer (Input_1) whose ingress point can support the temporal sequences of 80 features per input. This layer is of great help in the ingestion and processing of temporal data. Input layer the network employs LSTM layer (LSTM) after the first layer in order to access long-range dependencies that are observed in time series. This layer consists of 128 units capable of matching complicated time patterns that are useful to fine classification. A Dropout Layer (Dropout) is deliberately added to combat over-fitting and enhances generalization. In order to achieve the greatest possible connectivity during the training and resilience in learned representations, dropout regularization is applied without any of the units being discarded. After LSTM layer, more advanced features of the temporal representations that are learned through a Dense Layer (Dense) that comprises 64 units fully connected. This layer plays an essential role in the abstraction of vital attributes that are required to classify them correctly. The last layer of the architecture (Output Layer) contains three units (Dense_1) that regulate final output of the classification. This layer finds probability distribution of the three classes that represent the varieties of emergence vehicles based on the soft-max activation operation.

In the neural network architecture, there are 115,459 parameters, which have been designed to be trained. This painstakingly designed architecture combines state-of-the-art methods designed to identify and categorize emergency vehicles with an unmatched level of precision and effectiveness. The neural network architecture that used in this study is proposed to classify emergency vehicles accurately by extracting temporal patterns.

2.4 Post-processing and Evaluation

The modeling evaluation section in its entirety covers performance of the proposed YOLO model in determining the emergency vehicle. This section dives into the most important evaluation metrics, which include accuracy, precision, recall, and F1 score and are calculated based on the confusion matrix of the model. Effectiveness and reliability of the YOLO model in the accurate detection of emergency vehicles in a range of situations become explainable through precise analysis. The various performance measures are calculated with the following formula:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

$$Recall \ = \frac{TP}{TP + FN} \tag{10}$$

$$F1\,Score \ = \frac{2\ x\ (Precision\ x\ Recall)}{Precision \ + \ Recall} \tag{11}$$

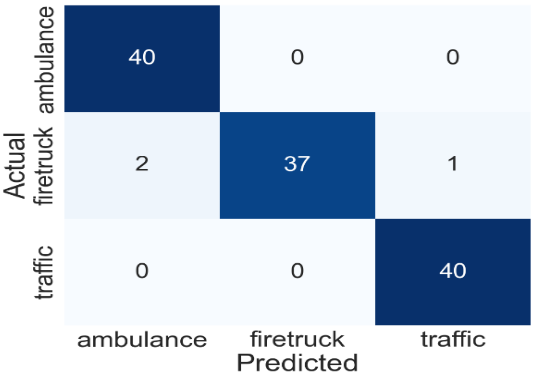

Fig. 5 *YOLOv8 model confusion matrix*

Referring to the provided confusion matrix, the results regarding the performance of the YOLO model showed interesting results in regard to the significant metrics. Fig. 5 is the confusion matrix of the trained YOLOv8 model. Having a great level of accuracy, about 95.6 %, the model showed that it could effectively identify emergency vehicles. The precision or how effective it can accurately predict a positive case was estimated about 97.6% highlighting the low false positive results produced by the model. Also, the recall (sensitivity) of the model (approximately 94.3%) indicated that it could effectively obtain most positive instances of the dataset. Balanced trade-off between precision and recall was realized as the F1 score which is the harmonic mean of precision and recall and was roughly 95.9%. The results illustrate that the YOLO model performs well to detect emergency vehicles and confirm that the model can be applied in real practice where accurate and reliable false identifications of objects are needed.

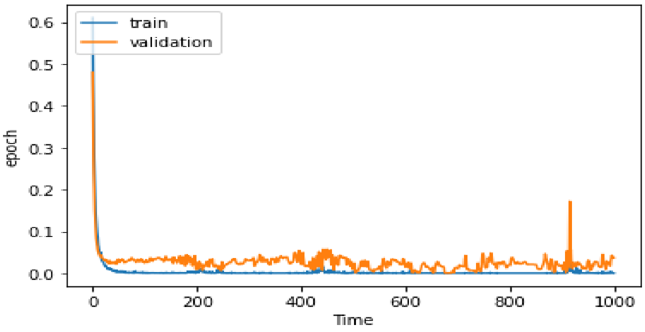

Fig. 6 *Training and validation loss (LSTM)*

Fig. 6 depicts training and validating loss per batch in the LSTM model. The model was evaluated using similarity measurement such as precision, recall, F1 score, and mean average precision (mAP). These measures assisted in the assessment of the accuracy of detecting emergency vehicles and falsely decreasing the alarms. The LSTM model of the siren data obtained a 95.56% accuracy. The Fig. 6 represents loss of the model during training and validation.

The model assessment indicates encouraging outputs in aspects of performance that are critical. The model demonstrates that it is capable of proper identification and classification of the emergency vehicles with a recall value lying between 95.238% and 100 % and with a precision of between 92.5% percent and 100 % in three classes of the emergency vehicles. Also, this model shows the overall accuracy of 97.5% indicating the effectiveness of it identifying emergency vehicles. These calculated F1 scores of about 98.75% and 96.15 % prove an appetizing trade-off between recall and precision, thus proving that the LSTM model is robust and reliable in emergency vehicle detection functions. The presented outcomes reveal the effectiveness of the given model in practical situations when it is required to detect the presence of emergency vehicles in the driving lane by means of correct and successful sound analysis. Fig.7 represents the confusion matrix obtained by testing the LSTM model.

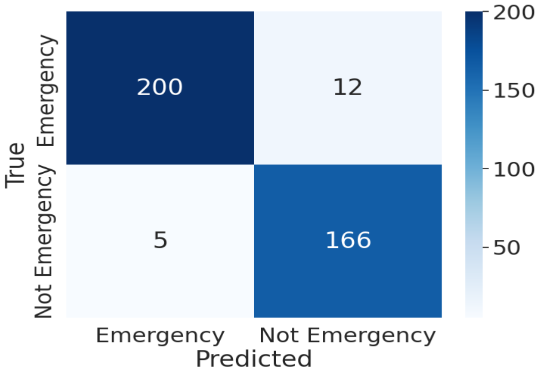

Fig. 7 *Training and validation loss (LSTM)*

3. Results and Discussion

The analysis is presented on three famous methods used in detecting emergency vehicles: YOLO, LSTM, and synergistic combination of the former two. The research critically looks at these procedures concerning major performance parameters, such as precision, performance, and feasibility under hostile conditions. Notably, XAI integration is also investigated in the work and is crucial to enhancing the interpretability and transparency of the decision-making frameworks.

Table 4 represents a comprehensive comparison of three distinct methods for emergency vehicle detection: YOLO, LSTM, and the combined approach leveraging both techniques. The evaluation framework encompasses key performance metrics including accuracy, efficiency, performance in challenging conditions, and integration of XAI techniques when testing the system against real-world data after training the models. YOLO demonstrates a commendable accuracy of 92.1%, effectively identifying emergency vehicles within the dataset. Its high efficiency in processing images facilitates real-time applications, although its performance under challenging conditions such as day-night transitions and low-visibility scenarios is low. Notably, YOLO does not integrate XAI techniques, thereby limiting the transparency and interpretability of its decision-making process.

Table 4 *Comparison of EVD methods*

Method	Accuracy (%)	Efficiency	Performance in Challenging Conditions	Integration of XAI
YOLOv8	92.1	High	Low	No
LSTM	85.2	Moderate	Moderate	No
YOLO + LSTM Combined	96.6	High	High	Yes

The LSTM network has an accuracy of 85.2 % and is very popular due to its ability to detect temporal patterns represented as sequences of data. It would be useful in detecting the presence of emergency cars because it has moderate processing capability on time data. It is particularly remarkable how it could learn about the time patterns and as such, better in tricky situations than YOLO. Similarly to YOLO, LSTM lacks the use of XAI techniques, which may impede the understanding of the stakeholders, and their trust in its decisions.

Last but not least, YOLO and LSTM work together to produce the best accuracy of 96.6%, demonstrating synergistic gains over separate techniques. When compared to standalone YOLO, the combined approach's efficiency is high while maintaining a respectable processing speed. Importantly, incorporating XAI techniques improves decision-making's interpretability and transparency, making it possible for stakeholders to understand the system's classifications. This guarantees safer interactions and builds trust.

The proposed system is an integrated structure consisting of several architectural components, and each component has a specialized role in the identification and classification of emergency vehicles. The initial stage of the operation is the Sensory Fusion which conceptually employs the notion of multi-modal sensory processing and intelligently separates video inputs of different formats into two distinct sensory feeds. Via visual and auditory sensors that are separate, this part of it sets the stage of parallel processing.

The first flow is the Visual Sensor; it embarks on a two-path journey which is achieved through the YOLOv8 model that is famous in object real-time recognition. As a staple of the Visual Cortex, at each frame of the video stream, YOLOv8 will in turn constantly modify its focus on what the video stream is "to see," adding bounding boxes to encapsulate the precise location of the emergency vehicle locations that the video stream is recognizing. Owing to the Deep Convolutional Neural Network (CNN), the YOLOv8 model displays incredible effectiveness in tricky scenarios and ensures high-precision detection at different environmental conditions.

The Auditory Sensor analyzes the audio segment in parallel and by using the librosa module, obtains all audio features. Features of complex audio signals are applied to the Temporal Pattern Recognition Unit, which is processed in the Librosa Acoustic Frontend. In this case, the LSTM network, capable of modeling time series data, identifies audio signatures by the multi-class classification and properly distinguishing the sounds of emergency vehicles. Since LSTM is recurrent, it is able to memorize the temporal dependencies which ensure proper classification regardless of potential audio fluctuations. Interaction and visualization of results are easy due to the many user-friendly features of the system. In this interface, the users are given an opportunity to submit the input files and this triggers execution of the processing pipeline. The system provides real-time information on the identified emergency vehicles as well as audio classification results when it develops. Users can easily visualize results from both visual and auditory sensor models thanks to this extensive interface, which promotes comprehension and confidence in the system's functionality.

The YOLOv8 model had a high level of accuracy in detecting the presence of emergency vehicles in video streams within a continuous real-time performance and displayed impressive object detection processes. The YOLOv8 model can not only find emergency vehicles well in various test scenarios but also under adverse conditions such as bad lighting, different situations on roads, and weather changes. The high accuracy and real-time performances of the model assure timely response of emergency services. The work proved the one-hundred-percent correct location of the sound of an emergency vehicle (voice recognition) in a complex of noise. The model proved successful in its audio classification by making difference between different audio signatures of the emergency and rescue vehicles, such as sirens and horns.

The combined use of the YOLOv8 object detection and LSTM-based audio classification modules forms the whole and trustworthy emergency vehicle detection system. This hybrid solution provides a complete solution, which significantly increases the emergency service efficiency and reliability as the visual as well as audio parts of the emergency vehicle identification is approached. Indeed, the system can show its capability to identify the signature acoustic sound of emergency vehicles under the influence of background noises and varying environmental factors through advanced audio processing methods.

Fig 8 shows the model outputs that have been tested on various emergency vehicles at various times of the day. This is evidenced in the system when emergence vehicle is detected, and the system is confident enough that it is analyzing an ambulance. The effectiveness of the system identifying different types of emergency vehicles relying on their sizeable shades of sound was also reflected in recognizing another emergency vehicle, however, now being a fire truck, and, in this case, saying that it would be a fire truck. The robustness of the system under condition of poor visibility is further exhibited when the system identifies an emergency vehicle at night and recognizes correctly that it is an ambulance.

Finally, the same figure reveals accuracy of the system in detection and classification of normal traffic as non-emergency vehicles. This characteristic achievement in numerous situations shows the excellent and conclusive nature of the proposed system in detecting and classifying emergency vehicles. The system is very flexible and reliable which makes it a cost worthy tool in enhancing the safety of the road as well as its brilliant performance in classification and detection of emergency vehicles. The system is capable of making rapid decisions on the type of vehicle out of identity and other background noise by using deep learning strategies and complex audio processing algorithms regardless of the complex and dynamic environment.

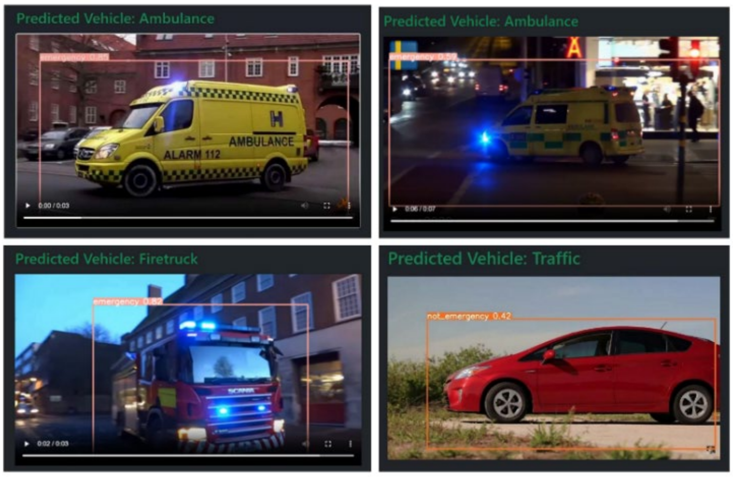

Fig. 8 *System inference and predictions*

4. Conclusion

The Emergency Vehicle Detection System is reliable because it uses both visual and aural information, hence offering an integrated, precise solution when it comes to emergency vehicle identification and classification. Although the LSTM model displays an equal success in categorizing the specific audio cues, YOLOv8 ensures their correct identification in video streams. The results show the efficiency of combining YOLOv8 and LSTM for multi-modal emergency vehicle detection system then single-modal methods. In operations involving emergency services, traffic management, and the general safety of the populace, this multi-modal synergy is insanely beneficial because more accuracy, reliability, and the situation awareness are enhanced. The system is the latest tool to improve emergency response and ensure safer and smarter cities due to its real-time effectiveness and efficiency of the system. The future work will be focusing on multimodal fusion techniques to improve the emergency vehicle detection to improvement. The transformer models will be good choice to combine visual and audio data to enhance the accuracy results in complex environments. In addition, more experiments and results would be evaluated and comparison with existing works/related works.

Conflict of Interest

Authors declare that there is no conflict of interests regarding the publication of the paper.

Author Contribution

The authors confirm contribution to the paper as follows: ***study conception and design:*** *E.P and M.V;* ***data collection:*** *N.S and N.R;* ***analysis and interpretation of results:*** *E.P, M.V and A.O.I;* ***draft manuscript preparation:*** *E.P, N.S, A.O.I and R.G. All authors reviewed the results and approved the final version of the manuscript.*

References

[1] Autonomous Vehicle Incidents Involving Emergency Vehicles. Available online: https://www.mercurynews.com/2023/03/09/map-these-16-tesla-crashes-are-part-of-a-federal-investigation-into-autopilot/ (Accessed on 20 March 2024)

[2] Agnew, D., Lüke, S., Fischer, M., & Krökel, D. (2017). Emergency vehicle detection with digital image sensor. U.S. Patent 9,576,208, issued February 21, 2017.

[3] Cantarini, M., Gabrielli, L., & Squartini, S. (2022). Few-shot emergency siren detection. Sensors, 22(12), 4338. https://doi.org/10.3390/s22124338.

[4] Carvalho Barbosa, R., Shoaib Ayub, M., Lopes Rosa, R., Zegarra Rodr\'\iguez, D., & Wuttisittikulkij, L. (2020). Lightweight PVIDNet: A priority vehicles detection network model based on deep learning for intelligent traffic lights. Sensors, 20(21), 6218. https://doi.org/10.3390/s20216218.

[5] Choudhury, K., & Nandi, D. (2023). Review of Emergency Vehicle Detection Techniques by Acoustic Signals. Transactions of the Indian National Academy of Engineering, 8(4), 535–550. https://doi.org/10.1007/s41403-023-00424-9.

[6] Dodia, A., Kumar, S., Rani, R., Pippal, S. K., & Meduri, P. (2023). EVATL: A novel framework for emergency vehicle communication with adaptive traffic lights for smart cities. IET Smart Cities, 5(4), 254–268. https://doi.org/10.1049/smc2.12068.

[7] Ghaffarian, S., Taghikhah, F. R., & Maier, H. R. (2023). Explainable artificial intelligence in disaster risk management: Achievements and prospective futures. International Journal of Disaster Risk Reduction, 98, 104123. https://doi.org/10.1016/j.ijdrr.2023.104123.

[8] Jagannathan, P., Rajkumar, S., Frnda, J., Divakarachari, P. B., & Subramani, P. (2021). Moving vehicle detection and classification using gaussian mixture model and ensemble deep learning technique. Wireless Communications and Mobile Computing, 2021, 1–15. https://doi.org/10.1155/2021/5590894.

[9] Kolekar, S., Gite, S., Pradhan, B., & Alamri, A. (2022). Explainable AI in scene understanding for autonomous vehicles in unstructured traffic environments on Indian roads using the inception U-Net Model with Grad-CAM visualization. Sensors, 22(24), 9677. https://doi.org/10.3390/s22249677.

[10] Li, G., Wang, Q., & Zuo, C. (2022). Emergency lane vehicle detection and classification method based on logistic regression and a deep convolutional network. Neural Computing and Applications, 34(15), 12517–12526. https://doi.org/10.1007/s00521-021-06468-8.

[11] Lisov, A. A., Kulganatov, A. Z., & Panishev, S. A. (2023). Using convolutional neural networks for acoustic-based emergency vehicle detection. Modern Transportation Systems and Technologies, 9(1), 95–107. https://doi.org/10.17816/transsyst20239195-107.

[12] Mahmud, U., Hussain, S., Sarwar, A., Toure, I. K., & others. (2022). A distributed emergency vehicle transit system using artificial Intelligence of Things (DEVeTS-AIoT). Wireless Communications and Mobile Computing, 2022. https://doi.org/10.1155/2022/9654858.

[13] Mankodiya, H., Jadav, D., Gupta, R., Tanwar, S., Hong, W.-C., & Sharma, R. (2022). Od-xai: Explainable AI-based semantic object detection for autonomous vehicles. Applied Sciences, 12(11), 5310. https://doi.org/10.3390/app12115310.

[14] Nayak, A. (2019). Development of vision-based response of autonomous vehicles towards emergency vehicles using infrastructure-enabled autonomy, PhD diss.

[15] Nwakanma, C. I., Ahakonye, L. A. C., Njoku, J. N., Odirichukwu, J. C., Okolie, S. A., Uzondu, C., Ndubuisi Nweke, C. C., & Kim, D.-S. (2023). Explainable artificial intelligence (xai) for intrusion detection and mitigation in intelligent connected vehicles: A review. Applied Sciences, 13(3),1252. https://doi.org/10.3390/app13031252.

[16] Omeiza, D., Webb, H., Jirotka, M., & Kunze, L. (2021). Explanations in autonomous driving: A survey. IEEE Transactions on Intelligent Transportation Systems, 23(8), 10142–10162. DOI: 10.1109/TITS.2021.3122865.

[17] Patel, R., Mange, S., Mulik, S., & Mehendale, N. (2022). AI based emergency vehicle priority system. CCF Transactions on Pervasive Computing and Interaction, 4(3), 285–297. https://doi.org/10.1007/s42486-022-00093-7.

[18] Rosayyan, P., Paul, J., Subramaniam, S., & Ganesan, S. I. (2023). An optimal control strategy for emergency vehicle priority system in smart cities using edge computing and IoT sensors. Measurement: Sensors, 26, 100697. https://doi.org/10.1016/j.measen.2023.100697.

[19] Uma, K., Sathya Bama, B., & Maheesha, M. (2021). Emergency Vehicle Detection in Traffic Surveillance Using Region-Based Convolutional Neural Networks. Advances in Automation, Signal Processing, Instrumentation, and Control: Select Proceedings of i-CASIC 2020, 561–567. https://doi.org/10.1007/978-981-15-8221-9_49.

[20] Wang, C., Song, Y., Liu, H., Liu, H., Liu, J., Li, B., & Yuan, X. (2022). Real-time vehicle sound detection system based on depthwise separable convolution neural network and spectrogram augmentation. Remote Sensing, 14(19), 4848. https://doi.org/10.3390/rs14194848.

[21] Shah, Arya; Singh, Amanpreet (2023), “sireNNet-Emergency Vehicle Siren Classification Dataset For Urban Applications”, Mendeley Data, V1, doi: 10.17632/j4ydzzv4kb.

[22] Jyotsna Rani Thota, Anuradha Padala, Human Remains Detection in Natural Disasters using YOLO: A Deep Learning Approach, Engineering, Technology & Applied Science Research, 14 (6), 17678-17682, 2024, https://doi.org/10.48084/etasr.8483

[23] NSC. 2020. Emergency Vehicles – Injury Facts. Accessed Jan 7, 2025. injuryfacts.nsc.org

[24] Sak W. (2022). Emergency Vehicle Detection Dataset. Roboflow Universe. Retrieved from https://universe.roboflow.com/project-sawkw/emergency-vehicle-detection-el8ej