Crowdsensing road quality using inertial measurement units on bicycles

SYBE DE OUDE, University of Twente, The Netherlands

Abstract. This work examines methods to enhance crowdsensing of road quality using Inertial Measurement Units (IMUs) on bicycles. The focus of this research has been on three main areas. Firstly, a centralized machine learning model was developed to classify road types based on different pavements. Secondly, a classical machine learning model was developed to classify participants and identify unique features of participants. Lastly, a federated learning model has been created to explore its potential to classify road quality while mitigating privacy issues.

Through this study, significant personalized factors were found that impact the accuracy and generalizability of road quality classification models. Whereas prior studies have often overlooked these personalized biases, this research highlights their importance in developing robust and universally applicable models. Although, the federated learning approach did not fully mitigate these biases, it offers promising direction for future research to achieve more universally applicable road quality insights.

Additional Key Words and Phrases: machine learning, time-series classification, federated learning, road quality, bicycle, IMU, classification algorithms, non-IID data, vibration

1 INTRODUCTION

In 2022, the Dutch government announced that it would invest 780 million euros in cycling infrastructure [7]. Combined with existing municipal and provincial commitments, this brings the total investment to 1.1 billion euros by the year 2030.

This financial investment underscores the growing demand for safe and pleasant cycling experiences. The funds will be allocated towards various initiatives, such as bicycle parking facilities, new bicycle roads, tunnels, bridges, and road maintenance. The effectiveness of cycling infrastructure is dependent on both quantity and quality. Factors such as pavement smoothness, potholes and irregularities impact the overall riding experience and safety. However, determining which roads (in general, not only bicycle roads) require maintenance can be challenging. Traditional pavement distress inspection methods are carried out manually, which may be subjective and require extensive effort [5].

Inertial Measurement Units (IMUs) attached to bicycles can be used to assess road quality through the vibrations and motions that they capture. Prior studies have examined the effectiveness of such methods with promising results. Despite previous studies that attached IMUs to bicycles to understand road conditions, the majority have developed solutions from a single probe bicycle. Several personalised factors such as tire pressure, cyclist's weight, bicycle type, suspension, and riding style affect the vibrations of bicycles. These personalized biases impact the accuracy and generalizability of road classification models.

TScIT 41, July 5, 2024, Enschede, The Netherlands

This research highlights the importance of recognizing personalized biases and investigates methods to classify road quality and identify unique participant features using machine learning techniques.

This study involves the development of a centralized machine learning model, a model for participant classification and feature extraction, and an exploration of federated learning. A combination of these techniques could facilitate a reliable mechanism for measuring road quality using IMUs on bicycles. To guide this investigation, the following research question has been formed:

- How can machine learning models be developed to extract generic road quality insights from data collected using IMU sensors on a bicycle?

This research question can be answered with the following subquestions:

(1) How can a centralized machine learning model be developed to classify different types of road quality from crowdsensed data collected using IMU sensors on a bicycle?
(2) How can personalized features that influence road quality assessment be identified and captured from IMU data of multiple cyclists?
(3) How can federated learning be utilized to address privacy concerns in the development of a bicycle road quality classification model?

In addressing these questions, this research aims to contribute to the existing knowledge in the field of road quality assessment using IMU sensors on bicycles. Firstly, a centralized machine learning model was developed using Edge Impulse [1] to classify different road types based on IMU data. This model enhances the reliability of road quality assessments. Secondly, another machine learning model was also developed with Edge Impulse to classify participants and identify features unique to each participant. Thirdly, a federated learning model was developed. The implementation of this model contributes to the existing knowledge by demonstrating how privacy-preserving techniques can be effectively applied to road quality classification using IMU data from bicycles. This paves the way for more secure and scalable crowdsensing applications.

These contributions help towards the development of a generic road quality classification model, which in turn can contribute to more informed decision-making in infrastructure planning and maintenance. The development of a model capable of extracting generic road quality insights from heterogeneous data is also expected to add to the field of crowdsensing. This could lead to a scalable and cost-effective approach to road quality assessment. For instance, one possible use case could be the integration of IMU sensors in public shared bikes to assess road quality. Lastly, this research seeks to offer valuable insights and methodologies that can be applied not only in the context of cycling, but also in other domains of transportation and urban planning. The focus of the research is on the use of IMU sensors on bicycles, but a similar approach could be adapted for use in other modes of transportation, such as cars or trains.

The structure of this proposal is as follows: In section 2 an overview of related works in the field of road quality assessment and the use of IMUs on bicycles is given. Section 3 will detail the methodologies used to answer the research question and sub-questions. Section 4 will discuss the results. Section 5 gives a brief discussion about the potential reasons for the lower accuracy of the new centralized model in Python. Finally, section 6 shows the planning of the research.

2 RELATED WORK

A substantial amount of research has already been done on road quality and road surface detection. In general, there are three popular approaches to monitoring road surfaces: 3D reconstruction, vibration/sensory-based, and computer vision-based [4]. The scope of this research is vibration/sensory-based monitoring. As this is still a broad field of research, a specification for IMUs on bicycles can be made. Both smartphones and dedicated IMUs have been used with bicycles to determine road quality. Prior research however often lacks personalized factors, such as tire pressure, cyclist's weight, bicycle type and suspension which affect the vibrations of the bicycle. Nagaraj et al. [9] and Vittorio et al. [15] assessed road quality using the IMU data processed by smartphones. Peng et al. [12] looked into the use of IMUs on dock-less shared bikes. In their experiment, two bikes were used. A dynamic calibration was needed as the IMUs on the bikes were mounted at different angles. A few years later, more extensive research continued on this work [6]. The work of Tazelaar [13] uses an IMU connected to a Raspberry Pi [2] and focuses on the labelling of road quality using two buttons on the handlebar. In contrast to the other studies, the participants themselves specify which roads are classified as 'good' or 'bad' in this study. López [8] also researched crowdsensing road quality and roughness using IMU sensors on bicycles. Where this research will focus on the personalized factors, their research had a focus on making estimations under various speed conditions. Heidt and Dorer [3] also predicted road quality using IMUs on bicycles and mentions that a variety of routes, bicycles, riders, speeds and tire pressure could be taken into account for future work. Peirens [11] analysed vibrations made by speed-pedelecs.

In conclusion, multiple studies have assessed road quality using bicycles. In general personalized factors are overlooked. This research will have an emphasis on these personal factors. How this research is conducted can be read in the methodologies below.

3 METHODOLOGIES

This section outlines the methodologies employed in this research to develop and evaluate models for road quality classification and participant identification using IMU data.

3.1 RQ 1 - Centralized road quality classification model

To address the first research question, a centralized machine learning model was developed to classify road quality based on data collected from IMU sensors on a bicycle.

3.1.1 The dataset. Prior and ongoing research at the University of Twente has involved collecting data from cyclists. This data has been examined for its utility in classifying road quality. The already collected data was deemed sufficient for this research. The obtained dataset consists of crowdsensed data from 17 participants. Only participants who cycled at least once per month in the past half-year were eligible to take part in the research. The participants were asked to cycle a specific route on the e-bike at a safe speed while following traffic regulations.

During the field trials, the bicycle and participants were equipped with various sensors. Four IMU sensors were placed on the helmet, handlebar, pedal and frame, respectively. Additionally, a sensor was used to capture the location data, a wristband captured biometric data and a sensor on the chest provided even more accurate heart-rate data. In some trials, button presses have been used to determine how pleasant the participant experienced their cycling experience. Furthermore, a camera was attached to the bicycle to record the cycled route.

For this research, only the data from the IMU sensors on the handlebar and frame, the camera, and the location data was used.

3.1.2 Data labelling and synchronization. This study requires the crowdsensed data of multiple participants who cycle on a bicycle with sensors attached to it. Firstly, roads of different qualities have been classified. The road quality was classified manually by looking at the camera footage in combination with the IMU data. Three different road qualities have been labelled: 'asphalt', 'good brick', and 'bad brick'. An approximately even amount of data has been labelled for each label. The scope of this research was determined to classify continuous data of the whole road. So, instead of an event-based classification, the overall pavement of a road is determined. This is done in segments of approximately 35 seconds.

In Excel, a 'master file' was created to label the data of the participants. At the start of a field trip, a synchronization movement was made by moving the handlebars of the bicycle left, right, and left again. The ProMove-mini sensors from Inertia Studio were used to capture acceleration, gyroscope, and compass values in x, y, and z directions. The Inertia Studio application visualises these values in graphs. The synchronization movement's timestamp was identified using this visualization.

3.1.3 Data preprocessing. Using the programming language Python [14], various functions were made to preprocess the data into a format that can be utilized and handled by Edge Impulse. The 'master file' is used to connect the filenames with the participants and the relevant segment times with different road qualities. The timestamps in the Inertia Studio application, the IMU data file, and the video files all have different time formats. Functions have been created to convert these timestamps. It was thoroughly checked whether the time conversion was correct. An example can be found in Figure 1, where a timestamp from a video file was used to find the corresponding visualisation in Inertia Studio.

The preprocessing workflow included the following steps:

(1) **Looping Through Participant Trips:** The 'master file' is used to iterate over all participant trips. The relevant data is accessed and processed systematically.

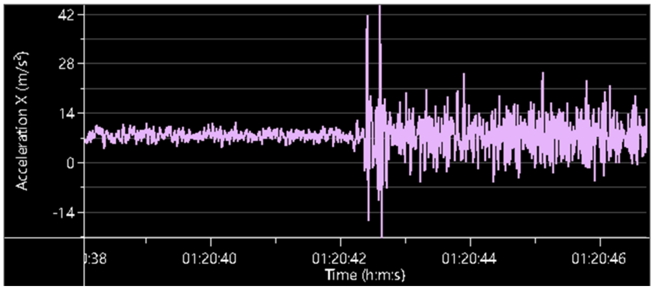


Fig. 1. A clear transition in road quality

(2) **Removing Metadata:** The metadata is removed from the raw files. This is done such that Edge Impulse can process the files.
(3) **Segmenting Road Quality Data:** The relevant columns of the IMU files are extracted. For both the handlebar and frame, these are the acceleration and gyroscope values in x, y, and z direction. Separate csv files are created for the different road quality segments. The synchronization movement at the start of each trip was used as a reference point. By calculating the time since this movement, data was segmented into different road quality categories (asphalt, good brick, bad brick).
(4) **Merging IMU Data:** The IMU data from the handlebar and frame are merged based on the timestamp in milliseconds.
(5) **Adding Velocity Data:** From the location data, the velocity of the participant is extracted. Since the IMU data was sampled at 200 Hz and the location data at 1 Hz, linear interpolation was used to estimate and insert the velocity values into the IMU data.
(6) **Updating the Master File:** The filenames of the newly created csv files were written back to the 'master file'. This ensures that all processed data files were systematically cataloged and easily referenced.

3.1.4 Model development. The preprocessed data was used as input for a centralized machine learning model using Edge Impulse. The model was trained to classify the three different road types (asphalt, good brick, bad brick).

The final model has a 67%/33% train/test split. This split was chosen, as for each participant two trips were allocated to the training set, and one trip is reserved for testing. This approach ensures that the model trains and tests on data from all participants. Thus a representative and unbiased evaluation is maintained.

In table 1 an overview of the most important model parameters is shown. Multiple combinations of parameters have been tested extensively to optimize the road quality classification model. The final model has a window size of 4 seconds with a window increase of 1 second. This window approach led to the best results as it gives the model enough time to accurately classify the road quality while producing enough samples. Zero-padding is applied if a sample in the dataset is shorter than the window size. The model performs spectral analysis on 13 input axes. Namely, the acceleration and gyroscope in x, y, and z direction for both the handlebar and frame, plus

Table 1. Road classification model parameters

Parameter	Value
Window size	4000 ms
Window increase	1000 ms
batch size	32
FFT length	64
Training cycles	200
Learning rate	0.0005

Table 2. Road classification neural network architecture

Model layer	Value
Input	481 features
Dense	20 neurons
Dense	10 neurons
Dropout	0.5 rate

the velocity data. The spectral analysis uses FFT (Fast Fourier Transform) to convert the signal from the time domain to the frequency domain. This reveals the strength of each frequency component. An FFT length of 64 was chosen. Through testing, this appeared as the most optimal balance between the time resolution and frequency resolution. A log base of 10 was applied to the FFT spectrum to compress the dynamic range. This makes it easier for the neural network to learn from the data. The setting to allow overlapping FFT Frames was checked. This works similar to a frame stride.

The training process involves 200 learning cycles with a learning rate of 0.0005. These values offer sufficient learning time while reducing the risk of overfitting.

In Table 2 the architecture of the road classification model is shown. Edge Impulse extracted 481 distinct features for the model input. The model has two dense layers with 20 and 10 neurons, respectively, followed by a dropout layer with a rate of 0.5. This relatively high dropout rate is chosen to prevent overfitting and improves the model's ability to generalize on new, unseen data.

Edge Impulse has a tool called the 'EON Tuner'. This tool analyzes the input data, processing blocks, and neural network architecture and gives an overview of possible architectures. One intended use for this is to create a model that fits the chosen target device's latency and memory requirements. Edge Impulse does not recognize the ProMove-Mini sensor from Inertia Studio as a deployment device. However, the platform provides an estimated processing time of 48 milliseconds and a peak RAM usage of 46 kilobytes. The 'EON Tuner' is also used to fine-tune the parameter values. This greatly helped to develop a model with the promising performance as described at the results in section 4.

3.2 RQ2 - Identifying personalized features

The second sub-question focuses on finding features to differentiate between the participants. Personalized factors can significantly influence the data collected by the sensors. Several personalized factors such as tire pressure, cyclist's weight, riding style, bicycle type, and suspension can alter the captured data. The Pervasive

Table 3. Participant classification model parameters

Parameter	Value
Window size	4000 ms
Window increase	1000 ms
batch size	32
FFT length	64
Training cycles	300
Learning rate	0.0005

Table 4. Participant classification neural network architecture

Model layer	Value
Input	481 features
Dense	40 neurons
Dense	20 neurons
Dense	10 neurons
Dropout	0.5 rate

Systems research group at the University of Twente has multiple studies that use data from the sensor-embedded bicycles. In this research, the road quality was assessed, where the free variable was the different participants that participated in the study. This variability was expected to influence the data captured by the sensors. For instance, through the different riding styles and the total weight of the system. To capture these personalized features, an additional machine learning model was developed in Edge Impulse to classify the participants based on the collected data.

3.2.1 The dataset and preprocessing. The raw dataset used for the participant classifcation model is identical to that of the road quality classification model. The preprocessing steps are largely similar with one key difference. Instead of solely focusing on segments of different road quality, also the intervening parts are used. In the dataset, this is most commonly asphalt pavement. This approach results in approximately 9 minutes of data being utilized per field trip. The model uses the data of 4 participants.

3.2.2 Model development. The participant classification model was also developed using Edge Impulse. The model has a 71%/29% train/test split.

In Table 3 the model parameters can be found. Interestingly, the optimal model selected by running the 'EON Tuner' exhibited near identical parameters for both the road classification and participant classification models. In comparison with the road classification model, the final participant classification model uses more training cycles (300), and it has an extra dense layer of 40 neurons.

In Edge Impulse, an option can be enabled during the model training to calculate the feature importance. This gives an overview of which features are the most relevant to uniquely identify a certain participant. This tool also shows what features are the most influential overall. In section 4 the results of the model and the feature importance will be analysed.

Table 5. Centralized model parameters (Python)

Parameter	Value
Window size	4000 ms
Window increase	2000 ms
batch size	16
Training cycles	200
Learning rate	0.001

3.3 RQ 3 - Federated learning

Federated learning is a decentralized approach to machine learning that aims to train a machine learning algorithm on multiple local datasets without explicitly exchanging raw data samples. Instead, multiple participants collaboratively train a model locally using their own data and only share the model weights and gradients with a central server.

Crowdsensed data involves the data of a lot of participants. As shown in the results section, the participant classification model is able to correctly classify participants to a certain extend. Federated learning is particularly useful to mitigate privacy issues as it allows data to remain on local devices.

In this research, federated learning is used to assess whether it can achieve a performance similar to traditional models. By training the models locally and subsequently aggregating these models, this approach aims to improve the generalizability and robustness of the road quality classification model while addressing privacy concerns associated with crowdsensed data.

3.3.1 Challenges in adapting the road classification model. Extracting the road classification model from Edge Impulse along with its weights proved to be more challenging than anticipated. As a consequence, an agreement was made with the supervisor of this research to utilize his Python code for federated learning. This code had been used for a similar project and has been adapted to fit the dataset used in this research.

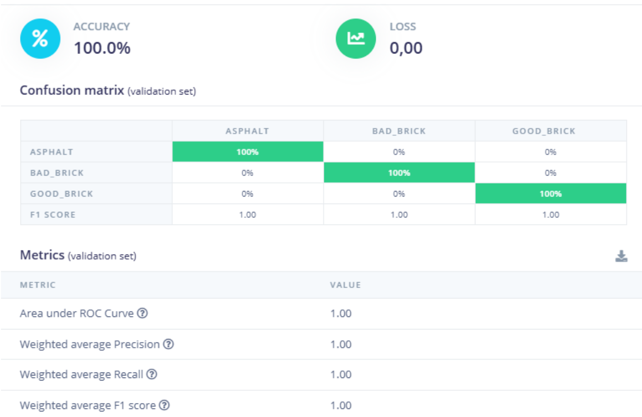
3.3.2 Data preparation. Several steps were needed to change the data before feeding it to the machine learning model. The road quality classification model in Edge Impulse uses the same data files as the federated learning approach. Even so, the structure is slightly different as data files were merged to have all training data of a participant in one file. The data is windowed into windows of 4 seconds with an overlap of 2 seconds. A function was created to generate the label files, corresponding to the data segments. Lastly, the filenames were changed to the format used in the existing code.

3.3.3 The centralized model (Python). Table 5 shows the model parameters used for the centralized model that is used for the federated learning model. These values are slightly different from the centralized model in Edge Impulse. Changing (part of) the parameter values and or model architecture seemed to worsen the accuracy.

3.3.4 The federated learning model. The received Python code was adapted to work with the dataset. Data from two participants was used to train the centralized model, while data from another two participants was used for federated learning. The final model is

Table 6. Centralized model neural network architecture (Python)

Model layer	Value
Input	13 features
Dense	128 neurons
Dense	64 neurons
Dropout	0.5 rate



ACCURACY 100.0%

LOSS 0,00

Confusion matrix (validation set)

	ASPHALT	BAD_BRICK	GOOD_BRICK
ASPHALT	100%	0%	0%
BAD_BRICK	0%	100%	0%
GOOD_BRICK	0%	0%	100%
F1 SCORE	1.00	1.00	1.00

Metrics (validation set)

METRIC	VALUE
Area under ROC Curve	1.00
Weighted average Precision	1.00
Weighted average Recall	1.00
Weighted average F1 score	1.00

Fig. 2. Road quality training perfomance metrics

trained over 3 rounds, with a batch size of 16 and 20 epochs per round.

4 RESULTS

This section presents the results of the three developed models in this research: the road quality classification model, the participant classification model, and the federated learning model. The performances and insights are written down in the subsections.

4.1 Road quality classification model

The centralized road classification model was trained using a 67%/33% train/test split, where two field trips per participant are used for training and one trip for testing. The final model achieved a training accuracy of 100% and a testing accuracy of 97.04%. This indicates that the model performs quite well with robust generalization to unseen data.

4.1.1 Road qualification training results. In Figure 2 an overview of the training performance metrics is shown. A training accuracy of 100% might indicate some overfitting of the model. Nevertheless, this is not necessarily the case here, since the testing accuracy is also quite high.

Figure 3 shows Edge Impulse's data explorer visualisation for the training set. Edge Impulse uses UMAP (a dimensionality reduction algorithm) to project the high dimensionality feature input data into a 2 dimensional space. With the exception of a singular node, all data point all points seem to be neatly separated into different clouds. This indicates that the model can accurately identify differences in features between the differently labeled data.

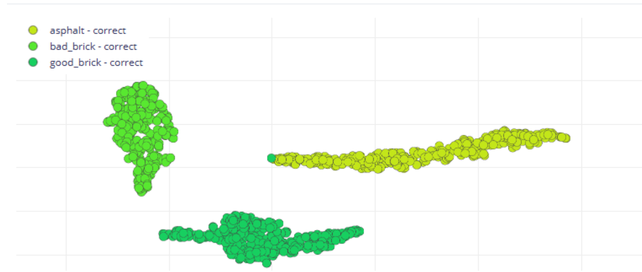


Fig. 3. Road quality training data explorer

ACCURACY 97.04%

Metrics for Classifier

METRIC	VALUE
Area under ROC Curve	0.99
Weighted average Precision	0.97
Weighted average Recall	0.97
Weighted average F1 score	0.97

Confusion matrix

	ASPHALT	BAD_BRICK	GOOD_BRICK	UNCERTAIN
ASPHALT	97.4%	0%	2.1%	0.5%
BAD_BRICK	0%	98.4%	1.6%	0%
GOOD_BRICK	0%	4.4%	94.9%	0.6%
F1 SCORE	0.99	0.97	0.95	

Fig. 4. Road quality test metrics

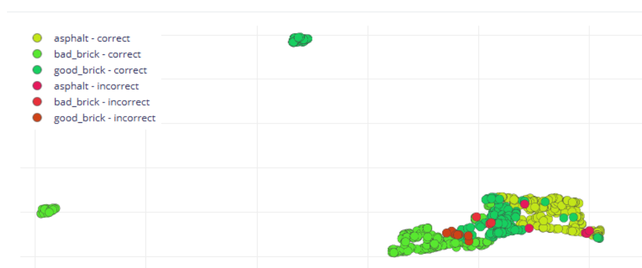


Fig. 5. Road quality test feature explorer

4.1.2 Road qualification test results. Figure 4 shows the metrics of the test set for classifying the road quality. The high precision, recall and F1 scores across all classes indicate the model's ability to accurately distinguish between the different road quality classes.

For the confusion matrix a minimum confidence rating of 0.6 was used. The confusion matrix shows that the model can effectively and correctly identify each road quality class.

Figure 5 shows the visualisation of the features for the test set. There is a separation between the different classes and the wrongly labeled classes are in the area between two classes. This indicates that the model is well-trained and shows the model is not likely to be overfitting.

Fig. 6. Participant classification training performance metrics

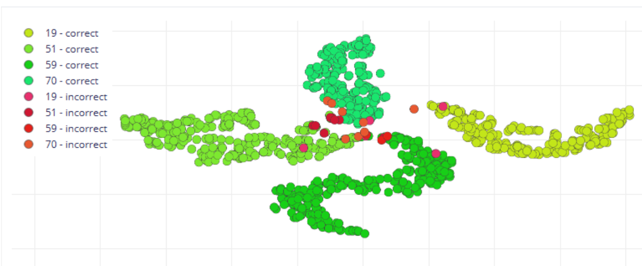


Fig. 7. Participant training data explorer

4.2 Participant classification model

In order to find personalized features a participant classification model was created. The shown model trains and tests on data of 4 participants.

4.2.1 Participant classification training results. Figure 6 shows the performance metrics for the participant classification model. There is a training accuracy of 95.6% and a loss of 0,21.

Figure 7 shows the participant classification data explorer. Just as for the other model, the data points seem to be clearly separated. The wrongly labeled data points are mainly at the edge of a class.

4.2.2 Road qualification test results. Figure 8 shows the test metrics for the participant classification. With 4 participants, the model is able to correctly classify a participant with a 73.92% accuracy. Some participants perform better than others. The accuracy, precision, recall, and F1 scores are somewhat lower than the corresponding training metrics. It's possible that the model is slightly overfitting.

Figure 9 shows the feature explorer of the test set for the participant classification model. In comparison with the other feature visualisations, it is much harder to humanly identify the different classes. This is also reflected in the lower testing accuracy of 73.92% in comparison with the training accuracy of 95.6%.

4.2.3 Personalized features. One of the main goals of the participant classification model is to find personalized features. Table 7 shows a subset of the on average most relevant features found for all data. The most important feature to classify a participant appears to be the Spectral power of the acceleration data of the frame, in the z direction, with a frequency of 4.69 - 7.81 Hertz. This suggests that the way the participant moves the bicycle in this direction within this frequency range contains unique characteristics that help to distinguish between the different participants. The full list of relevant features (per participant) can be found in the appendix.

Fig. 8. Participant classification test metrics

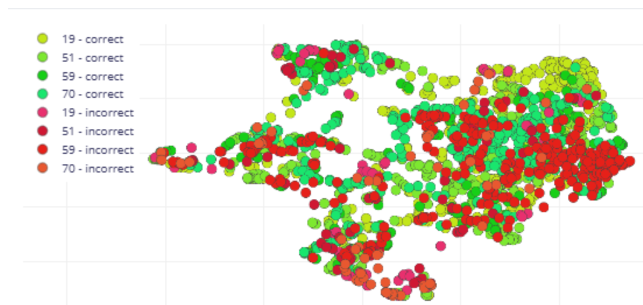


Fig. 9. Participant test feature explorer

Table 7. Feature importance (all data)

Feature	Importance score
az_f Spectral Power 4.69 - 7.81 Hz	3/100
ay_h Spectral Power 4.69 - 7.81 Hz	2/100
ay_h Spectral Power 1.56 - 4.69 Hz	2/100
ax_h Spectral Power 67.19 - 70.31 Hz	2/100
az_h Spectral Power 23.44 - 26.56 Hz	1/100
gz_h Spectral Kurtosis	1/100

Table 8. Centralized model metrics (Python)

Metric	Value
overall accuracy	69% (52334/75200)
Test loss	0.946618
Asphalt test accuracy	82% (20362/24800)
Good brick test accuracy	33% (8292/24800)
Bad brick test accuracy	92% (23680/25600)
Precision	0.77
Recall	0.69
F1-score	0.67

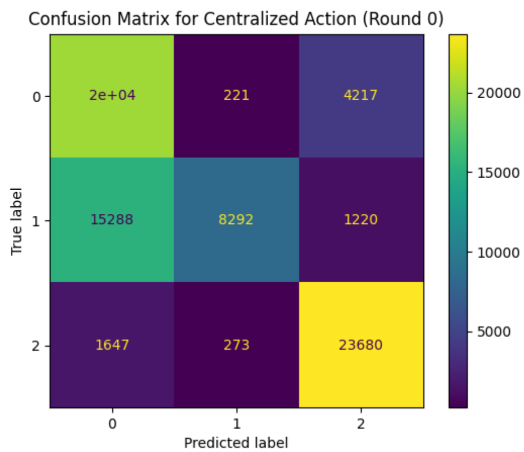


Fig. 10. Centralized confusion matrix (Python)

4.3 Federated learning

In this section the results of applying federated learning to classify road quality are presented. With the code of the supervisor of this research a new centralized model was created. This model serves as a benchmark to compare the performance of the federated learning approach.

4.3.1 centralized model performance (Python). The centralized model achieved an overall accuracy of 69% (52,334 correctly classified instances out of 75,200). The metrics can be found in Table 8 and a confusion matrix is shown in Figure 10. The overall accuracy of 69% is quite lower than the accuracy that was achieved with Edge Impulse. This is likely due to slightly different and less optimized parameters and a slightly different model architecture. Adapting part of the parameters to the ones used in Edge Impulse worsened the accuracy.

4.3.2 federated learning model performance. Different parameters were tested for the federated learning model. The final model was trained over three rounds with a batch size of 16, with 30 epochs per round. The results are shown in Table 9.

The results of the federated learning model are comparable to the centralized model, with an accuracy of 67% versus 69%. This indicates that federated learning can be a viable alternative to centralized learning. Especially considering the additional benefits that federated learning has on mitigating privacy issues.

Table 9. Federated learning model metrics

Metric	Value
overall accuracy	67% (49656/73600)
Test loss	0.797593
Asphalt test accuracy	51% (12864/24800)
Good brick test accuracy	84% (18995/22400)
Bad brick test accuracy	67% (17797/26400)
Precision	0.71
Recall	0.67
F1-score	0.67

5 DISCUSSION

With more time available this research could have continued by analysing why the new centralized model in Python has a lower accuracy than the model in Edge Impulse. The amount of labeled data could also be increased. The current model is only classifying 3 different road quality classes. Due to the nature of the received data, it was difficult to (manually) determine more different road quality classes. It would be interesting to see how the model performs with more road quality classes.

6 CONCLUSION AND FUTURE RESEARCH

This study aimed to develop a reliable approach to road quality assessment using inertial measurement units (IMUs) on bicycles. This is done with both centralized and federated learning models. Prior research has been done on road quality assessment using IMUs on bicycles. In comparison with the existing literature, this research focused on personalized factors in the data.

The centralized model developed in Edge Impulse achieved an overall testing accuracy of 97.04% with a weighted F-1 score of 0.97. The model performs well across all 3 different road quality classifications. These results underscore the effectiveness of the model to accurately identify the various road surfaces using data collected from the sensors.

In addition to the road quality classification model, a separate machine learning model was developed to classify participants based on personalized features derived from the IMU data. This gave valuable insights into which features have the biggest importance in uniquely identifying the different participants.

Federated learning was employed to mitigate the privacy concerns that come with crowdsensed data. A new centralized model was created as extracting model weights from Edge Impulse model turned out to be more challenging than expected. The new centralized model has an overall accuracy of 69% with a weighted F1-score of 0.673774. Subsequently, the federated learning model achieved an overall accuracy of 67% and a weighted F1-score of 0.671661. This slight decrease in performance demonstrates the potential of federated learning models as a viable privacy-preserving alternative to centralized models.

Future research could delve deeper into the personalized factors that affect the IMU data. Originally, this research aimed to mitigate personalized factors. For instance, through personalized federated learning. Due to the limited time available for this research, this was not achieved and is left for future research. Another interesting idea for future research is to combine more sensors in the classification of the road condition. For this research the camera was only used to determine the different road quality segments. Future research could combine video/image recognition with IMU data and other sensors such as LiDAR to form an even more comprehensive understanding of road conditions.

In conclusion, this research validates the feasibility of using IMU sensors on bicycles to effectively assess the road quality. Both centralized and federated learning approaches proved capable of accurately classifying different road surfaces. This study advances the current understanding of crowdsensing applications and lays the way for new research to mitigate personalized factors in crowdsensed road quality data.

REFERENCES

[1] EdgeImpulse Inc. (2023). 2024. *Edge Impulse - The Leading edge AI platform.* https://edgeimpulse.com/

[2] Warren Gay. 2014. *Raspberry Pi Hardware Reference.* https://doi.org/10.1007/978-1-4842-0799-4

[3] Johannes Heidt and Klaus Dorer. 2023. Classification and Prediction of Bicycle-Road-Quality using IMU Data. (2023).

[4] Mohammad Hijji, Rahat Iqbal, Anup Kumar Pandey, Faiyaz Doctor, Charalampos Karyotis, Wahid Rajeh, Ali Alshehri, and Fahad Aradah. 2023. 6G Connected Vehicle Framework to Support Intelligent Road Maintenance Using Deep Learning Data Fusion. *IEEE Transactions on Intelligent Transportation Systems* 24, 7 (July 2023), 7726–7735. https://doi.org/10.1109/TITS.2023.3235151 Conference Name: IEEE Transactions on Intelligent Transportation Systems.

[5] Mohammad Jahanshahi, Farrokh Jazizadeh, Sami Masri, and Burcin Becerik-Gerber. 2012. An Unsupervised Approach for Autonomous Pavement Defect Detection and Quantification Using an Inexpensive Depth Sensor. *Journal of Computing in Civil Engineering* 27 (Aug. 2012). https://doi.org/10.1061/(ASCE)CP.1943-5487.0000245

[6] Shuo Jiang, Zach Strout, Bin He, Daiyan Peng, Peter B. Shull, and Benny P. L. Lo. 2023. Dual Stream Meta Learning for Road Surface Classification and Riding Event Detection on Shared Bikes. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 53, 11 (Nov. 2023), 7188–7200. https://doi.org/10.1109/TSMC.2023.3295424 Conference Name: IEEE Transactions on Systems, Man, and Cybernetics: Systems.

[7] Ministerie van Binnenlandse Zaken en Koninkrijksrelaties and Ministerie van Infrastructuur en Waterstaat. 2022. Miljardeninvesteringen voor bereikbaarheid woonwijken in heel Nederland - Nieuwsbericht - Rijksoverheid.nl. https://www.rijksoverheid.nl/actueel/nieuws/2022/11/14/miljardeninvesteringen-voor-bereikbaarheid-woonwijken-in-heel-nederland Last Modified: 2023-10-05T16:58 Publisher: Ministerie van Algemene Zaken.

[8] Fátima González-Novo López. 2022. Bike lane quality estimation under variable speed conditions using off-the-shelf motion sensors. (2022).

[9] Deepak Nagaraj, Marcel Mutz, Supreeth Mysore Venkatesh, Lea Riebschlaeger, and Dirk Werth. 2022. A Practical Approach for Road Quality Estimation using Smartphone based Inertial Data: IMU data processing pipeline to estimate road quality. In *Proceedings of the 2022 7th International Conference on Machine Learning Technologies (ICMLT '22)*. Association for Computing Machinery, New York, NY, USA, 87–91. https://doi.org/10.1145/3529399.3529414

[10] OpenAI. 2024. *ChatGPT [Large language model].* https://chatgpt.com

[11] Kiran Peirens. 2023. Design and use of a modular measurement module for speed pedelecs. (2023).

[12] Daiyan Peng, Zach Strout, Shuo Jiang, and Peter Shull. 2019. A Road Condition Classifier via Lock Embedded IMU on Dock-Less Shared Bikes. In *Proceedings of the International Conference on Industrial Control Network and System Engineering Research*. ACM, Shenyang China, 32–36. https://doi.org/10.1145/3333581.3333597

[13] Steven Tazelaar. 2022. Automatic labeling of road quality using machine learning. (2022).

[14] Guido Van Rossum and Fred L. Drake. 2009. *Python 3 Reference Manual.* CreateSpace, Scotts Valley, CA.

[15] Astarita Vittorio, Vaiana Rosolino, Iuele Teresa, Caruso Maria Vittoria, P. Giofrè Vincenzo, and De Masi Francesco. 2014. Automated Sensing System for Monitoring of Road Surface Quality by Mobile Devices. *Procedia - Social and Behavioral Sciences* 111 (Feb. 2014), 242–251. https://doi.org/10.1016/j.sbspro.2014.01.057

A APPENDIX

A.1 On the use of AI

During the preparation of this work the author(s) used ChatGPT3.5 / ChatGPT4.0 [10] in order to create a structure for the paper, to review the written text, and to help in reformulating sentences. It was also used to help the coding process and was used as a base for the comments that explain the written functions. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the work.

A.2 Personalized features

Fig. 12. Feature importance participant 51

Fig. 13. Feature importance participant 59

Fig. 15. Feature importance: all data

Fig. 11. Feature importance participant 19

Fig. 14. Feature importance participant 70