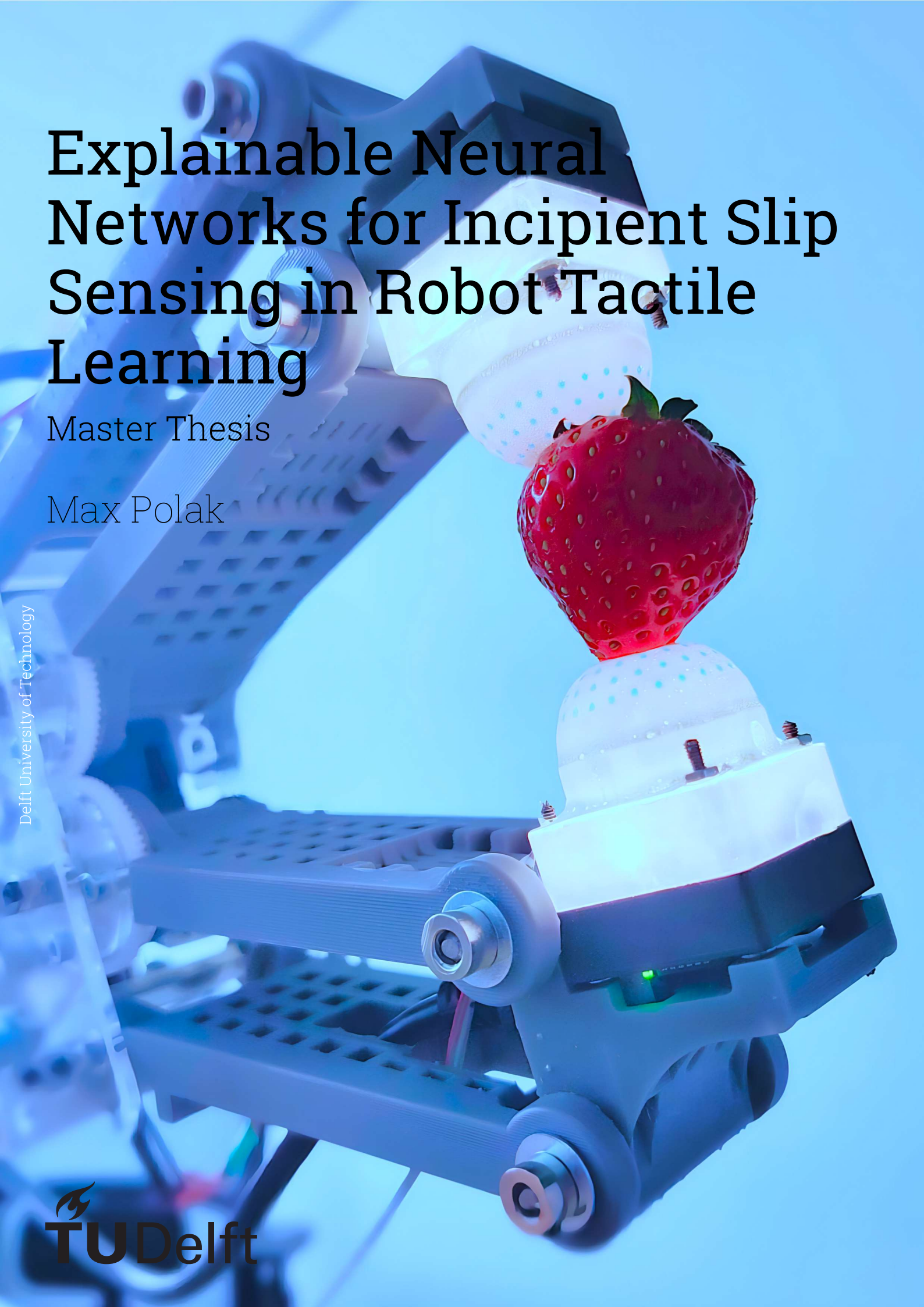
Explainable Neural Networks for Incipient Slip Sensing in Robot Tactile Learning

Master Thesis

Max Polak



Delft University of Technology

TU Delft

Explainable Neural Networks for Incipient Slip Sensing in Robot Tactile Learning

by

Max Polak

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Thursday 31 of August, 2023 at 9:30 AM.

Student number:	4570677	
Project duration:	September, 2022 – June, 2023	
Thesis committee:	Dr. M. Wiertlewski,	TU Delft, supervisor, chair
	Ir. G. Vitrani,	TU Delft, supervisor, PhD
	Dr. J. Kober,	TU Delft, external committee member

An electronic version of this thesis is available at `http://repository.tudelft.nl/`.

TU Delft

Explainable Neural Networks for Incipient Slip Sensing in Robot Tactile Learning

M.E.A. Polak

Abstract—Incipient slip detection plays an important role in human and robotic grasping. With the growing use of deep learning in vision-based tactile sensing, the black-box nature of these deep neural networks (DNNs) makes it difficult to analyze, debug, and validate their behavior and learned patterns. To fill this gap, eXplainable AI (XAI) methods have been introduced to shed light into the DNN's reasoning regarding incipient slip detection. These methods generate saliency maps, highlighting the relevant regions in the input tactile image that resulted in the predicted degree of incipient slip. Temporal difference images have been used to enhance the visualization of incipient slip and make saliency maps easier for human viewers to understand. Additionally, this research evaluates several XAI methods based on criteria such as high-resolution, smoothness, and faithfulness. The experiment examined 42 samples from the ChromaTouch tactile dataset, focusing on contact interactions with a flat object. The results showed that Poly-CAM satisfies all three criteria by accurately highlighting markers while emphasizing their relative importance in the DNN's decision-making process. Overall, through visual analysis of saliency maps, our findings confirm that DNNs have successfully learned to localize crucial deformation features for detecting incipient slip.

Index Terms—Explainable AI, Deep Learning, Vision-Based Tactile Sensing, Incipient Slip Detection, Frictional Safety Margin

1 Introduction

For many years, robotic manipulation has been extensively used in industries to perform repetitive tasks in controlled environments. However, maintaining a stable grasp during robotic manipulation is challenging, particularly when object contact parameters vary, as in the case of grasping fruits and vegetables, or in the presence of unexpected external perturbations [1, 2]. Loose grips with insufficient grasping forces can cause objects to slip, while excessive grasping forces can damage objects [3].

When humans grasp an object, mechanoreceptors located in their fingertips play a vital role in perceiving tactile information regarding incipient slip, which refers to the early stages of slippage that occur before an object fully slips out of the hand [4, 5]. Through skin deformation and vibrations, humans are capable of sensing incipient slip with high sensitivity [6]. The brain receives this sensory information and provides feedback to the muscle control system, allowing humans to maintain enough grip to prevent slippage, while being gentle enough to avoid damage to the object in various circumstances [7]. Hence, the integration of tactile sensors and feedback would be a valuable addition to robots, since it can improve the stability of grasping operations and enable them to exhibit human-like dexterity [8].

Recently, vision-based tactile sensors have emerged as a promising approach among various tactile sensing technologies. They leverage the advancements of computer vision to equip robots with a sense of touch [9]. These sensors use a camera to capture high-resolution tactile images, providing a visual representation of the deformation of contact between the soft fingertip and grasped object [10].

Model-based techniques have traditionally been used to detect incipient slip by inferring from the marker displacement when the sensor deforms under contact [11–15]. However, there has been a growing

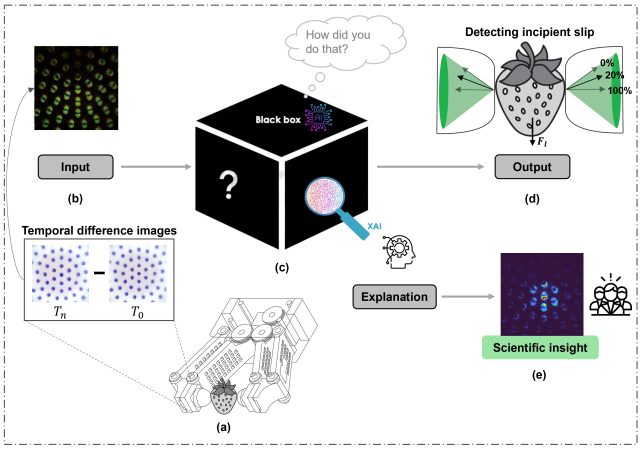


Fig. (1) Explainable tactile perception framework. (a) Grasping a strawberry using tactile sensors to capture raw images of contact deformation (b) Temporal difference between tactile images taken before (T_0) and after contact (T_n). (c) Black-box neural networks can't explain their decisions, unlike humans. (d) Detecting incipient slippage based on its degree where full stick (100%), incipient slip (100-0%) and full slip (0%). (d) Saliency maps are a form of visual explanation heatmaps that indicate which input features were used by the DNN to generate its output decisions.

trend towards employing Deep Neural Networks (DNN) for tactile sensing, driven by their remarkable ability to autonomously learn hierarchical representations from high-dimensional tactile data [16–20]. Consequently, robots can be trained more effectively to process tactile data and predict incipient slip.

Unlike model-based approaches, DNNs are criticized for their 'black box' nature owing to their intrinsic complexity (e.g. multiple hidden layers, non-linear activations, etc.). As a consequence, the predictions made by DNNs provide minimal or no understanding of the underlying decision-making process [21]. This limitation becomes particularly concerning when DNNs produce inaccurate predictions, as they often struggle to generalize to new data in real-world ap-

plications due to overfitting or susceptibility to adversarial attacks [22]. The opacity of DNNs poses challenges for developers in identifying, diagnosing, and addressing the causes behind these failures in such scenarios. In response, various explainability methods have been proposed to enhance transparency and foster trust in AI systems [23, 24]. These methods seek to provide insights into how these neural networks arrive at a particular decision.

This study introduces a novel eXplainable AI (XAI) framework specially designed to solve the black-box nature of DNNs in tactile sensing. The proposed framework (Figure 1) expands the application of XAI beyond its typical domains, such as object recognition tasks. The framework employs saliency maps as a key component to validate the learned patterns and reveal the reasoning of DNNs in estimating incipient slip from tactile images. The incorporation of temporal difference images enhances the interpretability of these saliency maps, making incipient slip more noticeable for human viewers. Furthermore, an in-depth explainability experiment is conducted to determine the most suitable XAI method for tactile sensing. The evaluation process considers criteria such as high-resolution, smoothness, and faithfulness of the generated saliency maps.

This research aims to gain insight into the behavior of DNNs and to verify their ability to recognize deformation features in tactile images. By doing so, a more profound comparison can be achieved between different trained deep learning models.

2 Related Work

This section provides an overview of the existing literature regarding incipient slip detection in tactile sensing, and highlights the necessity of eXplainable AI in understanding neural network decisions.

2.1 Contact Mechanics of Slip

When a soft finger contacts with an object and tangential force is applied, full slip does not occur immediately but gradually evolves from incipient slip, resulting in a stick and slip region within the contact area, as shown in Figure 2. Increasing the tangential force causes the slip region to spread from the boundary towards the center of the stick region, and as soon as it covers the entire contact area, full slip or gross sliding occurs [12, 13]. The deformation of the sensor surface can be used to analyze the incipient slip state [25]. Vision-based tactile sensors, such as GelSight [11, 14] and ChromaTouch [15, 26, 27], have been developed due to their ability to capture deformations in high-resolution tactile images, which makes them well suited for incipient slip detection.

2.2 Incipient Slippage Detection

Detecting incipient slip is crucial for controlling the grasp force and thus preventing grasp failures. However, solely classifying whether full slip has occurred or not is insufficient because it could only provide an early warning [28]. Consequently, feedback is only provided once full slip has already begun, leading to delayed grasp adjustment and the potential of dropping the object. Furthermore, it does not specify the amount of additional force required to prevent slippage. To address this issue, it is crucial to obtain fine and rich measurements of the slip state by quantifying the degree of initial slip (Γ). This function expresses the proximity of the current state to full slip, as shown in Figure 2, and could serve as a parameter for controlling grasp force [29, 30].

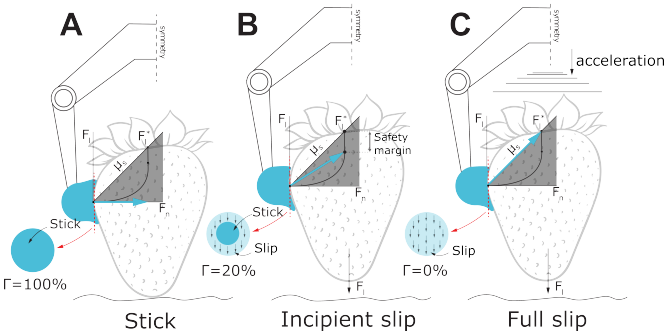


Fig. (2) This figure illustrates the contact mechanics and resultant forces present when grasping a fragile object for three scenarios, indicated by different values of Γ. Here, Γ correspond to either the stick ratio (related to area stick/slip region) or frictional safety margin (related to contact forces and friction cone).

2.2.1 Human grasping

The degree of incipient slip can be characterized by both the frictional safety margin and stick-ratio, as shown in Equation 1. The frictional safety margin describes the distance between the maximum available friction force of the contact F_l^* and the current external load forces acting on the object F_l [31]. From Coulomb's law, it follows that $F_l^* = \mu_s F_N$, where μ_s is static friction coefficient and F_N is the normal force. Similarly, the stick ratio is a measure of the proportion of the stick region A_{stick} in the contact area $A_{contact}$.

$$\Gamma = \frac{F_l^* - F_l}{F_l^*} = \frac{A_{stick}}{A_{contact}} \tag{1}$$

A high value of $\Gamma \rightarrow 100\%$ indicates full stick, where excessive grasping forces can potentially damage the object. Conversely, a low value of $\Gamma \rightarrow 0\%$ means that the contact is at the onset of slip, and even slight perturbations in load forces or contact surface could result in full slippage. While both measures are valuable for understanding and predicting systems subject to frictional forces, the frictional safety margin is particularly relevant to human grasping. To ensure secure grasp, humans generally apply additional grip force during grasping to maintain a safety margin of 20-40% of the minimum required grip force [32]. The specific margin range varies depending on the task and contact conditions, but serves as a balance between grasp stability and object safety [33].

2.2.2 Model-based vs learning-based

The available methods to detect incipient slippage from tactile images can be categorized into model-based and learning-based approaches.

Model-based techniques have predominantly been utilized to detect and track movement of markers in the soft fingertip, enabling the extraction of a displacement field. This displacement field serves as a representation of contact and can be utilized to estimate the incipient slip state, as illustrated in Figure 3. Researchers have explored various contact models to leverage the marker displacement field for this purpose. Obinata et al. [29, 34] and Sui et al. [25, 28] have investigated the estimation of the stick ratio. Furthermore, Yuan et al. [35] and Dong et al. [13, 14] proposed a novel method for incipient slip detection using the GelSight/GelSlim sensor, which detects slip based on the rate of change in the displacement field. Despite the interpretability of these methods by involving humans in the loop, their reliance on accurately modeling the contact mechanics imposes inherent performance limitations [36, 37].

Hence, DNNs are used for the purpose of end-to-end learning in the detection of incipient slip from tactile images (Figure 3). Given that slip is a spatio-temporal event, several studies have developed deep learning models that leverage temporal features in a sequence of tactile images [16–19]. Nevertheless, processing spatio-temporal signals is computationally expensive, posing a challenge for robots to react quickly and avoid full slip. Calandra et al. [20] applied a Convolutional Neural Network (CNN) to predict grasp outcomes based on difference images obtained from the GelSight sensor. This method efficiently captures temporal features by taking the time difference between pre- and post-contact images.

Although the aforementioned approaches achieve high accuracy in detecting incipient slip, they only classify whether incipient slip has occurred or not. On the contrary, a notable exception is the work by Dirk et al. [38]. They used the ChromaTouch tactile sensor along with a lightweight CNN architecture to predict the frictional safety margin as a continuous value ranging from 0% to 100%.

2.3 Explainable AI

Despite the remarkable success of CNNs in solving complex vision problems in various domains, their lack of interpretability has been a matter of concern. Although there are techniques to visualize activation maps and filters in CNNs, they provide limited insight into the overall decision-making process. Consequently, XAI has become increasingly important in unveiling the intricate mechanisms of DNNs by tracing the path from input to output. Several XAI techniques have been devised for computer vision, including feature visualization and feature attribution methods [39]. Feature attribution methods serve as 'post-hoc' explanations for trained black box models.

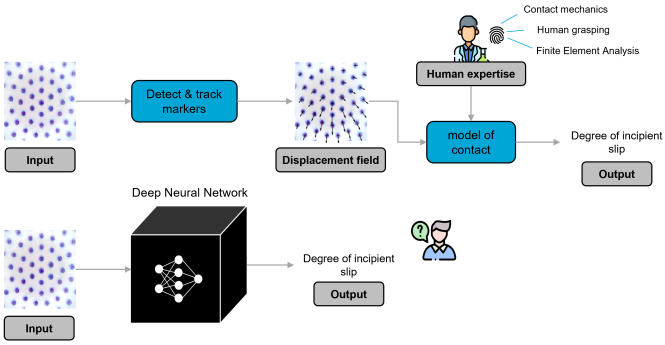


Fig. (3) Model-based vs learning-based.

They generate saliency maps that highlight the important regions of the input image that the network is "looking" at to make a certain prediction. The saliency maps serve as a method for weakly supervised object localization, aiding in the verification of whether predictions are based on relevant features [40]. Particularly in the field of tactile sensing, these methods could offer valuable insights into the prioritization of markers by the neural network. They aid in understanding how the DNN make decisions across different contact states (stick, incipient slip, full slip) and diverse contact conditions (e.g. forces, contact area, and object surface properties).

2.3.1 Feature attribution

Within the scope of post-hoc attribution methods, three families of methods can be distinguished, (1) Gradient-based, (2) Perturbation-based, and (3) CAM-based [41]. Gradient-based methods produce pixel-level saliency maps by backpropagating the gradient of the output through the network to the input image. Perturbation-based methods, on the other hand, introduce perturbations to the original input image and analyze the consequential alterations in the model's predictions to ascertain saliency regions. CAM-based methods employ a linear weighted combination of activation maps from the final convolutional layers and the predicted output. This is then upsampled to the size of the input image for overlaying purposes. Recent approaches in explainability often leverage the advantages of different methods while addressing their respective limitations.

2.3.2 XAI in tactile sensing

Certain studies have employed feature attribution techniques to validate the learned patterns of DNNs in touch processing tasks. Cao et al. [42] verified the effectiveness of their Spatio-Temporal Attention Model for tactile texture recognition through Grad-CAM visualizations. Similarly, Han et al. [19] used Grad-CAM visualizations to confirm the efficacy of self-attention mechanisms in Vision Transformers. This analysis demonstrated their capability to capture the global context of tactile images for slip detection and grasp outcome prediction.

3 METHODOLOGY

This section covers the fundamental requirements for the successful integration of XAI in the field of tactile sensing. It explores the necessity of enhancing the intuitiveness of the network's input for human viewers, the selection of appropriate XAI methods to generate saliency maps that are clear and detailed, and introduces a modified metric to assess the faithfulness of these saliency maps.

3.1 Temporal Difference Imaging

Temporal difference imaging captures changes in contact by taking the time difference between pre- (T_0) and post-contact (T_n) tactile images. It allows for the enhancement of visible changes that indicate the start of slipping (incipient slip) and improves the ability to differentiate between different contact states.

3.1.1 Spatio-temporal features

This study expands the previous research conducted on the ChromaTouch tactile sensor [15, 26, 27, 38]. Here, a lightweight CNN was employed to perform a single-output regression, aiming to predict the frictional safety margin from raw tactile images. Unlike other deep learning models that preserve the slip state in memory, this method doesn't utilize the temporal features of slip. This is because such models tend to be computationally expensive and aren't suitable for real-time applications. To address this issue, our study uses temporal difference images. This approach combines the benefits of both strategies by including temporal features while preserving the 2D spatial image. As a result, we establish a system that only captures changes in deformation during interactions with a soft fingertip and excludes irrelevant background information.

3.1.2 Visualizing incipient slip

In order to facilitate effective use of XAI in tactile sensing and provide meaningful assistance to both developers and end users, it is crucial for humans to understand how incipient slip can be recognized from tactile images. While XAI is intuitive for humans in image classification, where the network should highlight the object in the image, as humans can do as well. This is not the case in tactile sensing, where the deformations are often too small and subtle for humans to notice. Hence, temporal difference images are introduced to enhance the visibility of incipient slip and aid human understanding. By subtracting the initial tactile image (T_0) from the current tactile image (T_n), temporal difference images render changes in contact over time. This process accentuates the overlap of markers (depicted as blobs) and unveils their displacement caused by normal and/or frictional forces. As a result, both the initiation of contact and incipient slip become more visually distinguishable, as illustrated in Figure 4.

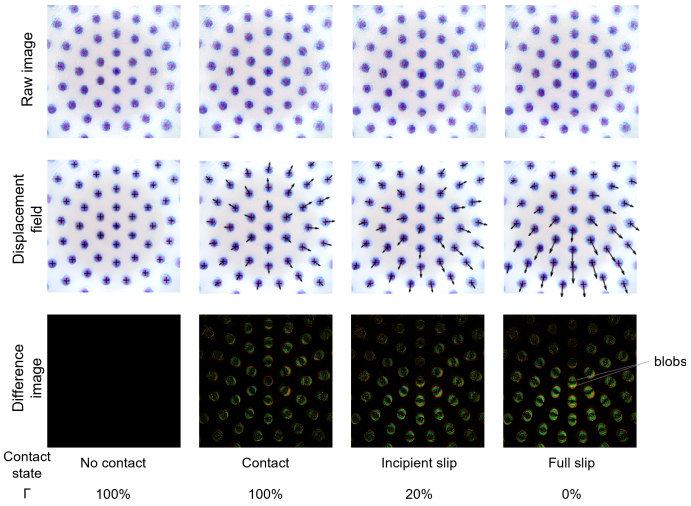


Fig. (4) The figure illustrates raw tactile images, marker displacement fields obtained through the method by Lin et al. [27], and temporal difference images among the four contact states.

The contact states can be classified into four distinct categories, each characterized by specific visual characteristics. These categories include: (1) "no contact," represented by a black image where the markers of both the initial and current tactile images perfectly align; (2) "contact," where all markers appear due to the indentation of the dome; (3) "incipient slip," identified by the disappearance of markers located above the center, while the bottom markers light up due to stretching in the shear direction; and (4) "full slip," which shares similarities with incipient slip, but with an intensified signal intensity.

3.1.3 Contact mechanics analysis

The visual representations of the four contact states are further analyzed to assess their conformity with the contact mechanics relevant to dome-shaped contacts. According to the Hertz contact model, the central region of the contact experiences the highest pressure distribution when subjected to a normal force. As a result, it also exhibits a higher signal-to-noise ratio (SNR).

During the incipient slip stage, the displacement at the bottom is more prominently visible due to the combined effects of normal indentation and slippage shear, while the opposite is observed for the top part. This phenomenon can be further understood by analyzing the marker displacement field (Figure 4). Upon initial contact, a divergent vector field emerges due to the peripheral stretching of the dome. With increasing slippage (a lower Γ), the displacement vectors in the bottom region progressively increase, while the displacement field in the top region tends towards zero. This occurs in the upper region due to the counteractive effect of slippage shear against the "divergence" resulting from the normal indentation. Conversely, in the bottom region, the shear arising from slippage accumulates with the "divergence" induced by normal indentation. Consequently, larger displacement vectors of the markers are observed in the bottom region.

3.2 Visual Explanations of Tactile Images

Previous research in vision-based tactile sensing has shown that comparing various DNNs based solely on performance metrics is inadequate [19, 42]. Instead, explainability methods should be employed to ensure that the model has learned the correct pattern and can take advantage of the new method effectively. However, no research has been conducted on comparing existing XAI methods in this field, as well as evaluating their saliency-based explanations on their faithfulness to truly reflect the network's decision. In this research, several XAI techniques are evaluated on the basis of their human readability and faithfulness.

3.2.1 Aesthetic Quality of XAI

To effectively apply XAI in vision-based tactile sensing, it is crucial that the generated saliency maps (1) are detailed by highlighting the individual markers that are used in the decision, and (2) are clear and easily understandable for humans. This ensures that researchers can draw meaningful conclusions about the DNN's ability to comprehend the deformation captured in tactile images.

Considering these criteria, three XAI methods will be compared in this study. Examples of saliency maps generated by these XAI methods for image classification can be found in Appendix A. Analysis of these samples has led to the formulation of certain expectations. The widely used Grad-CAM [43] and its variants may be limited by their low resolution and inability to localize multiple instances of an object. Consequently, these methods may not be suitable to localize specific markers (condition 1) since it requires highlighting precise details of the object features. Conversely, gradient-based methods like Integrated Gradients [44] can generate pixel-level saliency maps but are often visually noisy, likely failing to meet the human-readability criterion (condition 2). The state-of-the-art Poly-CAM technique [41] emerges as a promising choice, as it have shown to generate high resolution and noise-free saliency maps.

3.2.2 Faithfulness of XAI

In addition to the aesthetic quality of the explanations, it is essential to assess the correctness of XAI methods in accurately representing the decision-making process of DNNs. A faithful explanation should highlight the features the model actually used to make a prediction. If an explanation is not faithful, it could mislead us into drawing incorrect conclusions about the model's behavior.

Two commonly used metrics for assessing the faithfulness of XAI explanations are deletion and insertion [45]. To simplify, deletion is like removing pieces from a jigsaw puzzle – we gradually take out the most important pieces (or 'features') and see how it affects the overall picture (or the model's output). Insertion, on the other hand, is like starting with a blank canvas and gradually adding the most important pieces to see how the picture improves.

For our study, we'll be focusing on the deletion metric, as it can be easily adapted to suit the unique requirements of regression tasks. However, it is important to acknowledge that these metrics were primarily developed for image classification tasks, which differ fundamentally from regression tasks. A detailed summary of the deletion metric for image classification can be found in Appendix B. In the context of regression, the focus shifts from predicting probability scores associated with classes to estimating a continuous output.

$$\mathrm{D}_{\mathrm{AUC}} = \frac{1}{n_{\mathrm{mask}}} \sum_{i=0}^{n_{\mathrm{masks}}} \overbrace{\left(1 - \frac{|y_{\mathrm{base}} - y_{\mathrm{mask}}|}{100}\right)^2}^{\text{deletion score (D)}} \qquad (2)$$
$$\overline{\mathrm{D}}_{\mathrm{AUC}} \downarrow = \frac{1}{n_{\mathrm{data}}} \sum_{j=0}^{n_{\mathrm{data}}} \mathrm{D}_{\mathrm{AUC}}$$

To make the deletion metric more applicable to regression tasks, a modified version is proposed. In simple terms, this modified metric is designed to account for situations where removing important features can either decrease or increase the model's output. Both these changes indicate a deviation from the actual prediction. The metric is calculated using a mathematical formula (see Equation 2), which involves measuring the 'area under the deletion score curve' (D_{AUC}). This entails plotting a graph to observe how the prediction changes across different levels of feature deletion. The resulting area under this curve provides an overall score, with smaller values indicating a more 'faithful' saliency map. To make this clearer, a visual representation of the deletion process is shown in Figure 5. The metric takes two variables into account, namely the number of masked images generated (n_{masks}) and the number of baseline images (n_{data}). Further details on these parameters will be provided in the next section.

4 Experimental Setup

The experimental design aims to ensure the reproducibility of the study. It includes detailed information about the training of DNNs, the evaluation of their performance, and both qualitative and quantitative assessments of the XAI methods.

Hardware components The ChromaTouch sensor used in this study is based on the design by Scharff et al. [15]. The sensor setup remains unchanged, and an overview is shown in Figure 6.

Data Acquisition This study used an existing dataset [38] consisting of ChromaTouch tactile images. This dataset exclusively examined contact interactions between a flat object under distinct frictional conditions: high friction (dry surface) and low friction (surface sprayed with water). Incipient slip scenarios were simulated by applying a load force on the object

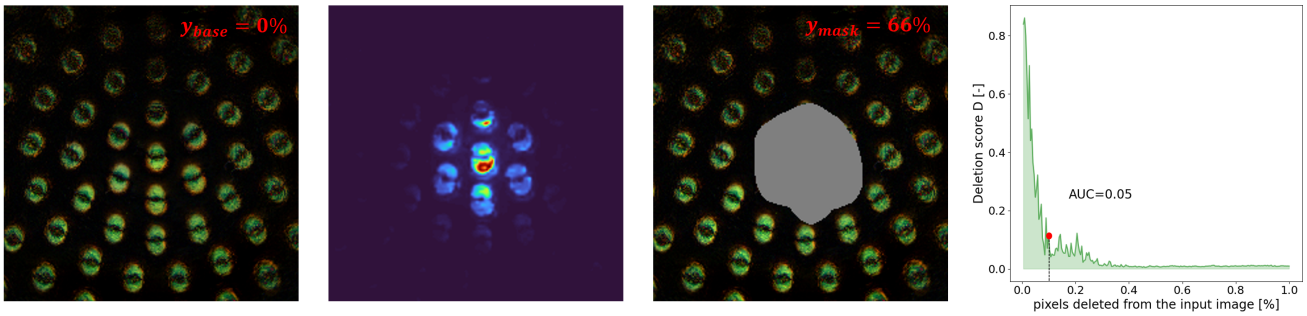


Fig. (5) Faithfulness of XAI. (a) Shows the baseline input image and output y_{base}. (b) The saliency map generated. (c) Demonstrates a sample of the process in gradually removing the most important pixels and the affected output y_{mask}. (d) The curve illustrates the deletion scores as a function of the percentage of pixels removed from the input image.

with a pulling mechanism. Each trial involved the acquisition of tactile images (roughly 800 images per trial) throughout the transition between contact states ($\Gamma : 100\% \rightarrow 0\%$). Grasp forces spanning 0.5-2N were randomly allocated in each trial to promote generalizability across various indentation depths. Tactile images were obtained from two sensors along with force measurements from an embedded force sensor. Using the formula described in Equation 1, the ground truth values Γ were calculated and assigned to the images. To analyze the DNN's reasoning independent of the effects of sensor fusion, our work has focussed on a single sensor.

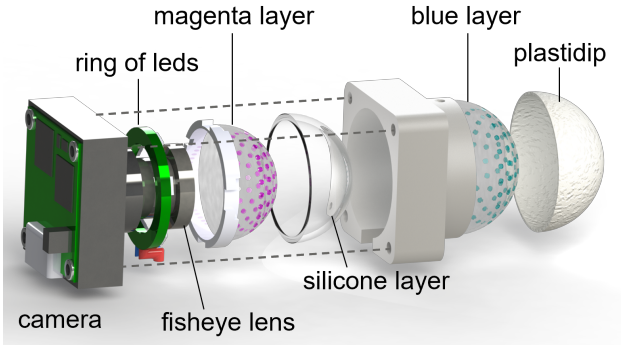


Fig. (6) ChromaTouch tactile sensor

Data splitting To mitigate prior generalization limitations [38], a more balanced dataset was created that uses data from four distinct days. The occurrence of overfitting on a single day could be attributed to fluctuations in contact conditions resulting from humidity-induced variations in friction, as well as potential changes in dome stiffness due to wear and tear. The dataset was divided into three subsets: training (128,043 samples), validation (15,839 samples), and testing (16,875 samples). The training set encompassed samples from flat dry objects collected on day one and a random 75% selection from days two and three. The remaining 25% of samples from these days formed the validation set. The fourth day was exclusively reserved for the two distinct test sets, involving dry and wet objects. This configuration enabled the evaluation of the DNN's robustness against fluctuations in the friction coefficient.

Training details To ensure a fair evaluation between CNNs trained on raw images and difference images, the study adopted the ShufflenetV2 architecture [46], as previously used by Dirk-Jan et al. [38]. The ShufflenetV2 model was loaded with the TorchVision library, and the original fully connected layer was replaced with a single output neuron that represents the value of the frictional safety margin.

Training was performed from scratch using an NVIDIA GeForce RTX 3060 Ti GPU, with an average training time of approximately 2 hours. The training process is based on the PyTorch Lightning [47] framework, which incorporates mixed-precision training and model checkpointing to save the best weights achieved during training. The learning curves for training and validation were logged to Weight & Biases[1] for experiment tracking.

This study used the novel Ranger21 [48] optimizer, which integrates the latest deep learning optimization techniques into a single optimizer, enhancing both validation performance and training speed. The default hyperparameters associated with Ranger21 were used in our experiments. Additional hyperparameters included the MSE loss function, 25 epochs, image size of 224x224, batch size of 124, and a learning rate of 8e-4. Hyperparameter tuning was not conducted in this study, as the chosen optimizer mitigates the sensitivity of the algorithm to its hyperparameters.

To prevent overfitting, a set of random image augmentations was applied by using the Albumentations [49] library. Slight rotations were incorporated to accommodate scenarios where the direction of slip deviates slightly from the gravitational direction. Moreover, the inclusion of Gaussian noise and blur was intended to improve the model's resilience to minor perturbations in the input images during inference. Max normalization was chosen over standardization in order to retain crucial details in the difference images, as standardization sometimes resulted in the loss of information.

1. https://api.wandb.ai/links/max-deep/l59koowi

Performance evaluation The performance of the neural network in predicting the frictional safety margins is assessed using unseen contact cases involving both dry and wet conditions on the same flat object. The evaluation metrics used are the coefficient of determination (R^2) and mean square error (MSE). The R^2 score is a vital metric that quantifies the regression model's goodness of fit. A higher value (ranging from 0 to 1) indicates a better fit between the model and the data. The MSE scores provide information on the accuracy of predicting the safety margin, where lower values indicate better performance.

XAI evaluation Both visual and quantitative assessments were used to evaluate the performance of three commonly used XAI methods (Grad-CAM, Integrated Gradients, and Poly-CAM). The visual assessment involved inspecting the saliency maps for conformity with the high-resolution and smoothness criteria from 3.2.1. Additionally, the quantitative assessment aimed to measure the faithfulness of XAI methods in accurately representing the importance of features, as mentioned in 3.2.2.

The visual and quantitative assessments were performed on a subset of 14 trials from the training set. This subset was strategically chosen to ensure that only the trials were considered where the predictions align closely with the labels (up to ±3% deviation of Γ). The visual assessment considered tactile images with Γ values ranging from 100% to 0% in steps of 20%. This selection aimed to thoroughly analyze saliency maps across each contact state. This made it possible to ascertain the learned patterns of trained DNNs in discerning between high and low Γ conditions. The quantitative assessment places further emphasis on contact states near full slip, considering only Γ labels of 0%, 20%, and 40%. This subset encompassed a total of $n_{data} = 42$ instances.

In terms of implementation, Grad-CAM was sourced from torchcam [50], Integrated Gradients from captum [51], and Poly-CAM from GitHub[2]. Our study used the +/- variant of Poly-CAM, which is known for producing the best results. The saliency maps produced by these XAI methods offered insights into the relative importance of input features in relation to the DNN's output. These saliency values were normalized between 0-1, and a Jet color bar was used for visualization purposes. Normalization and color mapping make it easier to identify and compare relevant regions, improving the overall understandability of the feature attribution results.

Regarding the masking process for the quantitative assessment, the number of masks created is equal to the size of the input image ($n_{masks} = 224$). When it comes to removing regions in an image, one common approach is to set the pixel values to a gray value [45]. However, it's important to note that alternative methods exist, each with their own advantages and disadvantages.

2. https://github.com/aenglebert/polycam

5 Results

This section presents the performance evaluation findings for both DNNs trained on raw and difference tactile images. Additionally, it offers a comprehensive review of the XAI assessment, considering both quantitative and qualitative aspects. Finally, the findings regarding the learned patterns of the DNNs to estimate Γ from tactile images are presented.

5.1 Performance evaluation

The experimental results presented in Table 1 provides a performance comparison between Diff-ShuffleNetv2 and Raw-ShuffleNetv2 models for both dry and wet objects. In particular, Diff-ShuffleNetv2 has a 23% lower MSE along with a 5% higher R²-score compared to Raw ShuffleNetv2. However, the findings reveal that both models face challenges in predicting Γ for wet objects, evident from the significantly higher MSE and lower R²-score. Despite this, Diff-ShuffleNetv2 maintains a 21% lower MSE and an even more substantial improvement of 16% in R²-score.

TABLE (1) The MSE and R^2 values for the two models considering dry and wet contact conditions of a flat object.

Model	Dry		Wet	
	MSE	$\mathbf{R^2}$	**MSE**	$\mathbf{R^2}$
Raw-ShuffleNetv2	$1.87e^{-02}$	0.85	$5.07e^{-02}$	0.57
Diff-ShuffleNetv2	$\mathbf{1.44e^{-02}}$	**0.89**	$\mathbf{4.02e^{-02}}$	**0.66**

Figure 7 illustrates the MSE scores for both dry and wet objects across specific intervals of Γ. The figures clearly show that the two models exhibit higher errors when Γ is within the range of 0-30% in both contact conditions. More specifically, concerning flat dry objects, Diff-ShuffleNetv2 demonstrates a significant lower MSE within the range of 0-30%, comparable MSE values within the ranges of 30-40% and 70-100%, and marginally higher MSE within the range of 40-70%. On the other hand, for flat wet contact surfaces, Diff-ShuffleNetv2 exhibits lower MSE in general. However, its performance notably degrades when Γ ranges from 0% to 10%, in comparison to dry contact surfaces.

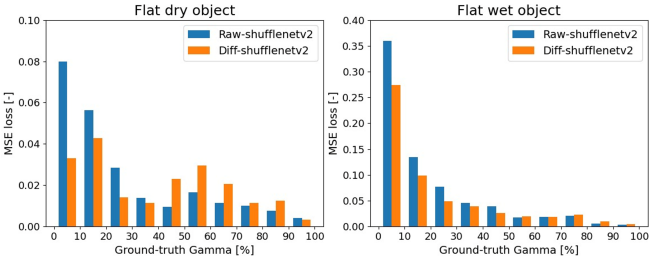


Fig. (7) The MSE scores across the binned range of Γ for both dry and wet contact conditions.

5.2 Visual Assessment of XAI

Visual inspection of saliency maps from XAI methods involved a review of their resolution and smoothness criteria. This evaluation is illustrated in Figure 8 for the contact state $\Gamma = 20\%$. A detailed comparison can be found in Appendix C. The results show that Grad-CAM produces saliency maps with smooth transitions, but it tends to emphasize a single large region rather than localizing specific markers of interest. On the other hand, Integrated Gradients is able to highlight the markers for Diff-ShuffleNetv2, yet it encounters noise-related challenges. Determining the relative importance among markers becomes arduous in this case. Meanwhile, the saliency map for Raw-ShuffleNetv2 is characterized by high levels of noise and is hardly readable. In contrast, Poly-CAM stands out by offering saliency maps that are easily readable and highlight the individual importance of each marker.

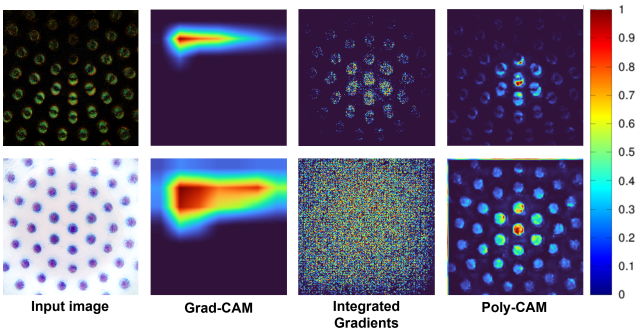


Fig. (8) Visual comparison of the XAI methods applied to Diff-ShuffleNetv2 (top) and Raw-ShuffleNetv2 (bottom). The saliency map highlight the regions used by the model to predict Γ (redder values indicate higher importance).

5.3 Quantitative Assessment of XAI

Table 2 presents a comparison of the faithfulness of saliency maps generated by various XAI methods. Among these methods, Poly-CAM achieved the best score for both models. Interestingly, the XAI methods applied to Diff-ShuffleNetV2 produced significantly lower $\overline{D}_{AUC}$ scores ($> 38\%$) than Raw-ShuffleNetV2. In fact, Integrated Gradients exhibited the largest disparity (60%) between the two models. While Integrated Gradients obtained the second-best score for Diff-ShuffleNetv2, Grad-CAM obtained the second-best score for Raw-ShuffleNetv2.

TABLE (2) $\overline{D}_{AUC}$ scores (lower is better) of the three XAI methods for the two models.

Model \ XAI method	GradCAM	Integrated Gradients	Poly-CAM
Diff-ShuffleNetv2	0.33 ± 0.11	0.27 ± 0.16	$\mathbf{0.20 \pm 0.146}$
Raw-ShuffleNetv2	0.54 ± 0.10	0.68 ± 0.12	$\mathbf{0.51 \pm 0.20}$

To further examine the variations in $\overline{D}_{AUC}$ values between the two models and three XAI methods, the faithfulness curves are presented in Figure 9. For a more detailed view, refer to Appendix D. In the case of Diff-ShuffleNetv2, the mean deletion scores exhibit a gradual decrease as more pixels are removed from the input image, regardless of the XAI methods. More specifically, both Poly-CAM (green) and Integrated Gradient (blue) exhibit a steep decline in deletion scores even with the removal of a small percentage of pixels. Conversely, Grad-CAM (red) shows higher deletion scores ($\overline{D} = 0.7$) in the initial stage when approximately 10% of the pixels are removed, compared to Integrated Gradients ($\overline{D} = 0.5$) and Poly-CAM ($\overline{D} = 0.3$). However, it is intriguing to observe that $\overline{D}$ of Integrated Gradients increases towards the end, when around 50-80% is removed.

Regarding Raw-ShuffleNetv2, the $\overline{D}$ scores of the XAI methods demonstrate a steep decline initially, then remain relatively constant, and eventually experience a decrease when a significant portion of the pixels are removed. It is curious to note that only Integrated Gradients (blue) remains at an average deletion score of 0.7 throughout the process (covering approximately 10-95% of pixel removal), after which it declines sharply. On the other hand, Grad-CAM (red) initially has higher deletion scores (0-40% pixel removal), but undergoes a more rapid decrease in the subsequent stage (40-100% pixel removal) compared to Poly-CAM. However, Poly-CAM (green) demonstrates lower average deletion scores in the early stages (0-55% pixel removal).

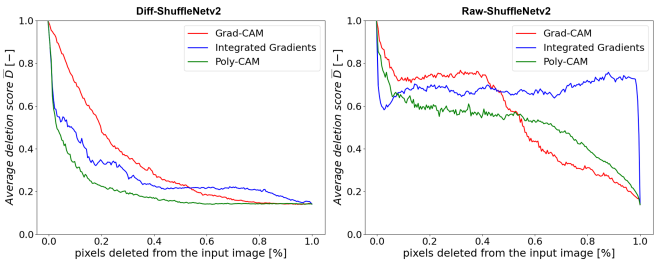


Fig. (9) Faithfulness curves of the XAI methods for Diff-ShuffleNetv2 (left) and Raw-ShuffleNetv2 (right).

A more holistic understanding of the faithfulness curves can be gained by visualizing the masking process. The impact of masking the most salient input features on predictions is shown in Figure 10 (refer to Appendix E for the complete process). This visual examination covers both trained DNNs and all three XAI methods. A careful analysis of the results reveals that in the case of Grad-CAM, the exclusion of the highlighted features does not affect the prediction y_{mask}. In contrast, Integrated Gradients, generates pixel-wise saliency maps, causing the masking process to delete individual pixels rather than entire regions. Despite the removal of a significant amount of information, most markers remain visible. Interestingly, Diff-ShuffleNetv2 achieves a lower deletion score of $D_{20\%} = 0.22$ compared to $D_{50\%} = 0.40$ of Raw-ShuffleNetv2, even though it removes a lesser amount of information. In the case of Poly-CAM, the removal of a relatively small but selective area leads to the lowest deletion scores for both models compared to the other XAI methods.

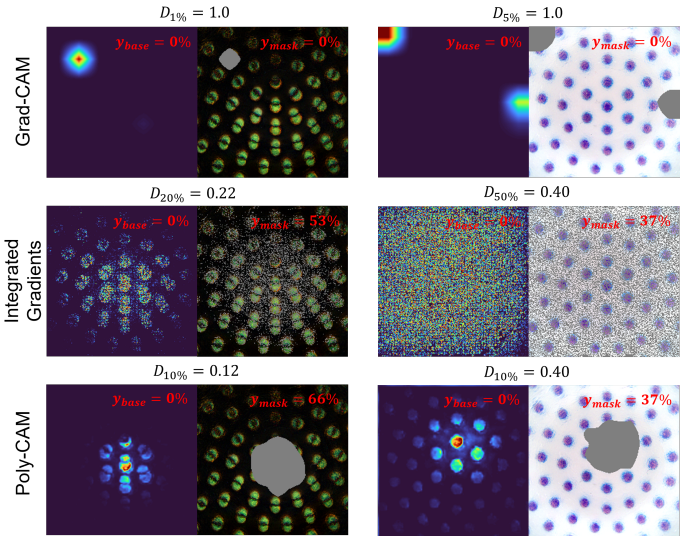


Fig. (10) A visual comparison of removing all highlighted features with saliency values > 0.5 from the input image for Diff-ShuffleNetv2 (left) and Raw-ShuffleNetv2 (right). Here, D_n is the deletion score where n% pixels are removed.

5.4 Neural Network Reasoning of Slip

Poly-CAM has demonstrated promising initial results and has been further used to explore how neural networks interpret and process deformation features in tactile images. Saliency maps for the difference tactile images are presented in Figure 11, and the saliency maps for the raw tactile images are displayed in Figure 12. These saliency maps unveil the regions within tactile images that attract the DNNs' interest, thereby shedding light on how they learned to discern between high and low values of Γ.

Upon analyzing the saliency maps for Diff-ShuffleNetv2, it is evident that the model focuses primarily on two specific markers. These markers alternately play a vital role in predicting the value of Γ. For contact states with high Γ (100-60%), only the upper marker is highlighted. However, as the system approaches full slip (60-0%), the importance of the upper marker diminishes while the prominence of the center marker increases. Interestingly, when examining the input image, the upper marker is clearly visible for high Γ values and becomes less visible as it decreases. The opposite holds true for the center marker. Furthermore, the edges of the overlaid markers, specifically its lower part, emerge as the most crucial feature that the model is looking at. This is clearly visible for both full stick and slip contact states. When Γ reaches the range of 20-0%, the model exclusively concentrates on the central region of contact, disregarding the other markers.

The saliency maps of Raw-ShuffleNetv2 reveal the model's focus on markers situated exclusively in the central contact region. This CNN demonstrates a distinct emphasis on an alternative pair of markers. The marker displacement field is used to analyze the learned pattern because the raw input image is less intuitive. Interestingly, for high values of Γ (100-60%), the displacement vector of the upper marker points towards the top side of the image, while it approaches zero for Γ values of 40%, and points downwards for low values of Γ (20-0%). In contrast, the displacement vector of the center marker is almost zero for high Γ values (100-60%), but points more downwards for low Γ values (40-0%). This behavior is also reflected in the saliency maps. Although the saliency maps of the model indicate a primary focus on the central region of contact, they do not completely disregard the information from other markers. Upon closer examination of the saliency maps, it is surprising to observe that when Γ ranges from 40% to 20%, the upper and left borders are also highlighted.

6 Discussion

Improved Predictive Performance Employing temporal difference images as input for DNNs, as opposed to raw images, has yielded improved performance across both dry and wet contact conditions. In fact, tactile difference images enhance the visibility of how incipient slip evolves. This feature is leveraged by the DNN to achieve more accurate and robust predictions, especially in contact cases where the system approaches full slip (Γ = 40-0%). However, the current performance of the system is still not accurate enough to be applied in realistic manipulation scenarios. This is primarily due to the limited generalization capability of the trained models. Notably, performance significantly degrades with a decreased friction coefficient, such as with wet objects.

Aesthetic Quality of XAI A visual assessment was conducted to evaluate various XAI methods. Among the examined methods, only Poly-CAM satisfied the criteria of producing high-resolution and seamlessly smooth saliency maps. Grad-CAM, on the other hand, produced smooth saliency maps but failed to localize the markers individually. Conversely, Integrated Gradients managed to localize the markers (only for Diff-ShuffleNetv2), but the pixel-level saliency maps were too noisy to make sense of them. The observed noise reduction for Diff-ShuffleNetv2 might be due to positive impact of background removal. Previous studies [19, 42] have used Grad-CAM to explain DNN's outputs in similar touch processing tasks. However, our findings show that only Poly-CAM can precisely highlight markers, thereby offering a more detailed explanation of the DNN's decision.

Faithfulness of XAI This study has adapted the original deletion metric [45] to suit regression tasks, and used it to evaluate the faithfulness of saliency maps. The findings revealed that Poly-CAM is able to accurately localizes salient features with a high degree of faithfulness. In contrast, Integrated Gradients yields slightly less faithful results, because of the noisy saliency map. This distinction is particularly evident considering Diff-ShuffleNetv2. Specifically, the faithfulness curve of Poly-CAM exhibits a sharper decline without reaching a plateau, leading to the lowest ($\overline{D}_{AUC}$) score. Yet, the relatively low $\overline{D}_{AUC}$ scores

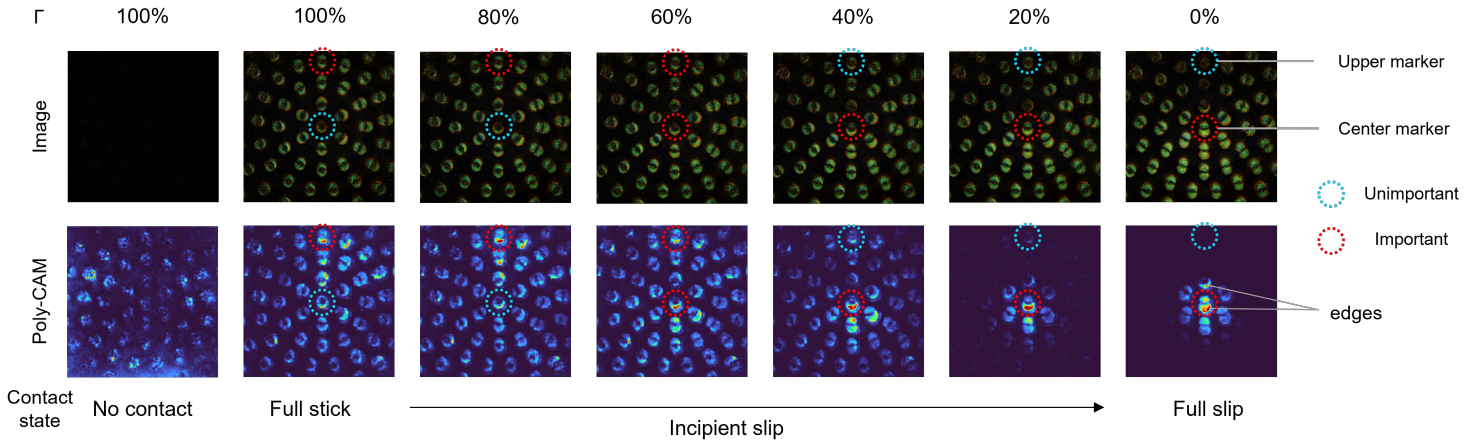


Fig. (11) Poly-CAM visualizations for Diff-ShuffleNetv2 model across different contact states.

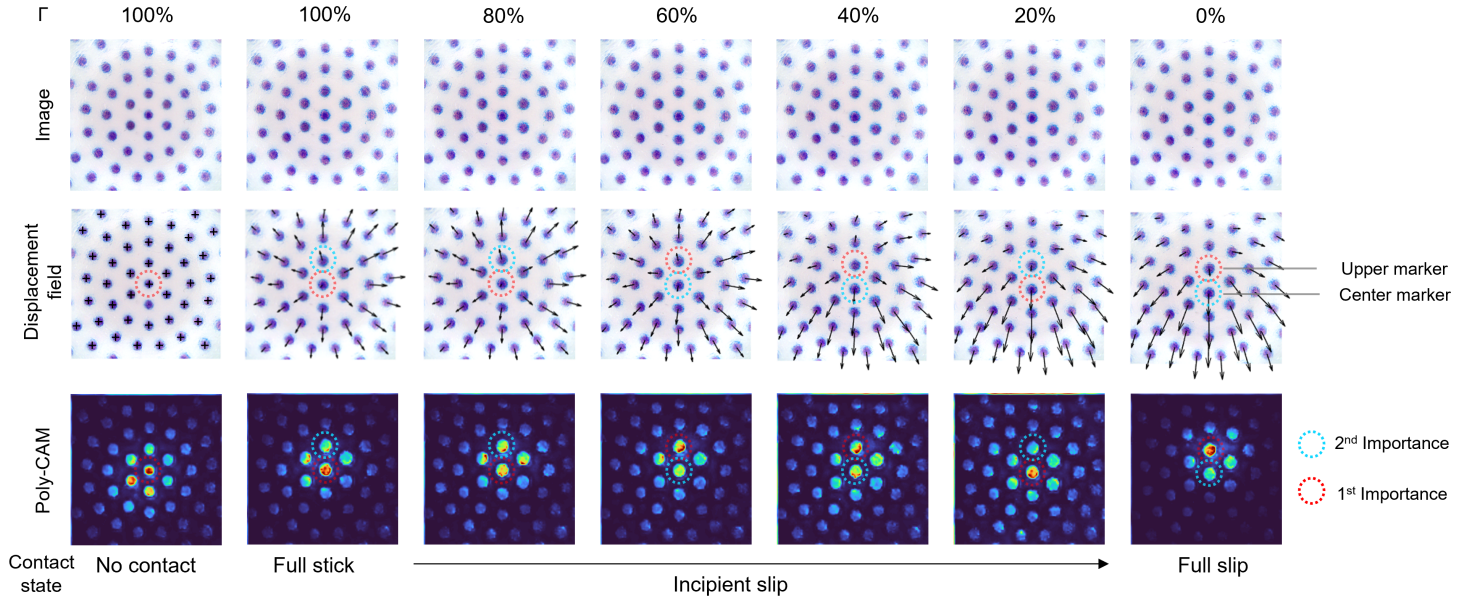


Fig. (12) Poly-CAM visualizations for Raw-ShuffleNetv2 model across different contact states.

of Grad-CAM for both DNNs are surprising, considering that the saliency maps failed to adequately highlight the markers. Further investigation of the deletion process has provided compelling evidence regarding the lack of faithfulness in Grad-CAM explanations. This is supported by the fact that removing the highlighted features from the input image has little to no significant impact on the prediction.

Limitations and Modifications of the Deletion Metric It is crucial to acknowledge the limitations of our metric. This has been noticed from the disparities in the faithfulness curves between the two DNNs, as well as the $\overline{D}_{AUC}$ scores obtained by Grad-CAM. These limitations stem primarily from the fundamental differences between image classification and regression tasks.

The faithfulness metrics were initially designed for image classification, where the gradual removal of important features of the object in the image would decrease the probability score of the target class from one to zero. However, for regression tasks, the removal of important features may potentially alter the output. The extent of this deviation remains ambiguous and depends upon the interplay between the model and data. For instance, in the case of Diff-ShuffleNetv2, progressively removing features from images labeled with low Γ would likely result in a prediction of $\Gamma = 100\%$. This is because an image lacking any markers is indicative of the "no contact" scenario. On the contrary, for Raw-ShuffleNetv2, removing the most important features causes some deviation in the prediction, but further removal of pixels from the image has a minor impact (i.e. constant $\overline{D}$ in the faithfulness curves).

An additional important aspect is the deletion procedure, which entailed progressively removing pixels until the entire image has been masked. This methodology enables an assessment of whether the explanations overlooked vital features or mistakenly emphasized irrelevant features. However, employing the AUC as an overall metric is considered an unfair comparison in our regression task. Particularly, regarding Grad-CAM, upon erasing all highlighted regions, it randomly eliminated other regions that were not highlighted as important. As a consequence, Grad-CAM attained a low $\overline{D}_{AUC}$ score because it

started to remove the central region (see Figure 16). Hence, it is recommended that the deletion process focus solely on computing the AUC for the initial stage by considering only removing the highest feature importance. For instance, the deletion should be stopped until all features with saliency values of > 0.5 have been removed from the input.

Moreover, the deletion method used might introduce outliers, leading to masked images that deviate from the distribution of natural input images. Hence, it is recommended to modify the masking procedure to align with the background, thus minimizing image distortion. Specifically, it is advised to employ white masking for raw images and black masking for difference images, as opposed to gray masking used in the current approach.

Finally, it is essential to acknowledge that the removal of specific markers results in the loss of information, but there may still be other markers within the image that could contribute to the accurate prediction of the incipient slip state. In fact, Poly-CAM visualizations for Raw-ShuffleNetv2 revealed that markers outside the central region were not completely ignored. This might have then played a role in the relatively slower decline and stagnation of the faithfulness curves for Raw-ShuffleNetv2.

Considering these limitations, it is not recommended to directly apply explainability metrics designed for classification to regression tasks without careful consideration. Future research could explore alternative faithfulness metrics that effectively address these constraints, or just simply translate regression problems into a multi-class classification task [52]. By doing so, a more appropriate framework can be established to evaluate the faithfulness of XAI methods in regression tasks. However, treating the task as a multi-class classification problem could provide a deeper understanding of the model's certainty, as opposed to relying solely on a continuous value that may be difficult to interpret.

Validating the Learned Patterns of DNNs Overall, our investigation has successfully identified the best XAI method that is capable of generating saliency maps characterized by high-resolution, smoothness, and faithfulness. In addition, we enhanced the visibility of deformations in tactile images to aid in the comprehension of input-related saliency maps. In fact, raw tactile images lack intuitiveness since the deformations are too subtle to notice for humans. To address this issue, we introduced tactile difference images, which significantly improved the visual rendering of deformations. Nonetheless, analyzing the marker displacement vectors has proven to be effective in interpreting saliency maps derived from raw tactile images. With these advancements, we can finally answer the research question of how these DNNs interpret the deformation captured in tactile images to predict incipient slip (Γ).

Interestingly, it can be concluded that Diff-ShuffleNetV2 successfully localized the deformation features that play a role in the discrimination between high and low Γ. The identified pattern indicates that the model gives significant importance to the edges of the marker stretching when exposed to normal and/or shear forces. This stretching represents the change in marker location of the current tactile image with respect to its reference. As a consequence, the visibility/invisibility of two specific markers emerged as a notable feature that has garnered significant attention from the CNN.

In contrast, Raw-ShuffleNetv2 shows a tendency to look at markers positioned within the central contact area. The CNN learned to recognize the relative displacements among markers, specifically characterized by the opposing directions of the displacement vectors. This attribute has been identified as the key feature for effectively discriminating between high and low Γ. Additionally, the upper and left edges of the image could have potentially functioned as reference points to indicate the degree of marker displacement along the x-y direction.

All things considered, both DNNs align with the human domain knowledge obtained from the contact mechanic analysis. They successfully detect deformations of the soft fingertip occurring within the central region of contact. However, it is important to note that their underlying reasoning is not the same. This happens because they render the deformations differently. To elaborate, Diff-ShuffleNetv2 shows a more localized attention towards the edges of marker stretching, whereas Raw-ShuffleNetv2 employs a more global strategy. To this end, it is anticipated that even a minor change in hyperparameters, such as a different network architecture, can lead to entirely new learned behaviors.

Future Research An intriguing avenue for future research involves leveraging XAI to delve into the fundamental factors contributing to the limited generalizability of DNNs when encountering unfamiliar contact conditions and objects. Through the analysis of saliency maps, a deeper understanding can be achieved regarding the influence of various contact conditions on the behavior of the model. This examination helps to clarify the underlying causes and specific scenarios where failures may arise. As a result, it enables the development of proactive strategies to mitigate these errors, leading to enhanced performance and robustness of DNNs. In order to substantiate this claim, it has been observed that humans exhibit remarkable accuracy in detecting early signs of slip, even when faced with diverse contact conditions or when confronted with unfamiliar objects [53]. Building upon this understanding, it is expected that the insights provided by XAI can lead to a more effective optimization of deep learning models and improving their predictive performance.

7 Conclusion

This study extensively explored the applicability of XAI in the field of tactile sensing. Specifically, our research aimed to enhance the transparency of DNNs in the context of predicting the incipient slip state from tactile images. To achieve this, we utilized feature attribution methods to create visual explanations in the form of saliency maps. These saliency maps shed light on the particular regions within tactile images that influenced the predictions. Consequently, an explainability framework has been developed to elucidate the decisions made by black-box models involved in processing tactile data.

Our methodology emphasized the crucial need to improve the intuitiveness of DNN inputs, allowing human comprehension of the saliency maps. In addition, our comparative analysis showed that Poly-CAM emerges as the most suitable XAI method, as it fulfills the criteria of high-resolution, smoothness, and faithfulness. In light of this, we highly recommend Poly-CAM for explaining the decisions of DNNs in other touch-related tasks (e.g. touch detection and texture recognition).

In conclusion, our explainability framework synergized intuitive input with the most effective XAI method, enabling a comprehensive analysis and understanding of the neural network's decision-making process. Our results confirmed that the network can effectively recognize markers correlated with incipient slip, while aligning with the contact mechanics of deformation. More importantly, the enhanced clarity and detailed explanations of Poly-CAM facilitated a more nuanced comparison of DNNs regarding their ability to identify salient deformation features.

The findings of our study make notable contributions to the progression of XAI in the field of tactile sensing. Going forward, XAI holds promising potential as a valuable tool for debugging AI systems and fostering human involvement in the machine learning process. It can assist in developing well-informed strategies to address the failures and limitations of DNNs in tactile sensing applications.

Acknowledgments

The success of this project can be attributed to the invaluable contributions made by a group of remarkable individuals. I would like to express my deep gratitude to Dr. Michael Wiertlewski for his expertise in tactile sensing, as well as to Ir. Joris Kuiper for his guidance throughout the thesis. I would like to acknowledge the collaborative efforts with Ir. Dirk-Jan Boonstra, Ir. Laurence Willemet, Ir. Giuseppe Vitrani, Ir. Julian de Wit, and Ir. Mehrdad Jahanbanifard. Additionally, I appreciate the valuable contributions made by Ir. Alexandre Englebert in the redesign of the faithfulness metric.

References

[1] J. Zhu, A. Cherubini, C. Dune, D. Navarro-Alarcon, F. Alambeigi, D. Berenson, F. Ficuciello, K. Harada, J. Kober, X. Li, J. Pan, W. Yuan, and M. Gienger, "Challenges and Outlook in Robotic Manipulation of Deformable Objects," IEEE Robotics and Automation Magazine, vol. 29, no. 3, pp. 67–77, 9 2022.

[2] C. Blanes, M. Mellado, C. Ortiz, and A. Valera, "Review. Technologies for robot grippers in pick and place operations for fresh fruits and vegetables," Spanish Journal of Agricultural Research, vol. 9, no. 4, pp. 1130–1141, 2011.

[3] A. Ikeda, Y. Kurita, J. Ueda, Y. Matsumoto, and T. Ogasawara, "Grip force control for an elastic finger using vision-based incipient slip feedback," 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), vol. 1, pp. 810–815, 2004.

[4] R. S. Johansson and G. Westling, "Roles of glabrous skin receptors and sensorimotor memory in automatic control of precision grip when lifting rougher or more slippery objects," Experimental brain research, vol. 56, no. 3, pp. 550–564, 10 1984. [Online]. Available: https://pubmed-ncbi-nlm-nih-gov.tudelft.idm.oclc.org/6499981/

[5] L. Willemet, N. Huloux, and M. Wiertlewski, "Efficient tactile encoding of object slippage," Scientific Reports 2022 12:1, vol. 12, no. 1, pp. 1–9, 8 2022. [Online]. Available: https://www-nature-com.tudelft.idm.oclc.org/articles/s41598-022-16938-1

[6] R. S. Johansson and G. Westling, "Signals in tactile afferents from the fingers eliciting adaptive motor responses during precision grip," Experimental Brain Research, vol. 66, no. 1, pp. 141–154, 3 1987. [Online]. Available: https://link.springer.com/article/10.1007/BF00236210

[7] R. S. Johansson and J. R. Flanagan, "Coding and use of tactile signals from the fingertips in object manipulation tasks," Nature Reviews Neuroscience 2009 10:5, vol. 10, no. 5, pp. 345–359, 4 2009. [Online]. Available: https://www-nature-com.tudelft.idm.oclc.org/articles/nrn2621

[8] W. Chen, H. Khamis, I. Birznieks, N. F. Lepora, and S. J. Redmond, "Tactile Sensors for Friction Estimation and Incipient Slip Detection - Toward Dexterous Robotic Manipulation: A Review," IEEE Sensors Journal, vol. 18, no. 22, pp. 9049–9064, 11 2018.

[9] K. Shimonomura, "Tactile Image Sensors Employing Camera: A Review," Sensors, vol. 19, no. 18, p. 3933, 9 2019. [Online]. Available: https://doi.org/10.3390/s19183933

[10] M. Li, L. Zhang, T. Li, and Y. Jiang, "Continuous Marker Patterns for Representing Contact Information in Vision-Based Tactile Sensor: Principle, Algorithm, and Verification," IEEE Transactions on Instrumentation and Measurement, vol. 71, 2022.

[11] W. Yuan, S. Dong, and E. Adelson, "GelSight: High-Resolution Robot Tactile Sensors for Estimating Geometry and Force," Sensors, vol. 17, no. 12, p. 2762, 11 2017. [Online]. Available: https://doi.org/10.3390/s17122762

[12] M. Bazrafshan, M. B. de Rooij, and D. J. Schipper, "On the role of adhesion and roughness in stick-slip transition at the contact of two bodies: A numerical study," Tribology International, vol. 121, pp. 381–388, 5 2018.

[13] S. Dong, D. Ma, E. Donlon, and A. Rodriguez, "Maintaining grasps within slipping bounds by monitoring incipient slip," Proceedings - IEEE International Conference on Robotics and Automation, vol. 2019-May, pp. 3818–3824, 5 2019.

[14] S. Dong, W. Yuan, and E. H. Adelson, "Improved GelSight tactile sensor for measuring geometry and slip," IEEE International Conference on Intelligent Robots and Systems, vol. 2017-September, pp. 137–144, 12 2017.

[15] R. B. Scharff, D. J. Boonstra, L. Willemet, X. Lin, and M. Wiertlewski, "Rapid manufacturing of color-based hemispherical soft tactile fingertips," 2022 IEEE 5th International Conference on Soft Robotics, RoboSoft 2022, pp. 896–902, 2022.

[16] J. Li, S. Dong, and E. Adelson, "Slip Detection with Combined Tactile and Visual Information," Proceedings - IEEE International Conference on Robotics and Automation, pp. 7772–7777, 9 2018.

[17] Y. Zhang, Z. Kan, Y. A. Tse, Y. Yang, and M. Y. Wang, "FingerVision Tactile Sensor Design and Slip Detection Using Convolutional LSTM Network," 10 2018. [Online]. Available: https://arxiv.org/abs/1810.02653v1

[18] M. Lambeta, H. Xu, J. Xu, P. W. Chou, S. Wang, T. Darrell, and R. Calandra, "PyTouch: A Machine Learning Library for Touch Processing," Proceedings - IEEE International Conference on Robotics and Automation, vol. 2021-May, pp. 13 208–13 214, 2021.

[19] Y. Han, R. Batra, N. Boyd, T. Zhao, Y. She, S. Hutchinson, and Y. Zhao, "Learning Generalizable Vision-Tactile Robotic Grasping Strategy for Deformable Objects via Transformer," 12 2021. [Online]. Available: https://arxiv.org/abs/2112.06374v4

[20] R. Calandra, A. Owens, M. Upadhyaya, W. Yuan, J. Lin, E. H. Adelson, and S. Levine, "The Feeling of Success: Does Touch Sensing Help Predict Grasp Outcomes?" 10 2017. [Online]. Available: https://arxiv.org/abs/1710.05512v1

[21] A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," IEEE Access, vol. 6, pp. 52 138–52 160, 9 2018.

[22] T. Eche, L. H. Schwartz, F. Z. Mokrane, and L. Dercle, "Toward Generalizability in the Deployment of Artificial Intelligence in Radiology: Role of Computation Stress Testing to Overcome Underspecification," Radiology Artificial intelligence, vol. 3, no. 6, 11 2021. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/34870222/

[23] B. H. van der Velden, H. J. Kuijf, K. G. Gilhuijs, and M. A. Viergever, "Explainable artificial intelligence (XAI) in deep learning-based medical image analysis," Medical Image Analysis, vol. 79, p. 102470, 7 2022.

[24] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable AI: A Review of Machine Learning Interpretability Methods," Entropy 2021, Vol. 23, Page 18, vol. 23, no. 1, p. 18, 12 2020. [Online]. Available: https://www.mdpi.com/1099-4300/23/1/18/htmhttps://www.mdpi.com/1099-4300/23/1/18

[25] R. Sui, L. Zhang, T. Li, and Y. Jiang, "Incipient slip detection method with vision-based tactile sensor based on distribution force and deformation," IEEE Sensors Journal, vol. 21, no. 22, pp. 25 973–25 985, 11 2021.

[26] X. Lin and M. Wiertlewski, "Sensing the Frictional State of a Robotic Skin via Subtractive Color Mixing," IEEE Robotics and Automation Letters, vol. 4, no. 3, pp. 2386–2392, 7 2019.

[27] X. Lin, L. Willemet, A. Bailleul, and M. Wiertlewski, "Curvature sensing with a spherical tactile sensor using the color-interference of a marker array," Proceedings - IEEE International Conference on Robotics and Automation, pp. 603–609, 5 2020.

[28] R. Sui, L. Zhang, T. Li, and Y. Jiang, "Incipient slip detection method for soft objects with vision-based tactile sensor," Measurement, vol. 203, p. 111906, 11 2022.

[29] N. Watanabe and G. Obinata, "Grip force control based on the degree of slippage using optical tactile sensor," 2007 International Symposium on Micro-NanoMechatronics and Human Science, MHS, pp. 466–471, 11 2007.

[30] I. Waters, D. Jones, A. Alazmani, and P. R. Culmer, "Utilising Incipient Slip for Grasping Automation in Robot Assisted Surgery," IEEE Robotics and Automation Letters, vol. 7, no. 2, pp. 1071–1078, 4 2022.

[31] M. Wiertlewski, S. Endo, A. M. Wing, and V. Hayward, "Slip-induced vibration influences the grip reflex: A pilot study," 2013 World Haptics Conference, WHC 2013, pp. 627–632, 2013.

[32] R. S. Johansson and J. R. Flanagan, "Tactile Sensory Control of Object Manipulation in Humans," The Senses: A Comprehensive Reference, vol. 6, pp. 67–86, 1 2008.

[33] A. M. Hadjiosif and M. A. Smith, "Flexible Control of Safety Margins for Action Based on Environmental Variability," Journal of Neuroscience, vol. 35, no. 24, pp. 9106–9121, 6 2015. [Online]. Available: https://doi.org/10.1523/jneurosci.1883-14.2015

[34] Y. Ito, Y. W. Kim, and G. Obinata, "Slippage degree estimation for dexterous handling of vision-based tactile sensor," Proceedings of IEEE Sensors, pp. 449–452, 2009.

[35] W. Yuan, R. Li, M. A. Srinivasan, and E. H. Adelson, "Measurement of shear and slip with a GelSight tactile sensor," Proceedings - IEEE International Conference on Robotics and Automation, vol. 2015-June, no. June, pp. 304–311, 6 2015.

[36] C. Sferrazza and R. D'Andrea, "Transfer learning for vision-based tactile sensing," IEEE International Conference on Intelligent Robots and Systems, pp. 7961–7967, 11 2019.

[37] C. Sferrazza, T. Bi, and R. D'Andrea, "Learning the sense of touch in simulation: A sim-to-real strategy for vision-based tactile sensing," IEEE International Conference on Intelligent Robots and Systems, pp. 4389–4396, 10 2020.

[38] D.-J. Boonstra, L. Willemet, J. Luijkx, and M. Wiertlewski, "Learning Gentle Grasps for Robust Robotic Manipulation using high-resolution Tactile Images," Not yet published, 2023.

[39] A. Das, G. Student Member, P. Rad, and S. Member, "Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey," 6 2020. [Online]. Available: https://arxiv.org/abs/2006.11371v2

[40] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning Deep Features for Discriminative Localization," Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2016-December, pp. 2921–2929, 12 2016.

[41] A. Englebert, O. Cornu, and C. De Vleeschouwer, "Backward recursive Class Activation Map refinement for high resolution saliency map," Proceedings - International Conference on Pattern Recognition, vol. 2022-August, pp. 2444–2450, 2022.

[42] G. Cao, Y. Zhou, D. Bollegala, and S. Luo, "Spatio-temporal Attention Model for Tactile Texture Recognition," 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 9896–9902, 10 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9341333/

[43] R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," Proceedings of the IEEE International Conference on Computer Vision, vol. 2017-October, pp. 618–626, 12 2017.

[44] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic Attribution for Deep Networks," 34th International Conference on Machine Learning, ICML 2017, vol. 7, pp. 5109–5118, 3 2017. [Online]. Available: https://arxiv.org/abs/1703.01365v2

[45] V. Petsiuk, A. Das, and K. Saenko, "RISE: Randomized Input Sampling for Explanation of Black-box Models," British Machine Vision Conference 2018, BMVC 2018, 6 2018. [Online]. Available: https://arxiv.org/abs/1806.07421v3

[46] N. Ma, X. Zhang, H. T. Zheng, and J. Sun, "Shufflenet V2: Practical guidelines for efficient cnn architecture design," Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 11218 LNCS, pp. 122–138, 2018. [Online]. Available: https://link-springer-com.tudelft.idm.oclc.org/chapter/10.1007/978-3-030-01264-9_8

[47] W. Falcon and The PyTorch Lightning team, "PyTorch Lightning," 3 2019.

[48] L. Wright and N. Demeure, "Ranger21: a synergistic deep learning optimizer," 6 2021. [Online]. Available: https://arxiv.org/abs/2106.13731v2

[49] C. S. Greene, "Albumentations: Fast and Flexible Image Augmentations," Information 2020, Vol. 11, Page 125, vol. 11, no. 2, p. 125, 2 2020. [Online]. Available: https://doi.org/10.3390/info11020125

[50] F.-G. Fernandez, "TorchCAM: class activation explorer," https://github.com/frgfm/torch-cam, 3 2020. [Online]. Available: https://github.com/frgfm/torch-cam

[51] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan, and O. Reblitz-Richardson, "Captum: A unified and generic model interpretability library for PyTorch," 9 2020. [Online]. Available: https://arxiv.org/abs/2009.07896v1

[52] S. Letzgus, P. Wagner, J. Lederer, W. Samek, K. R. Muller, and G. Montavon, "Toward Explainable Artificial Intelligence for Regression Models: A methodological perspective," IEEE Signal Processing Magazine, vol. 39, no. 4, pp. 40–58, 7 2022.

[53] J. D. Lieber and S. J. Bensmaia, "Emergence of an Invariant Representation of Texture in Primate Somatosensory Cortex," Cerebral Cortex, vol. 30, no. 5, pp. 3228–3239, 5 2020. [Online]. Available: https://dx-doi-org.tudelft.idm.oclc.org/10.1093/cercor/bhz305

Appendices

Appendix A Visual comparison of XAI methods in image classification

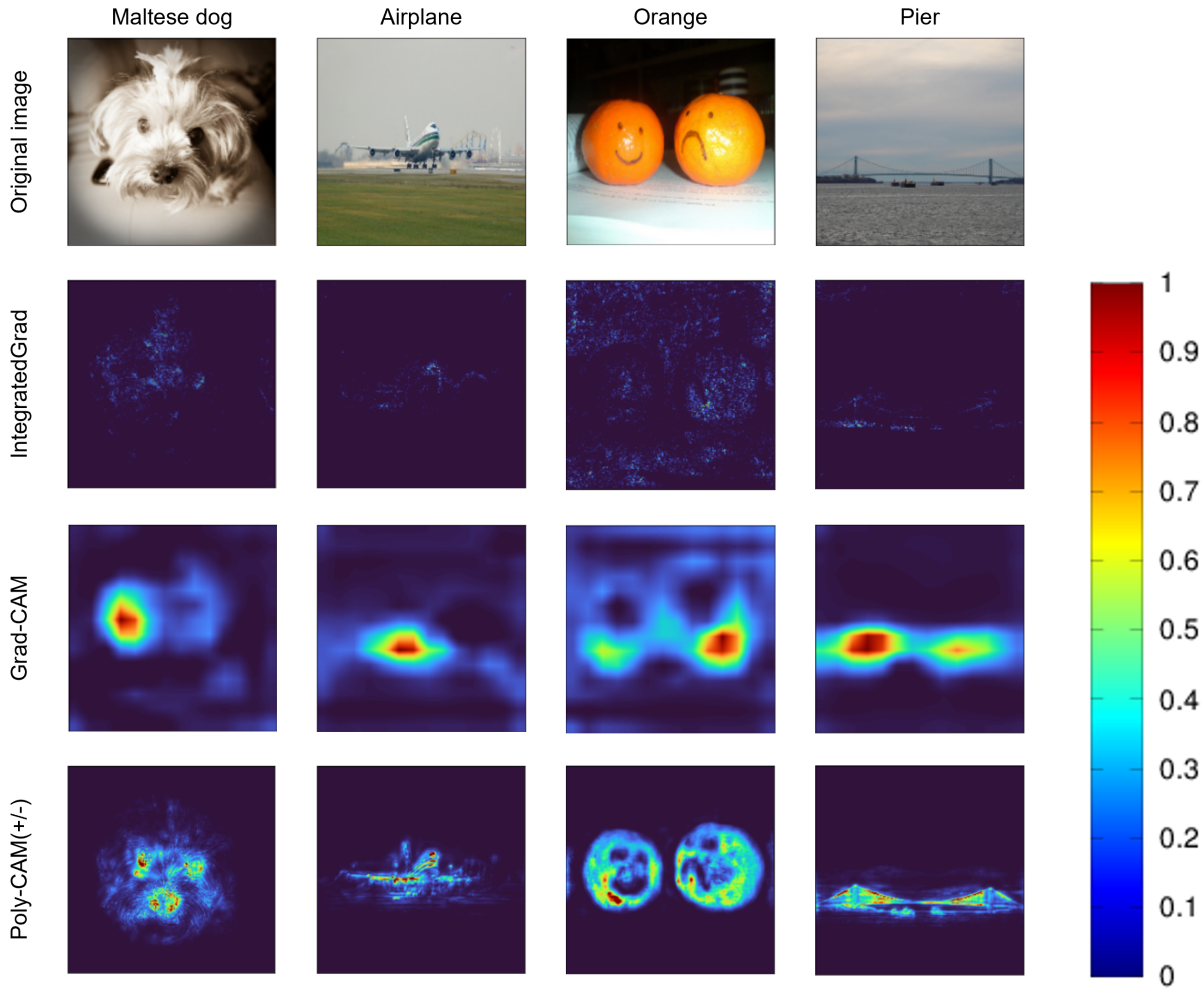


Fig. (13) Examples of saliency maps generated by three distinct XAI techniques, which were applied to a pre-trained ResNet50 model for the purpose of image classification on the widely-used ImageNet dataset. The full comparison is available in the study from Englebert et al. [41].

APPENDIX B FAITHFULNESS OF XAI METHODS IN IMAGE CLASSIFICATION

In the context of image classification, the deletion metric (lower better) involves the iterative removal of the most salient features from the input image, while observing the resulting impact on the model's output. The model's output is represented by the probability score assigned to the predicted target class in the image.

For instance, in Figure 14, the network confidently predicted the presence of the 'goldfish' class in the image with a probability of approximately 0.95. The XAI method generates saliency maps to highlight the importance of each region for the model's prediction. In the deletion metric, input features are masked starting with the most important ones based on the explanations provided. The deletion curve depicts how the probability score decreases progressively as more and more important features are removed.

A faithful saliency map should exhibit a substantial decrease in the probability score, ideally transitioning from approximately 1 to 0. The swifter this drop occurs, the greater the faithfulness and localization of the explanation. Conversely, a less faithful explanation would manifest a slower decline and may plateau above 0. Such a pattern suggests that certain important features have not been sufficiently emphasized or that irrelevant information has been highlighted. This observation is evident in the Grad-CAM [43] explanations, whereas RISE [45] exhibits superior localization of salient features, resulting in the lowest AUC. The limitations of Grad-CAM become apparent in this case, as it fails to effectively emphasize multiple objects that are actually important and present in the input image.

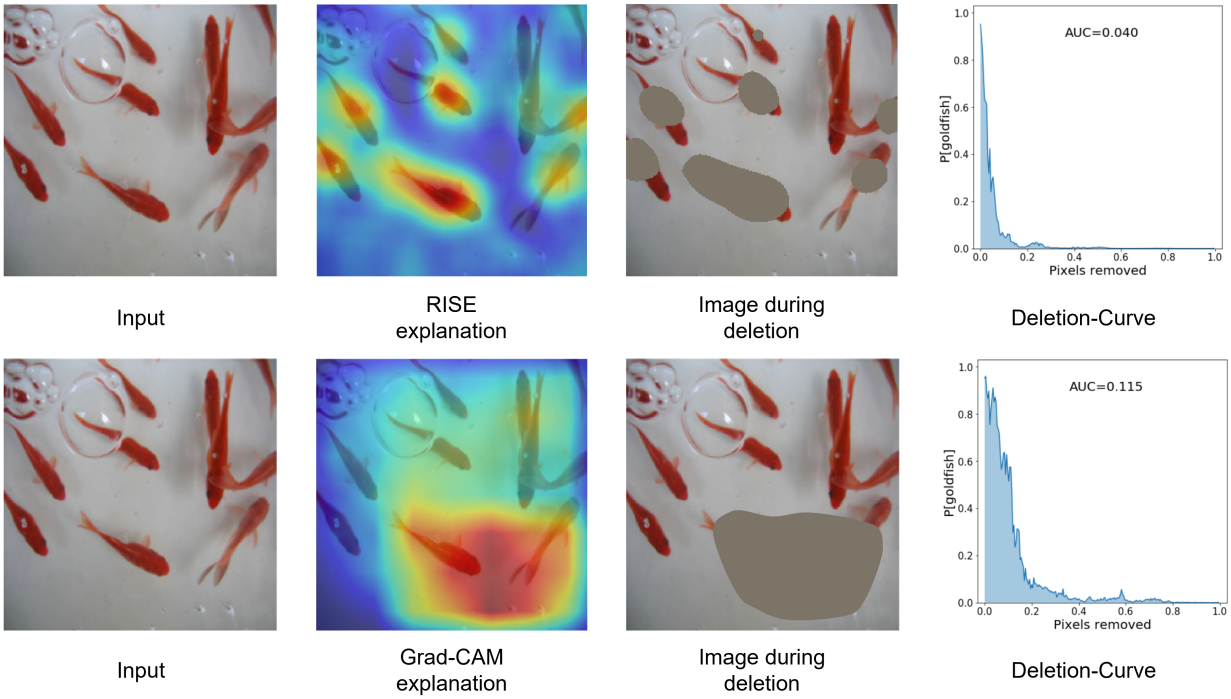


Fig. (14) Estimation of importance of each region by RISE and Grad-CAM for a base model's prediction, along with 'deletion' scores (AUC). See the full comparison here [45].

APPENDIX C VISUAL COMPARISON OF XAI METHODS IN TACTILE SENSING

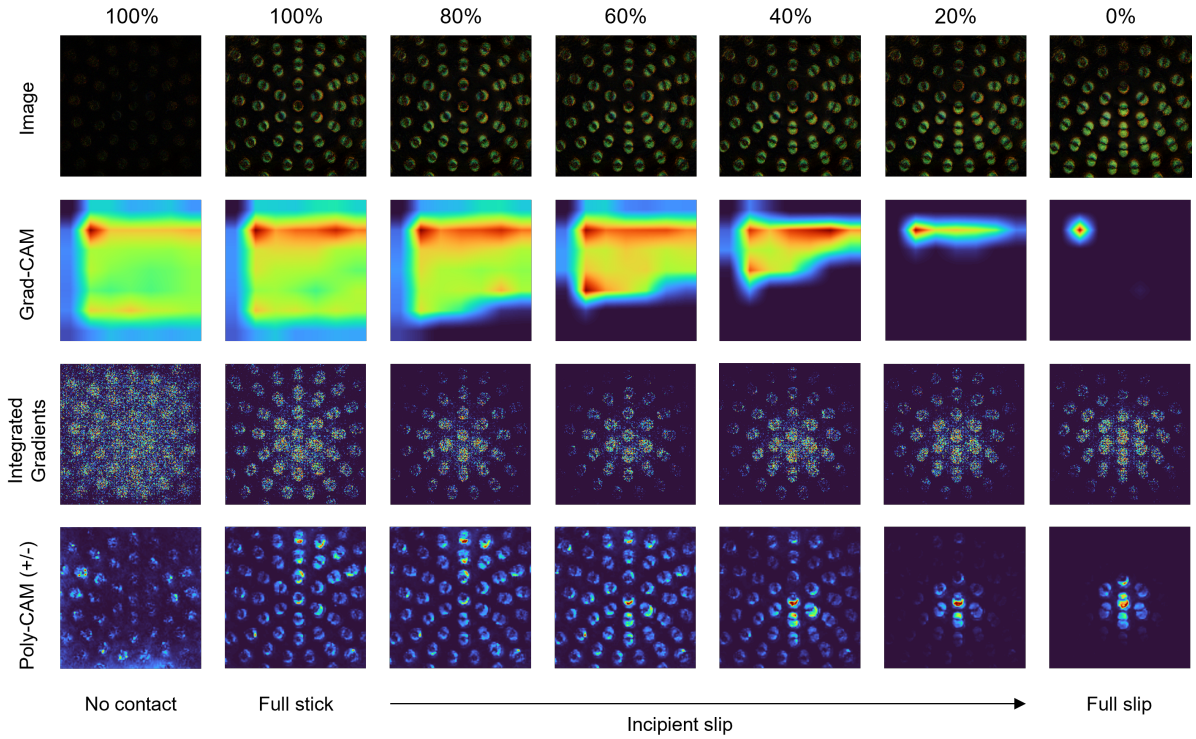


Fig. (15) Visual comparison of the XAI methods applied to Diff-ShuffleNetv2 across the contact states. The XAI methods under consideration include Grad-CAM [43], Integrated Gradients [44], and Poly-CAM [41].

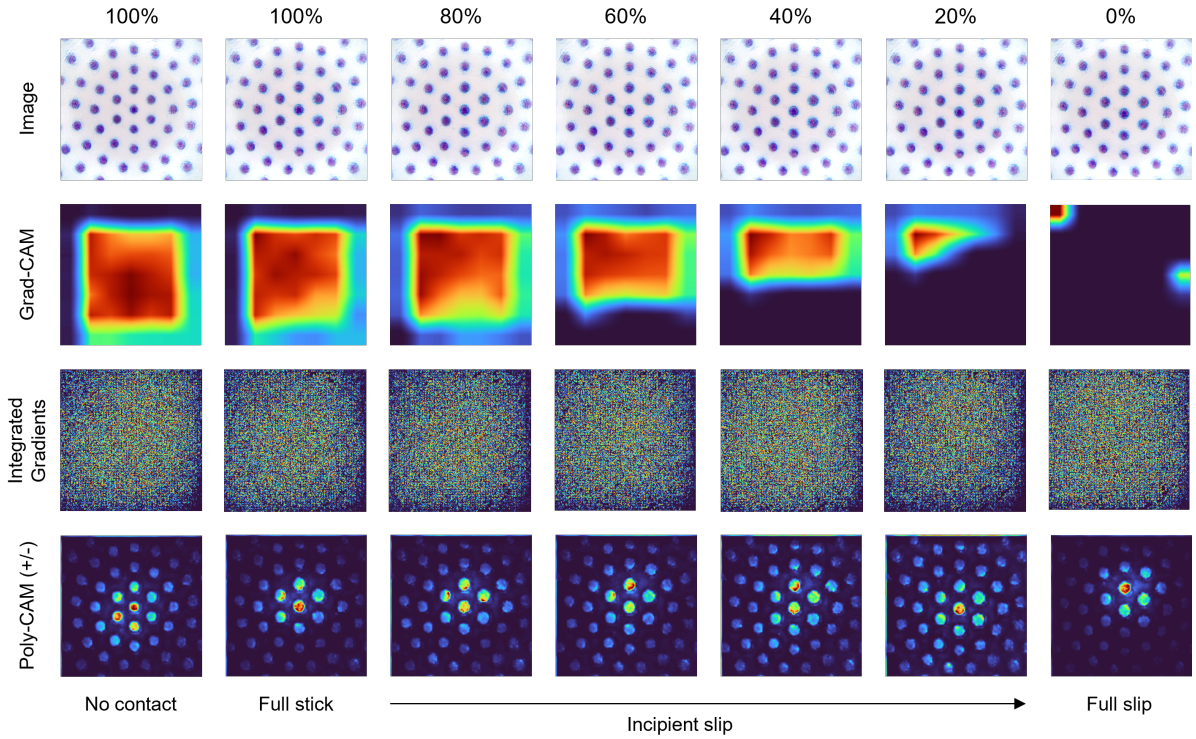


Fig. (16) Visual comparison of the XAI methods applied to Raw-ShuffleNetv2 across the contact states. The XAI methods under consideration include Grad-CAM [43], Integrated Gradients [44], and Poly-CAM [41].

The computation time for generating saliency maps has also been considered. In the XAI evaluation set, which consisted of 42 images, the average GPU processing time for generating a single saliency maps was as follows: Grad-CAM took 0.029±0.017s, Integrated Gradients took 0.27±0.0073s, and Poly-CAM took 4.9±0.47s. Poly-CAM explanations takes the most time, as it requires performing multiple input perturbations and monitoring their impact on the output. Nevertheless, the saliency maps are highly detailed and clear.

APPENDIX D FAITHFULNESS OF XAI METHODS IN TACTILE SENSING

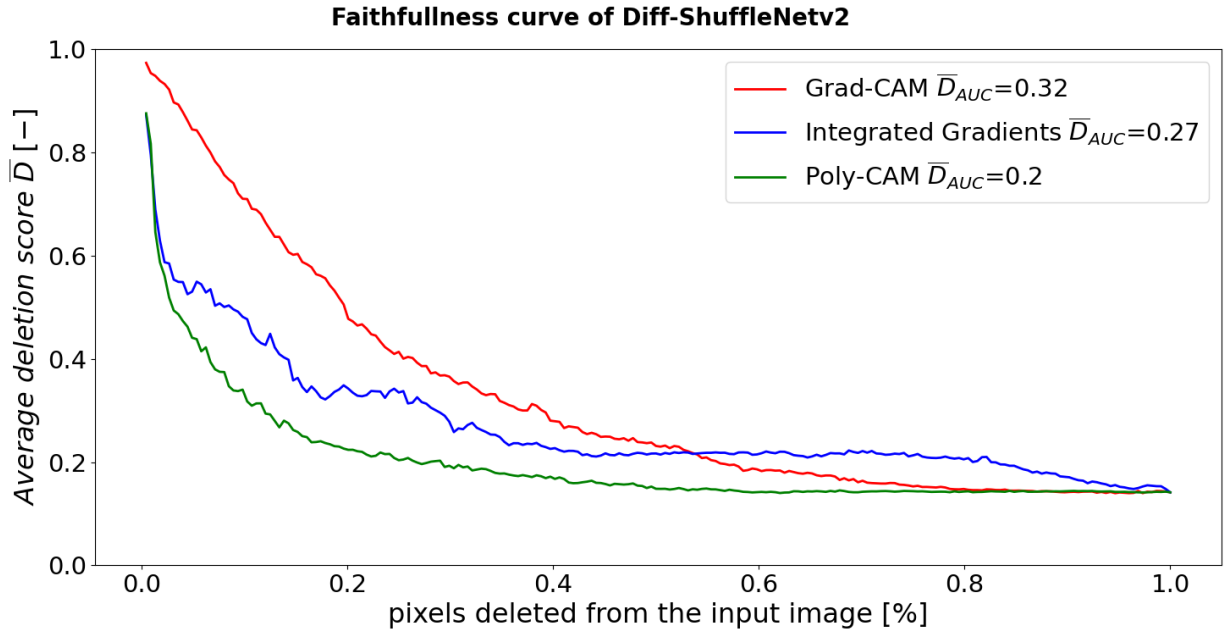


Fig. (17) Faithfulness curves of the XAI methods for Diff-ShuffleNetv2 considering 42 samples. Here, Poly-CAM shows a fast decline without stagnation; Grad-CAM shows a less steep decline but doesn't stagnate; and Integrated Gradients shows a fast decline but stagnates slightly before removing all the pixels.

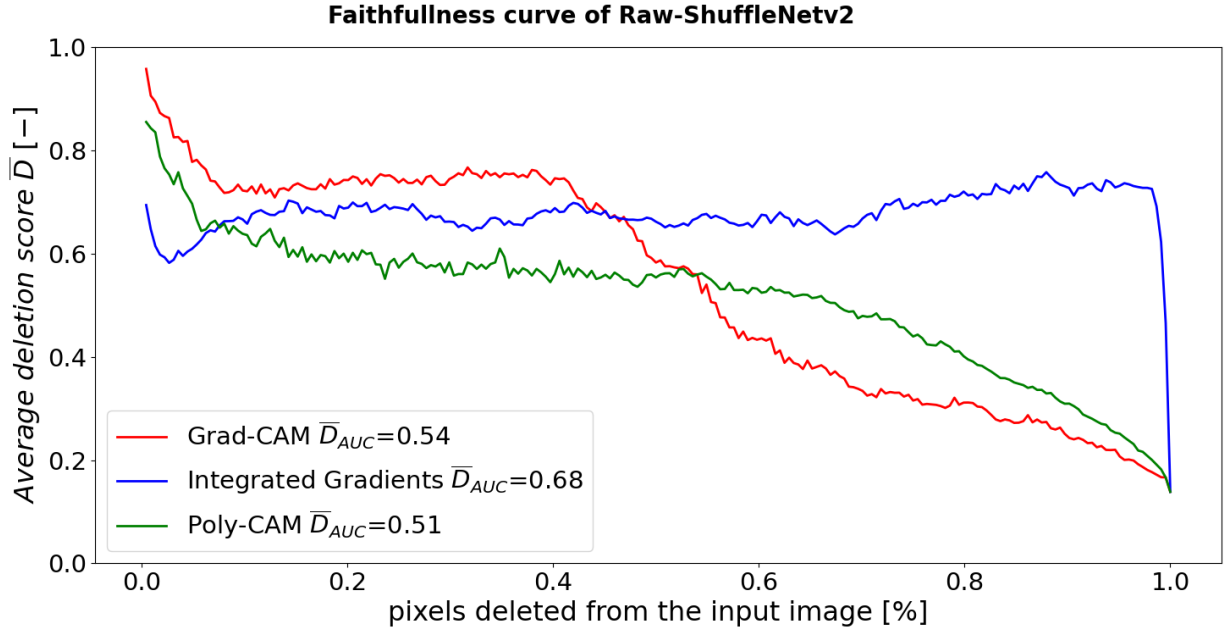


Fig. (18) Faithfulness curves of the XAI methods for Raw-ShuffleNetv2 considering 42 samples. All XAI methods shows a period of stagnation, with Integrated Gradients being the longest.

APPENDIX E VISUALIZATION OF DELETION PROCESS

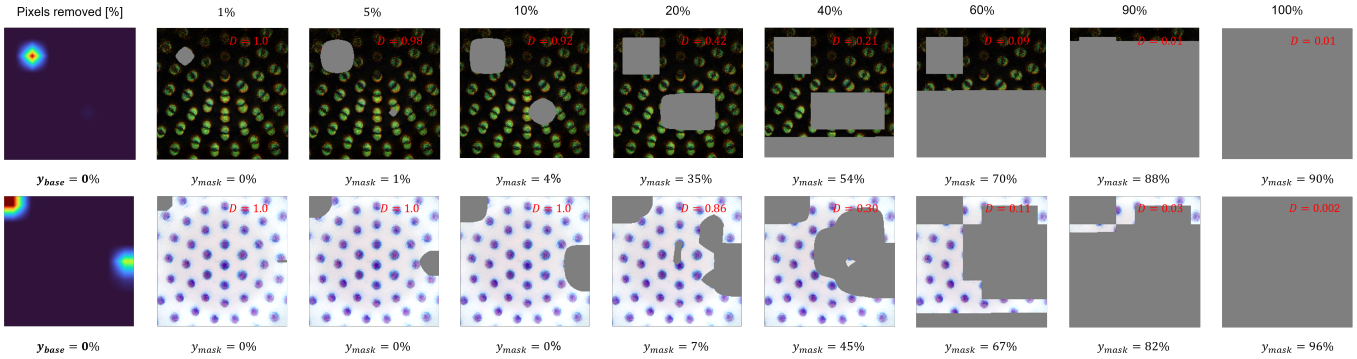


Fig. (19) Deletion process considering the saliency map produced by **Grad-CAM** for Diff-ShuffleNetv2 (top) and Raw-ShuffleNetv2 (bottom). Despite the removal of all the significant features (5% of the input image), the prediction y_{mask} remains unchanged. This suggests that Grad-CAM did not effectively identify the truly important features. As the deletion process progresses, random regions start to be removed. Notably, when the center region is removed for Diff-ShuffleNetv2 at 20% and Raw-ShuffleNetv2 at 40%, the prediction y_{mask} deteriorates, and the deletion score decreases accordingly.

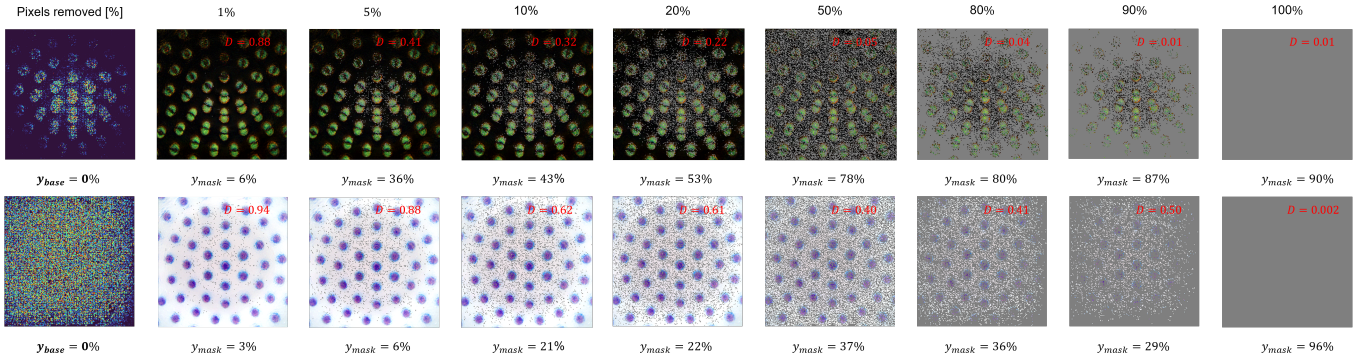


Fig. (20) Deletion process considering the saliency map produced by **Integrated Gradients** for Diff-ShuffleNetv2 (top) and Raw-ShuffleNetv2 (bottom). The saliency map provides pixel-wise importance, and thus the masking process involves removing pixels one by one. As more important features are removed, the predictions y_{mask} for Diff-ShuffleNetv2 deviate further from the correct outcome y_{base}, whereas the predictions y_{mask} for Raw-ShuffleNetv2 plateaus after removing 10% of the pixels. Despite the significant alteration of the image caused by masking, a majority of the markers remain visible, which still can be used to infer the incipient slip state. This explains why the prediction is not significantly deprecated (i.e. somewhat constant D).

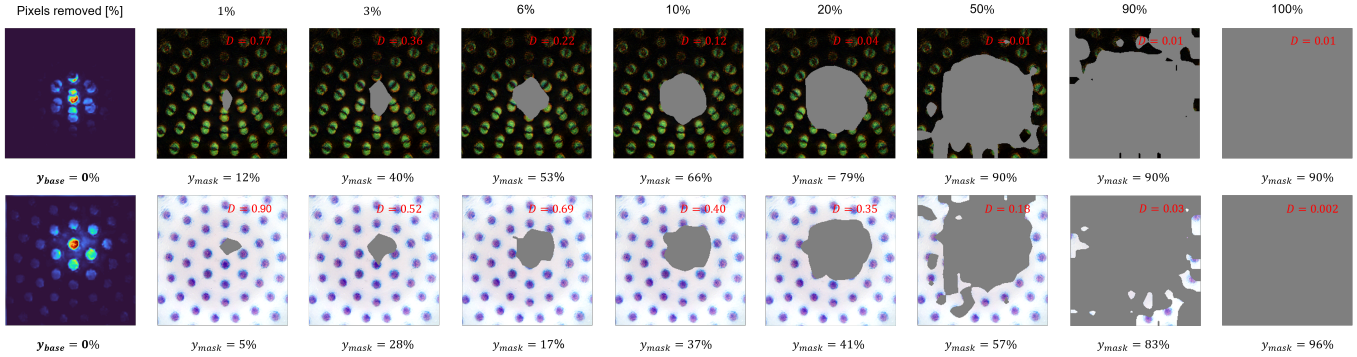


Fig. (21) Deletion process considering the saliency map produced by **Poly-CAM** for Diff-ShuffleNetv2 (top) and Raw-ShuffleNetv2 (bottom). The removal of all the significant features (10% of the input image), leading to a large deviation of the prediction, (lower D). This means that the highlighted input features were indeed important to make the accurate prediction. Notably, the prediction deviation decreases more rapidly for Diff-ShuffleNetv2 compared to Raw-ShuffleNetv2.