改进的车辆子类型分类：用于声学交通监测

Mohd Ashhad计算机科学与工程系 *Jamia Hamdard*新德里，印度
ashhad.ja
miahamdard@gmail.com

Umang Goenka信息技术系 勒克瑙印度信息技术学院勒克瑙，印度
lit2019033@iiitl.ac.in

Aaryan Jagetia信息技术系 勒克瑙印度信息技术学院勒克瑙，印度
lit2019045@iiitl.ac.in

Parwin Akhtari
机械工程系 *NIT Rourkeela*Roorkela，印度
parwinakhtari7@gmail.com

Sooraj K. Ambat
海军物理与海洋学实验室 国防研究与发展组织科钦，印度 sooraj.npol@gov.in

Mary Samuel
数学系 勒克瑙印度信息技术学院勒克瑙，印度
m
arysamuel@iiitl.ac.in

摘要——**道路上车辆的检测与分类是交通监测的一项关键任务。通常，计算机视觉（CV）算法主导着道路车辆分类任务，但计算机视觉方法在光照条件差时可能效果不佳，而且需要更强的计算能力。此外，在敏感和安全区域安装摄像头还会引发隐私顾虑。相比之下，声学交通监测成本低廉，且能提供更高的准确率，尤其是在光线不足以及无法安装摄像头的场所。本文研究声学车辆子类型分类任务，将声学信号分为 4 类：汽车、卡车、摩托车和无车辆。我们采用梅尔频谱图、梅尔频率倒谱系数（MFCC）和 GFCC 作为特征，并进行数据预处理，以训练一个简单且经过充分优化、能够胜任该任务的卷积神经网络（CNN）。使用 MFCC 作为特征并进行精细的数据预处理后，我们提出的方法在 IDMTTraffic 数据集上超越了已有的最先进基线，准确率达到 98.95%**。

索引词—**声学交通监测（ATM）、信号处理、音频分类、机器学习、卷积神经网络（CNN）**

I. 引言

由于城市化进程不断加快及其带来的挑战，近年来，交通监测已成为一个重要的研究领域。它可用于智慧城市中的多种应用，例如控制交通信号灯、检测交通密度、预测和检测道路事故等。声学交通监测（ATM）技术或可帮助解决其他交通监测方法中存在的某些问题。除成本低之外，ATM 方法还有助于保护驾驶员和车辆的隐私；这一点可能至关重要，尤其是在敏感的驻军区，与基于视觉的方法相比更是如此。

声学交通监测（ATM）算法面临着一系列独特的挑战。在声学车辆子类型分类中，仅凭音频来描述车辆极其困难，因为声音是由发动机、轮胎、风湍流以及道路以外的其他因素等不同部分混合而成的，例如道路状况。这有时会对模型产生不利影响，因为它可能会遇到各种复杂情况，其中某个声音成分会占据主导地位。例如，距离麦克风 3 米或 5 米处录制的汽车声音之间存在显著差异。此外，汽车在干燥路面或湿滑路面上行驶时发出的声音也可能存在差异。

本文研究声学车辆子类型分类任务。具体而言，我们探讨一种四类车辆分类场景，其中包括三种车辆类型——汽车、卡车和摩托车——以及一个非车辆类别（包含背景噪声）。本文的主要贡献如下：

- 我们表明，对于声学车辆分类任务，简单且经过良好优化的卷积神经网络，结合简单的特征提取方法，其表现可优于 VGG网络等复杂的最先进神经网络。
- 我们研究了开放基准数据集（IDMT Traffic）中汽车和卡车样本分类准确率较低的原因 [1] ，并进行了人类性能评估，以进一步找出问题所在。
- 最后，我们提出的处理流程在车辆子类型分类任务中表现更好，准确率达到 98.95%；与既定基线相比，卡车类别的 F1分数提高了 42% [1]。

第 II 节简要介绍了当前相关研究所涉及的声学交通监测模型和算法。研究所用的数据集已在第 III 节中介绍。第 IV 节阐述了本文采用的各种方法。第 V 节讨论了研究期间开展的多项实验，并对结果进行了全面分析。第 VI 节总结了这项工作，并对未来研究提出了建议。

II. 相关工作

车辆检测是交通监测中最基础的任务之一。Kazuo 等人 [2] 开发了一种超低功耗声学车辆检测系统，该系统使用两个麦克风。他们的检测器精确率和召回率分别达到 0.94 和 0.95。Shigemi Ishida 等人 [3] 开发了一种稳态噪声抑制方法，以提高车辆检测中的信噪比。他们的方法在正常降雨和暴雨情况下的 F值分别达到 0.92 和 0.90。Billy Dawton 等人 [4] 表明，在车辆检测和分类方面，基于立体声麦克风的系统优于单麦克风系统。因此，大多数情况下更倾向于使用立体声麦克风 [1] [6] 或多通道音频 [1] [7] [5]进行录音。G.Szwoch 等人 [5] 在平行于道路和垂直于道路的两个方向使用传感器，并通过多种特征实现了 0.95 的车辆检测 F1分数。

在车辆类型和子类型分类方面，Amir Y. Nooralahiyan 等人 [6]使用时延神经网络（TDNN）对在利兹市城市道路上录制的音频中的摩托车和轻型货车或厢式货车进行分类。他们取得了令人瞩目的 82.4% 测试准确率。Jobin George 等人 [11] 考虑了一个四分类场景：重型、中型、轻型和喇叭声。他们使用平滑对数能量检测车辆，然后从检测到的峰值周围固定区域提取梅尔频率倒谱系数，以训练人工神经网络，最终取得了 67% 的准确率。An-Chih Yang 等人 [7] 使用智能手机在一处孤立的停车标志处记录车辆加速度，以避免复杂多样的交通状况和迎面驶来的车辆，从而获得了音频清晰的数据集。他们提出的处理流程包括频谱和时域特征提取方法，以及噪声注入和音高变化等数据增强技术；随后将处理后的数据输入多个卷积神经网络模型，将车辆分为以下类别：混合动力车、轿车、皮卡、公交车和商用车。他们取得了 75% 的准确率。Billy Dawton 等人 [4] 使用线性核支持向量机，在踏板车、汽车和公共汽车三分类任务中取得了 95.01% 的准确率。Jakob Abeber 等人 [1] 提供了名为 IDMT Traffic for ATM 的公开数据集。他们使用梅尔频谱图进行车辆子类型分类，并采用 VGG网络和残差网络等先进的卷积神经网络进行分类。然而，他们的分类器在卡车分类方面表现不佳，F1分数仅为 0.5。

许多方法采用计算机视觉，而非声学特征。对于图像，模型可以获取前照灯、车辆设计和测量数据等视觉元素。已有多项研究利用照片或视频帧对车辆类型进行分类 [8] [9]。此外，红外热成像等现代技术 [10]，可根据车辆外部的热特性对车辆类型进行分类。图像能够取得良好效果，但通常成本高昂且难以安装。此外，在敏感区域使用图像还会引发隐私顾虑。

III. 数据集

本研究采用 IDMT- Traffic 数据集 [1]。IDMT Traffic 于 2021年 发布，是一个用于声学交通监测的开放基准数据集。该数据集包含在四个不同录音地点采集的移动汽车同步立体声音频录音，其中三个地点位于伊尔梅瑙及其周边的城市交通地点，一个位于德国伊尔梅瑙及其周边的乡村道路地点；录音使用高品质 sE8 麦克风和价格较低的微机电系统（MEMS）麦克风采集。录音场景涵盖不同限速（30、50 和 70 km/h），以及干燥和湿滑路面状况。数据集包含以下内容：汽车（3903 个事件）、卡车（511 个事件）、公共汽车（53 个事件）和摩托车（251 个事件），以及无车辆经过时的背景录音。

IV. 方法

A. 特征提取与音频预处理

在本研究中，我们将梅尔频谱图、梅尔频率倒谱系数（MFCC）和伽马音调频率倒谱系数（GFCC）作为从原始音频波形中提取主要特征的方法，随后将这些特征输入基于卷积神经网络的分类器。这些方法通常旨在模拟人类听觉感知。我们首先将立体声音频文件的左右声道取平均值，合并为单声道，且采样率保持不变，为 48 kHz；随后在进一步提取特征时，将采样率降至 22.05 kHz。具体实现细节将在以下章节中介绍。

1）梅尔频谱图： 作为我们的第一种特征提取方法，我们使用 Librosa Python 库 [12] 计算了梅尔频谱图。梅尔频谱图通过滤波器组划分频域值，并以对数方式描绘高于预设阈值（拐角频率）的频率，从而计算输出。我们使用了 2048 的快速傅里叶变换大小、1024 的窗长和 512 的帧移。梅尔频带数量设为 128，并采用对数幅度缩放。图 1 展示了数据集中属于不同类别的样本的梅尔功率频谱图。

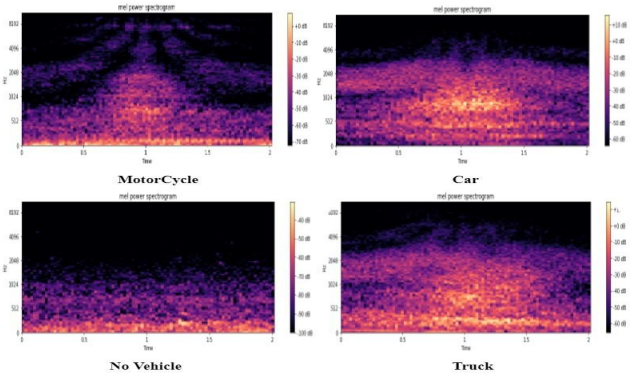


图 1：摩托车、汽车、无车辆和卡车的梅尔频谱图

2）梅尔频率倒谱系数：梅尔频率倒谱由若干个称为梅尔频率倒谱系数（MFCC）的系数组成。这些系数基于音频片段的倒谱表示（一种非线性的“谱的谱”）生成。梅尔频率倒谱与倒谱的区别在于，其频带在梅尔刻度上均匀分布；与传统频谱中采用的线性间隔频带相比，这种分布更接近人类听觉系统的响应。为计算数据集中样本的梅尔频率倒谱系数，我们使用了 Librosa Python 库 [12]。返回的梅尔频率倒谱系数数量设为 128。快速傅里叶变换大小、窗长和帧移均与上一节相同。

3) GFCC：在广泛采用梅尔频率滤波器之后，研究人员根据对听觉外周的心理物理学研究，创建了一个更全面的模型，以模拟耳朵的特性，因为梅尔频率滤波器与人耳模型具有相似性。Patterson 和 Smith [15]创建了 GFCC（伽马音频率倒谱系数）。人类听觉系统由伽马音滤波器组表示为一组相互重叠的带通滤波器。方程 (1) 给出了每个滤波器的脉冲响应。

$$g(t) = at^{n-1}e^{-2\pi bt}cos(2\pi f_c t + \phi) \quad \textbf{(1)}$$

其中，f_c 为中心频率，a为常数（通常等于 1），n为滤波器阶数，ϕ 为相位偏移，b 为带宽。滤波器的等效矩形带宽（ERB）可用于确定中心频率和带宽。我们使用 Spafe Python 库 [13]提取伽马音频率倒谱系数特征。我们提取了 64 个倒谱系数，并将滤波器数量设为 128。

B. 数据增强

我们采用时间拉伸作为主要的数据增强策略，以获得更多训练/测试样本，从而减少部分实验变体中的数据不平衡问题（见第 V 节）。

1) 时间拉伸和零填充：时间拉伸是一种简单的数据增强方法，在不影响音高的情况下延长录音时长（声门音频率不会改变）。播放时间最多可以增加 100%（即达到原始播放时间的两倍）。如果拉伸因子大于 1，则信号会加速；如果拉伸因子小于 1，则信号会减速。我们使用 Librosa 库来实现时间拉伸。根据实验变体的不同（参见第 V 节），我们使用了 1.5、0.9 和 1.2 的拉伸因子。对信号应用时间拉伸后，我们进行了零填充（后置填充），使每段录音的长度保持为 2 秒，以避免模型可能遇到的维度问题。

C. 分类模型

经过仔细调优和评估，我们构建了一个相对简单的卷积神经网络分类模型，该模型能够很好地完成这项任务。我们的模型由四个一维卷积层组成，每个卷积层后面都接一个一维最大池化层。最后一个最大池化层的输出随后传入展平层，再由该层将前一层的输出传递给后续三个全连接层，最终由 Softmax 层提供所需的预测结果。我们还采用了 Dropout 正则化，以提升模型的泛化能力（dropout 概率为 =0.3）。我们使用 Adam优化器训练模型 30 个训练轮次，并结合学习率衰减（最小学习率 = 1.0e-05），耐心值设为 2。图 2 展示了分类模型的计算示意图。

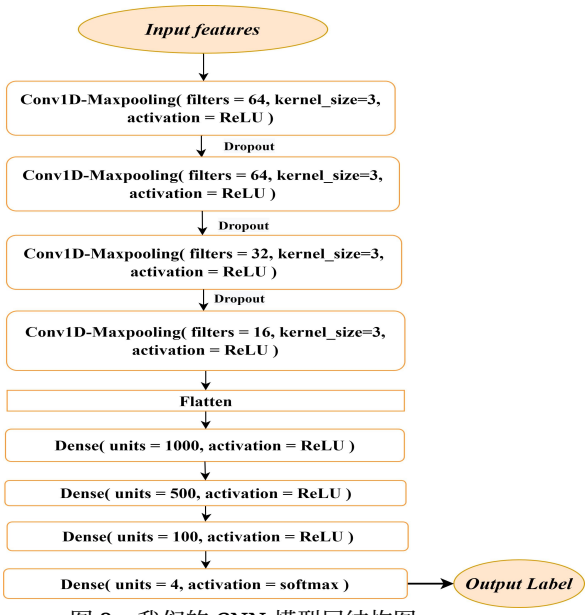


图 2：我们的 CNN 模型层结构图

V. 实验与分析

所有实验均在 Google Colab Pro 上进行，配备 32 GB 主内存、16 GB 显存的 Nvidia A100 GPU，以及 Intel Xeon CPU（视资源可用情况而定）。

A. 复现基线结果

由于 IDMT Traffic 研究中未提供代码 [1]，我们尝试尽可能准确地复现其方法和基线结果。我们将梅尔频带数设为 16，并使用梅尔频谱图提取特征。根据作者的训练测试划分（参见表 II），[1], 中以 30 km/h 和 50 km/h 行驶的车辆样本用于训练，而以 70 km/h 行驶的样本用于测试。在四个类别中的三个类别上，我们取得了相当接近的结果（参见表 I）；然而，对于卡车类别，我们未能达到作者所称的 0.5 F1分数。

B. 使用更简单的模型和不同的特征提取方法建立结果

我们使用基于CNN的分类模型，在数据集上获得了车辆子类型分类结果。此外，

表 I：基线及我们对结果的复现中各类别的 F1分数

模型	Car	卡车	摩托车	无车辆
Abeßer, J 等人 [1]	0.94	0.5	0.96	1.00
我们的复现	0.94	0.35	0.95	1.00

表 II：作者在 [1] 中使用的训练/验证/测试集划分

类别	训练样本	验证集样本	测试样本-ples
Car	2471	275	1157
摩托车	132	15	99
无车辆	2393	266	1412
卡车	290	32	189

我们还尝试了各种特征提取技术，以确定哪种最适合我们的应用。测试结果表明，梅尔频谱图的表现最佳，准确率为 94.24%（参见表 III）。尽管可训练参数数量不到 VGG 模型的一半，我们较为简单的 CNN 模型仍取得了与 [1]和第 V-A 节中使用的 VGG 模型相当的结果。然而，该分类器在卡车类别上的表现仍然不佳，F1分数仅为 0.35。

表 III：采用不同特征提取方法的模型分类结果

类别	特征	精确率	召回率	F1-得分	准确率
Car	MFCC	0.87	0.96	0.91	93.41%
摩托车		0.97	0.89	0.92	
无车辆		0.99	0.98	0.98	
卡车		0.65	0.21	0.31	
Car	GFCC	0.88	0.96	0.92	93.78%
摩托车		0.96	0.91	0.93	
无车辆		0.99	0.99	0.99	
卡车		0.65	0.23	0.33	
Car	梅尔频谱图	0.88	0.98	0.93	**94.24%**
摩托车		0.98	0.90	0.94	
无车辆		1.00	0.99	1.00	
卡车		0.69	0.22	0.35	

C. 采用数据平衡和数据增强的分类

无论采用何种特征提取方法，分类器在卡车类别上的表现都较差（参见表 III），这表明可能存在数据不平衡问题。因此，我们尝试采用平衡训练技术来解决数据集的类别不平衡问题。此外，我们使用时间拉伸作为数据增强方法，将车速为 30 km/h 和 50 km/h 时摩托车和卡车类别的训练样本数量分别从 147 和 322 增加一倍至 294 和 644。汽车和无车辆类别各随机选取了 700 个样本。使用的测试样本与第 V-A 节和第 V-B 节相同。采用这一技术后，卡车类别的 F1分数从 0.35 略微提升至 0.41，而摩托车或无车辆类别的表现未受到显著影响

类别的表现则明显下降，不过汽车类别的 F1分数显著降低（参见表 IV）。

表 IV：采用数据平衡和数据增强的分类准确率

类别	精确率	召回率	F1分数
Car	0.96	0.64	0.77
摩托车	0.89	0.84	0.86
无车辆	0.99	0.99	0.99
卡车	0.27	0.84	0.41

D. 人类性能评估

为了进一步了解 IDMT Traffic 数据集及其相关问题，我们开展了一项人类性能研究，让 20 名不同的受试者（年龄均在 20 至 25 岁之间）各自聆听该数据集中的 20 段音频（每个类别 5 个样本）。这些音频从 70 公里/小时车辆录音中随机抽取，以模拟前几小节中使用的神经网络测试集。我们要求受试者在未经任何训练的情况下，将每段未标注的音频归入四个类别之一。本研究结果总结于表 V。结果显示，超过 60% 的卡车样本被误分类为汽车，这表明数据集可能存在问题，尤其是卡车类别。

表 V：人工研究的混淆矩阵（数值以 % 表示）

	Car	卡车	摩托车	无车辆
Car	95	3	1	1
卡车	61	38	1	0
摩托车	3	1	95	1
无车辆	1	0	1	98

E. 误分类卡车样本研究

为了评估卡车被高比例误分类为汽车的原因，我们分析了神经网络和人类受试者误分类的卡车样本。研究发现，这些误分类之间存在相关性：人类受试者误分类的卡车样本中，超过 90% 也被神经网络误分类。对卡车类别中正确分类和误分类样本进行信号分析后，发现了一个共同趋势。正确分类样本的均方根能量明显高于相应的误分类样本（参见图 5）。同样，与正确分类的卡车样本相比，误分类样本中通过快速傅里叶变换计算得到的所有纯音频率的幅值都非常低（参见图 3 和图 4）。

这两种现象都可以用驶过的卡车与录音麦克风之间的距离来解释：当卡车驶过时距离麦克风较近，麦克风就能更好地录下振幅较大的高频频谱，因此均方根误差会

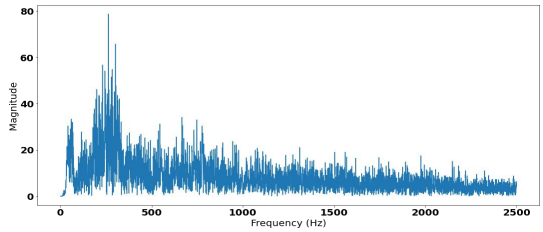


图 3：正确分类的卡车样本的频谱

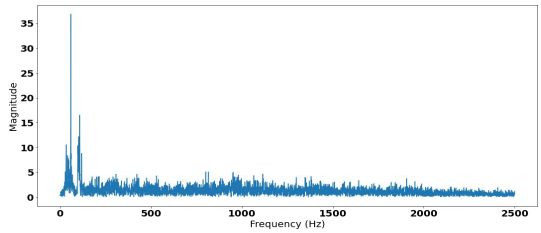


图 4：误分类的卡车样本的频谱

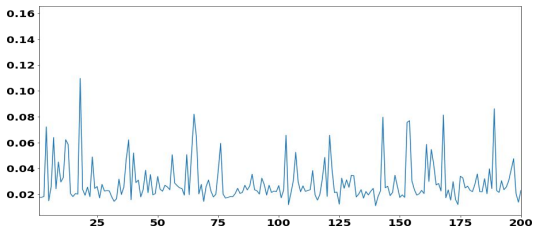


图 6：200 个随机卡车录音的 RSME 值

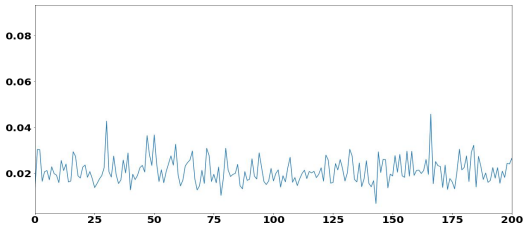


图 7：200 个随机汽车录音的 RSME 值

更高，纯音频率的幅值也更大；而卡车驶过时距离麦克风较远则相反。我们认为，均方根误差值较低的卡车样本不可靠，不能用于

这些样本不适用于模型训练和评估，因为它们无法为车辆子类型分类提供足够的特征。只有卡车样本中观察到属于同一类别的样本之间 RMSE 值的高阶差异（参见图 6 和图 7）。为进一步验证我们的分析，我们采用无监督方法，先将两类卡车样本分开，再进行车辆子类型分类。这两类样本如下：1）优质样本，因此 RMSE 值较高；2）劣质样本，因此 RMSE 值较低。我们使用方程（2）计算了卡车类别中每个样本的 RMSE 值。

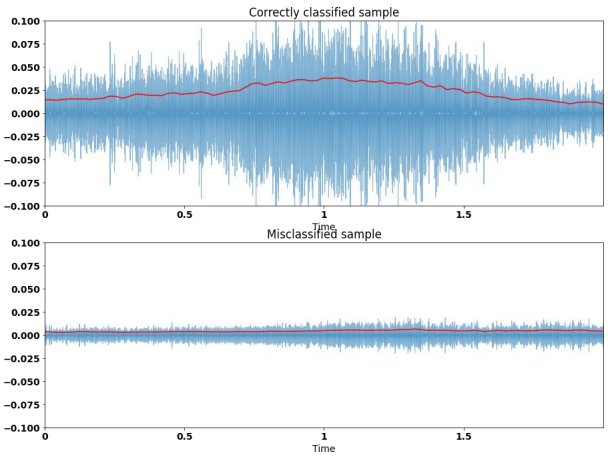


图 5：正确分类和误分类卡车样本的逐帧 RSME 值（以红色显示）

$$\sqrt{\frac{1}{K}\sum_{k=t.K}^{(t+1).K-1} s(k)^2} \qquad \textbf{(2)}$$

其中，K 是每帧中的样本数，t 是帧编号。我们使用这些值创建了一个长度为 87 的向量（所有录音时长均为 2 秒，采样率为 22.05 kHz）。随后，我们将这些向量输入 K 均值聚类算法，并将簇数设为 2。计算得出，簇 1 和簇 2 中的样本数分别为 444 和 67。信号分析表明，簇 1 包含了大多数均方根误差幅值较低的样本（劣质样本），而簇 2 包含了大多数幅值较高的样本（优质样本）。

F. 清理数据集后的车辆子类型分类

在移除卡车类别中的劣质样本后，我们再次进行了车辆子类型分类。由于卡车样本数量不足，无法进行有效的训练和测试，我们采用时间拉伸来增加高质量样本的数量，最终获得了总计 268 个样本，并对选定的样本应用了 1.5、0.8 和 1.2 的拉伸因子。为了针对车辆速度可能变化的现实应用场景对模型进行适当校准，我们随后从数据集中选取了所有摩托车样本，以及第 V-A 节和第 V-B 节中使用的所有汽车和无车辆音频样本，不受速度影响。随机打乱数据后，我们将其划分为 70:30 的训练测试集。将梅尔频率倒谱系数作为特征提取方法后，我们的模型总体准确率达到 98.95%，同时卡车类别的 F1分数提高了 42%（参见表 VI 和 VII）。我们所提出方法的处理流程如图 8 所示。

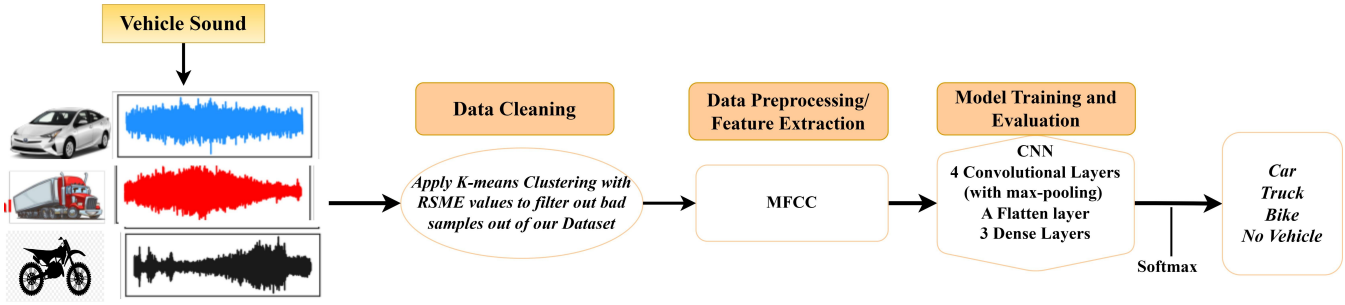


图 8：我们提出的方法的处理流程。首先对采集的数据进行清理，根据样本的均方根误差值剔除样本。然后从剩余样本中提取梅尔频率倒谱系数特征，并将其输入卷积神经网络进行训练和评估。

表 VI：剔除不良样本并采用不同特征提取方法后的分类结果

类别	特征	精确率	召回率	F1-得分	准确率
Car	MFCC	0.99	0.99	0.99	**98.95%**
摩托车		1.00	0.97	0.99	
无车辆		0.99	1.00	1.00	
卡车		0.95	0.89	0.92	
Car	GFCC	0.98	1.00	0.99	98.94%
摩托车		0.97	0.99	0.98	
无车辆		1.00	0.99	0.99	
卡车		0.97	0.84	0.90	
Car	梅尔频谱图	0.98	0.99	0.98	98.08%
摩托车		0.89	0.96	0.93	
无车辆		1.00	0.99	0.99	
卡车		0.86	0.83	0.84	

表 VII：我们的模型与 [1] 的各类别 F1分数比较

模型	Car	卡车	摩托车	无车辆
Abeßer, J 等人 [1]	0.94	0.5	0.96	1.00
我们的模型	**0.99**	**0.92**	**0.99**	**1.00**

VI. 结论与未来工作

声学交通监测是一个前景广阔的领域，随着智慧城市需求不断增长，该领域仍需进一步研究。通过本文及分析，我们证明，在声学车辆子类型分类任务中，简单且经过良好优化的卷积神经网络，配合审慎的数据预处理，往往能取得与更复杂、计算开销更大的方法相当甚至更好的结果，从而提高了我们的方法在实际部署中的实用性。使用梅尔频率倒谱系数、简单的卷积神经网络模型以及样本拒绝方法，我们在 IDMT Traffic 数据集的所有类别上都优于既有基线，其中卡车类别的 F1分数提升了 42%。在 V-F 节中，我们推测录音麦克风与车辆之间的距离可能和录制到的特征质量相关，进而影响分类准确率。在在进一步的研究中，我们希望探究这种相关性，以确定现实世界大规模部署所需的最佳麦克风质量和录音距离。

上述实验的代码已公开提供[15]。

参考文献

[1] Abeßer, Jakob 等人，《IDMT-traffic：用于声学交通监测研究的开放基准数据集》，2021年第29届欧洲信号处理会议（EUSIPCO）。IEEE，2021年。[2] Kubo, Kazuo 等人，《利用离散小波变换设计超低功耗车辆检测器》，ITS AP 论坛论文集，2018年。[3] Ishida, Shigemi 等人，《具有稳态噪声抑制功能的声学车辆检测器设计》，2019年 IEEE 智能交通系统会议（ITSC）。IEEE，2019年。[4] Dawton, Billy 等人，《使用路侧立体声麦克风识别车辆类型的初步评估》，2020年 IEEE 传感器应用研讨会（SAS）。IEEE，2020年。[5] Szwoch, Grzegorz 和 Józef Kotus，《基于声强的道路车辆声学检测器》，Sensors 21.23（2021）：7781。[6] Nooralahiyan, Amir Y. 等人，《用于车辆分类的声学特征分析实地试验》，Transportation Research Part C: Emerging Technologies 5.3-4（1997）：165-177。[7] Yang, An-Chih 和 Emmett D. Goodman，《加速车辆的音频分类》（2019）。[8] Kavitha, N. 和 D. N. Chandrappa，《面向智能交通系统的基于优化 YOLOv2 的车辆分类与跟踪》，Results in Control and Optimization 2（2021）：100008。[9] Butt, Muhammad Atif 等人，《面向智能交通系统的卷积神经网络不利照明条件下车辆分类》，Complexity 2021（2021）。[10] Tichý, Tomáš 等人，《城市区域车辆类别的热特征检测》，Sustainability 13.12（2021）：6873。[11] George, Jobin 等人，《探索利用人工神经网络进行车辆检测与分类的声音特征》，International Journal on Soft Computing 4.2（2013）：29。[12] McFee, Brian、Colin Raffel、Dawen Liang、Daniel PW Ellis、Matt McVicar、Eric Battenberg 和 Oriol Nieto，《librosa：Python 中的音频与音乐信号分析》，载于《第十四届 Python 科学会议论文集》，第18-25页，2015年。[13] Ayoub Malek、Stefano Borzì 和 Christian Heider Nielsen。（2022）。SuperKogito/spafe：v0.2.0（v0.2.0）。Zenodo。https://doi.org/10.5281/zenodo.6824667。[14] Patterson, Roy D. 等人，《基于伽马音调函数的高效听觉滤波器组》，RSRE IOC 语音组听觉建模会议，卷2，第7期，1987年。[15]Jagetia（2022）AcousticVehicleDetection [源代码] https://github.com/ajagetia2001/AcousticVehicleDetection.git