applied sciences

MDPI

Article

Evaluation of Infrared Thermography Dataset for Delamination Detection in Reinforced Concrete Bridge Decks

Eberechi Ichi * and Sattar Dorafshan

Department of Civil Engineering, College of Engineering & Mine, University of North Dakota, 243 Centennial Drive Stop 8115, Grand Forks, ND 58202-8115, USA; sattar.dorafshan@und.edu
* Correspondence: eberechi.ichi@und.edu

Abstract: Structural health monitoring and condition assessment of existing bridge decks is a growing challenge. Conventional manned inspections are costly, labor-intensive, and often risky to execute. Sub-surface delamination, a leading cause of deck replacement, can be autonomously and objectively detected using infrared thermography (IRT) data with developed deep learning AI models to address some of the limitations associated with manned inspection. As one of the most promising classifiers, deep convolutional neural networks (DCNNs) have not been utilized to their fullest potential for delamination detection, arguably due to the scarcity of realistic ground truth datasets. In this study, a common encoder–decoder semantic segmentation-based DCNN is adapted through domain adaptation. The model was tuned and trained on a publicly available dataset to detect subsurface delamination in IRT data collected from in-service bridge decks. The authors investigated the effect of dataset augmentation, class imbalance, the number of classes, and the effect of background removal in the training dataset, resulting in an overall number of seventy-five UNET models. Four out of five bridges were adopted for training and validation, and the fifth bridge was for testing. Most models averaged 80 iterations, and the training progress finally reached a training accuracy of 75% with a loss of about 0.6 without any overfitting. The result showed a substantial difference in the minimum and maximum values for the evaluated performance metrics (0.447 and 0.773 for global accuracy, 0.494 and 0.657 for mean accuracy, 0.239 and 0.716 for precision, 0.243 and 0.558 for true positive rate (TPR), 0.529 and 0.899 for true negative rate (TNR), 0.282 and 0.550 for F1-score. The results also indicated that the models trained on the raw annotated balanced dataset performed best for half of the metrics. In contrast, the models trained on raw data (with no dataset enhancement) performed better when only global accuracy was considered.

Keywords: bridge infrastructure; non-destructive evaluation (NDE); deep learning; artificial intelligence; unmanned aerial system (UAS); infrared thermography (IRT); deep convolutional neural networks (DCNN); semantic segmentation; concrete; bridge deck; delamination detection; encoder–decoder; uNET

Citation: Ichi, E.; Dorafshan, S. Evaluation of Infrared Thermography Dataset for Delamination Detection in Reinforced Concrete Bridge Decks. *Appl. Sci.* **2024**, *14*, 2455. https://doi.org/10.3390/app14062455

Academic Editors: Dwayne McDaniel and Cesar Levy

Received: 1 February 2024
Revised: 2 March 2024
Accepted: 5 March 2024
Published: 14 March 2024

1. Introduction

Effective bridge inspection and assessment techniques are a growing concern and challenge to stakeholders, investors, practitioners, highway users, and concerned government institutions. The deteriorating condition of the over 619,588 bridges across the United States, of which 7.5% are considered structurally deficient, has motivated the US Government to provide and deploy a USD 40 billion bridge fund program to states for bridge repairs through the year 2026. However, this fund still cannot cater to a backlog of nationwide bridge repair funding needs of about USD 125 billion [1].

Ratings of bridges are determined by human and visual inspection of the various parts of the bridges. However, traditional methods of inspections, such as physical and visual inspections, are hindered by several limitations, such as variability and inconsistencies in designated condition ratings by different inspectors. These inconsistencies are usually

affected by factors such as fear of traffic, near visual acuity, color vision, formal training on bridge inspection, dependency on skill or experience, accessibility, and complexity of structure [2–5]. In addition, an inspection of the bridge structure is usually by contact, which involves a significant setback, such as lane closure, traffic disruption, hard-to-reach areas (sub-surface damages), personnel safety, accident-related issues, cost, and time consumption [3].

Unmanned and autonomous inspections with artificial intelligence (AI) are potentially viable techniques aiding conventional and traditional bridge inspection methods at a fast-growing pace [6]. In recent decades, automation has gained ground in bridge inspection to help alleviate and offset the problems and challenges faced using traditional methods. Several non-destructive techniques have been used to investigate sub-surface defects in bridge decks, such as sound techniques (hammer sounding, chain dragging, and impact echo). Other methods are electric resistivity (ER), ultrasonic surface wave (USW), ground penetration radar (GPR), and visual sensor imaging [7–10]. However, these methods are not effective in the detection of sub-surface delamination. Unmanned aerial systems (UAS) mounted with sensors such as infrared thermography (IRT) in recent years have proven to be effective in collecting NDE data and assessing sub-surface defects of bridge structures, fatigue crack detection in steel bridges with fracture critical members (FCM), and corrosion detection in ancillary structures [11,12]. The UAS-IRT method assesses bridges in a timely, effective, and cost-efficient manner, revealing sub-surface defects, minimizing traffic closures, and ensuring the safety of inspectors, providing access to difficult areas inaccessible by traditional methods [13–17].

Autonomous inspections performed by platforms such as robots or UAS are usually image-based. The dataset collected during the inspection is either processed visually by an inspector or autonomously by a computer algorithm/model. Previous studies have mainly demonstrated image-based inspections using computer vision and deep learning techniques in crack detection, spalls, and corrosion in concrete bridges, bearing displacement, and bolt loosening, but very few on delamination [12,18–23]. Past studies have detected sub-surface delamination using image processing methods and deep convolutional neural networks (DCNN) models on laboratory-prepared specimens. Most of these studies are usually investigated mainly on laboratory specimens that do not depict the on-site and environmental constraints for IRT data collection [13,24]. The few studies on delamination detection of bridge decks are image-based only, where inspectors visually process and classify the images based on pixel intensity or visual characterization. The image-based method is usually based on temperature difference and pixel-contrasting thresholding techniques. Other studies adopt image-enhancement techniques for semantic segmentation to distinguish the delaminated from sound portions of the structure [13,14]. In studies conducted by Cheng et al. [24] for automatic delamination segmentation, the authors adopted an encoder–decoder architecture. The developed model was trained on augmented laboratory data, validated, and tested on non-augmented data. The model falsely detected defects, resulting in lower precision and recall.

To the best of the authors' knowledge, very few or no studies have been carried out extensively on evaluating delamination detection for in-service bridges using publicly available IRT datasets for training, validating, and testing DCNN models. Most DCNN applications for delamination detection are conducted on mostly laboratory-developed datasets. These datasets are usually collected in a controlled environment. In this study, we have used a publicly available dataset [13,14,25,26] to develop a DCNN model for delamination classification and prediction. We further evaluated the effect of different preprocessing and annotated datasets on the model's output. We have adopted an encoder–decoder UNET DCNN model in our study. This architecture had been successful for biomedical image segmentation [27] but had not been effectively deployed to evaluate delamination in bridge decks. The architecture has also been adopted in evaluating three classes of bridge structural conditions: delamination, rebar exposure, and non-damage [28]. Therefore, the aim of this study centers on deploying the UNET encoder–decoder DCNN

architecture to train the publicly available SDNET2021 dataset for pixel-wise semantic segmentation in bridge decks. The following objectives are highlighted in achieving the goal of this study:

i Adopted publicly available IRT dataset of in-service bridge deck for training, validating, and testing DCNN model;
ii Considered different levels of preprocessing on the performance of the model;
iii Compared the DCNN model's performance with the image-based method adopted on the same dataset.

1.1. Infrared Thermography (IRT)

Structural defects in concrete bridge decks are broadly categorized into surface and sub-surface defects. Inspection of surface defects such as cracks can be detected by visual inspection or cameras. On the other hand, delamination is a sub-surface anomaly not visible within the visible wavelength spectrum. IRT applications have grown rapidly in recent years and have been used in studies such as those investigating historic structures [29] and medical studies [30]. IRT has been significantly explored in assessing and detecting delamination in bridge decks [6,11,14].

The theory behind IRT technology is based on the thermal emission and absorption between two different material mediums with different thermal properties. The material's temperature and emissivity control the amount of radiation being emitted. This relationship was developed and shown by the Stefan–Boltzmann Equation in Equation (1) [31].

$$E = \varepsilon \, \sigma \, T^4 \quad (1)$$

where E = Total surface radiation ($W.m^{-2}$), ε = Emissivity, σ = Stefan–Boltzmann constant, and T = Temperature of material (in Kelvin). Material mediums with higher emission capacity are more suitable for IRT tests. Emissivity ranges from 0 to 1 and changes due to the variability of surface texture, temperature, and emission characteristics [29]. Therefore, it is common for defective regions to have different emissivity from sound regions. IRT applications are either conducted by active or passive thermography. Active means require introducing an artificial heating or cooling source to stimulate temperature differences before testing, while passive thermography does not require a heat source. Field tests are usually conducted during the available heating source of the sun. This is usually the case for large structures such as bridge decks. Other NDE methods such as microwave thermography, pulsed-eddy-current-simulated thermography, flash thermography, vibrothermography, sonic thermography, and laser thermography are passive-based and have been highlighted in Ichi et al. [14], where an external source of heat is introduced. Delamination, an internal defect, causes a thermal gradient (ΔT), resulting in surface temperature difference due to different heat conduction rates between the sound and defective portions. A 0.2–0.5 °C thermal contrast is suitable for possible detection [11,32,33].

Past studies have elaborated and established several factors and conditions that adversely affect the reliable outcome of delamination detection using IRT. Some of these are sizes and depth of delamination, materials preset within the delamination, concrete's thermal properties, the presence of overlays, data collection time, height of data collection, speed of UAS data collection platform, surface conditions, solar loading, ambient wind speed, ambient temperature, and sensor resolution [11,14,22,34,35].

1.2. DCNN Application in Segmentation

Recently, the use of convolutional neural networks (CNN) for image classification tasks has been on the rise. This is due to its ability to extract complex features autonomously from large datasets [24,36]. CNN has broad input, feature extraction, and classification layers. The feature extraction layers contain sub-layers such as convolutional, activation function, pooling, and batch normalization, while the classification layers have fully connected, dropout, SoftMax, and output layers. The function of the feature extraction layer is primarily

to determine a common pattern in the sets of images, such as edges, textures, shapes, and objects in the images. These features are, in turn, fed into the classification layer for the prediction or classification of the input image. Commonly adopted CNN models for image classification, AlexNet [36], VGG [37], GoogleNet, and ResNet [38,39], are appropriate for image classification tasks.

Deep learning has been broadly applied for defect detection, condition assessment, inspection, and evaluation of existing infrastructures. Studies have been conducted to investigate the performance of the DCNN for defect detection, such as delamination, cracks, spalling, and patches in visual and infrared thermography images of laboratory-modeled and in-service civil infrastructures. Past studies have conducted segmentation by deploying region-level or pixel-wise semantic segmentation procedures. Region-level segmentation in bridge inspections draws bounding boxes around damages. Cha et al. [40] adopted Faster R-CNN, a region-based segmentation model, to detect structural defects. Chen and Jahanshahi et al., in their study, boost the detection performance of the model by introducing transient information using video sequences. This model is, however, limited by detecting various defects at the grid-cell level. Pozzer et al. [41] investigated the MobileNetV2 model for multi-class damage detection in thermal images from an existing buttress dam. The classification model identified 79.7% of the defects with a reduction in false positives (FP) when the VGG 16 model was adopted. The encoder–decoder type of architecture has been used in past studies for pixel-wise classification in semantic segmentation tasks [42]. Fully convolutional networks (FCNs) have also been used with high prospects [43].

DCNN models have been shown to perform optimally and better with more datasets. The effect of data augmentation was investigated by Cheng et al. [24]. Their studies revealed a drop in the model's performance after augmentation when considering the intersection over union (IoU). A general drop in IOU ranged from 12% to 19%. The authors stated that the model's overall performance for on-site use can be improved with more robust in-service data. The lack of a model trained with in-service data is a significant roadblock in deploying DCNN for the autonomous assessment of existing bridge decks.

A review of the published literature indicates that autonomous delamination detection of in-service bridge decks is not common. Most models have been developed by training laboratory datasets subjected to active heating sources. These generated datasets do not depict the actual conditions of the bridge decks. The performance of models has been based on laboratory datasets. Therefore, these datasets are augmented to generate more datasets for training and validation. This study is carried out to develop a trained DCNN model with a publicly available IRT dataset for delamination detection and to evaluate the effect of augmentation, dataset balancing, and a combination of these on the performance of the developed models.

1.3. U-Net Encoder–Decoder Architecture

Convolutional neural networks (CNN) have been broadly applied in the semantic segmentation of cracks in bridge decks but have rarely been used for delamination detection for in-service bridge decks. The encoder uses convolution and pooling to down-sample the input images. The size of the original image is reduced, keeping the spatial features in place throughout the processing until the end. After a series of down-sampling has been completed, the images are thereafter passed through a deep neural network and fed into the decoder. In the decoder subnetwork, sampling techniques are used to upscale the images. Deconvolution operations are applied to the images for upscaling. The up-sampling and down-sampling blocks are connected to each other through skip connections. After completing a series of up-sampling, the network outputs an image with segmented masks on the features. A 2-by-2 max pooling layer comes after two sets of convolutional and ReLU layers in the U-Net encoder subnetwork. The decoder subnetwork comprises two convolutional and ReLU layers after a transposed convolution layer for up-sampling. Similarly, the bridge consists of two sets of convolutions and ReLU layers. The convolutional layer starts with a zero-bias term. Using the weight initialization technique created by He et al., the

convolution layer weights in the encoder and decoder subnetworks are initialized [38]. The uNET architecture framework adopted in this study is shown in Figure 1.

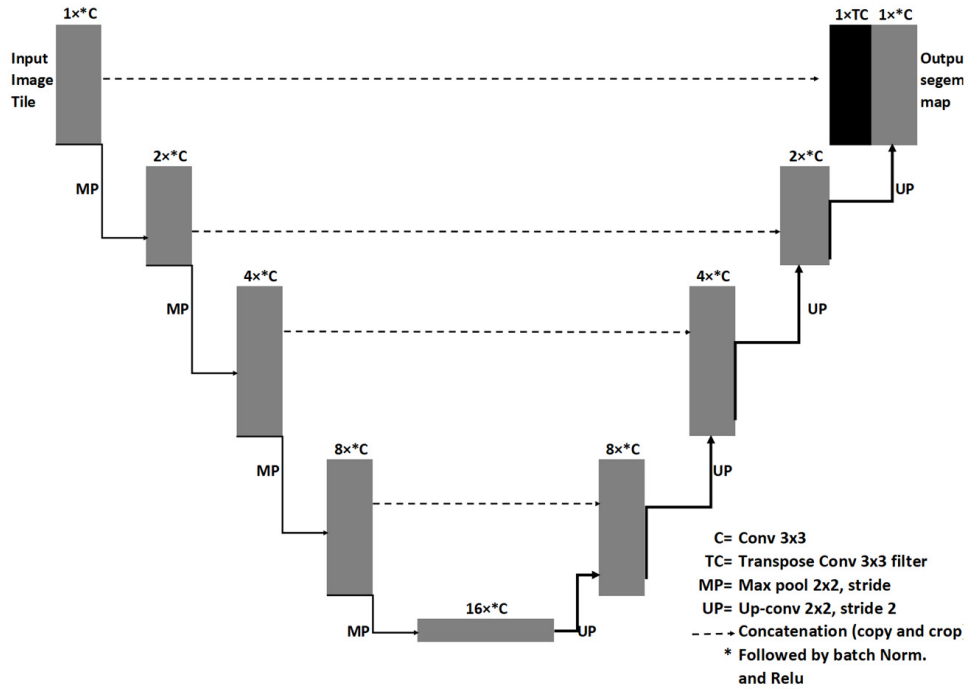


Figure 1. UNET encoder–decoder architecture. (Adapted from Wang et al. [44].)

2. Research Methodology and Dataset

The description of the research methodology and dataset adopted in this study is presented as follows.

2.1. Data Acquisition and Ground Truth for Validation

In this study, five existing reinforced concrete bridge deck slabs were inspected and evaluated for sub-surface anomalies such as delamination. The bridges were between 47 and 49 years old and, at the time of investigation, were supporting the I-29 traffic. The spans range from 64 m to 142 m for Forest and Park River bridges, respectively [13,18,24,25]. At an average height of 17 m above ground level (AGL), a DJI Matric 210 UAS equipped with a FLIR XT V2 infrared thermal camera was used for non-invasive IRT data collection (Table 1). A summary of data collection conditions, such as the data collection date, UAS specification, sensor specifications, ambient weather conditions, data quality, data preprocessing, and annotation, and the number of pixels, are discussed in detail in Ichi et al. [25]. Table 1 provides an overview of the FLIR XT V2 infrared thermal camera specifications for data collection, while Table 2 highlights an overview of the data collected from the bridge decks. The resolution of the stitched images is shown in Table 1. The set of image frames collected for each of the five investigated bridge decks was stitched to generate a single mosaic thermal image using Agisoft Metaphase 2021 © software. The annotated IRT dataset used in this study and its detailed description are discussed in SDNET2021 [25,26] and available via DOI at https://doi.org/10.31356/data019, accessed on 21 September 2023.

The methodology developed in this study is shown in Figure 2. The raw images from SDNET2021 were annotated based on the number of proposed classes. Initially, the raw data was annotated as a two-class problem (classes 0 and 1). The sound and background were annotated as class 0, while the delaminated pixels were annotated as class 1.

Table 1. Camera specifications for FLIR XT V2 thermal camera.

Characteristics	Specifications
Thermal Resolution	640 × 512 pixels
Full Frame Rates	30 Hz (NTSC) 25 Hz (PAL)
Spectral Band	7.5–13.5 µm
Pixel Pitch	17 µm
Thermal Imager/Detector Type	Uncooled VOx Microbolometer
Digital Zoom	2×, 4×
Field of View	24° × 19°
Focal Length	9 mm

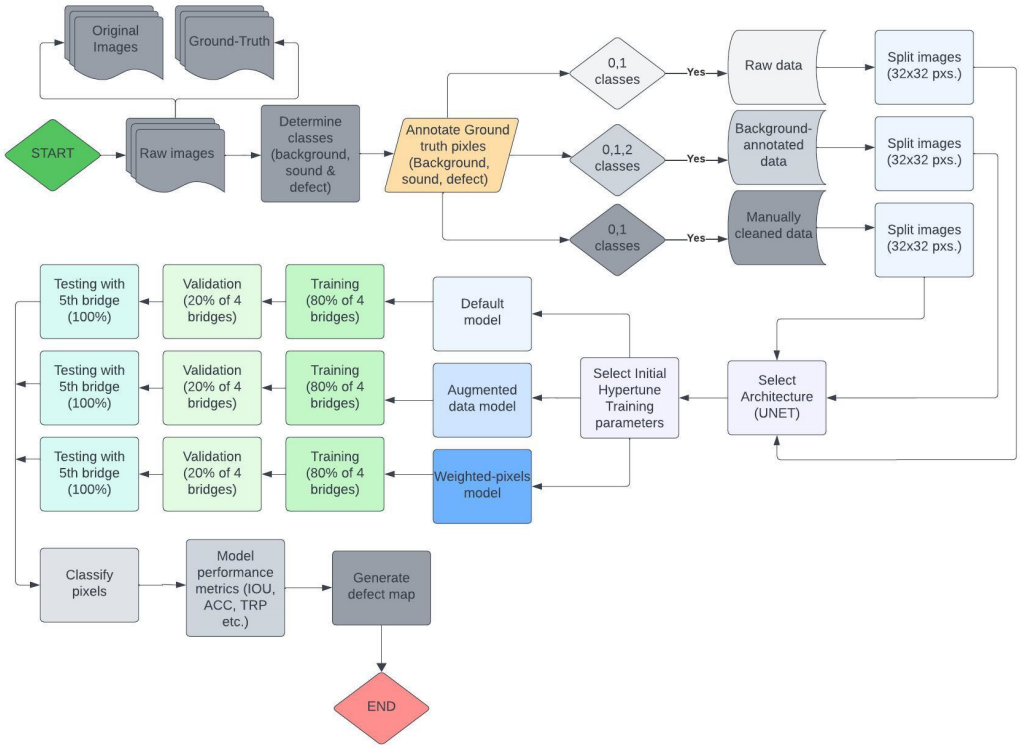


Figure 2. Methodology and workflow.

This developed dataset was called the raw annotated dataset (RD). To evaluate any likely effect of annotation on the dataset and examine effective background segmentation, we further annotated the dataset differently as a background annotated dataset (BD) and a manually cleaned dataset (MD). The MD was also annotated as two-class segmentation problems (classes 0 and 1), where class 0 signified sound and class 1 signified delaminated pixels. The BD was a three-class problem (classes 0, 1, and 2) for the sound, delaminated, and background pixels. The number of delaminated, sound, and background pixels for each bridge is shown in Table 2. The three datasets adopted in this study are summarized as follows.

i. RD—The developed ground truth images were preprocessed such that the sound and background pixels were labeled zeros (0), and the delaminated pixels were labeled ones (1), showing a binary pixel annotation and classification problem.
ii MD—After sub-dividing the image of 32 × 32 pixels, the blocks for the background pixels outside the bridge deck's ROI were manually hand-picked and excluded from the ground truth and corresponding dataset. This thereby reduces the number of

image blocks for each bridge deck. Similarly, the pixels were annotated as two classes (binary classification): sound (0) and delaminated (1).

iii BD—The dataset was preprocessed such that the sound pixels were labeled zeros (0), delaminated pixels were labeled as ones (1), and the background pixels outside the deck's ROI were annotated as twos (2). This presents three (3) classes with multiple classifications. The images were split into 32 × 32 sub-images based on the recommended least sizes for the selected UNET architecture [27].

A summary of the annotated data for training the DCNN is shown in Table 2.

Table 2. Annotated data types developed for training.

Datasets /Bridge	Image Resolution	RD				MD			MD			
		Images (32 × 32)	Total Pixels	Del.	Sound	Background	Del.	Sound	Images (32 × 32)	Total Pixels	Del.	Sound
FRNB *	1952 × 480	915	936,960	184,833	752,127	251,309	190,601	495,048	565	578,560	276,487	660,473
FRSB *	1632 × 416	663	678,912	135,597	543,315	183,487	139,790	355,633	411	420,864	202,864	476,048
PRMD *	4352 × 416	1768	1,819,432	354,409	1,456,023	496,993	366,114	947,324	1188	1,216,512	537,229	1,273,203
PRNB *	3008 × 384	1128	1,155,072	222,003	933,069	377,553	229,528	610,289	930	952,320	360,085	794,987
PRSB *	4192 × 480	1965	2,012,160	429,349	1,582,810	472,809	443,594	1,095,756	1036	1,060,084	593,603	1,418,557

* Forest River North Bound (FRNB); Forest River South Bound (FRSB); Park River Median (PRMD); Park River North Bound (PRNB); Park River South Bound (FRSB).

2.2. Model Selection and Training

Computations were performed on a desktop computer. The PC has a 64-bit operating system, 32 GB of memory, and a 3.80 GHz processor running an Intel ® Core™ i7-9800X CPU. The programming software for the operations and command was executed using MATLAB R2022.

Five models were developed from the UNET architecture. The encoder depth of the architecture was three (3), resulting in a total of 46 layers and 48 connections. At the onset of this study, the control model (CM) was first developed and trained with the base hyperparameters (Figure 1 and Table 3).

Table 3. Base learning hyperparameters for models.

Optimizer	Stochastic Gradient Descent with Momentum (sgdm)
InitialLearnRate	1×10^{-3}
Momentum	0.8
MaxEpochs	30
MiniBatchSize	100
LearnRateSchedule	piecewise
GradientThresholdMethod	l2norm
GradientThreshold	0.05
ValidationData	valid
ValidationFrequency	3
VerboseFrequency	3
Verbose	False

These hyper-tune parameters were selected after several tuning to yield optimal training and validation loss (Figure 3). There were 44 iterations per epoch, and the model was set to a maximum of 30 epochs, yielding 1320 iterations in total. The model converged at 80 iterations, and the training progress finally reached a training accuracy of 73% and a loss of about 0.67. The computation time was 9 min. 43 sec. The training and validation loss

were seen to converge before 10 epochs for the CM model (Figure 3). To assess the effect of augmentation and balancing on the dataset, four (4) models in addition were proposed and developed, keeping the base hyperparameters the same for all models. The models developed are (i) raw data (RD), unprocessed dataset (ii). Augmented level 1 models (AM1): translation of the data prior to training of the model (iii). Augmented level 2 models (AM2): translation, rotation, scale, and shear augmentation on the data prior to training of the model (iv). Balanced dataset model (BM): the unbalanced size of data is balanced prior to training (v). Augmented–balanced model (ABM): augmentation and balancing are both applied to the data prior to training. A total of 75 models were developed for the bridges, as shown in Table 4. The training and validation progress showing the training accuracy and loss function is shown in Figure 3.

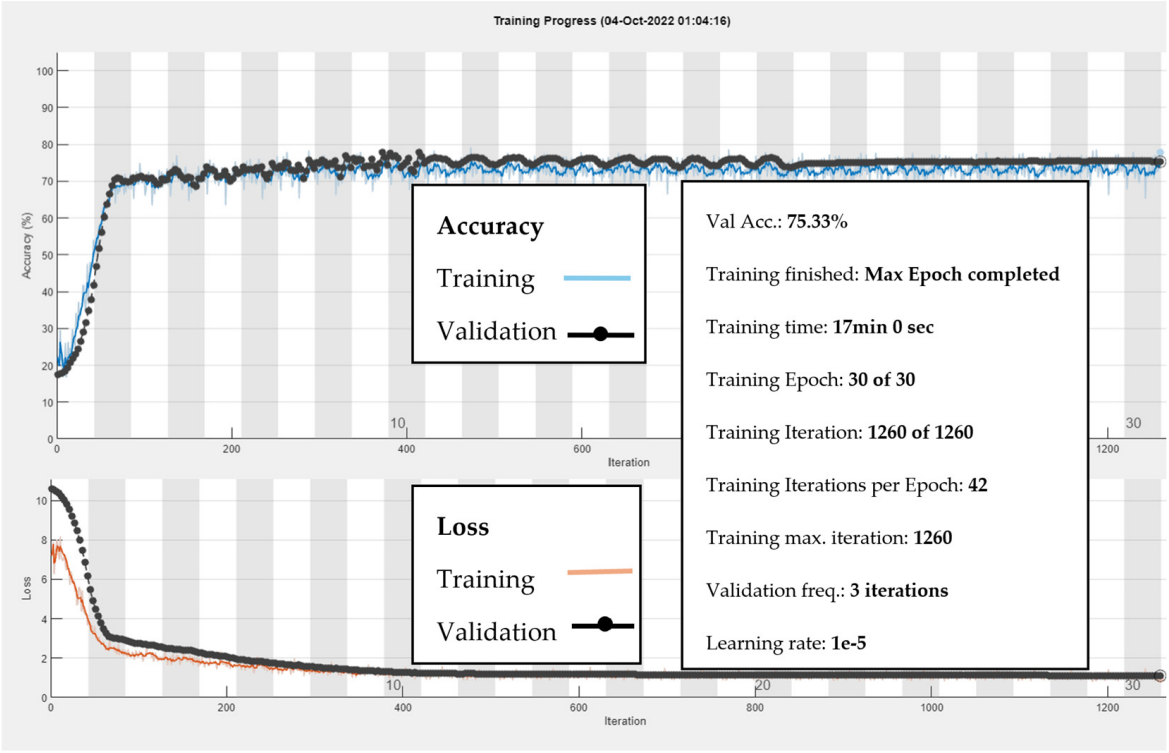


Figure 3. Training and validation—progress and parameters for FRNB MD preprocessed dataset.

Table 4. Training model matrix.

Preprocessing	Testing Bridge Name	Model Type				
		CM	AM1	AM2	BM	ABM
RD	FRNB	FN/R/D	FN/R/A	FN/R/A2	FN/R/B	FN/R/AB
	FRSB	FS/R/D	FS/R/A	FS/R/A2	FS/R/B	FS/R/AB
	PRMD	PM/R/D	PM/R/A	PM/R/A2	PM/R/B	PM/R/AB
	PRNB	PN/R/D	PN/R/A	PN/R/A2	PN/R/B	PN/R/AB
	PRSB	PS/R/D	PS/R/A	PS/R/A2	PS/R/B	PS/R/AB
BD	FRNB	FN/B/D	FN/B/A	FN/B/A2	FN/B/B	FN/B/AB
	FRSB	FS/B/D	FS/B/A	FS/B/A2	FS/B/B	FS/B/AB
	PRMD	PM/B/D	PM/B/A	PM/B/A2	PM/B/B	PM/B/AB
	PRNB	PN/B/D	PN/B/A	PN/B/A2	PN/B/B	PN/B/AB
	PRSB	PS/B/D	PS/B/A	PS/B/A2	PS/B/B	PS/B/AB

Table 4. *Cont.*

Preprocessing	Testing Bridge Name	Model Type				
		CM	AM1	AM2	BM	ABM
MD	FRNB	FN/M/D	FN/M/A	FN/M/A2	FN/M/B	FN/M/AB
	FRSB	FS/M/D	FS/M/A	FS/M/A2	FS/M/B	FS/M/AB
	PRMD	PM/M/D	PM/M/A	PM/M/A2	PM/M/B	PM/M/AB
	PRNB	PN/M/D	PN/M/A	PN/M/A2	PN/M/B	PN/M/AB
	PRSB	PS/M/D	PS/M/A	PS/M/A2	PS/M/B	PS/M/AB

Legend. Dataset: R = Raw, B = Background annotated, M = Manually cleaned dataset. Models: D = Base, A = Augmentation level 1, A2 = Augmentation level 2, B = Balanced, AB = Augmented and balanced models.

2.3. Data Balancing and Augmentation

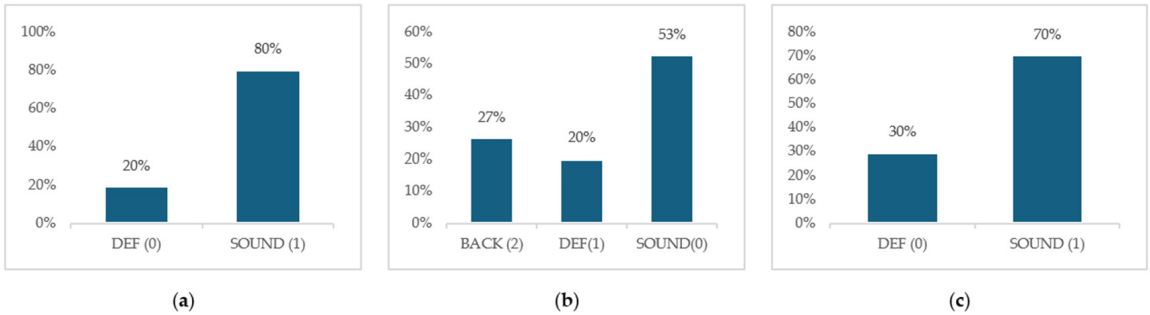
The preprocessing of the dataset before feeding into the model required that the images be normalized from the (0–255) uint eight image type to a grey image within the (0–1) range. Data augmentation was applied to the training dataset to evaluate the effect of augmentation on the model's performance. The augmentation adopted for the AM1 and AM2 models is shown in Table 5. The images were translated about the X- and Y-axis within the range [−10 10] pixels, causing the dataset to increase 20 times. The augmentation basis was from studies conducted by Nanni et al. [45]. The AM1 and AM2 models are compared with themselves to see the effects of scale, rotation, and shear on the performance of the model and thereafter compared to the control model without augmentation.

Table 5. Augmentation parameters for models.

Parameter	Axis	AM1 Models	AM2 Models
Translation (pixels)	X	[−10 10]	[−10 10]
Translation (pixels)	X	[−10 10]	[−10 10]
Rotation (degrees)			[−20 20]
Scale	X		[10 10]
Scale	Y		[10 10]
Shear	X		[10 10]
Shear	Y		[10 10]

An unbalanced dataset describes a dataset in which the classes of the training dataset are not equally represented. The defective pixels in the dataset adopted in this study are relatively smaller than the sound pixels. The loss function used in most semantic segmentation tasks represents the overall accuracy that is inappropriate for an unbalanced training dataset. Therefore, an unbalanced dataset may influence the model's results and bias towards the dominant class and still give a favorably high overall accuracy. Figure 4 shows the pixel distribution of FRNB datasets adopted in this study. In addressing this bias in classification, the class weighting factors are introduced to create a balanced dataset. In this case, the inverse of class frequency is applied [46]. Class weighting effectively balances classes when there are underrepresented classes in the training data. In this study, the RD and MD models' defect-to-sound pixels ratio was 1:4 and 3:7, respectively (Figure 4).

Similarly, the ratio of the defect-to-background-to-sound pixels is 2:3:5. The class weights in the classification layer for each class label, which is either one or none, were replaced with the inverse class frequency of the other class label. In balancing the RD dataset, the defected pixel class weight was replaced by the inverse of the class frequency of the sound pixels (1.33), while the sound pixel class weight was replaced by the inverse of the class frequency of the defected pixels (4.00). The total sum of the pixels remained the same.



(**a**) (**b**) (**c**)

Figure 4. Pixel distribution of FRNB for (**a**) raw annotated dataset (RD), (**b**) background annotated dataset (BD), and (**c**) manually cleaned annotated dataset (MD).

The images for each set of four (4) bridges set for training were split at every instance into five (5) k-folds and the fifth bridge was used for testing the model. Therefore, in every instance, 80% of the dataset for every four-bridge set is for the training and 20% for the validation set. The five bridges and k-folds were alternated for cross-validation to improve the model's effectiveness and reduce the possibility of overfitting.

2.4. Model Performance Evaluation Metrics

As mentioned earlier, the output of the classification is in a binary form but with three classes of annotation, where white segments of the images represent the delaminated pixels, the black regions represent the sound/non-delaminated, and the background pixels of the bridge decks. Several metrics can be used to evaluate the performance of the proposed architecture. The model's performance is determined by benchmarking the output predictions with the annotated ground truth pixel-wise.

The performance of the model was evaluated based on the following selected metrics: (i) Global and mean Accuracy (ACC), (ii) Precision/Positive Predictive value (PPV), (iii) F1-score, (iv) True positive rate (TPR)/Recall/Sensitivity, (v) True negative rate (TNR)/Specificity, (vi) mean Intersection of Union (IOU). The metrics are calculated by the following equations, as shown in Equations (2)–(8):

$$\textit{Global Acc.} = \mathrm{TP} + \mathrm{TN}/(\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}) \quad (2)$$

$$\textit{Mean Acc.} = [\mathrm{TP}/(\mathrm{TP} + \mathrm{FP})]/2 + [\mathrm{TN}/(\mathrm{TN} + \mathrm{FN})]/2 \quad (3)$$

$$\textit{Precision/PPV} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FP}) \quad (4)$$

$$\textit{F1-Score} = 2\mathrm{TP}/(2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}) \quad (5)$$

$$\textit{TPR/Recall} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN}) \quad (6)$$

$$\textit{TNR} = \mathrm{TN}/(\mathrm{TN} + \mathrm{FP}) \quad (7)$$

$$\textit{Mean IOU} = [\mathrm{TP}/(\mathrm{TP} + \mathrm{FP} + \mathrm{FN})]/2 + [\mathrm{TN}/(\mathrm{TN} + \mathrm{FP} + \mathrm{FN})]/2 \quad (8)$$

where TP refers to a true positive, which is when the model detects delamination correctly, FP refers to a false positive when defective pixels are falsely predicted as sound, TN refers to a true negative when defective pixels are correctly predicted as defective, and FN refers to a false negative when sound pixels are falsely predicted as defected pixels. IOU measures the percentage of the overlapped area between the predicted pixels and the actual pixels (ground truth) over their union. An IOU of 0 implies no overlapping, while a value of one means they are perfectly matched. The model's ACC indicates the correct detection rate with respect to the total number of detections. The PPV and TPR are often used to understand prediction outcomes further. PPV measures what fraction of

the detected delamination has been correctly detected given the ground truth, and TPR indicates what fraction of the real damages are correctly detected by the model among the actual defects [47]. Low PPV rates indicate many false defect detections, where many areas are incorrectly classified as delaminated. Models with high PPV results are preferred to minimize false defect detections. Low TPR rates indicate a high number of false negatives or where many of the existing delaminations were missed.

3. Results and Discussion

The results of the U-Net model developed for the semantic segmentation of the bridge decks are evaluated and discussed in this section. They are presented and discussed for each bridge based on the data annotation and preprocessing categories: (i) RD, (ii) BD, and (iii) MD.

3.1. Raw Annotated Data (RD)

The summary of the results for the five sets of bridges for RD is presented and discussed. The models developed by training the RD dataset for the FRNB showed that the precision values for the balanced models ABM and BM were 0.631 and 0.667, while that of the CM, AM1, and AM2 models were 0.269, 0.264, and 0.282, respectively. The precision increased by an average of 80% when the pixels were balanced (Figure 5). The least performing metrics are seen in the unbalanced models (CM, AM1, and AM2) except for GA and mean IOU. The BM model was seen to have better performance metrics than the ABM model, except for GA metrics.

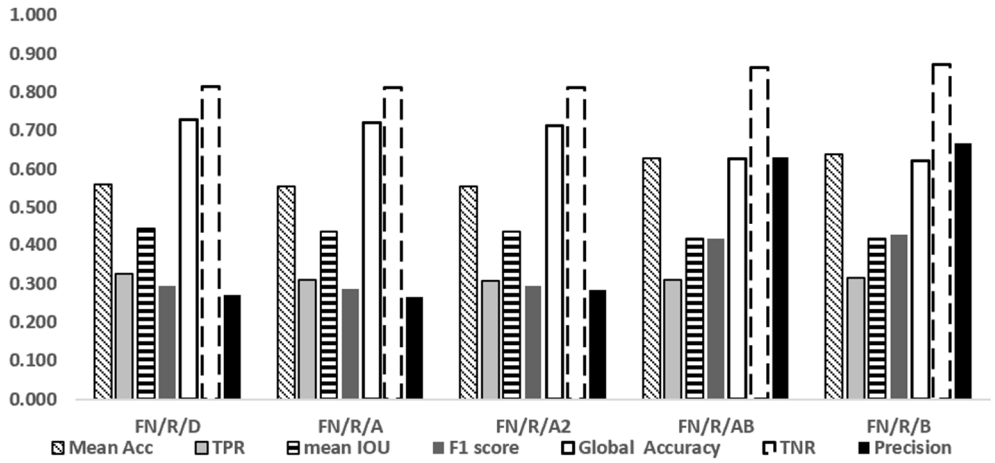


Figure 5. Evaluation metrics for FRNB for RD.

The F1 score for ABM and BM (balanced) models is 0.416 and 0.427, while that of the augmented models (AM1 and AM2) is 0.284 and 0.294, respectively. The F1 score of the CM is 0.295. The F1 score increased by an average of 47% after balancing the pixels, while it remained almost the same for the augmented models. The average GA for AM1 and AM2 is 0.720, while that of the BM and ABM is 0.623. Balancing the dataset caused a drop in the GA by 15%. Contrarily, by computing the mean accuracies (MA), the average values of the ABM and BM models increased by 14% compared to the AM1 and AM2 models. The TNR for the balanced models is the highest, with values of 0.872 and 0.863, respectively, compared to the base and augmented models, which have values of 0.813, 0.810, and 0.811, respectively. There was no significant change in the TNR values for the CM and the augmented models.

Considering the RD dataset for FRSB, the CM model showed the least performance considering precision, TNR, and F1 score (Figure 6).

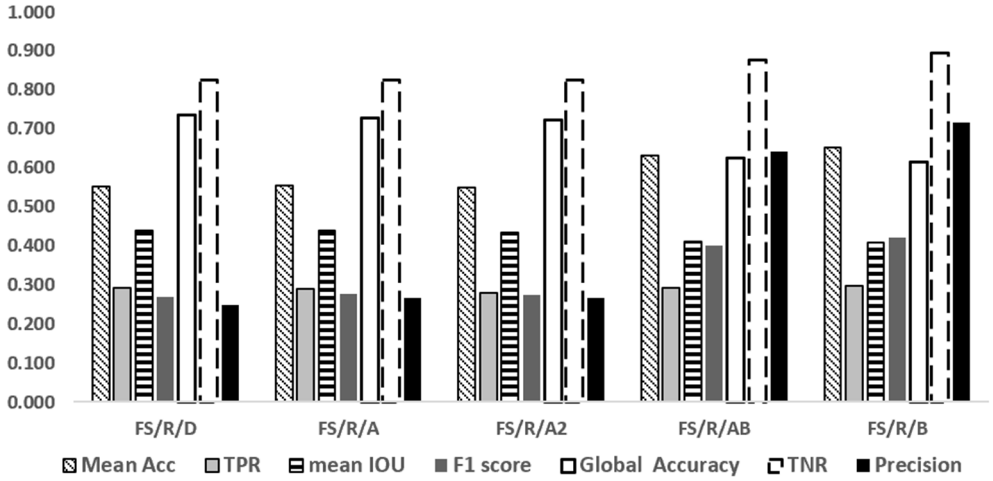


Figure 6. Evaluation metrics for FRSB for RD.

In contrast, AM1 and AM2 models showed the least performance for MA, TPR, and TNR metrics except for GA, precision, mean IOU, and F1-score. The results depicted that the balanced models, BM and ABM, developed from the RD increased the model's performance. For instance, the MA, precision, TPR, TNR, and F1-score for the balanced models increased by an average of 16%, 175%, 5%, 8%, and 53%, respectively, when compared to the least performing augmented model. CM had a higher GA value of 0.736 and outperformed the balanced dataset by 20% (Figure 6). This is contrary to when MA is considered. The MA of the balanced models outperformed the CM, AM1, and AM2 models by 19%. This shows that MA may be a preferred evaluation metric over GA. The GA does not account for the effect of the biased nature of the unbalanced dataset. The preference of MA over precision for a balanced dataset is due to the balanced models outperforming other models for most evaluation metrics. The scale, shear, and rotation augmentation parameters for AM2 had to significant effect on the performance of the models.

The BM and ABM models developed from the RD for the PRMD bridge showed a significant increase in the MA, precision, TPR, TNR, and F1 scores. In comparison to the least performing augmented model, results showed that MA, precision, TPR, TNR, and F1-score increased by an average of 17%, 143%, 11%, 8%, and 52%, respectively. Conversely, the AM1 models performed the least for all metrics except GA and mean IOU.

The models developed by training the RD dataset for the PRNB bridge show that the MA, precision, TPR, TNR, and F1-score of the balanced models increased by an average of 3.6%, 135%, 1%, 3%, and 37% compared to the least performing augmented (AM1) model (Figure 7).

In addition, the precision increased from 0.281 for the CM model to 0.674 and 0.261 for the AM1 model to 0.362 for the BM models. The maximum GA, MA, precision, TPR, TNR, and FI scores are 0.644 and 0.526, 0.674, 0.257, 0.793, and 0.362, respectively. The models developed by training the RD dataset for the PRSB bridge show that the MA, precision, TPR, TNR, and F1-score of the BM model increased by 17%, 150%, 3.5%, 5.5%, and 50%, respectively, when compared to the least performing augmented model. The results showed that MA increased from 0.572 for the AM1 to 0.655 for the BM model, precision increased from 0.239 for CM to 0.605 for the BM model, and F1-score increased from 0.261 for CM to 0.393 for the BM model. Considering GA metrics, CM shows the best performance value of 0.773.

Furthermore, the confusion matrix showing the outcome of the pixel distribution after classification for the FRNB raw annotated dataset is depicted in Figure 8. The effect of augmenting the model had caused no significant change. The FP and TN only increased and reduced by 1%, respectively. In contrast, the effect of balancing only caused a substantial change in the evaluation parameters. The TP increased by 6%, while the TN reduced by

19%. Similarly, FP reduced by 8% while the FN increased by 19%. Augmenting and then balancing the dataset had an insignificant effect on the balanced-only outcome.

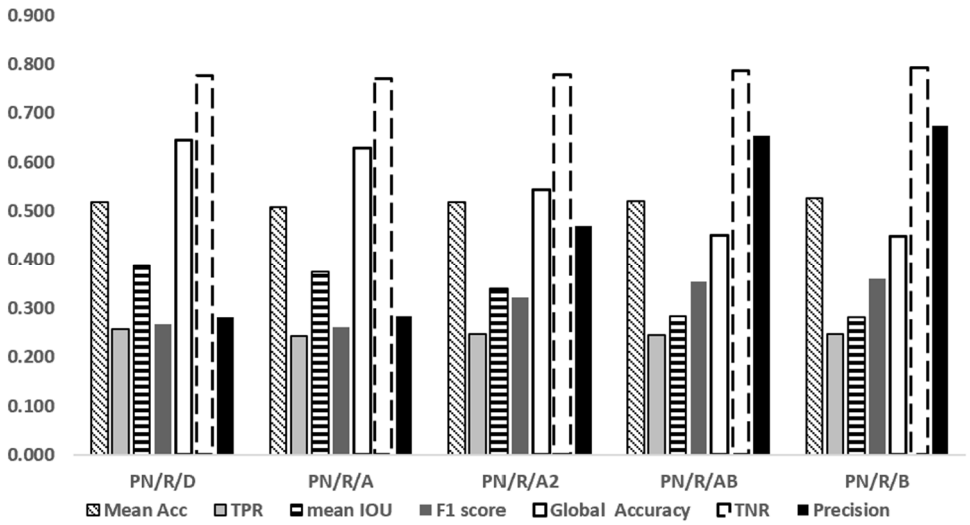


Figure 7. Evaluation metrics for PRNB for RD.

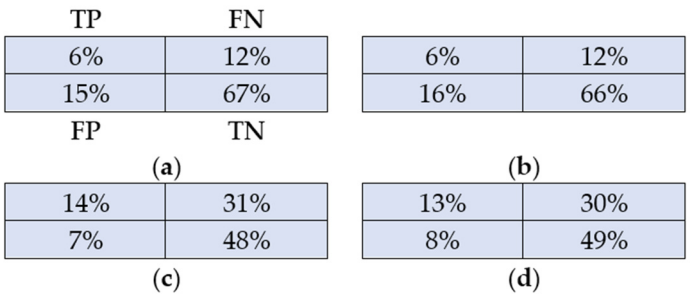


Figure 8. Confusion matrix for raw annotated dataset for FRNB for (**a**) Base, (**b**)Augmented, (**c**) Balanced, and (**d**) Augmented–Balanced models.

The evaluation metrics outcome shows that augmentation reduced the performance of the models while balancing the dataset alone significantly improved the model's performance. This corroborated studies by Cheng et al. [24], where IOU dropped after augmentation for the testing dataset. Lastly, the GA, precision, TNR, F1-score, and TPR are highest for all models developed with the BD dataset.

3.2. Background Annotated Data (BD)

The models developed by training the BD dataset for the FRNB showed that the average TNR value of 0.693 for the balanced models is lower than the average value of 0.842 for the base and augmented models. The F1-score increased by 36% from 0.291 for the augmented models to an average of 0.397 for the balanced models. There was no significant change in the precision and MA before and after data balancing, whereas the GA dropped by 8% for the augmented models compared to the balanced models from 0.693 to 0.636 (Figure 9).

Similarly, the least performing metrics were seen in the augmented models except for GA and mean IOU, while the balanced models showed higher performance except for the GA, precision, mean IOU, and TNR. The TPR for the balanced models is 110% higher than the unbalanced models, which have TPR values of 0.266 and 0.268, respectively.

The balanced models, BM and ABM, developed from the BD for the FRSB dataset outperformed the base and augmented models; CM, AM1, and AM2 when considering performance metrics; and MA, precision, TPR, and F1-score. These metrics increased on

average by 2.5%, 3.4%, 100%, and 38% compared to the least-performing augmented model (Figure 10).

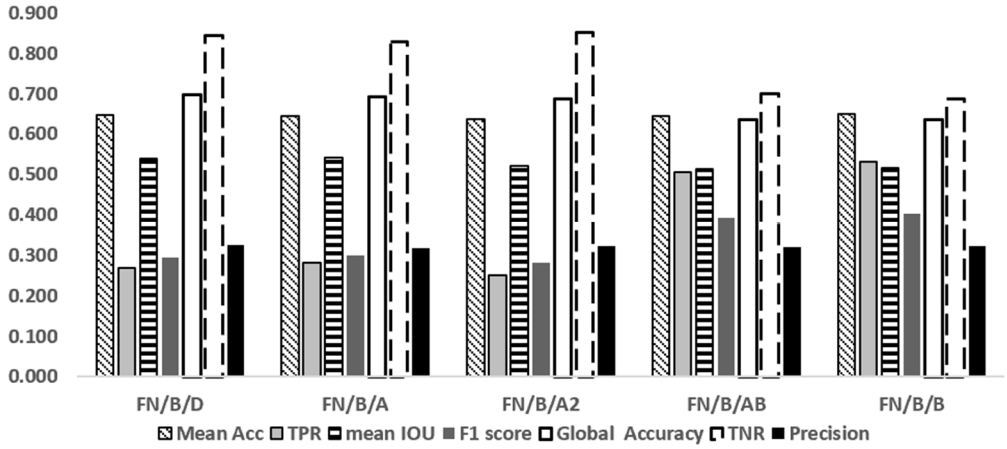


Figure 9. Evaluation metrics for FRNB for BD.

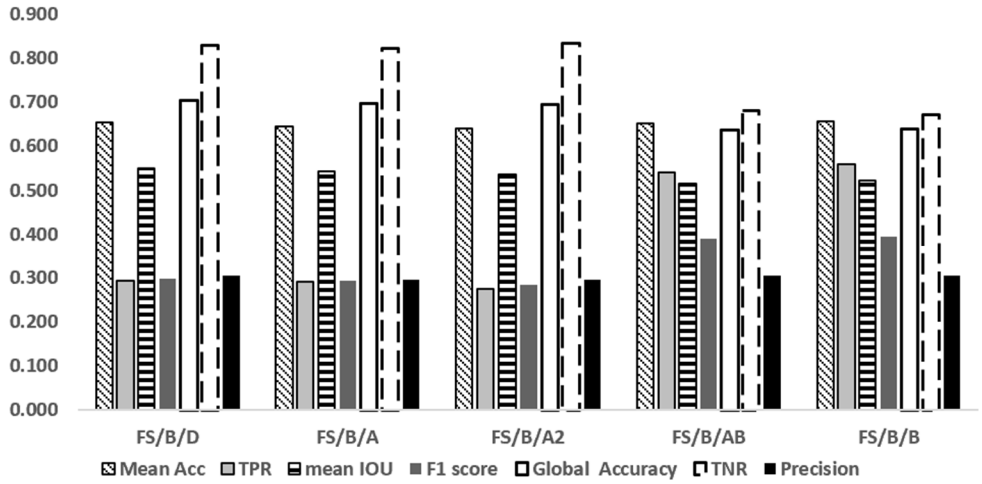


Figure 10. Evaluation metrics for FRSB for BD.

Conversely, the base and augmented models show lower performance for all metrics except for GA and mean IOU. This is a similar trend and pattern for all bridges. This further corroborates the preference of MA for performance evaluation instead of GA. Furthermore, the precision for both CM and balanced models showed the same value of 0.304, while that of the augmented models was 0.294. This implies that balancing the dataset did not affect the precision of the model for the FRSB.

Evaluating the performance metrics of the models developed from the BD dataset for PRMD, the results show that the TPR was almost doubled from 0.265 for the least performing augmented model (AM1) to 0.515 for the BM model. Balancing the dataset significantly improved the performance of the models. Similarly, the F1 score increased by an average of 40% after balancing the dataset. The TNR increased by 2% for the BM model compared to the ABM model. The GA for the CM, AM1, and AM2 increased by 12%, 15%, and 8% compared to the least-performing balanced ABM model.

The results of the models developed by training the BD dataset for the PRNB bridge show that the TPR, TNR, and F1-score of the balanced models increased by an average of 70% and 23%, respectively, compared to the least performing augmented (AM2) model (Figure 11).

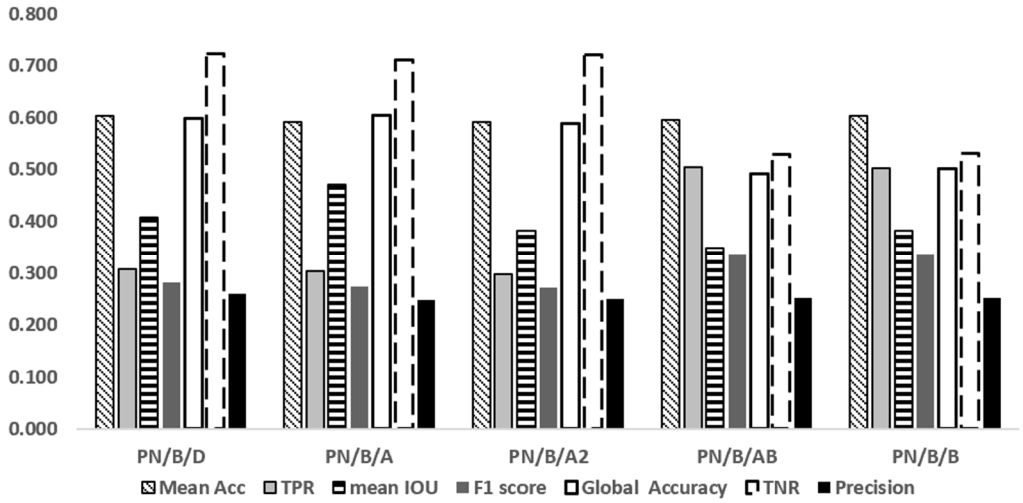


Figure 11. Evaluation metrics for PRNB for BD.

In addition, the TPR increased from 0.297 for the AM2 to 0.504 for the BM model, while the F1-score increased from 0.272 for AM2 to 0.336 for the ABM model. The maximum TPR, TNR, and F1 scores are 0.504, 0.724, and 0.336, respectively.

The models developed by training the BD dataset for the PRSB bridge show that the MA, precision, TPR, and F1-score for the BM model increased by 2.5%, 4%, 87%, and 34%, respectively, when compared to the least performing augmented model. The results showed that MA increased from 0.604 for AM2 to 0.622 for the BM model, TPR increased from 0.269 for CM to 0.499 for the BM model, and F1-score increased from 0.280 for CM to 0.374 for the ABM model. Considering GA metrics, AM1 shows the best performance value of 0.680.

The confusion matrix for the FRNB background annotated dataset is depicted in Figure 12. Similarly, augmentation of the data caused no significant change in the values of the metrics shown in the confusion matrix. The figure depicts the same trend as the previous; while the TP increased by 6%, the TN reduced by 12%. Consequently, while the FN reduced by 10%, the FP increased by 12%. This further confirms that balancing had a significant effect on the outcome of the model when compared with augmentation.

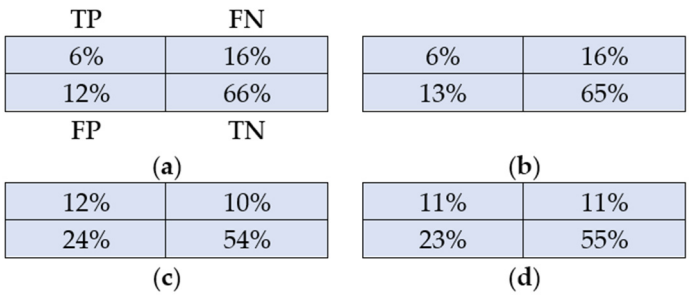


Figure 12. Confusion matrix for background annotated dataset for FRNB for (**a**) Base, (**b**)Augmented, (**c**) Balanced, and (**d**) Augmented–Balanced models.

The background annotated dataset (BD) results showed significant improvement in the TPR compared to other models developed from RD and MD datasets. The annotation of the BD in three classes distinguishes the background pixels from defective pixels. This may have improved the TPR of the model compared to others. However, the precision for all models remained almost the same without significant improvement. The MA and TPR are the metrics with the highest values for all models developed with the BD dataset.

3.3. Manually Cleaned Annotated Data (MD)

BM and ABM models developed from MD for FRNB show the least performance for all metrics except for precision and F1-score (Figure 13). The models developed from the MD had the lowest performance in comparison to the models developed from the RD and BD. For instance, the average values of the GA, MA, precision, TPR, TNR, and mean IOU for the BD reduced by 20%, 23%, and 16%, 11%, 19%, 34%, and 3%, respectively, when compared to the average values for best-performing models developed from the RD or BD dataset.

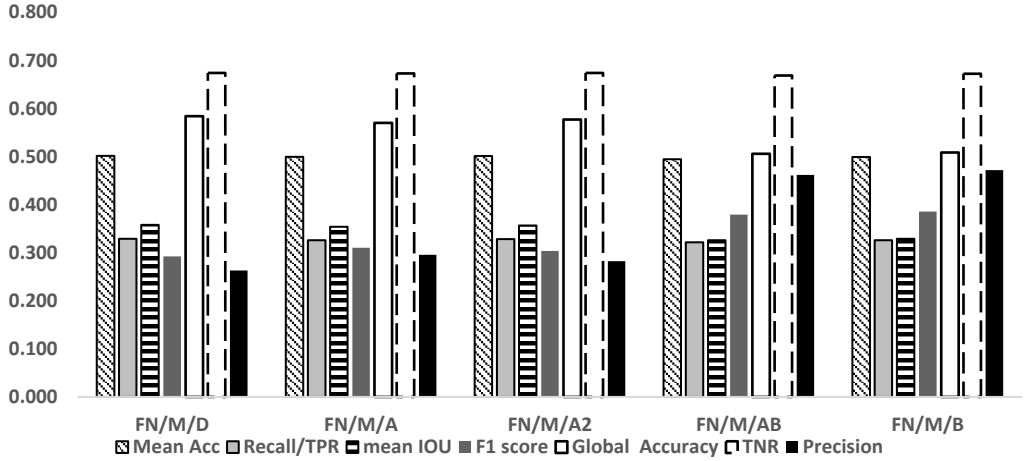


Figure 13. Evaluation metrics for FRNB for manually cleaned annotated data.

Considering the MD dataset for FRSB, a similar trend was observed (Figure 14). The augmented models showed lower performance considering all metrics except for GA and mean IOU. The ABM model showed the least GA and the mean IOU values. The effect of balancing was minimal when considering the MA metrics showing an increase of 1%. The effect of balancing was most prominent in the precision and F1 score metrics. The precision and F1 scores increased by 79% and 29%, respectively, compared to the least-performing augmented models. The models developed from the MD dataset for the FRSB showed the lowest average performance compared to the RD and BD models.

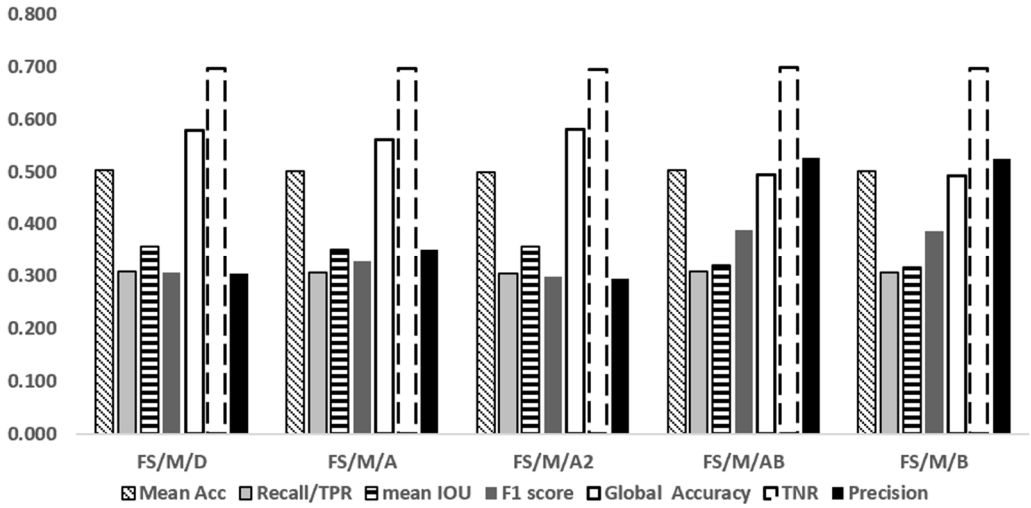


Figure 14. Evaluation metrics for FRSB for manually cleaned annotated data.

Evaluating the performance metrics of the models developed from the MD dataset for the PRMD, results show that the MA, precision, TPR, TNR, and F1-score of the BM and ABM models increased by an average of 3%, 90%, 4%, 2%, and 34% compared to the least

performing augmented (AM1) model. In addition, the precision increased from 0.27 for the CM model to 0.511 for the balanced models. The maximum GA, MA, precision, TPR, TNR, and FI scores are 0.597 and 0.513, 0.511, 0.317, 0.705, and 0.391, respectively.

Results for the MD-trained models for PRNB showed that the MA, TPR, TNR, and F1-score of the BM and ABM models increased by an average of 2%, 1.5%, 1%, and 25%, respectively, when compared to the least performing augmented model (Figure 15). However, the augmented models had better performance metrics than the balanced models for GA and TPR.

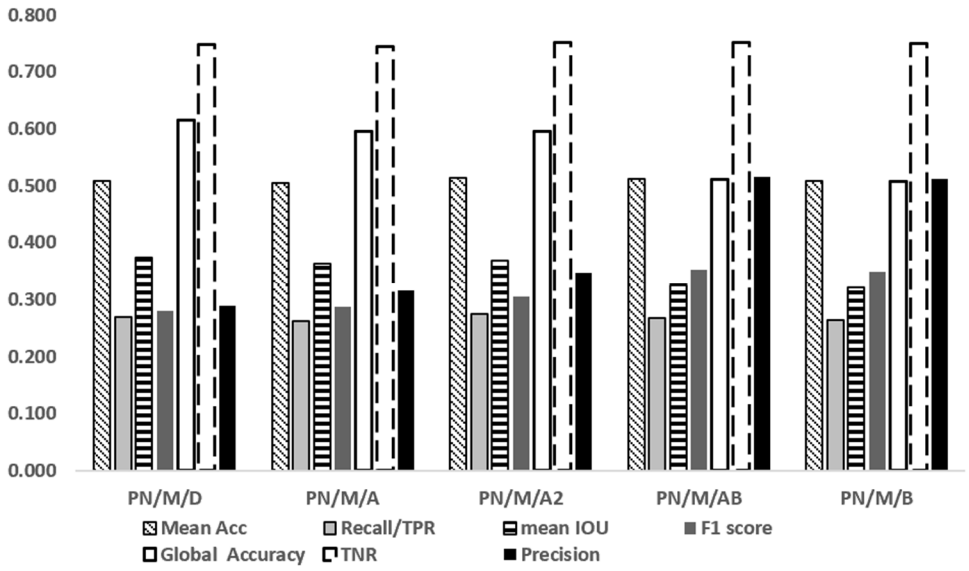


Figure 15. Evaluation metrics for PRNB for manually cleaned annotated data.

The models developed by training the MD dataset for the PRSB bridge showed that the MA, precision, TPR, and F1-score for the BM model increased average by 1.0%, 91%, 2%, and 34%, respectively, when compared to the least performing augmented model. The F1-score for the BD datasets had the same rate of increment. The results showed that precision increased from 0.285 for CM to 0.552 for the ABM model, and the F1-score increased from 0.299 for CM to 0.397 for the BM model. Considering GA metrics, the AM1 model showed the best performance value of 0.680.

Similarly, the confusion matrix shown in Figure 16 further depicts the significant effect of balancing the dataset, unlike augmentation, which had no significant contribution. The TP and TN were reduced when compared to the performance of the models developed from the MD and RD. This may be attributed to a reduction in the data size after removing the images with artifacts background from the dataset.

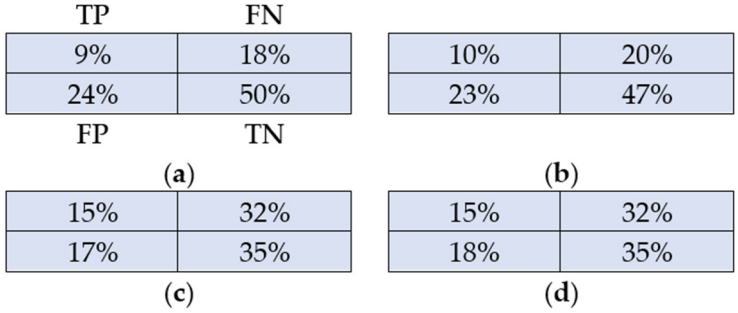


Figure 16. Confusion matrix for manually cleaned annotated dataset for FRNB for (**a**) Base, (**b**) Augmented, (**c**) Balanced, and (**d**) Augmented–Balanced models.

The models developed from the manually annotated dataset (MD) performed less. This may have been caused by significantly reducing the datasets by manually removing the images with artifacts and background noises.

The summary of the performance metrics showing the minimum and maximum values for the developed models is shown in Table 6. The RD models for all the bridges show the highest GA values ranging from 0.644 to 0.773, while the least-performing model was the PN/R/B with a value of 0.447. Contrarily, for the MA metrics, the balanced models showed the best performance ranging from 0.603 to 0.657, and the least performance value of 0.494 for the ABM model. Evaluating the models with the precision metrics, the BM trained with the RD dataset showed the best performance for all bridges with the exception of the PS/R/D model. The balanced models trained with the BD dataset had the highest TPR values ranging from 0.504 to 0.558, while the augmented models showed the least performance values. Similarly, the balanced models trained with the RD showed the best performance for the TNR and F1-score with the exception of the PS/R/D model.

Table 6. Summary of minimum and maximum performance metrics.

	Bridge	Global Accuracy	Bridge	Mean Acc	Bridge	Precision	Bridge	TPR	Bridge	TNR	Bridge	F1 Score
FRNB	FN/R/D	0.728	FN/B/B	0.650	FN/R/B	0.667	FN/B/B	0.532	FN/R/B	0.872	FN/R/B	0.427
	FN/M/AB	0.506	FN/M/AB	0.494	FN/M/D	0.263	FN/B/Aa	0.250	FN/M/AB	0.669	FN/B/Aa	0.282
FRSB	FS/R/D	0.736	FS/B/B	0.657	FS/R/B	0.716	FS/B/B	0.558	FS/R/B	0.895	FS/R/B	0.420
	FS/M/B	0.491	FS/M/Aa	0.500	FS/R/D	0.247	FS/B/Aa	0.273	FS/B/B	0.671	FS/R/D	0.268
PRMD	PM/R/D	0.719	PM/B/A	0.641	PM/R/B	0.672	PM/B/B	0.515	PM/R/B	0.876	PM/R/B	0.427
	PM/M/AB	0.506	PM/M/A	0.496	PM/M/Aa	0.270	PM/B/A	0.265	PM/B/AB	0.675	PM/B/A	0.273
PRNB	PN/R/D	0.644	PN/B/B	0.603	PN/R/B	0.674	PN/B/AB	0.504	PN/R/B	0.793	PN/R/B	0.362
	PN/R/B	0.447	PN/M/A	0.504	PN/B/A	0.248	PN/R/A	0.243	PN/B/AB	0.529	PN/R/A	0.261
PRSB	PS/R/D	0.773	PS/R/B	0.655	PS/R/D	0.239	PS/B/AB	0.509	PS/R/B	0.899	PS/M/AB	0.401
	PS/M/AB	0.485	PS/M/A	0.497	PS/R/B	0.605	PS/B/D	0.269	PS/M/A	0.685	PS/R/D	0.261
	Maximum											
	Minimum											

In general, the results for FRNB for the evaluated datasets are presented in Table 7 with a highlight of the highest and lowest performance metrics. The results depicted that the FR/R/D model had the lowest performance for one-out-of-six metrics, whereas the F/R/B model showed the highest performance for three-out-of-six metrics, and FM/M/AB showed three-out-of-six lowest metrics. This suggests that the model trained with background annotated and balanced pixels for the FRNB bridge gave the best performance, while the model trained with MD had the worst performance. This is expected as the number of original pixels had dropped by 38% after manually removing the artifacts and background blocks. Similarly, the other four bridges show the same pattern where the model trained with the RD and balanced pixels depicted the best performance metrics of at least three out of six of the highest metrics, and the worst models are the MD-trained models. If the GA were to be used as the presiding metric for evaluating models' performance, then the models trained with the RD with base hyperparameters showed the highest performance. The augmentation of the pixels had a downtrend effect on the performance of the models. This suggests that local augmentation, as adopted in the models, had no upside effect on the model's performance. This may imply that the pixel-wise augmentation adopted is not suitable and robust for the semantic segmentation of IRT images.

3.4. Semantic Segmentation Maps

Results for the semantic segmentation for the RD annotated dataset for FRNB are presented in Figure 17, showing the ground truth (Figure 17a,b) and the predicted outcome (Figure 17c–g) of the test bridge image after training. The output of the individual image blocks was combined to make a complete map for visualization without any further image processing. The probability outcome of the delamination detection was presented in a heat map to show the progress and gradient of delamination over the bridge deck area from

sound (zero) to severe delaminated (one) and mid-delamination level (0.5), represented by blue, red, and green colors on the map.

Table 7. Performance metrics for FRNB models (Highest and lowest values in *italics*).

	Global Acc.	Mean Acc.	Precision	TPR	TNR	F1 Score
FN/R/D	*0.728*	0.560	0.269	0.326	0.813	0.295
FN/R/A	0.720	0.553	0.264	0.311	0.810	0.285
FN/R/A2	0.713	0.555	0.282	0.306	0.811	0.294
FN/R/AB	0.625	0.628	0.631	0.311	0.863	0.416
FN/R/B	0.621	0.638	*0.667*	0.314	*0.872*	*0.427*
FN/B/D	0.697	0.648	0.325	0.268	0.844	0.294
FN/B/A	0.693	0.645	0.318	0.282	0.829	0.299
FN/B/A2	0.688	0.636	0.322	*0.250*	0.852	*0.282*
FN/B/AB	0.636	0.644	0.321	0.505	0.699	0.392
FN/B/B	0.636	*0.650*	0.324	*0.532*	0.687	0.402
FN/M/D	0.584	0.501	*0.263*	0.329	0.674	0.292
FN/M/A	0.570	0.499	0.296	0.326	0.673	0.310
FN/M/A2	0.577	0.501	0.282	0.328	0.674	0.304
FN/M/AB	*0.506*	*0.494*	0.462	0.322	*0.669*	0.379
FN/M/B	0.508	0.499	0.472	0.326	0.672	0.385

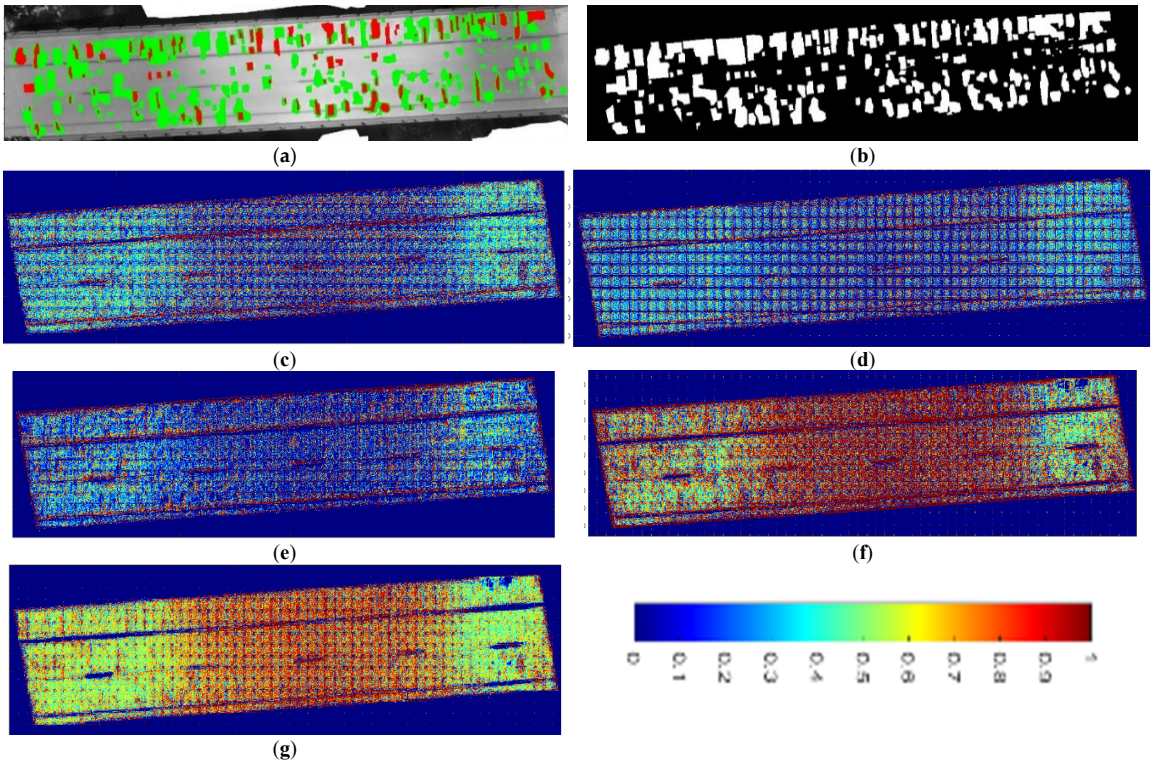


Figure 17. Semantic segmented heat maps for RD annotated dataset for (**a**) image ground truth, (**b**) binarized ground truth, (**c**) CM model, (**d**) AM1 model, (**e**) AM2 model, (**f**) ABM model, (**g**) BM model.

The output for the CM, AM1, and AM2 models barely distinguished between sound and delaminated regions. The maps generated from the BM and ABM models, compared

to the ground truth (Figure 17f,g), show the defective pixels concentrated around the center of the maps. This may not have depicted the exactness of the ground truth map, but there is an indication that balancing the data significantly improved the model's performance and the semantic segmentation output, as presented in the earlier discussion. The BM output, in comparison to the ABM model, showed a clearer distinction in the prediction and segmentation of sound and delaminated pixels. This segmentation map output justifies the preference of the FN/R/B model based on its performance metrics, as shown in Table 6. This also further suggests that GA may not be an appropriate metric for evaluating the performance of a model. The maps clearly show that the model distinctly and successfully segmented the background from the deck's boundary but would need further studies to improve the segmentation output map for the delamination.

A summary of the results of this study compared to earlier studies using the iterative image-based method [13], carried out on the same dataset, is shown in Table 8. The condition map from past studies is presented in Figure 18. GA was used to choose the best model for comparative analysis with previous studies by Ichi et al. [14]. The table shows that the uNet had better metrics than the iterative image-based method. Contrarily, the image-based method shows a better visual condition assessment map. This suggests that combining deep learning with image processing-based models for delamination assessment has prospects of yielding improved model and prediction outcomes.

Table 8. Comparison of image-based [14] against UNET model.

Bridge	FRNB			FRSB			PRMD			PRNB			PRSB		
Model	**Image-Based**	**uNet**	**Increase (%)**	**Image-Based**	**uNet**	**Increase (%)**	**Image-Based**	**uNet**	**Increase (%)**	**Image-Based**	**uNet**	**Increase (%)**	**Image-Based**	**uNet**	**Increase (%)**
GA	0.69	0.728	5.4	0.716	0.736	2.8	0.655	0.719	9.8	0.696	0.644	−7.5	0.731	0.773	5.8
F1-sc.	0.26	0.295	13.5	0.099	0.268	170.7	0.131	0.290	121.4	0.287	0.268	−6.6	0.148	0.261	76.4
TPR	0.248	0.326	31.5	0.077	0.292	279.2	0.123	0.307	149.6	0.254	0.257	1.2	0.136	0.286	110.3
TNR	0.816	0.813	0.0	0.879	0.824	−6.3	0.799	0.814	1.9	0.836	0.776	−7.2	0.853	0.852	0.0

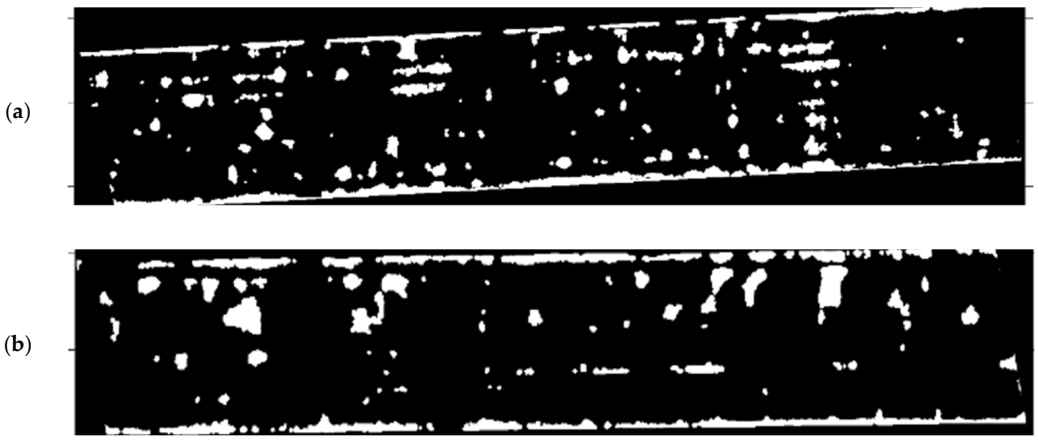


Figure 18. Adaptive image-based model for delamination detection for (**a**) FRNB and (**b**) FRSB bridges [14].

3.5. Limitations of Study

This study has shown that autonomous delamination detection in bridge decks using DCNN models that are trained, validated, and tested with publicly available IRT datasets is saddled with some challenges and limitations. The reliability and performance of the models to correctly predict delamination depends largely on several factors. The final condition map shows that an appropriate model, image processing and quality, hyper-tuning parameters, and others would largely determine the performance of the developed model and segmentation map.

DCNNs are required to be developed with large datasets for training and validation. Previous studies have only centered on developing and training models with laboratory model samples. This study adopted only datasets from in-service bridges, without any laboratory datasets. This invariably presented limited datasets for study. The TPR of the model is still seen to be lower than 35% for the FRNB bridge deck. This implies that less than this proportion is detected as defective by the model, resulting in higher false positives (FP). This may have been due to the quality of the image. Ichi et al. [14] highlighted the impact of image quality on the performance of the model. In addition, the feature extraction layer may not have extracted very prominent features for classification. This may be due to fewer features in thermal images when compared to visual images, which contain prominent and notable textural features for classification. Future studies will also include hyper-tuning parameters and modifying the models' architecture by adopting the backbone of other architecture.

4. Conclusions

This paper presents the findings of the study on sub-surface delamination detection in bridge decks. UNET deep convolutional neural network architecture was adopted. The network was trained with a first-of-its-kind publicly available bridge dataset SDNET2021. This study largely differs from previous studies where the investigation of delamination is carried out on mostly laboratory samples. In addition, the effect of different annotations and preprocessing approaches was assessed. The results were compared with iterative image-based methods conducted on the same dataset. The outcome of the investigation showed prospects in delamination detection. The summary of our findings is hereby highlighted herein:

i. Preprocessing and annotation approach adopted before model development had a significant effect on the outcome of the results.
ii The model trained with the raw annotated dataset (RD) with base hyper-tune parameters without pixel balancing and augmentation showed the best performance for all the bridges. This is the case when the global accuracy (GA) was used for the evaluation of the model's performance. The GA ranged from the least 0.644 for Park River North Bound (PRNB) to the highest value of 0.773 for Park River South Bound (PRSB).
iii The model trained with the raw annotated dataset (RD) with base hyper-tune parameters and pixel balancing showed the best performance for all the bridges. This is considered when at least three-out-of-six performance metrics were highest. All the bridges had the highest metrics for precision, true negative rate (TNR), and F1 score.
iv A combination of multiple metrics is more adequate for model evaluation. In this study, a combination of precision, true negative rate (TNR), and F1 score with the semantic segmentation maps were used to evaluate the models' performance and select the optimal model.
v The models trained with the balanced datasets showed distinctive segmentation of the background, sound, and delaminated pixels compared to other models trained with manually cleaned and background annotated datasets (MD and BD).
vi The scale, shear, and rotation augmentation parameters had no significant effect on the performance of the models. Considering the models being balanced alone or with augmentation, there was a performance improvement based on precision, TPR, and FI scores.

Further studies are required to achieve improved performance metrics and segmentation maps. Several models will be adopted to compare the results and outputs of the model to determine the most optimal model for delamination detection. It is important to assess the effect of image size and resolution on the performance of the models. It is noteworthy that the effect of reduction in image quality should be evaluated and investigated in future studies. A higher super-resolution quality image may improve the model's performance. Future studies should adopt image-based methods and deep learning for improved results

and output. In addition, since hyperspectral imagery combines spectral and spatial features over the visual and IR spectrum, this will undoubtedly yield better model performance and will be considered in future studies.

Author Contributions: S.D.: Conceptualization, Methodology, Acquired the Research Funding, and Supervised Preparation of First Draft. E.I.: Methodology, Software, Data Collection and Curation, Visualization, Writing—Original Draft, Data Quality, Annotation, and Validation. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data are available at the link: https://commons.und.edu/data/19. (Last accessed on 17 January 2023).

Acknowledgments: The authors would like to acknowledge Anna Crowell for assisting in editorial works.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Lehman, M. The American Society of Civil Engineers' Report Card on America's Infrastructure. In *Women in Infrastructure*; Springer International Publishing: Cham, Germany, 2022; pp. 5–21. [CrossRef]
2. Graybeal, B.A.; Phares, B.M.; Rolander, D.D.; Moore, M.; Washer, G. Visual inspection of highway bridges. *J. Nondestruct. Eval.* **2002**, *21*, 67–83. [CrossRef]
3. Dorafshan, S.; Maguire, M. Bridge inspection: Human performance, unmanned aerial systems, and automation. *J. Civ. Struct. Health Monit.* **2018**, *8*, 443–476. [CrossRef]
4. Dorafshan, S.; Thomas, R.J.; Maguire, M. Comparison of deep convolutional neural networks and edge detectors for image-based crack detection in concrete. *Constr. Build. Mater.* **2018**, *186*, 1031–1045. [CrossRef]
5. Rens, K.L.; Wipf, T.J.; Klaiber, F.W. Review of nondestructive evaluation techniques of civil infrastructure. *J. Perform. Constr. Facil.* **1997**, *11*, 152–160. [CrossRef]
6. Gucunski, N.; Kee, S.-H.; La, H.; Basily, B.; Maher, A. Delamination and concrete quality assessment of concrete bridge decks using a fully autonomous RABIT platform. *Struct. Monit. Maint.* **2015**, *2*, 19–34. [CrossRef]
7. Gucunski, N.; Maher, A.; Ghasemi, H. Condition assessment of concrete bridge decks using a fully autonomous robotic NDE platform. *Bridge Struct.* **2013**, *9*, 123–130. [CrossRef]
8. Oh, T.; Kee, S.-H.; Arndt, R.W.; Popovics, J.S.; Zhu, J. Comparison of NDT methods for assessment of a concrete bridge deck. *J. Eng. Mech.* **2013**, *139*, 305–314. [CrossRef]
9. Rathod, H.; Gupta, R. Sub-surface simulated damage detection using Non-Destructive Testing Techniques in reinforced-concrete slabs. *Constr. Build. Mater.* **2019**, *215*, 754–764. [CrossRef]
10. Yehia, S.; Abudayyeh, O.; Nabulsi, S.; Abdelqader, I. Detection of common defects in concrete bridge decks using nondestructive evaluation Techniques. *J. Bridge Eng.* **2007**, *12*, 215–225. [CrossRef]
11. Hiasa, S.; Birgul, R.; Catbas, F.N. Infrared thermography for civil structural assessment: Demonstrations with laboratory and field studies. *J. Civ. Struct. Health Monit.* **2016**, *6*, 619–636. [CrossRef]
12. Das, A.; Ichi, E.; Dorafshan, S. Image-Based Corrosion Detection in Ancillary Structures. *Infrastructures* **2023**, *8*, 66. [CrossRef]
13. Ichi, E.; Jafari, F.; Dorafshan, S. SDNET2021: Annotated NDE Dataset for Subsurface Structural Defects Detection in Concrete Bridge Decks. *Infrastructures* **2022**, *7*, 107. [CrossRef]
14. Ichi, E.; Dorafshan, S. Effectiveness of infrared thermography for delamination detection in reinforced concrete bridge decks. *Autom. Constr.* **2022**, *142*, 104523. [CrossRef]
15. Sony, S.; LaVenture, S.; Sadhu, A. A literature review of next-generation smart sensing technology in structural health monitoring. *Struct. Control Health Monit.* **2019**, *26*, e2321. [CrossRef]
16. Rakha, T.; Gorodetsky, A. Review of Unmanned Aerial System (UAS) applications in the built environment: Towards automated building inspection procedures using drones. *Autom. Constr.* **2018**, *93*, 252–264. [CrossRef]
17. Kim, I.H.; Jeon, H.; Baek, S.C.; Hong, W.H.; Jung, H.J. Application of crack identification techniques for an aging concrete bridge inspection using an unmanned aerial vehicle. *Sensors* **2018**, *18*, 1881. [CrossRef]
18. Voutetaki, M.E.; Naoum, M.C.; Papadopoulos, N.A.; Chalioris, C.E. Cracking Diagnosis in Fiber-Reinforced Concrete with Synthetic Fibers Using Piezoelectric Transducers. *Fibers* **2022**, *10*, 5. [CrossRef]
19. Vaghefi, K.; Ahlborn, T.M.; Harris, D.K.; Brooks, C.N. Combined imaging technologies for concrete bridge deck condition assessment. *J. Perform. Constr. Facil.* **2015**, *29*, 04014102. [CrossRef]
20. Qian, Y.; Huang, C.; Han, B.; Cheng, F.; Qiu, S.; Deng, H.; Duan, X.; Zheng, H.; Liu, Z.; Wu, J. Quantitative Analysis of Bolt Loosening Angle Based on Deep Learning. *Buildings* **2024**, *14*, 163. [CrossRef]
21. Wu, J.; He, Y.; Xu, C.; Jia, X.; Huang, Y.; Chen, Q.; Huang, C.; Dadras Eslamlou, A.; Huang, S. Interpretability Analysis of Convolutional Neural Networks for Crack Detection. *Buildings* **2023**, *13*, 3095. [CrossRef]

22. Tran, Q.H.; Han, D.; Kang, C.; Haldar, A.; Huh, J. Effects of ambient temperature and relative humidity on subsurface defect detection in concrete structures by active thermal imaging. *Sensors* **2017**, *17*, 1718. [CrossRef]
23. Sun, Z.; Sun, M.; Siringoringo, D.M.; Dong, Y.; Lei, X. Predicting bridge longitudinal displacement from monitored operational loads with hierarchical CNN for condition assessment. *Mech. Syst. Signal Process.* **2023**, *200*, 110623. [CrossRef]
24. Cheng, C.; Shang, Z.; Shen, Z. Automatic delamination segmentation for bridge deck based on encoder-decoder deep learning through UAV-based thermography. *NDT E Int.* **2020**, *116*, 102341. [CrossRef]
25. Ichi, E.; Dorafshan, S. SDNET2021: Annotated NDE dataset for Structural Defects [Data set]. In *UND Datasets*; The University of North Dakota: Grand Forks, ND, USA, 2021. Available online: https://commons.und.edu/data/19 (accessed on 21 September 2023).
26. Ichi, E.O. Validating NDE Dataset and Benchmarking Infrared Thermography For Delamination Detection In Bridge Decks. Master's Thesis, University of North Dakota, Grand Forks, ND, USA, 2021. Available online: https://commons.und.edu/theses/4170 (accessed on 30 October 2023).
27. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015*; Proceedings, Part III 18; Springer International Publishing: Berlin, Germany, 2015; pp. 234–241.
28. Deng, W.; Mou, Y.; Kashiwa, T.; Escalera, S.; Nagai, K.; Nakayama, K.; Matsuo, Y.; Prendinger, H. Vision based pixel-level bridge structural damage detection using a link ASPP network. *Autom. Constr.* **2020**, *110*, 102973. [CrossRef]
29. Avdelidis, N.; Moropoulou, A. Applications of infrared thermography for the investigation of historic structures. *J. Cult. Herit.* **2004**, *5*, 119–127. [CrossRef]
30. Lahiri, B.; Bagavathiappan, S.; Jayakumar, T.; Philip, J. Medical applications of infrared thermography: A review. *Infrared Phys. Technol.* **2012**, *55*, 221–235. [CrossRef] [PubMed]
31. Madding, R.P. Science behind thermography. In *International Society for Optics and Photonics*; Pratt, W.K., Ed.; Wiley & Sons. Inc.: Hoboken, NJ, USA, 1983; Volume 371, pp. 2–9. [CrossRef]
32. Washer, G.; Fenwick, R.; Bolleni, N.; Harper, J. Effects of environmental variables on infrared imaging of subsurface features of concrete bridges. *Transp. Res. Rec. J. Transp. Res. Board* **2009**, *2108*, 107–114. [CrossRef]
33. Tomita, K.; Chew, M.Y.L. A Review of infrared thermography for delamination detection on infrastructures and buildings. *Sensors* **2022**, *22*, 423. [CrossRef] [PubMed]
34. Hiasa, S.; Birgul, R.; Catbas, F.N. Effect of defect size on subsurface defect detectability and defect depth estimation for concrete structures by infrared thermography. *J. Nondestruct. Eval.* **2017**, *36*, 57. [CrossRef]
35. Washer, G.; Fenwick, R.; Nelson, S.; Rumbayan, R. Guidelines for thermographic inspection of concrete bridge components in shaded conditions. *Transp. Res. Rec. J. Transp. Res. Board* **2013**, *2360*, 13–20. [CrossRef]
36. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
37. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
38. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
39. Rosso, M.M.; Aloisio, A.; Randazzo, V.; Tanzi, L.; Cirrincione, G.; Marano, G.C. Comparative deep learning studies for indirect tunnel monitoring with and without Fourier pre-processing. *Integr. Comput. Eng.* **2023**, *31*, 213–232. [CrossRef]
40. Cha, Y.; Choi, W.; Suh, G.; Mahmoudkhani, S.; Büyüköztürk, O. Autonomous structural visual inspection using region-based deep learning for detecting multiple damage types. *Comput. Civ. Infrastruct. Eng.* **2018**, *33*, 731–747. [CrossRef]
41. Pozzer, S.; Azar, E.R.; Rosa, F.D.; Pravia, Z.M.C. Semantic segmentation of defects in infrared thermographic images of highly damaged concrete structures. *J. Perform. Constr. Facil.* **2021**, *35*, 04020131. [CrossRef]
42. Garcia-Garcia, A.; Orts-Escolano, S.; Oprea, S.; Villena-Martinez, V.; Martinez-Gonzalez, P.; Garcia-Rodriguez, J. A survey on deep learning techniques for image and video semantic segmentation. *Appl. Soft Comput.* **2018**, *70*, 41–65. [CrossRef]
43. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
44. Wang, L.; Li, R.; Zhang, C.; Fang, S.; Duan, C.; Meng, X.; Atkinson, P.M. UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS J. Photogramm. Remote Sens.* **2022**, *190*, 196–214. [CrossRef]
45. Nanni, L.; Maguolo, G.; Paci, M. Data augmentation approaches for improving animal audio classification. *Ecol. Inform.* **2020**, *57*, 101084. [CrossRef]
46. Kampffmeyer, M.; Salberg, A.-B.; Jenssen, R. Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 1–9.
47. Kumar, S.S.; Abraham, D.M.; Jahanshahi, M.R.; Iseley, T.; Starr, J. Automated defect classification in sewer closed circuit television inspections using deep convolutional neural networks. *Autom. Constr.* **2018**, *91*, 273–283. [CrossRef]