Metamaterial Bandgap Prediction using Convolutional Neural Networks

Ryan Lutz, Jiaxuan Zhang, Ukamaka Ezimora

Duke University

December 2025

Abstract

Metamaterials are synthetic materials with engineered microstructures that control wave propagation, but identifying geometries with desired bandgaps typically requires finite element simulations that take hours per design, severely limiting exploration of the design space. We develop a convolutional neural network using transfer learning to rapidly predict bandgap existence across five 1 kHz frequency ranges for two-dimensional metamaterial unit cells. Training a ResNet18-based model on 32,768 samples, we achieve 94.5% element-wise test accuracy, exceeding both the 91.48% interpretable baseline and our 93% design goal. Systematic evaluation shows that increasing input resolution from 64×64 to 128×128 yields the largest performance gain, while per-class threshold optimization provides modest additional improvement in higher frequency ranges. The trained model classifies a design in roughly 25 ms, enabling rapid screening of thousands of candidates.

1 Introduction

Metamaterials are synthetic materials that have properties rarely seen in nature. Their unique electromagnetic, acoustic, and mechanical properties enable diverse applications across optics, acoustics, and electronics. Metamaterials have micro- and meso-scale structures that result in these unique physical properties[3]. These properties are determined by bandgaps which are bands in the materials through which certain frequencies cannot pass [5]. Metamaterials are typically designed and evaluated using costly finite element simulations that evaluate the geometry of the material and a dispersion curve that contains information on the bandgap. As this method is costly, there is a large focus on utilizing machine learning models to aid in and design metamaterials with the desired bandgaps.

Chen et al. (2022) developed an interpretable machine learning approach for predicting bandgaps in 2D metamaterials using engineered shape-frequency features extracted from unit-cell geometries. They evaluated six models: Support Vector Machine (SVM), Logistic Regression (LR), Random Forest (RF), Multi-Layer Perceptron (MLP), Classification and Regression Trees (CART), and Light Gradient Boosting Machine (LightGBM). Their best model (LightGBM) achieved 91.48% testing accuracy across all frequency ranges. While Chen et al.'s approach provides a strong baseline performance with interpretability, this project explores whether deep learning with transfer learning can achieve higher accuracy by learning complex geometric patterns directly from images while still enabling rapid bandgap screening compared to finite element simulations. The advantage of the CNN over the interpretable methods is the ability to identify bandgaps in metamaterials without knowing the shape frequencies.

Accurate bandgap prediction is essential for designing metamaterials in applications such as acoustic filters, semi-conductors, and electromagnetic shielding [3]. While Chen et al.'s interpretable models provide a strong baseline performance, even modest improvements (2-3%) can substantially reduce the number of candidates requiring finite element evaluation. If successful, this project would demonstrate that deep learning can offer meaningful advantages over interpretable methods for metamaterial design, enabling more reliable computational screening and accelerating the development of novel metamaterial structures.

2 Data Description

The dataset, released by the Brinson Research Group at Duke University (Chen *et al.*, 2022) [2], consists of MATLAB files containing 10×10 unit–cell geometries of two-dimensional metamaterials and their corresponding dispersion curves generated via finite-element simulation. Each unit cell is a square of side length 0.1 m and contains two constituent materials: a soft, light polymer ($E = 2\,\text{GPa},\ \rho = 1000\,\text{kg}\,\text{m}^{-3}$) and a stiff, heavy steel-like material ($E = 200\,\text{GPa},\ \rho = 8000\,\text{kg}\,\text{m}^{-3}$). The unit cells exhibit four axes of symmetry: along the x- and y-axes and along the diagonals at $\pm 45°$. Under these symmetry constraints, the 10×10 patterns reduce to 15 irreducible pixels (from 100), yielding a 15-dimensional binary vector (0 = soft, 1 = stiff) and a total design space of 2^{15} states.

For use with a CNN, each 15-dimensional vector is expanded back to a full 10×10 unit cell by applying the four-axis symmetry, producing the image input. The full dataset contains 32,768 samples, each with dispersion data for 20 bands evaluated at 150 k-points (shape $32768 \times 20 \times 150$). Across samples, the dispersion values have mean 1935.68 ± 570.91.

3 Preprocessing

Before training the CNN, several preprocessing steps were applied. First, band gaps were identified by computing the frequency difference between adjacent dispersion bands across all k-points. Gaps smaller than 0.01 Hz were treated as negligible and set to zero.

This work aims to address the existence of bandgaps as a binary classification problem: whether a metamaterial structure exhibits a complete bandgap within a specified frequency range. A bandgap exists when wave propagation of certain frequencies is prohibited in all directions. This is either present or absent for a given structure, making it inherently binary.

To match the work of Chen et al., five target frequency ranges were defined (expanded from 3 in the progress report): $[0, 1]$, $[1, 2]$, $[2, 3]$, $[3, 4]$, and $[4, 5]$ kHz. Samples were labeled positive (1) for a range if any band gap overlapped that range (overlap threshold $s_{\min} = 0$); otherwise the label was 0. Thus, a sample might have the label vector $[1, 0, 1, 0, 0]$, indicating bandgaps in ranges $[0, 1]$ and $[2, 3]$ kHz, but not in the other ranges. This multi-label formulation allows the model to predict bandgap existence across all frequency ranges simultaneously.

To prepare inputs for the CNN, each 15-dimensional vector was reconstructed into a 10×10 unit-cell image using the four-axis symmetry, then resized to 64×64 with bilinear interpolation. Images were encoded as single-channel grayscale (binary geometry), converted to tensors, and normalized to $[-1, 1]$, which centers the inputs and improves numerical stability.

4 Methods

Metamaterial bandgap predictions are modeled from 64 x 64 grayscale unit-cell images as five-label multi-label classification. A transfer-learning CNN model based on ResNet18 was built for this project. ResNet18 is an 18-layer residual network with skip connections and approximately 11 million parameters, originally trained on ImageNet's 1.2 million images. In section 4.1 the baseline model is summarized, then in section 4.2 improvement modifications are discussed that systematically evaluate architectural and training modifications to optimize performance.

4.1 Baseline Model Architecture

The baseline model adapts ResNet18 with three modifications: (i) averaging the pretrained RGB weights in the first convolution to accept grayscale input ($7 \times 7 \times 3 \to 7 \times 7 \times 1$), (ii) replacing the final layer with a 5-neuron output for multi-label prediction, and (iii) removing the initial max-pooling layer to preserve spatial resolution for 64×64 inputs.

For each frequency range i, the model predicts a bandgap probability $p_{ij} = \sigma(z_{ij}) = 1/(1 + e^{-z_{ij}})$ from the logit z_{ij}. Training uses a class-balanced binary cross-entropy:

$$\mathcal{L} = \frac{1}{5N} \sum_{i=1}^{5} \sum_{j=1}^{N} \left[-w_i \, y_{ij} \log(p_{ij}) - (1 - y_{ij}) \log(1 - p_{ij}) \right],$$

where $w_i = n_{negative,i}/n_{positive,i}$ is the positive-class weight for frequency range i to address class imbalance.

The model was trained using a two-phase strategy. First, a 5-epoch warm-up trained only the adapted layers (grayscale conv and final head) while freezing all pretrained weights, using Adam with a learning rate of 1×10^{-3} to avoid destabilizing pretrained features. In the fine-tuning phase (epochs 5–19), all parameters were unfrozen, the learning rate was reduced to 3×10^{-4} with ℓ_2 regularization [1], and a ReduceLROnPlateau scheduler halved the learning rate if validation loss plateaued for two epochs. The dataset was split into 70% training (22,938 samples), 15% validation (4,915), and 15% testing (4,915) using a fixed random seed (42) and a batch size of 32.

4.2 Model Improvements

To systematically optimize the baseline model, individual modifications were evaluated in isolation, then successful improvements were combined iteratively. The modifications tested included: (i) class-weighted loss functions to further penalize under performing classes, (ii) increasing input resolution to 128×128, and (iii) utilizing a deeper architecture (ResNet50). Each variant was trained using identical data splits and hyperparameters, and performance was evaluated on the validation set. The performance of these modifications is discussed in Section 5. The final optimized model incorporated increased input resolution (128×128). In addition, per-class threshold optimization was added to help under performing classes. A dropout (rate 0.35) was added after the final fully connected layer for improved generalization, and training was extended to 30 epochs to allow the model more time to learn the parameters.

5 Results

5.1 Baseline Results

Figure 1 shows the performance of the baseline model described in Section 4.1. This model achieved 87.3% element-wise accuracy on the validation set, 89.6% on the training set, and 87.4% on the test set. The small gap between training and validation accuracy (2.3 percentage points) suggests the model generalizes well without overfitting, indicating that the two-phase training strategy and ℓ_2 regularization were effective. However, with performance below Chen et al.'s 91.48% benchmark, there is clear room for improvement through architectural and training modifications.

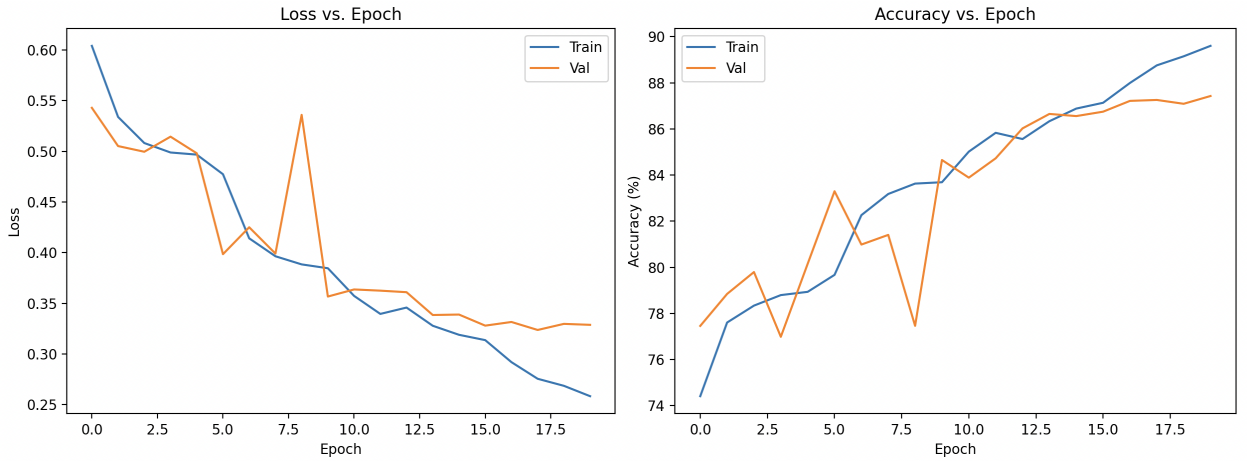

Figure 1: Baseline Model Performance

5.2 Improvement Results

To identify the most effective improvements, each modification described in Section 4.2 was individually evaluated against the baseline. Table 1 shows the impact of each modification on the training, validation, and testing accuracy.

Table 1: Impact of architectural modifications on model performance

Configuration	**Train**	**Val**	**Test**
Baseline (ResNet18, 64×64)	89.6%	87.3%	87.4%
(i) Class-weighted loss	87.9%	86.2%	86.4%
(ii) Higher resolution (128×128)	91.3%	89.7%	**89.7%**
(iii) Deeper arch. (ResNet50)	91.2%	89.4%	89.6%
(iv) ResNet50 + 128×128	87.9%	87.1%	86.4%
(v) (iv) + dropout (0.35)	86.2%	86.3%	86.2%

Higher resolution (128×128) (modification ii) provided the strongest individual improvement, increasing test accuracy by 2.3 percentage points. The increased spatial resolution enables the network to capture finer geometric details in metamaterial unit cells that are critical for bandgap prediction. Deeper architecture (ResNet50) (modification iii) yielded similar gains (2.2 percentage points), suggesting that additional model capacity enabled the model to better extract geometric details key for bandgap classification. Interestingly, combining modifications (ii) and (iii) yielded worse results than even the baseline model. This suggest the model was too deep for the training dataset of (22,938) samples to accurately learn and extract geometric features. Class-weighted loss (modification i) unexpectedly decreased test accuracy to 86.4%, below the baseline. The aggressive class weighting ([1.0, 1.0, 1.2, 1.8, 2.5]) over-penalized false negatives in higher frequency ranges, causing the model to over predict bandgaps to avoid the penalty, thus increasing the false positives, reducing precision and overall accuracy.

Based on these results, the final model incorporated modification (ii); increasing the resolution to 128×128. This model was then further optimized by adding per-class threshold

optimization and dropout (0.35). It was then ran for 30 epochs however, the epoch with the lowest loss achieved on the validations set was also saved. The results can be seen in table 2.

Table 2: Performance comparison at different training epochs

Configuration	**Val. Loss**	**Train Acc.**	**Val. Acc.**	**Test Acc.**	**Test Loss**
Lowest Val. Loss (26 epochs)	0.295	92.8%	90.1%	92.5%	0.269
Extended Training (30 epochs)	0.310	99.1%	94.5%	94.5%	0.357

The final model at 30 epochs achieved 94.5% element-wise test accuracy, a 7.1 percentage point improvement over the baseline and exceeding Chen et al.'s interpretable ML benchmark (91.48%) by 3.0 percentage points. After epoch 26, validation loss increased from 0.295 to 0.310 by epoch 30. While this is generally a sign of overfitting, validation accuracy continued to improve substantially over this period (90.1% at epoch 26 to 94.5% at epoch 30). The test results showed similar trends, with accuracy increasing from 92.5% to 94.5%. Since both validation and test accuracies exhibit similar improvement patterns, this indicates genuine learning rather than overfitting. The perfect agreement between validation and test accuracy at 30 epochs (94.5%) further confirms good generalization, not overfitting. Therefore, the 30-epoch model is selected as the final model for this project, as it provides the best predictive performance on both held-out datasets.

Figure 2 visualizes what geometric features each model focuses on using Grad-CAM activation maps. The 26-epoch model (top) shows activation spread across multiple geometric features, while the 30-epoch model (bottom) focuses more sharply on specific critical regions, particularly stiff-soft material interfaces. This increased selectivity suggests the 30-epoch model has learned to focus on the most diagnostic patterns for bandgap prediction rather than attending broadly to all features, explaining its superior accuracy (94.5% vs. 92.5%). The more focused attention indicates the model has refined its understanding of which geometric configurations are critical for bandgap formation.

The final model demonstrates substantially improved performance across all frequency ranges compared to the baseline model which underperformed in the higher frequency ranges particularly classes 3 and 4 ([3-4] kHz and [4-5] kHz). Table 3 compares per-class F1 scores across model variants.

Table 3: Per-class F1 scores: baseline vs. final model with optimized thresholds

Frequency Range (kHz)	**[0-1]**	**[1-2]**	**[2-3]**	**[3-4]**	**[4-5]**
F1 Score					
Baseline	0.897	0.910	0.811	0.698	0.358
Final Model w/o Optimized Thresholds	0.955	0.954	0.909	0.835	0.680
Final Model	0.952	0.956	0.919	0.853	0.686
Improvement from Baseline	+0.055	+0.046	+0.108	+0.155	+0.328
Classification Threshold					
Baseline	0.50	0.50	0.50	0.50	0.50
Final Model	0.50	0.40	0.50	0.60	0.70

The majority of improvement came from increased resolution (128×128), which raised

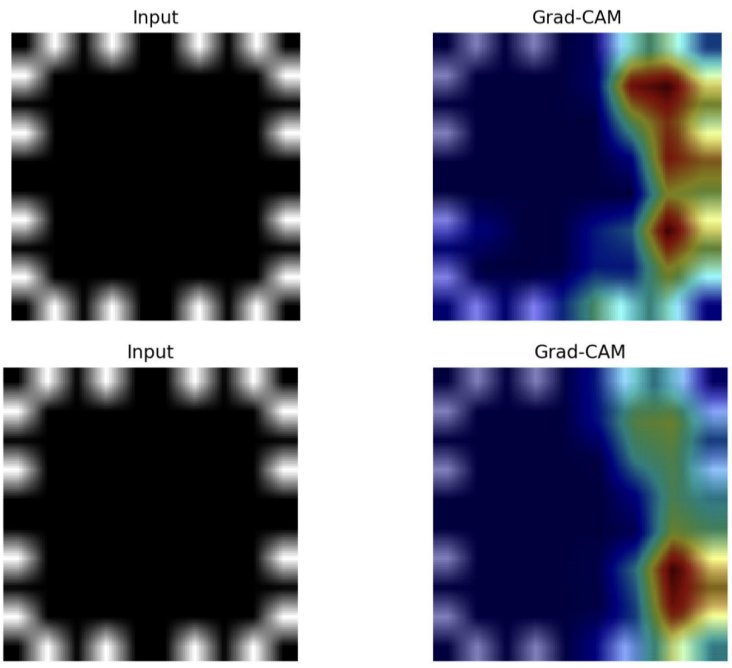

Figure 2: Grad-CAM activation maps comparing model attention patterns. Top: 26-epoch model (lowest validation loss) Bottom: 30-epoch model

class 4's F1 score from 0.358 to 0.680 by capturing finer geometric details. Per-class threshold optimization provided modest additional gains (+0.010 to +0.018 for classes 2-4), with optimized thresholds of 0.40, 0.60, and 0.70 for classes 1, 3, and 4 respectively. The higher thresholds for classes 3 and 4 address a tendency toward false positives in these ranges. By requiring higher confidence for positive predictions, the optimized thresholds reduce false positives and improve precision.

6 Current Conclusions, Future Work

This project successfully developed a convolutional neural network for bandgap prediction in metamaterials, achieving 94.5% accuracy on both testing and validation data. This exceeds Chen et al.'s interpretable LightGBM benchmark (91.48%) by 3.0 percentage points, and surpasses the project's success criterion of achieving at least 93% accuracy. For computational screening of metamaterial candidates, this increase in accuracy justifies the trade-off in interpretability.

The system meets expectations in accuracy and computational cost. Training the model takes 10-12 hours; however, once trained, the model classifies a sample in 25.54 ms (39 samples per second). This represents a dramatic speedup compared to the finite element simulations required to generate the training data, enabling the rapid screening of thousands of metamaterial candidates.

An unexpected pitfall of the model is uneven performance across frequency ranges. While this was addressed through resolution improvements and threshold optimization, classes 3 and 4 still underperform compared to the lower frequency classes (F1 scores of 0.853 and 0.686 vs. 0.95 for classes 0-1). One possible explanation is that geometric features causing

bandgaps at higher frequencies are more subtle and harder to learn than those at lower frequencies. A potential solution would be to generate a larger dataset with additional samples capturing the subtle geometries that produce higher frequency bandgaps, though this would require significant computational resources for finite element simulations.

The biggest area for improvement is the model's ability to generalize to metamaterials with different unit cell discretizations. All training data consisted of 10×10 grid representations, so the model's performance on finer discretizations (20×20, 50×50, 80×80 grids) is unknown. This limits applicability to metamaterial designs that require higher spatial resolution for accurate representation. To address this, samples with varying grid resolutions would need to be generated through finite element analysis, then either tested directly on the current model to evaluate transfer performance, or used to retrain the model on multi-resolution inputs. Ideally, the model would maintain accuracy across all resolutions, enabling flexible metamaterial design at any required level of geometric detail.

As of now, the CNN is a very useful tool for one task: identifying bandgaps in metamaterials given a 10×10 unit cell. An important extension of this work would be inverse design, i.e., property-to-structure sampling. In practice, materials scientists want to design metamaterials with specific bandgap properties, not just identify bandgaps in existing designs. Going forward, this model could be embedded in an optimization framework as a surrogate, with gradient-based methods used to search the design space for geometries that maximize the predicted likelihood of a desired bandgap. The model's learned selectivity for diagnostic geometric features (Figure 2) suggests it has identified physically meaningful patterns that could effectively guide optimization toward viable metamaterial designs. Chen et al. demonstrated high-precision property-to-structure sampling using interpretable models, and adapting a similar inverse-design objective to this CNN-based surrogate would be a natural next step toward turning this project into a full research project.

7 Broader Impacts

This paper introduces a method for rapid bandgap prediction in metamaterials with 10×10 unit cells that can vastly accelerate the design of metamaterials with desired bandgaps, potentially increasing the speed at which medical imaging devices are developed, advancing seismic protection research, and improving wireless communication systems. However, this accelerated discovery carries important risks and limitations. First, metamaterials have significant dual-use applications in military and defense systems, particularly in antenna design and radar signature reduction—metamaterials can manipulate electromagnetic waves to enable smaller, higher-performance antennas and reduce an object's visibility to radar detection systems [4]. Practitioners should carefully consider deployment contexts and prioritize beneficial societal applications over harmful uses. Second, the model sacrifices interpretability for predictive accuracy; unlike Chen et al.'s interpretable methods, this CNN approach does not reveal which geometric features drive bandgap formation, making it difficult for researchers to develop physical intuition or identify when the model might fail on novel designs. Finally, the model achieves 94.5% accuracy but is not perfect and is limited to 10×10 unit cell geometries. Predicted metamaterial performance should always be validated through finite element analysis or experimental testing before deployment in safety-critical applications.

References

[1] Sam Austin. Transfer learning in deep learning: Save time and boost accuracy, July 2025. Medium.

[2] Zhi Chen, Alexander Ogren, Chiara Daraio, L. Catherine Brinson, and Cynthia Rudin. How to see hidden patterns in metamaterials with interpretable machine learning. *Extreme Mechanics Letters*, 57:101895, 2022.

[3] Anastasiia O. Krushynska, Shahram Janbaz, Jae Hwan Oh, Martin Wegener, and Nicholas X. Fang. Fundamentals and applications of metamaterials: Breaking the limits. *Applied Physics Letters*, 123(24):240401, 2023.

[4] John M. McHale III. Executive interview: Metamaterials for military radar, invisibility cloaks, and more. *Military Embedded Systems*, February 2024. Accessed 2025-12-12.

[5] Andrew M. Smith and Shuming Nie. Semiconductor nanocrystals: Structure, properties, and band gap engineering. *Accounts of Chemical Research*, 43(2):190–200, 2010.

We as a group adhered to the honor code. AI was used for sentence structuring and proofreading parts of this report.