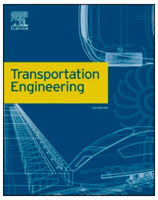
Contents lists available at ScienceDirect

Transportation Engineering

journal homepage: www.sciencedirect.com/journal/transportation-engineering



Full Length Article

A multimodal deep learning approach for gravel road condition evaluation through image and audio integration

Nausheen Saeed [*], Moudud Alam, Roger G Nyberg

School of Information and Engineering, Dalarna University, Röda vägen 3, Borlänge, Sweden

ARTICLE INFO

Keywords:
Gravel road maintenance
Data fusion
Sound analysis
Machine vision
Machine
Learning

ABSTRACT

This study investigates the combination of audio and image data to classify road conditions, particularly focusing on loose gravel scenarios. The dataset underwent binary categorisation, comprising audio segments capturing gravel sounds and corresponding images. Early feature fusion, utilising a pre-trained Very Deep Convolutional Networks 19 (VGG19) and Principal component analysis (PCA), improved the accuracy of the Random Forest classifier, surpassing other models in accuracy, precision, recall, and F1-score. Late fusion, involving decision-level processing with logical disjunction and conjunction gates (AND and OR) in combination with individual classifiers for images and audio based on Densely Connected Convolutional Networks 121 (DenseNet121), demonstrated notable performance, especially with the OR gate, achieving 97 % accuracy. The late fusion method enhances adaptability by compensating for limitations in one modality with information from the other. Adapting maintenance based on identified road conditions minimises unnecessary environmental impact. This method can help to identify loose gravel on gravel roads, substantially improving road safety and implementing a precise maintenance strategy through a data-driven approach.

1. Introduction

Loose gravel on gravel roads significantly challenges road safety and maintenance efforts. Loose gravel can lead to reduced traction, vehicle skidding, and increased dust emissions, potentially causing hazardous conditions for drivers and pedestrians alike. Accurate and timely detection of loose gravel is paramount for traffic agencies to initiate maintenance measures promptly and ensure the safety of road users.

Traditional methods of loose gravel detection on gravel roads have relied on manual inspections, often limited in scope and subject to human error. In recent years, machine learning and multimodal sensor fusion advancements have provided opportunities to revolutionise gravel road condition assessment, offering a more data-driven and precise approach to detecting loose gravel.

In [1,2], the possibility of objectively classifying the loose gravel conditions using audio and images independently was investigated, and the results were promising. This paper introduces an approach to detecting loose gravel on gravel roads, utilising the fusion of spectrograms from audio recordings and images captured from the road surface. This multimodal fusion aims to significantly enhance the accuracy and reliability of loose gravel detection, aligning closely with the standards set forth by road traffic agencies worldwide.

The proposed methodology harnesses the synergistic nature of audio and image data, recognising that each modality brings unique insights into loose gravel detection. Audio spectrograms capture acoustic signatures, such as gravel impacts, surface disturbances, and vehicle-induced vibrations, offering valuable acoustic signals indicative of loose gravel presence. Meanwhile, images provide high-resolution visual information about the road surface, enabling the detection of loose gravel patches, displacement, and surface irregularities.

This paper examines two fusion methods specifically designed for detecting loose gravel: feature-level fusion and decision-level fusion. Feature-level fusion involves combining features from two different sources-in this case, images and audio from gravel roads. On the other hand, decision-level fusion occurs at a later stage, combining decisions from models trained separately on images and audio. Fusion techniques offer a notable advantage in classification by enhancing accuracy and robustness. This advantage stems from their ability to effectively utilise complementary information from different sources or modalities, addressing the limitations of individual methods. Integrating data modalities or decision outputs improves accuracy, reliability, and adaptability, particularly in handling complex classification tasks [3,4]

[*] Corresponding author.
E-mail address: nse@du.se (N. Saeed).

https://doi.org/10.1016/j.treng.2024.100228
Received 6 November 2023; Received in revised form 30 January 2024; Accepted 31 January 2024
Available online 2 February 2024

The suggested framework for loose gravel detection aims to provide an objective method aligning with the Swedish Road Transportation Agency (Trafikverket) standards. Automating the loose gravel assessment would improve safety conditions on gravel roads and decrease maintenance response times.

2. Literature review

There are many studies utilising the multimodal approaches for classification tasks. Multimodal methodologies benefit classification tasks [3–5]. By seamlessly integrating information from diverse modalities, these approaches exhibit enhanced performance compared to their unimodal counterparts. The fusion of different modalities provides a robust and redundant framework, ensuring the system's resilience, even in noisy or incomplete data. Multimodal models excel in handling ambiguity and demonstrate improved generalisation, making them adaptable across various scenarios and datasets. Their versatility extends across domains, such as computer vision, natural language processing, and healthcare ([6–8]).

Moreover, these methodologies mirror the human-like perception that integrates multiple sensory inputs, aligning with the holistic nature of human cognition. Multimodal models prove valuable in scenarios where data may be incomplete or missing in one modality, and they facilitate transfer learning, enabling the transfer of knowledge between modalities or tasks. In essence, the advantages of multimodal methodologies lie in their ability to harness the strengths of different modalities synergistically, resulting in more robust, versatile, and practical solutions for classification tasks.

Considering the maintenance of gravel roads in Sweden by the Trafikverket, the current assessment methods involve subjective evaluations based on guidelines, incorporating factors such as crossfall, irregularities, loose gravel, and dust [9]. These are rated subjectively and, in some cases, involve manual measurements using specialised equipment. However, due to the high costs associated with alternative objective methods, such as laser scanners, they are typically not employed, prioritising the minimisation of gravel road maintenance expenses [10]. Fig. 1 illustrates loose gravel conditions and their grades, depicting Road Type 1 as well-maintained and Road Type 4 as severely deteriorated, following the [9] grading system. This visual reference provides a clear insight into the spectrum of gravel road conditions under assessment.

A dust classification algorithm was developed by [11] for the Gravel Roads Management System, using smartphone images to classify dust amounts on gravel roads accurately. The algorithm was validated against dustometer measurements. The results showed that the algorithm is a cost-effective and accurate alternative, offering potential assistance to local agencies in maintenance planning regarding dust evaluation on gravel roads. The study explores challenges with gravel pavement, noting its lower construction costs but inferior performance to asphalt. Dust emission, deformation, and deepening ripples impact vehicle vibrations, fuel consumption, and driving comfort. The research proposes a methodology for gravel pavement evaluation, measuring profiles and analysing the international roughness index (IRI). Findings stress the importance of timely maintenance. The study's objectives include adapting road roughness indicators for gravel pavement and evaluating dynamic responses, with specific speed ranges (30–45 km/h and 90 km/h) indicating the need for careful prediction of safe driving speeds.

In a recent study by [12], a semi-automated approach utilising UAV-captured images from a one-kilometre road segment was introduced to identify and extract parameters of unpaved road surfaces, such as potholes and rutting. This method addresses the crucial necessity for efficient road condition surveys. The research was conducted in the Ofirikrom Municipality, Ghana, showcasing the correlation between UAV imagery and conventional field methods, suggesting the potential for cost-effective road maintenance monitoring. Although the study was confined to a limited road length, it suggests future endeavours for a fully automated methodology to enhance road condition assessment further.

The literature review highlights a predominant emphasis on overall road roughness in existing research, revealing a relatively lesser focus on identifying distinct distress types on gravel roads. Notably, there is a research gap in the automation of loose aggregate assessment ([13–16]). There is potential in exploring avenues that involve integrating data from various sources, including sound and images, offering promise for the automated assessment of loose gravel on these roads.

3. Methodology

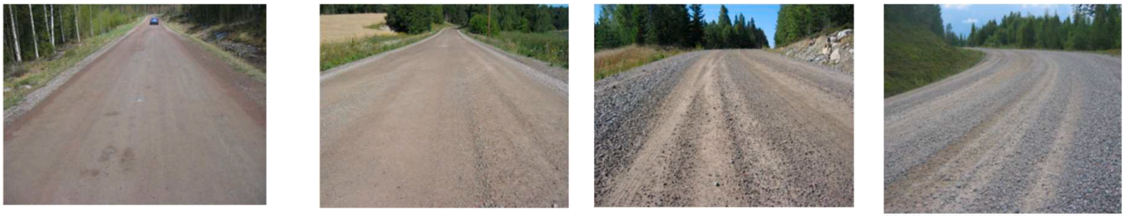
This section will discuss methodology, including an overview of the study's methodology, starting with an overview of multimodal fusion, the data collection approach, followed by a discussion of the preprocessing steps. Subsequently, the two distinct techniques utilised in this study, feature-level early fusion, and decision-level fusion, are introduced.

3.1. Overview of multimodal fusion techniques

This study incorporates multimodal fusion techniques. The subsequent section offers a brief technical introduction to general fusion methods.

Multimodal fusion techniques are methodologies used to combine information or data from multiple sources or modalities to enhance the understanding or performance of a system or application. These techniques are valuable for improving models used in affect recognition tasks, which are analysed based on data from various sources such as audio, visual, physiology, and more. Multimodal fusion holds significant merit in the realm of classification tasks. Fusing information from multiple modalities can enrich the feature space, enhance the discriminative power of models, and provide a more comprehensive understanding of complex phenomena. The literature discusses three joint fusion strategies: feature-level, decision-level (or score-level), and model-level fusion [17]. These are discussed below:

- Feature-level fusion: This strategy combines features extracted from different modalities by creating a single feature vector encompassing



Road Type 1. Road Type 2. Road Type 3. Road Type 4.

Fig. 1. showcases images of loose gravel conditions with their respective grades. Road Type 1 illustrates a well-maintained road, while Road Type 4 depicts a severely deteriorated gravel road [9].

information from all modalities. This approach mimics how humans process information, where features from various sources, such as audio and visual cues, are integrated before making predictions. Feature-level fusion often requires large training datasets because it captures more information than a single modality alone. Additionally, the modalities should have corresponding data for this strategy to work effectively. One major advantage is that predictions can still be made even if data from one modality are missing [18].

- Decision-level (Score-level) fusion*:* In this strategy, each modality is used independently to make predictions, and then the scores or results from each modality are combined. A drawback of this approach is that if data from one modality are missing, the full potential of that modality cannot be realised. Fusion can be as simple as a majority vote for classification tasks, but more sophisticated versions may be introduced, e.g., incorporating learning weights. For regression tasks, a linear regressor can be trained using the predictions from each modality, and its weights can be used for the fusion.
- Model-level (Hybrid-level) fusion: This strategy combines the strengths of both feature-level and decision-level fusion strategies. For instance, a model-level fusion might involve performing feature-level fusion for certain modalities and then combining those predictions with scores from other modalities that were processed independently. This approach offers flexibility and can adapt to the specific requirements of the task [19]. An example of model-level fusion is the method proposed by [20], which combines the results of feature-level fusion with scores from independently processed modalities. This hybrid approach aims to harness the benefits of integrating features from some modalities, while still considering the unique information provided by others. This fusion technique can improve performance in affect recognition tasks, especially when dealing with complex and diverse data from multiple sources [21].

3.2. Data collection

The data collection involved using two HERO7 GoPro cameras manufactured by GoPro Inc. based in San Mateo, CA, USA. One camera was positioned inside a vehicle to capture audio and video data, while a second camera was mounted on the car's bonnet to obtain recordings with an improved view of the roads. The recordings were made during the summer seasons of 2020 and 2022 along gravel roads in Dalarna, Sweden. The car maintained a constant speed of 50 km/h during these recordings under dry and sunny weather conditions. It is important to note that certain portions of the recorded videos were excluded from the dataset. These excluded segments contained activities such as travelling to the selected road, turning the car around, driving at varying speeds, and conversations between the data collectors. These marked segments did not represent the gravel road conditions the study aimed to analyse.

The dataset consisted of a total of 15 videos, with a combined duration of 1 h, 13 min, and 54 s (01:13:54). The purpose of this data collection was to investigate the gravel road conditions, utilising the audio and video recordings obtained from the GoPro camera. For more detailed information about the camera and vehicle specifications, refer to the publication by Saeed et al. [2]

3.3. Preprocessing

Audio and image data were extracted from recorded videos, resulting in separate datasets for both modalities. Preprocessing procedures were subsequently applied to each dataset. Roboflow's Annotation Tool was instrumental in highlighting the gravel roads by creating bounding boxes and isolating the road sections. This approach was applied so that images with only the crucial aspects of the gravel roads are obtained, while discarding unrelated elements, such as the sky and vegetation. As a result, a new dataset containing solely the road information was generated with the assistance of Roboflow.

Roboflow is a specialised platform tailored for developers and researchers dealing with visual data. It offers a comprehensive set of tools and services for tasks such as image annotation, dataset management, data preparation, and even model deployment in computer vision and image processing [22]. In Fig. 2(a), Roboflow illustrates its capability to detect roads and segment the gravel road, excluding vegetation, as shown in Fig. 2(b). This process was undertaken to ensure that during subsequent classification, the algorithms focus on learning features extracted specifically from road conditions.

A conversion process was employed for the audio data to transform the audio files into spectrograms, shown in Fig. 3. The audio data went through a conversion process, during which the audio signals were broken down into smaller segments, predominantly employing the Short-Time Fourier Transform (STFT) technique [23]. These temporal segments were subsequently translated into image representations, featuring time on one axis and frequency on the other. This transformation resulted in the creation of spectrogram images, effectively rendering the audio data in a visual format conducive to integration with the existing image-based processing pipeline.

3.4. Dataset

Following the preprocessing of both the image and audio datasets, each dataset was categorised through labelling into two classes: 1 & 2 and 3 & 4, aligning with Trafikverket's classification, where 1 represents good road conditions, and 4 indicates the worst road conditions. The former had a combined count of 487 instances, while Classes 3&4 had a sum of 398 for both images and audio. The size of each data set in total was 885. The class labelling adheres to the guidelines outlined in the Trafikverket Road Maintenance Gravel Road assessment manual [9]. Considering the limited size of the available dataset, we have combined Classes 1 and 2, as well as 3 and 4. These can be considered as roads in good condition and roads in poor condition, respectively. In the case of audio labelling, each audio clip received its label based on its extraction from a corresponding video segment. For example, if the video segment indicated Road Types 1&2, the audio extracted from that section was labelled Classes 1&2. Table 1 presents the details of the dataset.

Within this study, we have implemented both feature-level fusion and decision-level fusion. The following discussion elaborates on the particulars of the processes utilised in this study.

3.5. Feature-level fusion

In this study, features were extracted from road images and audio spectrograms using the VGG19, a pre-trained convolutional neural network architecture. VGG19 is recognised for its effectiveness in image classification and achieves feature extraction by guiding input data through its hierarchical layers. It progressively captures intricate patterns and details [24]. The extracted features from road images and spectrograms are later combined through concatenation, creating a unified representation. This comprehensive and integrated representation is valuable for enhancing subsequent stages of analysis.

After feature extraction and concatenation, feature reduction was applied using Principal Component Analysis (PCA), and the optimal number of components was determined using the elbow method. PCA transforms original features into orthogonal principal components, capturing maximum data variance. Projecting data onto a lower-dimensional subspace, PCA effectively reduces dimensionality while retaining crucial variance [25]. The elbow method identifies the "elbow point", where additional components cease to significantly increase explained variance [26].

After feature extraction, concatenation, and reduction, machine learning algorithms were trained on this feature set, specifically Random Forest, Multi-layer Perceptron (MLP), and XGBoost classifiers. Finally, the classification results were obtained. Fig. 4 illustrates how feature-level fusion works in this study, using gravel road images and audio spectrograms as inputs to produce a classification decision as the output.

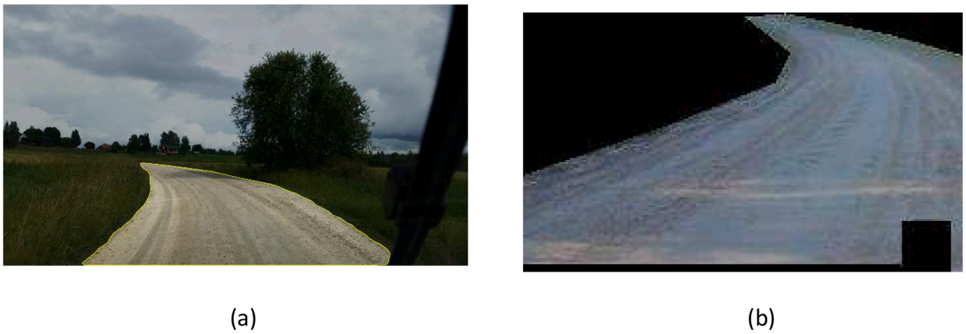


Fig. 2. Image (a) depicts gravel road detection, while image (b) exclusively displays the extracted roads from the image, omitting vegetation and sky.

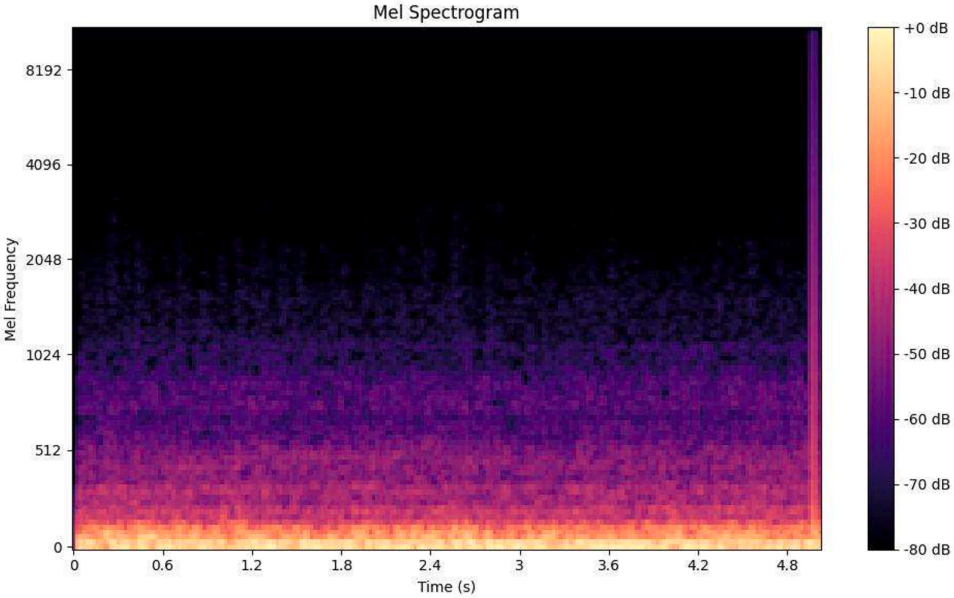


Fig. 3. Example of audio spectrogram obtained after audio preprocessing.

Table 1
Dataset summary: images and audio class distribution.

	1 &2	3&4	Total
IMAGES	487	398	885
AUDIO	487	398	885

It visually guides through the entire process.

These classifiers are widely recognised for their efficacy across diverse domains [19,27–29]. The Random Forest classifier operates as an ensemble learning approach, uniting numerous decision trees to generate precise and resilient predictions. This involves training individual trees on distinct data subsets and amalgamating their outcomes for the final predictions. Conversely, the Multi-layer Perceptron (MLP) is an artificial neural network tailored for intricate pattern recognition tasks. Comprising multiple layers of interconnected nodes, it undertakes data processing and transformation, each contributing to the network's adeptness in capturing complex data relationships. The Gradient Boosting XGBoost algorithm incrementally constructs a sequence of weak learners, often decision trees. Each new learner addresses the errors of its predecessors, fostering potent predictive capabilities [30]. This iterative strategy empowers XGBoost to manage intricate datasets proficiently. Each of these classifiers boasts unique merits, and their selection hinges on the specific attributes of the given problem (X. [31–33]).

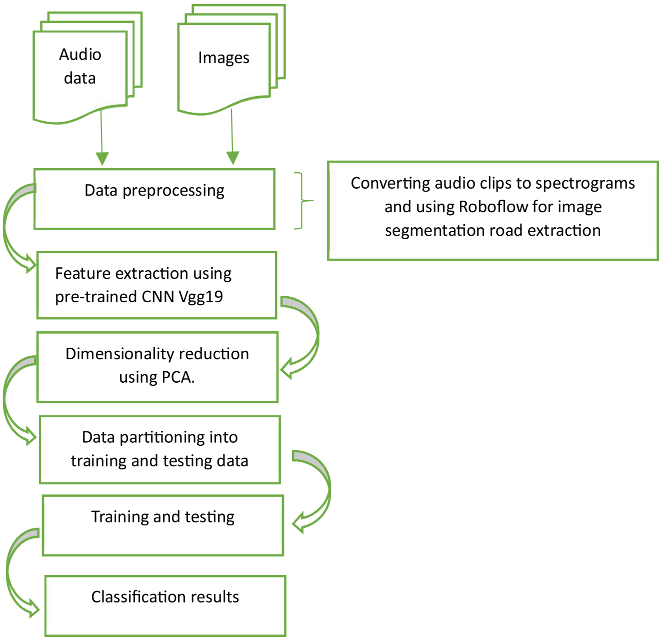

Audio data
Images
Data preprocessing
Converting audio clips to spectrograms and using Roboflow for image segmentation road extraction
Feature extraction using pre-trained CNN Vgg19
Dimensionality reduction using PCA.
Data partitioning into training and testing data
Training and testing
Classification results

Fig. 4. Methodology used in this study for Feature-level Fusion.

3.6. Decision-level fusion

The second fusion method employed in this study is decision-level fusion, and it is discussed below. It incorporates two variations using OR and AND rules. Fig. 5 illustrates the use of both decision-level and feature-level fusion used in this study. Fig. 5 gives a broad view of the decision fusion methods employed: both feature-level fusion and decision-level fusion. It focuses on explaining the key components of these fusion approaches Existing studies consistently demonstrate improved classification performance with decision fusion (K. [34–36]) emphasising the need for diverse techniques due to varied classifier outcomes [37].

3.6.1. Decision fusion with an OR rule

The technique commonly known as majority voting, logical disjunction, or voting with a logical OR, is widely used in various studies. This includes applications such as person recognition using imperfect face images alongside supporting gait images, as well as in spam detection through videos and images. The use of decision fusion in this approach consistently leads to better performance in classification results [4,38]. In majority voting, the final prediction is based on the majority decision of the individual models. If most models predict a positive outcome (Class 1), the fused prediction will be positive. Otherwise, if the majority predicts a negative outcome (Class 0), the fused prediction will be negative. Consider two binary classifiers, C1 and C2, where each classifier makes a binary decision (0 or 1). The final decision $\mathrm{D}_{\mathrm{final}}$ in the OR gate scenario is 1, if at least one of the decisions of classifiers D_{C1} or D_{C2} predicts 1.

$$\mathrm{D}_{\mathrm{final}} = D_{C1}\ V\ D_{C2} \tag{1}$$

Here, ∨ represents the logical OR operation.

3.6.2. Decision fusion with an AND rule

This method is often called unanimous voting, or voting with a logical AND or Logical Conjunction. In unanimous voting, the final prediction is positive only if all individual models predict a positive outcome [39]. If any one model predicts a negative outcome, the fused prediction will be negative. This approach ensures that all models agree before making a positive prediction.

$$\mathrm{D}_{\mathrm{final}} = D_{C1}\ \Lambda\ D_{C2} \tag{2}$$

Here, ⋀ represents the logical AND operation [4].

Both majority voting (OR rule) and unanimous voting (AND rule) are variations of voting-based ensemble methods commonly used to combine predictions from multiple models. The specific logical operations (OR and AND) determine how the predictions are aggregated to arrive at the final decision. These methods harness the collective knowledge of diverse models and enhance overall predictive performance [40].

4. Results and discussion

In this study, data extraction from video recordings encompassed both audio and image components. The audio segment specifically captured the auditory cues of gravel impacting the undersides of vehicles, serving as a significant source of information regarding road conditions. The dataset, categorised into binary Classes 1&2 and 3&4, aims to discern road conditions, especially in loose gravel scenarios, aligning with Trafikverket's standards, where Class 1 signifies well-maintained roads and Class 4 indicates poor conditions. Classes 1&2 are combined to denote good road conditions, while Classes 3 and 4 signify areas needing maintenance.

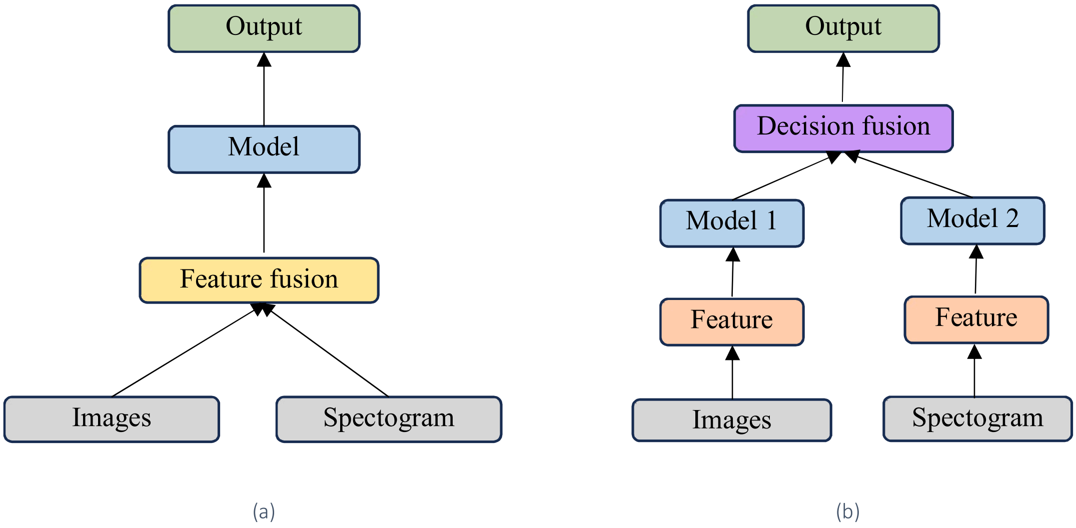
Audio and image data fusion were integrated to investigate the potential enhancement of the classifier's accuracy. Initially, early feature fusion was employed, involving the extraction of features from both audio spectrograms and images using the pre-trained convolutional neural network VGG19. These distinct features from both modalities were concatenated to create a unified feature space. Subsequently, PCA (Principal Component Analysis) was applied for dimensionality reduction.

The Random Forest classifier, Multi-layer Perceptron, and XGBoost classifier were then trained using 80 % of the dataset and validated on the remaining 20 %. The experimental outcomes, as presented in Table 2, highlight the Random Forest classifier's superior performance across metrics such as accuracy 0.9018, precision 0.9011, recall 0.9018, and F1-score 0.9014 compared to other models.

A late fusion methodology, also known as decision-level fusion, as discussed previously in the methodology section, was explored to

Table 2
Classification results of various machine learning algorithms using the early feature fusion method.

Algorithms	Accuracy	Precision	Recall	F1-score
Random Forest Classifier	**0.9018**	**0.9011**	**0.9018**	**0.9014**
MLP	0.8679	0.8841	0.8736	0.8654
XGBoost Classifier	**0.9075**	0.8641	0.8736	0.8654



(a)

(b)

Fig. 5. illustrates the application of two fusion techniques in this study: (a) feature-based early fusion and (b) decision-level fusion.

improve the results further. Two decision-level gates, namely AND and OR gates, were tested. Individual classifiers based on DenseNet121 were trained separately on each modality, i.e., images and audio. Subsequently, these classifiers were tested on the designated test dataset, resulting in individual accuracies of 0.95 for images and 0.92 for audio, respectively, as seen in Table 3. The fusion of their decisions was achieved through AND and OR gates. Notably, the OR gate demonstrated superior performance with an accuracy of 0.97. This accuracy was derived by comparing the test results with the ground truth labels.

Table 3
Classification results by decision-level fusion method.

Pretrained CNN		Accuracy	Decision-level Fusion AND	Decision-level Fusion OR
DenseNet121	Images	0. 95	0.90	**0.97**
	Audio	0.92		

The superiority of late fusion results is evident, and late fusion methods also demonstrate increased adaptability to diverse input conditions. These methods excel in scenarios where one modality may be afflicted by noise or incompleteness, as the other modality can effectively compensate for these limitations. Each modality typically possesses unique strengths and weaknesses; for instance, images might excel in capturing visual details, while audio can contribute additional contextual information. Late fusion, as an approach, enables the fusion of data from both modalities (images and audio), thereby augmenting the overall system's robustness through the utilisation of complementary information from distinct sources.

5. Conclusion

This study introduces a novel methodology that employs both audio and image data to detect loose gravel conditions on gravel roads. Audio clips from Road Classes 1&2 and 3&4, capturing varying degrees of gravel hitting the bottom of the car, were labelled into these two groups. The labelling process entailed a thorough examination of videos, extracting relevant segments, and categorising roads based on predefined classes. These classes adhere to the labelling system of Trafikverket, ranging from Class 1 to 4, where Class 1 represents a good road condition, and Class 4 indicates the worst condition. However, due to limitations in data volume, we combined Classes 1 and 2 into one category and Classes 3 and 4 into another. Subsequently, the audio segments were transformed into spectrograms. Using Roboflow annotation tool, roads from images were isolated to ensure that the classifier learned features relevant to road conditions, while disregarding irrelevant elements such as vegetation and sky. The fusion technique involved combining decisions from two classifiers trained on gravel images and corresponding audio segments to enhance road classification. Both feature-level early fusion and decision-level late fusion techniques were evaluated, incorporating OR and AND gates.

The decision-level approach using the OR gate exhibited superior accuracy in the classification process. The collected data from Swedish roads can be utilised to assess gravel road conditions in Sweden and similar terrains. Applications developed through this method can be deployed on cost-effective devices, such as smartphones for capturing data from gravel roads, and the classification results can be mapped on real maps, displaying the road profile. This could assist drivers in planning their trips and gaining knowledge of road conditions in advance. These applications empower road assessment agencies to conduct timely and unbiased evaluations of gravel conditions, particularly concerning loose gravel. With additional data, the study's scope could be broadened to classify gravel roads into four classes. The methodology is adaptable to other gravel road defects, contributing to a comprehensive system that provides insights into road status, including defects such as potholes, dust, and corrugations. The study advocates for data-driven decision-making by road maintenance agencies, streamlining prioritisation and resource allocation based on identified road conditions. The utilisation of audio and image data, particularly from smartphones, allows for remote monitoring, reducing the need for physical inspections and enhancing efficiency. Adapting maintenance strategies to the identified road conditions has the potential to minimise the unnecessary environmental impact associated with extensive road repair activities. The study's reliance on easily accessible devices, such as smartphones, creates opportunities for community engagement in data collection, involving residents as contributors to road maintenance efforts. The study's findings and methodology could serve as a catalyst for further research in the integration of audio and image data for assessing infrastructure, fostering continuous advancements in road maintenance technology.

Disclosure statement

No potential conflict of interest was reported by the author(s).

CRediT authorship contribution statement

Nausheen Saeed: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing – original draft, Writing – review & editing. **Moudud Alam:** Conceptualization, Project administration, Validation, Writing – review & editing. **Roger G Nyberg:** Conceptualization, Validation, Writing – review & editing.

Declaration of competing interest

The authors (Nausheen Saeed, Moudud Alam and Roger G Nyberg) declare that they have no conflict of interest.

Data availability

Data will be made available on request.

References

[1] N. Saeed, R.G. Nyberg, M. Alam, Gravel road classification based on loose gravel using transfer learning, Int. J. Pavement Eng. (2022) 1–8, https://doi.org/10.1080/10298436.2022.2138879.

[2] N. Saeed, R.G. Nyberg, M. Alam, M. Dougherty, D. Jooma, P. Rebreyend, Classification of the acoustics of loose gravel, Sensors 21 (14) (2021), https://doi.org/10.3390/s21144944.

[3] N.H. Alsaedi, E.S. Jaha, Dynamic audio-visual biometric fusion for person recognition, Comput. Mater. Contin. 71 (1) (2022) 1283–1311, https://doi.org/10.32604/cmc.2022.021608.

[4] K.P. Das, J. Chandra, Multimodal classification on PET/CT image fusion for lung cancer: a comprehensive survey, ECS Trans. 107 (1) (2022) 3649.

[5] X. Zhao, M. Zhang, R. Tao, W. Li, W. Liao, L. Tian, W. Philips, Fractional fourier image transformer for multimodal remote sensing data classification, IEEE Trans. Neural Netw. Learn. Syst. (2022).

[6] G. Cai, Y. Zhu, Y. Wu, X. Jiang, J. Ye, D. Yang, A multimodal transformer to fuse images and metadata for skin disease classification, Vis. Comput. 39 (7) (2023) 2781–2793.

[7] Y. Du, Y. Liu, Z. Peng, X. Jin, Gated attention fusion network for multimodal sentiment classification, Knowl. Based Syst. 240 (2022) 108107.

[8] Y. Wang, Y. Feng, L. Zhang, J.T. Zhou, Y. Liu, R.S.M. Goh, L. Zhen, Adversarial multimodal fusion with attention mechanism for skin lesion classification using clinical and dermoscopic images, Med. Image Anal. 81 (2022) 102535.

[9] Trafikverket. (2014). Bedömning av grusväglag (Assesment of gravel roads),TDOK 2014:0135 Version 1.0,Trafikverket. https://trafikverket.ineko.se/Files/sv-SE/10845/RelatedFiles/2005_060_bedomning_av_grusvaglag.pdf.

[10] A. Alhasan, D.J. White, K. De Brabanter, Quantifying roughness of unpaved roads by terrestrial laser scanning, Transp. Res. Rec. 2523 (1) (2015), https://doi.org/10.3141/2523-12.

[11] O. Abu Daoud, O. Albatayneh, L. Forslof, K Ksaibati, Validating the practicality of utilising an image classifier developed using TensorFlow framework in collecting corrugation data from gravel roads, Int. J. Pavement Eng. (2021), https://doi.org/10.1080/10298436.2021.1921773. May.

[12] W. Oppong Adu, G. Dumedah, A.C. Adams, Surface condition assessment of unpaved roads through the use of unmanned aerial vehicle, Int. J. Pavement Res. Technol. (2023), https://doi.org/10.1007/s42947-023-00374-z.

[13] K. Gopalakrishnan, S.K. Khaitan, A. Choudhary, A. Agrawal, Deep Convolutional Neural Networks with transfer learning for computer vision-based data-driven pavement distress detection, Constr. Build. Mater. 157 (2017) 322–330, https://doi.org/10.1016/j.conbuildmat.2017.09.110.

[14] A. Mednis, G. Strazdins, R. Zviedris, G. Kanonirs, L. Selavo, Real time pothole detection using android smartphones with accelerometers, in: Proceedings of the International Conference on Distributed Computing in Sensor Systems and Workshops (DCOSS), 2011, pp. 1–6, https://doi.org/10.1109/DCOSS.2011.5982206. June 2014.

[15] M.I. Rajab, M.H. Alawi, M.A. Saif, Application of image processing to measure road distresses, WSEAS Trans. Inf. Sci. Appl. 5 (1) (2008) 1–7.

[16] H.W. Wang, C.H. Chen, D.Y. Cheng, C.H. Lin, C.C. Lo, A real-time pothole detection approach for intelligent transportation system, Math. Probl. Eng. 2015 (2015) 869627, https://doi.org/10.1155/2015/869627.

[17] S.S. Kanhere, Participatory sensing: crowdsourcing data from mobile smartphones in urban spaces, in: Proceedings of the IEEE 12th International Conference on Mobile Data Management, 06-09 June, Lulea, Sweden, 2011, pp. 3–6, https://doi.org/10.1109/MDM.2011.16.

[18] A. Rattani, D.R. Kisku, M. Bicego, M. Tistarelli, Feature level fusion of face and fingerprint biometrics, in: Proceedings of the IEEE Conference on Biometrics: Theory, Applications and Systems, BTAS'07, 2007, https://doi.org/10.1109/BTAS.2007.4401919.

[19] E. Pei, D. Jiang, H. Sahli, An efficient model-level fusion approach for continuous affect recognition from audiovisual signals, Neurocomputing 376 (2020) 42–53, https://doi.org/10.1016/j.neucom.2019.09.037.

[20] L. Chao, J. Tao, M. Yang, Y. Li, Z. Wen, Long short term memory recurrent neural network based multimodal dimensional emotion recognition, in: Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge, 2015, pp. 65–72.

[21] S.K. Roy, A. Deria, D. Hong, B. Rasti, A. Plaza, J. Chanussot, Multimodal fusion transformer for remote sensing image classification, IEEE Trans. Geosci. Remote Sens. (2023).

[22] Q. Lin, G. Ye, J. Wang, H. Liu, RoboFlow: a data-centric workflow management system for developing AI-enhanced robots, A. Faust, D. Hsu, & G. Neumann (Eds.), in: Proceedings of the 5th Conference on Robot Learning, PMLR, 2022, pp. 1789–1794, 164, https://proceedings.mlr.press/v164/lin22c.html.

[23] C. Mateo, J. Antonio Talavera, Short-time Fourier transform with the window size fixed in the frequency domain, Digit Signal Process. 77 (2018) 13–21.

[24] T. Carvalho, E.R.S. De Rezende, M.T.P. Alves, F.K.C. Balieiro, R.B. Sovat, Exposing computer generated images by eye's region classification via transfer learning of VGG19 CNN, in: Proceedings of the 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), 2017, pp. 866–870.

[25] K. Annbuselvi, N. Santhi, S Sivakumar, A competent multimodal recognition using imperfect region based face and gait cues using Median-LBPF and Median-LBPG based PCA followed by LDA, Mater. Today Proc. 62 (2022) 4869–4879, https://doi.org/10.1016/j.matpr.2022.03.505.

[26] A.K. Gárate-Escamila, A. Hajjam El Hassani, E. Andrès, Classification models for heart disease prediction using feature selection and PCA, Inform. Med. Unlocked 19 (2020), https://doi.org/10.1016/j.imu.2020.100330.

[27] M. Ahmadlou, A. Al-Fugara, A.R. Al-Shabeeb, A. Arora, R. Al-Adamat, Q.B. Pham, N. Al-Ansari, N.T.T. Linh, H. Sajedi, Flood susceptibility mapping and assessment using a novel deep learning model combining multilayer perceptron and autoencoder neural networks, J. Flood Risk Manag. 14 (1) (2021), https://doi.org/10.1111/jfr3.12683.

[28] K. Budholiya, S.K. Shrivastava, V. Sharma, An optimized XGBoost based diagnostic system for effective prediction of heart disease, J. King Saud Univ. Comput. Inf. Sci. 34 (7) (2022), https://doi.org/10.1016/j.jksuci.2020.10.013.

[29] A.M. Walker, A. Cliff, J. Romero, M.B. Shah, P. Jones, J.G. Felipe Machado Gazolla, D.A. Jacobson, D. Kainer, Evaluating the performance of random forest and iterative random forest based methods when applied to gene expression data, Comput. Struct. Biotechnol. J. 20 (2022), https://doi.org/10.1016/j.csbj.2022.06.037.

[30] H. Jiang, Z. He, G. Ye, H. Zhang, Network intrusion detection based on PSO-Xgboost model, IEEE Access. 8 (2020), https://doi.org/10.1109/ACCESS.2020.2982418.

[31] X. Li, L. Ma, P. Chen, H. Xu, Q. Xing, J. Yan, S. Lu, H. Fan, L. Yang, Y. Cheng, Probabilistic solar irradiance forecasting based on XGBoost, Energy Rep. 8 (2022), https://doi.org/10.1016/j.egyr.2022.02.251.

[32] G.R. Sumsion, M.S. Bradshaw, K.T. Hill, L.D.G. Pinto, S.R. Piccolo, Remote sensing tree classification with a multilayer perceptron, PeerJ 2019 (2) (2019), https://doi.org/10.7717/peerj.6101.

[33] O.K. Toffa, M. Mignotte, Environmental sound classification using local binary pattern and audio features collaboration, IEEE Trans. Multimed. 23 (2021), https://doi.org/10.1109/TMM.2020.3035275.

[34] K. Li, W. Pan, Y. Li, Q. Jiang, G. Liu, A method to detect sleep apnea based on deep neural network and hidden Markov model using single-lead ECG signal, Neurocomputing 294 (2018), https://doi.org/10.1016/j.neucom.2018.03.011.

[35] H.D. Nguyen, B.A. Wilkins, Q. Cheng, B.A. Benjamin, An online sleep apnea detection method based on recurrence quantification analysis, IEEE J. Biomed. Health Inform. 18 (4) (2014), https://doi.org/10.1109/JBHI.2013.2292928.

[36] S.S.M. Noor, K. Michael, S. Marshall, J. Ren, Hyperspectral image enhancement and mixture deep-learning classification of corneal epithelium injuries, Sensors 17 (11) (2017), https://doi.org/10.3390/s17112644 (Switzerland).

[37] S.A. Singh, S. Majumder, Chapter one – Short and noisy electrocardiogram classification based on deep learning, in: Himansu Das, Chittaranjan Pradhan, Nilanjan Dey, Deep Learning for Data Analytics, Academic Press, 2020, pp. 1–19, https://doi.org/10.1016/B978-0-12-819764-6.00002-8.

[38] M. Kihal, L. Hamza, Robust multimedia spam filtering based on visual, textual, and audio deep features and random forest, Multimed. Tools Appl. (2023), https://doi.org/10.1007/s11042-023-15170-x.

[39] A. Gumaei, W.N. Ismail, M. Rafiul Hassan, M.M. Hassan, E. Mohamed, A. Alelaiwi, G. Fortino, A decision-level fusion method for COVID-19 patient health Prediction, Big Data Res. 27 (2022), https://doi.org/10.1016/j.bdr.2021.100287.

[40] Y. Chandola, J. Virmani, H.S. Bhadauria, P. Kumar, End-to-end pre-trained CNN-based computer-aided classification system design for chest radiographs, Deep Learn. Chest Radiogr. (2021) 117–140, https://doi.org/10.1016/b978-0-323-90184-0.00011-4.