2022-09-25

Learning pavement surface condition ratings through visual cues using a deep learning classification approach.

Waqar Shahid Qureshi
Technological University Dublin, waqar.qureshi@tudublin.ie

David Power
Pavement Management Services

Joseph McHale
Pavement Management Services

See next page for additional authors

Recommended Citation

Waqar Qureshi, David Power, Joseph McHale, Brian Mulry, Keiran Feighan, Dympna O Sullivan. Learning pavement surface condition ratings through visual cues using a deep learning classification approach. 18th IEEE International Conference on Intelligent Computer Communication and Processing, ICCP 2022, Cluj-Napoca, Romania September 22-24, 2022. DOI: 10.21427/msqw-ck18

Funder: Enterprise Ireland

Learning pavement surface condition ratings through visual cues using a deep learning classification approach

Waqar S. Qureshi
Department of Computer Science
Technological University
Dublin, Ireland
waqar.qureshi@tudublin.ie

David Power
Pavement Management Services
Dublin, Ireland
davidpower@pms.ie

Joseph McHale
Pavement Management Services
Dublin, Ireland
joemchale@pms.ie

Brian Mulry
Pavement Management Services
Dublin, Ireland
brianmulry@pms.ie

Kieran Feighan
Pavement Management Services
Dublin, Ireland
kfeighan@pms.ie

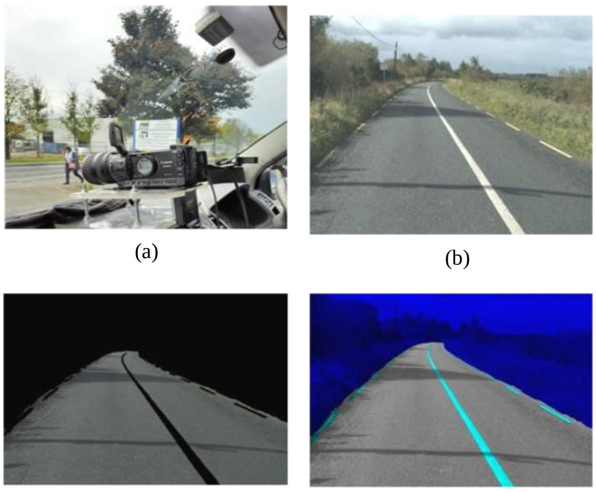
Dympna O Sullivan
Department of Computer Science
Technological University
Dublin, Ireland
dympna.osullivan@tudublin.ie

***Abstract*— Pavement surface condition rating is an essential part of road infrastructure maintenance and asset management, and it is performed manually by the data analyst. The manual rating requires cognitive skills built through training and experience, which is quantitatively challenging and time-consuming. This paper first analyses the complexity of the current manual visual rating system. This paper then investigates the suitability and robustness of a state-of-the-art convolutional neural network (CNN) classifier to automate the pavement surface condition index (PSCI) system used to rate pavement surfaces in Ireland. The dataset contains 3735 images of flexible asphalt pavements from Irish urban and rural environments taken from a video camera mounted in front of a van. The PSCI ratings were applied by experts using a scale of 1-10 to indicate surface conditions. The classification models are evaluated for different input pre-processing variations, image size, learning techniques, and the number of classes. Using 10 PSCI classes, the best classifier achieved a precision of 57% and a recall of 58%. Adjacent combination of classes (e.g., ratings 1 and 2 combined into a single class) to form a 5-class problem produced a classifier with a precision of 70% and recall of 77%. Given the complexity of the problem, classification using CNN holds promise as a first step towards an automated ranking system.**

Keywords—classification, pavement surface condition rating, PSCI rating, pavement distresses

I. INTRODUCTION

Pavement or road surface condition ratings are essential to road or pavement management systems in many countries, including the United States of America, the United Kingdom, China, Brazil, Taiwan, and Japan[1]. Pavement surface ratings and other measurements and information can help in future planning of budgets and maintenance priorities. The road or pavement consists of a subgrade layer at the bottom, then a foundation layer that consists of sub formation and capping. After the foundation layer, there is a subbase layer and then the base layer. After the base layer, there may be a binder course and surface course that makes the pavement surfacing. Visual distresses appear on the surface course; however, they may indicate a fault in the base or foundation layer (such as alligator cracking or rutting). According to the Transportation Information Centre at the University of WISCONSIN [2], pavement or road surface can be categorised into four general types, i.e., asphalt, concrete, gravel, and brick and block. Asphalt, also known as flexible pavement, is widely used to construct national, regional, or local roads across the road network of a country and has different sub-categories depending on its construction. Different countries use different standards for pavement ratings; a summary table is mentioned in [3]. The Irish Department of Transport (DoT) [4] and Road Management Office (RMO) have their standard known as the pavement surface condition index (PSCI), with associated manuals for carrying out pavement ratings. It is derived from the Pavement Surface Evaluation and Rating (PASER) [5] system to cater to local pavement distresses and make visual data collection easier for local authorities. The Irish PSCI [6] rating is on a scale from 1-10, similar to PASER, where index-1 is the lowest (surface wholly worn out or failed), and index-10 (no distress, new pavement) is the highest. It has three volumes to cover urban flexible roads, urban concrete roads, and rural flexible roads [7], [8]. The manual focuses on visual inspection through the naked eye in the field or from rating in the office forward view images recorded using a high-resolution video camera.



(a) (b) (c) (d)

Figure-1: A typical camera position, captured image, and the image after different pre-processing steps. (a) is the picture of a typical camera position mounted on the video van. (b) the output of the camera with a rating of 10. (c) In the 3-channel processed image, the first channel is a segmented road, the second channel is road plus marking, the last channel is the original intensity image, and (d) is the RGB segmented pavement image.

CareerFit Plus fellowship MF. 2021 0273 funded by EI and the European Union's Horizon 2020 Research and Innovation Programme under the Marie Sokolowski-Curie Co- Grant agreement No: 847402.

The PSCI rating in Ireland is done either directly by local authority staff or sub-contracted to private companies. Images are recorded with the video camera mounted on the front dashboard of a vehicle with a computer and GPS (Global Positioning System) sensor (see Figure-1(a)). Images are captured every approximately 5 meters, and the images are similar to those captured by an autonomous car (see Figure-1 (b). In practice, ratings are given to continuous stretches of roads with a similar condition, with 200 meters being the minimum length to have its' distinct rating. The main distresses found in flexible pavements are ravelling, bleeding, transverse and longitudinal cracks, alligator cracks, potholes, rutting, patching and surface breakup. When rating images captured by a video camera, a PSCI rating is given by a data analyst offline viewing images at a computer. The rating expert assigns a rating to the first 200 meters (~40 images) and then will adjust the rating as the pavement condition noticeably changes. The visual rating of pavements is a two-step distress identification followed by an estimate of the amount of distress in the image. The manual task is tedious, subjective, and prone to errors. As such, the overall procedure then requires a quality control loop through an experienced rating labeller.

An automated rating system may be able to ensure more consistent and accurate ratings and as well as reduce the overall time required. It can be implemented using machine learning as object detection, segmentation, or classification system. This paper analyses the automating pavement rating as a classification problem using a state-of-the-art CNN-based classifier (EfficientNet V2) with real-world images. We compare the output with a baseline classifier (Inception V3) and test the classifier with different augmentations (pre-processing) of the input images. Section III explains the image capture process and describes the classification problem methodology, including the evaluation metrics. Section-IV describes the results, Section-V discusses our observations, and we conclude in Section VI.

II. Related Work

A guide on data collection for pavement quality management is presented in [9]. The guideline presents standard procedures and practices to obtain data for pavement quality assessment and management. For visual pavement condition assessment through images, either of two views is recommended, i.e., a front-mounted camera placed orthogonal to pavement surface normal or a back-mounted camera placed inline to pavement surface normal. Different parameters are used to gauge the quality of pavements for service management and maintenance records [10]. Pavement Condition Index, as detailed in [5], [6], [11], is more focused on the visual estimate of pavement distress. International Roughness Index (IRI) measures pavement roughness [12] through vehicle vibration, and Rut Depth [10] measures the transverse deflection for computing the pavement quality index. In [13], the author measures the suitability of a major heavy-traffic road in Yemen for moving traffic loads using pavement evaluation rating PCI [11].

Deep learning architectures have recently been applied to pavement condition detection and classification [14]–[31]. These methods can be segregated into pavement condition rating through classification, pavement distress object detection, and segmentation approaches to pavement cracking. Researchers in [25] and [32] have used aerial images through drones as input and presented a convolutional neural network architecture for automated pavement distress detection and evaluation, respectively. In [23], an automated smartphone-based application is proposed to detect potholes and cracks, accelerometer, global positioning system (GPS) sensor, and compass are used to record the location of the potholes. Recall precision and accuracy are reported for eight distresses, with the lowest recall recorded as 5% for lateral linear cracks, 65% for alligator cracking, and the highest for crosswalk blur and white line blur at 95%.

Authors in [16], [33], [34] also used a smartphone to capture images from multiple countries and develop distress detectors, based on CNN, for alligator cracks, longitudinal cracks, transverse cracks, and potholes. The distress objects are similar to [23], and the measures reported for the three countries are F1-Score and mean average precision. Pavements in different countries have different F1-Score with a maximum score F1-score of 52% for alligator cracking and a minimum F1 score of 29% reported for linear transverse cracking for Japanese roads.

In [35], a CNN-based crack segmentation method is presented that consists of the novel architecture of five layers; the input layer is a line feature detector filter, followed by two convolutional layers and two fully connected layers to segment crack pixels in the 3D images of asphalt surfaces. The evaluation reported precision, recall and F1-score with an F1 score of 88%. This method is specifically for 3D data from the PaveVision3D laser system, which is mounted on a video van, viewing an orthogonal top view of the road. The maintenance and capital of sensing technologies used in the experiment are much more expensive than a camera mounted on the front of the vehicle. In [20], the authors presented a hybrid model of an object detector and segmentation for classifying and quantifying distress severity on pavements and predicted PASER indices for each patch. The images are collected from Google Street View maps, 70-degree wide-angle views and 90-degree birds-eye view images. Wide-view images are used for distress detection and birds-eye-view images to quantify crack severity. The results from the hybrid model are then fed to a linear and weighted regressor for predicting PASER indices to road patches. They trained YOLO to classify nine road defects. The U-Net, which is based on a fully convolutional layer, is used to segment road cracks by quantifying the density of pixel labels as cracks. The results from the two models are then combined to find the density of crack per road pavement defect. The results are then fed to a linear and a weightage regressor to label each image a PASER index. The images used are from U.S pavements, and the PASER calibration set is very small. The road condition, distress condition, and camera views differ from the current practices in Ireland. The predicted PASER model fits with a R^2 of 0.9382 or test data with a root mean square error of 10.45.

The authors in [21] have presented a pavement type and quality classification technique. The dataset used for the experiments is RTK [21], caRINE [36], and KITTI [37]. It classifies roads into three different pavement types and three different ratings. The images are first cropped to focus on the region of interest that contains the roads. Data augmentation is done to increase the robustness and avoid overfitting. The authors used three convolutional layers, a flattening layer, and then two fully connected dense layers to classify the road types into asphalts, paved, and unpaved. The classified images are then further passed through another classifier to estimate the quality of each road, as good, regular, and bad for each class. The surface type accuracy is 98% for three types, and the classification accuracy for the three quality types is 98% for good asphalt and 96% for bad asphalt [21]. The precision of classifying the good class is 86.7%, while the precision of classifying the bad asphalt class is 81%. The number of rating indices are only three, i.e., Good, Bad, and Regular, and they do not relate to the existing standard rating system.

Road pavement condition rating depends on the type of distress and its quantification, which changes (shape, size, and texture) with different factors, mainly environmental conditions and the pavement construction process. The environmental conditions and construction process changes with geographical locations. The variation in data in different regions is not only because of changes in shape, size, and texture of the distresses but also due to different imaging technology and sensor placement in the video van. The environmental condition and economic factors govern these variations and constraints on the use of imaging technologies. Therefore, smartly chosen training images are required to develop a deep learning-based solution, which can account for application domain constraints (environmental condition, construction processing, imaging technology, and sensor placement in the van) in an automated road rating system.

On the commercial side, a few companies in the U.S and Japan do provide automated solutions for pavement condition ratings. RoadBotics [38], working locally for U.S roads, use a limited version of PASER[5], i.e., they rate pavement from 1 to 5, with 5 being the lowest rating. An automated rating system from Ricoh [39] estimated the amount and location of cracks on a 50cm x 50cm patch and has adopted its own rating system for Japanese roads based on PASER.

A comprehensive analysis is essential to evaluate the classification-based convolutional deep architectures for automating the PSCI rating system. The first contribution of this research is to analyse the complexity of the current manual visual PSCI rating process. The second is to analyse the suitability, effectiveness, and robustness of state-of-the-art image-classification deep neural networks for automated rating across different input variations for rating a pavement from 1 to 10. The image-set for evaluation is of flexible asphalt pavements in urban and rural environments, from different Irish roads across the country, using a high-resolution camera mounted in front of a van. The images have been supplied by PMS (Pavement Management Services Ltd.), an Irish civil engineering company specialising in pavement evaluation with whom we are collaborating on this research. A state-of-the-art classification architecture EfficientNet V2 with different variants was evaluated along with an older Inception V3 for deep-features extractions. A dense neural network layer is added to the architecture to classify images into 1-10 PSCI ratings. The classification models were evaluated for different input pre-processing variations, image size, learning techniques, and classes.

III. Material and Methods

At first, we present how the images were captured and labelled for supervised learning. Then the pre-processing of images is explained. Then we present the deep neural network used for classification and the evaluation criteria used for quantitative analysis.

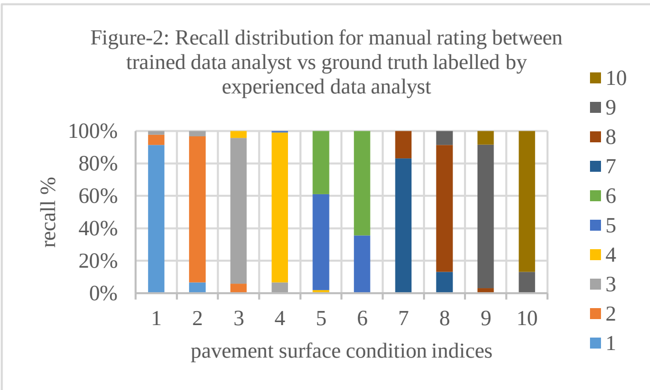
A. *Dataset and Labelling*

Images of flexible asphalt pavements from urban and rural environments across Ireland are acquired using a camera mounted on the front dashboard of a video van (see Figure-1(a)). The camera is attached to a server for recording images. A remote laptop accesses the server over the network to label each image stretch of the pavement—the server linked to the camera capture image every five meters. The size of each image is 720x576, with three channels (red, green, and blue). The images are labelled offline using a PSCI [6] scale of 1-10 by two data analysts; a data analyst (DA) and an experienced DA. For our experiments, the images are divided into classes 1-10 according to the PSCI ratings. Table 1 shows the total number of images for each class after removing images with moving wipers, low intensity, low contrast, and visually too blurry.

TABLE I. No. of Images in each class used for the experiment

class	Original images	Train (70%)	Test (30%)
1	472	330	142
2	391	273	118
3	344	240	104
4	310	217	93
5	296	207	89
6	191	133	58
7	360	251	109
8	534	373	161
9	353	247	106
10	484	229	146
Total	3735	2609	1126

As can be seen from Table 1, the classes are imbalanced; however, this is representative of a real-world dataset of images taken from Irish roads. A total of 3537 images were left for the experiment, and the dataset was divided randomly into training (70%) and test (30%) sets. Figure 2 shows the manual recall distribution between the trained and experienced DA (taken as ground truth). The overlapping between the adjacent recalls is evident, especially for the



central indices five, six, and seven. It is observed that extreme rating indices, i.e., one and ten are easier to distinguish. In contrast, adjacent ratings are harder visually—the difference in labelling between the two D.A.s highlights the complexity of the problem. The images in the dataset are not frames from a continuous video; they were selected from different stretches of Ireland's local and regional roads.

B. *Image Preprocessing*

To generalise the model's performance and remove the background biasness, we performed some pre-processing steps on the dataset similar to the concept explained in [40]. At first, we used pixel segmentation using the semantic segmentation CNN-based model from [41] to extract roads, marks, and background pixels. The mean accuracy of the model is 90.1%, with 97.4% accuracy for pavement. The masks are used for 3-channel segmented pavement images. Another 3-channel image we call an "augmented" image is computing by combining the pavement segmented intensity image, the pavement plus mark pixel intensity image, and the original intensity image. Image height is cropped 250 pixels from the top and 50 pixels from the bottom to remove the sky and pavement pixels further away from the camera and pavement pixels too close to the camera. Thus, there are three sets of training and test images; the original captured image set, named "original;" the pavement pixel segmented set, named "segmented;" three-channel "augmented" set. Figure-1 (b)-(d) shows the different 3-channel images used for training. Four different variants of training sets are assembled for evaluations (see Table-II) augmented set (2609 images), original plus segmented set (5218 images), original set (2609 images), segmented set (2609 images), original plus segmented plus augmented set (10436) images. After cropping, the image size is 700 x 330, which is further resized to 512 x 512, 480 x 480, and 384 x 384 as required by different model variants used as the baseline.

C. *Convolutional Neural Network Architecture*

Graphical abstract of the methodology used in our proposed automated rating system is shown in Figure 3. We selected EfficientNet V2 as a base model for computing deep features (1 x 1280). We added a fully connected dense layer (with l2 regularisation and softmax activation) and a dropout layer (with a 0.5 drop rate) for the classification. EfficientNet V2 is one of the state-of-the-art models [42] from Google Inc., with a classification accuracy of 90.2% on ImageNet-22K with 21841 classes. This makes it a perfect candidate for feature extraction for our image rating problem. We evaluated three different versions of EfficientNet V2, i.e., small, medium, and extra-large, with images of 384x384, 480x480, and 512x 512, respectively. The specific image size is chosen because the TensorFlow hub models are trained using these specific sizes as input. We used TensorFlow hub weights and biases to compute features from our images and then added a dense classification layer head to classify into different pavement condition ratings. The EfficientNet V2 small, medium and x-large weights and biases used in the experiment were trained on ImageNet 22K and ImageNet 1K (1000 classes). For the sake of comparison, we also used Inception V3 architecture to compute deep features for a classifier that can take an input image size of 299 x 299; the weights and biases were taken from the TensorFlow hub, which was trained on ImageNet 1K (1000 classes). The Inception V3 feature extractor follows the same dense classification layer.

Two types of learning methods are used - transfer learning

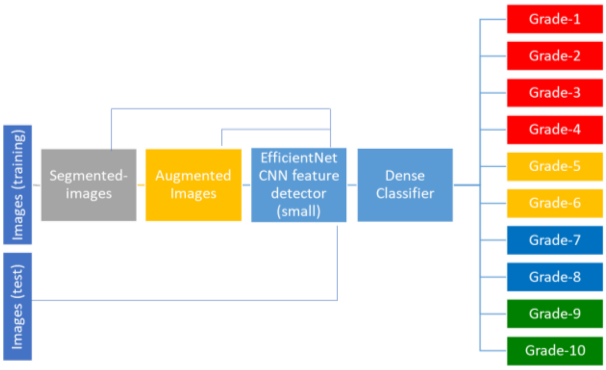


FIGURE-3: Block diagram of our evaluation model. The EfficientNet Deep Feature extraction is followed by a fully connected Dense layer

and fine-tuning. To determine the best model size of EfficientNet V2, we used an augmented set with transfer learning to train the model. Then, we used the overall accuracy of the augmented test set as a parameter to gauge model size selection. The best model size of EfficientNet V2 is then used to further evaluate the classifier by fine-tuning the hyperparameters on different training sets, decreasing the number of classes. For training, we use data augmentation techniques, including random vertical flip, random horizontal flip, random zoom up to 20%, and random contrast between 10% and 20%. A normalisation layer is added as TensorFlow hub models do not contain a built-in normalisation. A stochastic gradient descent optimiser with a sparse categorical cross-entropy loss function was used for model training. A learning rate of 0.0001 was set for all of our experiments with a batch size of 8, each training was repeated for 300 epochs: a total of 20,190,298 trainable parameters, and 153,872 non-trainable parameters.

EfficientNet V2 deep features are also compared to another deep feature extraction architecture, i.e., Inception V3. The ratings 1-10 are combined with their natural adjacent class to create five classes (1-5). This is based on expert knowledge about the PSCI acquired at PMS. Indices 1 and 2 are combined into one class as they usually have no surface layer or only traces of a surface layer. Indices 3 and 4 are combined as they only have potholes, alligator cracks, or deteriorated patches. Indices 5 and 6 are combined as they have only linear cracks or neat patches. Indices 7 and 8 are combined as they do not have potholes, cracks, or patches and only have ravelling and bleeding. Indices 9 and 10 are combined as they have minor or no ravelling or other distress. The models were retrained on the fused dataset for 5 classes instead of fusing the results for 5 classes.

Details of different parameter choices and evaluation criteria are given in Table II. The last two columns are the code-name followed by the given model's name for each variant of EfficientNet V2 or Inception V3.

D. *Evaluation Criteria*

The evaluation criteria are precision, recall and F1-score per class, as well as precision, recall and F1-score across all classes computed using the equations below:

We also plot the percentage prediction of each class with respect to the ground truth of each class.

$$precision = \frac{True\ positive}{True\ positive + False\ positive} \quad 1$$

$$recall = \frac{true\ positive}{total\ Images\ in\ \ class} \quad 2$$

$$F1\_score = 2 * \frac{precision * recall}{precision + recall} \quad 3$$

model uses a smaller image size and is less computationally complex, the inference will be faster than other models. Therefore, we choose this model to further evaluate for other variations in the input and number of classes.

Table-III compares classification models 'c' and 'd', based on EfficientNet V2 small architecture, which is trained using transfer learning, i.e., the ImageNet biases and weights are frozen, and only the dense classification layer is trained. Model 'c' uses the augmented dataset, and model 'd' uses the original plus segmented dataset as training. An augmented dataset test-set is used to test model 'c', and an original dataset test-set is used for model 'd'. Model 'c' shows a superior F1 score (0.46) on average.

TABLE-II: The first seven columns are the model parameters variation, and the last two column is the name given to each model and the code name

Dataset	Size	Learn	Model	Class	Evaluation	Training parameters	code	model-name
Augmented		Transfer	EfficientNet (small) (medium) (x-large)	10		Learning rate = 0.001 Optimizer = SGD Momentum=0.9, regularization l2=.0001 Activation=SoftMax Loss function = Sparce Cross Entropy Epoch=100 Dropout=0.5	a	EN-M-480-T-AUG
	480						b	EN-XL-512-T-AUG
	384						c	EN-S-384-T-AUG
	512						d	EN-S-384-T-ORIG+SEG
							e	EN-S-384-FT-ORIG+SEG
Original + Segmented	384	Transfer	EfficientNet (small)	10	recall, precision F1-Score	Learning rate = 0.001 Optimizer = SGD Momentum=0.9, regularization l2=.0001 Activation=SoftMax Loss function = Sparce Cross Entropy Epoch=200 Dropout=0.5	f	EN-S-384-FT-AUG
		Fine tune		5			g	EN-S-384-FT-ORIG+SEG-No-W
		Fine tune	Inception V3	10			h	EN-S-384-FT-ORIG+SEG-5
				5			i	EN-S-384-FT-SEG
Segmented	384	Fine tune	EfficientNet (small)	10			j	EN-S-384-FT-ORIG
Original	384	Fine tune	EfficientNet (small)	10			k	IN-V3-299-FT-ORIG+SEG
							l	IN-V3-299-FT-ORIG+SEG-5
Augmented + Segmented + Original	384	Fine tune	EfficientNet (small)	10			m	EN-S-384-FT-ALL
				5			n	EN-S-384-FT-ALL-5

IV. Result and Discussion

Table III summarises all the results from different model variations as listed in Table II. The first row (see Table-III) describes the test set used to evaluate the model in the second row. The second row lists the model code, while the rest list the individual F1 score for each class, the average precision, recall, and F1 score. The first three columns of Table III show the results of average recall, precision and F-1 score for models' a', 'b', and 'c', trained using transfer learning on an augmented set for different architecture sizes and input image sizes. An EfficientNet with an input image size of 384x384 has the best performance, with an F1 score of 0.46. As the

Table III also shows the F-1 score of the EfficientNet V2-based classifier for models 'e', model 'f', model i', and model 'j', which are trained using fine-tuning, i.e., the biases and weighted are fine-tuned along with added classification layer. Model 'e' is trained on the original plus segmented set, model 'f' on the augmented set, and model 'i' and model 'j' on only the original and segmented set. The results show slightly higher performance than transfer learning models (average F1-score: 0.53). The result also shows that augmenting the original set with a segmented set performs better than using only one dataset. Also, the augmented dataset achieves a similar performance but a lower F1 score for less

Figure-4: EfficientNet inference F1 Score for different pre-processing to test image, blue is augmented, orange is original, and gray is segmented.

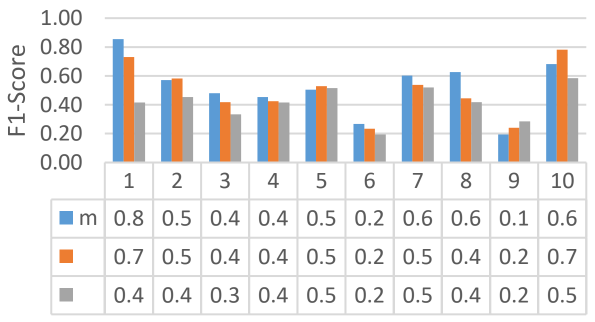
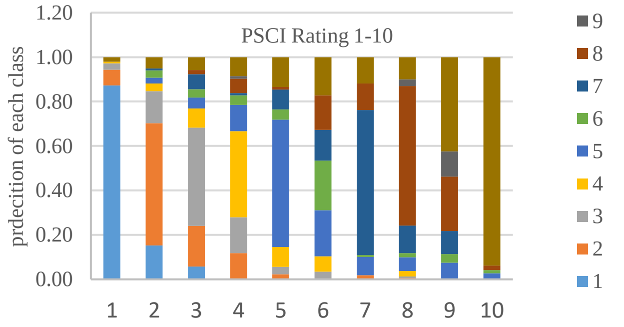


Figure-5: Prediction percentage of predicted class to ground truth for classifier trained on all input dataset with EfficientNet V2 deep features



discriminative classes 5-6, and 9-10. Table III shows the results of model ‘g’ (average F1-score:0.50 on the original test set), which is trained without using weights, for unbiased classes. The model does not perform as well as those that use weights while training for unbiased classes. Table-III highlights the F1-score comparison between model ‘k’ (average F1-Score: 0.50) that used inception v3 as deep features for classification vs EfficientNet V2 small size (average F1-score: 0.53), has better overall and in-class performance.

In summary, the model trained using all three datasets has higher average precision, recall, and F1-score, (see Figure-4), i.e., 57% precision with 58% recall and 0.57 F1-score. The model was tested on all test sets, i.e., original, augmented, and segmented test images.

Figure 5 shows the prediction percentage of each class to the ground truth class. The prediction shown is of the best classifier, i.e., model ‘m’, that is trained using all the three datasets on EfficientNet V2 small as a feature detector. The figure clearly shows a reasonable prediction rate for class-1 and class-10, even though images in both class 1 and class 10 have different backgrounds. Another observation is that many of the images rated as 9 are predicted as 10, which is also evident from manual labelling comparison among labellers. Similarly, some images labelled as 8 are predicted as 9 and vice versa. Classes from 10-7 are ratings with only ravelling and bleeding due to chip loss and have visible features that may be affected by intensity variation due to light reflectance. Classes 5-6 rated images only have neat patches and linear cracks; linear cracks have a very small change in intensity around the affected region and might be a good candidate for pixel-level classification compared to classification. Images rated 3 to 4 have either localised alligator cracking, potholes, or deteriorated patches, making them more distinguishable to other cracks. However, it is difficult to segregate between 3 and 4 as the quantity of the distress distinguishes between the two. Images rated between 1 and 2 usually do not have any pavement surface left and are easy to classify from others. One reason for poor performance on class-2 is that images in 2 might have no surface but a good intact base-layer surface.

Figure-6: Efficient Net vs Inception Model F1 Score bar for 5-class classifier. Blue is EfficientNet inference result for orginal image and orange are result of inception for original images

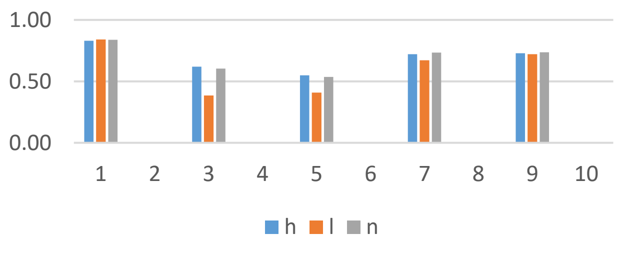


Figure-6 highlights the fact that if we merge classes 1 & 2, 3 & 4, 5 & 6, 7 & 8, and 9 & 10, then the overall precision reaches 70% with a recall of 77% (the model trained on all three training dataset as compared to a model trained on Inception V3 and Efficient V2 trained on the only original and segmented training set). Therefore it augments our observations above and a natural selection for combining ratings for an automated classification-based rating system. The current experiment has highlighted the difficulty in manual labeling of images. Deep learning features used for image classification are generally good in computing global features such as texture, shape, and color. In the case of pavement rating classification, much of the texture and color information is background and clutter. The image-wise result of the EfficientNet V2 small model ‘m’ (the best model) can be seen at [43]; a sample of true-positive for indices 1-9 is also shown in Figure-6. In summary, rating classification is suitable for classes 1-2, 7,8,9, and 10, where the distress is not localised and stressed across the whole road segment. Distresses that are localised in nature, such as patches and potholes, are more suitable to be detected using object detection, while cracks are more suitable to be identified using pixel level classification.

Table-III: Model comparison across different input training set, different set of classes, and different training parameters. Small letters represent models code as given in Table-II.

Test	aug	aug	aug	orig	seg	orig	seg	aug	orig	seg	orig	seg	seg	orig	orig	seg	orig	seg	aug	orig	seg	aug	orig	seg
S.No.	a	b	c	d		e		f	g		h		i	j	k		l		m			n		
1	0.81	0.78	0.81	0.78	0.78	0.77	0.78	0.81	0.53	0.77	0.83	0.87	0.79	0.78	0.82	0.79	0.84	0.87	0.86	0.73	0.42	0.84	0.79	0.57
2	0.55	0.52	0.54	0.48	0.42	0.60	0.53	0.55	0.70	0.61			0.63	0.49	0.65	0.60			0.57	0.58	0.45			
3	0.14	0.38	0.40	0.31	0.41	0.50	0.49	0.46	0.40	0.50	0.62	0.62	0.39	0.51	0.30	0.42	0.38	0.45	0.48	0.42	0.33	0.60	0.59	0.52
4	0.20	0.17	0.34	0.34	0.30	0.41	0.43	0.28	0.48	0.30			0.29	0.38	0.29	0.33			0.45	0.42	0.42			
5	0.33	0.35	0.37	0.24	0.29	0.50	0.43	0.32	0.47	0.46	0.55	0.50	0.45	0.39	0.37	0.50	0.41	0.49	0.50	0.53	0.52	0.54	0.48	0.43
6	0.18	0.03	0.14	0.13	0.26	0.20	0.26	0.12	0.20	0.15			0.27	0.22	0.18	0.18			0.27	0.23	0.20			
7	0.39	0.05	0.27	0.43	0.53	0.44	0.52	0.28	0.40	0.44	0.72	0.71	0.11	0.48	0.51	0.48	0.67	0.54	0.60	0.54	0.52	0.73	0.71	0.69
8	0.23	0.43	0.34	0.41	0.26	0.41	0.50	0.54	0.28	0.44			0.47	0.49	0.43	0.25			0.63	0.45	0.42			
9	0.32	0.13	0.30	0.18	0.18	0.27	0.33	0.02	0.08	0.17	0.73	0.74	0.14	0.27	0.25	0.26	0.72	0.66	0.19	0.24	0.28	0.74	0.72	0.64
10	0.38	0.61	0.58	0.64	0.61	0.70	0.72	0.69	0.76	0.77			0.72	0.68	0.67	0.66			0.68	0.78	0.58			
precsion	**0.40**	**0.41**	**0.46**	**0.43**	**0.44**	**0.50**	**0.52**	**0.56**	**0.52**	**0.55**	**0.70**	**0.70**	**0.46**	**0.50**	**0.50**	**0.50**	**0.65**	**0.66**	**0.57**	**0.52**	**0.48**	**0.70**	**0.67**	**0.60**
recall	**0.44**	**0.45**	**0.46**	**0.45**	**0.45**	**0.51**	**0.54**	**0.51**	**0.48**	**0.52**	**0.69**	**0.68**	**0.50**	**0.52**	**0.52**	**0.50**	**0.62**	**0.61**	**0.58**	**0.53**	**0.44**	**0.77**	**0.66**	**0.57**
F1-Score	**0.42**	**0.43**	**0.46**	**0.44**	**0.44**	**0.50**	**0.53**	**0.53**	**0.50**	**0.53**	**0.70**	**0.69**	**0.48**	**0.51**	**0.51**	**0.50**	**0.63**	**0.63**	**0.57**	**0.52**	**0.46**	**0.73**	**0.66**	**0.58**

CONCLUSTION

Automated pavement rating using standard pavement surface condition indices with acceptable accuracy and robustness is still challenging despite using state-of-the-art convolutional network architectures showing promise on the benchmarks dataset. The research literature and commercial products promise to solve the problem; however, they are limited to non-standardised rating approaches and support limited distress detections. This paper investigates an automated approach to predicting PSCI ratings using a classification approach on real-world images. The first contribution of this research is to analyse the complexity of the current manual visual PSCI rating task. It is observed that extreme rating indices, i.e., one and ten are easier to distinguish.

In contrast, adjacent ratings are harder visually—the difference in labeling between the two D.A.s highlights the complexity of the problem. The second is to analyse a state-of-the-art deep neural network's suitability, effectiveness, and robustness for automated rating across input variations. The images for evaluation are of flexible asphalt pavements with urban and rural environments, from other Irish roads across the country, from a camera mounted on the front of a van as practiced locally. The dataset contains 3735 images, manually rated by the PSCI rating expert, with a natural imbalance between classes. The EfficientNet V2 and Inception V3 architectures were evaluated for different input pre-processing variations, image size, learning techniques, and classes. Overall precision, recall, and per class F1-score are reported for quantitative evaluation. The best model for the 10-class PSCI rating achieved an overall recall of 58% with a precision of 57%, while for 5-classes, a recall of 77% with a precision of 70% is reported. The 10-class classifier classified class-1 with a F1 score of 86% and class 7,8,10 with 60%, 63%, and 68%, respectively. Figure -7 shows qualitative results of rating indices 1-9 for a sample of an image of each class.

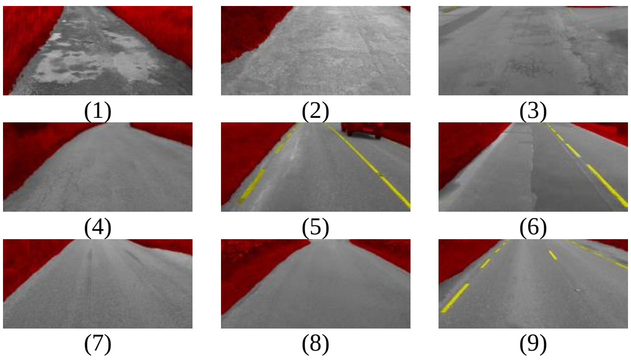


Figure-7: Images from classifier where the automated rating matches the ground-truth labels from augmented image test set for rating 1-9.

In summary, rating classification is suitable for classes 1-2, 7,8,9, and 10, where the distress is not localised and stressed across the whole road segment. The simple classification approach is affected by the cluttered background around the pavement; however, a segmented pavement may lead the CNN to learn the global feature such as edges and line shapes. An augmented image resulted in a better performance in most of the classes. A possible way forward is to evaluate the performance of the best model for predicting the rating over a 200-meter stretch of the road instead of using individual images and selecting the most common rating over that stretch as a result. In future, a hybrid approach to distress detection and quantification can also be evaluated; for example, distresses that are localised in nature, such as patches and potholes, are more suitable to be detected using the object detection approach, while cracks like alligator cracking, transverse cracks, and longitudinal cracks are more suited to be identified using pixel-level segmentation. In contrast, Gabor filters with a machine learning classifier can be used to explore surface defects such as ravelling and bleeding.

ACKNOWLEDGEMENT

W. S. Q would like to thank Justin Whyte for taking the time to rate each image as an expert and discuss his views on the rating given to each image during labelling. Waqar S. Qureshi likes to thank Enterprise Ireland (E.I.) for providing a research grant under Marie Curie Career Fit Plus Fellowship. No. M.F. 2021 0273.

REFERENCES

[1] N. S. P. Peraka and K. P. Biligiri, "Pavement asset management systems and technologies: A review," *Autom. Constr.*, vol. 119, p. 103336, Nov. 2020, doi: 10.1016/J.AUTCON.2020.103336.

[2] "Road pavement surface types," Feb. 03, 2022. https://interpro.wisc.edu/tic/?csis-search-options=site-search&s=paser&submit=Search (accessed Feb. 03, 2022).

[3] W. Cao, Q. Liu, and Z. He, "Review of Pavement Defect Detection Methods," *IEEE Access*, vol. 8, pp. 14531–14544, 2020, doi: 10.1109/ACCESS.2020.2966881.

[4] "Pavement Management - Road Management Office," Feb. 03, 2022. http://www.rmo.ie/pavement-management.html (accessed Feb. 03, 2022).

[5] "PASER Asphalt Roads Pavement Surface Evaluation and Rating PASER Manual Asphalt Roads," 2002. Accessed: Feb. 03, 2022. [Online]. Available: http://tic.engr.wisc.edu.

[6] J. Mccarthy, L. Fitzgerald, J. Mclaughlin, B. Mulry, D. O'brien, and K. Dowling, "Rural Flexible Roads Manual - Pavement Surface Condition Index, Vol 1 of 3, Department of Transport, Toursim and Sports, Dublin, Ireland October 2014," Oct. 2014.

[7] Brian Mulry and John McCarthy, "A Simplified System for Assessing the Condition of Irish Regional and Local Roads," in *Civil Engineering Research in Ireland 2016*, 2016, pp. 1–7, Accessed: Mar. 14, 2022. [Online]. Available: https://ceri2016.exordo.com/files/papers/97/final_draft/097.pdf.

[8] Brian Mulry, Dr. Kieran Feighan, and John McCarthy, "Development and Implementation of a Simplified System for Assessing the Condition of Irish Regional and Local Roads," in *9th International Conference on Managing Pavement Assets*, 2015, pp. 1–17, Accessed: Mar. 14, 2022. [Online]. Available: https://vtechworks.lib.vt.edu/handle/10919/56413.

[9] Fhwa, "Practical Guide for Quality Management of Pavement Condition Data Collection."

[10] A. Ragnoli, M. R. De Blasiis, and A. Di Benedetto, "Pavement Distress Detection Methods: A Review," *Infrastructures*, 2018, doi: 10.3390/infrastructures3040058.

[11] D. O. T. A. HEADQUARTERS, "PAVEMENT MAINTENANCE MANAGEMENT," *Technical Manual TM 5-623*, 1982. https://armypubs.army.mil/epubs/DR_pubs/DR_a/pdf/web/tm5_623.pdf (accessed Feb. 16, 2022).

[12] "Standard Practice for Roads and Parking Lots Pavement Condition Index Surveys." https://www.astm.org/d6433-09.html (accessed Feb. 16, 2022).

[13] F. M. A. Karim, K. A. H. Rubasi, and A. A. Saleh, "The Road Pavement Condition Index (PCI) Evaluation and Maintenance: A Case Study of Yemen," *Organ. Technol. Manag. Constr. an Int. J.*, vol. 8, no. 1, pp. 1446–1455, Dec. 2016, doi: 10.1515/OTMCJ-2016-0008.

[14] J. Wang, Q. Meng, P. Shang, and M. Saada, "Road surface real-time detection based on Raspberry Pi and recurrent neural networks," *Trans. Inst. Meas. Control*, vol. 43, no. 11, pp. 2540–2550, Jul. 2021, doi: 10.1177/01423312211003372.

[15] B. Prasetya, Y. C. S. Poernomo, S. Winarto, R. K. Dewanta, and F. M. Azhari, "Mengurangi Laju Kerusakan Jalan dengan Menggunakan Metode RCI (Road Condition Index) di Kabupaten Madiun," *J. Manaj. Teknol. Tek. Sipil*, vol. 4, no. 1, pp. 104–118, Jul. 2021, doi: 10.30737/JURMATEKS.V4I1.1722.

[16] D. Arya *et al.*, "Deep learning-based road damage detection and classification for multiple countries," *Autom. Constr.*, vol. 132, p. 103935, Dec. 2021, doi: 10.1016/J.AUTCON.2021.103935.

[17] A. Issa, H. Samaneh, and M. Ghanim, "Predicting pavement condition index using artificial neural networks approach," *Ain Shams Eng. J.*, May 2021, doi: 10.1016/J.ASEJ.2021.04.033.

[18] Q. Chen, Y. Huang, H. Sun, and W. Huang, "Pavement crack detection using hessian structure propagation," *Adv. Eng. Informatics*, vol. 49, Aug. 2021, doi: 10.1016/J.AEI.2021.101303.

[19] R. Stricker *et al.*, "Road Surface Segmentation - Pixel-Perfect Distress and Object Detection for Road Assessment," *2021 IEEE 17th Int. Conf. Autom. Sci. Eng.*, pp. 1789–1796, Aug. 2021, doi: 10.1109/CASE49439.2021.9551591.

[20] H. Majidifard, Y. Adu-Gyamfi, and W. G. Buttlar, "Deep machine learning approach to develop a new asphalt pavement condition index," *Constr. Build. Mater.*, vol. 247, p. 118513, Jun. 2020, doi: 10.1016/J.CONBUILDMAT.2020.118513.

[21] T. Rateke, K. A. Justen, and A. Von Wangenheim, "Road Surface Classification with Images Captured From Low-cost Camera-Road Traversing Knowledge (RTK) Dataset," *pdfs.semanticscholar.org*, vol. 26, no. 3, pp. 50–64, 2019, doi: 10.22456/2175-2745.91522.

[22] A. Zhang *et al.*, "Automated Pixel-Level Pavement Crack Detection on 3D Asphalt Surfaces with a Recurrent Neural Network," *Comput. Civ. Infrastruct. Eng.*, vol. 34, no. 3, pp. 213–229, Mar. 2019, doi: 10.1111/MICE.12409.

[23] H. Maeda, Y. Sekimoto, T. Seto, T. Kashiyama, and H. Omata, "Road Damage Detection and Classification Using Deep Neural Networks with Smartphone Images," *Comput. Civ. Infrastruct. Eng.*, vol. 33, no. 12, pp. 1127–1141, Dec. 2018, doi: 10.1111/MICE.12387.

[24] Y. Li, C. Liu, Y. Shen, J. Cao, S. Yu, and Y. Du, "RoadID: A Dedicated Deep Convolutional Neural Network for Multipavement Distress Detection," *J. Transp. Eng. Part B Pavements*, vol. 147, no. 4, p. 04021057, Dec. 2021, doi: 10.1061/JPEODX.0000317.

[25] Y. Jiang, S. Han, and Y. Bai, "Development of a Pavement Evaluation Tool Using Aerial Imagery and Deep Learning," *J. Transp. Eng. Part B Pavements*, vol. 147, no. 3, p. 04021027, Sep. 2021, doi: 10.1061/JPEODX.0000282.

[26] T. Nasiruddin Khilji, L. Lopes Amaral Loures, and E. Rezazadeh Azar, "Distress Recognition in Unpaved Roads Using Unmanned Aerial Systems and Deep Learning Segmentation," *J. Comput. Civ. Eng.*, vol. 35, no. 2, p. 04020061, Mar. 2021, doi: 10.1061/(ASCE)CP.1943-5487.0000952.

[27] I. Hashim Abbas and M. Qadir Ismael, "Automated Pavement Distress Detection Using Image Processing Techniques," *Eng. Technol. Appl. Sci. Res.*, vol. 11, no. 5, pp. 7702–7708, Oct. 2021, doi: 10.48084/ETASR.4450.

[28] T. Lee, Y. Yoon, C. Chun, and S. Ryu, "CNN-Based Road-Surface Crack Detection Model That Responds to Brightness Changes," *Electron. 2021, Vol. 10, Page 1402*, vol. 10, no. 12, p. 1402, Jun. 2021, doi: 10.3390/ELECTRONICS10121402.

[29] J. Menegazzo and A. von Wangenheim, "Road surface type classification based on inertial sensors and machine learning: A comparison between classical and deep machine learning approaches for multi-contextual real-world scenarios," *Computing*, vol. 103, no. 10, pp. 2143–2170, Oct. 2021, doi: 10.1007/S00607-021-00914-0.

[30] T. Rateke and A. von Wangenheim, "Road surface detection and differentiation considering surface damages," *Auton. Robots*, vol. 45, no. 2, pp. 299–312, Feb. 2021, doi: 10.1007/S10514-020-09964-3.

[31] S. Zhou and W. Song, "Crack segmentation through deep convolutional neural networks and heterogeneous image fusion," *Autom. Constr.*, vol. 125, May 2021, doi: 10.1016/J.AUTCON.2021.103605.

[32] J. Zhu, J. Zhong, T. Ma, X. Huang, W. Zhang, and Y. Zhou, "Pavement distress detection using convolutional neural networks with images captured via UAV," *Autom. Constr.*, vol. 133, p. 103991, Jan. 2022, doi: 10.1016/J.AUTCON.2021.103991.

[33] D. Arya, H. Maeda, S. K. Ghosh, D. Toshniwal, and Y. Sekimoto, "RDD2020: An annotated image dataset for automatic road damage detection using deep learning," *Data Br.*, vol. 36, p. 107133, Jun. 2021, doi: 10.1016/J.DIB.2021.107133.

[34] D. Arya *et al.*, "Transfer Learning-based Road Damage Detection for Multiple Countries," Aug. 2020, Accessed: Oct. 25, 2021. [Online]. Available: http://arxiv.org/abs/2008.13101.

[35] A. Zhang *et al.*, "Automated Pixel-Level Pavement Crack Detection on 3D Asphalt Surfaces Using a Deep-Learning Network," *Comput. Civ. Infrastruct. Eng.*, vol. 32, no. 10, pp. 805–819, Oct. 2017, doi: 10.1111/MICE.12297.

[36] P. Y. Shinzato *et al.*, "CaRINA dataset: An emerging-country urban scenario benchmark for road detection systems," *IEEE Conf. Intell. Transp. Syst. Proceedings, ITSC*, pp. 41–46, Dec. 2016, doi: 10.1109/ITSC.2016.7795529.

[37] J. Fritsch, T. Kuhnl, and A. Geiger, "A new performance measure and evaluation benchmark for road detection algorithms," in *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*, 2013, pp. 1693–1700, doi: 10.1109/ITSC.2013.6728473.

[38] "Roadway by RoadBotics." https://roadway.demo.roadbotics.com/map/wPJQ8Zc82QxFHBbswpYs/?assessmentType=normal (accessed Feb. 03, 2022).

[39] "Road Surface Inspection System | Global | Ricoh." https://www.ricoh.com/technology/tech/104_road_surface_monitoring (accessed Feb. 03, 2022).

[40] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel, "IMAGENET-TRAINED CNNS ARE BIASED TOWARDS TEXTURE; INCREASING SHAPE BIAS IMPROVES ACCURACY AND ROBUSTNESS," 2019. [Online]. Available: https://github.com/rgeirhos/texture-vs-shape.

[41] "road-segmentation-adas-0001 - OpenVINO Toolkit." https://docs.openvino.ai/2018_R5/_docs_Transportation_segmentation_curbs_release1_caffe_desc_road_segmentation_adas_0001.html (accessed Feb. 10, 2022).

[42] "ImageNet Benchmark (Image Classification) | Papers With Code." https://paperswithcode.com/sota/image-classification-on-imagenet?tag_filter=104%2C171%2C105 (accessed Feb. 21, 2022).

[43] W. S. Qureshi and (Technological University Dublin), "EfficientNet V2 Model 'M' test dataset qualitative results," 2022. https://tinyurl.com/yc2hs7wn (accessed Mar. 14, 2022).