Road Traffic Events Monitoring Using a Multi-Head Attention Mechanism-Based Transformer and Temporal Convolutional Networks

Selim Reza, Marta Campos Ferreira, J.J.M. Machado, and João Manuel R. S. Tavares

***Abstract*— Acoustic monitoring of road traffic events is an indispensable element of Intelligent Transport Systems to increase their effectiveness. It aims to detect the temporal activity of sound events in road traffic auditory scenes and classify their occurrences. Current state-of-the-art algorithms have limitations in capturing long-range dependencies between different audio features to achieve robust performance. Additionally, these models suffer from external noise and variation in audio intensities. Therefore, this study proposes a spectrogram-specific transformer model employing a multi-head attention mechanism using the scaled product attention technique based on $softmax$ in combination with Temporal Convolutional Networks to overcome these difficulties with increased accuracy and robustness. It also proposes a unique preprocessing step and a Deep Linear Projection method to reduce the dimensions of the features before passing them to the learnable Positional Encoding layer. Rather than monophonic audio data samples, stereophonic Mel-spectrogram features are fed into the model, improving the model's robustness to noise. State-of-the-art One-dimensional Convolutional Neural Networks and Long Short-term Memory models were used to compare the proposed model's performance on two well-known datasets. The results demonstrated its superior performance by achieving an improvement in accuracy of 1.51 to 3.55% compared to the studied baselines.**

***Index Terms*— Intelligent traffic monitoring, attention mechanism, mel-spectrogram, temporal convolutional networks, learnable spectrogram-specific positional encoding, deep linear projection.**

Received 19 February 2024; revised 20 August 2024 and 19 January 2025; accepted 30 June 2025. Date of publication 14 July 2025; date of current version 16 September 2025. This work was supported in part by the project "Sensitive Industry," co-funded by European Regional Development Fund (ERDF) through the Operational Program for Competitiveness and Internationalization (COMPETE 2020) under the PORTUGAL 2020 Partnership Agreement. The work of Selim Reza was supported in part by the "Fundação para a Ciência e Tecnologia" (FCT) through the Ph.D. Research Grant 2022.12391.BD. The Associate Editor for this article was H. Eldardiry. *(Corresponding author: João Manuel R. S. Tavares.)*

Selim Reza is with the Faculdade de Engenharia, Universidade do Porto, 4200-465 Porto, Portugal.

Marta Campos Ferreira is with INESC TEC and the Departamento de Engenharia e Gestão Industrial, Faculdade de Engenharia, Universidade do Porto, 4200-465 Porto, Portugal.

J.J.M. Machado and João Manuel R. S. Tavares are with the Instituto de Ciência e Inovação em Engenharia Mecânica e Engenharia Industrial and the Departamento de Engenharia Mecânica, Faculdade de Engenharia, Universidade do Porto, 4200-465 Porto, Portugal (e-mail: tavares@fe.up.pt).

Digital Object Identifier 10.1109/TITS.2025.3585801

I. Introduction

ROAD traffic event detection and monitoring are essential components of Intelligent Transportation Systems (ITS). There are various means to accomplish these tasks. For example, traffic flow and occupancy data can be used to monitor traffic incidents on a particular road section [1]. Developing acoustic-based traffic event detection and monitoring algorithms can also be a robust solution to enhance the safety and reliability of ITS. In this regard, some of the most popular research areas are predicting traffic noise to deal with noise pollution [2], [3], and classifying traffic events to enhance activity monitoring. The primary goal of the subsequent models is to detect the temporal activity of sound events in traffic auditory scenes and identify their classes in every instance. However, few challenges exist in developing an efficient acoustic model within the current context; for example, the nature of the traffic sounds to be recognized and how they manifest in natural settings, the presence of external noise, audio intensity variation, overlapping of audio signals, and lack of good quality datasets [4].

The current state-of-the-art solutions generally extract features such as spectrograms and Mel-scale Frequency Cepstral Coefficients (MFCCs) [5], [6] from the input audio signals and train the used models to learn those features for classification purposes. Conventional Machine Learning (ML) models such as Gaussian Mixture Model (GMM) [7], Support Vector Machine (SVM) [8], and Random Forest (RF) [9] are less effective to deal with the current challenges, particularly for polyphonic traffic sound events. On top of that, these models are not configured to instantiate multiple classes simultaneously. Deep Learning (DL) models such as Convolutional Neural Networks (CNNs) and Long Short-term Memory (LSTM) are the current state-of-the-art for efficiently solving these kinds of problems because of their ability to capture hidden features, determine interdependencies between the features and processing power to deal with them sequentially [10].

For CNNs, large receptive fields are needed to track long-range dependencies, resulting in computational and statistical efficiency loss. Furthermore, when large filters are used, CNNs only encode the relative position of different features,

hindering their performance. LSTMs are challenging to train, do not work well with noisy data, and may still suffer from gradient exploding and vanishing problems responsible for inefficient learning. The path lengths between the features grow linearly with the distance between them, hindering the learning of long-range dependencies [11]. The attention mechanism-based transformer models can handle these issues more efficiently than state-of-the-art algorithms. However, vanilla transformers lack an efficient means to capture positional information in audio spectrograms.

Zhang et al. [12], Guan et al. [13], and Tran and Tsai [14] proposed attention mechanism-based acoustic event detection and monitoring algorithms. A frame-level attention mechanism was proposed in [12] to generate discriminative characteristics and automatically concentrate on semantically relevant frames. The audio signals were converted into a log Gammatone spectrogram to feed into eight CNN layers, followed by two Bi-directional Gated Recurrent Units (Bi-GRU), combining the attention mechanism. Guan et al. [13] employed the self-attention mechanism with adjustable sparsity to eliminate irrelevant audio features, improving the model's robustness using CNNs and a transformer encoder to capture the local and temporal features from the log Mel-spectrogram as the input. In the final step, a fully connected layer was used for classification. Tran and Tsai [14] combined CNNs with an attention mechanism to propose an acoustic train arrival detection model, where Mel-spectrogram and MFCCs were merged to feed into the model to improve its accuracy and robustness. However, these models lack robustness to various sound durations, loudness, and noise levels. Although the accuracy achieved on datasets containing two classes has attained an acceptable level, the performance on multi-class classification still requires further improvements. The proposed model aims to address the aforementioned shortcomings in a more manageable way by presenting a new architectural paradigm. The input raw audio signals are uniquely preprocessed by (i) removing background noise, (ii) removing silent and near-silent segments from the edges of the audio signals, (iii) energy normalisation, (iv) audio augmentation, and (v) symmetric zero-padding with centring before being transformed to Mel-spectrograms. Currently available Positional Encoding (PE) layer treats all input dimensions equally with a static encoding scheme, resulting in less efficiency in dealing with audio spectrograms. This research proposes a Spectrogram-specific Positional Encoding (SPE) layer, which is learnable and considers that different dimensions benefit from different positional representations. Before feeding the preprocessed data as the encoder input, the feature values are linearly projected to lower their dimensions and improve the processing efficiency through a unique Deep Linear Projection (DLP) mechanism. The encoder consists of several layers using a multi-head attention mechanism to encode the data representation deeply. A Temporal Convolutional Network (TCN) block is then used to process them and is passed to the classification block consisting of a Dense layer with the $softmax$ activation function. Two state-of-the-art datasets were used to train and evaluate the proposed model and the baselines for comprehensive comparisons. Additive White Gaussian Noise (AWGN) [15] was added, and audio intensity was adjusted by modifying the signal amplitudes of the test datasets for robustness testing. The categorical cross-entropy function was used to calculate the loss and improve the learning efficiency of the proposed model during training. The schematic representation of the proposed model is illustrated in Figure 1.

The performance of the proposed model was compared with a state-of-the-art One-dimensional CNN (1D-CNN) model similar to the one suggested in [16], and an LSTM model proposed in [17], [18] and [19]. The results demonstrated its worthiness by achieving accuracy scores of 94.80 and 95.87% on the two used datasets, which are 1.51 to 3.55% higher than the baselines. The main contributions of this research are:

- It proposes a learnable SPE layer and a unique DLP mechanism to lower the dimensions of the input features before feeding them into the encoder to enhance the model's performance;
- The stereophonic raw audio data samples are uniquely preprocessed and transformed into Mel-spectrogram as the inputs instead of making them mono-channel signals to increase the model's robustness to noise;
- It uses a TCN block with residual connection to facilitate the construction of deep networks and capture the features' short- and long-term dependencies.
- It also proves the superior outcomes of the $softmax$ based attention compared to the $sigmoid$ based attention mechanism, contrary to [12]. The proposed model trained on $17K$ data samples can outperform the CNN-based model without any transfer learning approach. This is a contradictory outcome from the audio spectrogram transformer model proposed in [20].

This article is organized as follows: Section II presents a summary of state-of-the-art related works along with their performances and limitations; in Section III, the formulation of the proposed model is presented; the experimental setup and results are addressed in Section IV; Section V is devoted to discussing the overall performance and feasibility of the proposed model compared to state-of-the-art baseline models; and finally, the conclusions are drawn in Section VI.

II. Related Works

This section introduces recent DL-based models for solving acoustic road traffic event monitoring problems. The current models seek to identify auditory activity on road networks and predict classes in each case. However, with various noise components and other environmental phenomena, this task is challenging to accomplish efficiently.

In the literature, CNNs and LSTM-based models dominate in solving acoustic event monitoring problems because they can dexterously learn and process audio features. Wang et al. [21] proposed a depth-wise separable CNN model for road traffic sound event monitoring tasks. Before extracting the MFCC features, a spectrogram augmentation technique is applied to the Mel-spectrogram. This approach demonstrated a frame-wise classification accuracy of 94.64%. However, the accuracy deteriorated in the presence of noise

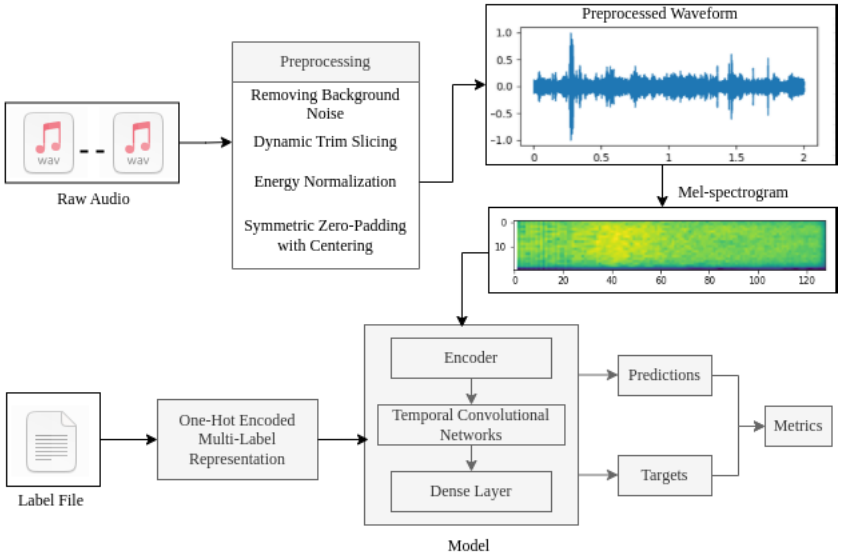

Fig. 1. Schematic representation of the overall concepts of the proposed model.

and hence lacked robustness. Kim et al. [22] combined the concepts of Residual Neural Networks (ResNet) and squeeze-and-excitation networks with the sample-level CNNs (SampleCNNs) architecture to solve audio classification problems. This model directly took raw audio waveforms as input instead of Mel's spectrograms and required a very small size of filters in all layers. The results demonstrated an Area Under the Receiver Operating Characteristic Curve (ROC-AUC) score of 90.83% on the MagnaTagATune dataset [23], which is more or less similar to the state-of-the-art result of 90.64%. Therefore, more robust algorithms are required to accomplish further performance improvements.

Zhang et al. [24] proposed a fast binary spectral features-based Deep Belief Network (DBN) for traffic event detection problems. The results showed better outcomes than the end-to-end CNNs and LSTM-based models. It achieved an average recognition rate of 79.0%, which is 2.9 and 5.6% higher than the CNNs and LSTM-based models, respectively. However, it lacked robustness, and further accuracy improvements are required. Marchegiani and Newman [25] proposed a multitasking learning scheme to classify and localise emergency vehicles from their sirens using a CNN-based U-Net [26] architecture. Considering the Gammatonegrams of the incoming sound signal as intensity images, a noise removal technique was implemented to capture inter-channel information, achieving an average classification rate of 94% across different sound classes. However, it lacks the robustness to deal with different noise sources and different natures of audible alarms.

Attention mechanisms have also been proposed to solve the problems under study. Lee et al. [27] proposed an attention-based multimodal sound event location and detection model using the DCASE2021 dataset [28]. The authors extracted Mel-spectrogram, log-chromatogram, and additional spectral information from the input raw audio signal to train the proposed model. The encoder was based on a CNN and was responsible for exchanging intermediate resources within its layers using the parameter-sharing approach. The decoder used an attention mechanism to improve the prediction efficiency. Tran and Tsai [14] combined a CNN with a temporal and frame-level attention mechanism to develop an acoustic model of train arrival detection by merging the UrbanSound8K [29] and ESC-50 [30] datasets with their private dataset for model training and testing purposes. Spectrograms and MFCCs were extracted and combined from the input audio signals to increase the robustness of the model in gaining robustness to various levels of environmental noise. The model achieved an average accuracy ranging from 91.13 to 95.11% at various noise levels.

Guan et al. [13] proposed a sparse self-attention mechanism for acoustic event detection to remove irrelevant features of various classes of sounds and background noise on the DESED dataset [31]. The results demonstrated a Polyphonic Sound Detection Score (PSDS) [32] of 66.7%, which is slightly lower than 66.9%, obtained from the CNN-based model presented in [33]. Besides, the self-attention mechanism is limited in modelling input dependencies unless the number of layers or heads increases with the input length. Hence, multi-head attention mechanism-based models are required to deal with these problems.

In summary, the models available in the literature lack robustness and accuracy to deal with background noise. Critical common difficulties are: (i) identifying efficient features to neutralise noise, (ii) addressing gradient vanishing and exploding phenomena, (iii) minimising the effect of audio intensity variation, and (iv) improving effective parallel processing. The proposed model aims to deal with these problems more flexibly and smoothly.

III. Methodology

Humans selectively focus on parts of the inputs to acquire relevant information. The concept of attention was brought into the Neural Network (NN) domain under this pretext. Thus, when attention is applied, some parts of the inputs are

improved, as they assume more weight in predicting the final output. Attention can be calculated using different approaches, such as dot product attention, additive attention, and scaled dot product attention. The intuitions between them are the same: knowing which parts of the inputs are most important to predict the final output.

Suppose that $[x_1, x_2, x_3, \ldots.., x_n]$ are audio data samples with $[y_1, y_2, \ldots., y_L]$ labels, where $n \in \mathbb{R}$ and L are the total number of samples and classes, respectively. This study modelled the two datasets as $L = 5$ and $L = 8$ class classification problems. The goal is to train the proposed model on the training dataset and classify the test data samples by providing all the labels' corresponding prediction probabilities.

A. Scaled Dot Product Attention

The query Q is a vector where each row corresponds to one query, meaning it is a simple set of queries. Its dimension is $[seq\_len\_q \times d_k]$, where seq_len_q and $d_q = d_k$ are the total numbers of queries and dimensions of each query, respectively. Similarly, the dimensions of the key K and value V vectors are $[seq\_len\_k \times d_k]$ and $[seq\_len\_v \times d_v]$, where seq_len_k and seq_len_v are the numbers of keys and values, and d_k and d_v are the dimensions of each key and value, respectively. When a query is made, the model finds which values are similar to that specific query. Values cannot be accessed directly but through the keys when a query is made. Each value has a corresponding unique key, and the number of keys must equal the number of values, but their dimensions do not have to be equal. In short, the query and value vectors are like two sets of questions and answers, where the keys are the access to the answers. The attention mechanism maps the set of queries into a set of key-value pairs using:

$$Attention(Q,\ K,\ V) = softmax(\frac{QK^T}{\sqrt{d_k}})V, \tag{1}$$

where QK^T is the dot-product between the query and key vectors. The reason for taking the dot product is that it gives a similarity score between two vectors. Now, let's assume that $\frac{QK^T}{\sqrt{d_k}}$ represents the scaled dot-product, which gives scaled similarity scores. Then, the $softmax$ operation is performed, which converts all the scaled similarity scores into probabilities so that all constraints are in the range between $[0, 1]$. Now, this operation outputs some weights, i.e., the attention weights, between 0 (zero) and 1 (one) and is multiplied by V. Hence, when a query is made, the model will ultimately output a weighted summation of all the values corresponding to that particular query. Figure 2 depicts the scaled dot-product mechanism.

B. Multi-Head Attention

The attention mechanism is called self-attention when Q, K, and V matrices are equal. In the multi-head attention mechanism, these matrices are passed through corresponding Dense layers to project them into lower dimensions. After that, the split heads mechanism transforms each Q_D, K_D and V_D according to the batch size using $(Q_D,\ Batch\_size)$, $(K_D,\ Batch\_size)$ and $(V_D,\ Batch\_size)$, where Q_D, K_D and V_D are the corresponding outputs of Q, K and V matrices of the Dense layers, respectively. These operations are performed several times, equal to the number of heads. Then, the scaled dot product attention mechanism is applied for each case. Finally, they are concatenated and passed through another Dense layer to obtain the weighted average of the values corresponding to a given query, as shown in Figure 3.

C. Encoder

The proposed model processes the raw inputs using a unique preprocessing mechanism and then applies a DLP transformation before entering further steps. First, the Mel-spectrogram features are passed through a Dense layer with a linear activation function to reduce their dimensions. They are then multiplied by the square root of the model dimension before being sent to the SPE layer.

1) Linear Projection: Let's consider a vector v in two-dimensional space and a point p not on the vector line but inside the vector space. Then, $\bar{p}$ is the projection of p onto the vector line. Now, let's assume that i is the dimension of the vector space and the linear projection must satisfy:

$$\begin{aligned} argmin_c \sqrt{\sum_i (\bar{p}_i - p)^2} &= argmin_c \sum_i (\bar{p}_i - p)^2 \\ &= argmin_c \sum_i (cv_i - p)^2, \end{aligned} \tag{2}$$

where cv_i is equal to $\bar{p}_i$, i.e., is a scalar multiplication between c and v_i. Equation 2 is minimised by differentiating it with respect to c. Now, $\frac{d}{dc}\sum_i (cv_i - p)^2$ would be equal to $2(\sum_i (cv_i{}^2 - \sum_i v_i p)$. After making some rearrangements, one can obtain:

$$\frac{d}{dc}\sum_i (cv_i - p)^2 = 2(c\acute{v}v - \acute{v}p) \Rightarrow 0. \tag{3}$$

Satisfying the condition: $2(c\acute{v}v - \acute{v}p) = 0$, where $\acute{v}$ is the transpose of v, will help calculate the value of c as:

$$c = (\acute{v}v)^{-1}\acute{v}p. \tag{4}$$

Thus, the linear projection matrix P of p point onto the v vector is idempotent and can be formulated as:

$$P = (\acute{v}v)^{-1}\acute{v}. \tag{5}$$

a) Implementation: Therefore, the point projection p on some vector v is a function that indicates the points closer to it along the vector. The closest one is the Euclidean distance $\sqrt{\sum_i (\bar{p}_i - p)^2}$ over dimension i, as suggested by Equation 2. Of course, there are other points further away from the Euclidean distance. Hence, the linear projection can effectively represent a higher-dimensional vector space onto a specific number of dimensions. A Dense layer with a linear activation function can mimic this operation.

Let's assume that the input is a vector of dimension i; thus, the goal is to construct a linear projection matrix of dimension j, where $j < i$. For simplicity, let's assume that the input is $\{X_{ijk} \in \mathbb{R}\ | i = 1, 2, \ldots, m;\ j = 1, 2, \ldots, n;\ k = 1, 2, \ldots, l\}$.

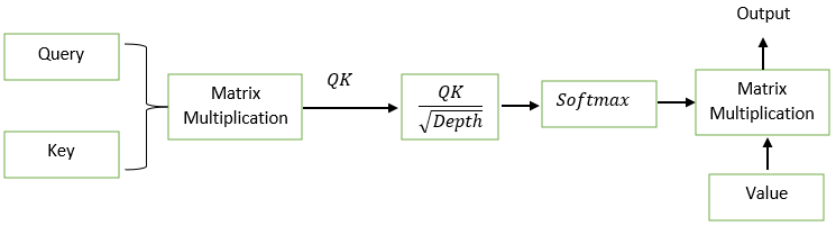

Fig. 2. Block diagram of the scaled dot-product attention mechanism: The Query, Key, and Value matrices are generated using three Dense layers with several neurons equal to the model's dimension.

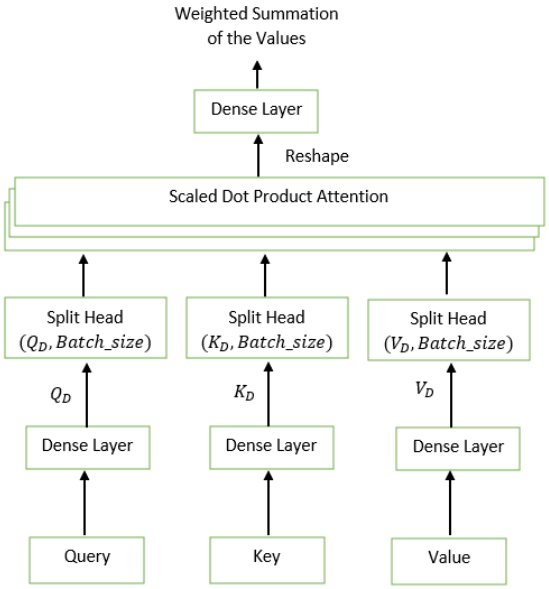

Fig. 3. Block diagram of the multi-head attention mechanism (after several trial-and-error tests, it was found that 4 attention heads provided the best outcomes).

If Y_{ijk}, W_{ijk}, and b_i represent the output vector, weight vector, and bias, respectively, then the output Z_{ijk}, is formulated as:

$$Z_{ijk} = W_{ijk}X_{ijk} + b_i. \tag{6}$$

Thus, the definition of Z_{ijk} becomes $\{Z_{ijk} \epsilon \mathbb{R} \mid i = 1, 2, \ldots, m; j = 1, 2, \ldots, n; k = 1, 2, \ldots, d\}$ with d being the model's dimension. Now, the output of the proposed DLP Layer is further modified by performing a scalar multiplication between Z_{ijk} and K using:

$$P_{ijk} = Z_{ijk}K, \tag{7}$$

where $K = \sqrt{d}$. In other words, each $(i, j, k)^{th}$ element of Z_{ijk} is multiplied by $\sqrt{d}$ for all possible values of i, j and k.

b) Effects: Current state-of-the-art attention mechanism-based models feed N-dimensional inputs directly into the PE layer, resulting in improper mapping of the positions of the vector elements. The above-mentioned technique can overcome this problem by representing higher-dimensional inputs in lower dimensions before transmitting them to the PE layer.

2) Spectrogram-Specific Positional Encoding: The relative or absolute positions of the sequence elements must be considered to enhance the model's efficiency. The PE layer is best suited for that purpose. Let's assume x is the desired position in an input sequence, and $\overrightarrow{PE}_x \epsilon \mathbb{R}^d$ is its corresponding encoding with d as the encoding dimension. Then, $f : \mathbb{N} \rightarrow \mathbb{R}^d$ produces the output vector as [34]:

$$\overrightarrow{PE_x}^{(i)} = f(x)^{(i)} = \begin{cases} sin(w_k \cdot x), & if\ i = 2k, \\ cos(w_k \cdot x), & if\ i = 2k+1, \end{cases} \tag{8}$$

where $w_k = \frac{1}{1000^{2k/d}}$ is the frequency component and considered as fixed according to currently available transformer-based models. However, this approach is most likely not suitable to deal with spectrogram-specific tasks for several reasons: (i) w_k is predetermined, which follows a fixed exponential decay mechanism, may not optimally represent the positional relationship, (ii) all dimensions are equally treated ignoring the fact that different dimensions may benefit from different positional representations, and (iii) the encoding scheme is static, does not evolve during training, and is not adaptable to the data characteristics. This study proposes a learnable SPE layer aiming to address these concerns. Hence, it introduces learnable amplitudes A_k and phases ϕ_k for the sinusoidal components using:

$$\begin{aligned} \overrightarrow{SPE_x}^{(i)} &= f(x)^{(i)} \\ &= \begin{cases} A_k \cdot sin(w_k \cdot x + \phi_k), & if\ i = 2k, \\ A_k \cdot cos(w_k \cdot x + \phi_k), & if\ i = 2k+1, \end{cases} \end{aligned} \tag{9}$$

where A_k and w_k represents amplitude and phase of k^{th} frequency component. The PE layer's outputs pass through a Dropout layer and are fed as inputs to the encoder multi-head attention layer, together with the encoder padding mask.

3) Encoder Layer: Each encoder layer consists of two parts. A multi-head attention layer is used in the first part according to the formulation presented in Section III-B. It produces a weighted average of V matching a specific Q. Now, let's suppose M_{ijk} is the output of the multi-head attention block, which can be obtained by concatenating $Attention(Q_D,\ K_D,\ V_D)$ of Equation 1 over the number of heads, H. The input of the encoder layer becomes $I_{ijk} = P_{ijk} + SPE$. Then, there is a Dropout [35] layer and a Layer Normalization (LayerNorm) [36], and during the forward is considered:

$$e_1 = M_{ijk}(I_{ijk}), \tag{10}$$

$$e_2 = Dropout(r)(e_1);\ \ r = rate, \tag{11}$$

$$e_3 = LayerNorm(I_{ijk} + e_2). \tag{12}$$

The second part contains multiple Dense layers with the Gaussian Error Linear Unit (GELU) [37] activation function and a LayerNorm. Here, e_3 is fed to two Dense layers followed by a Dropout layer with a dropout rate of r. Finally, a LayerNorm operation is performed on $(e_3 + e_6)$ to obtain the final embedded outputs of the encoder layer. During the forward pass, these can be formulated as:

$$e_4 = GELU(W^1_{ijk} e_3);\ \ W^1_{ijk} \in \mathbb{R}^{b\times n\times u}, \tag{13}$$

$$e_5 = GELU(W^2_{ijk} e_4);\ \ W^2_{ijk} \in \mathbb{R}^{b\times n\times d}, \tag{14}$$

$$e_6 = Dropout(r)(e_5), \tag{15}$$

$$e = LayerNorm(e_3 + e_6), \tag{16}$$

where b and n represent the batch size and time steps, with u being the number of neurons in the Dense layer.

D. Proposed Model

The architecture of the proposed model is illustrated in Figure 4. Instead of directly feeding the Mel-log filterbanks features as in the case of [20], the Mel-spectrogram features are passed to the model as inputs. It comprises four identical encoder layers within the encoder block. The outputs of the first encoder layer are fed back as input to the second one, and the process continues. Finally, the last encoder layer outputs are reshaped and passed through the TCNs block.

1) Temporal Convolutional Networks: The deeply encoded data representations are processed with TCNs to preserve the temporal order and learn both short and long-term relationships. A series of transformations is undertaken using 1D-CNN, Batch Normalisation, Dropout layer, and ReLU activation function. On top of that, a residual connection is used, aiming to enhance the model's stability and facilitate gradient flow. However, before the TCNs, the encoder output $e \in \mathbb{R}^{b\times n\times d}$ is passed through a Global Average Pooling Layer, which operates along the time dimension. It outputs $g(e) \in \mathbb{R}^{b\times d}$ by reducing the shape into $(b \times d)$. The used TCNs are constructed as:

$$C_t = \sum_{k=0}^{K-1} g(e_{t-k}) \cdot W_k;\ \ C_t \in \mathbb{R}^{b\times f}, \tag{17}$$

$$C_{t1} = ReLU(BatchNorm(C_t)), \tag{18}$$

$$C_{t2} = Dropout(C_{t1}), \tag{19}$$

$$R_t = ReLU(C_{t2} + g(e_t));\ R_t \in \mathbb{R}^{b\times f}, \tag{20}$$

where f denotes the number of filters, W_k refers to the weight of the k^{th} kernel filter, and e_{t-k} signifies the input element at time step $t - k$. The padding is causal to prevent information leakage from future time steps. The classification block contains a Dropout layer and a Dense layer. It uses a $softmax$ activation function and O_{size} neurons to provide probability scores of possible classes according to:

$$E_1 = Dropout(r)(R_t), \tag{21}$$

$$E_o = softmax(W^4_{ijk} E_1);\ \ W^4_{ijk} \in \mathbb{R}^{b\times O_{size}}. \tag{22}$$

E. Categorical Cross Entropy Loss

The proposed model uses the categorical cross-entropy function to calculate the loss. From an architectural point of view, predicted scores from the model's output are first passed through a $softmax$ operation. Then, the cross-entropy loss is calculated as illustrated in Figure 5.

Let's assume t_i and p_i are the base truth and model's score for i in the C classes, respectively. The categorical cross-entropy function includes a $softmax$ operation before calculating the cross-entropy loss (CE). If $f(p_i) = \frac{e^{p_i}}{\sum_j^C e^{p_i}}$ is the output of the $softmax$ operation, then CE can be formulated as:

$$CE = \sum_i^C t_i log(\frac{e^{p_i}}{\sum_j^C e^{p_i}}). \tag{23}$$

The labels are one-hot for multi-class classification, so only the positive class C_{pos} keeps its term in the loss. There is only one element of the target vector t, which is non-zero $(t_i = t_{pos})$. Thus, CE can be reduced as:

$$CE = -log(\frac{e^{p_{pos}}}{\sum_j^C e^{p_j}}). \tag{24}$$

The model needs to calculate its gradient for the output neurons to optimise the loss function and facilitate the back-propagation operation. Therefore, the derivative of the loss function with respect to the positive class is formulated as:

$$\frac{\delta}{\delta P_{pos}}\left(-log(\frac{e^{p_{pos}}}{\sum_j^C e^{p_j}})\right) = \left(\frac{e^{p_{pos}}}{\sum_j^C e^{p_j}} - 1\right). \tag{25}$$

Again, its derivative concerning the negative class has the following form:

$$\frac{\delta}{\delta P_{neg}}\left(-log(\frac{e^{p_{neg}}}{\sum_j^C e^{p_j}})\right) = \left(\frac{e^{p_{pos}}}{\sum_j^C e^{p_j}}\right). \tag{26}$$

IV. Experiments

An NVIDIA DGX Station with four NVIDIA Tesla $V100$ Tensor Core GPUs and $128GB$ of RAM was used for training the proposed model. The CUDA (version 11.2) was used for computing on the GPUs, and the open-source Tensorflow machine learning platform was used to develop the implementation code.

A. Preprocesssing

The input raw audio signals are preprocessed by (i) separating the percussive components such as background noise, (ii) removing silent and near-silent components from the edges based on the decibel threshold, (iii) energy normalization to ensure the same maximum amplitude of all the signals, (iv) audio augmentation by using the pitch shifting technique, and (v) applying symmetric zero-padding with centring when the length of the original data samples was greater or less than the expected sample, i.e., $sampling\ rate * duration$.

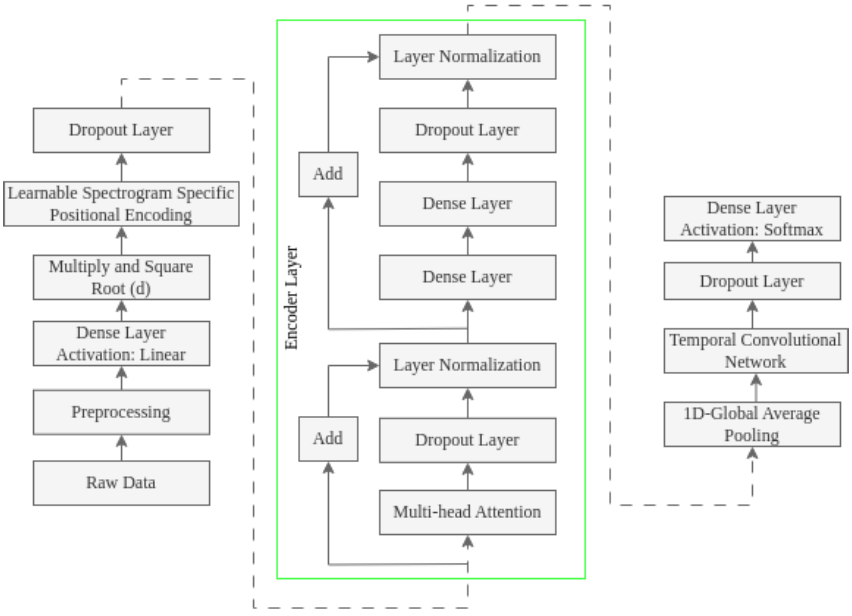

Fig. 4. Architecture of the proposed model for multi-label classification.

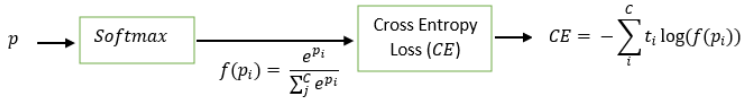

Fig. 5. Architecture of the adopted loss function: The model's scores are passed through a $softmax$ operation followed by a traditional cross-entropy function.

AWGN was added to the test datasets to examine the model's robustness using [38]:

$$R_{Signal} = \sqrt{\frac{\sum Signal^2}{N}}, \tag{27}$$

$$R_{Noise} = \sqrt{\frac{R_{Signal}^2}{10^{(SNR/10)}}}, \tag{28}$$

$$Noise = Gaussian(\mu,\ \sigma,\ number\ of\ samples), \tag{29}$$

$$Signal = Signal + Noise, \tag{30}$$

where R_{Signal} and R_{Noise} represent the root mean squared value of the signal and noise, respectively. The values of μ and σ were assumed to be 0 and R_{Noise}, respectively.

After that, the signals were transformed into Mel-spectrograms with a sampling rate, hop length, number of generated Mel bands, and Fast Fourier Transform (FFT) component numbers of 44100, 300, 128, and 2560, respectively. The lowest and highest frequencies were 20Hz and $(sampling\ rate/2)$. The output of the preprocessing step was a three-dimensional (3D) array of the form: (number of samples, time steps, Mel bands).

B. Implementation

The scaled dot product attention was developed according to the formulation in Section III-A using the *tf.matmul*, *tf.mat.sqrt* and *tf.nn.softmax* packages. The multi-head attention mechanism was built by defining a 'class' called 'MultiHeadAttention' using the *tf.keras.layers.Layer* package according to the formulation presented in Section III-B, combining the scaled dot product attention.

The proposed model consists of two distinct parts: (i) an encoder block with four identical attention-based encoder layers and (ii) an output block containing a TCNs block, a Dropout, and a Dense layer, as described in Section III-D. The *tf.keras.layers.Layer* package proved its worthiness in developing the encoder layers based on the proposed attention mechanism. The dimension of the input sequence was reduced using a DLP layer according to the formulation presented in Section III-C.1. After that, it was fed into a SPE layer built as described in Section III-C.2, followed by a Dropout layer. The output of the Dropout layer and an encoder mask, built to let the model know the actual data location, were finally fed into the encoder layer as inputs using the *tf.keras.Input* package.

The final phase of the code implementation contained the model's training and testing tools. The values and labels of features from the Mel-spectrogram were entered into the model for training. The categorical cross-entropy function of *tf.keras.losses.CategoricalCrossentropy* was used to calculate the loss and the adaptive moment (Adam) estimate of *tf.keras.optimizers.Adam* was used as an optimiser of the proposed model. Accuracy, recall and F_1 score of *sklearn.metrics*

were used to evaluate the proposed model. On the other hand, the Friedman Chi-Square statistical test mechanism was developed using the *scipy.stats.friedmanchisquare* module.

C. Dataset

This study used an open benchmark IDMT-TRAFFIC dataset [39] from Fraunhofer Institute for Digital Media Technology in Germany. It contains $17,506$ stereo audio snippets, each lasting two seconds. It also includes recorded vehicle passing and a variety of street-side background noise. The collection contains recordings from four distinct recording locations, four distinct vehicle types: bus, car, motorcycle, and truck, three different time limit areas, and dry and wet weather/road conditions, including movement direction. The recordings were made using high-quality $sE8$ small-diaphragm condenser microphones and medium-quality Microelectromechanical Systems (MEMS) microphones. The complete dataset was divided into training and testing sets using a ratio of 80:20 to meet the demands.

Another state-of-the-art Vehicle Interior Sound Classification (VISC) dataset [40] was used to examine the proposed model's generalizability. These data samples were collected from the driving point of view of eight different vehicle types: bus, minibus, pick-up truck, sports car, jeep, truck, crossover, and YouTube car. It contains $5,980$ samples, each lasting three to five seconds with $48KHz$ frequency. The dataset was divided into training and testing subsets using a ratio of 80:20 to verify the model's generalisation ability.

D. Model Training

The proposed model and baselines were trained using an NVIDIA DGX Station $V100$ for up to 100 epochs. It contains 1.86 to $4.5M$ trainable parameters depending on different hyperparameters. On the other hand, models based on CNN and LSTM networks only required $15K - 98K$ and $140K - 298K$ parameters for training. The choice of different hyperparameters profoundly impacted the model's performance. An appropriate set of hyperparameters was obtained after randomly choosing their values and performing many trial-and-error tests. In the preprocessing step, for Mel-spectrograms construction, the following sets of hyperparameters proved their worthiness: sampling rate, hop length, number of generated Mel bands, lowest frequency, and FFT component values of 44100, 300, 128, $20Hz$, and 2560, respectively. The following combination of hyperparameters during model training provided the best overall outcomes: $num\_head = 4$, $num\_layer = 4$, $units = 1024$, $p = 0.1$, $d = 128$, $lr = 0.0001$, and $time\_steps = 295$, where num_head, num_layer, $units$, p, d and lr represent the number of attention heads, number of encoder layers, number of neurons of the Dense layers, dropout rate, dimensionality of the representations used as input to the multi-head attention, and learning rate of the model, respectively. The proposed model took $1.5 - 2.0$ hours to complete the training for up to 100 epochs using the aforementioned computational platform. On the other hand, 1D-CNN and LSTM models took 1.2 and 1.5 hours, respectively, to finish training for the same number of epochs. In addition to Dropout layers, the k-fold cross-validation with $k = 5$ and EarlyStopping mechanism with a $patience = 10$ were used to address the overfitting problems.

E. Results

The proposed model and the two baselines under study were trained on the same system using the same datasets according to the aforementioned settings. It demonstrated good performance improvements on all used metrics compared to these baselines:

- 1D-CNN [16], [17]: The authors proposed a 1D-CNN architecture for environmental sound classification problem. A similar model was developed, trained and evaluated on the two used datasets.
- LSTM [18], [19]: The effectiveness of LSTMs in capturing long-term temporal dependencies was used to classify urban sounds. An identical model was implemented, trained, and evaluated using the same experimental setup.

1) Evaluation Metrics: Precision, recall, and F_1 score were used as evaluation metrics to assess the performance of the proposed model relative to the selected baselines. The first step was to get all the individual predictions in the test dataset. Assuming that the test dataset size was N, and the true and predicted labels were T_l and P_l, respectively. The model calculated the confidence scores for all labels. Then, the confidence scores were sorted. The True Positives (TP) and Negatives (TN) are the instances where the actual and predicted labels are positive and negative, respectively. On the other hand, False Positives (FP) and Negatives (FN) are the instances where the predicted labels are positive, but the actual labels are negative, and vice versa. The accuracy score reveals which fraction was predicted correctly among all the label predictions and calculates the fraction between TP and $TP + FP$. Their formulations are [41]:

$$Precision = \frac{TP}{TP + FP}, \tag{31}$$

$$Recall = \frac{TP}{TP + FN}, \tag{32}$$

$$F_1 = \frac{2(Precision * Recall)}{Precision + Recall}. \tag{33}$$

For multi-class classification tasks, the final F_1 score is the average of the F_1 score of each class, with weighting depending on the 'Average' parameter. Let's assume that y and $\hat{y}$ are the set of true and predicted pairs, respectively. Then, for the Micro Average, the employed formula is:

$$Precision(y, \hat{y}) := \frac{|y \cap \hat{y}|}{|y|}. \tag{34}$$

If L, y_l, and $\hat{y}_l$ represent the set of labels, the subset of y with label $l \epsilon L$, and the subset of $\hat{y}$ with label $l \epsilon L$, respectively, then the employed formula for Macro Average is:

$$Precision(y, \hat{y}) := \frac{1}{|L|} \sum_{l \epsilon L} Precision(y_l, \hat{y}_l). \tag{35}$$

TABLE I
PERFORMANCE COMPARISON OF THE PROPOSED MODEL AND THE BASELINES IN TERMS OF F_1 SCORE (BEST VALUES IN BOLD)

F_1 Score	IDMT			VISC		
	Proposed	CNN	LSTM	Proposed	CNN	LSTM
Macro Average	**76.19**	56.38	67.25	**93.88**	91.74	92.78
Micro Average	**94.00**	92.75	93.37	**94.14**	92.47	93.48
Weighted Average	**93.58**	89.96	91.28	**94.14**	92.56	93.46

TABLE II
PERFORMANCE COMPARISON OF THE PROPOSED MODEL AND THE BASELINES AS OF RECALL SCORE (BEST VALUES IN BOLD)

Recall Score	IDMT			VISC		
	Proposed	CNN	LSTM	Proposed	CNN	LSTM
Macro Average	**74.04**	56.87	64.47	**93.82**	91.65	93.00
Micro Average	**94.00**	92.75	93.38	**94.15**	92.47	93.48
Weighted Average	**94.00**	92.75	93.38	**94.15**	92.47	93.48

Lastly, the employed formula for Weighted Average is:

$$Precision(y, \hat{y}) := \frac{1}{\sum_{l \epsilon L} |y_l|} \sum_{l \epsilon L} |y_l| Precision(y_l, \hat{y}_l). \tag{36}$$

Tables I, II, and III summarise the results obtained for the proposed model and the two considered baselines. As it is a multi-class classification problem, the evaluation metrics considered three approaches: the macro average, which represents the unweighted mean of the metrics computed independently for each label; the micro average, which calculates metrics globally by aggregating the total counts of TPs, TNs, FPs, and FNs; and the weighted average, which computes the mean of the metrics for each label weighted by the number of instances for each label in the test dataset. The proposed model achieved a weighted average score of F_1 of 93.58 and 94.14% on the IDMT and VISC test datasets, which are 3.87 and 2.46%, and 1.68 and 0.72% higher than the corresponding scores from the 1D-CNN and LSTM baselines, respectively.

The proposed model surpassed the baselines under study by achieving a weighted average *Recall Score* of 94.00 and 94.15% on the two used datasets. Concerning statistics, an improvement of 1.33 and 1.78% compared to the 1D-CNN model was achieved. Likewise, considering the weighted mean *Precision Score*, the proposed model improved by 1.84% on the IDMT test dataset compared to the LSTM-based baseline.

Table IV summarises the comparison of the proposed model with the baselines under consideration regarding the accuracy of the IDMT and VISC test datasets, consisting of 1751 and 897 audio samples, respectively. It achieved an accuracy of 94.80 and 95.87%, while for LSTM and 1D-CNN models, the accuracy was 92.74 and 93.37%, and 92.47 and 92.68%, respectively. Regarding statistics, improvements from 2.17 to 1.51% and from 3.55 to 3.33% were achieved compared to the considered baselines on the two used test datasets.

2) Statistical Test: The main idea of performing statistical tests is to determine whether the prediction distribution of a model poses any significant difference in statistics concerning some specific property. A null hypothesis was defined, assuming no difference between the model's predictions and choosing a significance level of 5%. Using the Friedman Chi-Square [42] as the test statistic, the outputs for the VISC and IDMT datasets were: ($statistic = 7.36$; $p = 0.025$) and ($statistic = 49.77$; $p = 1.55e^{-11}$), respectively. Since the p-value was less than the significance level, it can be argued that the null hypothesis is invalid, and there are significant statistical differences between the model's predictions.

V. DISCUSSION

Monitoring traffic acoustic events is one of the most challenging tasks in ML due to its diverse acoustic characteristics. The audio clips are recorded using microphones where the target events are usually far away. Therefore, there are often overlapped simultaneous events, and the pressure/intensity of the target sound event may be less than the background/ambient noise. Consequently, traditional ML models fail to provide reliable performance, but DL models are a sought-after solution to these problems. Thus, models based on CNN and LSTM are the current state-of-the-art to solve the problems of monitoring acoustic events efficiently [43], [44], [45], [46]. However, these models demonstrate difficulties dealing with external noise, variation in sound intensities, and capturing long-range audio features; therefore, further improvements are needed.

Consequently, this study proposes a hybrid model for monitoring road traffic acoustic events using a multi-head attention-based transformer in combination with TCNs to learn long-range dependencies between audio features and overcome the above-mentioned challenges. Additionally, it introduces a unique preprocessing mechanism and presents a new DLP approach to downscale the feature dimensions before further processing. It also proposes a learnable SPE layer to capture spectrogram-specific positional information efficiently. It comprises an encoder block containing four identical encoder layers for deeply encoding the preprocessed input. The TCNs block handled the deeply encoded output to capture short and long-term dependencies. The classification block contains a Dense layer with the $softmax$ activation function, preceded by a Dropout layer to produce the probability scores of the predicted labels.

The state-of-the-art IDMT-TRAFFIC dataset, containing 17, 506 stereo audio clips, was used for model training and testing. It contains audio samples of five distinct classes, mainly bus, car, motorbike, and truck. Thus, this study

TABLE III
PERFORMANCE COMPARISON OF THE PROPOSED MODEL AND THE BASELINES REGARDING PRECISION SCORE (BEST VALUES IN BOLD)

Precision Score	IDMT			VISC		
	Proposed	CNN	LSTM	Proposed	CNN	LSTM
Macro Average	79.38	56.13	**81.69**	**94.08**	92.29	93.05
Micro Average	**94.00**	92.75	93.38	**94.15**	92.47	93.48
Weighted Average	**93.39**	87.49	91.67	**94.26**	93.01	93.82

TABLE IV
PERFORMANCE COMPARISON OF THE PROPOSED MODEL AND THE BASELINES WITH THE BASELINES ON THE TEST DATASET (BEST VALUES IN BOLD)

Models	**Accuracy (IDMT)** (%)	**Accuracy (VISC)** (%)
Proposed	**94.80**	**95.87**
1D-CNN	92.74	92.47
LSTM	93.37	92.68

TABLE V
PERFORMANCE OF THE PROPOSED MODEL WITH DIFFERENT LOUDNESS LEVELS ON THE VISC DATASET (BEST VALUE IN BOLD)

Mean Loudness (dB)	**Accuracy** (%) **(Proposed)**
Original (-17.64)	**95.87**
-25.66	95.83
-28.76	95.15
-31.68	91.78
-36.12	92.14

modelled the dataset as a multi-class classification problem with five classes: bus, car, motorcycle, truck, and non-vehicle. Another dataset called VISC, which contains 5890 audio samples of eight classes: bus, minibus, pick-up truck, sports car, jeep, truck, crossover, and car (automobile), was used to verify the model's generalisation ability. The two datasets were split into training and testing datasets. Many researchers convert multichannel audio signals to a single channel before further processing [47], resulting in poor noise robustness. Therefore, this study directly converted the stereophonic audio signals into Mel-spectrograms following the preprocessing steps to overcome this drawback. The categorical cross-entropy function was used as the loss function to facilitate the proper learning of the proposed model.

The results of the proposed model were compared with previously published 1D-CNN and LSTM-based models. State-of-the-art metrics were used to evaluate its performance, mainly $Precision$, $Recall$, F_1 score, and accuracy. Also, the Friedman Chi-Square statistical test was performed to determine the differences in the model's prediction distributions regarding statistics. Figures 6 and 7 represent the confusion matrices obtained for the models based on the IDMT and VISC test datasets, demonstrating the superior performance of the proposed model relative to the baselines under study. With a highly imbalanced dataset like IDMT, it achieved a higher accuracy score than its counterparts.

The impact of loudness and environmental noise on the model's performance was also examined. The amplitude was adjusted to achieve various loudness levels [48], and the loudness values were computed using the $ITU\ BS.1770$ algorithm [49]. The proposed model performed well at different loudness levels, as shown in Table V. The original audio data samples of the VISC dataset possess a mean loudness of $-17.64\ dB$, which led to the best accuracy score. The proposed model was also tested with four different mean loudness levels: $-25.66\ dB$, $-28.76\ dB$, $-31.68\ dB$, and $-36.12\ dB$, which were obtained by changing the amplitude of the original signals by a factor of 30%, 40%, 50%, and 70%, respectively. The accuracy decreased as the loudness from the alignment level (around $-20\ dB$) shifted further, except for $-36.12\ dB$.

TABLE VI
ACCURACY COMPARISON OF THE MODELS UNDER STUDY WITH DIFFERENT SNR LEVELS ON THE IDMT TEST DATASET (BEST VALUE IN BOLD)

Mean SNR (dB)	**Proposed** (%)	**CNN** (%)	**LSTM** (%)
-30	46.60	47.92	34.27
-20	46.60	58.99	40.32
Original (0.0)	94.80	92.74	93.37
10	93.49	92.06	92.58
20	**95.26**	92.00	92.92
30	95.09	92.06	92.92

AWGN, which can mimic natural random processes, was added to the original test datasets to examine the model's robustness on various noise levels. The proposed model exhibited good resilience against various Signal-to-Noise Ratio (SNR) levels, particularly for higher positive values, as shown in Table VI. Even more interesting was that the model outperformed the original audio signals in accuracy at a mean SNR of $-20\ dB$. However, it was also observed that the accuracy decreased significantly with negative SNR values.

Several ablation experiments were conducted to examine the effects of various structural designs and parameters on the proposed model, as shown in Table VII. Thus, the model's response was investigated without the proposed preprocessing mechanism, learnable SPE layer, TCNs block, and DLP mechanism. Without the proposed SPE layer, the model suffered a 9.29% reduction in prediction accuracy. Also, without the TCNs, its accuracy decreased by 3.71% from the original value. The effect of the number of encoder layers (4, 8 and 12) on its performance was also tested, indicating that a higher number of layers might not increase the performance. The computational cost also varied based on these attributes.

Dropout layers with a rate of r were used to avoid the model's overfitting. It is accomplished by randomly setting input units to 0 (zero) with a particular frequency specified by r at each step during training. The model was trained and tested with different values of r to observe its effect. It was found that gradually increasing its value from 0.1 to 0.3 decreased the model's accuracy on the IDMT test dataset.

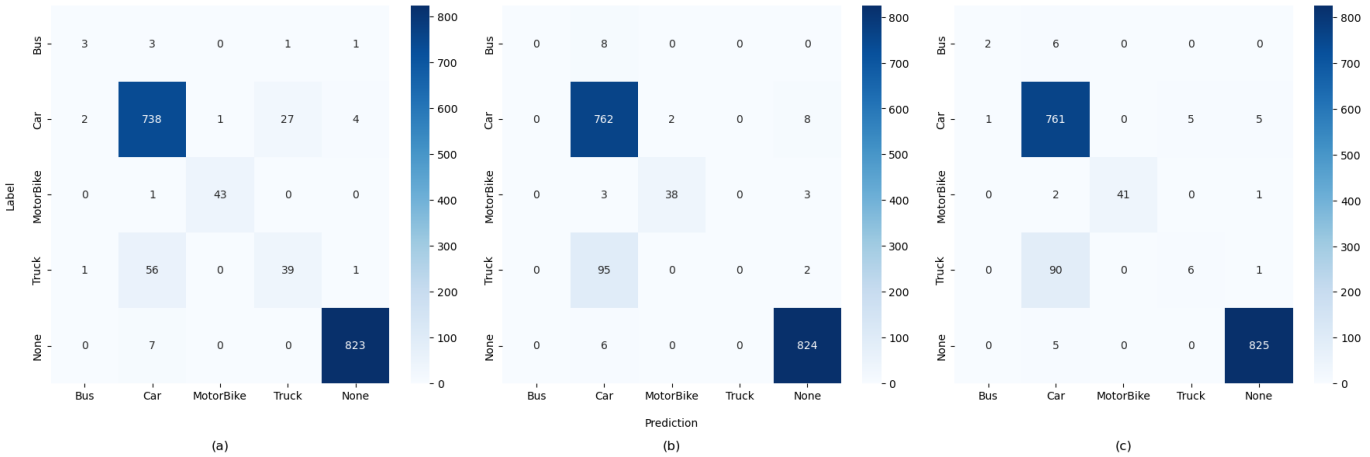

Fig. 6. Confusion matrices of the (a) proposed (b) 1D-CNN and (c) LSTM models on the IDMT dataset.

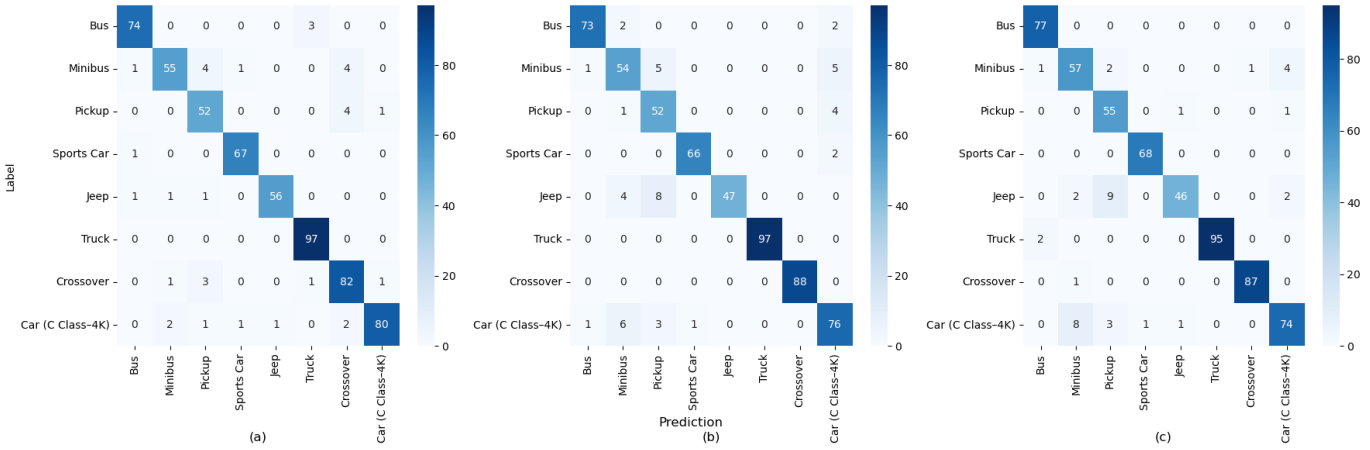

Fig. 7. Confusion matrices of the (a) proposed, (b) 1D-CNN, and (c) LSTM models on the VISC dataset.

TABLE VII

ABLATION STUDY ON THE VISC DATASET (W AND W/O REPRESENT WITH AND WITHOUT, BEST VALUES IN BOLD)

Attribute	Parameters (M)	Accuracy (%)
Original	1.86	**95.87**
w/o Preprocessing	1.86	91.81
w/o SPE	1.86	86.96
w/o TCNs	**1.32**	92.31
w/o DLP	1.85	87.79
w 8 layers	3.18	95.31
w 12 layers	4.50	91.81

This study also examined the $sigmoid$ and $softmax$-based attention mechanisms to justify their influences on the model's performance. The experimental results demonstrated the superior results of the $softmax$-based attention mechanism by improving the model's accuracy by 6.24% on the IDMT test dataset compared to its counterpart. The model presented in [12] led to an opposite conclusion, i.e., more effective performance of the $sigmoid$-based attention mechanism for sound classification tasks. The opposite finding may be attributed to using layered attention with CNNs instead of a full attention mechanism.

Gong et al. [20] suggested that a transformer model can only outperform CNN-based models with over $14M$ worth of training samples. However, publicly available audio datasets in this domain contain a much lower number of data samples. Therefore, the transfer learning approach, i.e., using a pre-trained model, is necessary to achieve superior results. The proposed model did not require any transfer learning technique and was trained on the IDMT dataset that contains only $17K$ samples. It still achieved superior performance compared to the CNN-based model under study. Although the proposed model may suffer from the overfitting problem, the previously mentioned requirement of training data volume for transformer models is problem-specific, which should be explored further.

Comprehensive comparisons with the considered baselines demonstrated the quality of the proposed model by outperforming them in most commonly used metrics. However, it has several shortcomings. It was modelled to classify only five classes, namely, Bus, Car, Motorbike, Truck, and non-vehicle, in the IDMT dataset, and eight classes were considered for the VISC dataset. A real road traffic environment may contain more classes, such as bicycles and ambulances. The problem is that publicly available state-of-the-art traffic audio datasets are

sparse compared to general-purpose acoustic datasets. Therefore, more attention should be given to building good-quality datasets in this area. Also, only AWGN was considered for examining the robustness of the model. Mixing the original signals with real-world noises for further robustness testing also needs to be considered. Another disadvantage of the proposed model is the computational cost and implementation in real-life practice. The 1D-CNN model only required $15K$ parameters for training and can be better suited for some real-world deployments than the proposed model.

VI. Conclusion

Models for monitoring traffic acoustics events can be crucial to improve the performance of ITS because video-based models lack efficiency in cases of occlusion, low light, snowy, and rainy conditions. Therefore, the research community has recently been focused on developing solutions based on acoustics. However, the current state of the art, mainly based on LSTM and CNN models, needs further improvement in capturing long-term relationships between different audio features to tackle background noise and improve efficiency. Therefore, this study presented a hybrid model combining the attention mechanism and the TCNs to solve these problems efficiently. It proposed a learnable SPE layer suitable for capturing spectrogram-specific positional information and a DLP mechanism to downscale feature values before passing them to the SPE layer. Attention mechanisms based on the $softmax$ function constituted the encoding block with four identical encoding layers to encode the preprocessed inputs. The deeply encoded output from the encoder block was processed by the TCNs block to extract short and long-term relationships between the features and a classification block to output the prediction distribution of the predicted classes. Two state-of-the-art baseline models were used for comprehensive performance comparison, and two benchmark datasets were used to train and evaluate the models. The results demonstrated the superior performance of the proposed model, achieving an accuracy improvement of up to 3.55% compared to the studied baselines. In the future, the proposed model will be tested on other state-of-the-art datasets containing more vehicle classes. Also, further robustness tests will be performed by mixing real-world noise with the original signals. In addition, research will be undertaken to study and improve real-world deployment skills.

References

[1] L. Zhu, R. Krishnan, A. Sivakumar, F. Guo, and J. W. Polak, "Traffic monitoring and anomaly detection based on simulation of Luxembourg road network," in *Proc. IEEE Intell. Transp. Syst. Conf. (ITSC)*, Oct. 2019, pp. 382–387.

[2] Y. Zhang, H. Zhao, Y. Li, Y. Long, and W. Liang, "Predicting highly dynamic traffic noise using rotating mobile monitoring and machine learning method," *Environ. Res.*, vol. 229, Jul. 2023, Art. no. 115896.

[3] A. Pascale, E. Macedo, C. Guarnaccia, and M. C. Coelho, "Smart mobility procedure for road traffic noise dynamic estimation by video analysis," *Appl. Acoust.*, vol. 208, Jun. 2023, Art. no. 109381.

[4] K. Imoto, S. Mishima, Y. Arai, and R. Kondo, "Impact of data imbalance caused by inactive frames and difference in sound duration on sound event detection performance," *Appl. Acoust.*, vol. 196, Jul. 2022, Art. no. 108882.

[5] H.-Y. Lai et al., "Mel-scale frequency extraction and classification of dialect-speech signals with 1D CNN based classifier for gender and region recognition," *IEEE Access*, vol. 12, pp. 102962–102976, 2024.

[6] K. K. Khine and C. Su, "Acoustic scene classification using deep C-RNN based on log mel spectrogram and gammatone frequency cepstral coefficients features," in *Proc. 3rd Int. Conf. Artif. Intell. Internet Things (AIIoT)*, May 2024, pp. 1–6.

[7] X. Zhuang, J. Huang, G. Potamianos, and M. Hasegawa-Johnson, "Acoustic fall detection using Gaussian mixture models and GMM supervectors," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Apr. 2009, pp. 69–72.

[8] A. Temko and C. Nadeu, "Classification of acoustic events using SVM-based clustering schemes," *Pattern Recognit.*, vol. 39, no. 4, pp. 682–694, Apr. 2006.

[9] H. Phan, M. Maaß, R. Mazur, and A. Mertins, "Random regression forests for acoustic event detection and classification," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 1, pp. 20–31, Jan. 2015.

[10] X. Xia, R. Togneri, F. Sohel, Y. Zhao, and D. Huang, "A survey: Neural network-based deep learning for acoustic event detection," *Circuits, Syst., Signal Process.*, vol. 38, no. 8, pp. 3433–3453, Aug. 2019.

[11] J. Li, X. Wang, Z. Tu, and M. R. Lyu, "On the diversity of multi-head attention," *Neurocomputing*, vol. 454, pp. 14–24, Sep. 2021.

[12] Z. Zhang, S. Xu, S. Zhang, T. Qiao, and S. Cao, "Attention based convolutional recurrent neural network for environmental sound classification," *Neurocomputing*, vol. 453, pp. 896–903, Sep. 2021.

[13] Y. Guan, J. Xue, G. Zheng, and J. Han, "Sparse self-attention for semi-supervised sound event detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 821–825.

[14] V.-T. Tran and W.-H. Tsai, "Acoustic-based train arrival detection using convolutional neural networks with attention," *IEEE Access*, vol. 10, pp. 72120–72131, 2022.

[15] J. Du, X. Qiao, Z. Yan, H. Zhang, and W. Zuo, "Flexible image denoising model with multi-layer conditional feature modulation," *Pattern Recognit.*, vol. 152, Aug. 2024, Art. no. 110372.

[16] S. Abdoli, P. Cardinal, and A. L. Koerich, "End-to-end environmental sound classification using a 1D convolutional neural network," *Expert Syst. Appl.*, vol. 136, pp. 252–263, Dec. 2019.

[17] M. Bubashait and N. Hewahi, "Urban sound classification using DNN, CNN & LSTM a comparative approach," in *Proc. Int. Conf. Innov. Intell. Informat., Comput., Technol. (3ICT)*, Sep. 2021, pp. 46–50.

[18] J. K. Das, A. Ghosh, A. K. Pal, S. Dutta, and A. Chakrabarty, "Urban sound classification using convolutional neural network and long short term memory based on multiple features," in *Proc. 4th Int. Conf. Intell. Comput. Data Sci. (ICDS)*, Oct. 2020, pp. 1–9.

[19] I. Lezhenin, N. Bogach, and E. Pyshkin, "Urban sound classification using long short-term memory neural network," in *Proc. Federated Conf. Comput. Sci. Inf. Syst. (FedCSIS)*, Sep. 2019, pp. 57–60.

[20] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio spectrogram transformer," 2021, *arXiv:2104.01778*.

[21] C. Wang et al., "Real-time vehicle sound detection system based on depthwise separable convolution neural network and spectrogram augmentation," *Remote Sens.*, vol. 14, no. 19, p. 4848, Sep. 2022.

[22] T. Kim, J. Lee, and J. Nam, "Comparison and analysis of SampleCNN architectures for audio classification," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 2, pp. 285–297, May 2019.

[23] E. Law, K. West, M. I. Mandel, M. Bay, and J. S. Downie, "Evaluation of algorithms using games: The case of music tagging," in *Proc. 10th Int. Soc. Music Inf. Retr. Conf.*, Oct. 2009, pp. 387–392.

[24] X. Zhang, Y. Chen, M. Liu, and C. Huang, "Acoustic traffic event detection in long tunnels using fast binary spectral features," *Circuits, Syst., Signal Process.*, vol. 39, no. 6, pp. 2994–3006, Jun. 2020.

[25] L. Marchegiani and P. Newman, "Listening for sirens: Locating and classifying acoustic alarms in city scenes," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 10, pp. 17087–17096, Oct. 2022.

[26] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[27] S.-H. Lee, J.-W. Hwang, M.-H. Song, and H.-M. Park, "A method based on dual cross-modal attention and parameter sharing for polyphonic sound event localization and detection," *Appl. Sci.*, vol. 12, no. 10, p. 5075, May 2022.

[28] A. Politis, S. Adavanne, D. Krause, A. Deleforge, P. Srivastava, and T. Virtanen, "A dataset of dynamic reverberant sound scenes with directional interferers for sound event localization and detection," 2021, *arXiv:2106.06999*.

[29] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proc. 22nd ACM Int. Conf. Multimedia*, Nov. 2014, pp. 1041–1044.

[30] K. J. Piczak, "ESC: Dataset for environmental sound classification," in *Proc. 23rd ACM Int. Conf. Multimedia*, Oct. 2015, pp. 1015–1018.

[31] R. Serizel, N. Turpault, A. Shah, and J. Salamon, "Sound event detection in synthetic domestic environments," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 86–90.

[32] Ç. Bilen, G. Ferroni, F. Tuveri, J. Azcarreta, and S. Krstulovic, "A framework for the robust evaluation of sound event detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 61–65.

[33] C.-Y. Koh, Y.-S. Chen, Y.-W. Liu, and M. R. Bai, "Sound event detection by consistency training and pseudo-labeling with feature-pyramid convolutional recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 376–380.

[34] Y. Lian, J. Wang, Z. Li, W. Liu, L. Huang, and X. Jiang, "Residual attention guided vision transformer with acoustic-vibration signal feature fusion for cross-domain fault diagnosis," *Adv. Eng. Informat.*, vol. 64, Mar. 2025, Art. no. 103003.

[35] P. Baldi and P. J. Sadowski, "Understanding dropout," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 26, Dec. 2013, pp. 2814–2822.

[36] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *Stat*, vol. 1050, p. 21, Jul. 2016.

[37] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," 2016, *arXiv:1606.08415*.

[38] G. Baldini, I. Amerini, and C. Gentile, "Microphone identification using convolutional neural networks," *IEEE Sensors Lett.*, vol. 3, no. 7, pp. 1–4, Jul. 2019.

[39] J. Abeßer, S. Gourishetti, A. Kátai, T. Clauß, P. Sharma, and J. Liebetrau, "IDMT-traffic: An open benchmark dataset for acoustic traffic monitoring research," in *Proc. 29th Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2021, pp. 551–555.

[40] E. Akbal, T. Tuncer, and S. Dogan, "Vehicle interior sound classification based on local quintet magnitude pattern and iterative neighborhood component analysis," *Appl. Artif. Intell.*, vol. 36, no. 1, Dec. 2022, Art. no. 2137653.

[41] M. Grandini, E. Bagli, and G. Visani, "Metrics for multi-class classification: An overview," 2020, *arXiv:2008.05756*.

[42] O. Rainio, J. Teuho, and R. Klén, "Evaluation metrics and statistical tests for machine learning," *Sci. Rep.*, vol. 14, no. 1, p. 6086, Mar. 2024.

[43] Y. Wu and T. Lee, "Enhancing sound texture in CNN-based acoustic scene classification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 815–819.

[44] M. Chavdar, B. Gerazov, Z. Ivanovski, and T. Kartalov, "Towards a system for automatic traffic sound event detection," in *Proc. 28th Telecommun. Forum (TELFOR)*, 2020, pp. 1–4.

[45] C.-C. Kao, M. Sun, W. Wang, and C. Wang, "A comparison of pooling methods on LSTM models for rare acoustic event classification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 316–320.

[46] S. Mohine, B. S. Bansod, R. Bhalla, and A. Basra, "Acoustic modality based hybrid deep 1D CNN-BiLSTM algorithm for moving vehicle classification," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 9, pp. 16206–16216, Sep. 2022.

[47] L. Lu, H.-J. Zhang, and H. Jiang, "Content analysis for audio classification and segmentation," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 7, pp. 504–516, Oct. 2002.

[48] R. R. Devi and D. Pugazhenthi, "Ideal sampling rate to reduce distortion in audio steganography," *Proc. Comput. Sci.*, vol. 85, pp. 418–424, Jun. 2016.

[49] S. Norcross, S. Nanda, and Z. Cohen, "ITU-R BS. 1770 based loudness for immersive audio," Dolby Laboratories, San Francisco, CA, USA, Tech. paper 9500, 2016. Accessed: Jul. 2025. [Online]. Available: https://aes2.org/publications/elibrary-page/?id=18199

Selim Reza is currently pursuing the Ph.D. degree with the Department of Informatics Engineering, Faculty of Engineering, University of Porto, Portugal. His Ph.D. study was supported through the Ph.D. Research Grant from the Fundação para a Ciência e Tecnologia (FCT), with reference 2022.12391.BD. He is the first author of six Web of Science-indexed journal articles. His current research focuses on developing deep learning models for enhancing the efficiency of intelligent transportation systems, such as multi-step regression models for large-scale traffic states and traffic acoustic events monitoring. He also works on integrating lightweight LLMs for traffic acoustic event reasoning and model deployment techniques, such as CI/CD. He is passionate about developing and deploying lightweight deep learning models in his research areas.

Marta Campos Ferreira received the bachelor's degree in economics from the Faculty of Economics, University of Porto, in 2007, and the M.Sc. degree in service engineering and management and the Ph.D. degree in transportation systems from the Faculty of Engineering, University of Porto, in 2010 and 2018, respectively. She is an Assistant Professor with the Department of Industrial Engineering and Management, Faculty of Engineering, University of Porto. She is the Vice Director of the M.Sc. degree in service engineering and management and a member of its scientific committee. She has served as an editor for five volumes of conference proceedings and published over 80 papers in books, conferences, and international journals. She has supervised more than 80 master's students and seven ongoing Ph.D. students. Her research focuses on user-centered mobility systems, participatory design of digital transport solutions, interoperable ticketing, behavioral adaptation, and accessibility. She also works on large-scale mobility data analysis to support decision-making, policy development, service optimization, often addressing complex problems in service integration, resource allocation, and network performance. She is the co-founder of International Symposium on Research and Entrepreneurship and the Congress on Service Engineering and Management (CESG). She has organized and co-chaired several conferences, including the Transport Research Arena (TRA) Lisbon 2022. She is an Associate Editor of *Expert Systems with Applications* and *International Journal of Management and Decision Making*. She is the Co-Founder and the Co-Editor-in-Chief of the collection Research and Entrepreneurship: Making the Leap from Research to Business in SN Discover Applied Sciences.

J.J.M. Machado received the master's degree in mechanical engineering and the Ph.D. degree in mechanical engineering, with a focus on adhesively bonded automotive structures from the University of Porto, in 2012 and 2019 respectively. His doctoral research was conducted in collaboration with Nagase ChemteX and Aston Martin Lagonda. He is an Assistant Professor with the Department of Mechanical Engineering (DEMec), Faculty of Engineering, University of Porto (FEUP), where he has been a Faculty Member since October 2021. From 2012 to 2016, he was a Mechanical Engineer, specializing in electric motors for hazardous areas with Oil and Gas industry, and in the maintenance of mining equipment. Then, he moved into academia, joined the Institute of Science and Innovation in Mechanical and Industrial Engineering (INEGI) as a Researcher from 2016 to 2019. He held a post-doctoral position with Tokyo Institute of Technology from 2019 to 2020, furthering his expertise in mechanical engineering and research innovation. As an academician, he has been deeply involved in mentoring, having co-supervised several M.Sc. theses, since 2017. He is also a co-author of more than 60 articles published in national and international peer-reviewed journals. His research contributions extend across multiple projects, where he has worked both as a Researcher and as a Scientific Coordinator. He has also actively contributed to the academic community through his participation with the organization and scientific committees of several international conferences. His primary research interests lie in design and manufacturing, computational vision, and new product development-areas, in which he continues to explore innovative solutions that bridge the gap between theoretical knowledge and real-world applications. He has served as a guest editor for various special issues of high-impact international journals.

João Manuel R. S. Tavares received the degree in mechanical engineering, the M.Sc. and Ph.D. degrees in electrical and computer engineering, and the Habilitation degree in mechanical engineering from the Universidade do Porto (UP), Portugal, in 1992, 1995, 2001, and 2015, respectively. He is a Senior Researcher with the Institute of Science and Innovation in Mechanical and Industrial Engineering (INEGI) and a Full Professor with the Department of Mechanical Engineering (DEMec), Faculdade de Engenharia da Universidade do Porto (FEUP). Since June 2023, he has been the Head of DEMec. He has made significant contributions to his field, serving as a co-editor for over 95 books and co-authoring more than 55 book chapters and 700 articles in international and national journals and conferences. He also holds three international and three national patents. He has also (co-)supervised several M.Sc. and Ph.D. theses and supervised several post-doctoral projects, participating in many scientific projects as both a Researcher and a Scientific Coordinator. His research focuses on computational vision, medical imaging, biomechanics, biomedical engineering, and new product development. He is a Co-Founder and the Co-Chair of several international conference series, including International Symposium on Computational Modeling of Objects Presented in Images, ECCOMAS Thematic Conference on Computational Vision and Medical Image Processing, International Conference on Computational and Experimental Biomedical Sciences, and International Conference on Biodental Engineering. His editorial roles are extensive; he is a committee member of several international and national journals and conferences, the Co-Founder and the Co-Editor of the "Lecture Notes in Computational Vision and Biomechanics" series (Springer), the Founder and the Editor-in-Chief of *Computer Methods in Biomechanics and Biomedical Engineering: Imaging and Visualization* [Taylor & Francis (T&F)], and the Editor-in-Chief of *Computer Methods in Biomechanics and Biomedical Engineering* (T&F). More details are available at www.fe.up.pt/ tavares