# Assisting human experts in the interpretation of their visual process: A case study on assessing copper surface adhesive potency

Tristan Hascoet
Kobe University
tristan@people.kobe-u.ac.jp

Deng Xuejiao
Kobe University
dengxuejiao1005@yahoo.co.jp

Kiyoto Tai
MEC Co., Ltd.
tai295@mec-np.com

Yuji Adachi
MEC Co., Ltd.

Sachiko Nakamura
MEC Co., Ltd.

Tomoko Hayashi
MEC Co., Ltd.

Mari Sugiyama
MEC Co., Ltd.

Yasuo Ariki
Kobe University

Tetusya Takiguchi
Kobe University

## Abstract

*Deep Neural Networks are often though to lack interpretability due to the distributed nature of their internal representations. In contrast, humans can generally justify, in natural language, for their answer to a visual question with simple common sense reasoning. However, human introspection abilities have their own limits as one often struggles to justify for the recognition process behind our lowest level feature recognition ability: for instance, it is difficult to precisely explain why a given texture seems more characteristic of the surface of a finger nail rather than plastic bottle. In this paper, we showcase an application in which deep learning models can actually help human experts justify for their own low-level visual recognition process: We study the problem of assessing the adhesive potency of copper sheets from microscopic pictures of their surface . Although highly trained material experts are able to qualitatively assess the surface adhesive potency, they are often unable to precisely justify for their decision process. We present a model that, under careful design considerations, is able to provide visual clues for human experts to understand and justify for their own recognition process. Not only can our model assist human experts in their interpretation of the surface characteristics, we show how this model can be used to test different hypothesis of the copper surface response to different manufacturing processes.*

## 1. Introduction

Humans are experts in communicating the reasoning process behind their answer to visual questions. For instance, on typical Visual Question Answering (VQA) samples, human annotators are often able to convincingly justify, in natural language, the reason behind their answer to a certain visual question using simple common sense reasoning. In contrast, deep Learning models are often viewed as black box predictors lacking interpretability in the sense that existing tools often fail to explain the decision making process behind the models predictions. For instance, a deep learning model trained end-to-end on a VQA dataset may be able to provide the same answer as its human counterpart, but the process through which the model reaches this answer is entirely opaque.

While it is true that humans can justify for their answers on high level reasoning tasks, humans also often fail to explain the process behind their low-level feature recognition ability: For example, precisely defining the nature of a specific texture (what are the defining features of a plastic or wooden surface?) or specific low-level part attributes exhibiting large intra-class variations (what is the defining features of a "leg" or a "wing"?) is a very difficult task. Humans constantly perform such low-level visual recognition tasks while being unable to precisely justify for their own recognition process.

In this paper, we present one very practical instance of such situation in the Printed Circuit Boards (PCB) industry, in which expert material scientists are tasked with assessing the adhesive potency of copper surfaces. We propose a model that, under careful design considerations, is able to provide visual clues for human experts to understand and justify for their decision process.

The proposed model is designed so that a subset of its internal representations carry semantically meaningful information that can be visualized and easily interpreted by humans. Providing these visual clues, however, comes with the cost of imposing additional constraints on the architecture, which we found to degrade the model accuracy: In-

deed, we found that networks with unrestricted architectures, (which do not provide interpretable features) perform better than network architectures restricted so as to provide semantically meaningful representations. This is because, as we restrict the architecture of the model, we formulate an assumption on the impact of the manufacturing process on the surface statistics which may not hold in reality. This result suggests an inherent trade-off between expressivity and the explainability in designing model architecture.

While the degradation of the model accuracy is problematic from a performance perspective, it offers an interesting opportunity from an explainability perspective: As the model accuracy degrades due to the inadequacy of the assumption made by the model architecture, we can use the model accuracy as a proxy metric for the adequacy of different assumptions. This allows us to quantitatively assess different assumptions regarding the impact of manufacturing processes on the copper surface. This may prove useful to quantify the impact of manufacturing process on cooper surface adhesive potency and eventually help optimize the manufacturing process.

In essence, the argument this paper is aiming for is as follows: although deep learning models lack the "common sense reasoning abilities and the powerful formalism of natural language to communicate and justify for their decision making process, they can provide powerful to explain low-level recognition processes.

In practice, the contribution of this paper is as follows:

1. We formalize a segmentation procedure based on a probabilistic weak label segmentation framework.

2. We introduce a formalism to show how the model accuracy can be used as a proxy metric to quantify the validity of different assumptions on the dataset.

The remainder of this paper is organized as follows: In Section 2, we present some background information on the motivation for this project: We start by discussing the importance of copper surface adhesive potency, and detail the dataset used in our experiments. Section 3 details our contribution. Section 4 briefly relates our work to different research topics and Section 5 presents the results of our experiments. Finally, Section 6 further discusses the relevance of our results, insisting on the limitations of our assumptions to conclude this paper.

## 2. Background

Printed circuit boards (PCBs) are integral part of a wide variety of electronic devices including industrial and household appliances (e.g. TV and PC), mobile communication devices and automobiles. PCBs play an important role in electrical connection between electronic components. Copper has been used in the PCBs industry as the conducting material, and the electric copper circuits are isolated from each other by insulators (solder resist, prepreg etc). A multilayered PCBs have a laminate having a plurality of electroconductive layers with insulating layers interposed therebetween. So there are many interfaces related to the copper and resins in PCBs. Figure 1 illustrates the organization of such an electric circuit.

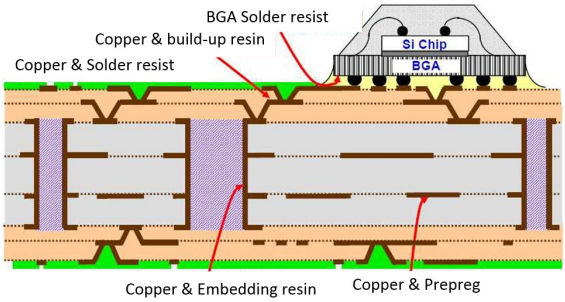


Figure 1. Illustration of a Printed Circuit Board. Copper circuits are made with various insulators for several purposes.

Since PCBs have been required to have higher heat-resistant properties in recent years, copper surface treatment technologies have performed a more and more important role in the manufacturing process. This is because they can enable PCBs to maintain high copper to resin adhesion even under harsh conditions. Copper surface treatment (copper surface roughening) offers one of the most effective ways to increase adhesion of the interfaces. Copper surface roughening has been widely used for the purpose of increasing adhesion of copper to resins. It produces a unique surface topography which enhances the mechanical bonding of copper to resins, as illustrated in Figure 2.

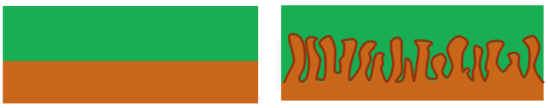

Figure 2. Illustration of a copper roughness surface. (Left) a perfectly smooth surface provides small adhesive surface as the interface between copper, in brown, and the resin, in green is minimal. (Right) a rough surface provides a larger surface at the interface of the resin. Larger contact surface areas provide higher adhesive potency.

In very broad terms, rough surface allow for stronger bonds as the asperities of the surface provide a wide range of adhesive surface area and an anchoring effect. In contrast, smooth surfaces dont provide such effects so that they have lower adhesive potential. In the remaining of this paper, we will refer to the potential bonding strength of a copper surface as its "adhesive potency". It is also important to note that the adhesive potency of a copper surface is related to its "roughness", which is observable at the microscopic scale.

Electronic substrate manufacturers have developed advanced manufacturing processes to shape the surface of copper sheets in order to increase their adhesive potency. This is typically achieved by applying an etching solution on the copper surface. Being able to accurately assess the adhesive potency of a copper surface would allow to further optimize manufacturing processes to increase the reliability of electronic devices. However, assessing the adhesive potency of a copper surface is a complex task, even for the most expert practitioners. Hence the motivations of this study is two fold: First we aim to automate the evaluation of a copper sheet adhesive potency from microscopic imaging of its surface. Second, we aim to better understand the process through which Towards this goal, we built a dataset of microscopic images of copper surfaces, which we detail below.

We imaged copper surfaces using Scanning Electron Microscopy (SEM) at a resolution of $xxx$ micro meters. To investigate the impact of different manufacturing processes on the copper surface, we applied 16 different etching solutions, with decreasing etching power, to the copper surface. For each of these solutions, we captured 50 SEM images of $960 \times 1280$ pixels so that the full dataset consists of 800 $(16 \times 50)$ images. Each image is annotated with a label $t$ corresponding to the etching solution used to shape the copper surface. Each solution was obtained by submitting the original solution $t = 0$ to an extreme stress test for a period of time $t$. Hence, we know that for all images of copper surfaces with label $t$ show higher adhesive potency than the images labelled with $t' > t$. However, we do not know the *exact* impact of the stress test on the surface adhesive potency.

## 3. Method

### 3.1. Dataset and Notations

We denote the dataset described above as $\mathcal{D} = \{(x_i, t_i)|i \in [1, xxx]\}$ where $x$ denote images $x \in \mathbb{R}^{H \times W}$ and labels $y \in [0, xxx]$ corresspond to the time of stress test applied. We split the dataset $\mathcal{D}$ into a training $\mathcal{D}_{tr}$, validation $\mathcal{D}_{val}$ and test $\mathcal{D}_{te}$ set so that the number of images $x$ per label $y$ in each set are 40 for training set, and 5 for the validation and test sets.

### 3.2. Baseline

We start by establishing a strong baseline for our study. The baseline architecture follows standard convolutional network designs for image classification. This architecture, illustrated in Figure 4, is made of several residual blocks sequentially interleaved with max pooling operations. Each residual block consists of $n$ repetitions of a sequence of $3 \times 3$ Convolution, Batch Normalization and ReLU layers, followed by a residual skip connection. We set $N$ residual

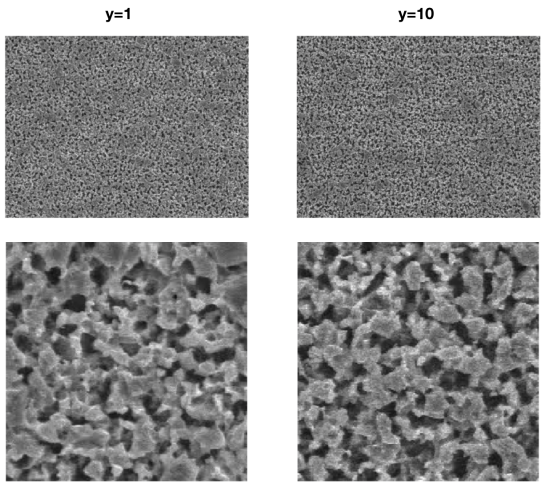


Figure 3. Illustration of a few images from the dataset. (Top) Full images. (Down) Zoomed-in areas of $200 \times 200$ pixels. (Right) Sample image of label $y = 1$. (Left) Sample image of label $y = 10$. The difference between both images are minimal to an untrained eye. Precisely defining the visible difference with words is a difficult task.

blocks between every max pooling layer and we denote by $d$ the number of pooling layers. Hence, the full depth $D$ of the network (in number of convolution layers) is given by $D = d \times n \times N + 1$, where the term 1 corresponds to the initial $3 \times 3$ Convolution layer happening before the first pooling operation. Finally, the top layer of the network is made of a global average pooling layer followed by a linear softmax layer with output dimension 15 corresponding to our number of classes. For simplicity and contrary to standard practices, we keep the number of channels $c$ constant in all layers of the network.

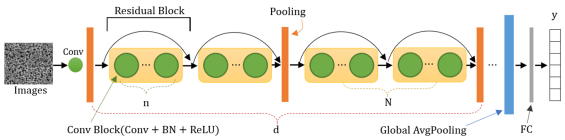


Figure 4. Illustration of our baseline architecture. With our modular architecture definition, the architecture is fully defined by parameters $n$,$N$,$c$ and $d$.

With this parameterization, our network is fully specified by the four hyper parameters $c$, $d$, $n$ and $N$. We performed a grid search over these hyper parameters to select the best performing architecture. The details of this architecture search are given in the experiment section, and, as we shall see then, the best performing architecture performs significantly better than human experts. However, the decision process through which this model reaches such a high accuracy is entirely opaque as the distributed nature of the model's internal representations provides little interpretability. The remainder of this section details our attempt to de-

sign an architecture that can provide useful explanations of the process through which high accuracy recognition can be performed.

### 3.3. Assumptions

The basic idea behind our method is similar to the visual insights provided by the visualization method of [] and the attention mechanisms in visual models for image captioning and question answering tasks: We would like to evaluate the contribution of each input pixel to the final classification decision.

To do so, we modify our initial problem formulation into a segmentation task: Given an input $x \in \mathbb{R}^{H \times W}$, we want to design a model that outputs a segmentation mask $s \in \mathbb{R}^{H \times W}$ assessing the contribution of each inidividual pixel to the output adhesive potency score (i.e., the output class $y$). Training a typical segmentation model for this task would require ground truth segmentation masks $s$ for each image $x$ of the training set. However, manually annotating ground truth segmentation masks for this task is not feasible, as human experts are not able to provide such fine-grained annotations. Instead, we have to train the segmentation given a single ground truth label $t$ per image. In order to do so, we make the following assumption:

**Assumption**: For a given image $x$ with label $t$, the *pixel-wise* values of the true (unknown) segmentation mask $s$ follow a spatially stationary binomial distribution whose expected value, averaged over the spatial dimensions of $x$, is given by $t$ following:

$$s_{hw} \sim \mathbb{B}(f(t)), \forall h, w, t \in H \times W \times T \tag{1}$$

in which we introduced a target function $f$, which we shall discuss in Section xxx.

In other words, we suppose that there exists a true binary segmentation map $s$ assigning to each individual pixel of $x$ a binary adhesive potency score: $s$ takes 0 values in regions of the surface providing low adhesive potency (i.e. smooth copper surface areas) and 1 values in regions of the surface providing high adhesive potency (i.e. rough copper surface areas). The adhesive potency score of an entire image $x$ is thus given by the average of the pixel-wise values, and this average value is uniquely defined by the image label $t$.

Figure 5 provides a visual illustration of this idea. For each image, we suppose the existence of such binary mask $s$ quantifying the adhesive potency of local areas of the surface. The average value of the binary mask $s$ amounts to the ratio of the copper surface covered in white (i.e. with value 1, corresponding to high adhesive potency). Our assumption means that this ratio stays constant for different images $x$ sharing a similar label $t$, and takes a value given by the target function $f$.

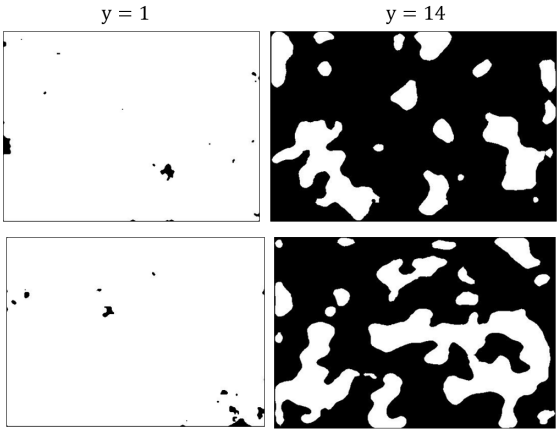


Figure 5. Illustration of binary segmentation masks. White pixels represent areas of high adhesive potency and black surface represent areas of low adhesive potency. We make the assumption that the ratio of white surface is constant for different samples (top and bottom) with equal label $t$. The exact ratio is defined by the target function $f$.

### 3.4. Architecture

In this section, we present the architecture used to compute binary segmentation masks from input images. Our architecture extends the baseline architecture presented in Section 3.1, with an ascending path that progressively upsamples the output of the descending path, similar to the UNet architecture [**?**] and illustrated in Figure 6.

Residual modules of the axcending path are the exact symmetric of their anolog in descending path. The ascending path uses bilinear upsampling layers (illustrated in xxx), instead of the max pooling layers of the descending path. Different from the UNet architecture, and following previous works [**?**], we merge the outputs of the descending path modules with the inputs of the ascending path by summation, instead of concatenation. We alos use valid convolutions to preserve the spatial resolution of the output. Finally, we add a sigmoid layer at the top of the network in order to bound the output values between 0 and 1.

### 3.5. Loss Function

Given a segmentation model $M_\theta$, with weight parameters $\theta$, and an input image $x$, we denote the average of the model output by $m_\theta(x)$:

$$M_\theta : \mathbb{R}^{H \times W} \to [0,1]^{H \times W} \tag{2a}$$

$$\tilde{s} = M_\theta(x) \tag{2b}$$

$$m_\theta : \mathbb{R}^{H \times W} \to [0,1] \tag{2c}$$

$$m_\theta(x) = \frac{1}{HW} \sum_{h,w} \tilde{s}_{hw} \tag{2d}$$

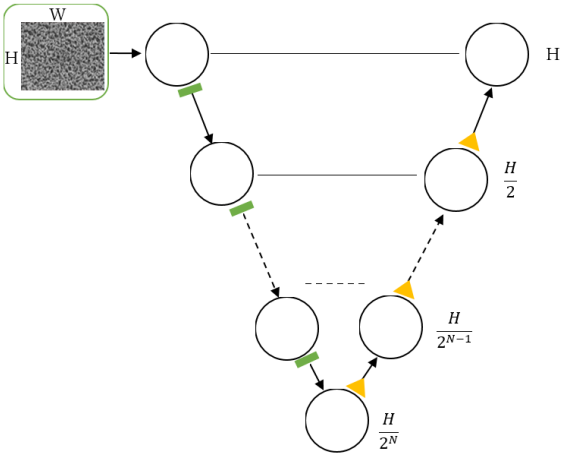


Figure 6. Illustration of the segmentation model architecture. This architecture follows standard practice in UNet-like segmentation architectures.

We can then train the model by regressing the average value of the segmentation mask to the target label given by $f(t)$. Given a training dataset of labeled samples $\mathcal{D}_{\sqcup\triangledown} = (x_i, t_i)$, learning is thus done by minimizing the following loss function over the model's parameters $\theta$:

$$\theta^* = argmin_\theta \mathbb{E}_{x,t\in\mathcal{D}_{tr}} ||m_\theta(x) - f(t)||^2 \qquad (3)$$

However, as we shall see in the next section, the target function $f$ represents an unknown ideal function, so we do not have access to the actual values of $f(t)$. Instead, we will approximate $f$ with a known hypothesis function $g \approx f$, so that the actual training loss used in our experiments is:

$$\theta^* = argmin_\theta \mathbb{E}_{x,t\in\mathcal{D}_{tr}} ||m_\theta(x) - g(t)||^2 \qquad (4)$$

### 3.6. Target Function

In section 3.3, we have made the assumption that labels $t$ uniquely define the expected value of the ground-truth binary segmentation masks $s$ through a target function $f$. In this section, we discuss the role of this target function.

The target function $f(t)$ describes the evolution of the copper surface adhesive potency with time $t$. More precisely $f$ defines the evolution of the *ratio of adhesive surface area* with time.

However, $f$ is an ideal, unknown function, of which we have only supposed the existence. We do not know the value taken by $f(t)$ for a given $t$ because we do not know the exact impact of the manufacturing process on the copper's surface characteristics. We thus introduce a known hypothesis function $g$ to approximate the ideal target function $f$. $g$ expresses our belief of what the values taken by the true function $f(t)$ are. Although we do not know the exact values taken by $f$, we know several of its characteristics, which we can use to reduce the search space of hypothesis functions $g$:

In Section 2, we have established that, for all $t$, copper surfaces with label $t' > t$ should have lower adhesive potency than copper surfaces of label $t$. Hence, $f$ should be a monotonically decreasing function of time, and we can thus restrict our search of hypothesis function $g$ to monotonically decreasing function. For example, the simplest of such function would be the linear function, taking linearly decreasing values from 1 to 0, which we analyse in the experiment section.

Second, as $f$ describes the ideal, true evolution of the copper surface statistics with $t$, so that training an ideal model with the supervision signal provided by $f$ should yield to zero test errors. Hence training a segmentation model with a hypothesis $g \approx f$ should lead to low test errors. This means that we can use the error of the model on the held out validation dataset as a proxy measure on the validity of the hypothesis function $g$.

In the experiment section, we evaluate the model errors with different hand-crafter hypothesis functions $g$. In future work, we aim to learn the hypothesis function $g$ together with the model parameters.

### 3.7. Learning the Hypothesis Function

## 4. Related Work

We identify three different lines of research that share similarities with our study: Explainability of learned visual representations, weak supervision of segmentation models and machine learning applications for material science. Our can be seen as a weakly supervised approach to provide material experts with strong supports for interpretable decision, which falls at the intersection of these three research lines. We briefly present some of these works in the following subsections

### 4.1. Explaining CNN Representations

Although deep neural networks have achieved a great success on a variety of challenging visualization tasks in recent years, our understanding of how these neural network models is far from enough to interpret. The pursuit of figuring out what is learned for each layer of models and how trained neural network models really "think" never stops.

Jason Yosinski et al. provide two useful tools for visualizing and interpreting neural nets. One is to visualize the activations produced on each layer of a trained convnet as it processes an image or video, and the other enables visualizing features at each layer of a DNN via regularized optimization in image space. Shan Carter et al. create an explorable activation atlas of features the network has learned, by using feature inversion to visualize millions of activations from an image classification net-

work, which can reveal how the network typically represents some concepts. There is a new dataset called Visual Question Answering(VQA) containing open-ended questions about images.These questions require an understanding of vision, language and commonsense knowledge to answer. Trevor Darrel et al. build models (such as Pointing and Justification-based explanation model) to explain their decisions, generating convincing explanations. TensorFlow provides an attention-based model, which enables us to see what parts of the image the model focuses on as it generates a caption.

### 4.2. Weakly Supervised Segmentation

An other point we should pay attention to is different labeling strategies in conventional supervised learning. Typically speaking, it's often assumed that each instance is associated with one single label. However, there should usually be more than one labels for one instance in real-world tasks, if it is multi-label learning.

Eugene et al. propose a semi-supervised framework that employs image-to-image translation between weak labels. Yuxing Tang et al. build a similarity-based knowledge transfer mode trying to investigate whether knowledge about visual and semantic similarities of object categories can help improve the peformance of detectors trained in a weakly supervised setting.

The most related work with ours is Rihuan Ke's[*] in which they present a semi-supervised learning strategy for segmentation with lazy labels and develop a multi-task learning framework to integrate the instance detection, separation and segmentation within a deep neural network.

### 4.3. Machine Learning for Material Science

Deep learning has showed remarkable success on many important learning problems in chemistry, drug discovery, biology and materials science.

In the field of materials science, deep neural networks have also been receiving inreasing attention and have achieved great improvements, for example, in material property prediction and new materials discovery for batteries.

## 5. Experiments

### 5.1. Classification results

We start by evaluating the baseline classification model described in Section 3.3 and illustrated in Figure 4. The hyper parameters $n$, $N$, $d$, and $c$ of the classifier architecture were obtained by a grid search within the limits of a 12GB Nvidia GPU memory size . The model was trained on the training dataset for 200 epochs using the Adam optimizer with default parameterization.

We compare the classification outputs of the baseline model to those of an expert material scientist on a blind test. The human expert was given the sample images of the training set to practice, and we evaluated the expert's answers on the samples of the test set.

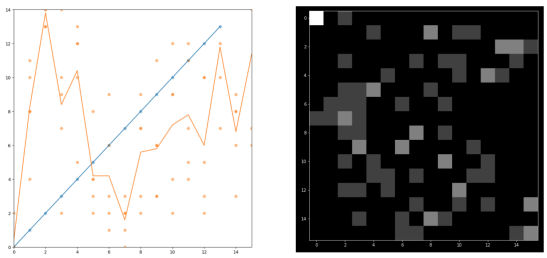

Figure 7. Results of the human evaluation.

Figure 7 shows the results achieved by the model and Figure 8 shows the results achieved by our baseline model. As can be seen in these figures, the model tends predict the manufacturing process more accurately than the human evaluator. However, these results should be taken with a grain of caution as the human expert was given little to practice on this specific dataset, while the model was selected as the best performing baseline from an extensive parameter search. We plan on re-conducting the human evaluation in the updated version of this paper.

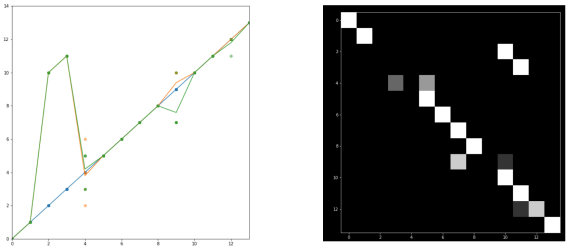

Figure 8. Results of the model evaluation.

Interestingly, however, the human expert seems to accurately recognize the surfaces of highest adhesive potency as seen in the upward trends of his results for low $y$ values. For example, he easily identified the surface for $y = 0$. On the other hand, his guesses for low adhesive potency are much more random.

This is in stark contrast with the classifier accuracy, which perfectly identifies the low adhesive potency surfaces (high $y$ values), but consistently misclassifies surfaces of high adhesive potency (i.e.; for $y = 1$ and $y = 2$).

### 5.2. Visualization of Segmentation Results

To motivate our results, Figure 9 illustrates an output of the segmentation model on a small patch of an input image $x$. To a novice observer, the model seems to assign lower

adhesive potential to smoother regions of the input. In future work, we plan on further exploring these visualizations with human experts to better understand the patterns characterizing the surface adhesive potency.

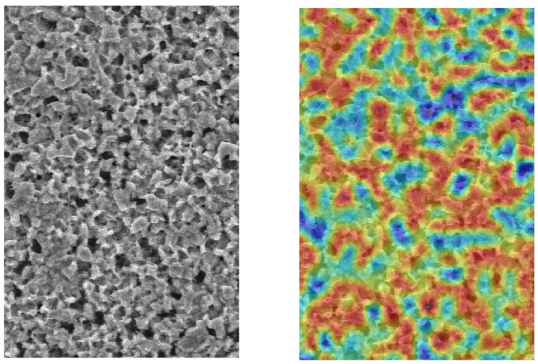

Figure 9. Visualization of the segmentation model output.

### 5.3. Investigation of the hypothesis function

In this section, we question the role of the hypothesis function on the accuracy of the model. As illustrated in Figure 10, we investigate the simplest possible hypothesis, $f(y) = y$. This hypothesis function formulates the idea that for $y = 0$, all the segmentation mask values should be 0, so that the whole surface area should have strong adhesive potency, while for $y = 14$, all the segmentation mask values should be 1, so that the whole surface area should have low adhesive potency. For each $y$ between 1 and 14, this hypothesis function defines a linear increase of the low adhesive potency surface ratio.

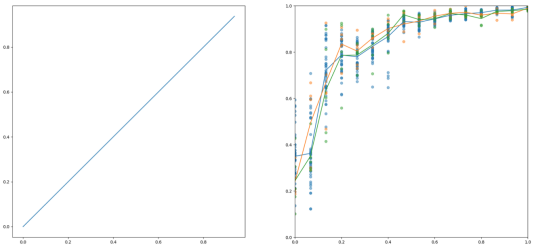

Figure 10. Results of the segmentation model evaluation.

Figure 10 compares the hypothesis function to the actual model output on the training, validation and test set. First, we observe no overfitting as all datasets yield similar average results. Second, we observe that instead of following a linear trend from 0 to 1, the surface ratio seems to evolve more similarly to a squared root or logarithmic function with a sharp increase in value for low $y$ and a slower increase in higher values of $y$. It is interesting to note that the same behavior is also observed on samples of the training set, for which the model was explicitly trained to reproduce the linear hypothesis function.

This result suggests that the surface ratio (i.e. the smoothing effect of the stress test on the copper surface) may increase sub-linearly with time, either as a square root or logarithmic trend. Investigating the impact of different hypothesis function is an interesting question we will consider in future work.

## 6. Discussion & Conclusion

In this paper, we have argued that deep neural networks can be used to help explain the process behind low-level recognition tasks. We have focused on the task of assessing the adhesive potency of copper surfaces in the context of PCB manufacturing.

This is an interesting task to showcase the limitations of human's ability to explain their visual process as it is a very low level recognition task for which trained experts can provide qualitative guesses while not being able to fully justify for their guess.

On this task, we have first shown that an unrestricted classifier architecture can outperform a human expert in accuracy. However, this architecture does not bring us any insight into *why* a given copper surface should be classified as a high or low adhesive potency.

To shed some light into the decision process of the model, we proposed to cast this problem as a a segmentation problem, in the model assigns a binary score to each pixel indicating wehter this pixel belongs to a high or low adhesive power local area of the surface. Visualizing the segmentation output of the model may prove useful for human experts to better understand the characteristics of high and low adhesive potency surfaces.

Finally, we developed a weak label training procedure to train this segmentation model. Our procedure relies on a hypothetical relationship between the manufacturing process and the copper surface roughness. By training the network on the hypothesis that the surface ratio of low adhesive potency increases linearly with the stress test time, we observed that the model output follows a log-like evolution.

However, the result of this last experiment should be taken with caution as this study is still in a very preliminary stage. In particular, our training procedure relies several assumptions that most likely do not hold in reality: In particular, we assumed that each individual area could be represented by a single binary value defining its adhesive potency while, in reality, the adhesive potency of local areas are most likely not binary in nature.

Nevertheless, these preliminary results are encouraging, and we will continue our analysis in future work.

## References