2024 IEEE 74th Electronic Components and Technology Conference (ECTC) | 979-8-3503-7598-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/ECTC51529.2024.00388

Demystifying Edge Cases in Advanced IC Packaging Inspection through Novel Explainable AI Metrics

Shajib Ghosh
ECE Department
University of Florida
Gainesville, USA
shajib.ghosh@ufl.edu

Antika Roy
ECE Department
University of Florida
Gainesville, USA
antika.roy@ufl.edu

Md Mahfuz Al Hasan
ECE Department
University of Florida
Gainesville, USA
mdmahfuzalhasan@ufl.edu

Patrick Craig
ECE Department
University of Florida
Gainesville, USA
pcraig1@ufl.edu

Nitin Varshney
ECE Department
University of Florida
Gainesville, USA
nitinvarshney1@ufl.edu

Sanjeev J. Koppal
ECE Department
University of Florida
Gainesville, USA
sjkoppal@ece.ufl.edu

Navid Asadizanjani
ECE Department
University of Florida
Gainesville, USA
nasadi@ece.ufl.edu

***Abstract*—Inspecting advanced packaging systems for integrated circuits presents difficulties due to the intricacy of contemporary semiconductor packages, which include small dimensions, dense structures, and diverse materials. The intricate nature of these packages necessitates unparalleled precision in identifying features such as microbumps and through-silicon vias (TSVs). Explainable AI (XAI) algorithms play a vital role in providing clarity in decision-making processes and facilitating the comprehension of anomaly detection. However, the challenge lies in the absence of standardized metrics to assess the reliability of XAI algorithms. This research endeavor aims to establish industry-standard metrics for evaluating XAI techniques in the physical inspection of advanced packaging systems, thereby enhancing transparency and dependability in complex semiconductor scenarios.**

***Index Terms*—Advanced Packaging, Explainable AI, Metrology, Physical Inspection, Edge Cases**

I. Introduction

Inspecting advanced packaging systems for integrated circuits (ICs) presents challenges in the semiconductor industry due to the intricate nature of these packages. As semiconductor packages decrease in size to accommodate higher component densities, accurate identification of small attributes such as microbumps and through-silicon vias (TSVs) becomes crucial. The presence of three-dimensional structures, stacked dies, and various materials adds complexity, resulting in variations that hinder visualization [1] [2]. Embedded components, such as capacitors and resistors, introduce an additional layer of complexity, necessitating specialized techniques for examination. The reliability of critical components like TSVs, which are essential for three-dimensional connectivity, requires careful scrutiny to identify flaws that affect electrical connectivity. Precise alignment is vital, especially when dealing with unusual shapes. Challenges also arise with advanced packaging materials such as low-k dielectrics, where the ability to detect defects is crucial. Dynamic imaging techniques are necessary to address potential failures over time, and continuous advancements in imaging technologies are essential for reliable inspection processes in advanced packaging systems [1].

Explainable AI (XAI) algorithms [3] can play a significant role in understanding the complexities of state-of-the-art packaging systems, providing transparency in decision-making processes. XAI algorithms possess the capability to not only identify but also explain variations in features such as microbumps and TSVs, which are crucial for establishing trust and understanding the significance of components like TSVs in three-dimensional connectivity. Despite the potential advantages of XAI, the lack of standardized assessment measures presents a challenge. This paper focuses on developing industry-standard measures to evaluate the reliability of XAI techniques in explaining exceptional cases in the physical examination of IC advanced packaging. The investigation utilizes cutting-edge imaging systems such as X-ray imaging on various advanced IC packages to demonstrate the impact of the proposed measures on improving accuracy and dependability. An initial examination of the impact of XAI assessment on PCB component detection based on X-ray images is conducted, with knowledge transferred to microbump detection on IC advanced packaging images. Despite the labor-intensive nature of acquiring and annotating advanced IC package images, this study establishes a framework for evaluating XAI techniques in IC advanced packaging inspection, including two unique metrics called model performance retention (MPR) that measure the degree to which XAI features retain model performance and the context relevance score (CRS), which measures how image context influences XAI fidelity.

Section III will explore XAI metrology in a more comprehensive manner, providing detailed insights. The subsequent sections of the paper are structured as follows: section II

2377-5726/24/$31.00 ©2024 IEEE
DOI 10.1109/ECTC51529.2024.00388

explains the challenges associated with X-ray and scanning acoustic microscopy (SAM) imaging modalities in inspecting advanced IC packaging. This section also discusses the potential and challenges of integrating XAI methods into this field along with highlighting the importance of XAI metrology in evaluating these methods. Section III presents the datasets used in experiments and defines the metrics in a systematic manner. The results obtained are presented in section IV, and section V explores key takeaways and proposes future research inspired by this study. Finally, section VI summarizes our approach, highlighting the incorporation of XAI methods and their evaluation modes in the context of IC advanced packaging inspection, emphasizing the crucial role in ensuring accuracy in emerging inspection techniques within this field.

II. Background

A. Inspection of IC Advanced Packaging

1) Challenges with SAM for Advanced Packaging Analysis: High-resolution acoustic imaging in 3D HI devices relies on factors like transducer frequency, aperture, sound velocity, focal length, and maintaining constant flight time during scanning [4]. A higher-frequency transducer enhances resolution but compromises the signal-to-noise ratio (SNR), introducing noise in SAM images. Time-of-flight (TOF) poses challenges in high-resolution imaging of 3D HI interconnections [5]. Surface triggers assist focus maintenance, yet certain areas remain prone to defocus, demanding consistent TOF compensation for artifacts. The application of SAM in interconnection detection faces challenges in interpreting defects from low-resolution acoustical imaging [6]. Even at higher frequencies, C-Scan resolution is inadequate for defect localization in advanced packaging. Stacked die complexities further complicate detection, as thinner layers with smaller spacing and shorter ultrasound delay times between interfaces make failure detection challenging, leading to potential false readings by CSAM [4]–[6].

2) Challenges with X-ray for Advanced Packaging Analysis: The semiconductor industry faces difficulties in dealing with the shrinking size of integrated circuit features and the widespread use of heterogeneous and wafer-level packaging [7] [1]. These challenges require higher resolution imaging techniques to handle larger samples such as advanced packages and wafers. However, current non-destructive 2D and 3D imaging methods struggle to detect sub-micron defects in interconnections smaller than 10 μm in diameter, especially in a timely manner. The larger size of modern high-density packages presents additional challenges for existing 3D X-ray tools, which are unable to provide enough detail for fault isolation in small-scale interconnections without extensive processing time. Even 3D X-ray microscopy, which is commonly used for chip analysis, faces difficulties when inspecting 300 mm wafers during wafer-level bonding or large system-in-package (SiP) end-products. This is due to practical limitations caused by the need for the sample to be positioned at a certain distance from the source for full rotation, as the Inverse Square Law prevents achieving sub-micron resolution in 3D tomography for a 300 mm wafer with current methods.

B. Prospects and Challenges of Explainable AI

The application and development of complex AI models have spurred innovation across many domains. Increases in work efficiency, refined decision-making, and streamlined automation represent key advantages stemming from the application of AI models. However, it is well-documented that AI models may produce erroneous outcomes when confronted with input data that deviates from the training dataset [8]. Such deviations, often not predictable before model deployment, can only be realized and adjusted for post-application with model refitting and the inclusion of more diverse samples in the training dataset. These inaccuracies not only undermine trust but precipitate adverse consequences when AI is prematurely deployed. Especially in critical domains like medical, military, and financial applications, small errors in an AI model's decision-making process can yield negative consequences, potentially affecting human lives. Consequently, research and development of explainable artificial intelligence (XAI) techniques serve to enhance the interpretability of results and provide model transparency to an AI model's decision-making. The benefits of increasing model transparency through XAI techniques are four-fold. According to Samek, XAI techniques can enhance the verification and improvement of AI systems, the derivation of knowledge from said systems and help with compliance legislation [3].

C. XAI Metrology for Physical Inspection

Explainable Artificial Intelligence (XAI) metrology systematically evaluates the clarity of XAI systems, with the goal of ensuring comprehensibility and reliability in AI decision-making. Cutting-edge metrics include faithfulness, stability, and comprehensibility. Faithfulness guarantees that explanations align with model behavior, stability ensures consistency across instances, and comprehensibility assesses the clarity of explanations. These metrics collectively foster transparency and trust in AI systems; however, a universal metric for all XAI applications is currently lacking [9]. Different contexts may necessitate diverse criteria. Despite the complexity of AI decision-making, there is a noticeable gap in utilizing XAI assessment to enhance comprehensibility in IC advanced packaging or PCB inspection. This study introduces innovative metrics, with a focus on the Local Interpretable Model-Agnostic Explanations (LIME) method [10]. Additionally, we apply various explanations to PCB and IC advanced packaging images obtained through X-ray imaging techniques.

III. Methodology

A. Datasets

This study utilizes a primary dataset comprising 44 X-ray images, consisting of 40 samples of printed circuit boards (PCBs) with six different types of components and 4 samples of integrated circuit (IC) advanced packaging with a single type of component. In order to facilitate explanation methods

based on features (e.g., Local Interpretable Model-Agnostic Explanations or LIME [10] and SHapley Additive exPlanations or SHAP [11]), an additional dataset (Dataset 1) was created. This dataset includes 15 texture and shape features, as well as relevant metadata such as file names, component ID, component types or classes, and sample type. Dataset 2 was derived by isolating four components (vias, solder balls, pin grid array or PGA, and microbumps) from the images. The goal was to address challenges in differentiation between these components, which arise due to shared structural and appearance similarities. The use of counterfactual explanations and saliency maps is crucial for revealing visual and contextual distinctions among these components, particularly considering the absence of benchmark datasets for training deep learning models capable of accurate differentiation. A concise overview of both datasets is provided below.

1) Dataset 1: As mentioned earlier for LIME [10] and SHAP [11] implementation and evaluation, we extracted 15 feature values (8 texture features and 7 shape features). Details of these feature defintions can be found in [12], [13]. We also have considered smaple type (e.g., PCB or IC) as an additional feature to predict class labels and explain contextual relevance in terms of sample type. Table I represents a brief overview Dataset 1.

TABLE I
DATA DISTRIBUTION FOR EXPERIMENTATION USING LIME AND SHAP

Features		Component Types		Data Distribution		
Texture Features	**Shape Features**	**Classes**	**Class ID**	**Total Data Samples**	**Training Set**	**Testing Set**
GLCM Contrast	Area	Solder Balls	0	7347	5877	1470
GLCM Correlation	Perimeter	PGA	1			
GLCM Energy	Circularity	Copper Planes	2			
GLCM Homogeneity	Solidity	Pads	3			
Tamura Contrast	Convexity	Traces	4			
Tamura Coarseness	Eccentricity	Vias	5			
Tamura Directionality	Aspect ratio	Microbumps	6			
Entropy						

2) Dataset 2: For counterfactual (CF) explanation, we cropped images of solder balls, PGA and vias from X-ray images of PCB and microbumps from X-ray images of IC advanced packaging. The data distribution for CF experiment is provided in Table II.

TABLE II
DATA DISTRIBUTION FOR THE COUNTERFACTUAL EXPERIMENT

Components	Train	Test
Solder Balls	60	21
PGA	60	17
Via	80	40
Microbump	50	30

B. XAI Algorithms

1) Local Interpretable Model-Agnostic Explanations (LIME): LIME is a method created to provide interpretable insights into the predictions made by complex machine learning models [10]. LIME generates a local surrogate model g that approximates the behavior of the black-box model f in the vicinity of a specific instance x from the input space. The surrogate model g is typically selected from a simpler and more interpretable class of models, often a linear regression model. The optimization problem can be defined as:

$$\min_{g \in G} \mathcal{L}(f, g, \pi_x) + \Omega(g) \tag{1}$$

In this equation, $\mathcal{L}$ represents a loss function that measures the dissimilarity between the predictions made by the original model f and the surrogate model g, calculated over instances sampled based on the proximity distribution π_x. The term $\Omega(g)$ introduces regularization to ensure the simplicity of the surrogate model. LIME achieves this by perturbing instances around x, sampling according to π_x, and fitting the surrogate model g to minimize the defined loss. The resulting g provides a locally faithful and interpretable approximation of f in the neighborhood of x, assisting in comprehending the decision-making process of the black-box model.

Discussions on the execution of LIME for various machine learning models and the analysis of their outcomes alongside the novel evaluation metrics are presented in section IV-A.

2) SHapley Additive exPlanations (SHAP): SHAP is an advanced approach that aims to elucidate the result of machine learning models through the assignment of a value referred to as the Shapley value to each feature, based on its contribution to the model's prediction. In the case of a given prediction $f(x)$, the SHAP values ϕ_i are determined for each feature x_i, representing the marginal contribution of said feature to the prediction. These values are computed by considering all possible combinations of features and calculating an average over all permutations [11]. The Shapley value is defined as:

$$\phi_i(f) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [f(S \cup \{i\}) - f(S)] \tag{2}$$

where N denotes the set of all features. Through the consideration of all potential interactions among features, SHAP values ensure an equitable distribution of credit among these features. The overall output of the model can be expressed as the summation of the contributions made by individual features:

$$f(x) = \phi_0 + \sum_{i=1}^{N} \phi_i \tag{3}$$

In this context, ϕ_0 represents the anticipated model output. SHAP values offer a comprehensive comprehension of the extent to which each feature contributes to the model's prediction.

3) Saliency Maps: Saliency maps enhance interpretability in deep learning CNNs by highlighting key areas in input images crucial for model predictions. These maps reveal which features the CNN prioritizes, aiding user understanding of the decision-making process. Although popular methods like Grad-CAM and Grad-CAM++ improve interpretability, they

have limitations, including incomplete feature capture and overly sensitive maps lacking contextual information.

This study employs Eigen-CAM as a model explainability tool for CNNs, pivotal in computer vision. Eigen-CAM utilizes class activation maps (CAM) to discern how models learn from visual data [14]. It computes primary components of learned features from convolutional layers, generating heatmaps superimposed on input images. Principal components vary across convolutional layers, reflecting evolving global representations. Heatmaps across layers or combined highlight visualization focus, enhancing model interpretability while addressing limitations of previous saliency map methods.

To calculate Eigen-CAM, we assume that X refers to the input image with size $(i \times j)$, $X \in \mathbb{R}_{i,j}$, and $W_{L=n}$ is the combined weight matrix of the first k layers of size (m, n) [15]. The class-activated output is the image X projected onto the last convolution layer $L = k$ and is given by:

$$O_{L=k} = W_{L=k} X \tag{4}$$

If we factorize $O_{L=k}$ using singular value decomposition to compute the principal components of $O_{L=k}$, we get:

$$O_{L=k} = U \Sigma V^T \tag{5}$$

Here U and V are orthogonal matrices and the columns of U are the left singular vectors, Σ is a diagonal matrix, the columns of V are the right singular vectors. The class activation map, $L_{Eigen-CAM}$ [15], is given by the projection of $O_{L=k}$ on the first eigenvector:

$$L_{Eigen-CAM} = O_{L=k} V_1 \tag{6}$$

Here V_1 is the first eigenvector in the V matrix.

4) Counterfactual Explanation: Counterfactual explanations have been introduced in the field of statistical learning following propositional logic analysis in analytic philosophy [16]. In the case of probabilistic learning, the problem is defined [17] as follows.

> For an input X, score p is returned because variable V has values $(v_1, v_2, ...)$ associated with them. If V instead held different values $(v'_1, v'_2, ...)$, then score p' would have returned provided all other variables had remained constant.

Following the above definition, the counterfactual explanation has been formulated for PCB component classification. Given a model F parameterized by w that classifies a component image x_i to class c_i with high probability, we tried to find a counterfactual x' that is as close as possible to the original point x_i and maps to a new target c'. x' is obtained by minimizing the following optimization problem.

$$\arg\min_{x'} \max_{\lambda} \left(\lambda (F_w(x') - c')^2 + D(x_i, x')\right) \tag{7}$$

Here D represents the distance between the original (x_i) and the counterfactual point (x'). So, maximization over λ is executed iteratively until x' is close enough to x_i. Following [17], we chose mean absolute deviation (MAD) over k features as the distance function D. Qualitative results on counterfactual representation for PCB component classification as well as microbump classification for IC advanced packaging are provided in section IV-D.

C. Evaluation Metrics

In the quest for improving the assessment of feature-based Explainable AI (XAI) models in the context of physical inspection, we present two innovative metrics: Model Performance Retention (MPR) and Context Relevance Score (CRS). The purpose of these metrics is to offer a thorough evaluation of the XAI algorithms by considering their influence on model performance and the significance of explanations within the particular inspection domain.

1) Model Performance Retention (MPR):

Definition: Model Performance Retention assesses the degree to which the predictive accuracy is maintained when the XAI algorithm selectively concentrates on a particular set of features. MPR measures the extent to which the model sustains its overall performance by solely considering the features emphasized in the XAI explanation. MPR can be defined as:

$$MPR = \frac{\text{Model Performance with XAI Features}}{\text{Model Performance with All Features}} \times 100\% \tag{8}$$

2) Context Relevance Score (CRS):

Definition: The Context Relevance Score evaluates the correspondence between explanations provided by XAI and the inherent context of the physical inspection task. CRS takes into account the significance of features mentioned in the explanation within the context specific to the domain, recognizing that not all influential features are equally pertinent in the particular inspection scenario. CRS can be defined as:

$$CRS = \frac{\sum_{i=1}^{N} \text{Contextual Weight}_i \times \text{Importance}_{\text{XAI},i}}{\sum_{i=1}^{N} \text{Contextual Weight}_i} \tag{9}$$

Where N is the number of features, $\text{Contextual Weight}_i$ represents the context relevance weight of feature i, and $\text{Importance}_{\text{XAI},i}$ is the importance assigned by the XAI algorithm.

CRS provides a nuanced understanding of the relevance of features highlighted by the XAI algorithm in the specific inspection context. It acknowledges that certain features may have greater importance within the domain, contributing to a more contextually informed evaluation of XAI performance.

D. Experimental Design

1) Implementation of LIME and SHAP Methods: We started our experiments by using Dataset 1 and the LIME package [18] to examine the importance of features in predicting each type of component. We utilized four different machine learning classifier models: random forest classifier, K-nearest neighbor (KNN) classifier, logistic regression classifier, and decision

tree classifier. After comparing their performances on the test set, we determined the best-performing model for further investigation and evaluation using Model Performance Retention (MPR) and Context Relevance Scores (CRS). Additionally, to validate the relevance of features, we also applied the SHAP method using the official PyPI package [19], [20]. All experiments with LIME and SHAP were carried out on an Intel Core i9-10980XE 3.00GHz CPU.

2) Counterfactual Explanation: Our training dataset for CF explanation consists of only 250 images which are pretty small to train deep neural network type architecture. So, we chose a simple model with two convolution layers and one fully connected layer to classify those four components from PCB and advanced packaging (see section III-A2) using 5-fold cross-validation. The model was run for 40 epochs with an ADAM optimizer on an NVIDIA RTX-2080 GPU with batch size 8. The learning rate was set to 0.0001 and cross-entropy loss was used as the objective function.

3) Visualization using Saliency Maps: In order to analyze the class activation maps using Eigen-CAM, we initially curated a dataset comprising of 40 PCB x-ray images and 4 advance packaging images(described in Dataset 1 in section III-A). We split the dataset with 32 images in training, 8 in validation, and 4 in the test set. The model was trained using an Intel Core i5-10300H 2.5GHz CPU with 10 epoches. The best performing model was saved and applied to predict on a set of test images, which consisted of 2 from PCBs and 2 from packaging samples. The next steps in our experiment involved loading an input image and the trained YOLOv8 model. Then, we specified the target layers for the computation of Eigen-CAM. By calculating Eigen-CAM for the specified layers, we were able to display the resulting CAM images superimposed on the original input image. This process allowed us to visually identify the regions of the input image that contribute most significantly to the model's predictions at different stages of the network.

IV. Results

A. Evaluation of LIME Results

Tables III, IV, and V consolidate the findings obtained from using the LIME method to clarify the performances of various classifier models. It is worth noting that the random forest classifier demonstrated higher accuracy and F1-score compared to other models when considering all features (Table I). As a result, subsequent LIME evaluation experiments were carried out using the random forest (RF) classifier as the base model. Table IV presents the top 5 essential features for predicting each of the four components of interest, as well as the overall top 5 LIME features for calculating model performance retention (MPR) scores. By utilizing these features, we conducted classification with the RF classifier and achieved accuracy and F1-scores that were close to those obtained with all features. The high MPR values (99.17% for accuracy and 99.84% for F1-score) confirm that LIME features effectively retained model performance.

TABLE III
Performance Summary of Classifier Models on Test Data considering All Feature Values

Classifier Models	Performance on Test Data (considering all features)	
	Accuracy (%)	F1-Score
Random Forest	90.68	0.9183
KNN	80.48	0.8263
Logistic Regression	34.29	0.2683
Decision Tree	88.91	0.9023

Furthermore, Table V reveals the top 3 contextually relevant features for classifying each component, along with their corresponding Context Relevance Scores (CRS). It is worth noting that despite the structural and appearance similarities among solder balls, PGA, vias, and microbumps, CRS values distinguish between them. This evaluation also emphasizes the contextual relevance of sample type (i.e., IC) for classifying microbumps in advanced packaging sample images, which aligns with our expectations.

In summary, these newly defined evaluation metrics offer valuable insights into identifying the distinct features that differentiate components with similar structures. This is particularly important in addressing edge cases in automated physical inspection.

B. SHAP Results

The experiments conducted using the SHAP explanation method in conjunction with the RF classifier model offer further validation for the findings obtained from the LIME evaluation experiments. An important discovery is the significant role played by shape features in the identification of the majority of both PCB and IC advanced packaging components. However, Tamura texture features, such as Tamura Coarseness, Tamura Directionality, and Tamura Contrast, emerge as crucial factors in differentiating microbumps from other components. These insights are visually depicted in Figure 1.

C. Saliency Maps Representation

Saliency Maps on PCB and Advance Packaging X-ray Samples. We generated different heatmaps from different layers of the model to understand in which portion of the image contributed more at that layer of the model. The heatmap is generated by overlaying it on top of the input image, where the red areas or spots indicate higher levels of magnitude. If the red areas are placed over places that are visually significant in differentiating the specific properties of the component, then this is a strong evidence of our model is learning properly [14]. In figure 2, we can see the heat map (fig 2 (b)) generated on consecutive layers of the trained model for the input image fig 2 (a). The heatmap contains yellow and red spots that indicate where the model is focusing to identify solder balls (class 0), vias (class 5), and some pads (class 3) in the input image. We compared the model's prediction result, fig 2 (c) from the test set, with the heatmap to demonstrate

TABLE IV
SUMMARY OF MODEL PERFORMANCE RETENTION (MPR) FOR LIME FEATURES

Component Type	Top 5 Features Obtained from LIME Method using RF Classifier for Each Component Type	Overall Top 5 LIME Features to Calculate Model Performance Retention (MPR) Scores	Model Performance (using LIME features only)		Model Performance Retention (MPR)	
			Accuracy (%)	F1-Score	MPR for Accuracy (%)	MPR for F1-Score (%)
Solder Balls	Circularity, Eccentricity, Aspect Ratio, GLCM Contrast, Perimeter	1. Circularity 2. Aspect Ratio 3. Perimeter 4. Solidity 5. Tamura Contrast	**89.93**	**0.9168**	**99.17**	**99.84**
PGA	Solidity, Circularity, Perimeter, Area, Entropy					
Vias	Circularity, Solidity, Aspect Ratio, Convexity, Tamura Contrast					
Microbumps	Tamura Directionality, Tamura Contrast, Tamura Coarseness, Aspect Ratio, Circularity					

TABLE V
CONTEXT RELEVANCE SCORES SUMMARY FOR EACH COMPONENT TYPE USING LIME METHOD WITH RF CLASSIFIER

Component Type	Top 3 Contextually Relevant Features	Corresponding Context Relevance Scores (CRS)
Solder Balls	1. Circularity	0.005090
	2. Aspect Ratio	0.001309
	3. Eccentricity	0.000927
PGA	1. Solidity	0.002458
	2. Circularity	0.001804
	3. Perimeter	0.000938
Vias	1. Circularity	0.018075
	2. Convexity	0.003437
	3. Solidity	0.003144
Microbumps	1. Sample Type (IC)	0.004541
	2. Convexity	0.002209
	3. Circularity	0.001295

that the model was able to predict some of these components. Furthermore, in Figure 2 (e), we observed red spots on the right side of the heat map , which indicate that the model assigned greater importance in that region to locate and predict PGAs (class 1).

We have shown our results in advance packaging samples with microbump class in figure 3. In figure 3 (b), the heatmap is showing dense red spots in the left side which may suggest bias in the training data towards microbump class in this image. Also, we found one interesting fact from figure 3(e) that there are some red spots in the lower side of the heatmap which indicates model's errorenus focus on the number value in the input image fig 3(d) and we found the model is predicting a microbump in the leftmost corner of the prediction image figure 3(f) which actually validates the reason why the model is predicting the number value 8 in the leftmost corner as a microbump.

D. Counterfactual Representation

PCB Components. We experimented with images of solder balls (class 0), PGA (class 1), and Vias (class 2) extracted from a few X-ray PCB images. As we had few data for each component, a small CNN consisting of 2 convolution layers and 1 fully connected layer was trained to classify the component images.

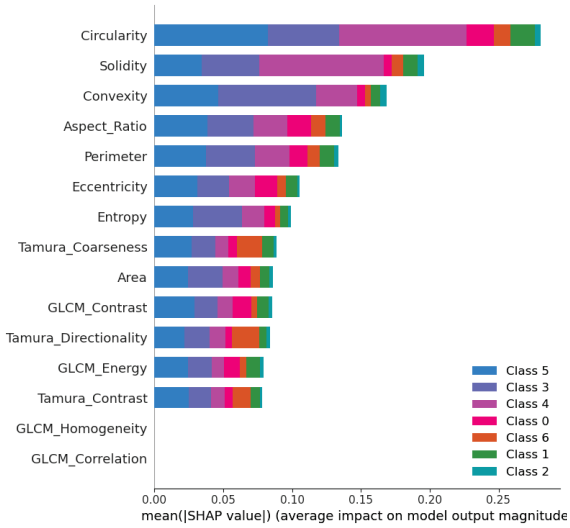

Fig. 1. Visualization of Feature-based Explanations employing SHAP for the Random Forest (RF) Classifier. The figure illustrates the average impact of individual features on the model output, quantified through mean SHAP values.

Before proceeding with counterfactual representation, heatmap was generated to visualize the representative regions for classification outcome as shown in figure 4.It can be seen from Figure 4 that Via is identified by the circular boundary while for PGA some values inside the circular boundary have also been activated. So, to generate a counterfactual representation of class via i.e., to make the model predict a Via sample as PGA, some pixels inside the circular boundary need to be activated too.

Figure 5 shows the counterfactual (CF) representation of a Via image that makes the model predict it as PGA. The CF representation was generated using the OmniXAI [21] library that used the MDA algorithm [17] to find the closest CF representation of the original input. It's visible that the MDA algorithm has been able to activate some of the values inside the circular boundary to fool the model into predicting it as

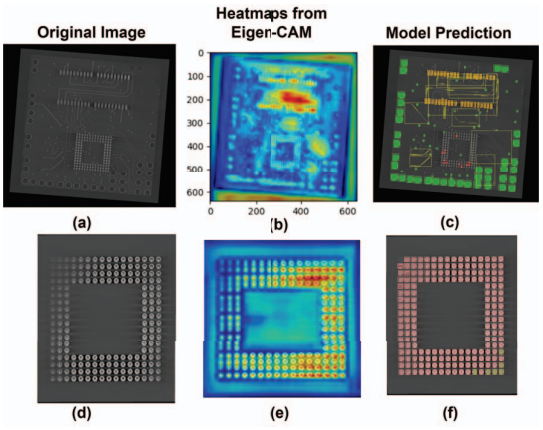

Fig. 2. Visual representation of heatmaps generated with Eigen-CAM on original input images from PCB samples and comparison of significant areas with model prediction on same images. Here, first column: figures (a) & (d) are input images. Second column: figures (b) & (e) are respective heatmaps. Third column: figures (c) & (f) are respective predictions from trained model.

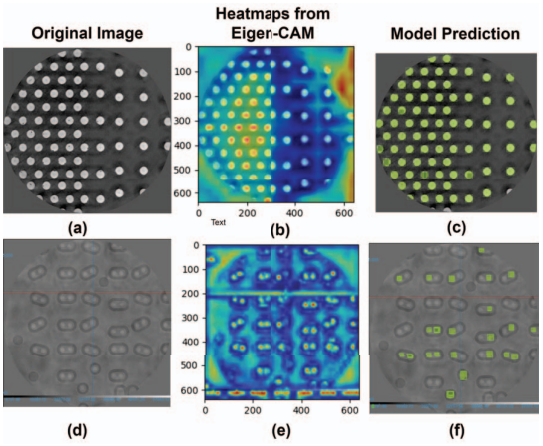

Fig. 3. Visual representation of heatmaps generated with Eigen-CAM on original input images from advance packaging and comparison of significant areas with model prediction on same images. Here, first column: figures (a) & (d) are input images. Second column: figures (b) & (e) are respective heatmaps. Third column: figures (c) & (f) are respective predictions from trained model.

PGA (class 1).

Advanced Packaging. We experimented with images of microbumps extracted from X-ray images of advanced packaging too. We conducted similar type of experiments by mixing up microbump images with PCB components and train a classifier to learn robust representation. The heatmap of classifier prediction for microbump is shown in figure 6.

While generating counterfactual representation using [21], the MDA algorithm has perturbed pixels on the bump to simulate the feature behavior like PGA class where few pixels inside the circular boundary are activated and rest are not as shown in figure 7.

Observation. The CF explanation in both PCB and Advanced packaging components provides insight into how

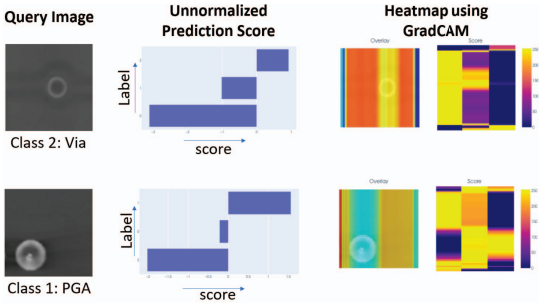

Fig. 4. Visual explanation using GradCAM. The middle column represents unnormalized prediction value from the model. The model has provided higher prediction values for respective classes as expected. The last column represents the regions highlighted by GradCAM responsible for a particular prediction. GradCAM was generated from the 2nd convolution layer of the model. So, different color in the score matrix just shows different values of the kernel on that layer feature.

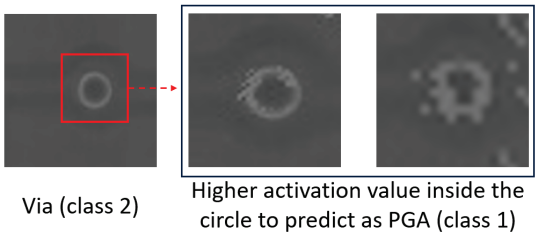

Fig. 5. Two counterfactual (CF) representations of class Via to make model predict it as PGA. OmniXai library [21] has been used to generate the CF representation.

model predictions can be altered through external noise or malicious alterations by any external entity. In particular, while developing a foundational model with the ability to detect counterfeit from x-ray images of any packaging level, it needs to be robust against such kind of subtle modifications. We believe with more data samples, it's possible to provide better CF representation as well as develop a robust model to resist those alterations. We are going to explore this in our future work.

V. Discussions

A. Key takeaways

This study has yielded valuable insights into advanced IC packaging and PCB inspection. Key takeaways from the experimental results include:

- The incorporation of model performance retention (MPR) and context relevance scores (CRS), in evaluating feature-based explanations like LIME, sheds light on crucial features distinguishing microbumps from visually similar PCB components such as vias, solder balls, and PGA.
- Additionally, the use of counterfactual explanations and saliency maps visually highlights key differences, marking a significant advancement in explaining edge cases in advanced IC packaging inspection.

B. Future Research Directions

Future research directions encompass several key areas:

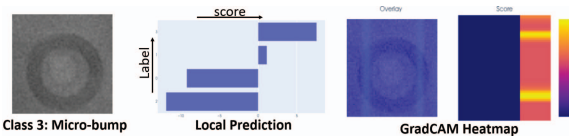

Fig. 6. GradCAM heatmap for classifier prediction for micro-bump image

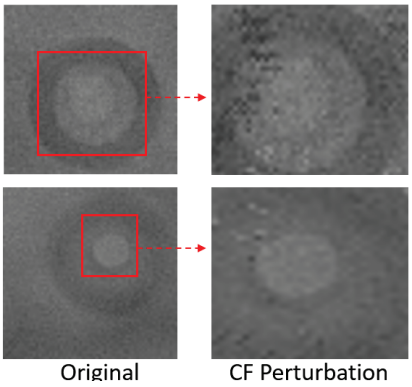

Fig. 7. Two counterfactual (CF) representations of class Micro-bump to make model predict it as PGA. The pixels on micro-bump are perturbed to make the feature similar to PGA.

- A significant avenue for future research involves expanding the dataset to include additional classes, like TSVs and C4 bumps. This expansion aims to explore the generalization capability of the introduced evaluation metrics.
- Deep-learning-based explanations necessitate substantial data, and gathering data through both X-ray and SAM methods is time-consuming. However, it's worthwhile to investigate the impact of imaging modalities and parameters by incorporating data from different classes of IC advanced packaging components. This exploration involves employing various explanation methods alongside their evaluation.

VI. Conclusion

In conclusion, the study underscores the complexity of inspecting advanced packaging systems for integrated circuits, emphasizing precision and transparency. Despite the valuable contributions of Explainable AI (XAI), the absence of standardized measurements continues to present a formidable obstacle. This research bridges the gap by establishing industry-standard metrics for XAI techniques, enhancing transparency in semiconductor scenarios. The insights gained from analyzing advanced IC packaging and PCB inspection, coupled with the proposed avenues for future research, hold great promise in advancing our understanding and evaluation of intricate semiconductor systems.

References

[1] M. S. M. Khan, C. Xi, N. Varshney, A. A. Khan, A. Serna, H. Dalir, V. Sorger, and N. Asadizanjani, "$CM_{\text{x-ray}}$: An X-Ray Compatibility Metric for Advanced Packages to Facilitate Design-for-Inspection," in *2023 IEEE Physical Assurance and Inspection of Electronics (PAINE)*. IEEE, 2023, pp. 1–7.

[2] R. Noor, H. R. Kottur, P. J. Craig, L. K. Biswas, M. S. M. Khan, N. Varshney, H. Dalir, E. Akçalı, B. Motlagh, C. Woychik *et al.*, "Us microelectronics packaging ecosystem: Challenges and opportunities," *arXiv preprint arXiv:2310.11651*, 2023.

[3] W. Samek, T. Wiegand, and K.-R. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," 2017.

[4] C. D. Hartfield, T. M. Moore, and S. Brand, "Acoustic microscopy of semiconductor packages," *Microelectronics Fialure Analysis Desk Reference*, p. 67, 2019.

[5] G.-M. Zhang, D. M. Harvey, and D. R. Braden, "Microelectronic package characterisation using scanning acoustic microscopy," *NDT & E International*, vol. 40, no. 8, pp. 609–617, 2007.

[6] T. Moore and C. Hartfield, "X-ray and sam—challenges for ic package inspection," in *ISTFA 2022*. ASM International, 2022, pp. q1–q52.

[7] C. Hartfield, C. Schmidt, A. Gu, and S. T. Kelly, "From pcb to beol: 3d x-ray microscopy for advanced semiconductor packaging," in *2018 IEEE International Symposium on the Physical and Failure Analysis of Integrated Circuits (IPFA)*. IEEE, 2018, pp. 1–7.

[8] W. Liang, G. A. Tadesse, D. Ho, L. Fei-Fei, M. Zaharia, C. Zhang, and J. Zou, "Advances, challenges and opportunities in creating data for trustworthy AI," *Nat. Mach. Intell.*, vol. 4, no. 8, pp. 669–677, 2022.

[9] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J. M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N. Díaz-Rodríguez, and F. Herrera, "Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence," *Information Fusion*, vol. 99, p. 101805, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253523001148

[10] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?": Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 1135–1144. [Online]. Available: https://doi.org/10.1145/2939672.2939778

[11] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 4765–4774. [Online]. Available: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

[12] W. Zhao, S. R. Gurudu, S. Taheri, S. Ghosh, M. A. Mallaiyan Sathiaseelan, and N. Asadizanjani, "Pcb component detection using computer vision for hardware assurance," *Big Data and Cognitive Computing*, vol. 6, no. 2, 2022. [Online]. Available: https://www.mdpi.com/2504-2289/6/2/39

[13] S. Ghosh, M. T. Mostafiz, S. R. Gurudu, S. Taheri, and N. Asadizanjani, "Pcb component detection for hardware assurance: A feature selection-based approach," in *2022 IEEE International Symposium on Hardware Oriented Security and Trust (HOST)*, 2022, pp. 109–112.

[14] L. So, "Understanding your YOLOv8 model with Eigen-CAM," https://www.datature.io/blog/understanding-your-yolov8-model-with-eigen-cam, accessed: 2024-2-21.

[15] M. B. Muhammad and M. Yeasin, "Eigen-cam: Class activation map using principal components," *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:220936532

[16] A. J. Ayer, *The Problem of Knowledge*. New York,: Harmondsworth, 1956.

[17] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: automated decisions and the gdpr," *Harvard Journal of Law and Technology*, vol. 31, no. 2, pp. 841–887, 2018.

[18] "Local interpretable Model-Agnostic explanations (lime) — lime 0.1 documentation," https://lime-ml.readthedocs.io/en/latest/, accessed: 2024-2-22.

[19] "Welcome to the SHAP documentation — SHAP latest documentation," https://shap.readthedocs.io/en/latest/index.html, accessed: 2024-2-22.

[20] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable ai for trees," *Nature Machine Intelligence*, vol. 2, no. 1, pp. 2522–5839, 2020.

[21] W. Yang, H. Le, T. Laud, S. Savarese, and S. C. H. Hoi, "Omnixai: A library for explainable ai," *ArXiv*, vol. abs/2206.01612, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:249375643