Application of K-Nearest Neighbors Algorithms for Void Classification in Composite Oriented Strand Board

Wenyue Hu¹, Yaser Eftekhari¹, Sam Callander², Xiaoxing Wang³, Christopher C. Bowland⁴, Frank Nguyen⁵, Jeremy McCaslin⁶, Christoph Schaal⁷, Grace X. Gu⁸, Carina Li⁹, Bo Jin¹

¹Advanced Composites Design Lab, University of Southern California
 ² Volume Graphics Software, Fiber Composite Material Analysis, Hexagon, Charlotte, NC
 ³ MATLAB, MathWorks Japan

⁴ Chemical Sciences Division, Carbon and Composites Group, Oak Ridge National Lab ⁵ The Gill Corporation, El Monte, California

⁶ ANSYS

Department of Mechanical Engineering, California State University, Northridge
 Department of Mechanical Engineering, University of California, Berkeley
 Department of Aeronautics and Astronautics, Massachusetts Institute of Technology

ABSTRACT

Composite Oriented Strand Board (COSB) is an aerospace-grade material that is manufactured using laminated, unidirectional carbon fiber-epoxy prepreg strands. Its manufacturing methodology allows fine-tuning, producing controllable thickness, flatness, and microstructure with quality assurance. To comprehend the microstructure of COSB from X-ray computed tomography (XCT) scan data, it is critical to accurately assess and quantify void content in the post-analysis. The conventional methods of post-analysis used for X-ray micro-computed tomography (micro-CT) have become inadequate due to their time-consuming nature and oftentimes imprecise measurements. In this paper, we present a new approach for enhancing void classification accuracy by utilizing the K-Nearest Neighbors Classifiers (KNN) with the assistance of convolutional kernels. KNN training based on two labels of greyscale thresholding images achieved a 1% error rate, a significant improvement over the three deep learning algorithms (Fully Convolutional Neural Network, U-net, SegNet) in our previous study. When classifying five different void labels, the multi-classifier KNN algorithm with convolutional kernels yields around 2% error rate. Encouraged by these results, we propose using convolutional neural networks (CNNs) to make more complex reasoning and decisions when classifying voids, improving the accuracy of void characterization.

1. INTRODUCTION

The application and demand for Carbon Fiber Reinforced Polymers (CFRPs) have increased steadily in the aerospace, defense, automotive, and wind energy industries because of superior mechanical performance levels, low density, durability, and resistance to fatigue and corrosion. In 2020, the CFRP market was valued at 7 billion dollars and had a market share of 45% of the composites market; an annual growth rate of 8.3% was forecast from 2021 to 2030, reaching 15.4 billion dollars by 2030. Voids in CFRPs are considered undesirable defects that stem from the manufacturing process. They can degrade a broad range of mechanical properties, such as

interlaminar shear stress, but usually remain hidden and undetected [1]. Researchers are increasingly interested in investigating porosity evolution. By understanding the evolution and distribution of porosity, researchers can limit/eliminate formation by modifications to the manufacturing process, such as optimizing cure parameters (pressure, temperature, duration, etc.) and protocols and designing processes and materials that restrict void formation and growth. Presently, trained professionals perform most void detection/measurement and quantification, causing a heavy burden of time, labor, and cost. To reduce this burden and respond to the growing sector, autonomous analysis solutions are needed and offer approaches to increasing automation in manufacturing [2 - 6].

The process of void study in Carbon Fiber Reinforced Polymers (CFRPs) is traditionally performed in two steps. Firstly, tomography is used to detect porosity in CFRPs through C-scan visualization of porous regions with lower density. Secondly, human experts intervene to identify and assess void morphology, location, volume fraction, and distribution. However, the manual identification and evaluation of voids are time-consuming and require significant human input, especially when dealing with large micro-CT data files. To overcome these challenges, machine learning algorithms can be employed in void classification, offering the potential to improve the accuracy, speed, and objectivity of the process. Once trained, machine learning algorithms can fully automate the void classification process, reducing the need for manual inspection and minimizing the risk of human errors. By analyzing large datasets, machine learning algorithms can identify patterns that may not be immediately recognizable to human eye, leading to improved accuracy in void classification. Furthermore, the algorithm provides more objective results, reducing the influence of personal bias or subjective interpretations in the void classification process.

In the world of machine learning, mimicking the human mind has been a long-standing goal. This is especially true when it comes to image recognition. In order to achieve this, many researchers have turned to convolutional neural networks (CNN), which are inspired by the human brain's visual cortex. In this regard, this study aimed to extract additional features from more complex neighborhood values in an image and use them to classify images. They employed convolutional kernels with various sizes to create these features and fed them into the K-Nearest Neighbors (KNN) algorithm for classification. This choice of algorithm was made based on its generality and ability to solve classification problems, especially when samples of different classes are distributed unevenly in the feature-set domain.

2. EXPERIMENTATION

2.1 Manufacturing of COSB

In the study, panels of composite-oriented strand board (COSB) were produced using fresh unidirectional carbon fiber-epoxy prepreg strands. Two layup methods were utilized in the manufacturing process, namely, mat-stacking and mechanical agitation layups. The void content and morphology of the panels were examined through the use of microscopy of polished cross-sections and stitched high-resolution micro-CT images.

To produce the panels, prepreg sheets were cut into strands using a razor blade. The aspect ratio (AR) of each strand, defined as the strand length-to-width ratio, was varied to examine the mechanical performance and durability of the resulting panels. Strands with AR values of 1, 2, 5, and 10 were prepared, with a width of 10 mm for all. While lower AR strands are expected to yield

inferior tensile strength due to limited fiber continuity, they are suitable for molding contoured parts with complex shapes. Conversely, higher AR strands yield superior mechanical performance due to longer fibers and fewer strand ends.

The two layup methods used to produce COSB panels were mechanical agitation and mat-stacking. The former method involved placing unidirectional (UD) strands of 250 g in an open box (215.9 mm x 215.9 mm x 200 mm) and manually shaking the box. This method is rapid and suitable for mass production, but it results in uneven fiber distribution. The latter method produced more uniform distributed layouts by forming multiple thin layers (mats of UD prepreg strands) as a COSB prepreg sheet, stacking them, and pressing them. The final mat-stacked COSB consisted of six layers of prepreg strand mats (each 215.9 x 215.9 mm mat weighing 41.7 g) and had a total weight of 250 g, regardless of strand AR.

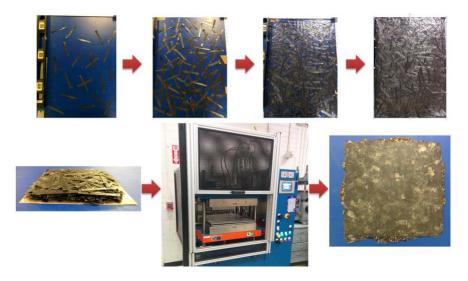


Figure 1. From left to right, the manufacturing process of the "mat-stacking" panels.

The prepreg strands were cured by compression molding using a Wabash hot press, resulting in a flat COSB panel measuring 215.9 mm \times 215.9 mm. The resulting panel resembles a wood-based oriented strand board, as shown in Figure 2 on the right.





Figure 2. Composite Oriented Strand Board (COSB, left), similar to the wood OSB (right), which is widely used in civil constructions.

2.2 Micro-CT Analysis

The potential of X-ray computed tomography (micro-CT) for detecting defects, measuring internal structural configurations, and determining mass distribution has been well-established. In line with previous research [7], this study utilized a Phoenix Nanotom tomographic machine with a Hamamatsu C-7942 detector and Molybdenum (Mo) anode to perform micro-CT scans on COSB samples. The settings of the tomographic machine were 80 kV and 150 uA, with a 500 ms exposure time, and the attained resolution was 13.04 µm/px. Multiple scans were carried out on different regions of a single sample to cover the entire volume, and the obtained volumes were reconstructed using a user-defined routine. Stitching and cropping techniques were employed to generate a complete sample volume, as shown in Figure 4. The micro-CT dataset consisted of 3195 images, with each image having a resolution >3400 x 1900 pixels, occupying ~24GB of memory. It comes from four COSB panels using two layup methods (mechanical agitation and mat-stacking) produced with four prepreg strand aspect ratios (1:1, 1:2, 1:5, and 1:10) (Figure 3). Due to limited storage space and micro-CT sample size, a global view of void distribution at the panel level was unobtainable. Nevertheless, image stitching was utilized to produce an animation-like continuous frame that accurately characterized the prepreg strands and void content.

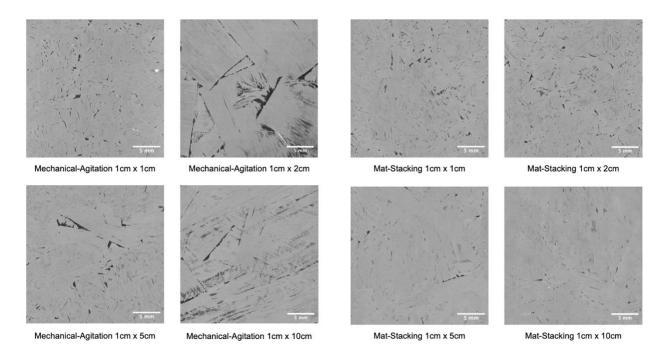


Figure 3. Micro-CT scans of COSBs made of prepreg strands of four Aspect Ratio respectively, with mechanical-agitation and mat-stacking.

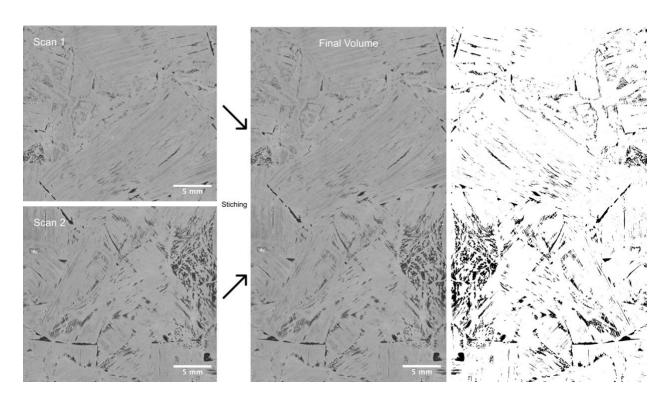


Figure 4. Stitching of multiple scanned volumes to obtain completed final micro-CT scanned volume.

2.3 Evaluations of Grey Scale Thresholding

Thresholding is a type of image segmentation method commonly used when selecting and segregating regions of interest (ROI) while leaving the remaining parts unconcerned. In cases where multiple objects with different grey levels need to be separated, multi-thresholding may be employed. It is also a commonly implemented tool in commercial software for image processing. In CT images, each pixel is assigned a grayscale value between 0 and 255, which corresponds to the X-ray beam attenuation of the object. Researchers in the field of material science often use global grayscale thresholding as a simple technique for the basic post-analysis of CT scans. It works on the idea that a CT scan can turn into a bimodal histogram based on the greyscale, and the object of interest has a considerable grayscale difference from the background.

While grayscale thresholding can achieve good results to some extent, it may not be suitable for images with complex and overlapping structures, where objects of interest have similar grayscale values to the background or other structures. Greyscale thresholding relies solely on setting a fixed threshold for image pixel values, which may not accurately capture the variations in the image that could be useful for classification purposes. To overcome these limitations, more advanced segmentation techniques incorporating human intelligence, such as machine learning-based approaches, make it more effective for images with complex object boundaries. Human operators can consider the image's overall context, such as the material's fiber direction and target objects' distribution, deduct based on local image grayscale values, and use their experience and intuition to accurately judge and identify the ROI. This study aims to develop a new approach to image segmentation that incorporates human-like deduction capabilities into the image processing process.

2.4 Design Logic Behind the Modified KNN

This study aims to improve the accuracy of void classification in COSB CT scans by mimicking the human mind's ability to make complex decisions based on local image features. We recognized that CNN had the ability to extract useful features from images. Incorporating additional features extracted from more complex neighborhood values in the image, the algorithm mimics the local pixel lens of a human observer and makes better classification decisions. To accomplish this, we used convolutional kernels of various sizes to extract additional features from the image.

The KNN classifier is a machine learning algorithm that classifies data points based on their proximity to other data points. One of the key advantages of using the KNN classifier is its ability to handle images with complex and variable backgrounds. In the context of image segmentation, the KNN classifier can be used to classify each pixel in an image based on its similarity to the neighboring pixels. This approach can provide more accurate results than traditional segmentation methods like grey scale thresholding because it considers the spatial relationship between pixels in addition to their intensity values. The choice of using KNN was primarily based on the generality of this algorithm in solving classification problems. KNN works best when the samples of different classes are distributed unevenly in the feature-set domain. By creating better distinguishing features, the accuracy of KNN in classifying different void types has been improved. While the choice of KNN was arbitrary, our results demonstrate its effectiveness in combination with convolutional kernels in improving the accuracy of voids classification.

These features generated by convolutional kernels are provided to KNN alongside the original pixel values. We found that normalizing the feature values before training KNN helped to improve

the algorithm's performance. By incorporating these additional features, we achieved better classification results than using the original pixel values alone.

2.5 Algorithmic Pipeline and Implementation of KNN Classifier

2.5.1 Data Labeling Process

The supervised clustering algorithm employed in this study requires manual labeling. To this end, a set of 200 images was randomly sampled from 7 datasets comprising two COSB layup methods with four prepreg strand aspect ratios. After thoroughly examining the images, a highly trained professional with a deep understanding of the manufacturing process of COSB identified 8 images that contained the most representative voids information. Utilizing their extensive experience and reference to 3D models, the professionals categorized the voids into four distinct categories based on their morphology and location relative to other voids. Additionally, the "background" category was added to represent the remaining areas of the image that did not contain voids. Consequently, the selected images were carefully labeled at the pixel level using the following steps:

- 1. A comprehensive observation of the 200 images led to the identification of four void types, which were then categorized into four labels, with one more label assigned to the background.
- 2. Pick out the voids and label voids first. Compare the 2D image with a rendered 3D model to identify the voids.
- 3. The morphology and location of each void were examined to determine the category to which it belonged. Repeat the process until there are no obvious voids found in this image.
- 4. Label the remaining area of the image as "background."

2.5.2 Labels Explain

Each category (label) in this study is defined by a set of specific criteria, which are based on the expertise of professionals with extensive experience in COSB manufacturing and void study in material science. When identifying and labeling each void, the professional evaluates various aspects, including size, morphology, and location relative to other voids, to assign the most appropriate label.

The first category, labeled Background Figure 5 (a), represents areas with no voids. According to the standardized labeling procedures, the background is assigned after all void labeling is complete. The second category, labeled Cigar-like voids Figure 5 (b), is defined as voids that are relatively longer than other types of voids (Aspect Ratios \geq 10), giving them a cigar-like shape. Their 2D geometry is distinct from other void types, and they may exhibit varying shades of color, ranging from very dark to greyish. Cigar-like voids are typically found in gaps between two prepreg strands. The third category, labeled Clustered voids Figure 5 (c), is defined as a group of relatively small voids (radius \leq 1 mm) clustered together, like an archipelago. These voids are usually interconnected with one another and cannot be distinguished as individual voids from a 2D image perspective. They are often found on the surface of the material, caused by unevenness. The fourth category, labeled Cloud-like voids Figure 5 (e), is defined as large voids with a clear boundary and a radius of \geq 5 mm. These voids are typically very dark in color and are most likely air

bubbles that have become trapped in the material. Cloud-like voids are easily distinguishable from other types of voids and are considered the most catastrophic. The fifth and final category, labeled Dot voids Figure 5 (d), is an umbrella term that includes voids that do not fit into the other three categories. These voids are small in size and have clear boundaries but may not necessarily be dot-shaped. The distribution of dot voids may be random.

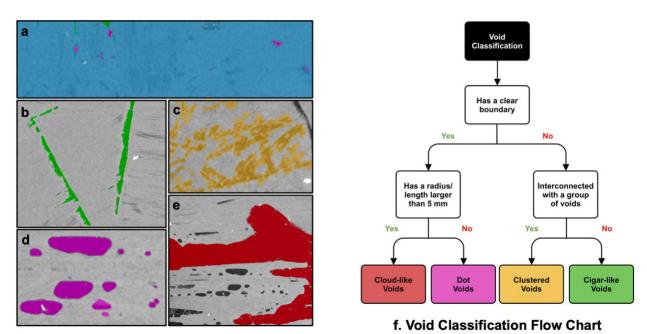


Figure 5. (a) Background, (b) Cigar-like voids, (c) Clustered voids, (d) Dot voids, (e) Cloud-like voids, (f) Flow chart of standardized procedures of void classification

2.5.3 Label Post-processing

As with any manual labeling process, there is a possibility that some pixels have not been assigned any of the pre-defined labels. This can be attributed to the fact that the labeling process involved first identifying and assigning labels to voids and subsequently categorizing the remaining area as "background" manually. Due to the complex nature of labeling a high-resolution image and the potential for human error, a small number of pixels may have been overlooked in this process. It is essential to assign appropriate labels to these omitted pixels to ensure that all data points are accounted for and accurately reflect the underlying microstructure. This will prevent misinterpretation of data points that do not conform to the pre-defined labels.

To address the issue of unlabeled pixels in our dataset, we implemented a 3x3 voting scheme. Specifically, we utilized a window of 3x3 of surrounding label values to conduct a vote and assigned the majority label to the unlabeled pixel. For example, a 3x3 neighborhood comprises three labels 5, one label 4, two labels 1, and two labels 2 surrounding the unlabeled center pixel. Based on the majority label in the neighborhood, we assigned the corresponding label value to the unlabeled pixel. In this particular example,

we assigned a label 5 to the center pixel since it appears more frequently than any other label in the neighborhood. Although this approach may seem simplistic, it effectively resolves the issue of unlabeled pixels, providing a practical solution that works well in most scenarios.

2.5.4 KNN Supervised Clustering Algorithm and Training Process

The K-Nearest Neighbors (KNN) classifier is a supervised classification algorithm in machine learning that aims to group data points based on their similarity. When classifying a new data point, the KNN algorithm searches through the training data to find the K-nearest neighbors (i.e., K data points closest to the new data point). The distance between data points is typically measured using a distance metric such as Euclidean distance. The class of the new data point is then determined by a majority vote of its K-nearest neighbors. After training, the algorithm assigns a class label to a new data point based on the class of its neighbors.

Our method involves downsizing the images by a factor of four and applying F-1 different convolutional kernels to the resized images to generate F features. Kernels, called convolution matrices or masks, are matrix structures with m x n dimensions. They are designed to perform various tasks, such as blurring, sharpening, and edge detection. During convolution, the kernel matrix moves through the entire image matrix from left to right and top to bottom. At each step, it returns a single pixel value, which is the average of the neighborhood pixels in an m x n matrix grid. Finally, the resulting pixel value at each step is used to generate the output image matrix. By applying different kernels, new features are extracted, which can be utilized for training purposes. We then combine the resulting features with the original pixel values, resulting in F features with WxH points for each input image. To improve the algorithm's training, we normalized the feature values before feeding them to the K-Nearest Neighbors (KNN) algorithm. By flattening the 2D images into a dataset of points with F features, we then shuffle the data and split the dataset into train and test sets. We run the KNN algorithm on the training data and test it on the test data. We repeated this process for each parameter combination and found the best parameter set by tuning the algorithm.

In order to improve model performance, hyperparameter tuning is a crucial step. The training process for the KNN algorithm involves selecting an optimal value for hyperparameters, such as the number of neighbors (k), through cross-validation. In our study, we selected five parameters in KNN: the number of neighbors, the number of samples, the power, the algorithm, and the weights. We plugged in different choices for each parameter during each iteration of cross-validation. The data is sampled during cross-validation to create both a training set and a testing set. The final testing score is used as a performance measure.

 Table 1: Hyperparameter Tuning Example

Parameters	Tuning Selection
Number of Neighbors	3, 5, 7, 9, 15
Number of Samples	35, 80, 150, 300, 1000, 2000
Power	1, 2, 3
Algorithm	'brute', 'ball tree', 'auto', 'kd tree'
Weights	Uniform, distance

3. RESULTS AND DISCUSSION

3.1 Results of KNN Classifier on Binary Labels Data

Our previous study implements 3 neural networks, Fully Convolutional Network (FCN), SegNet, and U-net, for automatically identifying voids in micro-CT data of COSBs using binary grayscale thresholding labels as training data [8]. The Fully Convolutional Network (FCN) is a type of neural network that can perform image segmentation in a fully convolutional manner. This means that it can classify each pixel in an image independently of the other pixels, making it more efficient than traditional neural networks. SegNet is another type of neural network that has been used for medical image segmentation. U-net consists of an encoder and decoder architecture, with the encoder used to extract features from the image and the decoder used to reconstruct the segmented image.

In this study, we evaluated the performance of a method for classifying voids in images based on grayscale thresholding and the k-nearest neighbor (KNN) algorithm. We found that the KNN algorithm achieved a 1% error rate in the case of binary label classification, significantly improving over the previously tested deep-learning techniques. This indicates that the KNN algorithm is a reliable and effective method for classifying voids in images.

3.2 Results of KNN Classifier on Five Labels Data

We expand the scope of classification by increasing the number of labels to five. After running the method again, we achieved a 5% error rate. Further analysis showed that the algorithm had some shortcomings, particularly in distinguishing between cloud shapes and cigar shapes of voids. We identified that the main reason for this was that the features created using the convolutional kernels were limited to rudimentary calculations such as edge detection, local minima, and local Gaussian averages. As a result, the algorithm could not make complex human-like reasoning, which is required for the accurate classification of voids with different labels.

We further optimized the algorithm by running multi-classifier KNN with random convolutional kernels and human-labeled images and achieved an error rate of around 2%. We evaluated the performance of our method using the micro F1 score and the macro F1 score. The micro F1 score is a measure of the overall performance of the classifier, while the macro F1 score is a measure of the classifier's ability to perform well in individual classes. Our method achieved a micro F1 score of 0.998, indicating that it can accurately classify images across all classes. However, the macro F1 score was comparably lower, at 0.377 for 5 classes, indicating that there is still much room for improvement in the classifier's ability to distinguish between individual classes. Through continued experimentation, optimization, and more epochs of training, we believe that our method has the potential to become a highly accurate and robust image classification method that can be applied to a variety of real-world applications.

Table 2: Comparison of Error Rate Among Different Algorithms

	Binary Classification (Two Labels)	Five Labels Classification
KNN	1%	5%
FCN	7%	-
SegNet	18%	-
U-net	10%	-
KNN + Convolutional	-	2%
Kernels		

4. CONCLUSIONS

In the study, we utilized the K-Nearest Neighbors (KNN) classifier to categorize greyscale images into two labels, managing to significantly lower the error rate to a mere 1%. This performance not only surpassed but demonstrated a significant enhancement over other models like the Fully Convolutional Neural Network, U-net, and SegNet when applied to the same dataset. Further, we overcame the problem of distinguishing between labels due to limitations in the feature extraction process and achieved a roughly 2% error rate when classifying five distinct void labels by applying a multi-classifier KNN equipped with random convolutional kernels. This performance, coupled with an impressively high micro-F1 score of 0.998, hints at the significant potential of employing convolutional kernels in such applications.

Considering the current method's limitations, we propose to design a semi-complex convolutional neural network (CNN) that can extract more sophisticated features from local pixel information. In particular, we aim to create a CNN that can extract features similar to the convolutional kernels used in the current method but with greater complexity and sensitivity. We believe that a CNN can make more complex reasoning and decisions when classifying a pixel because it has multiple convolutional layers stacked on top of previous layers. This allows it to extract increasingly complex features from the input image, which can lead to more accurate classification results.

To design the CNN, we propose leveraging a large dataset of labeled images to train the network via a supervised learning technique. Additionally, we plan to experiment with transfer learning—recycling pre-trained networks that have already acquired useful features from similar tasks. Once trained, the CNN could efficiently classify voids in images with a high degree of accuracy, offering value to a range of fields such as materials science, biology, and dermatology. Overall, given the existing KNN classifier with convolutional kernels' success in reducing the error rate to 2%, we are optimistic that our proposed CNN can substantially improve void classification accuracy. This could lead to new research directions in this domain.

5. CASE STUDY

The application of Computed Tomography (CT) technology is unique in that it enables the visualization of the internal structure of the COSB material in a 360-degree view. The CT scan data is then processed in Volume Graphics to create a highly accurate 3D volume composed of the scanned images. The output is a representation of the scanned part in voxels, where each voxel represents a gray value determined by the material's density. Analyzing the CT scan data involves separating these gray values to determine different materials or air, which provides highly detailed insight into the microstructure of the COSB material. This includes calculating local fiber

orientations to generate orientation tensors, displaying fiber orientation in color code or as vectors/tensors in 3D, detecting porosity or voids within the material, and determining local volume fractions.

We explored the porosity analysis using Volume Graphics and generated very insightful results. The use of an Integration Mesh applied in determining fiber orientations and defining tensors enables the visualization of how the fibers follow the relative orientation of the strand board. This is shown in Figure 6 (a). The amount of porosity present is a crucial metric for comparing the two manufacturing techniques. This is determined by overall volume, as seen in Figure 6 (b). Additionally, local volume fractions can be determined by considering the porosity present, with each cell indicating the percentage of the ratio of the volume of fibers present to the total volume of the layer. This is shown in Figure 6 (c).

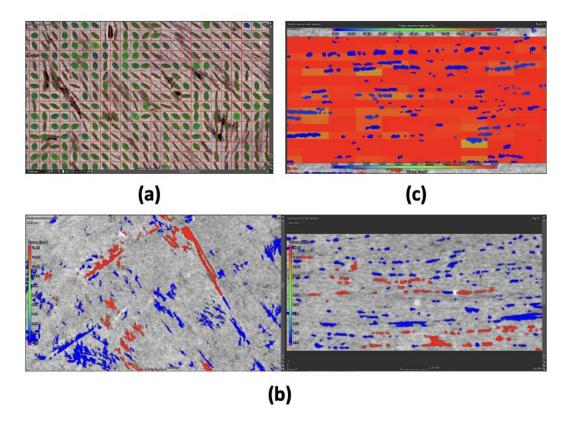


Figure 6. (a) Tensor flow network. (b) Void segregation based on greyscale thresholding. (c) The volume of fibers present to the total volume of the layer.

6. REFERENCES

- [1] R. A. Smith. Composite Defects and Their Detection. Materials Science and Engineering. Vol. III. Encyclopedia of Life Support Systems (EOLSS). 2009.
- [2] Reed Kopp, Joshua Joseph, Xinchen Ni, Nicholas Roy, and Brian L. Wardle. Deep Learning Unlocks X-ray Microtomography Segmentation of Multiclass Microdamage in Heterogeneous Materials. Advanced Material, 2022, 2107817.
- [3] Reed Alan Kopp. X-ray Micro-Computed Tomography and Deep Learning Segmentation of Progressive Damage in Hierarchical Nanoengineered Carbon Fiber Composites. Massachusetts Institute of Technology. Department of Aeronautics and Astronautics. 2021.
- [4] Brian L. Wardle, Roberto Guzman de Villoria, Antonio Miravete, Southborough. Systems and methods for structural sensing. United States Patent, 2014.
- [5] Wardle B, et al., Fabrication and Characterization of Ultrahigh-Volume-Fraction Aligned Carbon Nanotube-Polymer Composites. Advanced Materials, vol. 20, pp. 2707-2714 (2008)
- [6] Dhimiter Bello, Brian L Wardle, Namiko Yamamoto, et al. Exposure to nanoscale particles and fibers during machining of hybrid advanced composites containing carbon nanotubes. Journal of Nanoparticle Research, Springer Netherlands, 11-1, 2009.
- [7]. Djordjevic BB. Advanced Ultrasonic Probes for Scanning of Large Structures. Ultrasonic International; 1993.
- [8] Hu W, Wang X, Bowland C, Nguyen P, Li C, Nutt S, Jin B. Deep Learning for Void Detection in Composite Oriented Strand Board. In: Conference proceeding. CAMX Anaheim, United States, October 17 20, 2022.
- [9] Bo Cheng Jin, Xiaochen Li, Atul Jain, Carlos González, Javier LLorca, Steven Nutt. Optimizing microstructures and mechanical properties of composite oriented strand board from reused prepreg. Composite Structures, Volume 174, 2017, Pages 389-398, ISSN 0263-8223.