National College of Ireland

Project Submission Sheet

Student Name:	Yash Bhargava
Student ID:	x22220861
Programme:	MSc In Data Analytics
Year:	2023-24
Module:	Data Mining and Machine Learning 2
Lecturer:	Prof. Athanasios Staikopoulos
Submission Due Date:	19-05-2024
Project Title:	Data Mining and Machine Learning 2 TABA
Word Count:	4863

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the references section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	Yash Bhargava
Date:	19-05-2024

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. Projects should be submitted to your Programme Coordinator.
3. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
4. You must ensure that all projects are submitted to your Programme Coordinator on or before the required submission date. **Late submissions will incur penalties.**
5. All projects must be submitted and passed to successfully complete the year. **Any project/assignment not submitted will be marked as a failure.**

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

AI Acknowledgement Supplement

[Insert Module Name]

[Insert Title of your assignment]

Your Name/Student Number	Course	Date

This section is a supplement to the main assignment, to be used if AI was used in any capacity in the creation of your assignment; if you have queries about how to do this, please contact your lecturer. For an example of how to fill these sections out, please click here.

AI Acknowledgment

This section acknowledges the AI tools that were utilized in the process of completing this assignment.

Tool Name	Brief Description	Link to tool

Description of AI Usage

This section provides a more detailed description of how the AI tools were used in the assignment. It includes information about the prompts given to the AI tool, the responses received, and how these responses were utilized or modified in the assignment. **One table should be used for each tool used**.

[Insert Tool Name]	
[Insert Description of use]	
[Insert Sample prompt]	[Insert Sample response]

Evidence of AI Usage

This section includes evidence of significant prompts and responses used or generated through the AI tool. It should provide a clear understanding of the extent to which the AI tool was used in the assignment. Evidence may be attached via screenshots or text.

Additional Evidence:

[Place evidence here]

Additional Evidence:

[Place evidence here]

DATA MINING AND MACHINE LEARNING 2 TABA

Yash Bhargava
x22220861@student.ncirl.ie
MSc in Data Analytics
National College of Ireland

Question No.1 - Case Study
Topic: Prediction of volcanic eruption

Introduction

In the study of geophysics, the prediction of volcanic eruptions is the most important because of the need to protect infrastructure and individuals from the terrible consequences of volcanic activity. To classify volcanic seismic events, this study focuses on the use of modern machine learning models by employing recurrent neural networks (RNNs) and long-short-term memory (LSTM) networks. The complex nature of the temporal dependencies present in volcanic activity has not always been fully captured by conventional seismic data analysis techniques. By utilizing LSTMs behavioural modelling capabilities, which are particularly effective at processing data in sequence and detecting trends across time.
The proposed methodology presents a comprehensive approach, that begins with Exploratory Data Analysis (EDA) and data cleaning to ensure the consistency and quality of the seismic datasets. Various feature engineering and extraction techniques will be used to convert raw seismic signals into meaningful representations to improve the performance of the model. The study also shows the importance of dimensionality reduction and feature selection to enhance model efficiency and also the input parameters of the LSTM model for optimized performance. Scalability issues are also covered, by considering the requirement for scalable systems to manage huge amounts of data and real-time monitoring. The ethical implications will also be taken care of with a focus on the need to maintain privacy, maintain public safety, and promote equal allocation of resources.

Structure of the report

This report consists of various sections which discuss each step of our proposed methodology including Exploratory Data Analysis and data cleaning, section 2 consists of dimensionality reduction and feature selection, section 3 consists of feature extraction, section 4 consists of the choice of modelling techniques, section 5 consists of hyperparameters that were optimised, section 6 consists of evaluation metrics for assessing the model's performance, and last two sections include the scalability issues and ethical implications of this proposed methodology.

Exploratory Data Analysis / Data Cleaning

Data cleaning is an important step in this proposed methodology as it ensures better accuracy, consistency and performance of the model. By doing proper data cleaning, it can provide accurate classification, and reduce the error rate of the classifier which can further lead to accurate decision-making. Data cleaning consists of imputing missing values, treating outliers, and removing duplicates, in the case of images it is resizing, flipping, augmentation etc. If the data is cleaned, it improves the reliability of the classification model which can identify the seismic patterns in the images and decreases the risks of incorrect classification which can lead to better plans for saving the lives of people and avoiding damage. We have reviewed several papers in which some of the papers have done data-cleaning as we can see in the study (Titos et al. 2018), firstly the author divided the data into training i.e. 75% and test sets i.e. 25% respectively and the balanced shuffle was done randomly so that the correlation between the training data was avoided and cross-validated with 4 batches of the test data was used to verify the capability

of the model for generalization purpose. In the study (Lara et al. 2020), which presents a supervised classification system in which the dataset of signals was considered, they did signal conditioning by analysing the PSD signals and setting the frequency of all the signals to 50 Hz. Also, the instrumental correction was done i.e., converting the seismic counts into m/s for standardization into unity. The study by (Venegas et al. 2019*b*), utilised the dataset of 668 seismic events in which they have 587 long periods (LP) and 81 volcano-tectonic (VT) events, and the dataset was normalized using the min-max method for standardization. (Titos et al. 2019) used a dataset of 9332 volcanic seismic signals which were bandpass filtered in the range of 1 to 25 Hz and got a spectrogram by employing a short-time fast Fourier transform of 512 points which were converted into 32 X 32 sized and transformed into greyscale images. In the study (Shyamala et al. 2022), which used the MicSigV1 dataset that consisted of 1187 seismic records, they converted the seismic signals into grayscale images which reduced the original image sizes by 25% and the values of the pixels were normalized using the min-max method. In the study (Tempola et al. 2018), they utilized the seismograms that were of the separated files at 12000 s duration, the authors did the labelling and identification of the events manually by taking care of the waveform, spectrogram, and the location. However, these studies show the importance of thoroughly cleaning data before utilizing it in machine learning models and having seismic event datasets appropriately structured.
Although the steps that were done for EDA and data cleaning in previous studies, we have not seen any data augmentation that has been done such as introducing the time shifts, and noise reduction which can further improve the robustness of the proposed model for this study and generalize it. Also in the proposed methodology, we can employ some data resampling and add synthetic data to increase the size of the training datasets.

Dimensionality Reduction/ Feature Selection

Reducing the dimensions or the selection of significant features are very important step and plays a vital role in this methodology i.e., in building a classification model for the prediction of volcanic seismic events. The process of reducing the dimensions involves determining the most relevant variables if the dataset has a large number of variables and this step also reduces the computational complexity of the analysis. By selecting significant features, this step reduces the risk of overfitting and helps the classification model to become generalizable which results in more accurate prediction. For the classification of volcanic seismic events, where accurate prediction and on-time are necessary, this step ensures that the most significant features are prioritized first as compared to irrelevant features. After carrying out the literature review of the several papers, there are only some papers which have utilised this step as we can see in the study Lara et al. (2020), due to the very high number of dimensions i.e., 649 dimensions, there was a need to reduce the number of dimensions so that the dispersion of data may not be caused. So, as in the study, the authors have found out that the first 200 components were obtained after the implementation of principal component analysis which was 99.7% of the feature vector significance. Similarly, the study (Venegas et al. 2019*a*), the study has employed the five filter-based feature selection methods i.e., Information Gain (IG), One Rule, RELIEF, Chi2 Discretization, and u-filter and have got uFilter+5f as the best feature selection methods with an accuracy of 96.71%. in a category of 25:75 training and test split, uFilter+5f with an accuracy of 95.88% in the case of 50:50 training and test split and 1R+5f with an accuracy of 94.47% in the case of 75:25 training and test split in the classification of long- period and volcano-tectonic seismic events. These findings show the importance of feature selection and dimensionality reduction in reducing the side effects of dimensionality and improving the efficiency and accuracy of models.
However, from the previous studies, we can utilize ensemble feature selection in which we can combine 3-4 feature selection methods to improve the feature selection performance and optimize the performance of the classification model. Also, we can integrate the dimensionality reduction techniques and feature selection methods to improve the performance of the proposed model.

Feature Engineering/ Feature Extraction

Extraction of features is an important step in this methodology for the classification of volcanic seismic events as this can increase the performance of the model. Using this step, we can convert the raw data into important features that better present the patterns that were important to the volcanic seismic event. Also, feature engineering can help us to determine essential features and how much they are correlated with the dataset. After the successful feature engineering, these features provide some useful insights about the data and help in building a sophisticated predictive model which can classify volcanic seismic events efficiently. These extracted features can be able to build a better model which can be

more accurate and can classify the volcanic events effectively which results in early warning for volcanic eruption, assessment of risks to individuals and damages to property. After a thorough review of several papers, as we can see in the paper (Titos et al. 2018), the study employed a feature extraction process in a pipeline in which the input for 9332 seismic signals was bandpass filtered between the range of 1 Hz to 25 Hz and sampled at an average of 50 Hz and the raw data was transformed into a domain where each event has the same length by encoding each signal using LPCs. This results in a set of 21-dimensional input vectors, containing temporal and frequency features. In the paper (Lara et al. 2020), the study presented a proposed feature extraction block in which the seismic signals were decomposed using the Empirical Mode of decomposition (EMD) which is used for the analysis of non-linear and non-stationary signals. Then the study chose the number of IMFs which was based on each seismic signal. The authors utilised the variance contribution ratio for the IMF w.r.t variance and generated the most significant attributes. In the study (Venegas et al. 2019*b*,*a*), the authors have employed 668 seismic events in which 587 LP and 81 VT seismic records were considered and a set of 84 features were extracted which consisted of 13 features from the time domain, 21 features from the frequency domain and 50 features of the domain of scale. In the study (Titos et al. 2019), the author has used the concept of transfer learning for extraction of the features, in which the study utilises the LeNet architecture which was used to process the spectrogram images and provides a feature vector which was further used for training the model. In the study (Ramirez Jr & Meyer 2011), the study employs the use of multi-scale feature extraction for the clustering of the seismic data on manifolds of low dimensions and proposed information theoretic measures for merging regression scores on multi-scale feature manifolds. In the study (Manley et al. 2020), the study processes the raw seismic data for the detection of seismic events and then the authors have got the time series such as the event count of the attributes on a day-to-day basis, then extracted the features from the time series data of the seismic events and trained and tested the models on these extracted features.
After reviewing these papers that have employed different feature extraction techniques, to improve the performance metrics of our proposed model, we can employ feature extraction from the seismic data using deep learning such as Convolutional Neural networks (CNN) which allows us the model to train on hierarchical representations of the features of the seismic records for the accurate and efficient classification of the seismic events.

Choice of Modelling Techniques

The choice of models for the prediction of volcanic eruption cannot be underestimated as it is hard and changeable to predict seismic events. The precision in the prediction model is based on how the chosen models can find the significant features from every seismic record based on time, frequency, and scale domain. Therefore, future changes in seismic records that may occur should be adaptive to the selected models as well as the behaviour of previously used models so that they would become more effective. This is because if wrong modelling techniques are chosen, then the volcano eruption will not be reliable which affects the decision-making process. Although we have reviewed papers in which the authors have utilised the potential of machine learning and deep learning modelling techniques for accurately predicting the volcano eruption using seismic records dataset. After a review of the many research studies that focus on the classification of volcanic seismic events, a variety of different modelling techniques were employed by the previous studies. Within the range of machine learning, several different models were proposed which include stacked denoising autoencoders and deep belief networks (DBNs) by (Titos et al. 2018), conventional techniques such as multilayer perceptron (MLP), linear discriminant analysis (LDA) by (Lara et al. 2020), support vector machine by (Venegas et al. 2019*b*, Manley et al. 2020), k-Nearest Neighbour (KNN) by (Venegas et al. 2019*b*), Random Forest by (Lara et al. 2020, Venegas et al. 2019*b*, Manley et al. 2020), Naïve Bayes (NB) by (Venegas et al. 2019*b*, Tempola et al. 2018), logistic regression (LR) by (Manley et al. 2020), Gaussian process classifier by (Manley et al. 2020), and Gaussian Mixture model (GMM) by (Venegas et al. 2019*a*). Moreover, if we consider the deep learning-based techniques, the authors have employed convolutional neural networks (CNN) by (Calderón et al. 2020), transfer learning-based classification by (Titos et al. 2019), radial basis function neural network (RBFNN) by (Shyamala et al. 2022) and AlexNet architecture (Anantrasirichai et al. 2019) which was based on convolutional neural network (CNN). Every study has a different approach, often using various techniques and fine-tuning models to obtain the best results in classifying seismic occurrences connected to volcanic eruptions.
After the review of papers i.e., for the classification of volcanic seismic events, instead of using the conventional techniques, we will utilise the recurrent neural networks (RNNs) which can be advantageous

particularly Long-Short Term Memory Networks (LSTM) because of its ability to identify temporal dependencies that are important for the classification of the volcanic seismic event.

Hyperparameter Optimisation

Optimizing the hyperparameters is essential for building models with high-performance, generalizable, and efficient machine learning models. By tuning the hyperparameters, the researchers could take full advantage of the machine learning and deep learning models and get excellent results in their domains. As we have thoroughly reviewed the many papers, as we can see in paper (Titos et al. 2018), the performance of deep neural models was evaluated using a total of 250 x 100 various combinations of hyperparameters. The number of hidden units in the first, second, and third levels varies between 50 and 1250, increasing by steps of twenty-five hidden units. In this study, learning rates were evaluated within the range of 0.000001 to 0.01. The dataset was normalized by changing the mean and variance, and a sigmoid function was used as the nonlinearity function. In the study, the employed 2 models, both the models i.e., sDA and DBN, include a SoftMax probability layer, with 7 target outputs, related to each class of the dataset. These models were trained with a batch size of 10 training samples. Similarly, in the study (Venegas et al. 2019*b*), the model proposed was based on the combination of five machine learning classifiers in which the machine learning classifiers were configured except for one Naïve Bayes classifier and the model's parameters were fine-tuned by the 10-CV method. In the study (Calderón et al. 2020), the number of epochs, learning rate, max-pooling kernel size, convolutional kernel size, and padding types were the hyperparameters that were optimized The study optimizes the convolutional kernel with different dimensions, adjusts the padding types to no padding and same, increases the steps to optimize the learning rate, uses different dimensions for the max-pooling kernel size with a stride of 2 or 3, and increases the steps by 50 to reach a maximum of 500 iterations.
After the review of the papers, we have found out that many researchers have not fine-tuned the hyperparameters, so to achieve the best performance of our proposed model and for the accurate classification system, we will fine-tune some of the important hyperparameters of the LSTM model such as trying over a different number of LSTM units, length of the input sequence, the learning rate of the model, dropout rate to prevent overfitting of the model, number of layers and many more etc to get better performance from the models that were proposed in previous studies for the classification of volcanic seismic events.

Model Evaluation

Model evaluation is a crucial step of any methodology in building and training any machine learning or deep learning model, especially in classifying volcanic seismic events. The process of evaluation of the model's performance and effectiveness using some key metrics such as accuracy, precision, recall, F1 score, ROC, and AUC etc to ensure that the model performs accurately and efficiently. There are distinct reasons why evaluating these models' performance is important. An accurate classification of volcanic seismic events is important for keeping track of volcanic activity and the prediction of its eruption which is essential for public safety and management of volcanic hazard. As the nature of the seismic event data has noise and various patterns identifying and classifying based on these features needs a robust model which can be generalized to the test data to prevent overfitting of the data. It is also important to evaluate the model's performance to guide further improvements in the model's performance and informed decision-making. After a thorough review of many papers in this domain i.e., for the classification of volcanic seismic events such as in the papers (Titos et al. 2018, Anantrasirichai et al. 2019), the evaluation metrics that were used in these papers were Accuracy, Precision, Recall, F1 Score, Mean Squared Error (MSE) and the processing time in Venegas et al. (2019*a*) which were the most common performance metrics for the classification, regression and deep learning models. In the studies (Lara et al. 2020, Titos et al. 2019, Shyamala et al. 2022, Ramirez Jr & Meyer 2011), the researchers have only considered accuracy as an evaluation metric for the classification model which was not sufficient for evaluating the performance and efficiency of the model. Similarly, in the study (Venegas et al. 2019*b*, Calderón et al. 2020), the researchers have only taken AUC score and accuracy as a key performance metric which was again not sufficient for assessing the performance of the proposed model. Additionally, in the paper (Tempola et al. 2018), the study only considered the accuracy of the model when the classification was done on the test data in two different scenarios and the standard deviation was considered as a measure of the dispersion of accuracy values from different scenarios which were done using three cross-validations. In the study (Manley et al. 2020), we observed that the study does not mention any of the model evaluation metrics for the classification of seismic time series data by using machine learning

models.
After the review of the papers in this domain, most of the studies have not considered some important model evaluation metrics as these will cause incorrect and appropriate evaluation of the model's effectiveness. In this study, we will consider the important key performance metrics such as Accuracy, Precision, Recall, and F1 Score as these metrics can significantly evaluate the model's performance and we can be able to understand the model's performance and decide whether the model is accurate and reliable for the prediction of volcanic eruptions.

Scalabilty Issues

We have observed that scalability issues were not mentioned in these papers, and no one has talked about them when discussing the prediction of volcanic eruptions after reviewing all the previous studies. But when working in this problem domain, it's very important to address the scalability issues. There are several reasons which are important like firstly, the complex nature of volcanic activity secondly its potential effects on large geographical areas, and scalability problems in the prediction of volcanic eruptions. To ensure the warnings and action plans, effective prediction models need to take into consideration the large time and space scales of volcanic activity. The most important thing is scalability allows for real-time monitoring and risk assessment while handling the many different data sources and volumes required for accurate forecasts. Scalable systems also encourage data sharing and connectivity among stakeholders, which helps to promote international collaboration and security for both sides. In the end, I would mention that the scalability issues will help us protect infrastructure and human lives in dangerous regions by improving the ability to predict and reduce the risks caused by volcanic eruptions.

Ethical Implications

The ethical implications of predicting volcanic eruption have not been discussed in the above-mentioned papers that we reviewed until now, and none of the authors discussed the challenges of this area. But we are looking into this area because it is important to find the complex areas where improvements in research fulfil the needs of people, society, and the environment. There are many different ethical factors to consider in our mind while working in this domain, such as protecting the public by ethically informing them of potential risks or taking to the needs of minority groups that are most affected by volcanic hazards. The people and different small communities who live in the nearby volcanic eruption areas need to have access to accurate information to make decisions about their safety and daily life. Also, inequality issues must be discussed to give priority to those who were at the most risk to decrease differences in risk-taking through prediction. Ecosystems and biodiversity can accidentally be affected by monitoring activities, so environmental effects are also important to consider. Critical ethical concerns about data security and privacy also surface, which require careful data handling to protect people's right to privacy. More importantly, solving ethical issues involving resource allocation, data sharing, and fair participation among working together nations and groups is necessary to promote international collaboration. Volcanic prediction persons can make sure that their work is both ethically and scientifically accurate by recognizing and addressing these ethical implications.

Question No.2 - Paper Review

Topic: Augmenting roadway safety with machine learning and deep learning: Pothole detection and dimension estimation using in-vehicle technologies

Structure & Title

This article is presented in a well-structured format and consists of all the necessary elements such as Abstract, Introduction, Methodology, Results and Conclusion. Now the title of this article i.e., "Augmenting roadway safety with machine learning and deep learning: Pothole detection and dimension estimation using in-vehicle technologies" states that it uses machine learning and deep learning techniques, the research, article has only proposed a deep learning algorithm so according to me it should be only the deep learning which another aspect of artificial intelligence. But it is fine as the title focuses on pothole detection and estimating the dimensions of potholes using in-vehicle technologies.

Abstract

The abstract reflects all the main aspects of this research article and why it is necessary to do this research it states that pothole detection is necessary to ensure the safety of the vehicle and to improve the safety of the individual's property. The abstract clearly states the study in which it presents a deep learning model which was used to detect the potholes and estimate the dimensions of the potholes. The abstract also presents the results of this research study in which the performance of the proposed model was determined based on certain metrics such as Precision, Recall, F1 Score and Map. Although, the abstract summarizes the research article by considering all the main aspects of the paper.

Highlights

- Pothole detection and dimensional estimation using deep learning.
- You Only Look Once (YOLOv5) model used in this research.
- Combination of computer vision and deep learning techniques.
- Precision: 93%, Recall: 91.6%, F1-Score: 87%, mAP: 96.3%.
- Used Python's OpenCV library for the dimensions labelling of the potholes.

Introduction

In the research article, why potholes should be detected and their sizes estimated is clearly stated in the introduction of this research paper because it is an essential component of road maintenance. The writer has highlighted in this regard that what this research aims to achieve is to detect the potholes using a deep learning algorithm which can estimate their size so that we stop the road from being dangerous. The introductory part shows the context under which potholes have caused problems and the challenges experienced by traditional approaches to pothole detection and dimension estimation.

In this section, the author discussed the issues that are caused by delaying the maintenance of the roads such as severe injuries and deaths. The author has also discussed today's road conditions which can discuss how these problems can be solved by this research to detect the potholes using computer vision and deep learning techniques and estimate the size of the potholes using the in-built vehicle technologies. The author also discussed the importance of this research as it can contribute to the transportation agencies. This study aims to detect real-time potholes detection for road condition monitoring and proposes a model based on deep learning for the detection of potholes and the estimation of the dimensions of the potholes.

Methodology

In the methodology section, the study presented a detailed description of each of the steps of this methodology such as where the datasets were collected from, what type of data preprocessing was done on the dataset and what the models were chosen for the prediction and evaluation of this model.

The study used the dataset that was collected from Smartphone-Based Road Damage Detection and Classification which were 26,336 images that were taken from smartphones, and another was from pavement distress from Roboflow which consisted of 665 images. This study presented where the data was collected from, and a random sample of 1876 images of potholes was used for data preparation which includes the types of potholes images like dry potholes, wet potholes, daytime, and night-time images etc. And for the estimation of the pothole dimension the study has used the videos from American Honda Motor Co. of 104 h of real driving scenarios of California.

Also, this study presented a suitable design for the detection of potholes and estimating the dimensions of the potholes after the data collection step was performed successfully then the study presented various data preparation steps in which the image annotations and marking of the bounding boxes in the images was done using onlineakesense.ai tool, data augmentation, flipping of the images, rotation of the images, image blurring and random zoom, vertical and horizontal shifts, brightness of the image was adjusted. This study was consisted of two parts, the first was of pothole detection and the second estimation of the pothole size. This study proposed the models for the detection of the pothole in which they employed the use of the You Only Look Once (YOLOv5) model which is a CNN architecture used for the object detection a detailed description of this model was mentioned in the research paper. For using this model, the authors have trained this model for 2.05 h on 140 epochs. In this study, the authors followed a clear structure of procedure that would be followed which can be validated by the flowchart of the proposed model for pothole detection and estimation of sizes using the OpenCV library in Python. All the steps were followed in a meaningful way as it can finally provide the answer to the purpose of the study.

In the study, random samples of images were taken for the preparation of the data for modelling as it is appropriate for the model building to detect all distinct types of potholes to ensure road safety. All the information about the architecture of the YOLOv5 model used was explained clearly in detail. The authors have mentioned the size of training and testing sets for the second part of this study i.e., estimation of the dimension of the detected potholes can be achieved using Hough lines in the OpenCV library, which was mentioned, and the pseudocode was also provided clearly.

Finally, this study presented two parts of this study, the one in which diverse images of potholes on which the YOLOv5 model was trained and performed very well on the images datasets and the second one in which the images of the potholes and videos were employed for labelling the dimensions of the potholes using the Hough lines from OpenCV library and for determining the measurement scale of the pothole size marking the article have utilized the lane keeping assistance systems.

Results

In this paper, the results that were interpreted in the paper were clear and understandable. In this paper, the author conducted a statistical analysis which was for the detection of loops which was appropriate, and this can be validated by the p-value of 0.037 which indicates that the results were significant. According to me, I would recommend the editor of this research paper to ensure the accuracy of the statistical analysis. In the study, the results were good and explained the potential of the in-built vehicle technologies that were utilized by the study for pothole detection and dimension estimation of the pothole. In the future, the editor is required to investigate and validate their statistical analysis in this study.

Conclusion/ Discussion

The author's conclusions in this study have been confirmed by the findings of the research. The author has demonstrated that the results were related to pre-study assumptions and compared them with proposed models in earlier studies. This research validates previous theories about detecting potholes and estimating their dimensions using computer vision approaches. In this regard, a model that was suggested in the paper uses in-vehicle technologies to identify potholes and give an estimation of their sizes. However, the conclusion part of this study provides more insights into what kind of theoretical significance can be drawn from its results as well as how, according to the author's claims, the proposed one can outperform other models with better performance based on this idea, neither has the writer discussed how road maintenance or safety will be affected if his proposal is implemented.

Language

The language used in this paper is clear and concise, which is easy to understand for everyone. However, some grammatical errors can be improved for better readability. Otherwise, the language of this paper is clear and the paper format is well-structured and the figures and tables in this paper accurately explain all the important parts of the article they are consistent and in logical order as well but some improvements can be made for better understanding and readability.

Previous research

The article appropriately builds upon previous research and references the previous works accurately and comprehensively. Additionally, the article shows recent advancements, including using YOLOv5, a state-of-the-art object detection algorithm, explaining its commitment to keep it with current technological developments. The references are clear and accurate, giving complete information on the cited works. The article doesn't leave out any important studies, even though they appear to be a wide variety of studies, ensuring a thorough overview of the area. This complete referencing shows clear progress from previous research to the article's original contributions, proving the novel approach to real-time, cost-effective pothole detection using in-vehicle technologies.

References

Anantrasirichai, N., Biggs, J., Albino, F. & Bull, D. (2019), 'A deep learning approach to detecting volcano deformation from satellite imagery using synthetic datasets', *Remote Sensing of Environment* **230**, 111179.

Calderón, F., Pérez, N., Benítez, D. S. & Riofrío, D. (2020), Volcanic seismic event classification based on cnn architectures, *in* '2020 IEEE ANDESCON', IEEE, pp. 1–6.

Lara, P. E. E., Fernandes, C. A. R., Inza, A., Mars, J. I., Métaxian, J.-P., Dalla Mura, M. & Malfante, M. (2020), 'Automatic multichannel volcano-seismic classification using machine learning and emd', *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **13**, 1322–1331.

Manley, G. F., Pyle, D. M., Mather, T. A., Rodgers, M., Clifton, D. A., Stokell, B. G., Thompson, G., Londoño, J. M. & Roman, D. C. (2020), 'Understanding the timing of eruption end using a machine learning approach to classification of seismic time series', *Journal of Volcanology and Geothermal Research* **401**, 106917.

Ramirez Jr, J. & Meyer, F. G. (2011), Machine learning for seismic signal processing: Phase classification on a manifold, *in* '2011 10th International Conference on Machine Learning and Applications and Workshops', Vol. 1, IEEE, pp. 382–388.

Shyamala, K., Raju, Y. D. S., Swarnalatha, C., Saidulu, D. & Tha, M. P. (2022), Leveraging machine learning to predict volcanic eruptions by function neural networks, *in* '2022 4th International Conference on Inventive Research in Computing Applications (ICIRCA)', IEEE, pp. 1220–1225.

Tempola, F., Muhammad, M. & Khairan, A. (2018), Naive bayes classifier for prediction of volcanic status in indonesia, *in* '2018 5th International Conference on Information Technology, computer, and Electrical Engineering (ICITACEE)', IEEE, pp. 365–369.

Titos, M., Bueno, A., Garcia, L. & Benitez, C. (2018), 'A deep neural networks approach to automatic recognition systems for volcano-seismic events', *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **11**(5), 1533–1544.

Titos, M., Bueno, A., García, L., Benítez, C. & Segura, J. C. (2019), 'Classification of isolated volcano-seismic events based on inductive transfer learning', *IEEE Geoscience and Remote Sensing Letters* **17**(5), 869–873.

Venegas, P., Pérez, N., Benítez, D., Lara-Cueva, R. & Ruiz, M. (2019*a*), 'Combining filter-based feature selection methods and gaussian mixture model for the classification of seismic events from cotopaxi volcano', *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **12**(6), 1991–2003.

Venegas, P., Pèrez, N., Benítez, D. S., Lara-Cueva, R. & Ruiz, M. (2019*b*), Building machine learning models for long-period and volcano-tectonic event classification, *in* '2019 IEEE CHILEAN Conference on Electrical, Electronics Engineering, Information and Communication Technologies (CHILECON)', IEEE, pp. 1–6.