IEEE HOST 2023 - CERBERUS

Andrew MacGillivray, Faith Hedges, Tanvir Hossain, Viet Le, Tamzidul Hoque, and Sumaiya Shomaji

Abstract—**Due to the involvement of several untrusted entities in semiconductor supply chain and field operation, the increase of the counterfeit integrated circuit (IC) components has become an alarming issue to the genuine parties. However, the techniques used to create counterfeits – blacktopping, relabeling, using inferior/out-of-spec materials, and tampering – often create significant visual differences between the authentic piece and a counterfeit. Therefore, the creation of an image classification model designed to distinguish between counterfeit and authentic components would yield significant improvements for the genuine entities, especially the manufacturer in terms of safety, performance, and profit.**

I. Introduction

The increasingly globalized supply chain for integrated circuit components has opened the door for counterfeiting as a highly profitable venture, with severe consequences for manufacturers caught unawares [1]. A variety of methods have been used in attempt to identify these counterfeits, including physical and electrical tests [2]. Due to the destructive nature of physical tests and the challenges associated with electrical ones [2], a means of visually identifying counterfeit IC components would constitute a major improvement to the status quo and be invaluable to IC manufacturers.

II. Dataset Analysis

The dataset provided by the HOST 2023 Supply Chain Security competition was used. The provided data for Phase 1, in ”Phase-1_Original_Data”, contains a test set and validation set. Both sets contain photos of counterfeit and authentic IC components, taken via DSLR or Stemi 508 Stereo Microscope. The labels of each image are provided in a CSV file, and are also indicated by the first letter of each image file name (“A” for authentic, “C” for counterfeit), with corresponding encodings of 0 and 1 respectively.

III. Image Preprocessing

A. Features

To obtain a greater understanding of the subject, our team researched indicative characteristics of counterfeit IC components. We tailored our feature extraction methodology to preserve and amplify such features in the data that is input to our model.

We found that counterfeit IC components often have irregular or “noisy” surfaces, malformed labels, incorrect logos, bent and rusty pins, unusual indentations and cutouts, and other visually identifiable flaws. To ensure that our CNN has a high chance of detecting such features, we need to maximize the detail that’s captured and apply useful transformations to the image itself.

B. Feature Extraction Methodology

Using OpenCV, each input image is first converted to grayscale. Then, low-data regions at the peripheral (whitespace) are identified and removed. The resulting image is resized to a fixed resolution that has been selected to balance the time complexity and accuracy of our model.

To extract features, the Laplacian and Sobel operators are then applied. The Sobel operator is computer for both the x- and y-axis. This yields a total of 3 distinct images: the Laplacian, Sobel in the x-direction, and Sobel in the y-direction. Then, the three images are combined and stored as a single RGB image, where the Laplacian is encode in the blue channel, Sobel X in the green channel, and Sobel Y in the red channel.

IV. Counterfeit Detection

To distinguish between circuits in a generalizable form, the Cerberus team employs a Convolutional Neural Network (CNN). A CNN is a Deep Learning Model that convolves sequentially over windows in an image to extract local features. The feature extraction is achieved via layered filters; for example, the first layer may detect an edge or line, while the next may detect a rectangle, and the next detecting an IC pin. Our model contains seven convolutional layers, so this representation would be repeated for each layer.

This model is ideal for the circuit classification task because it can learn the signs of a counterfeit circuit without manual identification of these features; the CNN simply learns by example. A sample of known data is divided into folders representing each category, “Genuine” or “Counterfeit,” and once the images are read, they are assigned a corresponding label. This allows the model to learn from the validation data and adjust accordingly. The weights are set to an initial random state, which causes the model to give a random prediction for an image. If this value differs from the true value of an image, the weights are adjusted via backpropagation, which optimizes the model for the best combination of weights. This enables accurate classification of new images that correspond to the classification of the training images.

A. Training

The CNN was trained on the processed image data. The metrics used were accuracy, loss, validation accuracy, and validation loss. The goal is to minimize loss and maximize accuracy while adequately fitting the data. We modified our training method until we were able to observe a high degree of accuracy ($>60\%$) and reasonable loss (<1) for both the training set and the validation set. By tuning until we achieve similar scores across the training and validation sets, we ensure that the model isn't overfit and can be generalized and applied to novel data.

B. Testing

The unseen test data was given to the model and the resulting metrics were analyzed. To test the model, we used it to predict the labels of all of the images in both the training and testing data sets. The predictions are given as probabilities between 0 and 1. We began by iterating over the results and assigning any probability >=0.5 to 1 (counterfeit), and any probability <0.5 to 0. This method yielded accuracy identical to that observed during training. While inspecting the predictions, we noticed that many of the mislabeled counterfeit components had ambiguous probabilities (values closer to 0.5 than to 0). We altered our algorithm such that any probability greater than 0.15 will be mapped to a counterfeit label. This threshold was used for our final results, and the corresponding confusion matrices are shown in the following section.

V. Results

The final, tuned algorithm was used to predict labels on all of the images from both the "train" and "test" subfolders. The confusion matrix for the training and test sets was computed using tf.math.confusion_matrix().

The original results for the basic model, with a 0.5 probability threshold, are shown in the tables below:

Training Set Confusion		
	True Authentic	True Counterfeit
Predicted Authentic	60	35
Predicted Counterfeit	0	5
Accuracy:	65%	

Test Set Confusion		
	True Authentic	True Counterfeit
Predicted Authentic	10	8
Predicted Counterfeit	0	2
Accuracy:	60%	

The final results, which apply a probability threshold of 0.15 for authentic labels, are shown below:

Training Set Confusion		
	True Authentic	True Counterfeit
Predicted Authentic	60	23
Predicted Counterfeit	0	17
Accuracy:	77%	

Test Set Confusion		
	True Authentic	True Counterfeit
Predicted Authentic	9	4
Predicted Counterfeit	1	6
Accuracy:	75%	

We can see that by requiring a high confidence for the "authentic" label, we dramatically reduce the number of false negatives and improve accuracy overall. Furthermore, the model itself appears robust, with a total accuracy of about 76% across the 120 total predictions.

VI. Discussion

The largest problem our model faces is the false-negative case, where a counterfeit is labeled as authentic. We may be able to address this in the near future by better preparing input, and by tuning the algorithm to give more weight to counterfeit characteristics.

Additionally, we suspect that a more accurate model could first classify whether the image is of the front or back of a chip, and whether the image is from a digital camera or microscope, and then choose a model trained on an appropriately specialized set of data. Such specialized models could allow for higher accuracy while mitigating overfit. However, given the small size of our data set, this would likely be too restrictive. Furthermore, due to time constraints, we were unable to implement popular methods such as data augmentation by, for example, rotating or adding artifacts to the training data.

VII. Conclusion

Image classification will play an important role in addressing supply-chain issues. In this work, we use Laplacian and Sobel operators to extract features from a variety of images, and used those to train a generalized CNN for the recognition of counterfeit integrated circuit components. We then optimized the CNNs predictions by adding a high confidence threshold for the labeling of authentic parts. Across 120 images, our algorithm produced 92 correct labels - an accuracy of 76%.

References

[1] Enahoro Oriero and Syed Rafay Hasan. Survey on recent counterfeit ic detection techniques and future research directions. *Integration*, 66:135–152, 2019.

[2] Ujjwal Guin, Daniel DiMase, and Mohammad Tehranipoor. Counterfeit integrated circuits: Detection, avoidance, and the challenges ahead. *Journal of Electronic Testing*, 30:9–23, 2014.