Proceedings of the ASME 2025
International Mechanical Engineering Congress and Exposition
IMECE2025
November 16-20, 2025, Memphis, Tennessee

IMECE2025-167248

MACHINE LEARNING AND SENSORY INTEGRATION FOR REAL-TIME ROAD SURFACE ASSESSMENT

Artem Abzaliev[1], Rutchanon Hatasen[2], Hussein Kokash[1,2], Linda Zhu[2], Jun Chen[3], Mihai Burzo[2]

[1]Computer Science and Engineering, University of Michigan, Ann Arbor, MI 48109
[2]Mechanical Engineering, University of Michigan, Flint, MI 48502
[3] Electrical and Computer Engineering, Oakland University, Rochester, MI 48309

ABSTRACT

Road safety and vehicle efficiency are important concerns in our society and are heavily influenced by the condition of the road surface. The National Highway Traffic Safety Administration indicates that 22% of vehicular accidents are weather-related, resulting in substantial economic and human losses. Slippery roads due to rain, snow, and ice, can affect a vehicle's functionality and the safety of its occupants. Knowing the type of road surface (rough, smooth, etc) the car is traveling on is important for emergency braking and stability control. In this investigation, we address these issues by effectively detecting and analyzing various road conditions in real-time, providing valuable information to the vehicle's control systems, which is essential for enhancing vehicle control systems, including braking and stability controls, as well as better range prediction of electric vehicles.

We developed a system that utilizes a combination of machine learning algorithms and hardware to deliver road surface assessments. The system classifies the road condition such as rough, or slippery surfaces through tire-pavement sound and vehicle vibrations. Our methodology includes developing machine learning models optimized for processing multimodal data from microphones and accelerometers, resulting in accurate and timely road condition classification. The system consists of hardware embedded into an existing Lincoln MKZ platform. The test data was collected using a data acquisition system consisting of a custom-made National Instruments LabVIEW interface connected via four channels to the following sensors: an accelerometer attached to the wheel well under the front hood, and three microphones—one next to the front wheel, one inside the vehicle, and one near the back wheel of the vehicle. Real-time signals were collected while driving on various highways and roads in Michigan that have different road conditions, and at different ranges of speeds.

Given the raw waveform of one of the 4 sensors (accelerometer or microphones), we train a baseline machine learning model on the collected dataset to classify road type. Our architecture is inspired by human speech processing: it combines convolutional encoders to learn feature representations and a deep transformer model to classify the compressed sequence. Our preliminary test results indicate that the model can distinguish between arterial and dirt roads with 92% accuracy.

Embedding this technology into vehicles will enhance the vehicle's control systems, facilitate V2V and V2X communication for traffic control and improved safety, and provide both the onboard systems and drivers with crucial road condition information for better decision-making and automated responses.

Keywords: Road surface condition, Machine Learning, vibration and acoustic signal.

1. INTRODUCTION AND RELATED LITERATURE

The condition of the road surface is a foundational element influencing the safety, efficiency, and longevity of transportation systems. Accurate and timely assessment of road characteristics, ranging from material type and texture to the presence of contaminants or anomalies, has become increasingly critical with the advent of advanced vehicle technologies and the ongoing need for optimized infrastructure management. Real-time knowledge of the surface condition significantly impacts vehicle control and safety, particularly given that adverse conditions like wet pavement, snow, and ice are major contributors to accidents. Advanced Driver-Assistance Systems (ADAS) and Autonomous Vehicles

(AVs) require precise surface information (e.g., type, condition, friction) to adapt control strategies for braking, traction, steering, and path planning [1, 2]. Furthermore, automated monitoring using vehicle-mounted sensors offers a scalable alternative to traditional manual surveys for detecting pavement distress (e.g., potholes, cracks) and assessing overall road quality, facilitating proactive maintenance and extending infrastructure lifespan [3, 4].

The task of classifying road surfaces relies on capturing information that distinguishes different surface types and conditions. A diverse array of sensing technologies has been explored, studying different physical principles to probe the road surface or the interaction between the vehicle and the road. These modalities vary significantly in terms of cost, resolution, the specific surface properties they are sensitive to, and their robustness under challenging environmental conditions. One prominent approach involves measuring the vibrations induced in the vehicle as it travels over the road surface, vibration-based sensing, primarily using accelerometers and gyroscopes often within Inertial Measurement Units (IMUs), measures vehicle vibrations induced by road texture, roughness, and anomalies [1]. While effective for detecting localized defects and assessing roughness [1, 5], vibration signals are highly sensitive to vehicle speed, suspension characteristics, weight, and sensor placement, requiring careful preprocessing like filtering and coordinate reorientation to mitigate noise and confounding factors [1, 6]. Studies report high classification accuracy for surface types [7, 5], but achieving consistent real-world generalization remains challenging. Another modality focuses on the sound generated at the interface between the vehicle's tires and the road surface, commonly referred to as Tire-Pavement Interaction Noise (TPIN), captured by microphones near the wheel wells, inside the tire, or within the cabin [8, 9]. TPIN characteristics reflect surface macrotexture, microtexture, porosity, and contaminants [9]. This modality can be robust in poor visibility [9] and has shown high accuracy in classifying conditions like dry vs. wet asphalt or asphalt vs. snow [1, 9, 10]. However, TPIN is susceptible to ambient noise and strongly influenced by vehicle speed and tire properties [1, 9]. Distributed Acoustic Sensing (DAS) using roadside fiber optics presents an alternative for large-scale monitoring [3]. An alternative framework employs vision-based sensors, including cameras and Light Detection and Ranging (LIDAR), which provide rich visual or geometric information. Cameras capture texture, color, and reflectivity for identifying materials, conditions, and defects [2], while LIDAR creates 3D point clouds for precise profile and roughness measurements. Implementations range from vehicle-mounted cameras and LIDAR to specialized survey vehicles, drones, and fixed roadside cameras [2, 3, 11]. Deep learning, especially Convolutional Neural Networks (CNNs) and object detectors (e.g., YOLO), excels at analyzing image data for crack/pothole detection and surface classification [4, 3, 2]. However, vision systems are highly vulnerable to adverse weather (rain, snow, fog) and poor lighting conditions, and can suffer from occlusion and motion blur [9]. Given the complementary strengths and weaknesses of individual sensor types—vision's detail vs. weather sensitivity, vibration/acoustic interaction capture vs. speed/noise sensitivity—multimodal sensor fusion is crucial for robust, all-weather systems [1]. Combining data from sensors like cameras with LIDAR, or vision with vibration/acoustic sensors, aims to create a perception more reliable than any single modality, essential for safety-critical ADAS/AV applications [12].

Raw data acquired from sensors monitoring road surfaces is rarely suitable for direct input into classification algorithms. It is often contaminated by noise, influenced by extraneous factors like vehicle speed or sensor orientation, and may not be in a format conducive to analysis by standard machine learning models. Therefore, preprocessing and feature representation are critical steps to enhance signal quality, isolate relevant information pertaining to the road surface, and transform the data into a suitable format for subsequent classification. Vibration signals often undergo high-pass filtering and coordinate reorientation [6], while acoustic data is typically filtered, segmented into frames, and windowed [8, 10]. Image data preprocessing may involve region-of-interest extraction, resizing, and enhancement [13, 9]. Classical machine learning approaches relied on handcrafted features extracted from these preprocessed signals in the time domain (e.g., mean, variance, RMS, crest factor [1]), frequency domain (e.g., FFT, PSD, MFCCs [1]), or time-frequency domain using methods like the Short-Time Fourier Transform (STFT) producing spectrograms [10] or the Continuous Wavelet Transform (CWT) yielding scalograms with variable resolution [9, 6]. Deep learning (DL) has largely superseded classical methods due to its ability to learn relevant features automatically from data. CNNs are widely used, processing camera images directly [2] or analyzing 1D sensor data (vibration, acoustic) transformed into 2D image-like representations such as spectrograms or scalograms [9, 10, 14]. Additionally, CNN architectures have been widely employed to classify road surfaces into categories such as dry, wet, icy, rough, snowy, muddy, and curvy through various image-based datasets, both public and proprietary [15–19]. Some models extend this work by incorporating more complex architectures like ResNet, InceptionNet, and ConvNeXt for improved feature extraction and classification across more diverse surface types, including asphalt, gravel, concrete, sand, grass, and others [20, 19, 21, 22]. Recent efforts have increasingly focused on using fusion models or hybrid CNN-transformer frameworks to combine multi-sensor data, specifically from lidar and cameras. These approaches aim to enhance the recognition of road textures and conditions across diverse and variable environments. By integrating lidar's distance measurements with the visual details from cameras, these models achieve a more robust understanding of the scene. As highlighted in references [23, 24], these techniques effectively advance autonomous systems by improving their accuracy and adaptability in challenging road and weather conditions.

Table 1: Summary of recent deep learning studies for road surface classification

Reference	Sensor Modality(ies)	Data Representation	DL Model	Classification Task	Reported Accuracy/Metric	Key Limitation Noted
Yoo et al. [9]	Acoustic (TPIN)	CWT Image (2D)	CNN	Asphalt/Snow (with tire types)	>90% Acc	Limited speed range (~40 km/h)
Abdić et al. [10]	Acoustic (Tire-Surface)	Mel Spectrogram +Δ	BLSTM (RNN)	Wet/Dry	93.2% UAR	Performance dips at low speeds (<3mph)
Ozoglu [16]	Vibration (Accel+Gyro)	Pixel Map (6x4000)	CNN (5 variants tested)	Pothole Detection	93.2% Val Acc; 80-87% Field Acc	Tested in consistent conditions (dry, asphalt, <50 km/h)
Maeda et al. [12]	Smartphone camera	RGB images (600x600)	SSD (Single Shot MultiBox Detector)	Multi-class road damage classification	High accuracy (validated on smartphone & GPU)	Single modality; performance varies with lighting and road context
Menegazzo [7]	Vibration (Accel+Gyro)	Time-Series	CNN	Surface Type (Dirt/Cobble/Asphalt)	93.17% Val Acc	Generalization across vehicles/drivers/ envs challenging
Šabanovič et al. [2]	Vision (Camera)	Image	CNN (Modified AlexNet)	Type (3) & Condition (2)	88.3% Test Acc (6 classes)	Misclassifications between similar types/conditions
Wang et al. [42]	Camera (UAV, virtual &real images)	RGB images (synthetic + real)	Enhanced YOLOv8	Multi-class pavement distress	MAP of 94.8%	Heavy reliance on synthetic data
Pakkala et al. [33]	Acoustic (Sound) + Vib.	Spectrogram	CNN + BLSTM	Road Anomaly Events (Potholes)	83% Acc (ADAM optimizer)	Accuracy sensitive to optimizer choice
Moroto et al. [11]	Vision (Camera) + Aux Data	Image + Time-Series	Multimodal Transformer (MMT)	Winter Conditions (7 classes)	Improved over MLP (fusion)	Focus on fixed cameras, specific winter conditions
Trinh [13]	Vision (Camera) + LIDAR	Image + Point Cloud	CNN-Transformer Hybrid	Surface Type (4 classes)	High accuracy on private dataset	Lack of public benchmark data
Shi et al. [23]	Camera + Intelligent tire (piezoelectric tactile sensor)	Image + CWT-transformed tactile spectrogram	CNN-Transformer	Road surface classification (4 types)	99.48% accuracy	Only 4 road types considered
Current work	Acoustic (3 microphones) + Vibration	Raw waveform	wav2vec2	Surface Type (Asphalt, Dirt road, Concrete)	92.9% accuracy	Generalization across all road types and weather conditions

Transfer learning has also proven effective in this domain, where pre-trained models such as VGG16, MobileNetV2, ResNet18, and ResNet34 are fine-tuned on road-specific datasets to improve generalization and reduce training time, particularly when working with limited or private data [25–29]. These approaches have been tested across various countries and conditions, covering classifications like good, bad, medium, unpaved, or pavement distress indicators such as bleeding, raveling, and cracks [30, 31, 29]. Supplementary modalities, such as acoustic signal processing and accelerometer data, have also been used for classifying road quality, typically relying on either CNNs or self-organizing maps (SOMs) for processing non-visual features [32]. Recurrent Neural Networks (RNNs), particularly LSTMs and GRUs, are suited for sequential data, used for tasks like audio-based wetness detection [10] or identifying anomalies from sound [33]. Some studies suggest RNNs can be efficient. Transformers, leveraging self-attention for long-range dependencies [34], are an emerging trend applied to traffic forecasting [35], acoustic classification [36], and multimodal fusion [11]. Research suggests Transformers are achieving competitive performance [37], sometimes in hybrid CNN-Transformer architectures [13].

Table 1 summarizes several key deep learning studies, highlighting their methodologies and reported performance. Sensor fusion strategies combine data at the data, feature, or decision level [38, 9]. Deep learning facilitates advanced feature-level fusion using techniques like attention mechanisms and dedicated fusion layers. Multimodal Transformers (MMTs), for instance, use cross-attention to dynamically weigh features from different modalities, enhancing integration [11]. Common fusion combinations include vision with inertial/acoustic data [7], acoustic with vibration data [14, 33], and vision with auxiliary data like temperature or weather forecasts [11]. While fusion promises enhanced robustness [12], its implementation adds complexity, and rigorous benchmarking comparing state-of-the-art single-modality systems against various fusion approaches on standardized datasets is needed to quantify the true benefits.

Despite progress, persistent challenges remain. The lack of large-scale, diverse, standardized public datasets severely hinders model development, benchmarking, and reproducibility [11]. Generalization across different vehicles, speeds, sensor placements, tire types, and environmental conditions is a major hurdle [9, 7]. Real-time performance is essential for safety applications, requiring research into model compression, efficient architectures, and edge computing [9]. Robustness against environmental factors like weather and noise necessitates better sensors, sophisticated fusion, and advanced signal processing [1, 39]. The "black-box" nature of complex DL models raises interpretability concerns, demanding research into explainable AI (XAI) for safety-critical systems [6]. Future work should also move beyond discrete classification towards quantitative estimation of parameters like friction coefficient [2], roughness indices [1], or anomaly severity [6].

This paper investigates real-time road surface assessment using machine learning combined with acoustic and vibration data. Implemented on a Lincoln MKZ with microphones and an accelerometer, the system classifies surfaces like asphalt, concrete, and dirt, aiming to improve vehicle safety and efficiency. This multimodal approach offers a robust solution that supports the future of autonomous driving technologies.

2. DATA COLLECTION

Data was collected using a Lincoln MKZ vehicle, shown in Fig. 1. The sensors were mounted near the front and rear wheels and inside the car (microphones) and under the hood (vibration sensor). The data acquisition process was enhanced by using a custom-made National Instrument LabVIEW VI application.

Audio signals and vibrations were recorded on three common road surfaces: asphalt, concrete, and dirt roads. Speed during the recording process was consistently maintained to avoid being a confounding factor. For each road type, four different speeds were operated: 30, 35, 40, and 55 mph.. For each combination of speed and road, approximately 10 minutes of driving data was recorded. Figure 2 shows mel-spectrograms for the dirt road and the concrete road. Note that the spectrograms are for visualization purposes only; we use raw waveform data in our experiments.

All of the road vibration and acoustic recordings were collected by the authors in Ann Arbor, Michigan, USA.

FIGURE 1: Lincoln MKZ platform used for the data collection.

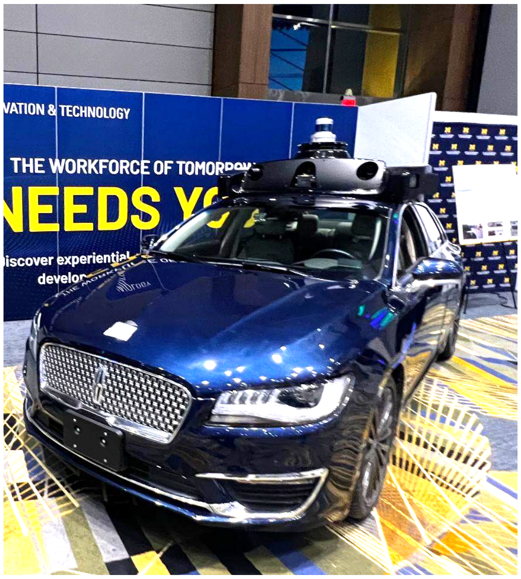
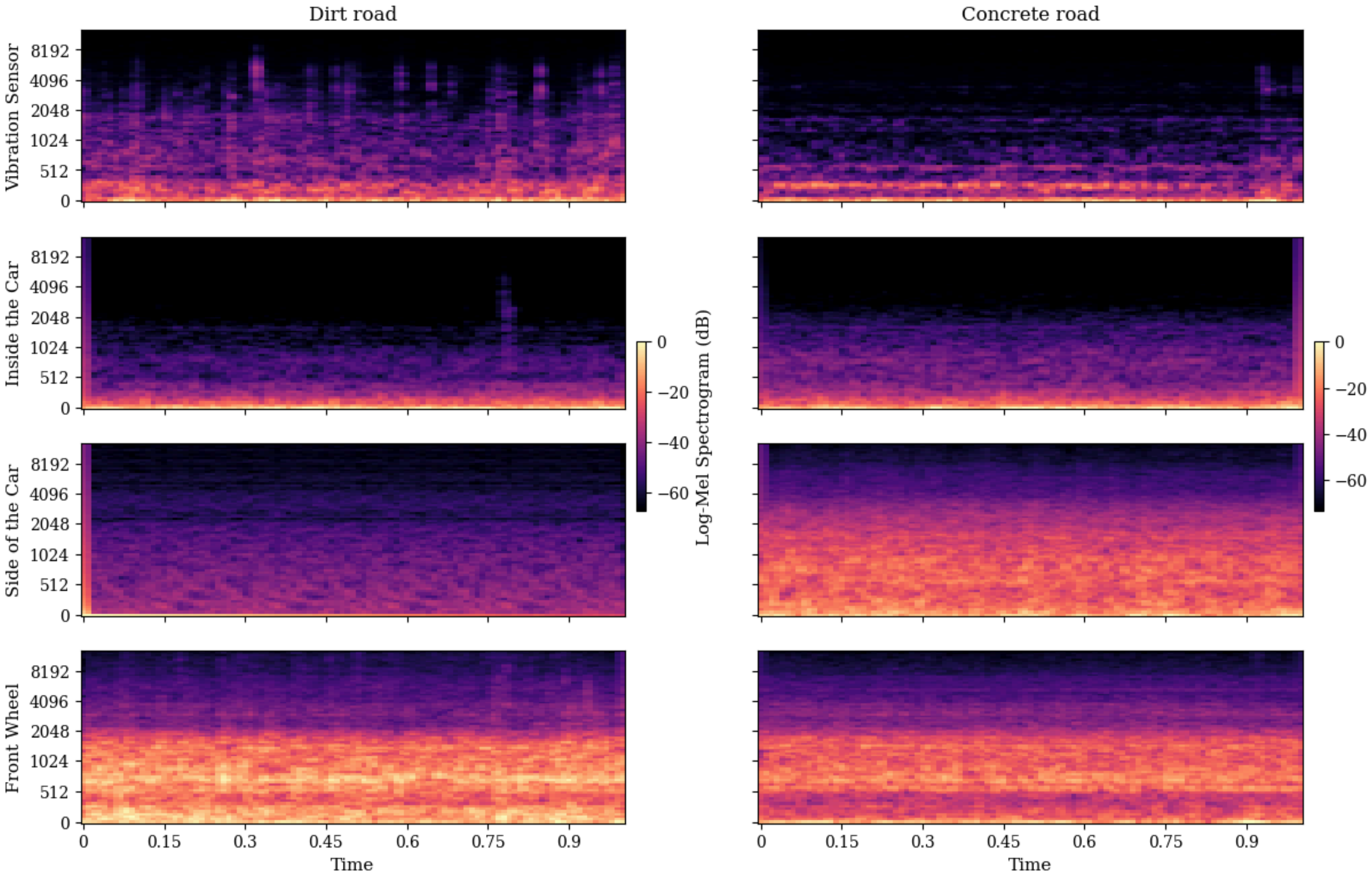


FIGURE 2: Mel-Spectograms for 2 random 1-second chunks, for all 4 inputs used in our experiments. Dirt road vibrations have higher frequencies.



3. EXPERIMENT SETUP

The data was split into random 1-second audio clips, resulting in a total of 5232 chunks. This segmentation approach ensured that the machine learning model had access to manageable and uniform data inputs, facilitating effective training and validation processes. Label distribution is depicted in Fig 3. Figure 3 visually represents the label distribution across these audio clips, shedding light on the diversity and balance of the dataset in terms of road surface types and driving conditions. During data collection, the vehicle was operated at varying speeds ranging from 30 to 55 miles per hour (mph) to simulate a range of realistic driving scenarios. However, the vehicle was not driven at the highest speed of 55 mph on the dirt road surface, which could have posed potential safety hazards.

We considered two types of experiments:

1. We split the data into an 85% training dataset and a 15% test set.
2. We set one speed aside as a test set, while training on all other speeds.

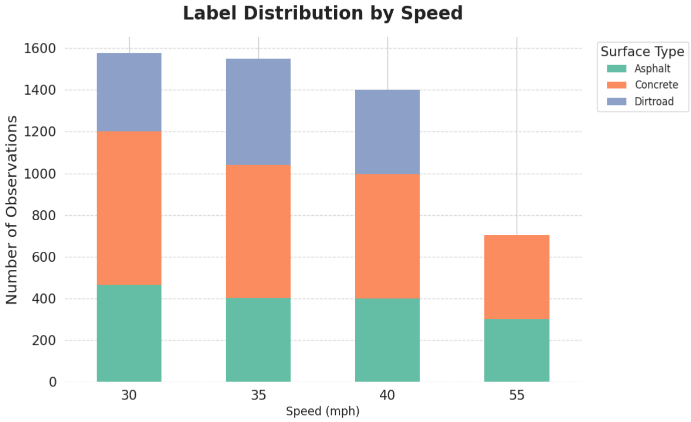


FIGURE 3: Label distribution across different speed conditions. The collected dataset has approximately balanced class distributions within each speed group.

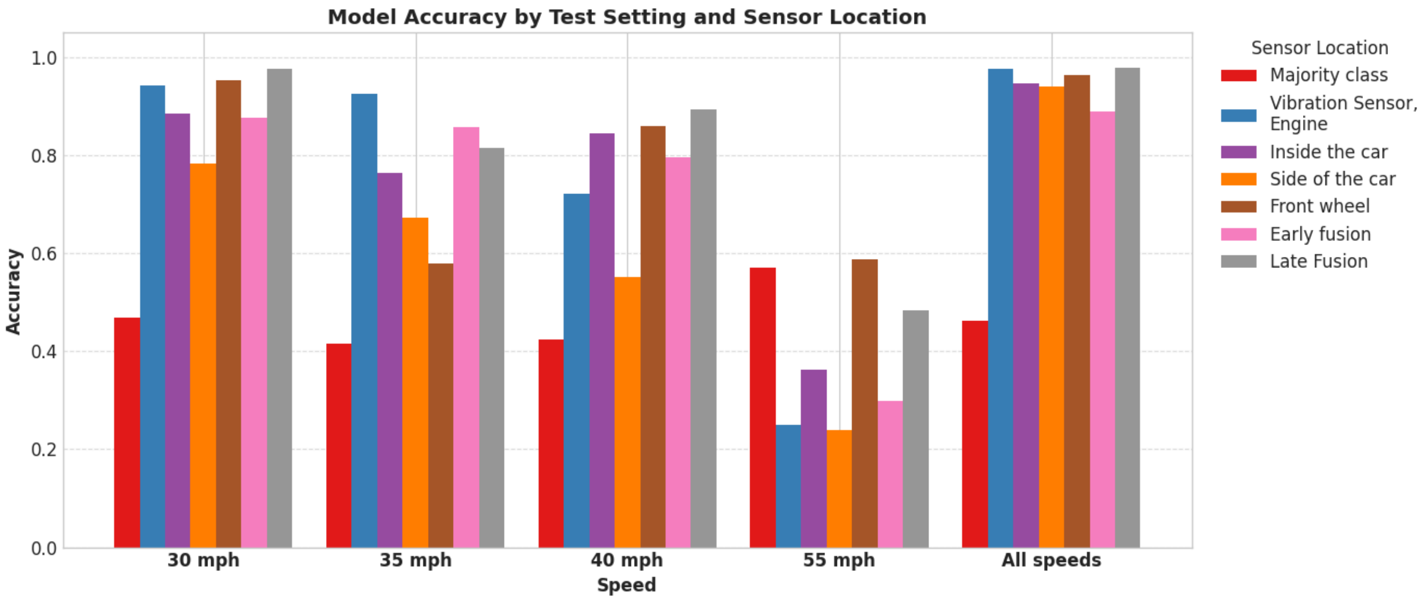


FIGURE 4: Results for models trained on different sensors/fusion of sensors. We report accuracy for various test scenarios. Sensor performance is scenario-dependent; no configuration is universally best.

In our experiments, we examined the impact of four different input types on road surface classification accuracy: a vibration sensor placed under the hood to capture detailed vehicle vibrations, and three microphones strategically positioned—one inside the car, one at the side near the front wheel, and one at the front wheel itself. Each sensor provided unique insights into the vehicle's interactions with the road surface. To optimize data use, we employed two fusion strategies: early fusion, where data from all sensors were combined before being input into the machine learning model, and late fusion, where each sensor's data was processed independently before integration for the final classification. This approach enabled us to assess whether merging diverse data streams at different stages could enhance classification accuracy, helping to identify the optimal configuration for a robust, real-time road surface assessment system.

3.1 Model

We conducted the experiments using the wav2vec2 model [43]. While the model is initially developed for speech recognition, we found that it can be a good initialization checkpoint of the weights of the model, resulting in better performance, which is in line with the previous findings. [44, 45]

For single input, we used the identical preprocessing as in wav2vec2 - normalized mono audio channel sampled at 16 kHz. For the fusion, we first normalize each channel independently and then concatenate them.

For early fusion, we treated each audio channel independently. We first normalized each channel and then stacked them together. We replaced the first convolutional layer of wav2vec2 to handle 4 channels instead of one, keeping the rest of the architecture. For the late fusion experiment, we ran the prediction of each channel independently, and then averaged the resulting features before passing it into the classification head.

We fine-tuned the model for 5 epochs with a learning rate of 5e-5 using the AdamW optimizer. The total training time for 5 epochs with batch size 16 is approximately 1 hour on a single Nvidia L40S GPU. During inference, it takes 0.59 seconds to predict a single batch of size 16, i.e. 37 milliseconds per 1-second audio. In a non-batched mode (batch size=1, simulating real-time inference scenario) it takes 73 milliseconds to predict 1-second audio (excluding preprocessing time).

RESULTS AND DISCUSSION

The results for our experiments are shown in Table 2 and Figure 4. We also show the majority class as a baseline, i.e., simply predicting the majority class. Almost all the experiments significantly outperform the majority approach, showing the feasibility of our approach. The observed patterns are described next.

First, no single sensor location consistently outperforms the others across all test scenarios, indicating that performance varies depending on the test data setup. Interestingly, the accelerometer, when treated as a microphone, performs on par with traditional audio sensors, suggesting that no additional modality-specific treatments are necessary.

Second, as the speed increased for the test data, the model's performance deteriorated. We hypothesize that this decline is due to a decreasing signal-to-noise ratio at higher speeds; specifically, at 55 mph, the microphones primarily picked up wind noise instead of road-specific acoustic

Table 2: Results for different validation sets and different sensors. We consider 4 scenarios: train on all but withhold speed, and evaluate on all the observations for this particular speed, and additionally a random 85% train-15% test split.

Accuracy					
Sensor location	Test dataset specification				
	30 mph	35 mph	40 mph	55 mph	15% random split
Vibration Sensor, Engine	95.1%	93.1%	75.3%	57.3%	97.7%
Inside the car	90.1%	79.8%	86.4%	57.0%	95.0%
Side of the car	85.0%	72.7%	70.5%	55.8%	94.7%
Front wheel	95.4%	68.3%	86.3%	88.3%	96.5%
Majority class	46.8%	41.6%	42.4%	57.0%	46.2%
Early fusion	90.1%	87.1%	82.0%	59.7%	89.9%
Late Fusion	98.1%	83.4%	90.3%	75.17%	97.9%

Table 3: Results for different validation sets and different sensors. We consider 4 scenarios: train on all but withhold speed, and evaluate on all the observations for this particular speed, and additionally a random 85% train-15% test split.

F1 scores					
Sensor location	Test dataset specification				
	30 mph	35 mph	40 mph	55 mph	15% random split
Vibration Sensor, Engine	94.2%	92.4%	72.1%	24.9%	97.7%
Inside the car	88%	76%	84%	36%	94%
Side of the car	78%	67%	55%	23%	93%
Front wheel	95%	57%	85%	58%	96%
Majority class	21%	19%	19%	36%	20%
Early fusion	87%	85%	79%	29%	88%
Late Fusion	97%	81%	89%	48%	97%

information. This was anticipated to some extent, as the model was trained on data from lower speeds (30, 35, and 40 mph), which had less wind noise. Consequently, the lack of wind noise data in the training set means the model encounters difficulties accurately classifying road surfaces when confronted with the wind-dominated audio profiles present at 55 mph. Conversely, at 35 mph, the microphones captured more road-specific sounds with less interference, contributing to better classification accuracy at lower speeds.

Third, late fusion consistently outperformed early fusion. This is likely because early fusion introduces substantial modifications throughout the whole wav2vec2 fine-tuning process, whereas late fusion only requires the final classification layer to learn how to combine outputs from each sensor-specific model. Nevertheless, it is noteworthy that late fusion performs on par with the best single-sensor models, which already exhibit very high accuracy, suggesting that the individual sensors effectively capture most of the relevant signals on their own. As a result, the benefits of fusing multiple signals are limited and should be carefully weighed against the additional computational overhead introduced by late fusion.

4. CONCLUSION

In this paper, we conducted an early study on using machine learning to analyse the sound and vibration from a vehicle to predict the road surface that the vehicle is travelling on. The road surface classification can be directly associated with the coefficient of friction for that specific road section, providing real-time information to the car's embedded systems about the level of adherence to the road. The ability to link road surface classification with the coefficient of friction provides vehicles with critical real-time information, enhancing control systems and decision-making processes. The data was acquired on three types of road surfaces: asphalt, concrete, and dirt roads in Ann Arbor, Michigan, USA. Our results showed that machine learning can be successfully used to predict road surface conditions from the audio and vibration signals alone. We tested various speeds/road conditions and showed that our approach achieves up to 96% accuracy. We also experiment with fusing different signals. Our results indicate that road conditions possess distinct patterns that machine learning models can learn and recognise. Through the integration of multiple sensor inputs and fusion strategies, the research underlined the importance of multimodal data for improving the reliability of road surface assessments. Looking ahead, future experiments will expand this study by incorporating additional sensor modalities, such as video data and measurements of tire pressure and temperature distribution. This expanded approach will not only refine the system's accuracy but also enable a comprehensive analysis of more varied weather and road conditions, including dry, wet, and icy surfaces, thereby broadening the applicability and robustness of the technology. These advancements will further strengthen the role of machine learning in the ongoing development of autonomous and advanced driver-assistance systems, paving the way for safer and more efficient transportation.

5. LIMITATIONS

In this setting, one road is equal to one road type. While we think the road conditions are significantly different, it is possible that the model learned individual features of the specific **section of the road** that was tested, instead of general features for that **type of road**. I.e., if the dirt road has a lot of potholes, the model might exploit this information. If there is a specific section of the road with a lot of potholes that is different from what was tested, the performance might degrade. We didn’t explicitly test this limitation as it was both impractical and beyond the scope of the work at this relatively early stage. However, we anticipate that expanding to a much larger and more varied dataset in future studies will mitigate this issue by enhancing the model's generalization capabilities.

ACKNOWLEDGEMENTS

This work is supported by the ADVANCE Grant Program sponsored by Michigan Economic Development Corporation (MEDC) on behalf of all public universities throughout the State of Michigan.

REFERENCES

[1] Andrades, Ignacio Sánchez, et al. "Low-cost road-surface classification system based on self-organizing maps." Sensors 20.21 (2020): 6009.

[2] Šabanovič, Eldar, et al. "Identification of road-surface type using deep neural networks for friction coefficient estimation." Sensors 20.3 (2020): 612.

[3] Choo, Kan Yeep, et al. "Machine Learning-Based Classification of Hand Tool Vibrations Using Distributed Fiber Optic Sensors for Road Health Monitoring." 2024 Multimedia University Engineering Conference (MECON). IEEE, 2024.

[4] Botezatu, Adrian-Paul, Adrian Burlacu, and Ciprian Orhei. "A review of deep learning advancements in road analysis for autonomous driving." Applied Sciences 14.11 (2024): 4705.

[5] Surblys, Vytenis, et al. "Accelerometer-Based Pavement Classification for Vehicle Dynamics Analysis Using Neural Networks." Applied Sciences 14.21 (2024): 10027.

[6] Martinez-Ríos, Erick Axel, Martin Rogelio Bustamante-Bello, and Luis Alejandro Arce-Sáenz. "A review of road surface anomaly detection and classification systems based on vibration-based techniques." Applied Sciences 12.19 (2022): 9413.

[7] Menegazzo, Jeferson, and Aldo Von Wangenheim. "Road surface type classification based on inertial sensors and machine learning: A comparison between classical and deep machine learning approaches for multi-contextual real-world scenarios." Computing 103.10 (2021): 2143-2170.

[8] Ramos-Romero, Carlos, et al. "Urban road surface discrimination by tire-road noise analysis and data clustering." Sensors 22.24 (2022): 9686.

[9] Yoo, Jinhwan, et al. "Classification of road surfaces based on CNN architecture and tire acoustical signals." Applied Sciences 12.19 (2022): 9521.

[10] Abdić, Irman, et al. "Detecting road surface wetness from audio: A deep learning approach." 2016 23rd International Conference on Pattern Recognition (ICPR). IEEE, 2016.

[11] Moroto, Yuya, et al. "Multimodal Transformer Model Using Time-Series Data to Classify Winter Road Surface Conditions." Sensors 24.11 (2024): 3440.

[12] Khanmohamadi, Masoud, and Marco Guerrieri. "Advanced sensor technologies in CAVs for traditional and smart road condition monitoring: A review." Sustainability 16.19 (2024): 8336.

[13] Trinh, Linh, Ali Anwar, and Siegfried Mercelis. "Improving classification of road surface conditions via road area extraction and contrastive learning." arXiv preprint arXiv:2407.14418 (2024).

[14] Ozoglu, Furkan, and Türkay Gökgöz. "Detection of road potholes by applying convolutional neural network method based on road vibration data." Sensors 23.22 (2023): 9023.

[15] V. Pereira, S. Tamura, S. Hayamizu, and H. Fukai, “Classification of paved and unpaved road image using convolutional neural network for road condition inspection system,” in 2018 5th International Conference on Advanced Informatics: Concept Theory and Applications (ICAICTA), 2018, pp. 165–169.

[16] D. K. Dewangan and S. P. Sahu, “Rcnet: road classification convolutional neural networks for intelligent vehicle system,” Intelligent Service Robotics, vol. 14, no. 2, pp. 199–214, 2021.

[17] L. Cheng, X. Zhang, and J. Shen, “Road surface condition classification using deep learning,” Journal of

Visual Communication and Image Representation, vol. 64, p. 102638, 2019.

[18] M.-H. Kim, J. Park, and S. Choi, “Road type identification ahead of the tire using d-cnn and reflected ultrasonic signals,” International Journal of Automotive Technology, vol. 22, pp. 47–54, 2021.

[19] M. Nolte, N. Kister, and M. Maurer, “Assessment of deep convolutional neural networks for road surface classification,” in 2018 21st International Conference on Intelligent Transportation Systems (ITSC), IEEE, 2018, pp. 381–386.

[20] J. Balcerek, A. Konieczka, K. Piniarski, and P. Pawłowski, “Classification of road surfaces using convolutional neural network,” in 2020 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA), 2020, pp. 98–103.

[21] H. Zhang, Z. Li, W. Wang, L. Hu, J. Xu, M. Yuan, Z. Wang, Y. Ren, and Y. Ye, “Multi-supervised bidirectional fusion network for road-surface condition recognition,” PeerJ Computer Science, vol. 9, p. e1446, 2023.

[22] T. Zhao, J. He, J. Lv, D. Min, and Y. Wei, “A comprehensive implementation of road surface classification for vehicle driving assistance: Dataset, models, and deployment,” IEEE Transactions on Intelligent Transportation Systems, vol. 24, no. 8, pp. 8361–8370, 2023.

[23] R. Shi, S. Yang, Y. Chen, R. Wang, M. Zhang, J. Lu, and Y. Cao, “Cnn-transformer for visual-tactile fusion applied in road recognition of autonomous vehicles,” Pattern Recognition Letters, vol. 166, pp. 200–208, 2023.

[24] C. Tian, D. Jin, B. Leng, and L. Xiong, “Reliable identification of road surface condition considering shadow interference,” in 2021 IEEE International Intelligent Transportation Systems Conference (ITSC), 2021, pp. 251–257.

[25] A. A. Hasanaath, A. Moinuddeen, N. Mohammad, M. A. Khan, and A. A. Hussain, “Continuous and realtime road condition assessment using deep learning,” in 2022 International Conference on Connected Systems and Intelligence (CSI), 2022, pp. 1–7.

[26] A. Chaudhary and P. Dr., “Road surface quality detection using light weight neural network for visually impaired pedestrian,” Evergreen, vol. 10, no. 2, pp. 706–714, 2023.

[27] Y.-A. Hsieh and Y. J. Tsai, “Automated asphalt pavement raveling detection and classification using convolutional neural network and macrotexture analysis,” Transportation Research Record, vol. 2675, no. 9, pp. 984–994, 2021.

[28] A. Riid, R. Louk, R. Pihlak, A. Tepljakov, and K. Vassiljeva, “Pavement distress detection with deep learning using the orthoframes acquired by a mobile mapping system,” Applied Sciences, vol. 9, no. 22, 2019.

[29] R. Stricker, M. Eisenbach, M. Sesselmann, K. Debes, and H.-M. Gross, “Improving visual road condition assessment by extensive experiments on the extended gaps dataset,” in 2019 International Joint Conference on Neural Networks (IJCNN), 2019, pp. 1–8.

[30] T. Rateke, K. A. Justen, and A. von Wangenheim, “Road surface classification with images captured from low-cost cameras – road traversing knowledge (rtk) dataset,” Revista de Informática Teórica e Aplicada (RITA), 2019.

[31] F. M. N. Sajad Ranjbar and H. Zakeri, “An image-based system for asphalt pavement bleeding inspection,” International Journal of Pavement Engineering, vol. 23, no. 12, pp. 4080–4096, 2022.

[32] A. Gagliardi, V. Staderini, and S. Saponara, “An embedded system for acoustic data processing and ai-based real-time classification for road surface analysis,” IEEE Access, vol. 10, pp. 63073–63084, 2022.

[33] Pakkala, Permanki Guthu Rithesh, et al. "Road safety analysis framework based on vehicle vibrations and sounds using deep learning techniques." International Journal of System Assurance Engineering and Management 15.3 (2024): 1086-1097.

[34] Khan, Salman, et al. "Transformers in vision: A survey." ACM computing surveys (CSUR) 54.10s (2022): 1-41.

[35] Chang, Ande, Yuting Ji, and Yiming Bie. "Transformer-based short-term traffic forecasting model considering traffic spatiotemporal correlation." Frontiers in Neurorobotics 19 (2025): 1527908.

[36] Whitaker, Steven, et al. "Through-ice acoustic source tracking using vision transformers with ordinal classification." Sensors 22.13 (2022): 4703.

[37] Saki, Mahdi, et al. "Precision Soil Quality Analysis Using Transformer-based Data Fusion Strategies: A Systematic Review." arXiv preprint arXiv:2410.18353 (2024).

[38] Rathee, Munish, Boris Bačić, and Maryam Doborjeh. "Automated road defect and anomaly detection for traffic safety: a systematic review." Sensors 23.12 (2023): 5656.

[39] Gong, Cihun-Siyong Alex, et al. "Deep learning with LPC and wavelet algorithms for driving fault diagnosis." sensors 22.18 (2022): 7072.

[40] Salazar, Addisson, et al. "On training road surface classifiers by data augmentation." Applied Sciences 12.7 (2022): 3423.

[41] Maeda, Hiroya, et al. "Road damage detection and classification using deep neural networks with smartphone images." Computer‐Aided Civil and Infrastructure Engineering 33.12 (2018): 1127-1141.

[42] Wang, Sicheng, et al. "Automated detection of pavement distress based on enhanced YOLOv8 and synthetic data with textured background modeling." Transportation Geotechnics 48 (2024): 101304.

[43] Baevski, Alexei et al. “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations.” ArXiv abs/2006.11477 (2020)

[44] Papadimitriou, Isabel and Dan Jurafsky. “Pretraining on Non-linguistic Structure as a Tool for Analyzing Learning Bias in Language Models.” ArXiv abs/2004.14601 (2020): n. pag.

[45] Jabbour, Sarah et al. “Deep Learning Applied to Chest X-Rays: Exploiting and Preventing Shortcuts.” *Machine Learning in Health Care* (2020).

Appendix A

Sensor location	Class	Test dataset specification				
		30 mph	35 mph	40 mph	55 mph	15% random split
Vibration Sensor	Dirt	0.9	0.99	0.62	0	0.99
	Asphalt	0.96	0.84	0.62	1	0.97
	Concrete	0.97	0.96	1	0.57	0.97
Engine Inside	Dirt	0.81	1	0.75	0	0.94
	Asphalt	0.93	0.57	0.84	0.57	0.95
	Concrete	0.94	0.97	0.97		0.96
Side of Car	Dirt	0.98	0.73	0.5	0	0.89
	Asphalt	0.68	0.65	0	0	0.95
	Concrete	0.98	0.78	1	0.57	0.98
Front Wheel	Dirt	0.95	0.86	0.99	0	0.96
	Asphalt	0.9	0.47	0.91	0.96	0.94
	Concrete	1	0.95	0.78	0.85	0.98
Early fusion	Dirt	0.92	0.85	0.76	0	0.95
	Asphalt	0.83	0.78	0.75	0.87	0.85
	Concrete	0.94	0.95	0.89	0.6	0.91
Late fusion	Dirt	0.98	1	0.9	0	0.98
	Asphalt	0.96	0.63	0.9	0.86	0.95
	Concrete	0.99	0.92	0.91	0.72	0.99

Table 4: Precision scores for different validation sets and different methods/sensors. We consider 4 scenarios: train on all but withhold speed, and evaluate on all the observations for this particular speed, and additionally a random 85% train-15% test split.

Sensor location	Class	Test dataset specification				
		30 mph	35 mph	40 mph	55 mph	15% random split
Vibration Sensor	Dirt	0.95	0.8	1	0	0.99
	Asphalt	0.88	0.99	0.42	0.01	0.95
	Concrete	1	1	0.81	1	0.99
Engine Inside	Dirt	0.96	0.42	0.95	0	0.97
	Asphalt	0.72	0.96	0.65	1	0.88
	Concrete	0.98	1	0.95		0.98
Side of Car	Dirt	0.38	0.27	1	0	0.95
	Asphalt	1	0.88	0	0	0.88
	Concrete	1	0.99	0.98	0.98	0.99
Front Wheel	Dirt	0.96	0.07	0.91	0	0.96
	Asphalt	0.98	1	0.66	0.77	0.96
	Concrete	0.94	0.97	0.97	0.97	0.97
Early fusion	Dirt	0.68	0.76	0.68	0	0.8
	Asphalt	0.96	0.84	0.7	0.09	0.86
	Concrete	0.98	0.98	0.99	0.98	0.98
Late fusion	Dirt	0.95	0.61	0.94	0	0.97
	Asphalt	0.98	0.86	0.76	0.51	0.98
	Concrete	1	1	0.98	0.94	0.99

Table 5: Recall scores for different validation sets and different methods/sensors. We consider 4 scenarios: train on all but withhold speed, and evaluate on all the observations for this particular speed, and additionally a random 85% train-15% test split.

Sensor location	Class	Test dataset specification				
		30 mph	35 mph	40 mph	55 mph	15% random split
Vibration Sensor	Dirt	0.92	0.89	0.77	0	0.99
	Asphalt	0.92	0.91	0.5	0.02	0.96
	Concrete	0.99	0.98	0.89	0.73	0.98
Engine Inside	Dirt	0.88	0.59	0.84	0	0.95
	Asphalt	0.81	0.71	0.73	0.73	0.91
	Concrete	0.96	0.99	0.96		0.97
Side of Car	Dirt	0.55	0.4	0.67	0	0.92
	Asphalt	0.81	0.75	0	0	0.91
	Concrete	0.99	0.87	0.99	0.72	0.98
Front Wheel	Dirt	0.96	0.14	0.95	0	0.96
	Asphalt	0.93	0.64	0.76	0.86	0.95
	Concrete	0.97	0.96	0.87	0.91	0.98
Early fusion	Dirt	0.78	0.8	0.72	0	0.87
	Asphalt	0.89	0.81	0.73	0.16	0.86
	Concrete	0.96	0.96	0.94	0.74	0.94
Late fusion	Dirt	0.96	0.76	0.92	0	0.98
	Asphalt	0.97	0.73	0.82	0.64	0.97
	Concrete	1	0.96	0.94	0.81	0.99

Table 6: F1 scores for different validation sets and different methods/sensors. We consider 4 scenarios: train on all but withhold speed, and evaluate on all the observations for this particular speed, and additionally a random 85% train-15% test split.