2022-28-0004 Published 05 Oct 2022

Multi-Sensor Information Fusion for Determining Road Quality for Semi-Autonomous Vehicles

Trisanu Bhar and Hrishikesh Venkataraman IIIT Sricity

Jaswanth Nidamanuri Project Scientist, IHub-Data, IIITH

Citation: Bhar, T., Venkataraman, H., and Nidamanuri, J., "Multi-Sensor Information Fusion for Determining Road Quality for Semi-Autonomous Vehicles," SAE Technical Paper 2022-28-0004, 2022, doi:10.4271/2022-28-0004.

Abstract

Pothole detection in Intelligent Transportation Systems (ITS) vehicles has been a part of Advanced driver-assistance systems (ADAS) for a long time. Various sensors have been used for this purpose so far: Accelerometer, Gyroscope, etc. However, the fusion of multiple modalities of information from different sensors remains a challenge, mainly owing to the different sampling rates and varying frame rates used by each sensor. Other sensor types like Radar and LIDAR, though precise, are difficult to use, thus forcing us to look for low-cost solutions. Our proposed work uses Accelerometer and Gyroscope sensor fusion to predict pothole presence in Indian scenarios. Previous works have mainly dealt with predicting potholes with data collected using either traditional machine learning techniques like Decision trees, (Support Vector Machines (SVM)'s and Light Gradient Boosting Machine (LGBM) and deep learning methods using neural networks and attention mechanisms. In this work, the main focus is on using Convolution Neural Network-based methods to extract information from the sensor data after appropriate preprocessing. Notably, time series models such as Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), and Transformers integrated with lightweight attention units are then used to classify and predict the presence of potholes based on the information extracted. Further, different classes of neural networks and transformers like Involution Neural networks are investigated for reliable predictions. Subsequently, this information is used to predict potholes and may be used to develop a Road Quality index for Indian Roads which will indicate the quality of a given stretch of road. This paper demonstrates how the proposed method would provide greater accuracy for the prediction of potholes when a vehicle passes through a particular road.

Keywords

ADAS, Attention, Fusion, IMU, INN

1. Introduction

Potholes on Indian roads have become very common. Due to this reason, driving has become a menace. Vehicles also suffer a lot of damage from sudden movement over potholes. It is also the cause of many accidental deaths related to vehicles overturning after coming in contact with potholes. An estimated total of 3,564 road accidents occurred in India due to potholes in the year 2020, and 4,775 deaths in the year 2019 [1]. Given these statistics, it is imperative to design an ADAS system for pothole prediction in Indian Road Traffic environments.

Existing systems have been using vision-based solutions for detecting potholes. For this purpose, many deep learning algorithms, namely R-CNN, Faster R-CNN, and YOLO [2] have been used. Camera-based solutions, though working well, aren't sufficient. Additionally, camera-based solutions may not work well in adverse conditions like heavy rainfall, in which vision may get blocked or get partially covered, giving faulty information. Hence, we consider the IMU sensor's information for the prediction of pothole presence. IMU sensor consists of accelerometer and gyroscope readings along 3 axes (X, Y, and Z). We fuse this information before proceeding with our model architecture.

The key contributions of this proposed work are as follows:

1. Multi-sensor fusion of information from Accelerometer and Gyroscope.
2. Proposed a new architecture with INN and Transformers.
3. Detailed ablation study comparing the performances of models.

Section II discusses the prior art. Section III explains the design and proposed deeper architectures. Section IV illustrates the results obtained with a detailed ablation study. Finally, section V concludes the proposed work by providing the possible research directions.

2. Related Work

Detecting the potholes is challenging in unstructured road traffic environments. Many works have tried to classify pothole detection with traditional Machine learning approaches like SVMs [3] or either used Deep learning methods which consist of using 1-D CNNs with a Multi-Layer Perceptron (MLP) head [4] to classify the resulting time series data we get from data collection or with Attention mechanisms using LSTMs and GRU [5]. Pawar et al. [6] have dealt with data collection by collecting raw accelerometer data, over the X, Y, and Z-axis using a smartphone application. Five trips were made over a given selection of roads with sufficient pothole presence. The resulting corpus consisted of 24 input features and a resulting label for each row of data corresponding to whether the point is a pothole or not. For preprocessing the dataset, grouping over every 2-second interval was done, along with normalizing the dataset.

Bhatt et al. have [3] tried to use Machine learning approaches, mainly the Support Vector Machines (SVM) based approach to classify accelerometer and gyroscope measurements. Feature engineering was done manually wherein the data were grouped and a set of 26 features were extracted from each of the resulting groups. Using this approach, an SVM was trained on the resulting data. Using the Radial basis function (RGF) kernel and gradient boosting, it achieved a test accuracy of 93.4% on their dataset. Zhang et al. [5] have proposed another method for the collection and preprocessing of the raw accelerometer and gyroscope data. Their work primarily deals with driver behavior detection, but the same method is quite useful for other tasks such as pothole detection as well. In their method, the raw data collected from the sensors have been rotated and corrected under "standard posture", as the device placement was not in the position during data collection and also to counter the effect of gravity along the axes. This has been done by defining 2 rotational matrices and multiplying the accelerometer values with them. The authors have noted that when theta = ±90°, they might be faced with the Gimbal Lock problem, but the vehicle can't reach ±90°, so this problem has been ignored. Furthermore, they have also proposed an attention-based module for their work in Driver activity recognition. Nidamanuri et al. [7] proposed the hybrid CNN-LSTM model for analyzing the driver distraction and driving behavior from the tr-axial sensor data measurements. Notably, the proposed hybrid model provides 99% of test accuracy on the test data values. The authors in [8] proposed a novel architecture for designing safe drive assistance considering both in-vehicle driver's distracted behavior and external road traffic environmental scene perception. Additionally, a new dataset is collected for Driver In-vehicle distraction analysis and is used for real-time validation on challenging unstructured traffic scenes.

Machine learning approaches do not show greater accuracy, and feature extraction from sensor data also becomes tedious, hence we look at deep learning techniques which again pose a problem with regards to increasing complexity while maintaining greater accuracy. In this regard, we propose a Transformer-based model along with using a 1D - Involution Neural Network (INN) - INN-former to improve accuracy for the prediction of potholes while maintaining low complexity.

3. Proposed Work: INN-former

This paper proposes using INN and Transformers (namely, the INN-former) to improve the accuracy of detection of potholes after the appropriate collection of data is needed. Further, this data is tested on already established baselines to get an estimate of the performance of the models which the proposed model can compare against. These have been done experiments keeping in mind that the raw data is time-series in nature.

3.1. Data Collection and Preprocessing

For data collection of tri-axial raw sensor readings for both Accelerometer and Gyroscope, a smartphone application was used. Data collection was done in a 2-wheeler vehicle while holding the device in a frontal position, as shown in Fig 1. The trial was completed for a total duration of 3 minutes on a road with sufficient pothole presence. During this process, the vehicle was traveling at a constant speed and all the readings were recorded as the vehicle passed over the potholes. Fig. 2 shows the raw sensor readings as collected from the trials along with the filtered data. For data preprocessing, a lowpass Butterworth filter has been used for filtering both the Accelerometer and the Gyroscope values. Mathematically, this can be described as follows:

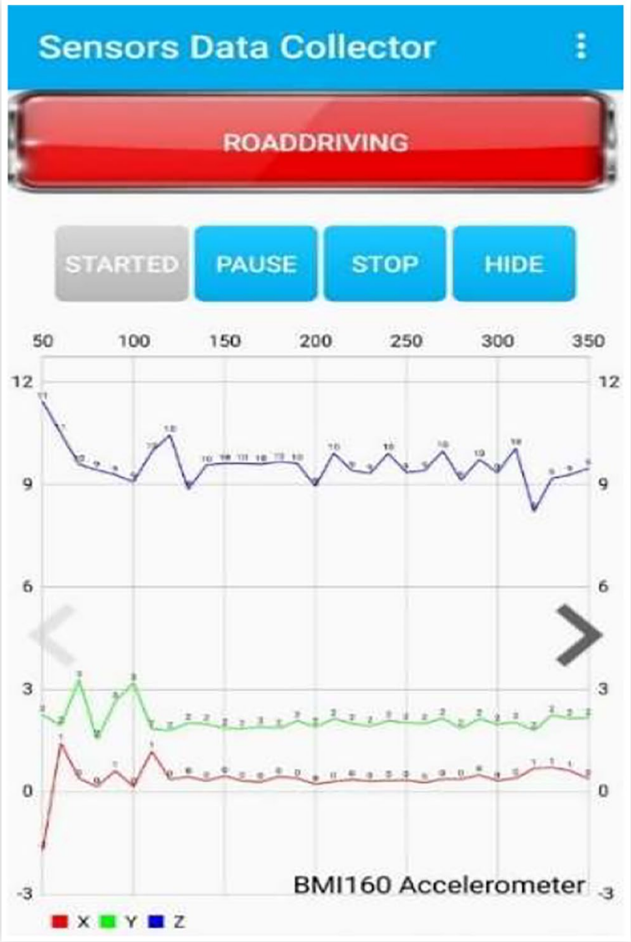
$$a'(x) = Butterworth(a(x)) \tag{1}$$

$$g'(x) = Butterworth(g(x)) \tag{2}$$

$$d(x) = a'(x) + g'(x) \tag{3}$$

In the above equations (1) and (2), $a(x)$ and $g(x)$ represent the raw accelerometer and gyroscope data respectively. *g'(x)* and *a'(x)* are the filtered data. The order of the Butterworth filter is taken as 2 and the cut-off frequency as 0.5 Hz. Further, the data of the Accelerometer and Gyroscope has been concatenated, hence fusing the information from the sensors. This is represented by *d(x)* in equation (3). The dataset has been annotated using the video that was recorded during the collection of the raw data. The presence of potholes has been annotated as follows:

FIGURE 1 Raw sensor collection from Android application. The above image shows data collected from a random event. The app is “Sensor data collector”.



- POTHOLE = 1
- NOT_POTHOLE = 0

Followed by which windowing has also been done to achieve uniformity concerning the sampling. A window length of 256 and a stride length of 6 has been chosen. The same strategy has been applied to the labels, with one change such that each row of windowed sensor readings consists of a target label. This has been achieved by using majority voting among the occurrences of both the label values. Finally, the dataset with labels that will be used for the prediction of potholes is prepared.

3.2. Model Architectures

In this section, various model architectures that have been experimented with have been described, followed by a detailed ablation study on the results that have been recorded from the experiments. First, a baseline model has been established for the proposed model to compare against. Following this, other various attention mechanisms have been explored to increase our accuracy and beat the baseline accuracy. Finally, Transformers, as well as Involution Neural Networks, have been explored for our tasks.

FIGURE 2 Raw Accelerometer and Gyroscope readings were collected using the application.

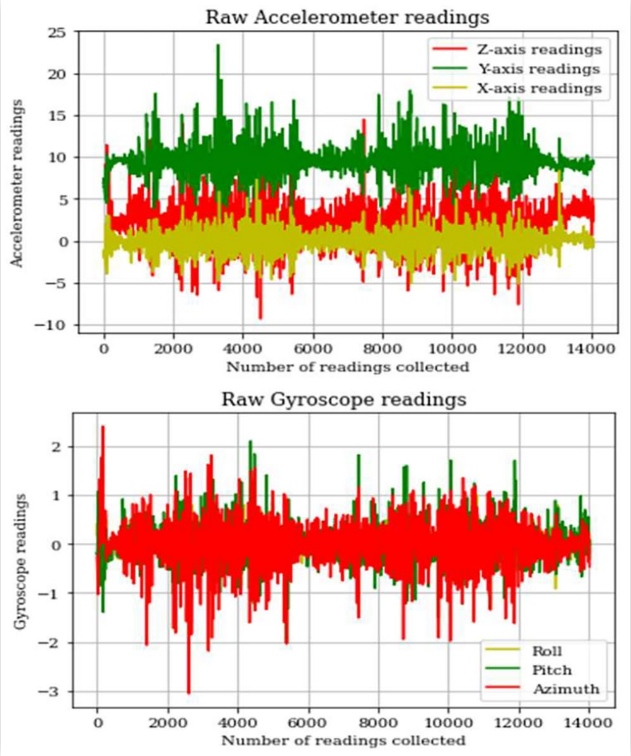
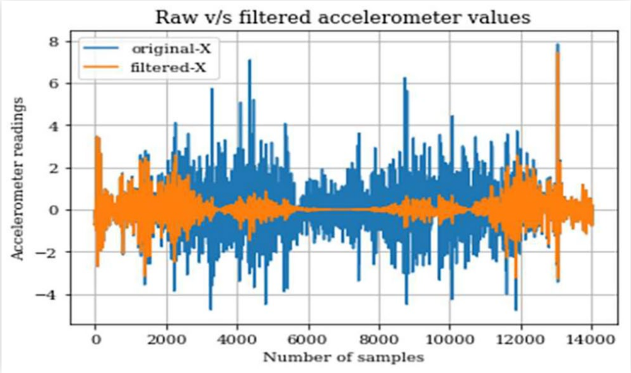


FIGURE 3 Raw Accelerometer readings compared to Filtered Accelerometer readings for one axes of Accelerometer readings. The same filtration strategy was applied for all 3 axes and also on the gyroscope readings.



Wang et al. have suggested a 1-D CNN-based model - FCN (Fully Convolution Networks). The basic block consists of a Conv1D layer followed by a BatchNormalization and ReLU (Rectified Linear Unit) activation function. There are 3 blocks of this in the proposed FCN network, followed by a GlobalPooling1D layer, and a Dropout layer. The model ends with a Dense layer with a SoftMax function, to get the per-class probabilities. The class with the maximum probability has been selected as the predicted class. The number of classes is kept as two (as defined before). Dropout layers have been added to counter the class imbalance, as the ratio of the road compared to potholes is largely on the side of the road, except in extremely bad road conditions.

3.2.1. Attention Model with CNN - Transformers

The next model explored has been proposed by Zhang et al. A 1D CNN backbone has been adopted that is going to serve as a feature extractor. After this, the output is going to be split into the number of classes we need to predict. These chunks are then fed into attention units like LSTMs and GRUs. The model ends with a couple of Dense layers with the final layer having the number of classes to predict (2 in our case) as the number of neurons. It has a SoftMax function to give per class probabilities. Finally, the class with the maximum probability is chosen as the predicted class. For recording results, LSTM and GRU have been used separately in the attention part of the model and recorded the results.

Transformers have also been studied for the prediction of potholes to classify pothole prediction. In the model explored with the Transformer, only the encoder was considered and the decoder has been disregarded as this is not a sequence-to-sequence translation task, the Dense layers after the Encoder perform the task of predicting the class outputs. Further, the encoder consists of Multi Headed Attention, which provides the attention mechanism in the encoder network.

Involution Neural Network (INN) was initially introduced [9] for image classification, as compared to Conv2d layers that were being used. It was a method to develop an operation that was both location-specific and channel-agnostic. This is achieved by generating each kernel based on the special location the kernel is currently covering. This is then reshaped and then multiplied with patches extracted from the images and then the final output is cast. In the proposed work, a 1D INN has been adopted for the prediction of sensor readings. The collected dataset consisting of raw sensor data has 6 channels corresponding to each of the axes of the sensor readings. First, the 1D INN-based Baseline model has been implemented, in which the Conv1D layers have been replaced with INN-1D layers. The same for the CNN backbone in the attention modules. Finally, an INN-based Transformer model has been proposed which will give greater accuracy as well.

FIGURE 4 The Attention Model with CNN backbone. The dotted line shows the interior of the CNN backbone. It consists of chaining Conv1D and MaxPooling1D layers before being split.

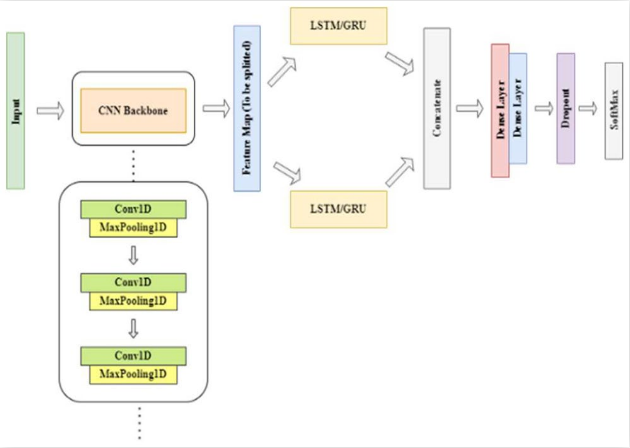
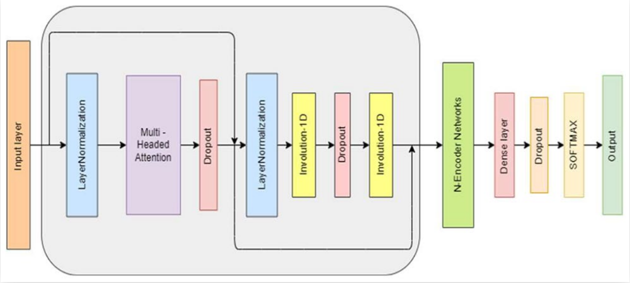


FIGURE 5 The INN-based Transformer network. We have used the Transformer encoder. The final layer gives the prediction outputs.



3.2.2. The INN-Former

In this network, the 1D-INN has been adopted in the encoder. Our encoder architecture consists of a LayerNormalization layer followed by MultiHeadedAttention and another LayerNormalization layer. The INN layer is followed by a Dropout and another Involution-1D layer. If need be (for improving accuracy), multiple such encoder layers are chained before passing them through a Dense layer and predicting the class outputs.

The Involution operation serves here as a method of extracting information in a channel-agnostic way. Conv-1D layers were channel-specific, in which regard, we hope to extract information that the kernel extracts in a spatial agnostic and channel-specific manner. It is to be noted here that the Encoder part can be stacked, i.e., we can use multiple encoders to improve accuracy if needed. The Multi- Headed Attention provides the attention mechanism, similar to the one provided in the previous attention networks. The input to the second LayerNormalization layer is the summation of the original Input and the output from the Dropout layer. Similarly, the output from the Transformer is the summation of the output of the Involution-1D layer and the input that goes into the LayerNormalization layer. The output then serves as an input to another Transformer encoder (if needed) or it passes through Dense layers, before another Dense layer with SoftMax activation to give the final output.

4. Results

The experiments have been done by comparing the proposed model with the same collected dataset after appropriate filtering and annotation. For experiment purposes, 70% of the filtered dataset has been chosen for the training corpus and 30% rest for the testing corpus. Following the experiments, the number of parameters used in each of the models has been recorded, and a detailed ablation study on each of the models and the parameter space covered has been done. Table 1 records the train and test accuracies as well as the test loss achieved by each model.

TABLE 1 Collection result of all the accuracy of the models which were tried on the dataset.

Model name	Train accuracy (%)	Test accuracy (%)	Test Loss
Baseline (FCN)	89.98	92.45	0.22
Attention - (Using CNN backbone) -LSTM	91.88	92.87	0.23
Attention - (Using CNN backbone) -GRU	91.68	92.88	0.25
Transformer - (Using CNN)	95.03	93.73	0.16
Baseline - Using INN	84.33	82.90	0.37
Attention (Using INN backbone)-LSTM	89.83	89.88	0.24
Attention (Using INN backbone)-GRU	94.80	94.15	0.12
INN-former (Proposed Model)	97.94	97.00	0.08

TABLE 2 The number of total parameters used by every model.

Model name	Number of total parameters
Baseline (FCN)	1,018
Attention - (Using CNN backbone) - LSTM	1,442
Attention - (Using CNN backbone) - GRU	1,442
Transformer - (Using CNN)	1,083
Baseline - Using INN	809
Attention with INN backbone-LSTM	1,001
Attention with INN backbone-GRU	1,001
INN-former (Proposed Model)	6,554 (Using dual encoder)

4.1. Analysis

From Table 1, it is clear that the accuracy of the proposed INN-former shows greater accuracy than the other models. The number of parameters and the hyperparameters used with each model has also been recorded. The hyperparameters have been tuned (using the Keras tuner [10]) to achieve the highest accuracy of each model.

The proposed model, INN with Transformer has a training accuracy of 97.94% and a test accuracy of 97.00%, and it easily beats the Baseline model (FCN network) by a good margin. The Attention with CNN backbone also does well compared to the Baseline model, but the test accuracy of Attention with CNN models (both the LSTM and GRU ones) are pretty close to the Baseline test accuracy. There is not much change in test accuracy between using GRU and LSTM for the attention network using the CNN backbone.

The vanilla Transformer encoder network shows good performance, compared to both the Baseline and Attention networks, beating both the train and test accuracies. The Baseline model with INN does not show immediate improvement over the first baseline model we used, which consisted of the Conv1D layers. This is because the INN is channel-agnostic, and does not capture the complexity between the channels, however, this model consists of lesser parameters which can be a big deciding factor considering deployment. Considering the INN backbone with the Attention model, both the GRU and LSTM show an improvement over the Baseline INN model. The Attention model with GRU and INN backbone shows greater accuracy compared to the Attention with CNN backbone-based models. Comparing Table 1 and Table 2 we can see that the INN-former model, though having more parameters, shows greater accuracy as well.

4.2. Model Parameters

INNs are also effective in reducing model parameters, which can be useful for deployment purposes. While deploying, the main goal is to make the model as light as possible and to get inference as quickly as could be. In Table 2, it can be observed that stacking transformer encoder layers (with INN and CNN backbone) lead to better accuracy. But with increasing layers and adding Multi Headed Attention, complexity also increases. From the conducted experiments, it has been found that increasing the size of Dense layers preceding the output layers can improve accuracy but it leads to extreme overfitting as well as it does not generalize well over new data. Due to this reason, methods to reduce complexity are preferred. INN show less complexity with increasing model architecture, even with our baseline model, we can see a difference between the FCN network and the adapted 1D-INN network. There is a slight increase in parameters while using INN with attention using GRU or LSTM but this is mainly due to the change in the size of the Dense layers. The INN-former shows greater parameters compared to the other models, but this is mainly due to stacking a greater number of encoder networks to increase the accuracy.

For the Baseline model and the CNN backbone used in the Attention models, experimentation has been done concerning both increasing and decreasing the number of Conv1D layers. It has been observed that with increasing layers, there is a danger of overfitting a lot, especially the more the number of such layers, the more the overfitting. The number of layers has been kept to a balanced number of 3, to avoid such problems. Further experimentation has been done with increasing the number of filters. Using a greater number

FIGURE 6 A. Train Accuracies of all the models over 100 epochs.

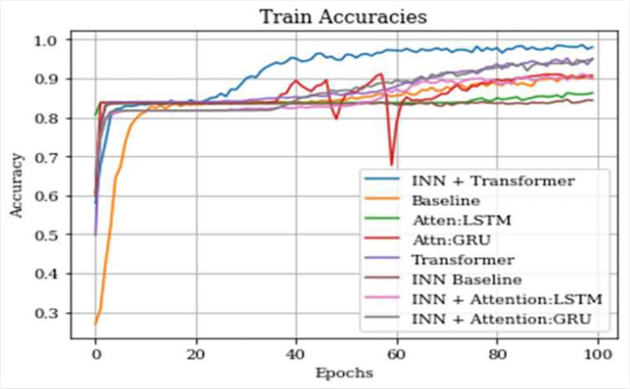
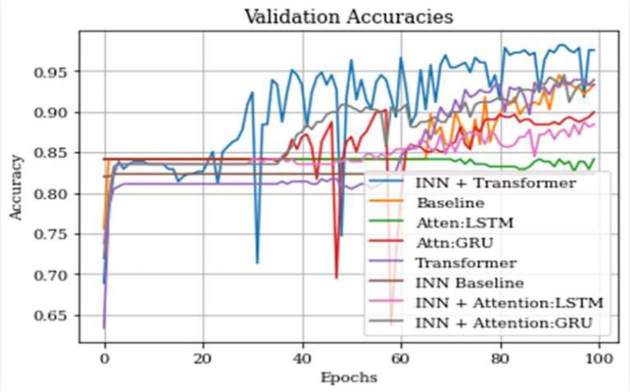


FIGURE 6 B. Validation Accuracies of all the models over 100 epochs.



of filters (around 64 or more) shows signs of overfitting, the training accuracy is much greater than the test accuracy. Increasing the kernel size allows the Conv1D to capture more information, but increases the complexity as well, which leads us to choose a tradeoff between the two.

In the case of using Involution neural networks, changing the kernel size and the stride leads to a change in the accuracy. Using a big kernel size and a big stride leads to information loss. Due to this reason, we have kept the number of kernel size 5 and the stride to 2. This balances the loss of information to overfitting. Increasing kernel size also leads to increasing parameter size, but it is compensated by increasing accuracy as well. In the case of using the transformer encoder layers, using 1 encoder layer for the CNN model and 2 encoders stacked for the INN model shows the desired results. Fig 6. shows the accuracies of train and validation for the models after we have tuned the hyperparameters.

It is to be noted that pothole occurrences might not be very high as compared to the rest of the road, leading to a class imbalance in the dataset, which might contribute to overfitting. Such overfitting has been tried to be reduced by tuning the hyperparameters, reducing the Dense layers, and adding Dropout layers. The probability of the Dropout layers has been kept as a hyperparameter, and after tuning and experimenting with various values, it has been found that the appropriate dropout probability is 0.2. Increasing this value leads to a loss of information and decreasing the value increases overfitting.

FIGURE 7 A. Training losses of all the models over 100 epochs.

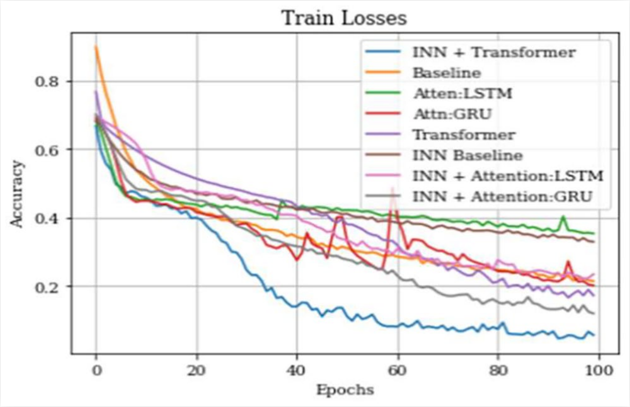
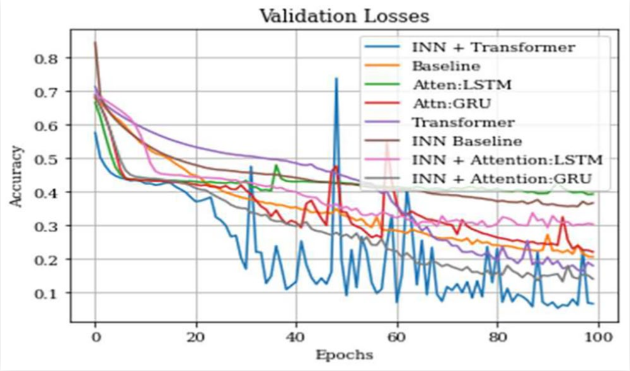


FIGURE 7 B. Validation losses of all the models over 100 epochs



5. Conclusion

In this work, we have proposed a method of using Involution Neural Networks along with Transformers for improving accuracy in detecting potholes in Indian Roads after fusing Accelerometer and Gyroscope sensor readings. Various models consist of non-attention-based models like the Baseline model (FCN) and attention-based models, in which we use CNN and INN-based backbones and record our observations. Furthermore, we have experimented with Transformer-based networks, where we have used the Transformer encoder (and using Multi headed attention) and using INN with Transformers. We have noted that our proposed model - the INN-former, shows a test accuracy of 97.00%. With this work, we hope to make the detection of potholes easier for ADAS systems so that it can navigate Indian Roads better.

In view of further work, information from vision-based sensors can be also fused along with the currently implemented IMU sensors. Along with this, GPS information fusion can also be considered for increased accuracy as well. One possible future extension to this proposed work would be deploying the model on edge devices and validating it in unstructured Indian road scenarios.

References

1. Ministry of Road Transport and Highways, Government of India, "Ministry of Road Transport & Highways, Government of India," Road Accidents in India, https://morth.nic.in/
2. Ahmed, K.R., "Smart Pothole Detection Using Deep Learning Based on Dilated Convolution," *Sensors* 21, no. 24 (2021): 8406, https://doi.org/10.3390/s21248406.

3. Bhatt, U., Mani, S., Xi, E., and Zico Kolter, J., "Intelligent Pothole Detection and Road Condition Assessment," arXiv.org, October 10, 2017, https://arxiv.org/abs/1710.02595
4. Wang, Z., Yan, W., and Oates, T., "Time Series Classification from Scratch with Deep Neural Networks: A Strong Baseline," [1611.06455v2] Time Series Classification from Scratch with Deep Neural Networks: A Strong Baseline, November 22, 2016, https://arxiv-export-lb.library.cornell.edu/abs/1611.06455v2
5. Zhang, J. et al., "Attention-Based Convolutional and Recurrent Neural Networks for Driving Behavior Recognition Using Smartphone Sensor Data," *IEEE Access* 7 (2019): 148031-148046, doi:10.1109/ACCESS.2019.2932434.
6. Pawar, K., Jagtap, S., and Bhoir, S., "Efficient Pothole Detection using Smartphone Sensors," *ITM Web of Conferences.* 32 (2020): 03013, doi:10.1051/itmconf/20203203013.
7. Nidamanuri, J. et al., "Auto-Alert: A Spatial and Temporal Architecture for Driving Assistance in Road Traffic Environments," *International Journal of Intelligent Transportation Systems Research* 20 (2022): 64-74, https://doi.org/10.1007/s13177-021-00272-3.
8. Nidamanuri, J., Mukherjee, P., Assfalg, R., and Venkataraman, H., "Dual-V-sense-Net (DVN): Multi-sensor Recommendation Engine for Distraction Analysis and Chaotic Driving Conditions," *IEEE Sensors Journal* (2022), doi:10.1109/JSEN.2022.3184983.
9. Li, D., Hu, J., Wang, C., Li, X. et al., "Involution: Inverting the Inherence of Convolution for Visual Recognition," arXiv.org, April 11, 2021, https://arxiv.org/abs/2103.06255.
10. O'Malley, T., Bursztein, E., Long, J., Chollet, F. et al., "KerasTuner," 2019, https://github.com/keras-team/keras-tuner

Contact Information

For further information, please contact the authors from the Smart Transportation Research Group (https://www.iiits.ac.in/research/research-groups/smart-transportation-group), IIIT Sri City:

Trisanu Bhar
trisanu.b19@iiits.in

Jaswanth N.
jaswanth.n@ihub-data.iiit.ac.in

Dr. Hrishikesh Venkataraman
hvraman@iiits.in

Acknowledgment

The authors acknowledge the support of IIIT Sri City and of different organizations such as the Department of Science and Technology (DST-ICPS), Navel Research Board (NRB), and IHUB Mobility Research.

Positions and opinions advanced in this work are those of the author(s) and not necessarily those of SAE International. Responsibility for the content of the work lies solely with the author(s).

ISSN 0148-7191