Real-Time Traffic Data Analysis and Deep Learning-based Traffic Volume Classification for Congestion Mitigation at Urban Intersections

Omar Abdullah Hasan[1], **Duraid Y. Mohammed**[2]

[1] College of Engineering, Al-Iraqia University, Iraq
Email: Omar.a.Hasan@aliraqia.edu.iq

[2]Department of Computer, College of Engineering, Al-Iraqia University, Iraq
Email: duraidyehya19@gmail.com.

Article History

Received: Sep. 17, 2024

Revised: Jan. 29, 2025

Accepted: Feb. 13, 2025

Abstract

Managing traffic at intersections in crowded cities is highly dependent on understanding crowding level, which can be assessed through factors such as traffic volume, vehicle count, and signal timing and control. Traditional Intelligent Traffic Systems (ITS) methods, including detector loops, GPS, and camera-based solutions, often present complexities and high costs. This study proposes an alternative approach using Acoustic Traffic Monitoring(ATM) technology to detect abnormalities traffic patterns. A new model was developed to detect traffic volume as crowded (abnormal flow) or non-crowded (normal flow) based on an acoustic dataset collected using Raspberry Pi. The collected data underwent analysis through signal processing techniques, followed by detection using machine learning models: Support Vector Machine (SVM), Random Forest (RF), Artificial Neural Network (ANN), Long Short-Term Memory (LSTM). Evaluating tow datasets, RTD and IDMT-Traffic, across four frame sizes, the results demonstrate the effectiveness of the proposed method. Notably, the LSTM model achieved accuracy of 98.25% on the RTD dataset and 98.69% on the IDMT-Traffic dataset, highlighting it is potential for accurate traffic jam detection.

Keywords- Real-Time Traffic, Data Analysis, Deep Learning, Congestion Mitigation, Urban Intersections.

I. Introduction

In big cities, traffic control is really difficult. Some countries across the world have adopted Intelligent Transportation Systems (ITS) to reduce the expenses related to traffic congestion. Prediction models are advantageous for the advancement of Intelligent Transportation Systems. ITS is a control and information system that uses integrated communications and data processing technology to improve the movement of people and goods. It does this by improving road safety, reducing traffic, and efficiently managing traffic incidents. It also helps transportation policy achieve its goals and objectives, including demand management and prioritizing public transportation[1],[2].Traffic congestion affects the economy of a nation and the health of its citizens both directly and indirectly. There are also personal effects[3] of traffic congestion. In addition to time loss, particularly during rush hours, traffic congestion also contributes to mental stress and increases global warming pollution.

Traffic flow prediction (TFP) has multiple applications in the management of areas and municipal transportation. The TFP problem is a time series (TS) problem that involves forecasting the traffic flow on urban roads for a future time using data collected from one or more observation locations in previous times. The objective of this study is to elucidate traffic forecasts using an acoustic dataset. Traffic congestion is the result of multiple factors that interact dynamically. These factors include alterations in road design, volume of traffic over a period of time, meteorological data, accidents, road maintenance, and other related aspects [4].Also, in this study, the general public will benefit by providing access to the most recent research on traffic forecasting using machine learning algorithms and acoustical data. Within the field of artificial intelligence (AI), ML focuses

on developing computer algorithms that become more accurate over time as they analyze and absorb large volumes of data. Many applications can benefit from machine learning's flexible ability to learn from previous datasets. Predicting traffic jams can be done by using ML principles and implementing related applications.

The methods of today are costly or intended for particular purposes, including. Instead of keeping track of traffic conditions, induction loops are made to track the speed of moving cars. They are expensive, involve some roadside assistance and are difficult to relocate once installed. Video analysis is frequently used to track traffic conditions, although it has certain drawbacks due to occlusions, changing lighting conditions, and other environmental factors (the analysis differs from daytime traffic). By analyzing the audio from the horns and processing GPS and mobile device data from the cars, it is also possible to indirectly infer the traffic conditions. However, these methods depend on the cooperation or habits of the drivers [3]. Based on these, building a prediction model that adopts new innovation strategies that can overcome the drawbacks of old methods and give a practical solution for forecasting traffic congestion is essential. The use of machine learning algorithms with auditory data based on vehicle engine sound for traffic flow predictions on various urban city roads is the main topic of this study. Through the use of auditory data for training, machine learning techniques enable the machine to automatically forecast traffic congestion.

II. Related work

The audible sound on a road is a combination of several sound sources, such as engines, exhausts, wheels, and air turbulence, which occur when vehicles pass by [4]. As a reasonable way to break down this complex audio analysis task, researchers approached traffic monitoring from different perspectives. In this section, we categorize existing ATM algorithms based on approaches for audio data acquisition and statistical modeling. Baghdad city, the capital of Iraq, has suffered mainly from the influences of transportation highway modes, especially at intersections. Baghdad's city experienced severe traffic congestion. Especially in the present few years, and this is normal in the absence of any modern techniques and traffic management studies to relive or alleviate some of the adverse consequences of congestions. The traffic management system at most intersections in Baghdad considers it F type according to the level of LO Sservice[5].

M. Ashhad[6]They considered the problem of classifying auditory signals into four categories: no vehicle, automobile, truck, and bike. This task is known as acoustic vehicle sub-type classification. As characteristics of the audio dataset, they employed GFCC, MFCC, and Mel-Spectrograms. Vehicle type classification has also been done using CNN-based classifiers. Their accuracy in the IDMT Traffic dataset is 98%.

AP Perdana[7] They were suggested in this study. The online application uses audio to detect traffic congestion and uses a trained model to reliably identify congestion and indicate probability. To distinguish between congested and non-congested traffic in a streamlit online application, they also implemented the SVM and Nave Bayes algorithms and evaluated how well they performed. The results clearly show that the support vector machine (SVM), with accuracy, precision, recall, and F-measure of 95%, 91%, 100% and 89%, respectively, surpasses Nave Bayes in the identification of congestion.

CY Chiang[8] Intelligent transport systems (ITS) employ distributed acoustic sensors (DAS) for traffic analysis. To analyze DAS signals for vehicle classification and occupancy detection, a deep learning technique was used. 92% precision in vehicle categorization and 92% to 97% accuracy in occupancy detection were achieved.

RC Gatto[3] An audio-based machine learning algorithm was implemented to detect traffic congestion. Their algorithm evaluates traffic conditions using machine learning and acoustic sensors. evaluation of the state of traffic using these auditory signals and audio data analysis. The two classes, Non-Congested/Free (NC) and Congested (C), were used to categorize the traffic. Isolated audio samples with no further information other than what was taken from the audio file were used to train the classifiers. However, temporal coherence information can be provided in real-world application settings. This allows the estimation of subsequent traffic samples to be aided by the classification of earlier ones. When doing this type of estimate filtering, Markov models are typically used. The model's output might identify non-congested traffic has great accuracy and a 0.89 average accuracy with a 0.10 standard deviation in crowded traffic.

Yang et al.[9] used a Smartphone to record vehicle acceleration at an isolated stop sign to avoid various traffic and oncoming vehicles, resulting in a data set with clear audios. Their proposed pipeline comprises both spectral and temporal feature extraction approaches, as well as techniques like noise injection and pitch change to augment the data, which is then sent into several CNN models for classification of vehicles into the following classes: Hybrid, sedan, pickup, bus, and commercial. They achieved an accuracy of 75%.

TABLE I: Methods, data sets, and accuracy results in five previous studies.

Methods	Dataset	Accuracy	References
CNN	IDMT data set	98%	[10]
SVM and Nave ayes	Acoustic dataset	95%	[7]
Deep Learning	Acoustic dataset	92%	[8]
Machine Learning	Acoustic dataset	89%	[3]
CNN	Acoustic dataset	75%	[9]
LSTM, ANN,RF, and SVM	RTD, IDMT Dataset	98.25% by LSTM	

III. Method and Materials

The block diagram shows the proposed model according to sequence of work as shown in figure1.

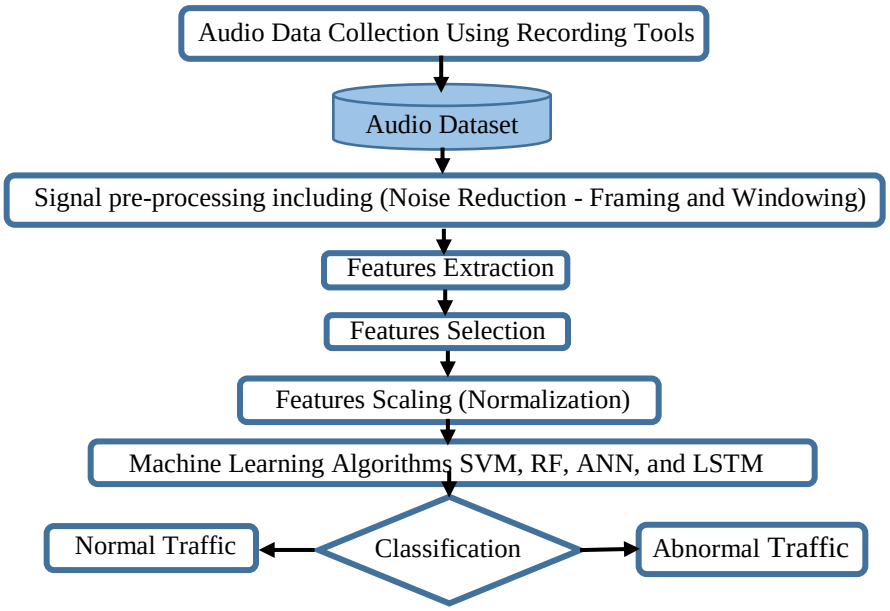

Figure 1: Block diagram of Traffic Jam Categorization system.

3.1 Recording Hardware Kit:

The designed hardware components for data collection and transmit it over internet is outlined in this section. The hardware kit designed for this study is a robust and efficient system tailored for continuous traffic data collection during working hours. Central to this system is the Raspberry Pi 4, which functions as the primary processing unit. The kit includes a high-fidelity microphone to capture traffic-related acoustic data. To ensure the system's reliability and uninterrupted operation, an uninterruptible power supply (UPS) is incorporated, mitigating the risk of data loss due to power failures. Moreover, the hardware is equipped with a Wi-Fi module, enabling seamless data transmission to a remote server for real-time processing and analysis. The detailed configuration of the hardware setup is depicted in the figure 2.

Figure 2: recording Hardware kit diagram for data collection.

3.1.1 Raspberry Pi 4

The Raspberry Pi 4 Model B, equipped with 4 GB of RAM, is a good choice for the data collection where it has features that can work as computer with all application that are required for recording. The new Raspberry Pi delivers significant advancements in CPU speed, multimedia performance, memory, and connectivity compared to the previous-generation Raspberry Pi 3 Model B+. It maintains compatibility with older versions and has similar power consumption. The Raspberry Pi 4 model B offers desktop performance that is equivalent to entry-level x86 PC systems, making it suitable for end users.

3.1.2 Mini UPS

The WB FSP1-3 18w mini UPS model was used in order to ensure constant availability of electricity when there is a power outage. It is construct from high capacity lithium batteries, provides long backup to the loads. Also provide high compatibility for most digital products where it is consisting of many types of output power supply interfaces such as USB (5V), three types of DC interfaces (5V, 9V, 12V), and power over Ethernet (POE) ports. The mini UPS can feed all devices that are needed for data collection with power for a minimum of two hours without the main electricity source.

3.1.3 Zain Router

Huawei Mobile Wi-Fi 3G E5336Bs-2 model device that supported from Zain Iraq telecom company was used as internet service source. It is providing good solution due it is characterized by its small size, ease of setup, and provides good service. The Zain router provides internet to a raspberry Pi 4 so we can access remotely to it from our PC for check the recording audio data status if there is any problem to solve it and store a copy from audio data in our PC.

3.1.4 Tenda Router

The Tenda Wireless Router F3 is a 300Mbps wireless router designed for home use. It is known for its ease of use, affordability, and decent performance for basic internet needs. The Tenda F3 supports Wireless Internet Service Provider (WISP) mode, which allows it to act as a client to connect to a wireless network provided by an ISP or another router. We used (WISP) feature to make the Tenda router as extender for Zain router which is act as (ISP). Due the Linux operating system which is default OS of raspberry Pi has some limitation with some of software that are needed in our work so we were installed Windows 10 operating system. This led to the Raspberry's wireless connection service could not be installed with Win10 OS, and only LAN port driver can be installed on raspberry. Therefore, we have been adding another router as internet service extender equipped with a number of LAN ports that can take internet service from the Zain router and deliver it to the Raspberry Pi through its LAN port.

3.1.5 Dual Wireless Microphone

The wireless dual microphone model utilized in this work incorporates a Digital Signal Processing (DSP) technique to optimize and improve the quality of recorded audio signals. The Bluetooth 5.0 technology enables wireless connectivity between smartphones, tablets, or PCs. Approximately 20 meters of covering area that is easily accessible. In addition, there is a built-in rechargeable battery that offers a power up to 6 hours of uninterrupted usage. Capable of functioning with a diverse range of audio and visual applications. The Bluetooth microphone used in this setup is a compact and wireless device designed to capture high-quality audio signals. It operates by picking up sound waves through its built-in diaphragm, which converts these acoustic signals into electrical signals. These electrical signals are then digitized within the microphone's internal circuitry. Once the sound signal is digitized, the Bluetooth microphone encodes and transmits it wirelessly via Bluetooth technology. The signal is sent as data packets to the Raspberry Pi 4, which is equipped with a Bluetooth receiver. The Raspberry Pi then receives these data packets, decodes them, and processes the audio signal for storage, analysis, or further transmission. This wireless transmission method reduces the need for physical connections, allowing for greater flexibility in the placement of the microphone and ensuring a clean setup free from cable clutter.

All of the above devices are placed inside a box equipped with a 12-inch fan for cooling, this is to protect the hardware kit from weather bad conditions such as the raining and high temperature. The USB hub is also connected to a Mini UPS to distribute power to all devices as shown in Fig. 3. The data collected devices box was putted beside road about 300 meters before Al Dora-Saydia intersection (Latitude 33.25310085681531, Longitude 44.3665956200852). The data set was collected over 5 days from Sunday to Thursday about 8 hours a day from 7 am to 4 pm to cover normal and abnormal traffic situations over time. Figure 3 shows the recording hardware kit.

Figure 3: Recording hardware kit for data collection

IV. Dataset:

This thesis leveraged two primary datasets: Institute for Digital Media Technology (IDMT-TRAFFIC) and the collected real-time dataset (RTD). The IDMT-TRAFFIC dataset has been extensively utilized in numerous research endeavors and formed a cornerstone of this study. the IDMT-TRAFFIC dataset is used in its original form, without any additional noise reduction processing. Moreover, this dataset was already annotated with labels, facilitating the analysis process. The IDMT-TRAFFIC dataset is utilized to evaluate the proposed methodology. In addition to the IDMT-TRAFFIC Dataset, the RTD dataset captures real-world scenarios, reflecting authentic environmental conditions. It's used for validate the model. Furthermore, manual labelling was conducted to ensure accurate annotations within the RTD dataset.In this section the TWO groups of dataset that used in the proposed model are described. First one is the Real-Time Collected Dataset (RTD) through this study and the other dataset is the Institute for Digital Media Technology (IDMT-TRAFFIC).

4.1 The Collected Real Time Dataset (RTD)

The RTD is collected at one of Baghdad intersections (Al Dora-Saydia intersection (Latitude 33.25310085681531, Longitude 44.3665956200852)) for validation purpose. The data set was collected over 5 days from Sunday to Thursday about 8-9 hours a day from 7 am to 4 pm for simulating and cover all scenarios and factors that causing traffic congestion such as peak time, accidents, and road infrastructures in order to training the model on all the traffic patterns for classification abnormalities in the traffic. The dataset collected by this study that fed the model was retrieved from sound signals. The sound signals are emitted from vehicles on the side of the road. The total length of recorded sound signals (raw dataset before pre-processing) is above of 40 hours' stereo sound. The recording audio signal format is waveform (WAV) and The sampling rate (Sr) is 48 KHz. The recording audio dataset is classified auditory into two groups: normal traffic and abnormal traffic. The labeling process was achieved manually where 0 is represent normal traffic and 1 for abnormal traffic. The dataset is segmented into 4 frame lengths (40ms, 60ms, 80ms, and 100ms) to choose the optimal frame length that serves the classification model.

The framing process is essential because most audio analysis techniques, especially in machine learning, require processing the audio signal in short, manageable pieces rather than as a continuous stream to check best frame length that give high accuracy.

Next step, the dataset undergoes a necessary signal pre-processing step aimed at eliminating unwanted artefacts. The table1 below shows recording process of the RTD.

TABLE II: Size, Duration, time and No. of Samples for RTD over five days.

Day	Sunday	Monday	Tuesday	Wednesday	Thursday
Recording size	5.32 GB	6.09 GB	6.31 GB	6.15 GB	4.83 GB
Recording duration	8 Hours	9:30 hours	9:45 hours	9:35 hours	7:30 hours
Recording time	8AM-4PM	7AM-4:30PM	7AM-4:45PM	7AM-4:35PM	8AM-3:30PM
Samples No.	288,000	342,000	351,000	343,800	270,000

4.2 Institute for Digital Media Technology (IDMT-Traffic) Dataset

The IDMT-TRAFFIC a novel open benchmark dataset consisting of stereo audio recordings for vehicle passing events, captured with high-quality sE8 and medium-quality MEMS microphones [60]. The dataset is designed to evaluate the use-case of deploying audio classification algorithms. In this study we deployed this dataset for evaluate our model. It's comprises 17,506 stereo audio samples lasting 2 seconds each, 48KHz samples rate (Sr) capturing recorded vehicle passing, as well as various background sounds present on streets. The recording scenarios of IDMT-Traffic include different speed limits (30Km, 50Km, and 70 km/h) as well as wet and dry road conditions. In this study, the speed limits 30Km is considered as abnormal traffic (crowded) otherwise is normal traffic (non-crowded).Also these dataset undergoes to same pre-processing steps detailed in 2.3 except the noise reduction step. Where it was also divided into four group frame lengths 40ms, 60ms, 80ms, and 100ms This makes voice analysis and pattern recognition more accurate.

V. Preprocessing

In this section we show the proposed method and technique used for preprocessing, features engineering, and classification of the acoustic data. It also discusses the machine learning algorithms used for classification and evaluating the proposed approach according to metrics to assess it. The first step in processing sound data is framing the recording datasets. The second step is to extract most related features from audio data that are dealing with traffic components by using appropriate ML. It is a crucial stage in machine learning procedures using acoustic data, such as voice recognition, sound event detection, and music classification. The typical pre-processing processes for acoustic data involve extracting the audio files and loading them from their storage locations using a Python library such as Librosa or pyDub. Subsequently, the retrieved data are subjected to the resampling procedure to guarantee that all sample rates are identical. The normalization procedure involves addressing inconsistencies in sample amplitude by using min-max normalization. The process of noise reduction is essential to eliminate background noise of the audio dataset. The estimation for frequency amplitude is used to achieve this stage. The steps below show the sequence of pre-processing process:

5.1 Framing and Windowing process

Some studies that deal with audio datasets, such as speech recognition, divide the dataset into frame sizes between 10 and 50 ms [12].In this studythe audio dataset is segmented into 4 frame sizes (40 ms, 60 ms, 80 ms, and 100 ms) to choose the optimal frame size that serves the classification model. The framing process is essential because the audio dataset is not constant with the time so for analysis it must achieved stationarity. Also most audio analysis techniques, especially in machine learning, require processing the audio signal in short, manageable pieces rather than as a continuous stream to check best frame length that give high accuracy.In this study the 100ms frame size is the optimal and it's achieves highest accuracy. The hop length is used equal 50% of frame length. These parameter values were determined through trial-and-error and were found to be the most optimal for the proposed method. This Prove a fact a large hop size lead to reduce the overlapping and gives butter frequency resolution but might miss temporal details. For the small hop size increase overlapping and gives butter time resolution but effect the frequency details. So medium hop size is providing a trade-off between time and frequency resolution.

5.2 Windowing

To solve spectral leak issue at the boundaries ofdiscontinuities frames, the widowing is applying to smooth these boundaries and reducing spectral leakage.

In this study the Hamming window was applied because to its superior accuracy compared to the other windows that were tested such as hanning, rectangular, and balckman window.

$$\begin{cases} 0.54 - 0.46\cos(2\pi n/(L-1)) & 0 \le n \le L-1 \quad (1) \\ 0 & \text{Otherwise} \end{cases}$$

5.3Noise Reduction

The next step is to apply noise reduction or audio signals enhancement techniques to correct any spectral distortions in the audio signals caused by the presence of additional noise such as birds noise, and Pedestrians on the side of the road. The audio dataset enhancement consider challenge due the variety of background noise and limitations of current tools for noise

reduction [13]. Algorithms developed with the purpose of diminishing or eliminating background noise to a certain degree are generally referred to as noise suppression algorithms, which aim to reduce noise or audio enhancement. Noise can be found in several types, such as birds sounds, wind, and Pedestrian sounds. This study utilizes a spectral subtraction algorithm to reduce background noise, the graphic below depicts the spectrum subtraction algorithm.

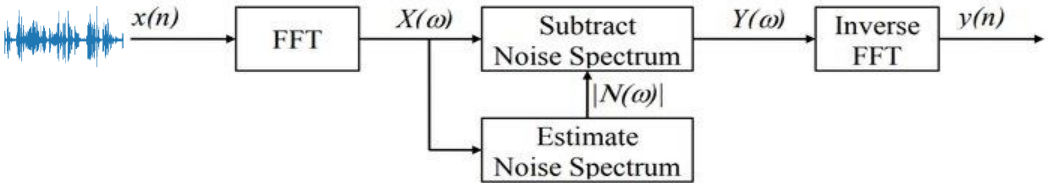

Figure 4: A block diagram illustration of spectral subtraction spectral subtraction.

Where $x(n)$, X(w), $Y(w)$, and y(n) are the noisy signal, the Fourier transforms of the noisy signal $x(n)$, Fourier transforms of the original signal, and y(n) the enhanced signal.

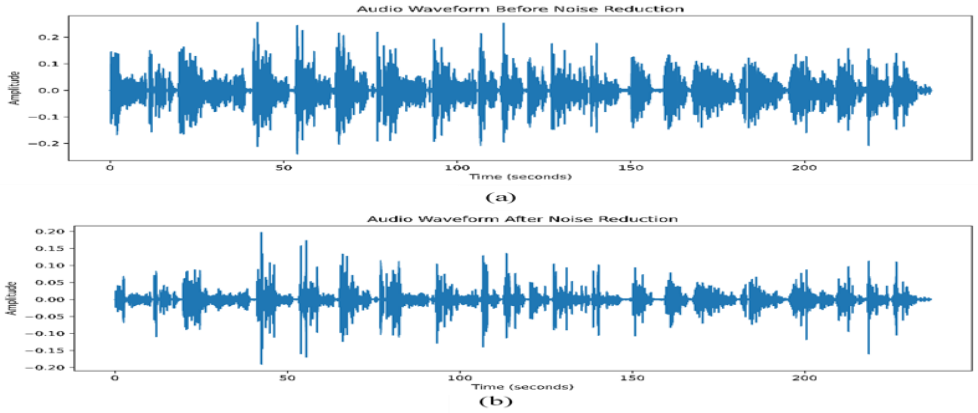

Figure 4: Difference between waveform (a) before noise reduction (b) after noise reduction.

In addition, the Signal to Noise Ratio (SNR) was computed for the real-time dataset both before and after the application of noise reduction techniques. The signal-to-noise ratio (SNR) was determined using equation 2.14 as stated in reference [11].

$$\text{SNR} = 10\log_{10}\frac{p\,x(n)}{p\,noise}(2)$$

Px represents the power of the input signal x(n), while $Pnoise$ represents the power of the noise.

The $Pnoise$ can be approximated by utilizing the subsequent equation:

$$\text{p}_{\text{noise}} = \frac{1}{\text{N}}\sum_{\text{k=1}}^{\text{N}}|\text{n[k]}|^2(3)$$

The n[K] is denote the noise sample at Kth time step, N is the total number of samples, |n[K]| is the power of the Kth noise sample.

Table III: SNR before and after noise reduction.

Real Time Dataset	
State	**SNR**
Before	-0.088

After	2.896

VI. Feature extraction

Feature extraction is an important stage in the proposed methodology, contributing to the comprehensive analysis of the preprocessed acoustic data. The objective is to derivea collection of features from the relevant dataset. These qualities must be informative regarding the intended properties of the source data. Since feature extraction aims to build our analytical algorithms on a limited set of features, it can also be considered a data rate reduction process [42]. Features derived from the time domain and features obtained from the frequency domain. In this study, the features were obtained from the time domain and frequency domain using FFT to generate a spectrum representation of the signal.

VII. Features selection

Following feature extraction, the obtained features undergo optimization to address potential data dimensionality issues. High-dimensional data can pose challenges in terms of computational efficiency and the risk of overfitting. To mitigate these challenges, the feature selection method has chosen to eliminate irrelevant information in order to reduce dimensionality issues and decrease the complexity of the final model. This process produces two-dimensional matrices for each frame length. Every row corresponds to a sample, whereas each column represents a distinct feature. The feature importance technique, using random forest (RF), was utilized in this study for feature selection. The number of features that are extracted (16), Mel Frequency Cepstral Coefficients MFCC (13), spectral centroid (1), spectral bandwidth (1), and zero crossing rate ZCR (1). Figure 5 displays a complete set of features arranged in descending order of importance.

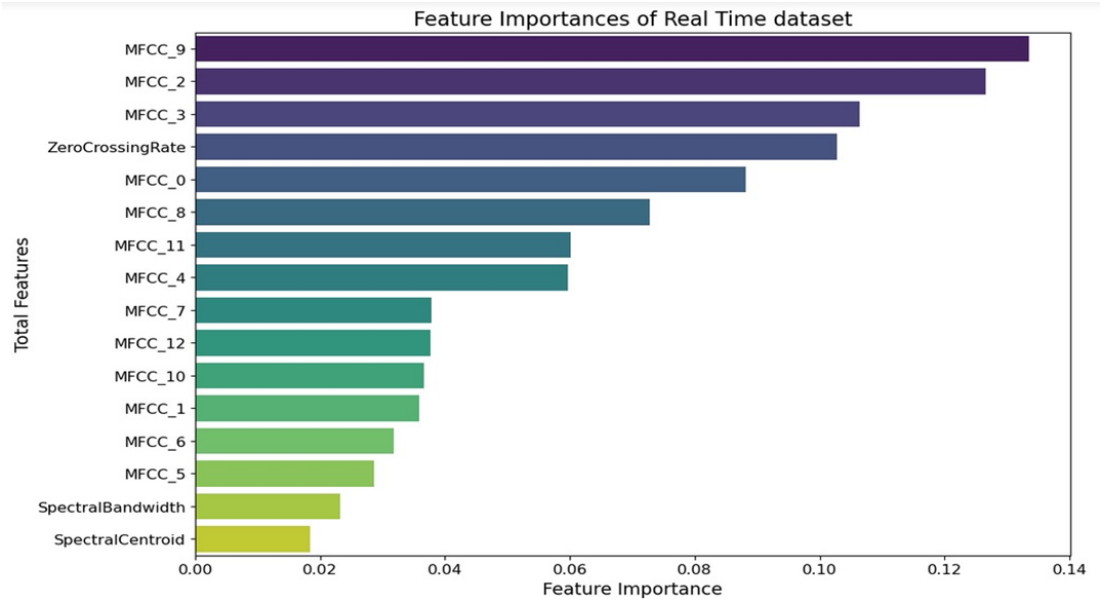

Figure 5: Total important features using RF.

VIII. Scaling(Normalization)

Normalization is a process applied in various contexts to scale data into a specific range or format. The purpose of normalization is to ensure that different data points or features are on a similar scale, which can improve the performance and convergence of algorithms, particularly in machine learning and signal processing. Normalization does not affect the signal-to-noise ratio or relative dynamics of the recording because the same amount of boost is applied uniformly across the entire track. The signal-to-noise ratio or relative dynamics of the audio recording are unaffected by audio normalization. This is due to the consistent distribution of gain applied during normalization throughout the track. As a result, the original dynamics is preserved and uncompromised sound quality is guaranteed. Both the audio signal and the background noise maintain their proportional amplitudes. Being non-destructive, it enables modifications and reversibility as necessary and does not change the actual audio data. Audio normalization is essential for maintaining constant volume levels fordiverse audio signals acquired from vehicles during traffic condition monitoring. The accuracy and dependability of the traffic condition categorization are ultimately improved by this method, which prevents abrupt changes in loudness and allows a seamless transition between the audio parts. Additionally, audio normalization maintains audio fidelity by avoiding any clipping or distortion caused by excessive volume levels, which enhances the experience of monitoring the state of traffic in general.

The formula for normalizing an audio signal x is as follows:

$$\text{standardized\_ x[i]} = \frac{x[i]-m}{\sigma} (4)$$

where x[i] is the i-th sample of the audio signal, μ is the mean of the audio signal, σ is the standard deviation of the audio signal, and Calculate mean M

$$M = \frac{1}{N}\Sigma_{i=1}^{N} x[i] (5)$$

where N is the number of samples in the audio signal.Calculate the standard deviation σ

$$\sigma = \sqrt{\frac{1}{N}\Sigma_{i=1}^{N}(x[i] - m)^2} (6)$$

IX. Machine Learning

In this study four classifiers have been used for categorize the datasets. The selection of these classifiers was based on their widespread adoption and broad utilization in numerous research papers. These algorithms include Support Vector Machine (SVM), Random Forest (RF), Artificial Neural Network (ANN), and Long Short-term Memory (LSTM).

X. Training and Testing

This section provides a comprehensive explanation of the training and testing processes of the model. Once the pre-processing procedure was completed and feature selection algorithms were applied, the data was prepared for training and testing the model. In this study the we utilised two datasets, RTD for validation of the model and IDMT-Traffic, to evaluate the model. K-fold cross-validation technique was used where it's employs several train-test partitions rather than a singular split. The complete dataset was partitioned into five equal-sized folds based on its total size. In each cycle, one of the 5 subsets is designated as the test set, while the remaining 5-1 subsets serve as the training set. This procedure is executed five times, with each of the K folds utilised precisely once as the test set. Upon concluding five iterations, the accuracy, precision, recall, and F1 score of each fold are summed to yield a comprehensive performance metric. This results in more dependable assessments of model efficacy, as the model is evaluated on various data subsets, with the ultimate performance indicator being averaged across these divisions. Also using Cross-Validation prevent overfitting that may be occur by guaranteeing that the model demonstrates robust performance over several datasets. The model is trained and evaluated on distinct subsets of the data, offering a more accurate assessment of its performance on unfamiliar data.

During the validation and evaluation processes, we employed four 100,000 x 16 matrices based on the frame sizes used to feed our models.

XI. Classification

The final stage of the proposed methodology involves the application of the four models: LSTM, ANN, SVM, and RF. The model is based on a supervised learning technique. The dataset was labeled manually with a value of 1 assigned to crowded (abnormal traffic flow) and 0 to non-crowded (regular traffic flow).

The long short term memory LSTM layers with dense (completely connected) layers of a neural network for binary classification. The method uses the advantages of LSTM networks for capturing temporal relationships, as well as the proficiency of dense layers to detect abnormalities in road conditions based on the patterns and features present in the input data. The LSTM layer consists of 64 units, which corresponds to the number of LSTM cells or memory units in the layer. LSTMs are a variant of recurrent neural networks RNN that proficiently capture long term dependencies in sequential data. They are especially proficient in activities where the sequence and context of data points are significant, including time series analysis, natural language processing, and audio processing, the sigmoid activation function is employed in the output layer for binary classification tasks.

The Artificial Neural Network (ANN) employs two densely connected layers, one for each layer. The 64 neurons do not select randomly; they are selected based on experimental findings that indicate that 64 neurons frequently achieve an optimal equilibrium between model capacity and computing efficiency. It is large enough to capture intricate patterns but not so large as to render the model unnecessarily complex or susceptible to overfitting. In these layers, the ReLU( ) activation function is used for the neurons. The ReLU provides an easy-to-use and effective activation function that helps neural networks learn by reducing issues like vanishing gradients and encouraging sparsity, which could lead to better generalization. The 'Sigmoid'

was used as an activation function of the output layer for the ANN model. It is used in an artificial neural network's output layer for binary classification, mapping the output to a value of 0 or 1.

The Random Forest (RF) constructed 100 estimators, meaning the model consists of 100 decision trees which are independent decision trees. The accuracy is calculated using the majority vote of these trees on the test data. The choice of 100 estimators was by trial and error, where when used with less than 100, the performance was effective and led to decreased accuracy. For more than 100 estimators, lead to minimal improvement and high computation. The predictions from these trees were then combined to reach the final classification decision.

For SVM in order to create a linear decision boundary that can effectively divide the classes, the linear kernel was used throughout training the model. The SVM model is classified the data using a hyperplane in a high-dimensional space. For hyperparameter tuning, the performance is improved by tuning there regularization parameter C value, which is set to 1.0 where it getting the highest accuracy after testing several C values by using the GridSearch CV tool from the sklearn library.

XII. Results and Discussion

In this section summarises and analyses the results obtained from all the experiments conducted in this study. As previously stated, 4 classifiers were used in the experiment, and two datasets were used for the analysis. As stated in the preceding chapter, the significance of the traits was determined by selecting only the relevant ones. Additionally, it is crucial to analyse the model's performance by employing the confusion matrix.

4.1 Model Validation using Real-Time dataset (RTD)

The RTD dataset was used for training and testing the model in the validation phase. Different frame sizes were also utilized during the training and testing phases as specified in Table 4.

Table III: Accuracy of four classifiers that are used with RTD dataset according to frame size.

Frame size	SVM	RF	ANN	LSTM
40ms	0.8262	0.9061	0.9211	0.9271
60ms	0.8406	0.9251	0.9406	0.9440
80ms	0.8619	0.9394	0.9530	0.9533
100ms	0.9450	0.9725	0.9805	0.9825

According to the data presented in Table III, the highest level of accuracy during validation stage was achieved by employing LSTM model and choosing 100ms as frame size. Due longer frame sizes capture more information in each frame, which can make the model’s predictions during validation seem more accurate. A frame size of 100ms captures more temporal information from the audio stream than lesser frame sizes such as 40ms or 60ms. In numerous audio classification tasks, including speech recognition and acoustic event detection, this further temporal information might be essential. It enables the model to more effectively capture extended auditory patterns, pauses, and subtle frequency variations that are crucial for differentiating various classes.

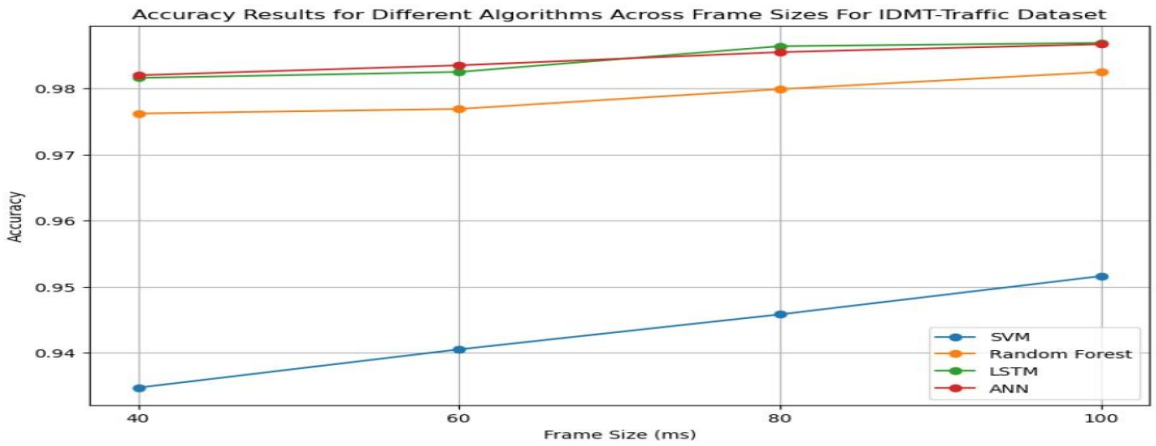

Figure 6: the accuracy results of different algorithms that has been used according to frame size for RTD Dataset.

4.2 Model Evaluation using IDMT-TRAFFIC Dataset

The IDMT-TRAFFIC dataset was used to evaluate the model, with IDMT serving as evaluation dataset. The dataset was also divided into four frame sizes (100ms, 80ms, 60ms, and 40ms). During the validation stage, the frame size of 100ms achieved the highest accuracy. However, in the evaluation stage the 100ms frame size also achieved the highest accuracy. As mentioned, longer frame sizes capture more information in each frame, which can make the model's predictions during validation seem more accurate. Table IV shows the results of algorithms applied to the IDMT-Traffic dataset.

Table IV: Accuracy of the four classifiers used with the IDMT-Traffic dataset according to frame size.

Frame size	SVM	RF	ANN	LSTM
40ms	0.9347	0.9762	0.9820	0.9816
60ms	0.9406	0.9769	0.9835	0.9825
80ms	0.9458	0.9799	0.9855	0.9864
100ms	0.9516	0.9825	0.9867	0.9869

The results demonstrate a correlation between the frame size and the improvement in accuracy. The Figure 7 shows the accuracy results of different algorithms that has been used according to frame size.

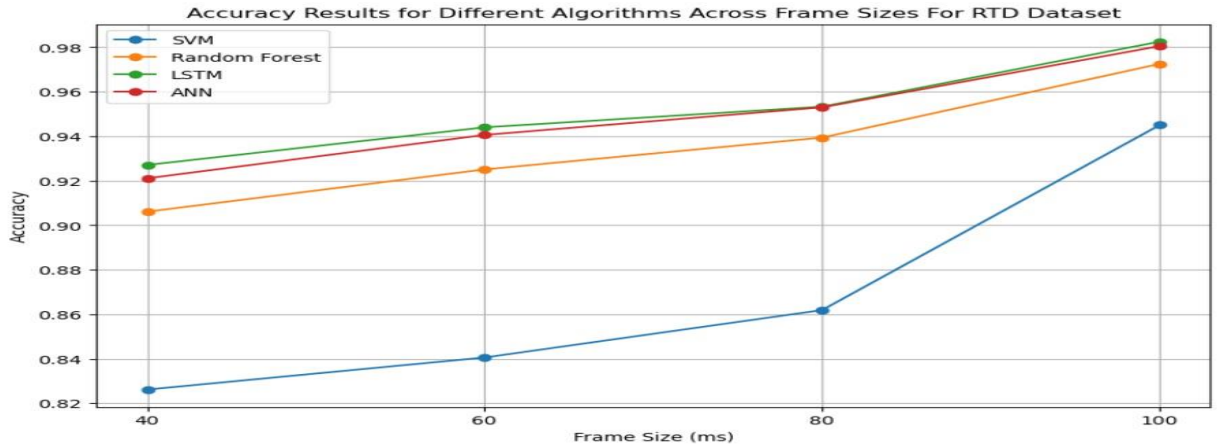

Figure 8: The accuracy results of different algorithms that has been used according to frame size for IDMT dataset

4.4 Confusion Matrix

The confusion matrix had been generated for the LSTM model using a frame size of 100ms. This frame size and classifier combination yielded the highest accuracy, as depicted in Figure 8 for the IDMT-Traffic and RTD datasets.

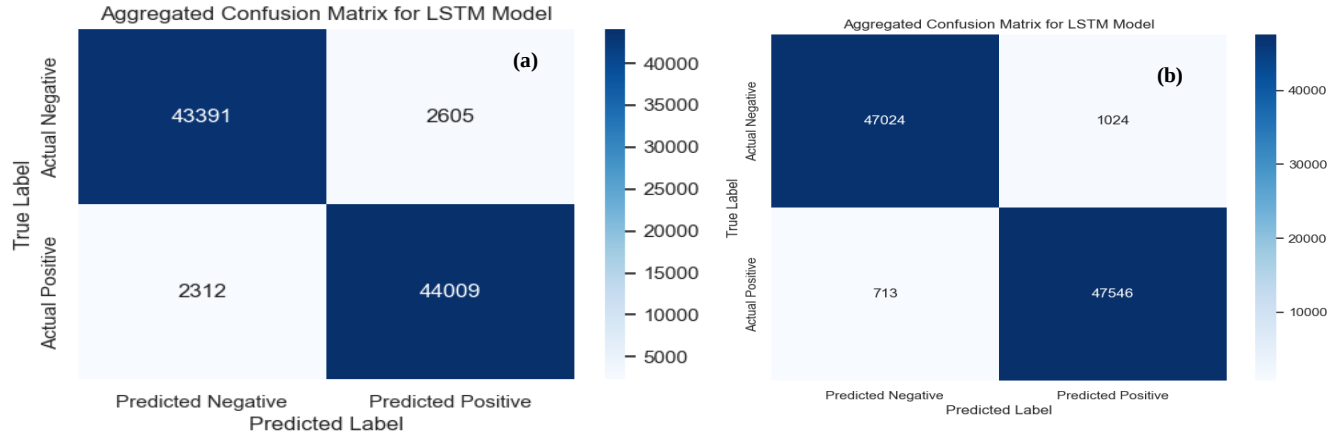

Figure 8: Confusion Matrix of 100ms frame size (a)RTD and (b)IDMT-Traffic datasets using LSTM model.

The confusion matrix shows a lot of information, but frequently a more straightforward metric may be required and these metrics are precision, recall, and F1 score.

4.4.1 Precision

Precision is a metric used to evaluate the performance of a classification model, particularly in binary classification, but it can also be extended to multi-class classification. It measures the accuracy of the positive predictions made by the model.

The precision can be calculated according to 4.1 equation:

$$Precision = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{7}$$

4.4.2 Recall

Recall is another important metric used to evaluate the performance of a classification model, particularly in the context of binary classification. It measures the ability of the model to correctly identify all relevant instances within the data. Recall is computed according to equation 4.4.2.

$$Recall = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negative}} \tag{8}$$

4.4.3 F1 Score

The F1 score is a metric used to evaluate the performance of a classification model by balancing both precision and recall. It is especially useful when you want to consider both false positives and false negatives, making it a good measure in cases where the class distribution is imbalanced.

Equation 8 is used to define F1 Score.

$$F1\ Score = 2\ \text{X}\ \frac{\text{Precision X Recall}}{\text{Precision} +\ \text{Recall}} \tag{9}$$

Table V shows the performance measurements using LSTM for IDMT-TRAFFIC and RTD datasets.

Table V: Performance matrices using LSTM (a) IDMT-TRAFFIC Dataset (b) RTD Dataset

RTD dataset		
Matric	**Normal Traffic**	**Abnormal Traffic**
Precision	0.95	0.94
Recall	0.94	0.96
F1-score	0.95	0.95

(a)

IDMT-TRAFFIC dataset		
Matric	**Normal Traffic**	**Abnormal Traffic**
Precision	0.99	0.98
Recall	0.98	0.99
F1-score	0.98	0.98

(b)

When comparing the confusion matrices from Figure 8 with the performance measurements in Table V for the RTD and IDMT-Traffic datasets, these can be drawing several insights:

4.4 Confusion Matrices Evaluation Discussion:

4.4.1 For Figure 4-1:

1. **RTD Dataset:**
    - True Negative (TN): 43,391
    - False Positives (FP): 2,605
    - False Negative (FN): 2,312
    - True Positive (TP): 44,009

2. **IDMT-Traffic Dataset:**
    - True Negative (TN): 47,024
    - False Positives (FP): 1,024
    - False Negative (FN): 713
    - True Positive (TP): 47,546

4.4.2 Performance Measurements Table 4-1:

1. **RTD Dataset:**
    - Precision: Normal Traffic (0.95), Abnormal Traffic (0.94)
    - Recall: Normal Traffic (0.94), Abnormal Traffic (0.96)
    - F1 Score: Normal Traffic (0.95), Abnormal Traffic (0.95)

2. **IDMT-Traffic Dataset:**
    - Precision: Normal Traffic (0.99), Abnormal Traffic (0.98)
    - Recall: Normal Traffic (0.98), Abnormal Traffic (0.96)
    - F1 Score: Normal Traffic (0.98), Abnormal Traffic (0.97

4.4.3 Comparison Insights:

- **Accuracy:** The confusion matrices indicate that the IDMT-Traffic dataset has a higher number of true positives and true negatives, suggesting better overall classification performance.
- **Precision and Recall:** The performance metrics show that the IDMT-Traffic dataset has higher precision and recall values for both traffic types compared to the RTD dataset, aligning with the confusion matrix results.
- **F1-Score:** The F1-scores for the IDMT-Traffic dataset are also higher, indicating a better balance between precision and recall.

4.4.4 Evaluation Conclusion:

Both the confusion matrices and performance measurement indicate that the IDMT-Traffic dataset in term of classification accuracy, precision, recall, and F1-score. The results are consistent across both representations, confirming the superior performance of the LSTM model on the IDMT-Traffic dataset.

Additionally, a commonly utilised tool in conjunction with binary classifiers is the receiver operating characteristic (ROC) curve. Figure 9 depicts the Receiver Operating Characteristic (ROC) curves of LSTM for the IDMT-TRAFFIC and RTD datasets.

According to Figure 9, the area under the curve (AUC) of the classifier was higher for the first dataset compared to the first dataset. This difference can be attributed to the distinct circumstances between the IDMT-TRAFFIC and RTD.

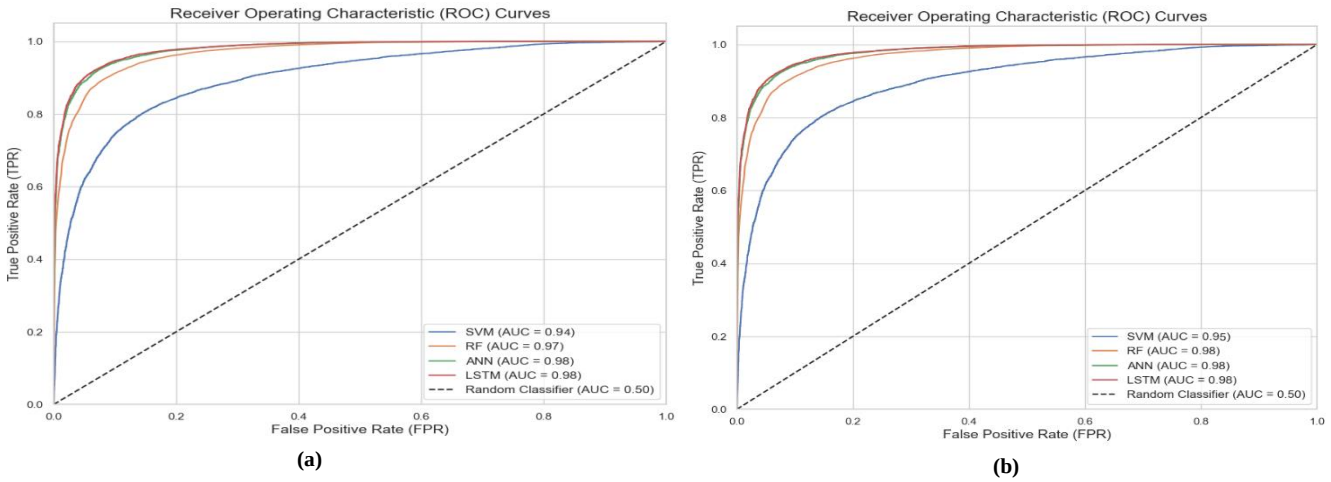

Figure9: The ROC curve for selected classifiers using (a) RTD and (b) IDMT-TRAFFIC datasets

It can be observed from Figure 9 (a) that the SVM has less value for AUC compared to the rest of the classifiers, while the LSTM model obtains an excellent result, which represents the greatest value of AUC using the RTD dataset.

In Figure 9 (b), The also (Support Vector Machine) SVM has less value for AUC compared to the rest of the classifiers, while the LSTM model obtains an excellent result, which represents the greatest value of AUC using IDMT-TRAFFIC dataset.

XIII. General comparison with related works

Acoustic traffic monitoring is a promising field, and further research is needed to meet the ever-increasing demands of cities. In our proposed research, we demonstrated that it is possible to expand the utility of acoustic datasets by introducing tasks that have not been explored in related work, such as congestion classification. We achieved this by using time-domain and frequency-domain features to extract more patterns from audio signals based on the RTD recording dataset, obtaining an accuracy of 0.9825. When evaluating our proposed method on benchmark dataset (IDMT-Traffic), we achieved an accuracy of 0.9869. Our findings illustrate the efficacy and promise of the suggested model for audio classification algorithms in accurately classifying traffic congestion within the realm of traffic monitoring. The table VI shows the comparison between the proposed model and the related work.

TABLEVI presents a comprehensive comparison between the proposed model and the related work.

Title and year	Dataset	Features	Classifier	Accuracy	Ref.
Web-Based Machine Learning Bi-Class Congestion Identification Based on Audio Data Application, 2022	Recording the road sounds	MFCC, Ci	SVM & Naïve Bayes	95% by SVM	[7]
Audio-Based Machine Learning Model for Traffic Congestion Detection, 2021	data from audio analysis of traffic registered	MFCC	ML	89%	[3]
Proposed method 2024	RTD	MFCC, Ci, SBW, and ZC	ANN, LSTM, SVM, RF	98% by LSTM	-
Proposed method 2024	IDMT-TRAFFIC	MFCC, Ci, SBW, and ZC	ANN, LSTM, SVM, RF	98% by LSTM	-

XIV. Conclusions

In recent years, the increasing number of active vehicles on city roads has resulted in severe traffic congestion, such that traditional traffic lights cannot meet a condition that has made them necessary as soon as AI techniques are used to make intelligent recommendations and improve traffic efficiency. Using different data collection methods, we can analyze and extract important traffic features from available road data. Based on these features, robust algorithms for traffic congestion detection and prediction can be developed. This paper reviewed recent developments in AI techniques, focusing on five existing papers that use acoustic datasets for traffic classification, and discussing data collection methods and preparation methods needed to use these datasets in AI-based systems. These datasets provide a valuable platform for testing new algorithmic approaches to domain optimization and provide opportunities to further improve traffic management systems. We utilized two types of datasets, RTD and IDMT-Traffic, to validate and evaluate our proposed model. Also, we have tested

these datasets by using four types of algorithms (ANN, LSTM, SVM, and RF); the LSTM model a achieved the highest accuracy, where it is achieved 98.25% and 98.69% of accuracy for RTD and IDMT datasets alternately. For future research on the development of traffic jam categorization systems recording more general dataset that include variety of samples such as weather conditions (dry, wet, and rainy). Additionally, road quality for instance (Rough, smooth, and grassy road) and its impact on traffic and thus the occurrence of congestions based on audio data analysis for a more robust transportation management system.

References:

[1] J. Zhang, F.-Y. Wang, K. Wang, W.-H. Lin, X. Xu, and C. Chen, "Data-driven intelligent transportation systems: A survey," IEEE Trans. Intell. Transp. Syst., vol. 12, no. 4, pp. 1624–1639, 2011.
[2] Y. Kim and J. Hong, "Urban traffic flow prediction system using a multifactor pattern recognition model," IEEE Trans. Intell. Transp. Syst., vol. 16, no. 5, pp. 2744–2755, 2015.
[3] R. C. Gatto and C. H. Q. Forster, "Audio-Based Machine Learning Model for Traffic Congestion Detection," IEEE Trans. Intell. Transp. Syst., vol. 22, no. 11, pp. 7200–7207, 2021, doi: 10.1109/TITS.2020.3003111.
[4] P. Borkar, L. G. Malik, and M. V. Sarode, "Acoustic signal based traffic density state estimation using adaptive neuro-fuzzy classifier," WSEAS Trans. Signal Process., vol. 10, no. 1, pp. 51–64, 2014.
[5] M. K. J. Al-Obaidi, "Improvement of the Traffic Management of Deactivated Al-Faris Al-Arabi Signalized Roundabout in Baghdad City," IOP Conf. Ser. Mater. Sci. Eng., vol. 518, no. 2, 2019, doi: 10.1088/1757-899X/518/2/022016.
[6] M. Ashhad, U. Goenka, A. Jagetia, P. Akhtari, S. K. Ambat, and M. Samuel, "Improved Vehicle Sub-type Classification for Acoustic Traffic Monitoring," 2023 Natl. Conf. Commun. NCC 2023, 2023, doi: 10.1109/NCC56989.2023.10067994.
[7] A. P. Perdana, P. H. Gunawan, and N. Aquarini, "Web-Based Machine Learning Bi-Class Congestion Identification Based on Audio Data Application," in 2022 2nd International Conference on Intelligent Cybernetics Technology & Applications (ICICyTA), IEEE, 2022, pp. 278–282.
[8] C.-Y. Chiang, M. Jaber, and P. Hayward, "A distributed acoustic sensor system for intelligent transportation using deep learning," arXiv Prepr. arXiv2209.05978, 2022.
[9] A.-C. Yang and E. D. Goodman, "Audio Classification of Accelerating Vehicles." Stanford University: Stanford, CA, USA, 2019.
[10] J. Abeßer, S. Gourishetti, A. Kátai, T. Clauß, P. Sharma, and J. Liebetrau, "IDMT-Traffic: An Open Benchmark Dataset for Acoustic Traffic Monitoring Research," Eur. Signal Process. Conf., vol. 2021-Augus, no. April, pp. 551–555, 2021, doi: 10.23919/EUSIPCO54536.2021.9616080.
[11] L. Galleani, L. Cohen, and D. Nelson, "Local signal to noise ratio," in Advanced Signal Processing Algorithms, Architectures, and Implementations XVI, SPIE, 2006, pp. 201–209.
[12] F. K. Al-Dhaher, D. Y. Mohammed, and M. Khalaf, "The Most Important Features of Lie Detection Using Voice Stress ", IJSER, vol. 3, no. 1, pp. 93–110, Mar. 2024.
[13] T. Ali Abdalkareem, K. A. Zidan, and A. S. Albahri, "A Systematic Review of Adversarial Machine Learning and Deep Learning Applications", IJSER, vol. 3, no. 4, pp. 14–40, Dec. 2024.