Hybrid Transfer Learning and Support Vector Machines Models for Asphalt Pavement Distress Classification

Alex Apeagyei
Associate Professor
School of Architecture, Computing and Engineering
University of East London, E16 2RD, UK
Email: a.apeagyei@uel.ac.uk

Toyosi Elijah Ademolake
PhD Student
School of Architecture, Computing and Engineering
University of East London, E16 2RD, UK
Email: a.apeagyei@uel.ac.uk

Joseph Anochie-Boateng
Associate Professor
Faculty of Engineering, Built Environment and Information Technology
Room 12-26, Level 12, Engineering 1
University of Pretoria, Private Bag X20
Hatfield 0028, South Africa

Word Count: 7459 words + 0 table (250 words per table) = 7,459 words

Submitted 1August 2023

Asphalt pavement distress classification; Support Vector Machines; Transfer learning; Hybrid models; Pavement distress classification

ABSTRACT

Pavement condition evaluation plays a crucial role in assisting with the management of the highway infrastructure. However, the current methods used for assessing pavement conditions are costly, time-consuming, and subjective. There is a growing need to automate these assessment tactics and leverage low-cost technologies to enable widespread deployment. This study aims to develop robust and highly accurate models for classifying asphalt pavement distresses using transfer learning (TL) techniques based on pretrained deep learning (DL) networks. This topic has gained considerable attention in the field since 2015 when DL became the mainstream choice for various computer vision tasks. While progress has been made in TL model development, challenges persist in terms of accuracy, repeatability, and training cost. To tackle these challenges, the study proposes hybrid models that combine DL networks with support vector machines (SVMs). Three strategies were evaluated: single DL models using transfer learning (TLDL), hybrid models combining DL and SVM (DL+SVM), and hybrid models combining TLDL and SVM (TLDL+SVM). The performance of each strategy was assessed using statistical metrics based on the confusion matrix. Results consistently showed that the TLDL+SVM strategy outperformed the other approaches in terms of accuracy and F1 score, regardless of the DL network type. On average, the hybrid models achieved an accuracy of 95%, surpassing the 80% accuracy of the best single model and the 55% accuracy for DL+SVM without TL. The results clearly indicate that employing transfer-learned models as feature extractors, in combination with SVM as the classifier, consistently achieves exceptional performance.

Keywords: Asphalt pavement distress classification; Support Vector Machines; Transfer learning; Hybrid models; Pavement distress classification

INTRODUCTION

The impact of pavement distress on the road network's maintenance, both in terms of cost and user safety, is significant. A key first step in the maintenance of roads is the pavement distress data collection and analysis. There are currently two main methods of pavement distress data collection and analysis. Current data collection is undertaken either automatically with the use of dedicated vehicles or manually with visual surveys. Numerous highway agencies emphasize a significant advantage of automating pavement distress data collection over manual methods, particularly with regard to personnel safety. According to McGhee [1] , various U.S. transportation agencies have expressed concerns about the risks associated with manual data collection on roadways, highlighting the potential hazards for individuals. In contrast, modern automated equipment enables the safe and efficient collection of data at traffic speeds. However, existing automated distress collection and identification systems require dedicated vehicles equipped with a variety of sensors, such as high-definition cameras and laser scanners. The limitation of these vehicles is their high purchase ($800,000) [2] and operational (approx. $50 per mile) costs (McGhee 2004). Hence, many US states, for instance, own just a few or none and operate them only once a year. Method of analysis of the data collected also varies depending on the agency. Some agencies rely on multiple operators to view and analyze the digital images captured by the survey vehicle [3] , a tedious undertaking that could be fraught with subjectivity. The reported benefits of such a system include better consistency as only selected operators are used, and comfort and safety, as ratings are done in the office. A major disadvantage, which deep convolutional neural network (DCNN) techniques such as those proposed in this study can address, is the significant amount of time involved in data analysis, thereby impeding the routine collection of continuous distress surveys. Thus, currently for some agencies, only a fraction (about 10%) of each road section is rated and surveys are limited to about once annually. With DCNN, data can be collected continuously with inexpensive vehicle mounted cameras and data analysed automatically.

In this regard, automated road distress identification can play a crucial role in reducing maintenance costs by helping to detect and repair pavement distresses in a timely manner. Consequently, machine learning approaches, particularly transfer learning (TL) models based on deep learning (DL) networks such as GoogLeNet, DenseNet, and ResNet50, have been widely suggested for the automatic classification of asphalt pavement distresses. However, there are still significant hurdles to overcome, including computational complexity, model-dependent accuracy, repeatability, and training cost. Fine-tuning a pre-trained DL model, the commonly used approach requires substantial computational resources. Moreover, the ability to develop highly accurate, repeatable, and robust models involving the hybridization of deep convolutional neural networks (DL) and shallow networks for pavement distress classification is an ongoing challenge.

Deep learning models (DLs) such as convolutional neural networks (CNNs) are well-suited for image classification as they can capture complex patterns and features hierarchically through their layers. They excel at learning discriminative features that differentiate between different image categories. However, DL models can have limitations when it comes to handling small training datasets or scenarios with class imbalance. This is where SVMs may come into play. SVMs are a type of supervised learning algorithm that is effective at binary classification tasks. They can handle data imbalance, robustly handle outliers, and generalize well too when applied to small datasets. Hybrid models have the potential to synergistically combine the best of DLs and SVMs for classification tasks.

In the hybrid models, the DL networks are used for feature extraction and learning high-level representations from the input images. In the context of image classification, a hybrid approach typically involves using a pre-trained DL model, such as the eight pretrained networks used in the current study, to extract features from the input images. These extracted features are then fed into an SVM, which performs the final classification based on the learned feature representations. The combination of DL and SVM in hybrid models aims to leverage the strengths of both approaches. DL models excel at learning complex and hierarchical representations, while SVMs provide robust classification and generalization capabilities. By integrating these two techniques, hybrid models can achieve improved accuracy, especially in scenarios with limited training data or class imbalance, such as pavement distress data.

Pavement distresses are typically caused by load-related, non-load-related or a combination of both. Thus, depending on location, distresses could be attributed to say environmental or traffic or both, thus causing distress imbalance even on the same section of a road pavement.

This study focuses on investigating the performance of hybrid machine learning models compared to single DL-based TL models. Three strategies were considered: DL-based TL, DL and SVM hybrids, and DL-based TL plus SVM hybrids. The experiments involve hyperparameter optimization, feature extraction from DCNN layers, and training with SVM classifiers. The results demonstrate that the hybrid models incorporating DL-based TL plus SVM consistently improve the classification accuracy across all DL models considered.

The remaining sections of the paper include an overview of related work, the methodology, the results and discussion, and the conclusions and recommendations for future work.

Background and related work

Image classification versus image detection

In the context of convolutional neural networks (CNNs), image detection and image classification refer to two different tasks that can be performed on input images. Thus it is important to explicitly distinguish between the two to prevent confusion and facilitate appropriate evaluation. Image classification is the task of assigning a single label or category to an input image. Image detection, on the other hand, is the task of identifying the presence and location of objects of interest in an input image. The current study is limited to image classification of eight selected asphalt pavement distresses as the authors believe the process of assigning labels to an entire image is the first step important step in both image classification and object detection. Once a model(s) has been developed for the classification task, object detection can be easily performed by passing features extracted from the classification stage as inputs for the detection stage.

Classification for asphalt pavement distress

In most cases, road surface conditions are typically assessed visually, either manually or automatically using specially-equipped vehicles. While some data collection aspects have been automated, classifying pavement distress remains a tedious and subjective task. Existing methods are often semi-automated, requiring further analysis by experienced technicians or expensive proprietary systems. Additionally, reported costs for some automated systems are high averaging about $1.2m to acquire a unit and about $70k/year to operate (Vavrik et al. [4]), making them less accessible. Consequently, many agencies still prefer manual inspection, considering it more convenient [5] or only surveying intermittently. Continuous monitoring for distress initiation and propagation is essential for cost-effective maintenance, which existing systems struggle to achieve. The challenge of replicating the expertise of trained technicians, reducing survey cost, and the need for continuous monitoring could be addressed by leveraging deep learning techniques, particularly hybrid transfer learning and support vector machines.

.

Transfer learning for asphalt pavement distresses

There have been several studies that have explored the use of transfer learning (TL) for the classification of asphalt pavement distresses in the attempt to automate asphalt pavement distress inspections. The studies have generally used deep learning models such as convolutional neural networks (CNNs) and transfer learning techniques such as fine-tuning and feature extraction. Most of these previous studies have focused on the development of TL classification networks based entirely on single DL models such as Inception [6-27]. Overall, these studies demonstrate the effectiveness of transfer learning for improving the accuracy of deep learning models for asphalt distress classification. However, the accuracy of transfer learning models can depend on various factors, such as the quality and size of the labelled dataset, the choice of pre-trained model, and the specific type of asphalt distress being classified. Few studies have attempted hybrid models involving multiple machine learning techniques. A key focus of the current study is to fill the gap related to the choice of pre-trained models for transfer learning since previous studies have shown significant differences in the performance of various pretrained DL models [9].

Apeagyei et al. [9] evaluated various state-of-the-art pretrained DCNNs for developing pavement distress classification models using TL approaches. Results indicated significant differences in the performance of selected pretrained in terms of accuracy. They identified the need for future studies to focus on image quality and quantity, exploration of hyperparameter variability, and consideration of synergistic effects from combining multiple DCNNs for improved predictive performance.

Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised machine learning procedure that is frequently employed for tasks involving classification and regression. While deep learning approaches have gained significant attention in recent years, SVMs were popular for image classification tasks before the rise of deep learning architectures. There have been many studies that have utilized Support Vector Machines (SVMs) for the detection of highway distresses. For example, Lin and Liu [13] investigated the possibility of detecting potholes on roads from digital images using SVMs. The researchers extracted colour and texture features from digital road images and used an SVM classifier to distinguish between potholes and non-pothole areas. The SVM model showed promising results in accurately identifying potholes from the road images. Others include Gavilán et al. [14] who used multi-class SVM to classify 10 pavement surface types so that features could be manually selected for the subsequent road distress detection module. The authors identified the need for a more sophisticated and automated approach to feature extraction.

Carvalhido et al. [15] proposed a crack classification technique using SVM. The approach required three different pre-processing configurations to smoothen the texture and enhance potential cracks in images. The authors suggested the need for additional pre-processing techniques, such as median and morphological filters to improve the robustness of their models. The study was limited to cracks whereas a typical highway will involve multiple distresses.

Prasanna et al. [16] proposed a histogram-based classification algorithm and applied it together with SVM to detect cracks on the concrete deck surface; the results on bridge data highlighted the need to improve the accuracy of practical predictions.

Gavilán et al. [14] proposed an automated crack detection system to distinguish between cracked and non-cracked areas on up to ten different types of pavements using a linear SVM-based classifier ensemble. The Gray-Level Co-occurrence Matrix (GLCM), Maximally Stable Extremal Regions (MSER) and Local Binary Patterns (LBP) were used to obtain the components of the feature vector. The authors recommended future studies that involve new performance indexes to differentiate between diverse types of cracks such as longitudinal, transverse and alligator cracking.

Ai et al. [17] proposed a SVM-based method to calculate the probability maps using the information of multi-scale neighborhoods to develop a fused map, which can detect cracks with accuracy higher than any of the original probability maps. The models were evaluated using performance metrics including precision, recall, F1-score, and receiver operating characteristic. The study was limited to the pixel level pavement crack detection problem at the expense of all other common pavement distresses. Furthermore, feature extraction was obtained in part using a probabilistic generative model-based method designed to calculate the probability of a crack for each pixel.

Hadjidemetriou et al. [18] proposed an automated vision-based method for detecting and quantifying pavement patches, which are vital for evaluating and rating pavement surfaces. The proposed system utilizes video frames captured from either a smartphone or an external camera positioned inside or outside a moving passenger vehicle. Support Vector Machine (SVM) classification is employed on feature vectors extracted from the images, utilizing two texture descriptors and the histogram of nonoverlapped square blocks. These feature vectors enable the characterization of image blocks as either patch or non-patch areas. The output of the system provides block-based and image-based classifications. The method demonstrated a detection accuracy of 82.5% for image-based classification.

Hoang et al. [19] compared classification algorithms using machine learning and image processing techniques such as steerable filters, the projective integral of the image, and an enhanced method for image thresholding. Feature extraction was based on image processing and texture computation. The results showed that SVM had the highest level of classification accuracy (87.50%), followed by artificial neural network, ANN (84.25%), and random forest, RF (70%).

Sari et al. [20] examined an automation method of classification and segmentation of asphalt pavement cracks to classify asphalt pavement cracks using the SVM algorithm and segmentation method of the OTSU algorithm. The approach involved classification of distress data into two groups – with crack and no crack. The models were evaluated using multiple performance metrics including accuracy, precision, recall, area under ROC curve (AUC), and ANOVA statistical test. The authors suggested that the SVM algorithm combined with OTSU segmentation and GLCM feature extraction could be used for the classification of asphalt pavement cracks.

The review of existing studies shows that the majority of transfer-learning based studies on pavement distress classification were limited to the evaluation of one or two existing models. Additionally, some existing models do not transfer accurately when applied to new learning so the use of one or two models for training is a major limitation for the pavement distress identification area. The number and definition of distress classes varied widely from two to nine. The studies reviewed demonstrate TL's effectiveness in improving deep learning models' accuracy for pavement distress classification. However, the need for future studies to focus on the choice of pre-trained models for TL, considering significant differences in performance among various DCNN models. Furthermore, exploration of image quality and quantity, investigation of hyperparameter variability, and consideration of synergistic effects from combining multiple distress classification models require further investigation. Reviews of multiple studies using SVM for highway distress detection show potential of the technique, however none of the studies reviewed explored the application of hybrid SVM and TL to address issues such as limitation of training, class imbalance, automated feature extraction, accuracy, and variability in performance of pre-trained DCNN model. Even though some of the selected models have been used in previous TL applications for pavement distress identification, very few, if any, have used the more robust graphical performance measures such as ROC, AUC, and t-SNE.

The paper's significance lies in its development of a hybrid model that effectively addresses challenges in automating pavement distress evaluation, utilizing DL for feature extraction and SVMs for classification, thereby improving accuracy and offering a potential solution to reduce costs and enhance safety in pavement management.

METHODOLOGY

This section presents the main steps used to undertake the study including acquisition of asphalt pavement distress data, selection of pretrained DL models, transfer learning and development of strategies for the proposed hybrid deep and shallow convolutional neural networks for asphalt distress image classification.

As previously discussed, TL is a technique in DL image classification where a pre-trained model that has been trained on a large dataset is used as a starting point to solve a new classification task. For the current study involving asphalt pavement distresses, the approach followed included the following steps: data collection and pre-processing, pre-trained model selection, feature extraction, training a classifier, fine-tuning the pre-trained model, and validation of the trained model. The section concludes with brief descriptions of the architecture and key operational parameters of each experimental strategy.

Data collection

The data collection and preprocessing step consisted of assembling a data store of distress images with category labels. The eight distress labels used in this study have been assigned by taking the names of the folders that contain the image files. The images were automatically resized to the correct image size based on the pretrained model.

Around 400 pictures of asphalt pavement distresses were obtained from various publicly accessible sources, which included Google Street, a commonly used method in this field. The 400 images were manually categorized into eight class labels which are block cracking, distress free, fatigue cracking, longitudinal cracking, patching, pothole, rutting, and transverse cracking by trained technicians. To process each classified image for each network type, user-defined functions were utilized for pre-processing into the required input size. The images were then randomly divided into training (85%) and validation (15%) groups. Samples of the images used and the distribution of images in each distress class are presented in Figure 1. As can be seen in Figure 1, the distribution of images is imbalanced, with the number of images per class ranging between 30 and 60. It is common to have such imbalanced or skewed class distributions with most pavement distress datasets due to climatic factors, traffic loading, and functional classification of the road. Additional distress images were obtained from an actual project located in southern Africa for verification of the most promising models in order to verify the models' performance and assess their generalization ability.

Class imbalance is a fundamental problem in DL classification problems, where some classes have significantly fewer samples than others. In such cases, a DL trained on an imbalanced dataset may be biased towards the majority class and have inferior performance on the minority class. The imbalanced class distribution can lead to overfitting of the model to the majority class, which can result in poor generalization to new and unseen data. Furthermore, the standard performance metrics, such as accuracy, precision, and recall, which are commonly used to evaluate the model's performance, may not provide an accurate representation of the model's performance in such scenarios. To overcome this problem, the technique of undersampling the majority class was found to lead to better model performance than either oversampling or generative adversarial networks (GAN)-generated synthetic images. Thus, all the models were developed by randomly undersampling the majority classes so that each class has the same number of images equal to the number of images in the minority class (i.e. 31 images).

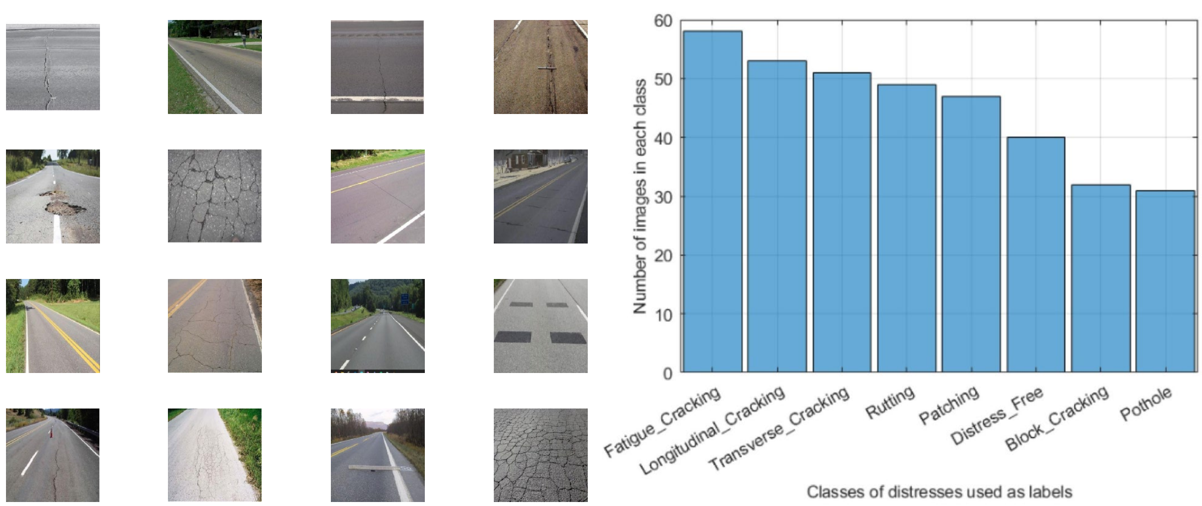

Figure 1 Sample images (left) utilized for training and validating the networks, along with their distribution into the eight distress classes (right)

Pretrained Deep Learning Models

Eight pre-trained DL models were selected including Alexnet, Densenet, Googlenet, Mobilenet, Resnet50, Squeezenet, VGG19, and Xception. The selected models have been pre-trained on the large-scale ImageNet dataset for image classification. ImageNet is a dataset that contains millions of labelled images, and it has been widely used as a benchmark for testing and comparing the performance of deep learning models for image classification. The eight models were chosen because they all achieved state-of-the-art performance on various image classification tasks. Further details of the selected DL models including a description of model performance on asphalt pavement distress classification can be found in [9].

The feature extraction step involved using the pre-trained model to extract features from the input images. This step also involves removing the classification layers of the pre-trained model and using the remaining layers to extract features from the images. Following the feature extraction step, training of the selected pre-trained models followed. This step involved tuning model hyperparameters to optimize performance and comprised of using the Stochastic Gradient Descent with Momentum (SGDM) as the optimizer, selecting the initial learning rate, specifying mini batch size depending on model size, and setting maximum epochs of 12 to optimize training time and accuracy. It should be noted that the selection of the aforementioned parameters was based on an extensive trial and error approach as well as the authors experience in the field.

The pre-trained models were fine-tuned by unfreezing some of their layers and retraining them on the new dataset. This step involves selecting the layers to be unfrozen and adjusting the learning rate to avoid overfitting. In this study, the models were fine-tuned by replacing the last three layers, which included the fully connected layer, the softmax layer, and the classification layer. The rest of the layers were kept frozen, which is a common method used by previous researchers to achieve the best predictive accuracy for their models. However, the number of layers to freeze or train can vary among different investigators. Some studies have replaced more layers than the three used in this study, but it should be noted that this does not always result in better models. The ultimate goal for all researchers is to achieve the best possible accuracy in their models.

For the current study, three strategies were evaluated: the traditional single transfer learning-based DL models (TLDL), hybrid DL+SVM, and hybrid TLDL+SVM. A schematic of the approach is depicted in Figure 2. A brief description of each strategy is discussed next.

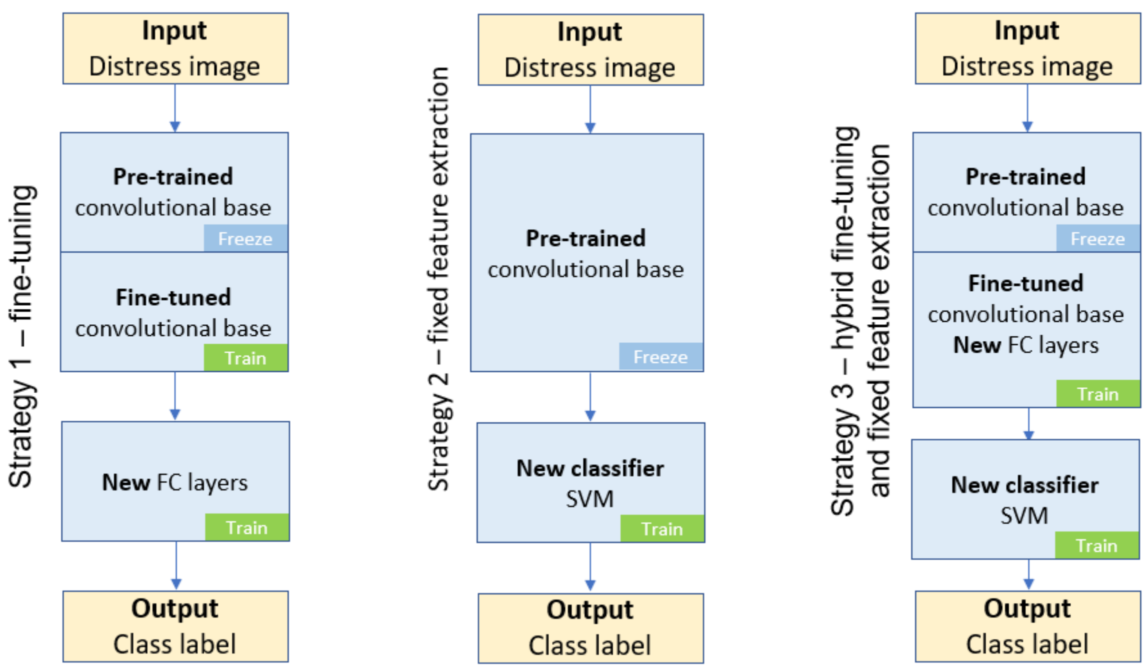

Figure 2. Strategies for developing and evaluating hybrid DL and SVM models for asphalt pavement distress classification

Strategy 1 – Transfer Learning, Fine-tuning Pretrained Model (TLDL)

A DCNN network at the basic level consists of three-layer sets: convolutional layers, pooling layers and fully connected layers. Even though each layer of a DCNN produces a response, or activation to an input image, only a few of the layers are suitable for image feature extraction. While the layers at the beginning of the network capture basic image features, such as edges, brightness or blobs, deeper layers detect more complex features that uniquely define the object.

Stochastic gradient descent method was used as the optimization routine for TL of the eight selected networks because of its accuracy and efficiency. Similar hyper-parameters were selected for use to ensure realistic comparisons among different DCNN architectures. In a DCNN, hyperparameters are used to control the learning process that determines model parameters that a network eventually learns. For this study, the model hyperparameters used included, among others, the following: 1) initial learning rate of 0.001, 2) momentum of 0.9, 3) L2 regularization of 0.0001, 4) epochs of 12, and 5) minibatch size of 5.

A graphical processing unit (GPU) with an NVIDIA® T1000 based on Turing architecture, and an Intel Core i-7 CPU at 2.6 GHZ operating on a Windows 10 Pro 64-bit operating system were used. A common mini-batch size of 8 was used for all the eight pretrained DCNNs. Mini-batches are samples of the training dataset that are processed on the GPU at the same time and therefore can impact the speed of training and the accuracy of a network. The larger the mini-batch, the faster the training in terms of computational efficiency. However, larger mini-batch sizes may lead to longer training times per epoch due to the accumulated time needed to process a larger batch before updating the model. The networks were compiled using the stochastic gradient descent (SGD) optimisation technique. To fine-tune the selected models for the transfer learning process for each model, the last fully connected layer of the original network was replaced with a new fully connected layer, which classified the features into the eight pavement distress categories.

Each network was retrained to identify eight categories of flexible pavement distresses. The steps used to accomplish the transfer training of each network included: 1) importing the pre-trained network, 2) configuring selected layers to perform a new recognition task, 3) training the network on a pre-processed pavement distress dataset and 4) using the results to predict and assess network accuracy.

Strategy 2 – fixed feature extraction DL+SVM

Similar to strategy 1, the fixed feature extraction approach used involved a process of replacing the last three layers including the fully connected, the softmax, and the classification layers from a pretrained network while maintaining the convolutional base consisting of a series of convolutional and pooling layers. The choice of which deep layer to choose to extract the image feature is a design choice. In this study, the layer right before the classification layer was used. For example, for Resnet50, the feature layer used was the fully connected layer 'fac1000.' The selected features together with training options such as minibatch size of 6 were used to train a multiclass SVM classifier with Error Correcting Output Codes (ECOC). SVM with ECOC is a technique used to extend the binary classification capability of SVM to handle multiple classes. In the context of the current study, SVM with ECOC was used to classify pavement distresses into eight different classes. A fast Stochastic Gradient Descent solver was used to train the multiclass, ECOC function with the 'learners' parameter set to linear and the coding parameter set to one versus all. The foregoing procedure used earlier to extract image features from a testing set of images was then passed to the classifier to measure the accuracy of the trained classifier. Predictions using the classifier were made and model performance was evaluated using the confusion matrix measures, and other performance metrics such as AUC and t-SNE.

Strategy 3 – hybrid fine-tuning and fixed feature extraction

This strategy involved all the steps described in strategy 1 leading to a trained DCNN capable of classifying distress images into eight class labels. From this stage, the process was similar to strategy 2 in that the layer right before the classification layer was selected as the extraction layer. Next, the selected features together with training options such as minibatch size were used to train a multiclass SVM classifier. A fast Stochastic Gradient Descent solver was used to train a multiclass, error-correcting output codes (ECOC) function with the 'learners' parameter set to linear, and the coding parameter set to one versus all. The performance of the resulting hybrid models was evaluated using confusion matrix metrics as described next.

Evaluation of Retrained Models

The prediction performance of each retrained DCNN hybrid model was evaluated by comparing confusion matrix statistics and accuracy measures such as precision, overall accuracy, and recall, which are commonly used to assess how well TL-based DCNNs perform by most previous investigators. In addition, combined measures such as F1-score and graphical measures including ROC, AUC, t-SNE, etc., that are more robust against class imbalance and overfitting were used.

Confusion matrix

The predictive performance of a deep learning (DL) model can be visualized using a confusion matrix (CM), which presents the results in a tabular format. Each element of the CM represents the number of predictions made by the network and whether they were classified correctly or incorrectly. Evaluating the diagonal entries of the CM is a common method to assess the success of a DCNN classifier. To correctly interpret the CM results, it is important to consider some fundamental concepts, which are further explained by Düntsch and Gediga [21]. In a two-class classification problem, where typically a positive and a negative class are involved, four metrics are commonly used. These metrics include true positives (t_p), false positives (f_p), true negatives (t_n), and false negatives (f_n).

These metrics can be utilized to estimate various key performance measures such as accuracy, F1-score, precision, recall, and specificity (Eqs. 1-5). Others include receiver operating characteristic curve (ROC), and area under the curve (AUC) for a given network. The overall accuracy is evaluated using the F1-score, which is the harmonic mean of recall and precision. The F1-score balances the trade-off between precision and recall. A value of 0 for the F1-score indicates that either the precision or the recall is 0.

$$\mathrm{Accuracy} = \frac{t_p + t_n}{t_p + t_n + f_p + f_n} \qquad (1)$$

$$\mathrm{Precision} = \frac{t_p}{t_p + f_p} \qquad (2)$$

$$\mathrm{Recall} = \frac{t_p}{t_p + f_n} \qquad (3)$$

$$\mathrm{Specificity} = \frac{t_n}{t_n + f_p} \qquad (4)$$

$$\mathrm{F1-score} = \frac{2}{1/_{Precision} + 1/_{Recall}} = \frac{2t_p}{2t_p + f_p + f_n} \qquad (5)$$

Receiver Operating Characteristic (ROC) Curves

A receiver operating characteristic curve (ROC) is a true positive rate (TPR) versus a false positive rate (FPR) plot that can be used to display the performance of a network at all classification thresholds. It is considered as one of the most robust measures of the predictive performance of a DCNN classifier. The magnitude of the classification threshold controls the number of items classified as positive. Thus, a network operating at lower classification thresholds will classify more items as positive than a network operating at a higher threshold. An ROC can be used to determine other useful performance metrics including the optimal operating classification threshold (OPROCPT) and the areas under the ROC curve (AUC). Both OPROCPT and AUC have values between 0 and 1, with values closer to 1 associated with better performance.

Area Under the ROC Curve (AUC)

The area under the ROC curve or AUC is a measure of the area with coordinates ranging from (0,0) to (1,1) and therefore has a magnitude of 1. A model with an AUC of 0 will be expected to make predictions that will be 100% wrong. A model with an AUC of 1.0 will be expected to be correct 100% of the time and will rank all positives higher than all negatives. In practice, it is expected that a reliable classification model will rank a random positive example higher than a random negative example more than 50% of the time and have an AUC in the range of 0.5-1.0. AUC is considered a more robust measure of performance than some of the previously reviewed measures such as accuracy, F1-score and recall as it is not affected by class imbalance

RESULTS AND DISCUSSION

Accuracy

The performance of the trained networks using the three strategies was evaluated using multiple performance measures including overall accuracy and F1-score (Figure 3) obtained from confusion matrices. The results obtained highlight the impact of the training strategy on the performance of the models. It is evident from Figure 3 that the choice of strategy significantly influenced the overall performance, with the models trained using strategy 2 demonstrating comparatively poorer results. Conversely, all models exhibited strong performance under strategy 3. A comprehensive analysis of the outcomes indicates that there exist statistically significant differences in performance between strategy 1 and strategy 2 (p-value = 0.001). Furthermore, the results unequivocally indicate that strategy 3 outperforms both strategy 1 and strategy 2 by a significant margin (p-value < 0.000011). Specifically, the accuracy of models trained under strategy 3 was notably superior to the accuracy achieved by the other two models. This finding aligns with our expectations, as it substantiates the superior capabilities of the models trained using the hybrid DL and SVM models. The detailed performance metrics of the models are presented in Figure 3 where the average performance for the three strategies were 0.6945±0.066, 0.5486±0.069, and 0.9383±0.019, respectively for strategy 1, strategy 2 and strategy 3.

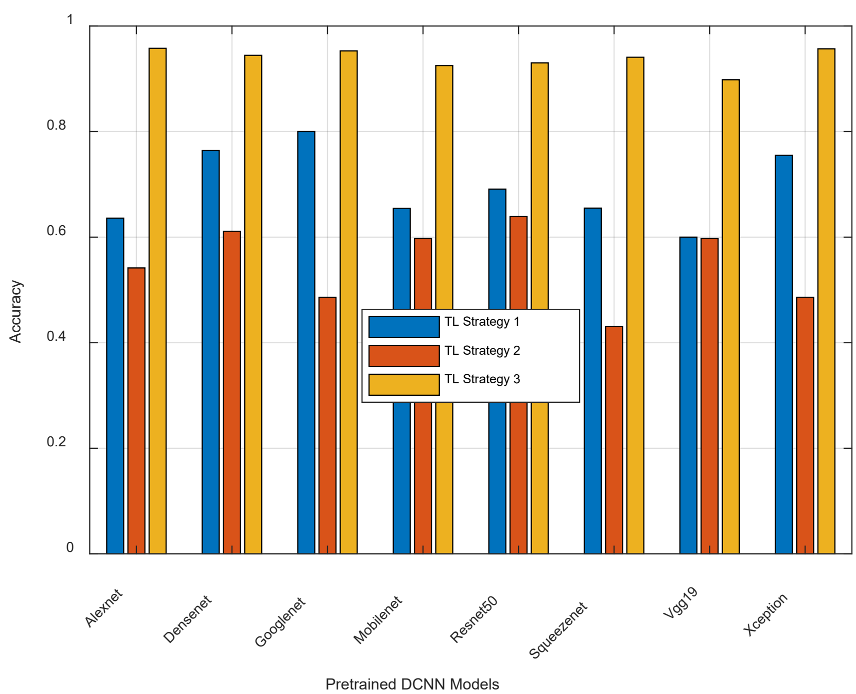

FIGURE 3 Evaluation of Pretrained DCNN Models' Performance (F1) with Different Training Strategies

Computational time and model complexity

Figure 4 shows a comparison of the three strategies in terms of training times and model complexity (assumed to be related to the size on file of each trained model). In the presented bubble plot depicting the performance of three strategies, the y-axis spans from 0 to100%, representing the accuracy of each model. while the x-axis extends logarithmically from 0 to 1460. The bubble sizes correspond to the respective sizes of the models on disk, providing a visual representation of their complexities. Strategy 1 models are centered on the graph, showcasing a balanced trade-off between model size, training time and accuracy. Strategy 2 models are positioned near the left-middle corner of the plot, indicating a modest model size, moderate accuracy and shorter training time. In contrast, Strategy 3 models are situated near the upper right corner, signifying superior accuracy and a larger size on disk compared to the other models. The distinct locations of the three models on the plot offer a clear illustration of their trade-offs between accuracy and storage efficiency, with Strategy 3 models emerging as the top performers in terms of predictive accuracy. The training times for Strategy 2 and strategy 3 can be considered as comparable, with the difference in most cases differing by a few dozens of seconds.

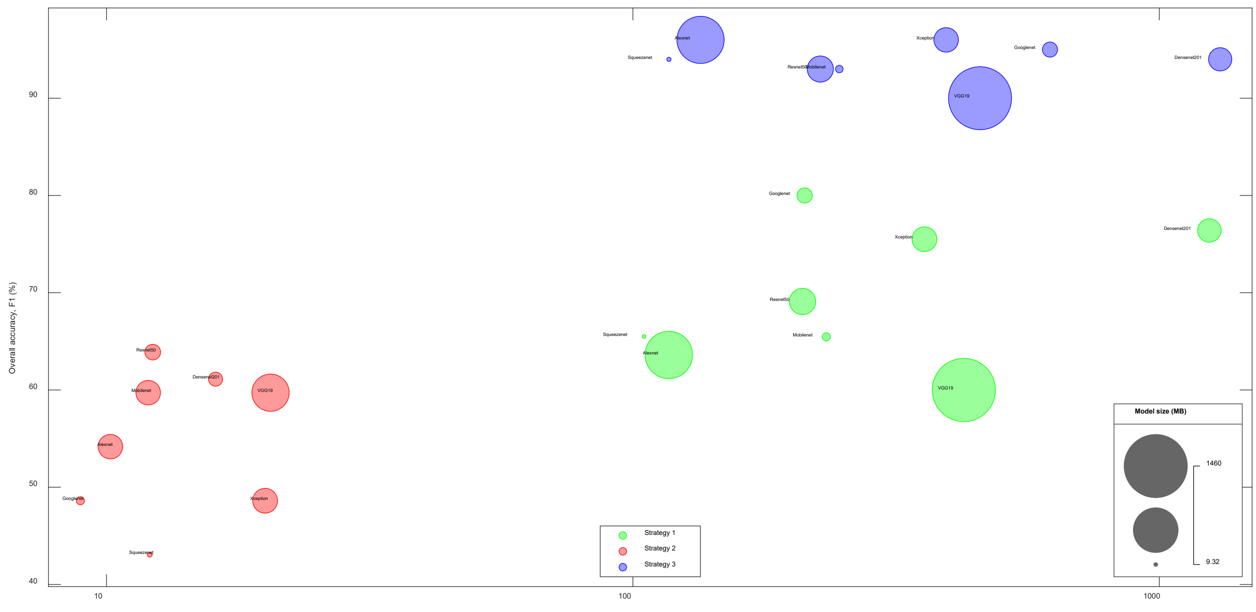

FIGURE 4 Comparison of the training speeds versus module complexity (size on disk) for eight pretrained SVM-DCNNs hybrid asphalt pavement distress classification models.

Robustness of hybrid models

As previously discussed, during training, distress images were randomly sampled for both training and testing. This necessary step naturally introduces randomness into the development of the models. Thus, it was deemed necessary to assess the robustness and repeatability of the SVM hybrid models' performance. Experiments were conducted that involved 10 separate runs of model training. For each experimental run, the F1 parameter was recorded. Figure 5 provides a summary of the results in the form of box-and-whisker plots. The comparative analysis of the experiments revealed interesting insights into their performance. It can be seen from Figure 5 that the performance of the models varied widely. For instance, considering Figure 5 (left), Xception achieved the highest median F1 score of 0.67, followed closely by Alexnet (0.63) and Squeezenet (0.61). Large variability in F1 was observed across all the modules with some models exhibiting data that can be considered as outliers. None of the literature reviewed reported this variability in training data as many authors only reported data for the best modules only. Thus, this study contributes to our understanding of this important phenomenon. A classifier with an average F1 score of about 0.51 can be considered as performing slightly better than random guessing, especially if there are eight different classes in the dataset, as is the case in this study. If a classifier's accuracy is consistently around 0.51, it suggests that the model is not able to capture the underlying patterns or features that differentiate the classes effectively. In other words, the classifier is making incorrect predictions more often than correct ones. While it may not be accurate to describe the classifier as a random number generator, as it still shows some capability to differentiate between classes, it is clear that there is room for improvement in its performance. Overall, the performance of strategy 2 was only slightly better than a random model as the average F1-score for all pretrained models combined was slightly more than 0.51.

For Strategy 3 (see Figure 5, right), among the tested models, Xception consistently exhibited the highest median F1 score of 0.9633, closely followed by Densenet (0.9630) and Googlenet (0.9519). These models consistently demonstrated strong classification capabilities across multiple test runs, indicating their effectiveness in distinguishing different distress types in asphalt pavement. On the other hand, models such as Squeezenet (0.9056) and VGG19 (0.9023) achieved slightly lower average F1 scores, suggesting comparatively lesser performance in classifying pavement distresses. It is worth noting that while these models had relatively lower average F1 scores, they still showed acceptable performance levels. Furthermore, the standard deviations for all models were relatively low, indicating consistency in their performance across different test runs. This consistency implies that the Strategy 3 models' classification performance is reliable and robust.

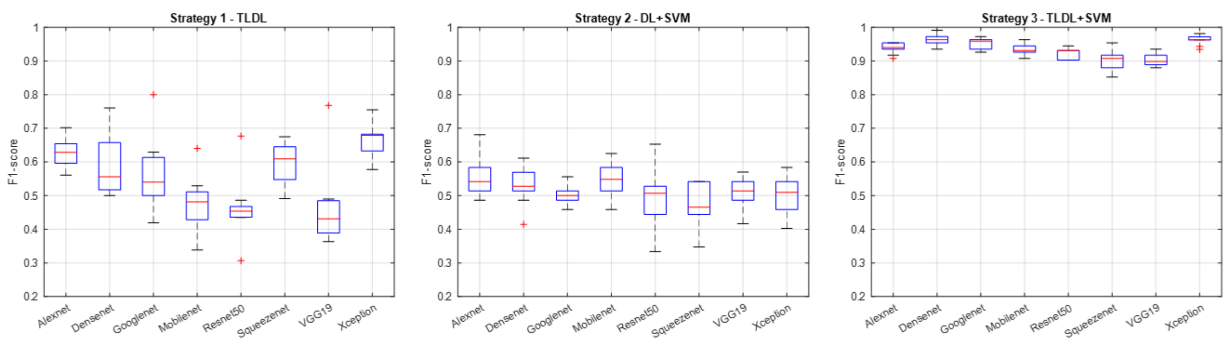

FIGURE 5 Quantifying model reliability and stability: F1-score distribution across 10 repeat runs

Receiver Operating Characteristic (ROC) analysis

Receiver Operating Characteristic (ROC) analysis was used as an evaluation metric to assess the performance of the classification models and their ability to discriminate between the different classes of pavement distresses. In image classification tasks, a model predicts the probability or confidence of an image belonging to a certain class. The output of the model can be interpreted as a continuous score or a probability. To construct an ROC curve, the model's output scores are used to calculate the true positive rate (recall) and the false positive rate (1 - specificity) at various classification thresholds. A classification threshold is applied to these scores to determine the predicted class labels. By varying the threshold, the trade-off between the true positive rate and the false positive rate can be adjusted. A lower threshold leads to more positive predictions, while a higher threshold results in fewer positive predictions. Based on the selected threshold, the model's predictions are compared with the ground truth labels of the images. The true positive rate (TPR) is the ratio of correctly predicted positive samples (e.g., images from the positive class) to the total number of positive samples. The false positive rate (FPR) is the ratio of incorrectly predicted positive samples to the total number of negative samples (e.g., images from the negative class). On an ROC, the true positive rate is plotted on the ordinate, and the false positive rate is plotted on the abscissa. By varying the threshold and calculating the TPR and FPR at each point, multiple (TPR, FPR) pairs are obtained. Connecting these points creates the ROC curve. The ROC curve provides valuable insights into the trade-off between recall and specificity in the classification task. A classifier with a higher ROC curve, closer to the top-left corner, indicates better performance in distinguishing between classes (Figure 6). The ROC curve shows a solid circle symbol at the operating point of the model for each class. The dotted line on the graph represents a random model. A perfect model is indicated by TPR equal to 1 and FPR equal to 0. On the other hand, the worst model is represented by TPR equal to 0 and FPR equal to 1. In the legend, the class name and corresponding AUC value are provided for each curve. In Figure 6, we compare the ROC curves obtained from strategy 2 and strategy 3 when using the Mobilenet pretrained network. The ROC curves for strategy 3 (hybrid TLDL+SVM) are plotted higher and closer to the top-left corner compared to strategy 2, indicating superior performance. Moreover, Figure 6 shows that the AUC for strategy 2 ranged from 0.6843 to 0.9824, while for strategy 3 it ranged from 0.9757 to 1.0000. The ROC curve and its associated AUC demonstrate that the hybrid TLDL+SVM method possesses superior predictive capability in distinguishing distressed pavements across multiple classes. Similar findings were observed across all the other models analyzed. These results suggest that strategy 3 consistently outperforms strategy 2 when using the eight pretrained networks.

The area under the curve (AUC) is calculated as the integral of the ROC curve. It represents the overall performance of the classifier across all possible threshold values. An AUC value of 1 indicates a perfect classifier, while an AUC value of 0.5 suggests a random or no-discrimination classifier.

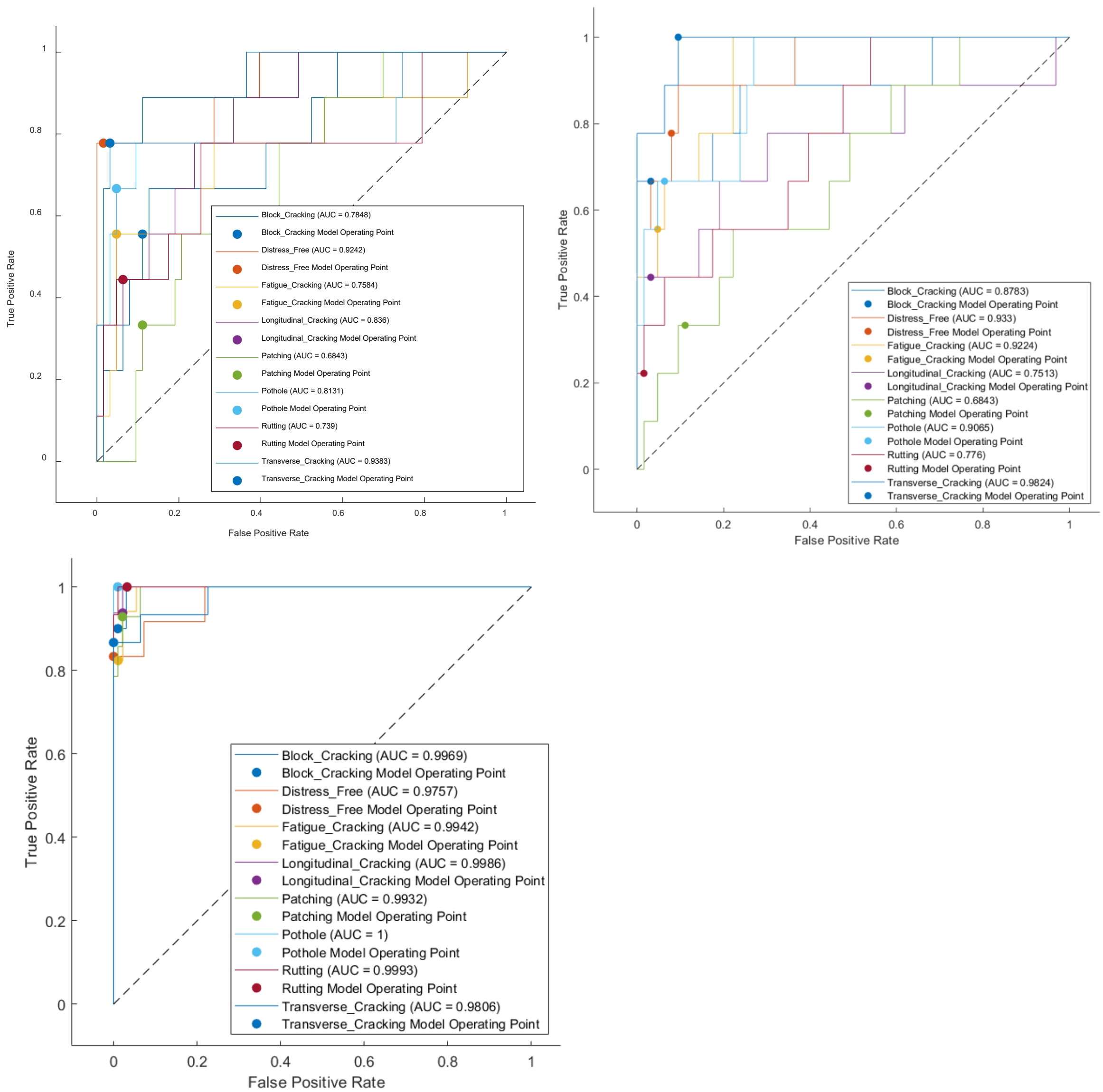

FIGURE 6 Comparison of ROC curves for hybrid SVM and DL pavement distress classification models. Top Left: Strategy 1; Top Right Strategy 2; Bottom Strategy 3

Comparing the Statistical Significance of the Performance of Hybrid TL and SVM Classifiers

The results obtained from our study demonstrate that employing transfer-learned models as feature extractors, combined with SVM as the classifier, yields superior performance (F1-score greater than 0.90, AUC greater than 0.98, and ROCs that plot in the top lefthand corner) regardless of the pretrained network used. To determine if significant differences exist among the different models developed under Strategy 3, we conducted the Kruskal-Wallis test followed by the Dunn's test. The Kruskal-Wallis test is particularly valuable for analyzing data when the assumptions of parametric tests, such as normality and equality of variance, are not met or not known, especially in scenarios involving multiple groups. In Dunn's test, the mean ranks are employed to assess the statistical significance of the differences between groups. A significant difference in mean ranks indicates that the groups are statistically distinguishable from each other, without providing information about the direction or magnitude of the difference. The results of the Dunn's test conducted on the pairwise comparisons between the eight different network groups (Alexnet to Xception) showed significant differences exist between certain pairs of groups. For example, Densenet vs Resnet50, Densenet vs Squeezenet, Densenet vs VGG19, Googlenet vs Squeezenet, Googlenet vs VGG19, Googlenet vs Xception, Resnet50 vs Xception, Squeezenet vs Xception, and VGG19 vs Xception show p-values less than 0.05, suggesting statistically significant differences. The Dunn's test results reveal both significant and non-significant differences between the groups, providing valuable insights into the comparative performance of the various groups under investigation. It can be inferred from the results that even though all Strategy 3 models performed very well, models based on the three pretrained networks, Densenet, Googlenet, and Xception, are outstanding as compared with previous studies [8-27].

Extracted Features

To gain insight into the distribution, relationships, and separability of the extracted features across the different classes of images, the dimensionality reduction technique known as t-Distributed Stochastic Neighbor Embedding (t-SNE) was applied to the extracted features comprising eight classes and 255 columns of data. The results of the t-SNE visualization are depicted in Figure 7. When conducting a t-SNE visualization of extracted features from a DL model, if the classes plot on separate parts of the figure with no overlaps, similar to Figure 7, it suggests that the extracted features have distinct and well-separated representations for each class. This separation indicates that the DCNN has successfully learned discriminative features that can effectively differentiate between the different classes. In other words, the t-SNE visualization shows that the DL network has been able to capture meaningful patterns and structures in the data, resulting in a clear separation of the classes in the feature space. This is desirable as it implies that the DL network has learned to encode relevant information specific to each class, enabling accurate classification or recognition of the different categories. The absence of overlaps between the classes in the t-SNE plot indicates that the extracted features have high intra-class similarity and low inter-class similarity. It suggests that instances belonging to the same class are closer to each other in the feature space, while instances from different classes are farther apart. This separation can be seen as an indication of the effectiveness of the DCNN in capturing the distinctive characteristics and discriminative information associated with each class. The results justify the two-step approach proposed in this study.

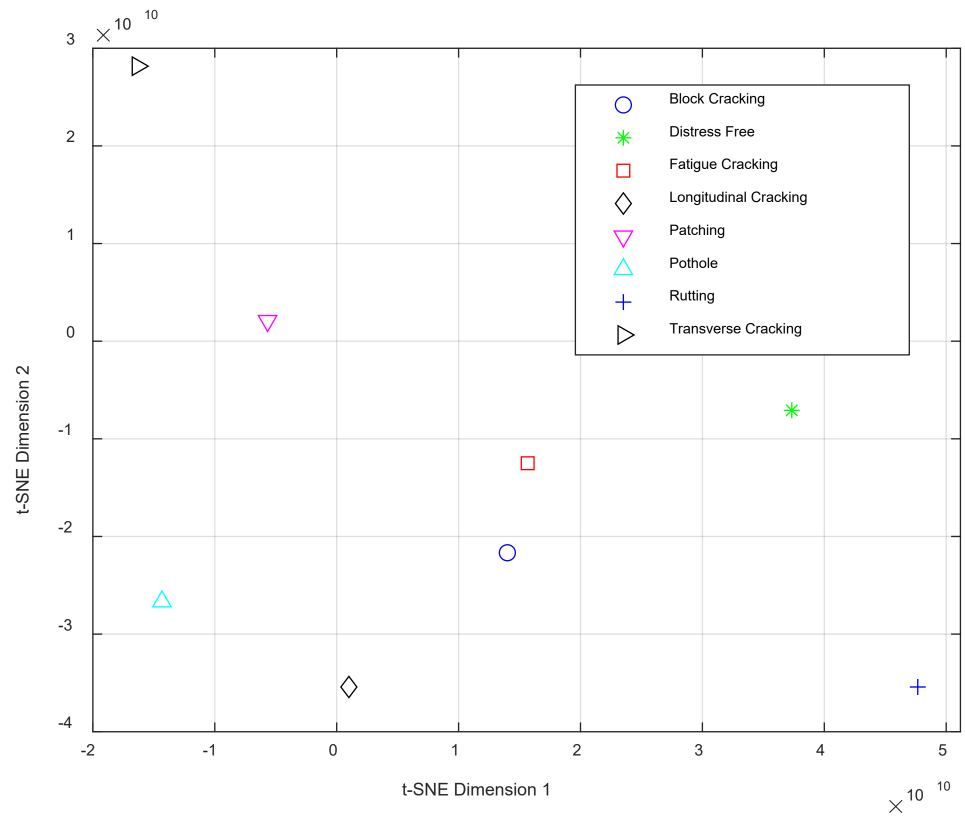

FIGURE 7 Sample t-SNE visualization of extracted features from retrained Mobilenet for Strategy 3. The symbols represent the eight distress classes arranged in alphabetical order from Block cracking to Pothole

Verification of models based on the most promising strategies

In addition to the training and validation process previously discussed in the development of the models, it is also common to have a final testing set that remains completely separate from both the training and validation sets. This testing set is designed to provide an unbiased assessment of the model's performance after the entire development process. It helps evaluate how well the model generalizes to completely new data and provides a more reliable estimate of its real-world performance. This step thus serves to verify the model's performance and assess its generalization ability. Therefore, as a further check on the performance and generalization ability of the models developed using the promising approach (Strategy 1 and Strategy 3), a new set of distress images which were not part of the dataset used in the training and validation of the models was acquired from a project completed in southern Africa by one of the co-authors of the current paper [27]. That project involved detailed forensic investigation to determine the causes of premature failure of asphalt concrete in Tanzania and contained a large dataset of distress images which provided an ideal opportunity for verifying the performance accuracy of the models developed in this.

Figures 8-11 show the results of analysis based on strategy 3; similar results were obtained for strategy 1. Also shown in Figures 8-11 are confidence attached by the model to the classification of each image. For example, the model assigned 100% confidence to the image in Figure 11, classifying it as patching. In Figure 8, the model was conflicted assigning a label transverse cracking to the image with about 58% confidence but at the same time providing evidence that the distress could be longitudinal cracking. It should be noted that this is a common problem even with human experts and in line with previous studies [20-26].

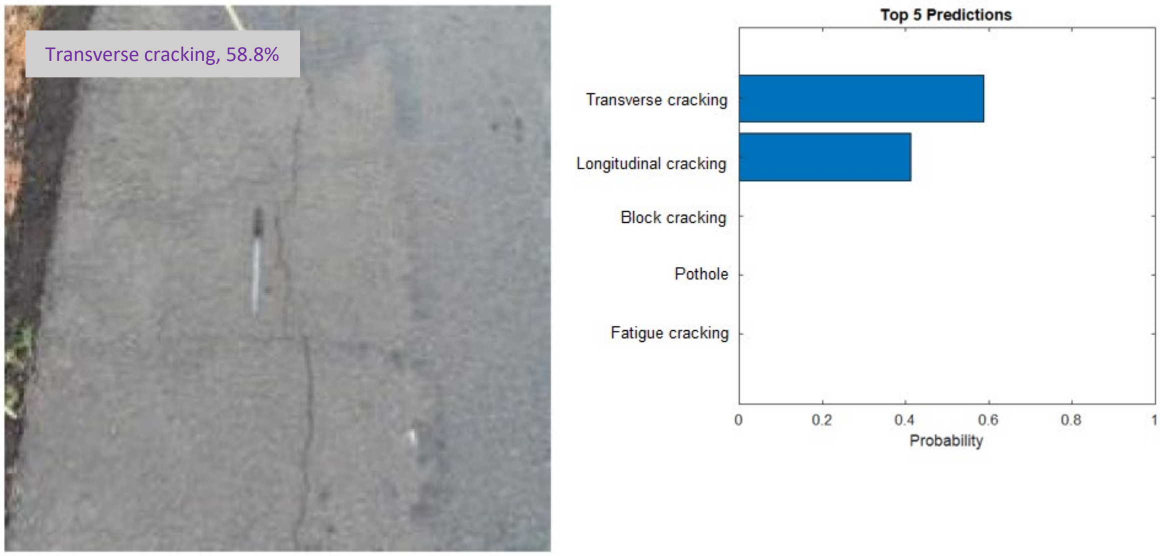

FIGURE 8 Verification of most promising models based on Strategy 3 for transverse cracking

FIGURE 9 Verification of most promising models based on Strategy 3 for rutting

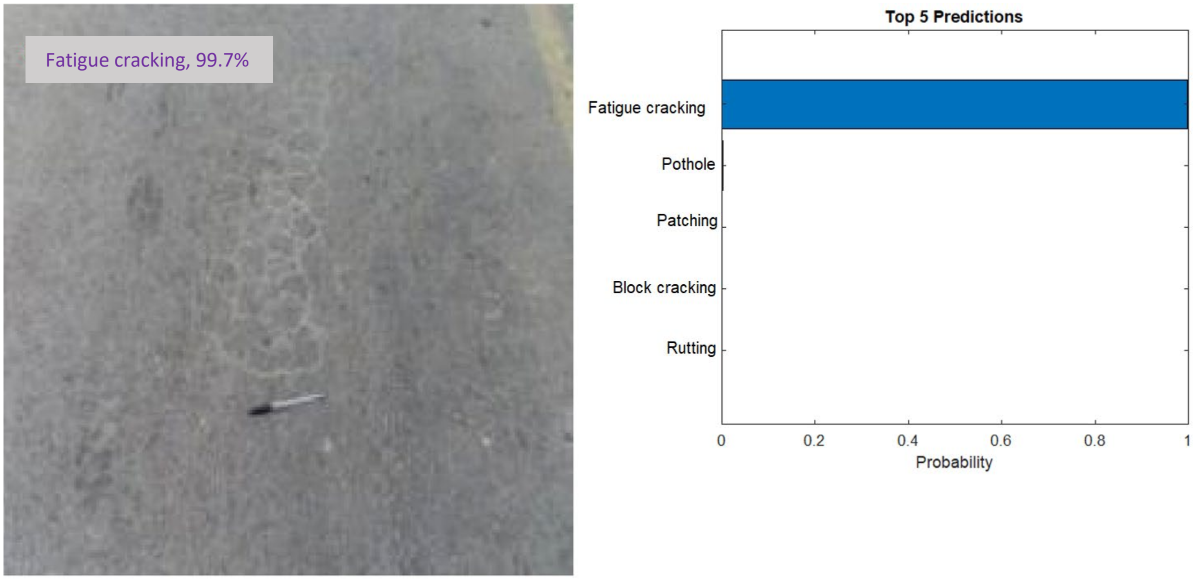

FIGURE 10 Verification of most promising models based on Strategy 3 for fatigue cracking

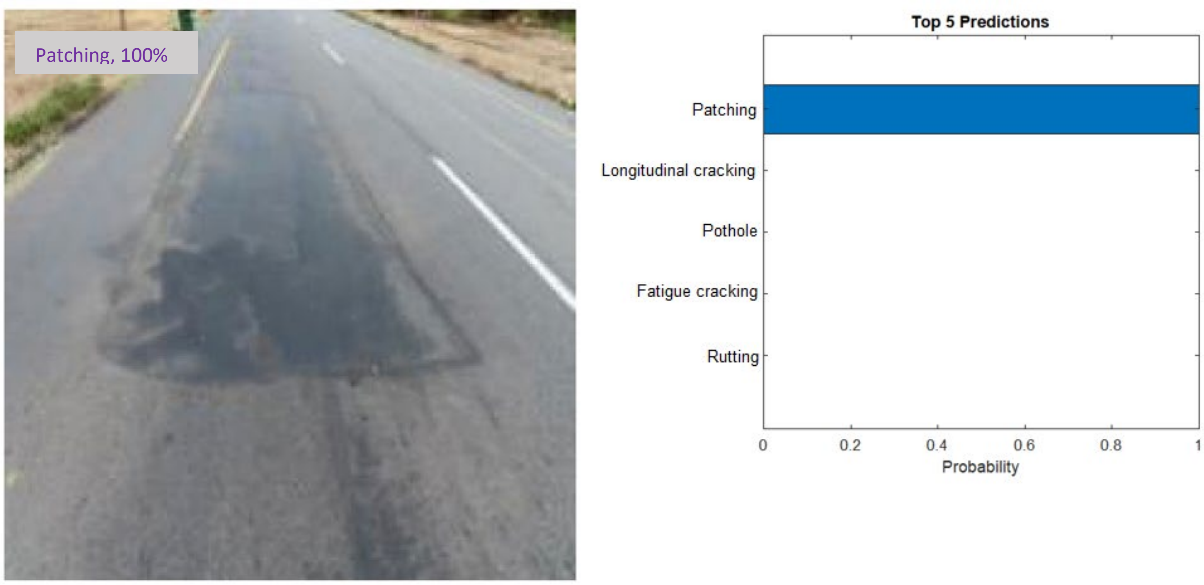

FIGURE 11 Verification of most promising models based on Strategy 3 for patching

CONCLUSIONS

We present a novel approach to classify distresses in asphalt pavement by utilizing a combination of deep learning and support vector machine models. The study focused on investigating the performance of hybrid machine learning models compared to single DL-based TL models. Three strategies were considered: DL-based TL, DL and SVM hybrids, and DL-based TL plus SVM hybrids. The experiments conducted involved hyperparameter optimization, feature extraction from DCNN layers, and training with SVM classifiers. The following conclusions were made based on the results presented in this paper:

1. Hybrid models incorporating DL-based TL plus SVM consistently improve the classification accuracy across all DL models considered, although the computation times of SVM models were longer for some models. Specifically, at the level of pre-trained models, our hybrid approach yielded an impressive F1 score of 0.96. In contrast, the alternative strategies, regardless of the pre-trained

model or hyperparameter optimization employed, failed to surpass an F1 score of 0.80. The findings strongly support the integration of hybrid transfer learning and SVM classifiers into the automated asphalt distress classification models, as they consistently enhance their overall accuracy.

2. The t-SNE visualization of extracted features from the TL models showed the different classes plotting on separate parts of the figure with no overlaps. The absence of overlaps between the classes in the t-SNE plot indicates that the DL-extracted features have high intra-class similarity and low inter-class similarity. This suggests that the extracted features have distinct and well-separated representations for each class. The robust separation of classes in the t-SNE plot demonstrates the capability of the transfer-learned pretrained models to extract discriminative features that capture the unique characteristics of distress images, allowing for accurate classification and analysis.
3. The results clearly indicate that employing transfer-learned models as feature extractors, in combination with SVM as the classifier, consistently achieves exceptional performance.

One limitation of the current study is that crucial details about the pavement distress dataset, including image resolution, weather conditions, and geographic locations, were unavailable to the authors. Subsequent studies should prioritize capturing these vital details as road pavement images are usually obtained under uneven weather/lighting conditions. Secondly, the current study is limited to image classification of eight selected asphalt pavement distresses, additional studies focused on object detection is warranted. Furthermore, based on our study's outcomes, future research should aim to 1) explore additional pretraining strategies, such as fine-tuning or self-supervised learning, to evaluate their impact on feature extraction and classification accuracy, 2) investigate key factors influencing the training time discrepancy between models, including network architecture, and 3) assess how the proposed methodology could estimate the severity of classified distresses.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the support of the University of East London for the research reported in this paper.

AUTHOR CONTRIBUTIONS

The authors confirm their contribution to the paper as follows: study conception and design: A. Apeagyei; data collection: A Apeagyei, TE Ademolake and J Anochie-Boateng; analysis and interpretation of results: All authors; draft manuscript preparation: All authors reviewed the results and approved the final version of the manuscript. The authors do not have any conflicts of interest to declare.

REFERENCES

1. McGhee K (2004). NCHRP Synthesis of Highway Practice 334: Automated Pavement Distress Collection Techniques, Transportation Research Board, National Research Council, Washington, D.C.
2. Radopoulou SC, Brilakis I, Doycheva K and Koch (2016). A Framework for Automated Pavement Condition Monitoring, Proceedings, Construction Research Congress 2016, ASCE, https://doi.org/10.1061/9780784479827.078. Accessed 26 November, 2023.
3. MnDOT (2015). An Overview of Mn/DOT's Pavement Condition Rating Procedures and Indices, Minnesota Department of Transportation. https://www.dot.state.mn.us/materials/pvmtmgmtdocs/Rating_Overview_State_2015V.pdf. Accessed 26 November, 2023.
4. Vavrik W, Stefanski J, and Sargand (2013). PCR evaluation–Consider transition from manual tosemi-automated pavement distress collection and analysis .Ohio Department of Transportation. Office of Statewide Planning and Research. https://rosap.ntl.bts.gov/view/dot/26795. Accessed on 28 November 2023.
5. Siriborvornratanakul, T. 2018. An automatic road distress visual inspection system using an onboard in-car camera. Advances in Multimedia, 2018: 1-10.
6. Maeda H, Sekimoto Y, Seto T, Kashiyama T, and Omata H (2018). Road damage detection using deep neural networks with images captured through a smartphone. https://doi.org/10.1111/mice.12387.
7. Ranjbar, S., Nejad, F.M. & Zakeri, H. An image-based system for pavement crack evaluation using transfer learning and wavelet transform. Int. J. Pavement Res. Technol. 14, 437–449 (2021). https://doi.org/10.1007/s42947-020-0098-9.
8. Nie M and Wang K (2018). Pavement Distress Detection Based on Transfer Learning. 2018 5th International Conference on Systems and Informatics (ICSAI), 435-439.
9. Apeagyei AK, Ademolake TE, and Adom-Asamoah M (2023). Evaluation of deep learning models for classification of asphalt pavement distresses, International Journal of Pavement Engineering, 24:1, DOI: 10.1080/10298436.2023.2180641.
10. Gopalakrishnan K (2018). Deep Learning in Data-Driven Pavement Image Analysis and Automated Distress Detection: A Review. Data. 2018; 3(3):28. https://doi.org/10.3390/data3030028.
11. Gopalakrishnan K, Gholami H, Vidyadharan A, Choudhary A, and Agrawal A (2018). Crack damage detection in unmanned aerial vehicle images of civil infrastructure using pre-trained deep learning model. Int. J. Traffic Transp. Eng. 1–14.
12. Gopalakrishnan K, Khaitan S, Choudhary A and Agrawal A (2017). Deep convolutional neural networks with transfer learning for computer vision-based data-driven pavement distress detection. Construction and Building Materials, 157: 322-330.
13. Lin J and Liu Y. (2010). Potholes Detection Based on SVM in the Pavement Distress Image. International Symposium on Distributed Computing and Applications to Business, Engineering and Science. 544-547. 10.1109/DCABES.2010.115.
14. Gavilán M, Balcones D, Marcos O, Llorca DF, Sotelo MA, Parra I, Ocaña M, Aliseda P, Yarza P, Amírola A (2011). Adaptive Road Crack Detection System by Pavement Classification. Sensors 2011, 11, 9628-9657. https://doi.org/10.3390/s111009628.
15. Carvalhido AG, Marques S, Nunes, FD, and Correia PL (2012). Automatic Road Pavement Crack Detection using SVM. https://fenix.tecnico.ulisboa.pt/downloadFile/395144950971/resumo.pdf. Accessed on 07 April 2023.
16. Prasanna P, Dana K, Gucunski N, and Basily B (2012), "Computervision based crack detection and analysis," Sensors Smart Struct. Technol. Civil, Mech. Aerosp. Syst. 2012, vol. 8345, p. 834542.
17. Ai D., G. Jiang, L. Siew Kei, and C. Li. 2018. "Automatic pixel-level pavement crack detection using information of multi-scale neighborhoods." IEEE Access, vol. 6, pp. 24452–24463.

18. Hadjidemetriou GM, Vela PA, Christodoulou SE (2018). Automated Pavement Patch Detection and Quantification Using Support Vector Machine, Journal of Computing in Civil Engineering, 32(1), 04017073, doi: 10.1061/(ASCE)CP.1943-5487.0000724.
19. Hoang ND and Nguyen QL (2019). A novel method for asphalt pavement crack classification based on image processing and machine learning. Engineering with Computers 35, 487–498 (2019). https://doi.org/10.1007/s00366-018-0611-9.
20. Sari, Y., Prakoso, P.B., & Baskara, A.R. (2019). Road Crack Detection using Support Vector Machine (SVM) and OTSU Algorithm. 2019 6th International Conference on Electric Vehicular Technology (ICEVT), 349-354.
21. Duntsch, I., and Gediga, J,2019. 3rd international Conference on Machine Vision and Information Technology (CMVIT 2019).Journal of Physics: Conference Series, 1229, 011001.
22. Chen C, Chandra S, Han Y and Seo H (2022). Deep Learning-Based Thermal Image Analysis for Pavement Defect Detection and Classification Considering Complex Pavement Conditions, Remote Sens. 2022, 14(1), 106; https://doi.org/10.3390/rs14010106.
23. Majidifard, H., Jin, P., Adu-Gyamfi, Y. and Buttlar, W. 2020. Pavement image datasets: A new benchmark dataset to classify and densify pavement distresses. Transportation Research Record: Journal of the Transportation Research Board, 2674(2): 328-339.
24. Mandal, V., Uong, L. and Adu-Gyamfi, Y. 2018. Automated road crack detection using Deep Convolutional Neural Networks. 2018 IEEE International Conference on Big Data (Big Data).
25. Peraka NSP, Biligiri KP, Kalidindi SN (2021). Development of a Multi-Distress Detection System for Asphalt Pavements: Transfer Learning-Based Approach. Transportation Research Record. 2675(10):538-553. doi:10.1177/03611981211012001.
26. Zhu J, Zhong J, Ma T, Huang X, Zhang W, Zhou Y (2022). Pavement distress detection using convolutional neural networks with images captured via UAV, Automation in Construction, Vol. 133, https://doi.org/10.1016/j.autcon.2021.103991
27. Anochie-Boateng, JK, Mataka, MO, Malisa, JT, and Komba, JJ. 2015. In: 32 - Road pavements of the XXVth World Road Congress in Seoul, Seoul, South Korea, November 2015, 15pp, https://proceedings-seoul2015.piarc.org/en/documents/individual-papers/s-2589.htm.