

Insights from Inside: Towards Explainable WiFi Sensing

Mina Shahbazifar¹, Dirk Schumacher², Zolfa Zeinalpour-Yazdi³, Mohammad Zoofaghari⁴, Matthias Hollick⁵,
Member, IEEE and Arash Asadi⁶, Senior Member, IEEE

Abstract—WiFi sensing relies heavily on blackbox machine learning (Machine Learning (ML)) models due to the large feature space and complexity. Despite achieving very high accuracies in complex scenarios, the blackbox nature of these ML-based sensing techniques is commonly criticized. This is in fact a major source of mistrust as these models provide very little explanation supporting their decision, while often handling critical applications (e.g., elderly monitoring). In this paper, we investigate explainable artificial intelligence (eXplainable Artificial Intelligence (XAI)) techniques to shed light on the decisions and behaviors of such blackbox models. Specifically, we propose eXSense, a workflow designed based on state-of-the-art XAI techniques to analyze the behavior of blackbox models both locally and globally. To demonstrate its potential, we conduct an extensive analysis on two case studies from the recent sensing literature. Finally, leveraging the insights obtained from our analysis, we propose and evaluate changes to these models, thus enhancing their efficiency and reliability. This includes reducing the feature space by at least 80% with no/minimal loss ($\leq 1\%$) to the model accuracy.

Index Terms—WiFi sensing, Explainability, Activity recognition, CSI.

I. INTRODUCTION

WiFi Sensing research has seen major progress addressing a wide span of applications from activity recognition (e.g., sitting [1], walking [2]) to fine-grained gesture recognition (lips [3], fingers [4]), surveillance and security [5], health/elderly monitoring (fall detection [6], vital sign monitoring [7]), and even emotion detection [8]. The relevance of this field is further emphasized by the ongoing standardization (i.e. 802.11bf) on the way which enables native WiFi sensing [9]. The majority of these applications tap into advanced Machine Learning (ML) algorithms, particularly Deep Learning (DL) due to their capability to extract features from data automatically. The workflow often consists of:

(i) preparing a labeled dataset with WiFi signal samples (e.g., Channel State Information (CSI), Received Signal Strength Indicator (RSSI)) for the targeted sensing tasks and (ii) training a model using the labeled data to learn the mapping function between the task and the measurements. The training can be performed directly on the raw CSI samples [10], [11] or on the features extracted from them [12], [13].

Motivation. The reliance on neural networks in WiFi sensing introduces two primary challenges. First, data collection and feature selection are inherently difficult due to the dynamic nature of WiFi environments and hardware limitations. Factors like multipath effects, interference, and changes in signal characteristics [13] can make it challenging to collect a dataset that accurately represents the real-world scenarios the models will encounter. Furthermore, the selection of meaningful features from high-dimensional CSI data is non-trivial, as it often requires significant preprocessing and domain expertise. These issues can impact the level of domain independency. Second, despite deep learning's success in recognizing complex patterns, its black-box nature limits transparency, which is critical for sensitive applications such as elderly care and security monitoring. This lack of explainability can lead to mistrust, especially when models are used to make high-stakes decisions. Therefore, understanding the decision-making logic of deep learning models in WiFi sensing is crucial not only for validation but also for gaining insights that could enhance algorithmic design and efficiency. In addition, considering the threat of adversarial machine learning, it is crucial to understanding model logic and feature selection [14]. Furthermore, considering the significant number of studies that utilize deep learning models, particularly Convolutional Neural Network (CNN), for automatic pattern extraction, it is crucial to validate these models by examining the logic behind their feature selection. While existing studies have attempted to address the first challenge through data augmentation [15], few-shot learning [16], and environment-independent feature extraction [1], these approaches predominantly target data quality and generalization without addressing model explainability. It is therefore essential to go beyond these traditional methods and incorporate XAI tools that offer deeper insights into the model's internal logic.

Challenge. To the best of our knowledge, there is no prior work aiming at explaining blackbox models in WiFi gesture/activity recognition. eXplainable Artificial XAI has been a subject of interest in the ML community with extensive application in computer vision [17] and natural language processing [18]. However, applying these methods to WiFi

M. Shahbazifar, Z. Zeinalpour-Yazdi and M. Zoofaghari are with Department of Electrical Engineering, Yazd University, Yazd, Iran (e-mail: minashahbazifar, zeinalpour, zoofaghari@yazd.ac.ir). M. Shahbazifar is also with the Resilient Communication Systems Group, Technical University of Darmstadt, 64283 Darmstadt, Germany (e-mail: m.shahbazifar@rscs.tu-darmstadt.de) and M. Zoofaghari is also with the Intervention Center, Oslo University Hospital, Oslo, Norway (e-mail: mohzoo@ous-hf.no).

Zeinalpour-Yazdi's work has been supported in part by Iran National Science Foundation (INSF) under Grant number 4025264.

D. Schumacher and M. Hollick are with the Secure Mobile Networking Lab, Technical University of Darmstadt, 64289 Darmstadt, Germany (e-mail: dschumacher, mhollick@seemoo.tu-darmstadt.de). M. Hollick is also with the IMDEA Networks Institute, Madrid, Spain (e-mail: matthias.hollick@networks.imdea.org).

A. Asadi is with the Faculty of Electrical Engineering, Mathematics and Computer Science (EEMCS), Delft University of Technology, 2628 CD Delft, The Netherlands (e-mail: A.Asadi@tudelft.nl).

sensing is challenging, as the unique nature of WiFi data and its hardware limitations differ significantly from other domains, often requiring domain-specific adaptations. Moreover, XAI methods vary widely in data compatibility, computational demands, and explanatory power. To address these challenges, our selection of XAI methods considers both domain-specific needs and the variability of XAI capabilities:

- **Domain Adaptability:** The chosen methods must be adaptable to the WiFi sensing domain, which involves complex and noisy CSI data while preserving both accuracy and interpretability.
- **Computational Efficiency:** Given the high-dimensionality of CSI data and the need for domain-specific adaptations, the methods should be scalable and computationally efficient for large-scale datasets.
- **Breadth of Explainability:** Effective XAI should provide both local and global explanations, offering insights into individual predictions as well as the overall model behavior across temporal and frequency domains.
- **Flexibility in Access Requirements:** The chosen techniques should support both model-specific (white-box) and model-agnostic (black-box) approaches, enabling reliable explanations under varying levels of model accessibility.

eXSense. To address these challenges, we propose eXSense, an Explainable AI-based workflow that seeks to bridge the gap between high-performing deep learning models and the need for transparency in WiFi sensing by transferring XAI to this domain. eXSense is designed to analyze black-box models at both a local and global level, providing detailed insights into how and why decisions are made. By doing so, we aim to highlight critical patterns within the data that can drive both validation and future algorithmic enhancements. In this paper, we validate eXSense through two representative case studies, demonstrating how explainability can enhance model efficiency and reliability without significant accuracy loss. The following summarizes our contributions and results:

(i) eXSense is the first attempt to shed light inside the black-box ML models in the area of wireless gesture recognition. We identify and adapt two classes of XAI methods: (a) Integrated Gradients (IG) [19] which is an attribution-based method suitable for whitebox testing, and (b) Counterfactual Instance (CF) [20] which is a counterfactual explanation method suited for blackbox testing. The eXSense workflow is depicted in Fig. 1. (ii) We extensively analyze two representative case studies using CSI samples (ARIL [10] and SignFi [11]). eXSense reveals that the models mainly rely on a small subset of the features, about 20%, despite their large feature space, i.e., roughly 10^4 for ARIL and 3.6×10^4 for SignFi. We validate eXSense by retraining the model using these specific features to examine the accuracy deviation. We observe no accuracy loss in SignFi while reducing 93% of features and only 0.82% loss in accuracy for ARIL with 80% feature reduction. These results confirm the reliability of eXSense. *We note that the majority of existing fML-based literature use the same feature structure: (subcarriers $\times$ RX antennas $\times$ TX antennas $\times$ packets) as our case studies [21].* (iii) Leveraging these insights,

we delve into physical analysis of the selected features to interpret the model's logic and answer the question "Why do certain features participate in the model's prediction?". Since CSI, unlike inherently interpretable data such as images and text, requires an additional stage to derive meaningful insights from XAI results, we employ feature selection and extraction techniques including point correlation, Doppler spectrum, and variance analysis across frequency and time domains. Notably, point correlation enables us to bridge the spatial shift-invariance property of standard CNNs with the physical propagation and system distortion effects of wireless signals. These metrics enable us to pinpoint informative features, distinguish relevant patterns from noise, and understand temporal signal changes influencing model behavior. Considering all results, we confirm the reliability of both ARIL and SignFi models, wherein the selected subcarriers exhibit the highest cross-packet correlation, the highest between-class and lowest within-class variance, and the highest Doppler-based speed. Beyond reliability checking, we discover that the Doppler spectrum is highly sensitive to hardware noise levels, and that the between-class to within-class variance ratio serves as the most effective feature selection metric, where a ratio below one indicates noise.

The rest of the paper is organized as follows. Related work is reviewed in Section II. In Sections III and IV we explain the background of WiFi sensing and XAI, respectively, which form the basis of our study. Section V presents the building blocks of eXSense. The experimental setup is detailed in Section VI. The results and performance of the proposed approach are evaluated in Section VII. Discussions and Conclusions are provided in Sections VIII and IX respectively.

II. RELATED WORK

The use of XAI in wireless sensing goes back to 2018 [22] where the authors visualized the model by converting the Photoplethysmography signals into pictures and using class activation map (CAM) to show the pixels' importance. Since then, some other articles have explored the use of XAI algorithms for device-based wireless sensing [23], [24]. While most of these articles use XAI for trustworthiness, the authors in [24] used the Shapley additive explanations (SHAP) algorithm, during the training phase for dimension reduction by removing the less important features. The authors in [25] used the XAI algorithms for layman's explanations on the end-user side to understand the model decisions in smart homes. In the field of device-free wireless sensing, the authors in [26] proposed PRF-PIR, passive, multi-modal sensing and used SHAP to show that their proposed fusion framework works better than software-defined radio (SDR) and passive infrared (PIR) sensor systems separately, by showing that the attributions in the combined use of sensors are higher than their individual use.

While the above-mentioned articles used secondary algorithms/techniques for obtaining explanations, some articles train their models in a way to be interpretable by humans [27], [28]. For example, in [29] and [30], hierarchical models for RF sensing and mm RADAR are introduced. The authors

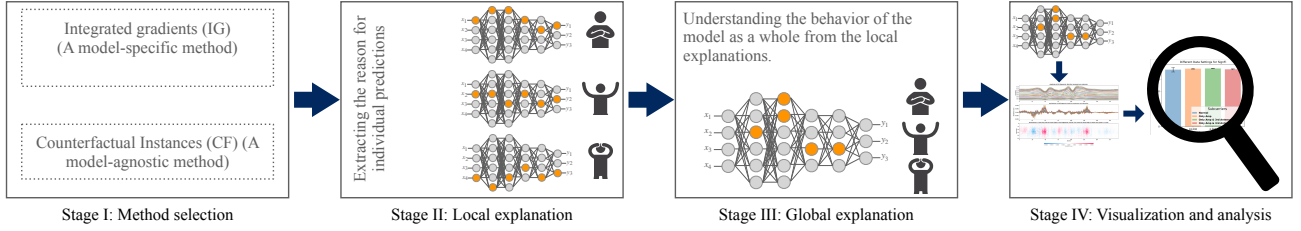


Fig. 1: Overview of eXSense workflow: (i) choosing the suitable explanation methods based on the access to the model, (ii) computing local explanations for individual predictions (e.g., a hand gesture), (iii) obtaining global behavior of the model for all predictions (e.g., all available gestures), and (iv) visualizing and analyzing the explanations.

of [31], proposed re-identification model based on 2D Angle of Arrival (AoA) of reflected WiFi signal from the body. They make their model interpretable by transforming 2D AoA into 3D digitizing picture based on the walking and body shape.

Unlike prior work, that uses the XAI outputs directly, there is no effort in analyzing and post-processing the XAI results to explain the blackbox nature of WiFi sensing models. To the best of our knowledge, this is the first attempt at using gradient-based XAI methods for the insensible data, CSI. Transferring to a new domain, presents unique challenges, including: (i) the usage of an adequate baseline as well as the confirmation of the results, and (ii) suitable means of interpreting CSI, which lacks inherent human interpretability compared to images. It is worth mentioning that our visualization strategy is not aimed at creating new techniques; rather, it adapts established visualization principles to meet the specific needs of WiFi sensing. Unlike traditional time-series visualizations like TSViz [32], LSTMVis [33], and Seq2Seq-Vi [34] which handle single-dimensional signals, our approach addresses the multi-dimensional nature of CSI data. This adaptation is crucial for visualizing and interpreting features unique to WiFi environments, such as amplitude and phase characteristics that are essential for accurate analysis of WiFi sensing models.

III. BACKGROUND: WIFI SENSING

WiFi sensing analyzes the impact of environments on radio waves. Several factors can affect the received signal, including multipath propagation and the movement of the transmitter/receiver or the objects in the vicinity of the communication path. The effect of these factors is generally manifested in the CSI, which is in essence the channel frequency response; Considering an OFDM signal with the main frequency f_c and the sub-channel spacing Δf , the channel frequency response at the k th subcarrier and n th received packet for a pair of transmitter and receiver is [13]:

$$H(k, n) = \sum_{p=1}^P a_p(n) e^{-j2\pi(f_c + k\Delta f - f_d(n))\tau_p(n)}, \quad (1)$$

where P is the total number of multipath components, $a_p(n)$ is the complex attenuation factor, $\tau_p(n)$ is the propagation delay for the p -th path, and $f_d(n)$ is the doppler frequency shift. However, due to undesired hardware artifacts, the estimated CSI is [35]:

$$H'(k, n) = H(k, n) e^{j\theta_{\text{off}}(k, n)} + W(k, n), \quad (2)$$

where $\theta_{\text{off}}(k, n)$ denotes the phase offset introduced by hardware imperfections such as carrier frequency offset and synchronization errors (i.e., Cyclic Shift Diversity (CSD), Sampling Frequency Offset (SFO)). This offset is modeled as $\theta_{\text{off}}(k, n) = \theta_C(n) + k\theta_L(n)$, where $\theta_C(n) \sim \mathcal{N}(0, \sigma_C^2)$ is the common part and $\theta_L(k, n) \sim \mathcal{N}(0, \sigma_L^2)$ linear part, across the subcarriers. $W(k, n) \sim \mathcal{CN}(0, \sigma_w^2)$ represents Gaussian noise causes by Inter-Carrier Interference (ICI), Inter-Symbol Interference (ISI) [35] and Co-Channel Interference (CCI) [36]. Very often signal processing algorithms are applied to the CSI to mitigate the effect of unwanted components, thus enhancing the *quality* of the CSI measurements.

Typically, if sensing is focused on recognition tasks, next processing steps include machine learning techniques, particularly neural networks, to classify the targeted tasks. While many proposals just use the raw CSI amplitude and phase [21], there are works which extract other parameters by post-processing CSI, e.g., AoA, Doppler shift [12]. These parameters are particularly helpful in achieving domain independence. In general, different events that happen in the sensed environment have different impact on the CSI, be it on the amplitude or the phase. The CSI can be processed through various means, e.g. DL algorithms, to solve detection, recognition and estimation tasks.

IV. EXPLAINABLE ARTIFICIAL INTELLIGENCE

As the applications of deep learning models increase, so does the necessity of understanding/explaining their behavior for scientific and legal purposes. Providing such insights helps scientists/practitioners to verify the decision logic of blackbox models and to learn from them by finding constellations in the data that were not previously apparent. Traditionally, the latter is performed using convolution layers but *the extracted features remain opaque to the user*. In contrast, XAI techniques project the model's decision back onto the input space, thus revealing the important input features. *To date, there is no one-fits-all solution for obtaining explainability*. In the following, we provide an overview of XAI techniques.

A. Explainability techniques

Explainability techniques can be divided into three classes [37]: (i) *attribution-based methods*, (ii) *surrogate models*, and (iii) *counterfactual explanations*.

Attribution-based methods are intended for estimating the contributions of the input features for individual predictions of the model [17], [19], [38], [39]. This is a non-trivial task since features may depend on/influence each other [40]. Many attribution methods have roots in Shapley values [41], whose exact computations is an NP-hard problem [42]. Consequently, the state-of-the-art attribution methods rely on gradients to extract the feature importance [17], [19], [38]. Attribution-based methods are easy to understand and visualize for humans, as they assign different importance levels to distinct features in the decision making process.

Surrogates are interpretable models (e.g., decision tree) trained to mimic the behavior of another (blackbox) model. The new interpretable model is then used to explain the blackbox model. Surrogates are model agnostic but they suffer from two major drawbacks. First, the blackbox models are often too complex to have an interpretable alternative. Second, the surrogates are just an approximation of the blackbox model. Thus, they introduce additional errors to the predictions, reducing the reliability of their explanations.

Counterfactual explanations do not try to explain the model's inner workings to the end-user but instead give an example on how the outcome of a particular prediction can be changed without opening the blackbox. Counterfactual instances describe the smallest change to the feature values that changes the prediction [43]. Coming from social sciences, counterfactuals describe casual situations in human understandable terms in the form: 'If x had not occurred, y would not have occurred' [20].

B. Global and local explainability

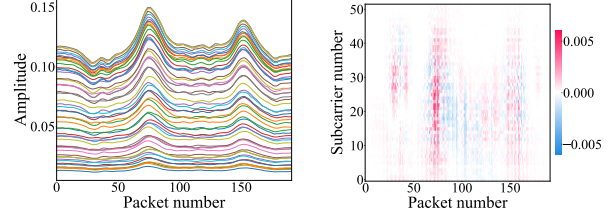
Global explainability provides a general understanding of the model based on the features, learned parameters and the structures. Global explanations encompass the whole behavior of the model, independent of a specific prediction. In contrast, *local explainability* provides insights into individual predictions of the model. Hence, it does not describe the model's overall behavior, but explains some local areas. From our point of view, both global and local explanation are necessary for understanding the behavior of the blackbox models. Local explainability is key to identify the root cause of misclassifications and anomalies, whereas global explainability is important to gain a general understanding of the model's behavior. We chose to prioritize local explanations because they provide detailed insights into individual predictions, allowing us to identify potential anomalies, biases, or unexpected patterns at the instance level. This forms a reliable basis for constructing a more accurate and comprehensive global explanation.

V. EXSENSE

Next, we elaborate on eXSense, an XAI workflow designed to obtain insights into blackbox ML-based WiFi sensing solutions.

A. Stage I: Selection of explanation method

Among the XAI approaches mentioned in Section IV, we choose Integrated Gradients, which is an efficient attribution-



(a) CSI amplitude (b) Local explanation
Fig. 2: The result of local explanations for action *circle*.

based method; and counterfactual instances, which is a model-agnostic explainability method. We refrained from using surrogates as they are not faithful to the original model and introduce additional approximation errors which are hard to identify and quantify. Moreover, we choose Integrated Gradients as model-specific attribution method, since it is one of the only two attribution methods that fulfill all of the five axioms presented in [19], ensuring high-quality explanations. The other method fulfilling the axioms, i.e., χ -Gradient [17], is only applicable to a specific type of DL, hence it is not considered here. Furthermore, we choose Counterfactual Instances as a model-agnostic technique, since they are easy to understand and applicable on all kind of models.

Another important consideration when selecting an explainability method is the required level of access to the model and the specificity of the method to the structure of the model. Model-specific techniques require access to the model, as they utilize specific properties (e.g., gradients, weights). Model-agnostic techniques, on the other hand, work on any kind of model, as they only need to access the input and the output (classification) of the model [44].

B. Stage II-a: Integrated Gradients(model-specific method)

IG [19] calculates the importance of each feature for individual classifications of the model. Let us consider a WiFi sensing approach that uses a DL model to classify activities using CSI amplitudes such as one of our case studies, ARIL [10]. In Fig. 2 we can see how the feature importance is picked up by eXSense at this stage for the action *circle*, where the user moves her hand in a circle. Given the model, an input and the class of that input, e.g., *circle*, IG computes the contribution of each amplitude value for the class *circle*. Furthermore, the attributions can be positive or negative, meaning that they correlate positively or negatively to the class. We can observe this in Fig. 2 where the CSI pattern around packet 75 positively contributes to class *circle*, whereas the fluctuating CSI patterns shortly afterwards contribute negatively to the classification, that is, removing such a pattern increases the probability of the sample being classified as *circle*.

Formal Definition: Let $F : \mathbb{R}^n \rightarrow \mathbb{R}$ be the function that represents a Deep Neural Network (DNN), $x \in \mathbb{R}^n$ be the input and $x' \in \mathbb{R}^n$ be the baseline input. Considering a straight line path from x' to x (in $\mathbb{R}^n$) and computing the gradients at all points along the path, the integrated gradients is acquired by cumulating these gradients. Specifically, integrated gradients are the path integral of the gradients along the straight line path from the baseline x' to the input x .

The integrated gradient along the i th dimension is given as:

$$IG_i(x) ::= (x_i - x'_i) \int_{\alpha=0}^1 \frac{\partial F(x' + \alpha \times (x - x'))}{\partial x_i} d\alpha, \quad (3)$$

with $\frac{\partial F(x)}{\partial x_i}$ being the gradient of $F(x)$ along the i th dimension [19]. For efficient computation, IG can be approximated as follows:

$$IG_i^{approx}(x) ::= (x_i - x'_i) \frac{1}{m} \sum_{k=1}^m \frac{\partial F(x' + \frac{k}{m} \times (x - x'))}{\partial x_i}, \quad (4)$$

with m being the number of steps in the Riemann approximation, which in practice lies between 20 and 300 [19]. The attributions are calculated for individual classifications of the model and have the same shape as the input, as each feature of the input is assigned an attribution value. This value is a real number and has two essential properties: the sign and the magnitude. The higher the magnitude of the attribution value, the higher the importance of the feature. As seen in Fig. 2, the sign indicates whether the feature contributes positively or negatively to the predicted class.

The calculation of the attribution values starts with a baseline, which is a neutral point in the input space, such that it contains no information for the model classification. With the baseline as starting point in the input space, IG takes a defined number of steps from the baseline to the input. In each step, IG calculates the gradients of the model to measure how the change of a feature impacts the model's classification.

What is a good baseline for WiFi sensing? An important input to the integrated gradient method is the baseline, which by definition should carry *zero information*. Defining this zero information baseline in some applications is more trivial such as a black image for computer vision and zero embedding vectors for natural language processing. However, zero CSI values for WiFi sensing are not reasonable. In WiFi sensing, the information is the variation of CSI due to changes in the environment, be it location or movement of an object. However, the CSI values are not zero in the absence of such changes due to channel noise and hardware impairments. In contrary to other applications, we propose to create the baseline from averaging the data of all data samples for each subcarrier, resulting in a static value for each subcarrier.

Theoretical Proof. Let's separate the static component ($H_S(k)$) (including the static backscatters and the LoS signal) and the dynamic ($H_D(k, n)$) part of the channel in Eq. 2

Considering that the amplitude and phase of the $H'(k, n)$ used as distinct input, let us consider averaging separately for each of them. Let $b_A(k)$ denote the baseline value for the amplitude in the k -th subcarrier:

$$\begin{aligned} b_A(k) &= \frac{1}{N} \sum_{n=1}^N |H'(k, n)| \\ &\approx \mathbb{E}[|(H_S(k) + H_D(k, n)) e^{j\theta_{\text{off}}(k, n)} + W(k, n)|] \\ &= \mathbb{E}[|(H_S(k) + H_D(k, n) + \tilde{W}(k, n))|], \end{aligned} \quad (5)$$

where $\tilde{W}(k, n) \sim \mathcal{CN}(0, \sigma_w^2)$ denotes the phase-rotated version of $W(k, n)$. Applying Jensen's inequality and the triangle inequality yields:

$$|H_S(k) + \mathbb{E}[H_D(k, n)]| \leq \mathbb{E}[|H'(k, n)|], \quad (6)$$

$$\mathbb{E}[|H'(k, n)|] \leq |H_S(k)| + \mathbb{E}[|H_D(k, n)|] + \mathbb{E}[|\tilde{W}(k, n)|].$$

Assuming the uncorrelated scattering channel with a uniformly distributed phases, the dynamic component can be modeled as a sum of independent complex random variables. By the Central Limit Theorem, the resulting channel becomes a circularly symmetric complex Gaussian random variable ($H_D(k, n) \sim \mathcal{CN}(0, \sigma_D^2)$) and consequently, the amplitude of the dynamic component follows a Rayleigh distribution ($|H_D(k, n)| \sim \text{Rayleigh}(\sigma_D)$). Consequently Eqs. 6 simplifies to:

$$|H_S(k)| \leq \mathbb{E}[|H'(k, n)|] \leq |H_S(k)| + \sigma_D \sqrt{\pi/2} + \sigma_W \sqrt{\pi/2}. \quad (7)$$

As the averaging interval increases, the contributions of the zero-mean stochastic dynamic component and noise diminish due to destructive interference, leading to $\mathbb{E}[|H'(k, n)|] \rightarrow |H_S(k)|$. In a similar way, let $b_P(k)$ denote the baseline value for the phase in the k -th subcarrier:

$$\begin{aligned} b_\theta(k) &= \frac{1}{N} \sum_{n=1}^N \angle(H'(k, n)) \\ &= \mathbb{E}[\angle(H_S(k) + H_D(k, n) + \tilde{W}(k, n))] + \mathbb{E}[\theta_{\text{off}}(k, n)]. \end{aligned} \quad (8)$$

Since the static component $H_S(k)$ contains the dominant LoS path, its magnitude is much larger than those of the dynamic and noise contributions. Consequently $\angle(H_S + H_D + \tilde{W}) \approx \angle(H_S)$, yielding to $b_\theta(k) \approx \angle(H_S(k))$. Therefore, baseline averaging largely captures the static channel component, approximating the empty-room response and serving as a practical baseline.

Validation test. To test our proposal, we tested multiple baselines on the ARIL model with our own measurements collected via two ASUS RT-AC86U routers which run nexmon CSI extractor [45]. The results can be seen in Fig. 3 where we compare the attribution of IG using the *zero baseline* and our proposed baseline for eXSense with the attributions computed by EG [38] which is a baseline-free XAI solution. We can observe that eXSense's baseline provides the closest results to EG, thus confirming our hypothesis that zero baseline is a poor choice.

Ablation Evaluation. To evaluate the reliability of the

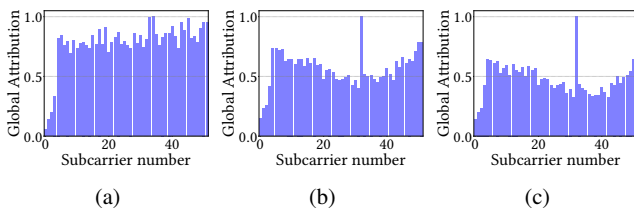


Fig. 3: The figure illustrates the estimated attributions by: (a) IG using zero baseline, (b) IG using eXSense baseline, (c) Expected Gradients (EG), a baseline-free method.

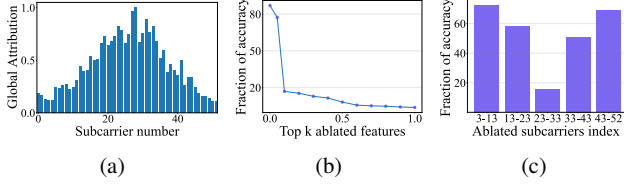


Fig. 4: The ablation test for evaluating the reliability of the averaging baseline. (a) the IG results of subcarrier importance, (b) Top K Ablation test, (c) Ablation test. In (a) subcarriers are gradually ablated to see accuracy fraction slope, while in (b), subcarriers are divided into five categories to evaluate the most accuracy fraction upon ablation.

averaging baseline, we conducted ablation studies using IG on the ARIL dataset. First, a Top-K ablation test [46] was performed by progressively removing the K most important features ranked by IG, where a sharper decline in accuracy indicates stronger attribution quality, as shown in Fig. 4b. The results demonstrate a rapid performance drop, confirming the effectiveness of the proposed baseline. In addition, we performed a group-wise ablation by partitioning subcarriers into five ranked groups. As illustrated in Fig. 4c removing higher-ranked groups leads to a significantly larger degradation in accuracy, further validating the consistency of the IG-based attribution ranking.

C. Stage II-b: Counterfactual Instances (model-agnostic method)

Let us describe the CF with an intuitive example in the context of WiFi sensing such as the recognition of waving a hand. To generate a CF, we simply tweak the feature values (e.g. CSI amplitude) of an instance and then we observe how the forecast changes. For example, if our CF analysis shows that changing the CSI amplitude of subcarriers 10 to 15 results in misclassification of the correct activity (e.g., waving), this indicates that perhaps those subcarriers have higher importance to the model's classification. The actual implementation of CF is more elaborate. Essentially, instead of randomly changing the input feature, we define a loss function based on the distances between the instances (x, x') and classes (y, y') . As shown in Fig. 5 the goal is minimizing these losses. There are several algorithms for this purpose amongst which we opt for the original method in [20] and its open-source implementation in [47].

Formal definition. Consider model $f(x) = y$, where an instance x produces class y . An ideal CF x' should lead to a different class y' while maintaining the difference between x and x' as small as possible. Therefore, the search for x' is an optimization problem with the following general loss:

$$L(x'|x) = (f_t(x') - p_t)^2 + \lambda L_1(x', x), \quad (9)$$

where t is the desired target class of x' , p_t the target probability of that class, f_t the model prediction/probability on class t , and L_1 is the distance function measuring the distance between x and x' [47]. The first term of the loss $(f_t(x') - p_t)^2$ ensures that the predicted class of x' is the target class t . As

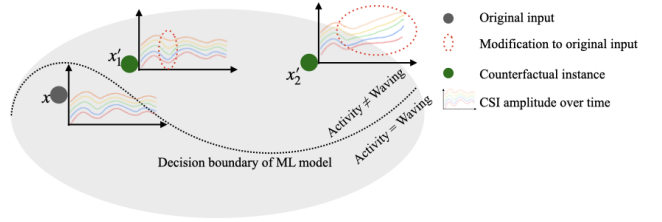


Fig. 5: An illustrative example of CF explanations. The goal of the algorithm is to find instances that are close to the decision boundary. This can be achieved by minimizing the loss function, Eq (9).

it is the main goal of the search, it is weighted quadratically to penalize deviations. Moreover, the L_1 distance function is chosen such that x' is a sparse CF, i.e., the number of altered features is minimized. As the optimal λ is hard to find and can vary from instance to instance, it is learned as part of the optimization algorithm:

$$\min_{x'} \max_{\lambda} L(x'|x), \quad (10)$$

subject to

$$|f_t(x') - p_t| \leq \epsilon \text{ (CF constraint)}, \quad (11)$$

where ϵ is the tolerance parameter. In practice, λ is found in two steps: First, a broad range, e.g. $\lambda \in \{10^{-1}, \dots, 10^{-10}\}$ is examined to find lower and upper bounds $\lambda_{lb}, \lambda_{ub}$ where CFs exist. Second, the maximum $\lambda \in \{\lambda_{lb}, \lambda_{ub}\}$ is found using bisection subject to the CF constraint. This provides us with a set of CFs with varying distance to x . From this set, we use the CF with the lowest distance to x , as it requires the smallest changes from the original instance/input.

D. Stage III: Global explainability

In Stage II, we only obtained local explanation, which essentially explains the reason behind classification of individual instances, e.g., why a specific measurement was recognized as waving. As mentioned, local explanations are important to verify the model and understand the root cause of anomalies. But we still need the complementary view that global explanations provide. Since the mechanics of IG and CF differ, separate techniques are employed to derive global explanations from their respective local interpretations.

1) *IG*: To derive a global explanation, we aggregate local attributions by computing the mean of their absolute values across the entire dataset [40]. Mathematically, this global attribution map, $\mathbf{G} \in \mathbb{R}^{I \times J}$, is defined as the empirical approximation of the expected attribution magnitude:

$$G_{i,j} = \frac{1}{N} \sum_{k=1}^N |\text{attrib}_{i,j}^k| \approx \mathbb{E}_x [|IG_{i,j}(x)|]. \quad (12)$$

Taking the absolute value of attributions prevents sign cancellation during aggregation, as both positive and negative contributions indicate feature importance. Furthermore, the

proposed global explanation is grounded in a principled statistical and physical interpretation rather than heuristic averaging. To justify this interpretation, we first consider the nature of CSI data. As shown in Eqs. 1 and 2 the clean measurements across subcarriers and antennas share the same activity-specific fingerprint, including path attenuation $a(n)$, Doppler frequency shift $f_d(n)$, and path time delay $\tau(n)$. However, due to multipath fading, system distortions, and hardware imperfections, the quality of the extracted amplitude and phase varies across subcarriers and antennas. Accordingly, the input data can be modeled as $x = x^c + \eta$, where x^c denotes the clean data and η represents additive noise. Assuming that the model is robust and trustworthy, the prediction remains essentially unchanged under small perturbations, i.e., $F(x) \approx F(x^c)$. Consequently, based on Eq. 4 the Integrated Gradients can be approximated as

$$IG_{i,j}(x) \approx IG_{i,j}(x^c) + \epsilon_{i,j}, \quad (13)$$

where $\epsilon_{i,j}$ denotes a perturbation term induced by noise. By averaging over all samples, the effect of this noise diminishes, yielding

$$\mathbb{E}_x[|IG_{i,j}(x)|] \approx \mathbb{E}_x[|IG_{i,j}(x^c)|]. \quad (14)$$

This holds because either (i) the noise η follows a zero-mean Gaussian distribution, or (ii) any nonzero-mean bias is captured by the baseline x' (the dataset mean), thereby neutralizing systematic bias.

2) *CF*: CF computes the feature importance implicitly as we seek for the minimal change that can alter the classification. Consequently, the modified features have a significant impact on the classification. Therefore, we utilize the differences between the original and the CF to highlight the modified features and the magnitude of their modification. A limitation of CF is that the quality of the discovered solutions may vary since it is a limited search and it might not always find a high-quality solution for all samples. With a high-quality solution, we refer to CFs that have a small distance to the original sample, i.e., a smaller number of features have been altered and hence they have a higher contribution to the classification. We obtain these high quality solutions through a so-called filtering operation. Considering the set $\mathbb{X}$ which contains the CF with the lowest distance, for each sample of the dataset we further create the set of filtered CFs:

$$CF_f = \{x | x \in \mathbb{X} \wedge d_x < d_{th}\}, \quad (15)$$

where d_x returns the distance between the original- and counterfactual instance x , and d_{th} is the distance threshold used to filter them. Next we move from local to global explainability by leveraging how often a feature is modified across all counterfactual instances, as shown in Algorithm 1. The input of the algorithm is the dataset *Data* as well as the set of filtered counterfactuals CF_f Eq. (15), whereby both have the same number of samples. Further it initializes the counter matrix *globalCounter*, which has the same shape as the data samples. Then, it computes the absolute difference of each pair in *Data* and CF_f . Inspired by [48], for each entry in the *diff* matrix greater than t_{diff} , it increases the counter at the corresponding position in *globalCounter*. Unlike d_{th} ,

Algorithm 1: Counter method

Input: Dataset *Data*, set of filtered counterfactuals CF_f ; Threshold

t_{diff}
Output: Matrix with same shape as data/counterfactual
globalCounter

```

1  $k_{max} \leftarrow |Data| = |CF_f|, k \leftarrow 0;$ 
2  $globalCounter \leftarrow \{0\}^{I \times J};$ 
3 for  $k < k_{max}$  do
4    $diff \leftarrow |Data^k - CF_f^k|;$ 
5   for  $i < I$  do
6     for  $j < J$  do
7       if  $diff_{i,j} > t_{diff}$  then
8          $globalCounter_{i,j} \leftarrow$ 
9            $globalCounter_{i,j} + 1;$ 
10        end
11      end
12    end
13  end

```

which operates at the sample level, t_{diff} acts at the feature level by filtering out small deviations in CFs (e.g., those introduced by data-type conversion in the *alibi* framework). Smaller d_{th} and larger t_{diff} enforce stricter filtering and may suppress meaningful variations, whereas the opposite settings increase sensitivity to minor changes, leading to a larger number of features being identified as significant. The selection of d_{th} and t_{diff} is inherently data-dependent and requires empirically calibration based on dataset characteristics, such as scale and statistical properties. For example, we set $t_{diff} = 0.0024$ and $d_{th} = 11$ for the ARIL dataset (9,600 features), whereas for the SignFi dataset (36,000 features), $t_{diff} = 0.4$ and $d_{th} = 1800$ is used.

E. Stage IV: Visualization and analysis

We analyze the results from both communication and statistical points of view. We leverage point correlation and Doppler spectrum from the communication point of view and variance from the statistical point of view. As the most natural way of finding interpretable explanations for humans is through visualization, the final step consists of finding suitable methods to represent the results. Unlike the previous stages which are rather generic, this stage is highly dependent on the structure of the model and the parameters. In Section VII we provide a customized visualization and analysis framework specifically tailored for WiFi data, focusing on multi-dimensional signal characteristics that are unique to WiFi signals.

1) *Correlation*: The system's CSI correlation can change in the time and frequency domain due to several factors. These factors include two main aspects: (i) channel conditions and environmental factors, and (ii) system design parameters and channel estimation offsets. Here's a detailed explanation of how the first aspect influences the correlation in the frequency domain.

- **Multipath Fading**: The presence of multipath fading causes frequency-selective fading, leading to different fading effects across subcarriers. Subcarriers within the channel's coherence bandwidth experience similar fading and higher correlation, while those outside have lower correlation [49].

• **Propagation Environment:** The physical environment, including obstacles, buildings, and terrain, affects the multipath profile and the correlation among channel estimations. Complex environments with many reflectors may increase correlation due to similar multipath components affecting adjacent subcarriers [50]. We delve into the impact of the second aspect on CSI analytically in the following. Consider the CSI's correlation specifically through point correlation as follows:

$$\hat{\mathbf{R}}_{HH} = \frac{1}{N} \mathbf{H} \mathbf{H}^H, \quad (16)$$

where $\mathbf{H}$ is a matrix of CSI with the dimension *subcarriers* $\times$ *packets* ($K \times N$) for the frequency domain. Considering phase offsets and system imperfections as discussed in Eq. 2, we can rewrite Eq. 16 as follows:

$$\hat{\mathbf{R}}_{H'H'} = \frac{1}{N} \sum_{n=1}^N \mathbf{h}'_n \mathbf{h}'_n{}^H, \quad (17)$$

in which

$$\begin{aligned} \mathbf{H}' &= \overbrace{\mathbf{H} \odot \boldsymbol{\Theta}_{\text{off}}}^{\tilde{\mathbf{H}}} + \mathbf{W}, \\ \mathbf{h}'_n &= [H'(1, n) \cdots H'(K, n)]^T, \\ \mathbf{h}_n &= [H(1, n) \cdots H(K, n)]^T, \\ \boldsymbol{\theta}_n &= [e^{j\theta_{\text{off}}(1, n)} \cdots e^{j\theta_{\text{off}}(K, n)}]^T, \\ \mathbf{w}_n &= [W(1, n) \cdots W(K, n)]^T. \end{aligned} \quad (18)$$

where $\odot$ is Hadamard Product. As discussed in [51], Eq. 17 is equal to:

$$\hat{\mathbf{R}}_{H'H'} = E\{\mathbf{H}'\mathbf{H}'^H\} = \underbrace{\mathbf{R}_{\tilde{\mathbf{H}}\tilde{\mathbf{H}}}}_{\textcircled{1}} + \underbrace{\sigma_W^2 \mathbf{I}}_{\textcircled{2}}. \quad (19)$$

As more noise affects the estimated channel, such as measuring offsets (i.e., ICI, ISI) or co-channel interference (i.e., CCI) expression ② masks expression ①, and the correlation decreases [36]. Consider the (i, k) -th entry of the correlation matrix $R_{\tilde{\mathbf{H}}\tilde{\mathbf{H}}}$:

$$R_{\tilde{\mathbf{H}}\tilde{\mathbf{H}}}(i, k) = \mathbb{E}[H(i, n)H^*(k, n)e^{j(i-k)\theta_L(n)}], \quad (20)$$

where $\theta_L(n)$ is the linear part of phase offset to the subcarrier as discussed in Section III. Without loss of generality, under the assumption of uncorrelated scattering (US) channel, wherein the multipath components are statistically independent, (i.e., $a_p(n) \perp a_q(n)$ and $a_p(n) \perp \tau_q(n)$ for $p \neq q$), Eq. 20 simplifies to:

$$\begin{aligned} R_{H'H'}(i, k) &= \mathbb{E}\left[\sum_{p=1}^P |a_p(n)|^2 \cdot e^{-j(i-k)\Delta f \tau_p(n)} \cdot e^{-j(i-k)\theta_L(n)}\right] \\ &= \sum_{p=1}^P (\mathbb{E}[|a_p(n)|^2] \cdot \mathbb{E}[e^{-j(i-k)\Delta f \tau_p(n)}] \cdot \mathbb{E}[e^{-j(i-k)\theta_L(n)}]) \\ &= \left(\sum_{p=1}^P \underbrace{\mathbb{E}[|a_p(n)|^2]}_{\text{Power Info}}\right) \cdot \underbrace{\Phi_{\tau_p}((i-k)\Delta f)}_{\text{multipath similarity}} \cdot \underbrace{e^{-\frac{1}{2}(i-k)^2 \sigma_\theta^2}}_{\text{phase offset effect}}, \end{aligned} \quad (21)$$

where Φ_{τ_p} is the characteristic function of the $\tau_p(n)$ distribution, quantifying the similarity of multipath contributions between subcarriers i and k . The Gaussian term models correlation decay due to hardware-induced phase offsets. Consequently, subcarriers with high correlation (small $|i - k|$) and less noise exposure preserve more reliable channel information. As shown in Eqs. 19 and 21, impairments such as phase offsets, ICI, ISI, and CCI reduce inter-subcarrier correlation. Therefore, a standard CNN should focus on highly correlated subcarriers, where multipath components combine more coherently over time.

In the time domain, mobility introduces time variations in the channel. This affects the coherence time of the channel. If channel estimation is done over periods shorter than the coherence time, the estimates will be highly correlated. For longer periods, the correlation decreases as the channel changes more significantly. To compute the time correlation (the correlation among the packets), we use the transpose of $\mathbf{H}$ in Eq. 16 and average over K .

2) *Doppler:* Movement significantly impacts CSI, primarily through the Doppler effect. The Doppler effect introduces frequency shifts due to the relative motion between the transmitter and the receiver, leading to time-variant channel characteristics. To analyze the Doppler spectrum, we utilize a state-of-the-art method outlined in [13]. If we consider H_i as the i th sliding window of the CSI with dimensions $(K \times N)$, then the Doppler vector of size N_D , D_i , can be computed as follows:

$$D_i(u) = \sum_{k=0}^{K-1} |\mathcal{F}\{H_i\}(k, u)|^2, \quad (22)$$

where $-N_D/2 \leq u \leq N_D/2$ is the Fourier transform index and

$$\mathcal{F}\{H_i\}(k, u) = \sum_{n=0}^{N-1} H_i(k, n) w_H(n) e^{-j2\pi n u / N_D}, \quad (23)$$

in which w_H is the Hanning function.

3) *Variance:* Variance is a simple yet powerful metric for feature selection in machine learning. According to the statistical pattern classification, a good feature has the least within-class variance and the most between-class variance. This is because the data belonging to a class should have the most similarity while the data belonging to different classes should have the least similarity. The term within-class variance refers to the average of all class variances and is given as:

$$\sigma_W^2 = \frac{1}{N} \sum_{j=1}^N \frac{1}{n} \sum_{i=1}^n (x_{ij} - m_j)^2, \quad (24)$$

where N is the number of classes, n is the number of samples in each class, $x_{i,j}$ indicates the i th sample of j th class and m_j is the average of j th class. For the term between-class variance, we consider the whole data which is given as:

$$\sigma_B^2 = \frac{1}{N} \sum_{j=1}^N (m_j - m)^2. \quad (25)$$

VI. EXPERIMENTAL SETUP AND CASE STUDIES

In our evaluation, we use publicly available datasets. In the following, we elaborate on our experimental setup and choice of hyper-parameters and provide a short overview to our case studies.

A. Parameters used for XAI

In the description of eXSense in Section V, we introduced certain hyper-parameters whose exact value is reported here for reproducibility. The IG attributions are calculated with the python library *alibi*^[2]^[47] using $n_steps=100$, the number of approximation steps m in Eq. (4), and our proposed eXSense baseline. For CF, the hyperparameters are set as follows: max_iter is set to 150, and early stopping is applied with a patience of 50. We set $p_t = 0.9$ and initialize $\lambda = 0.1$, using 10 bisection steps. All other parameters follow the default settings provided in^[3]. Explanations are generated using a fixed random seed of 7, and both training and test samples are included. For model training, random seeds ranging from 1 to 100 (with a step size of 10) are used, and the best results for each experiment are reported. When not predefined, an 80:20 train-test split is applied. Cross-validation is used in scenarios involving domain shifts (e.g., different subjects). All classes are used for training unless otherwise specified.

B. Case studies

We have selected two representative case studies that differ in terms of features and methodology: (i) ARIL^[10], a gesture recognition approach based on the CSI amplitude and (ii) SignFi^[11], a sign language recognition utilizing multiple antennas as well as the CSI amplitude and phase.

1) *ARIL*: This approach performs gesture recognition based on the CSI amplitude. It can differentiate between six different gestures: *arm up*, *arm down*, *arm left*, *arm right*, *circle*, and *cross*. For each gesture, they perform a measurement over 192 packets and record the CSI amplitudes over 52 subcarriers. The result of the measurement is then fed into a 1D-ResNet. The choice of 1D-ResNet is to enable sliding over the packets in the time domain^[4].

2) *SignFi*: SignFi uses WiFi sensing to differentiate 276 different signs of American Sign Language (ASL). Their proposal brings about a significant improvement over prior works which were only capable of classifying 25 different gestures. Additionally, it works across environments for the same user. The approach utilizes the amplitude and the preprocessed phase of the CSI of 30 subcarriers and three transmit antennas. A sample, that gets later fed into the model, consists of 200 packets, resulting in a 3D-matrix with shape $200 \times 60 \times 3$, where 60 results from the concatenation of 30 amplitude and 30 phase values, extracted from the same 30 CSI values. For classification, it uses a 2D-CNN, which can further extract features on its own.

^[2]<https://github.com/SeldonIO/alibi>

^[47]<https://docs.seldon.io/projects/alibi/en/stable/methods/CF.html>

^[4]In addition to the six gestures, ARIL performs localization in 16 locations in a room. In this work, we only analyze their gesture recognition solution.

VII. UNDERSTANDING AND ENHANCING BLACKBOX MODELS

In this section, we apply eXSense on the two case studies. Specifically, for each case study, we analyze the local explanations, followed by the global explanation. For better comparison between model-specific and model-agnostic XAI, we show the explanations offered by IG and CF. Utilizing the insights from eXSense, we improve the efficiency and reliability of the case studies.

A. ARIL

1) *Local Explainability*: As a first step, eXSense will calculate IG attributions and counterfactual instances for individual predictions of the model, to achieve local explainability. Fig. 10 illustrates the local explanations obtained from the classification of ARIL for two gestures (arm up and arm right). The graphs on the top show the measured CSI amplitude over all 52 subcarriers whereas the accompanying heat map shows the IG attributions of the features across subcarriers and over different packets. This form of visualization clearly shows the dependency on the subcarriers and further facilitates comparing the attributions of different samples at different times/packets. We can observe in Fig. 6a that the highest attributions for the *arm up* gesture are mainly located at the first change of the amplitude, which occurs due to the movement of the arm (around Packets 50-75). For the *arm right* gesture in Fig. 6b, the main part of the attribution lies in the range from Packets 100-125, where we observe the highest variations in the CSI amplitude. In both cases, we observe that the model mostly relies on the information of the middle subcarriers (between 23-33) for classification. We elaborate on this choice in global explanations.

The local explanations provided by CF are visualized in Figs. 6c and 6d by showing the absolute difference between the normal and the counterfactual instance. Both heat-maps reveal that the CF altered features that are in areas with high positive attribution (compare IG). This complies with the findings of IG, as changing those features should be the optimum way for changing the prediction of the model for that class and sample.

2) *Global Explainability*: Fig. 7 illustrates the normalized global importance of each feature where the results for IG are obtained from Eq. (12). For CF's global explainability, we use the method proposed in Algorithm 1 which correlates the frequency of changes for each feature to its importance. Figs. on the left show the global importance as a heat-map for each subcarrier within each recorded packet over the measurement time. Figs. on the right accumulate the attributions of each subcarrier over all packets, thus highlighting the key subcarriers for the classification of the model.

Figs. 7a and 7b show that: (i) ARIL relies on two periods, around Packet 50 and Packets 100-140. (ii) ARIL mostly relies on the subcarriers between 20-35. Figs. 7c and 7d illustrate the explanations provided by CF where we observe similar trends as IG, i.e., important packets clustered around Packets 50 and 100, and the most important subcarriers are between 27- 33. However, we can clearly see that in this case

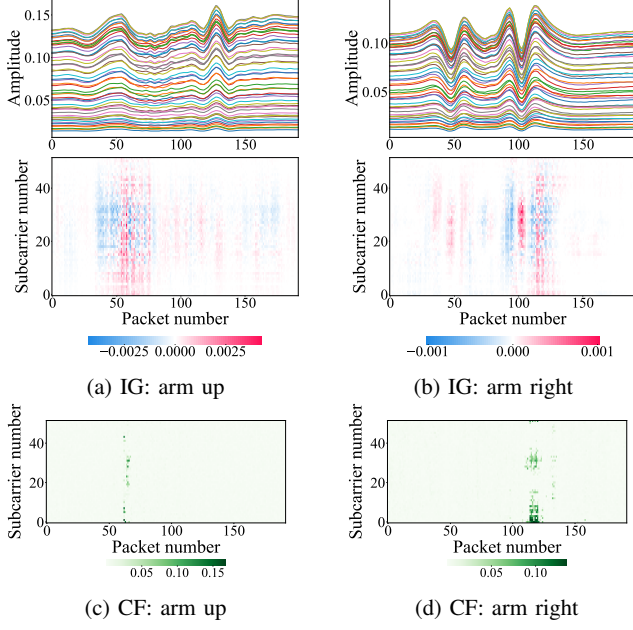


Fig. 6: The result of local explanations of ARIL model using IG and CF. The heat-map for CF visualizes the absolute difference between the two instances.

the *quality* of CF’s explanations suffers due to lack of access to the model internals. *Nonetheless, when dealing with blackbox models, CF can still shed light on important internal decisions of the model.*

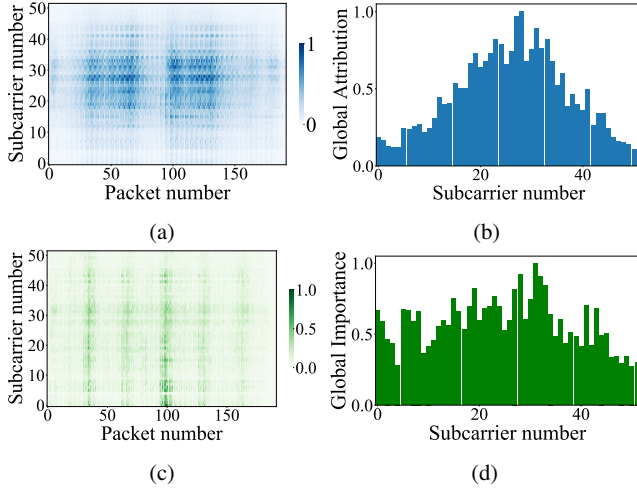


Fig. 7: The global explanations of ARIL: (a) over different packets and subcarriers using IG; (b) for each subcarrier using IG; (c) over different packets and subcarriers using CF; (d) for each subcarrier using CF.

3) *Validation: Reducing feature space:* The global explanations reveal that only a handful of the 52 subcarriers are critical to the classification of ARIL. We train ARIL with only the 1, 5, and 10 most important subcarriers and provide global explainability to validate our results, see Figs. 8 and

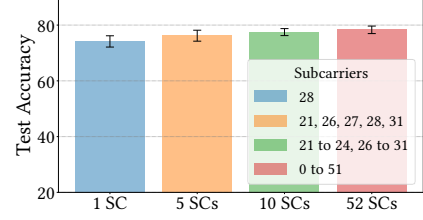


Fig. 8: Impact of number of used subcarriers in the CSI fingerprints on the classification accuracy of the ARIL model.

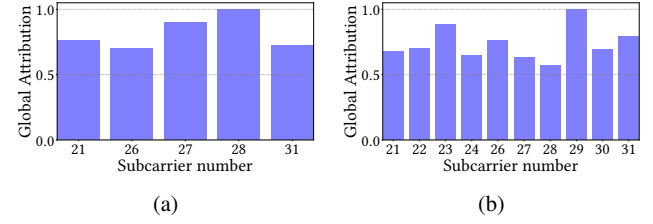


Fig. 9: Global importance of each subcarrier after feature space reduction to: (a) 5 subcarriers, and (b) 10 subcarriers.

9 For each setting, we trained 15 models and compared their accuracies. The results confirm the reliability of eXSense as the accuracy of the model only increases slightly with the number of used subcarriers. Using only the ten most important subcarriers, the model achieves nearly the same results (78.9%) as with 52 subcarriers while reducing the feature space by 80%. Moreover, if we use the most important subcarrier (i.e., 28), the accuracy only decreases by 4%, reducing feature space by nearly 98%. *This can be especially interesting in sensing with IoT devices, as these devices are constrained in computing and data storage.* Next, we evaluate importance from the time domain perspective. Using eXSense’s insights (identifying the packets around 40-70 and 100-150 as the most important ones), we only focused on those parts of the data which consequently reduced the feature space by 80% and 98%, while incurring only 0.82% and 4.16% loss in accuracy compared to the original model, respectively.

4) *Interpretation: Why some features are more important:* As mentioned, ARIL mostly relies on the subcarriers between 20-35. To shed light on the reason behind that, we examined the frequency correlation. We see among all packets and different gestures both locally and globally, Subcarriers 20-35 are the most correlated as shown in Fig. 10a. The very low correlation in the outer subcarriers can be due to noise or interference from the neighboring channels as discussed in Section V-E1. To clarify more, looking into Fig. 10b, we observe that the outer subcarriers are received with very low power. We interpret it as the CSI is more affected by a co-channel with high power than the person movement interferences.

Looking into the time domain, we observe that ARIL relies on two periods, around Packet 50 and Packets 100-140. By leveraging the time correlation, we see around these two periods, the correlation is the weakest. The correlation in the time domain indicates a lack of entropy/information. In essence, the channel changes more in the presence of movement compared to its absence. As a result, packets with high correlation

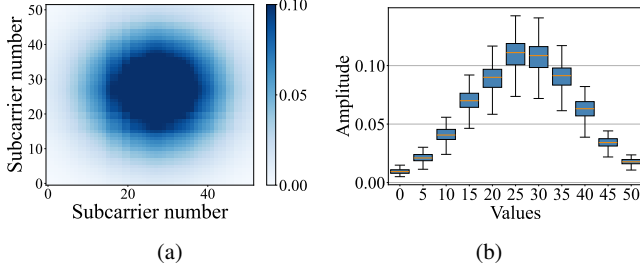


Fig. 10: (a) Frequency correlation and (b) subcarrier amplitude range

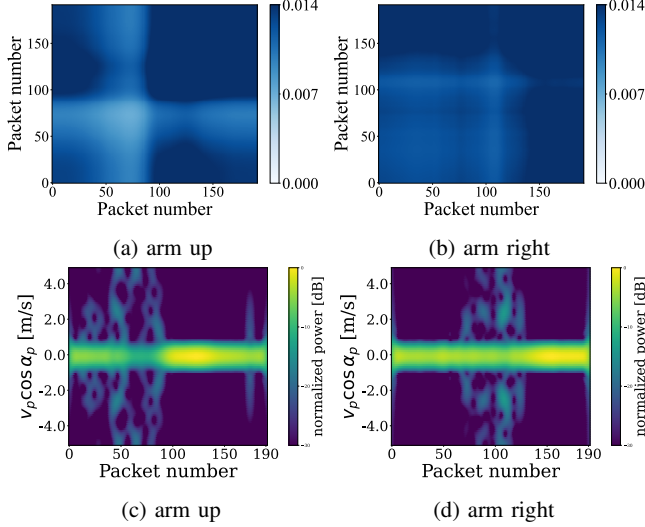


Fig. 11: Time correlation and Doppler spectrum for gestures (a)-(c) arm up, and (b)-(d) arm right.

observe little changes in the channel and consequently are considered less important by the NN, see Figs. 11a and 11b. To further investigate, we analyzed the Doppler spectrum. As shown in Figs. 11c and 11d the Doppler spectrum characterizes the frequency shifts and the proportion speed caused by the motion between the transmitter and receiver, which directly affects the time correlation and confirm our hypothesize. From the statistical point of view, we leverage the within-class and between-class variance according to Section V-E3 both in the time and frequency domain. As we can see in Fig. 12 the most important subcarriers have the highest between-class variance. There was no significant difference between the graphs of within-class variance and between-class variance, except that the maximum value of within-class variance is half of the between-class variance in the frequency domain. However, the very low variance in the outer subcarriers indicates the low SNR and confirms the previous observations. In the time domain, the packets around 50 and 100 have the least within-class and the most between-class variance which is preferable according to the pattern classification theory (see Figs. 12b and 12c).

Takeaway. The explanations of eXSense revealed that the model makes decisions based on two critical times (the packets around 50 and 100) and the subcarriers between 23-33.

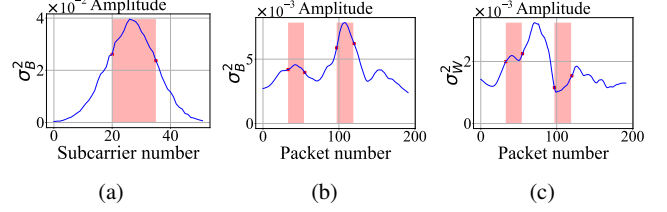


Fig. 12: Between-class variance in (a) the frequency domain and (b) the time domain, and within-class variance in (c) the time domain. The plots illustrate the corresponding variances, where the pink bars highlight the most important features identified by IG.

We shed light on the reasons behind this by considering correlation, Doppler spectrum and variance.

B. SignFi

SignFi uses 36,000 features ($3 \text{ antennas} \times 30 \text{ subcarriers} \times 200 \text{ packets}$) to differentiate among 272 ASL signs. Looking at the amplitude (Fig. 13a) and phase (Fig. 14a) of the CSI samples for the ASL sign *help*, it is evident that there are no visually observable patterns.

1) *Local Explainability*: Next, we follow the workflow of eXSense to understand the decision mechanism of SignFi. The local explanations from IG reveal that the information of the third antenna is more important than the other two, see Fig. 13b. However, a similar trend is not seen for this particular example when using CF, as observed in Fig. 13c. The sample could also be an outlier for the CF. To confirm this, we need to analyze the global explainability.

Fig. 14 shows local explanations for the phase. The interesting finding here is that the algorithm barely uses phase information for classification. This is in particular evident by comparing Figs. 13b and 14b where the importance of the amplitude values (scaled with e^{-4}) is three orders of magnitude higher than that of the phase (scaled with e^{-7}). We can observe the same for the explanations provided by CF. Figs. 13c and 14c show that more changes are made to the amplitude than to the phase, which indicates the amplitude is more important.

2) *Global Explainability*: Given that SignFi's features are multi-dimensional, we study the global importance from three aspects. In Figs. 15a and 15b, we observe the importance of information from each antenna for CSI phase and amplitude. Both IG and CF confirm that the information from the third antenna is more important than the ones of the other two. Furthermore, phase plays a small role in the decision making. For example, IG ranks the amplitude 158 times more important than the phase. Next, we analyze how importance evolves over time/packets. Unlike ARIL, we see in Figs. 15c and 15d that all packets contribute to a large extent to the overall classification. This can be due to the high number of gestures, which force the algorithms to better exploit the temporal variance of CSI amplitude. Last we observe the importance of subcarriers in Figs. 15e and 15f. Here the findings of both

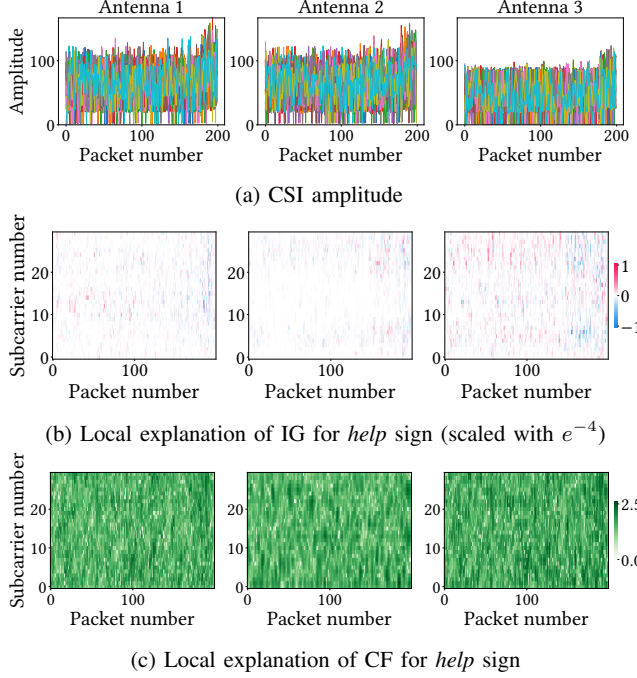


Fig. 13: The local explanation of sign *help* with focus on the CSI amplitude.

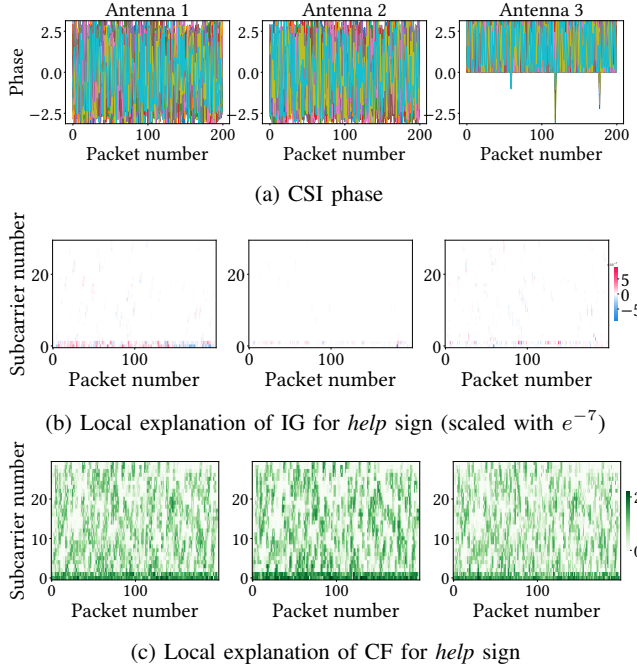


Fig. 14: The local explanation of sign *help* with focus on the CSI phase.

methods strongly correlate: The model mainly utilizes three subcarrier bands around Subcarriers 4-7, 13-15, and 19-22.

3) *Validation: Reducing feature space:* From the global and local explanations, we observed that phase has negligible importance compared to the amplitude, and the third antenna

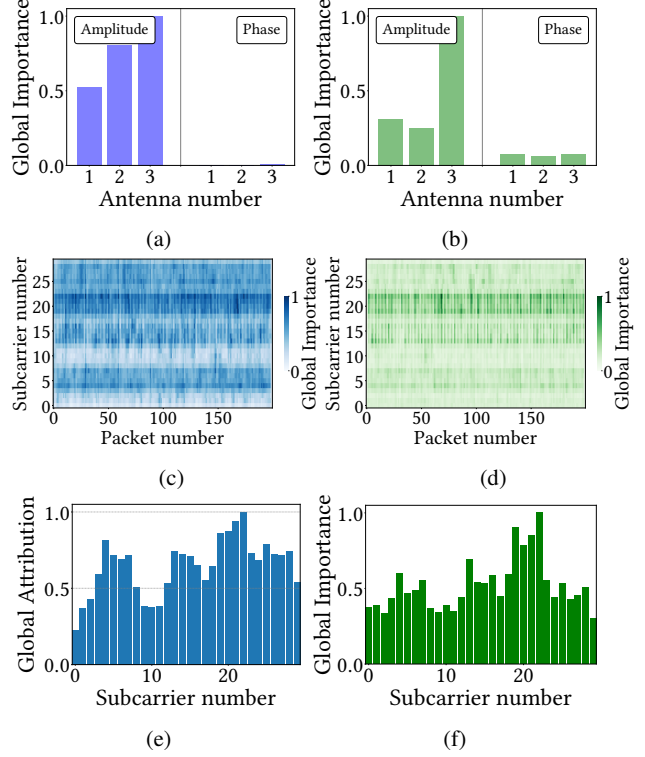


Fig. 15: The global explanation SignFi using (a)-(c)-(e) IG; (b)-(d)-(f) CF.

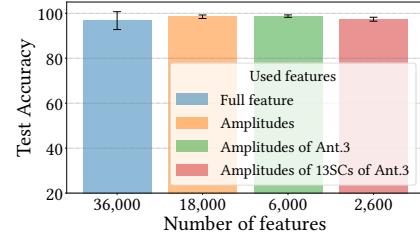


Fig. 16: Different data settings and their impact on the accuracy of the SignFi model. Using less and meaningful features can increase the performance and efficiency of the model.

is the most important among antennas. Lastly, we saw that the model mainly utilizes three distinct subcarrier bands: ranging from 4 to 7, 13 to 15, and 19 to 22. Given this information, we should be able to reduce the feature space by using only the amplitude of Antenna 3 and a subset of highly attributed subcarriers. To verify this, we train three new models with (i) amplitude values by removing the phase (18,000 features), (ii) amplitude values of Antenna 3 by removing the phase and the first and second antennas (6,000 features), and (iii) amplitude values of 13 subcarriers of Antenna 3 (2,600 features).

The first interesting observation in Fig. 16 is that by excluding the phase and the two antennas not only we increase the accuracy of the models (by 2%), but also make them more stable during training (the variance goes down from 18% to 0.3%), while simultaneously compressing the feature space by 83%. The feature space can be further optimized by selecting a subset of important subcarriers which results in a feature space

compression of up to 93% by using 13 subcarriers while still having a small improvement in accuracy. The reason for these improvements is that the new models are trained on less but more meaningful data. Hence, the models can concentrate on more critical parts of the data and further adapt the weights better on them. This exhibits that eXSense not only fosters understanding of the models but also improves them.

4) Interpretation: why some features are more important?:

As previously mentioned, SignFi primarily relies on the Subcarriers 4-7, 13-15, and 19-22, with the third antenna contributing the most significant data. To elucidate this preference, we employ point correlation analysis across subcarriers from different antennas. Examining Fig. 17a, it is evident that none of the subcarriers from Antenna 1 exhibit correlation, potentially attributed to the absence of cyclic shift diversity (CSD) in this antenna and interference from the adjacent antenna. In contrast, the subcarriers from Antenna 3 display a higher correlation than those from Antenna 2. Notably, the Subcarriers 4-7, 13-15, and 19-22 in both antennas emerge as the most correlated, as illustrated in the figure. For further investigation, we leverage the Doppler spectrum. Considering Figs. 17a, 17d and 17g, the Doppler spectrum of Antenna 1 is masked by noise. To ensure that this spectrum is white noise, we have computed the spectrum of a random white Gaussian noise and get the same result. Among Antennas 2 and 3, while they have similar spectrums, the Doppler spectrum of Antenna 3 is stronger than Antenna 2 (Consider the threshold in Figs. 17h and 17i). To analyze the results, we consider the real part of CSI as it contains the information of both amplitude and phase(see Figs. 17e and 17f).

From the statistical point of view, interesting results were obtained. We compute the within-class and between-class variances for the phase and amplitude of each antenna, both in the time and frequency domain. In the frequency domain, as we can see from Figs. 18a and 18b, the amplitude of Antenna 3 has the least within-class variance and the highest between-class variance in Subcarriers 4-7, 13-15, and 19-22 (the pink bars in the plot indicate these subcarriers as the most important feature). However, looking at Figs. 18c and 18d for the CSI phase the between-class is less than the within-class variance across all the antennas which is not preferable as discussed in Section V-E3. Furthermore, the first antenna has the highest within-class variance and the least between-class variance. In the time domain, we observe the same behavior for the antennas across all the packets. As depicted in Figs. 19a and 19c, the first antenna exhibits the highest within-class variance, and from Figs. 18b and 19d it is revealed that it has the lowest between-class variance, which is not a desirable characteristic. However, the third antenna has the least within-class and the highest between-class variance for the amplitude which are signs of a good feature. Consequently, these results confirm our previous hypothesis that the first antenna is noisy. Looking into Figs. 19c and 19d for the phase domain, the within-class variance is greater than the between-class variance across all the antennas, the packets and subcarriers that confirm that the phase has offsets and is not a good feature.

As a consequence, Figs. 17, 18 and 19 answer these questions very well, Why does the third antenna have higher

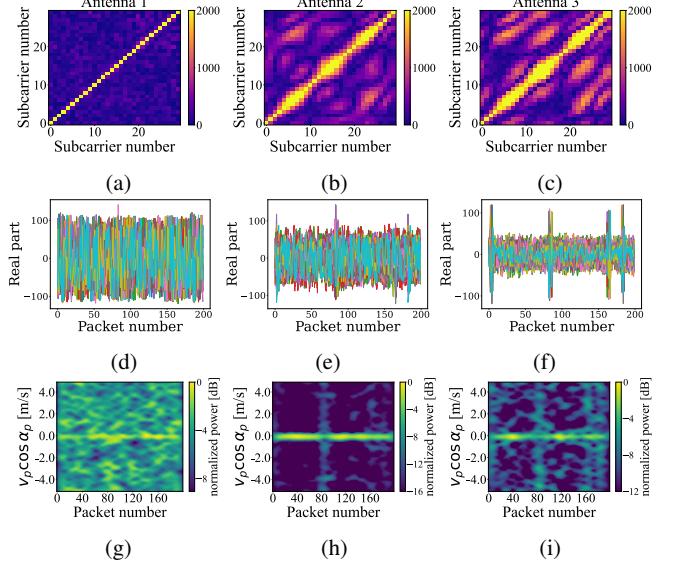


Fig. 17: Frequency correlation, real part and Doppler spectrum of the CSI among different antennas. (a)-(d)-(g): Antenna 1 (b)-(e)-(h): Antenna 2 (c)-(f)-(i): Antenna 3.

importance than the other two? Why is phase information not utilized by the model? And why are some subcarriers more important than others?

Takeaway. The explanations of eXSense revealed that CSI phase and the first and second antennas are less important. Further investigation demonstrated that SignFi's feature space can be compressed by 83% (omitting 30,000 features). This improvement enhances the data efficiency of the model and results in a significantly more stable model (lower variance) with a small 2% improvement in accuracy.

VIII. DISCUSSION

In this section, we discuss some other important aspects of XAI in general and eXSense in particular.

A. Stability, Fidelity and Efficiency of the Explanations

To rigorously evaluate the quality and fairness of the proposed eXSense framework, we compare explanation methods using four criteria: stability, fidelity, hyperparameter sensitivity, and efficiency (Table I), and benchmark them against two state-of-the-art methods. Specifically, Local Interpretable Model-agnostic Explanations (LIME) and Layer-wise Relevance Propagation (LRP) are implemented using the Python libraries *lime* [39] and *Zennit* [52], respectively. All experiments are conducted on ARIL dataset (Class arm right) under consistent system settings. A unified evaluation metric based on the Spearman rank correlation is used across criteria to assess the consistency of feature importance rankings. Unlike cosine similarity, which is sensitive to magnitude and sparsity, Spearman correlation focuses on preserving the relative ordering of important features [53].

<https://github.com/marcotcr/lime>

<https://github.com/chr5tphr/zennit>

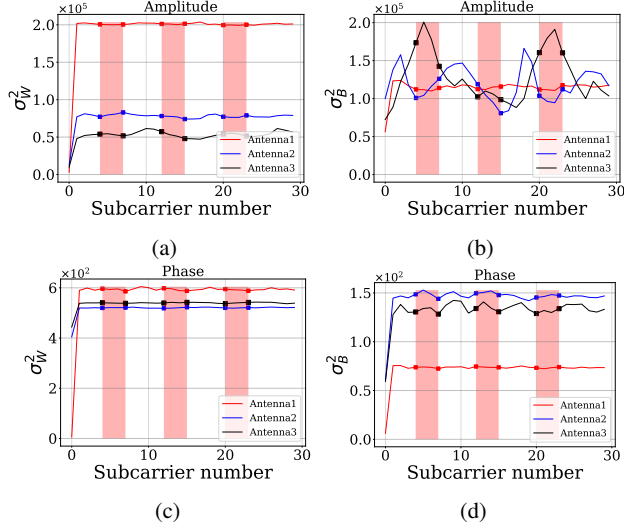


Fig. 18: Computed variance in the frequency domain: (a) within-class and (b) between-class variance of amplitude, (c) within-class and (d) between-class variance of phase. Important features exhibit lower within-class variance and higher between-class variance.

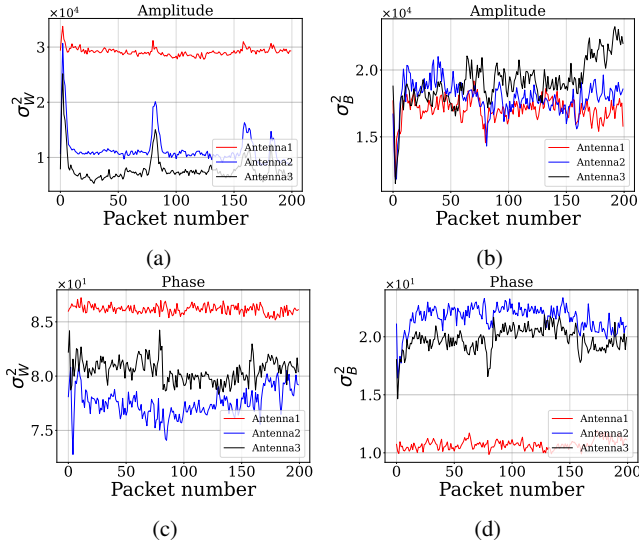


Fig. 19: Computed variance in the time domain: (a) within-class and (b) between-class variance of amplitude, (c) within-class and (d) between-class variance of phase. The results indicate that the amplitude of the third antenna exhibits the lowest within-class variance and the highest between-class variance in the time domain.

Stability is evaluated across (i) samples belonging to the same activity, and (ii) specific sample across different random initializations. For the latter, the random seed is varied from 1 to 100 with a step size of 10, resulting in model accuracies between 78% and 84.5%. IG and CF demonstrate stable results.

Fidelity is assessed via an ablation-driven analysis. Following

TABLE I: Comparison of explainability methods

Method	Stability over samples	Stability over Seeds	Ablation Fidelity	Hyperparameter Sensitivity	Time per Sample
IG	1	0.72	0.88	Baseline: 0.32 / Step: 0.99	0.15s
CF	0.97	0.86	0.82	p_t : 0.99 / d_x : 0.91	20min
LRP	0.69	0.65	0.78	γ : 0.62 / ϵ : 1	0.11s
LIME	0.13	0.09	-0.07	Number of feature: 0.1	5.4s

the ablation test conducted on Fig. 4c subcarriers are grouped into five categories and ranked from 1 to 5 based on their impact on performance. The explanation outputs are quantized into the same ranking levels, and their agreement with the ablation-based ordering is measured via Spearman correlation. IG, CF, and LRP show acceptable results, with IG achieving the highest performance.

Hyperparameter sensitivity is analyzed by varying one parameter at a time while fixing all others at their optimal values. The explored ranges include IG baselines (zero, Gaussian $\mathcal{N}(0, \sigma^2)$, Gaussian $\mathcal{N}(\mu, \sigma^2)$, and averaging) with steps 50, 100, 150, 300; CF parameters $p_t \in 0.6, 0.7, 0.9$ and $d_x \in 6, 11, 30$; LRP parameters $\gamma \in 0.01, 0.25, 0.5, 1$ and $\epsilon \in 10^{-6}, 10^{-4}, 10^{-3}, 10^{-2}$; and LIME feature ratios $N \in 1\%, 5\%, 10\%, 20\%$. LIME yields the worst results, as the ranking of important features is highly dependent on the number of selected features, this is in case the number of important features are not predetermined.

Efficiency is measured as the average computation time per sample⁷. CF incurs the highest computational cost due to its optimization-based formulation and hyperparameter tuning (i.e., λ). Technically, model-specific methods such as IG and LRP are faster than model-agnostic approaches such as LIME and CF, as they operate directly on the model parameters. Furthermore, methods based on Shapley additive values are not computationally feasible under the same hardware configuration due to their high memory and processing requirements. IG achieves a favorable trade-off between efficiency and reliability.

B. Feature selection and noise reduction

Based on the results from eXSense and considering the theoretical relations in Section V-E the metric that best fulfills the feature selection by these models is the between-class and within-class ratio. As depicted in Fig. 20 the most important features have the highest ratio. The amplitude has a higher ratio than the phase in both datasets. However, it is essential to consider both the amplitude and the phase to identify the most important packets or subcarriers (e.g., the packets around 50 in the ARIL dataset) as they are part of one CSI sample.

In the case of the Doppler spectrum, our observation shows that omitting the less important (e.g. antenna or subcarrier) increases the spectrum resolution. For the theoretical reason behind that, substituting Eq. 18 in Eq. 23 then we have:

$$\mathcal{F}\{H'_i\}(k, u) = \mathcal{F}\{H_i\}(k, u) + \underbrace{\mathcal{F}\{W\}(k, u)}_{(*)}. \quad (26)$$

⁷All experiments in this part are conducted on a system with an 11th Gen Intel Core i5 CPU and 32 GB RAM.

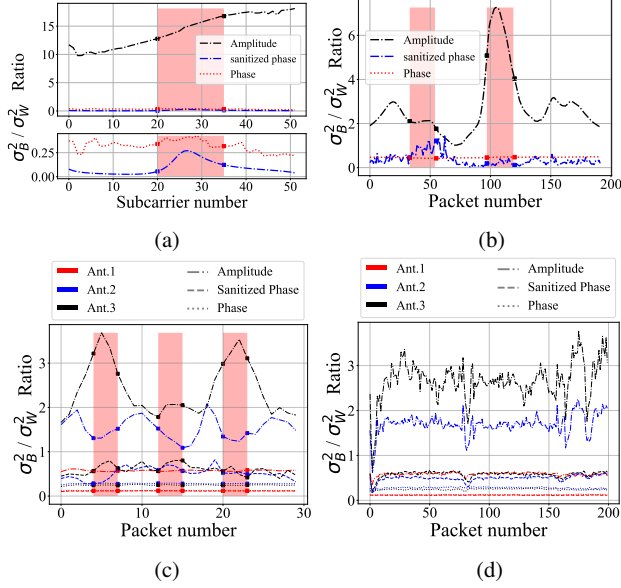


Fig. 20: The ratio of between-class to within-class variance across (a) subcarriers and (b) packets for ARIL, and (c) subcarriers and (d) packets for SignFi. The most important features identified by eXSense exhibit the highest ratio.

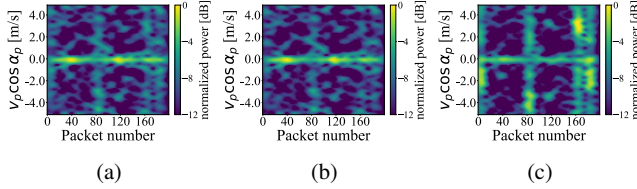


Fig. 21: Doppler spectrum from CSI observed by Antenna 3 with (a) all, (b) the least, and (c) the most important subcarriers.

By excluding the less important subcarriers in Eq. 22, we can reduce the impact of noise on the spectrum ($\odot$ in Eq. 26). As can well be inferred from Figs. 17g and 21b, the derivation from H adds noise to the spectrum. It is important to note that although phase sanitizing can help reduce offsets, it may not completely sanitize the phase, especially in the presence of co-channel interference or hardware IQ imbalance. In such cases, disregarding the noisy subcarriers is more reasonable (see Fig. 21). To validate the robustness of eXSense, we trained both the ARIL and SignFi datasets using the SHARP [13] and SLNet [54] workflows, which are based on Doppler spectrum representations. Experiments were conducted under two conditions: (i) using all available features (full features), and (ii) using only the most important subcarriers and antennae (reduced features), namely the Subcarriers 27–33 for ARIL and the Subcarriers 4–7, 13–15, and 19–22 of Antenna 3 for SignFi. As shown in Fig. 22a, the reduced-feature configuration improves both training and testing accuracy in the SHARP workflow. In the SLNet workflow, instead of summing over subcarriers in Eq. 22, Principal Component Analysis (PCA) is applied and only the first principal component is

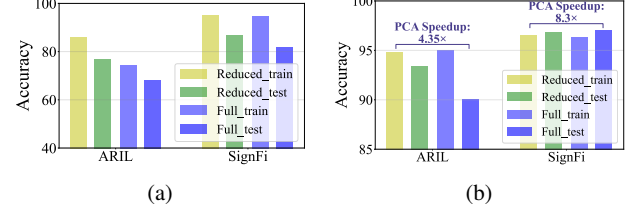


Fig. 22: Training ARIL and SignFi with (a) SHARP and (b) SLNet under two settings: using all features, and using the most important subcarriers and antennae (i.e., subcarriers 27–33 for ARIL, and subcarriers 4–7, 13–15, and 19–22 of Antenna 3 for SignFi). The SHARP workflow improves both training and testing accuracy. For SLNet, ARIL achieves higher test accuracy, while SignFi achieves up to an 8.3× speedup due to accelerated PCA computation.

retained. Under this setting, ARIL demonstrates improved performance, achieving higher test accuracy. Moreover, dimensionality reduction substantially accelerates PCA computation (e.g., reducing the runtime from 36.4 minutes to 8.36 minutes, while the computation of the between-class to within-class ratio takes only 0.03 seconds).

C. Dimension reduction and model efficiency

Although we used dimension reduction to validate eXSense, the outcome of between-class to within-class ratio can also be used as a practical tool for reducing the feature space. As we see in Table III, this *guided* approach for feature removals has several advantages: (i) reducing the training time,⁸ (ii) increasing the classification speed, (iii) reducing the data size itself. This is critical for adaptation of ML-based sensing on constrained IoT devices. We also observed optimizing the feature selection can enhance the accuracy and the stability of the model hence making the model more efficient. We applied the variance between-class to within-class ratio and exclude features with ratio below 1 to all SignFi datasets, including *Home-276*, *Lab-276*, and *Lab-150*⁹. The first part of each dataset name indicates the location where data were collected, while the second part denotes the number of classes. The *Lab-150* dataset has been performed by five different individuals, which we concatenate and treat as a single dataset. As shown in Fig. 23a, this approach reduces training time while increasing test accuracy. Notably, for *Lab-150*, accuracy improved from 42.34% to 72.16%, and training time decreased from 50 seconds to 27 seconds. As another example of improvement in model efficiency is an human independency levels. Considering *Lab-150* trained the model, we trained the model on data from users 3, 4 and 5, and evaluated it through cross-validation on data from users 1 and 2. As shown in Fig. 23b, both test accuracy and cross-validation accuracy have improved. This includes improvement in cross-validation

⁸The experiments are conducted on a Windows machine with an Intel i5-7400 processor, 16GB RAM, and an NVIDIA GeForce GTX 1060 graphic card.

⁹The explanation was conducted on *Home-276* dataset and *Lab-276* and *Lab-150* are considered as unseen data

accuracy from 43.4% to 57.9%. We also evaluated the Widar3 [12] dataset using the SignFi model. The Widar3 dataset is organized into folders corresponding to different experimental settings, varying in participants, activity types, and room environments. To assess cross-location performance, we used data from folders "20181109", "20181118", and "20181211", where a fixed participant (Subject 3) performed experiments in Rooms 1, 2, and 3. For cross-person evaluation, folder "20181130," which contains the largest number of participants, was selected. For consistency across all experiments, only activities 1 to 4 and receiver number 1 are considered. Each receiver is equipped with three antennas, which aligns well with the SignFi dataset structure. Fig. 24a illustrates the between-class to within-class ratio for Antenna 1. The ratios corresponding to Antennas 2 and 3 are below 1 across all rooms. As shown in the figure, Room 3 exhibits the highest ratio across all subcarriers, followed by Room 1, while Room 2 shows the lowest ratio, consistently remaining below 1. Correspondingly, as depicted in Fig. 24c, Room 3 achieves the highest classification accuracy (100%), whereas Room 2 attains the lowest accuracy (65%). Furthermore, by considering only the amplitude and the Subcarriers 15–30 from Antenna 1 as a reduced feature set, the classification accuracy in Room 2 improves to 75%, while the accuracies for Rooms 1 and 3 remain unchanged. Fig. 24b illustrates the results for different subjects in Room 1 using Antenna 3. As shown, the between-class to within-class ratio exceeds 1 for Subjects 5, 10, 12, and 16. Moreover, amplitude features consistently outperform phase features, and Subcarriers 15–30 provide superior performance compared to other subcarrier ranges. By selecting amplitude features associated with Subjects 5, 10, 12, and 16, as well as the corresponding antennas and subcarriers whose ratios exceed 1, specifically, amplitude of Subcarriers 0–15 from Antenna 2 and Subcarriers 15–30 from Antenna 3, the performance improves significantly. In this configuration, Subjects 5, 10, and 16 are used for training and testing, while Subject 12 is treated as an unseen subject for cross-validation. As shown in Fig. 24d, the classification accuracy increases substantially across training, testing, and cross-validation. The summary of human independency improvement is given in Table III. The other example of this enhancement is an improvement in resistance to noise levels in ARIL. Here, we compare the full feature model with the model with 5 and 10 subcarriers. To test the model's efficiency, we introduce several noise levels (i.e., from 5%-30%) to the samples of the highest attributed subcarrier (i.e., 28). We can observe in Fig. 25 that the models with our proposed feature optimization tolerate the noise significantly better than the original full-feature model. The accuracy of the model trained on five subcarriers is not even affected. Moreover, we can observe that after reaching a specific noise level the accuracy is not further decreased, as the loss of information is saturated. Note that using eXSense on DL for dimension reduction appears to be more advantageous than the state-of-the-art feature selection methods. Unlike state-of-the-art techniques, eXSense eliminates the need for expert intervention. Moreover, certain aspects of automatic feature selection by DL models, such as CNNs, may lack clarity for

TABLE II: Summary of improvements in the case studies

Case Study	Features		Training time (s)	Classification time (s)	Data Size
ARIL	9,960	full feature	101	0.14	89MB
	1,920	81% reduction	90	0.11	17MB
	960	93% reduction	90	0.11	8.5MB
	192	98% reduction	90	0.11	1.7MB
SignFi	36,000	full feature	17	0.29	1.27GB
	18,000	50% reduction	10	0.2	635MB
	6,000	83% reduction	8	0.12	211MB
	2,000	95% reduction	5	0.11	63MB

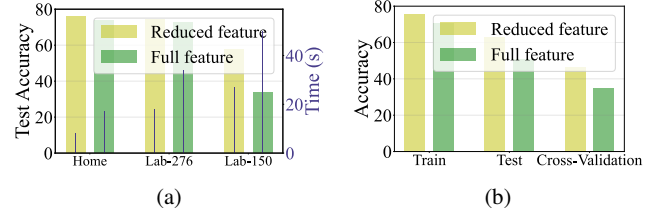


Fig. 23: Enhancement of the SignFi model through feature selection based on the between-class to within-class ratio. (a) Training optimization: improved test accuracy and reduced training time across different environments, with the largest gain in the *Lab-150* dataset. (b) Human-independency improvement: training with three individuals in the *Lab-150* dataset improves both training and cross-validation accuracy on two unseen individuals.

TABLE III: Summary of improvements in human independency

Case Study	Training Accuracy	Test Accuracy	Cross-Validation	Training Time
SignFi / <i>lab-150</i> full feature reduced feature	Human ID: 3,4,5 88.5% 94.5%	Human ID: 3,4,5 63.67% 74.78%	Human ID: 1,2 43.4% 57.9%	(s) 23 15
	Human ID: 5,10,16 85% 97%	Human ID: 5,10,16 58% 100%	Human ID: 12 50% 70%	(s) 180 32.6

manual consideration by experts. Essentially, DL models prove adept at identifying meaningful features within vast datasets, and eXSense serves as a valuable tool for scrutinizing and validating these features. Consequently, this knowledge can be harnessed to develop leaner and more efficient models. Additionally, eXSense facilitates the comparison of model efficiency and the impact of different datasets.

D. Environment independency

One of the main challenges in WiFi sensing is to create a model that is environment-independent. We have shown with eXSense that the information lies in the CSI's amplitude. Unfortunately, the amplitude is heavily affected by the environment. Given that tools used for CSI extraction transmit constant packets and considering Eq. 27 the use of CSI's amplitude is similar to using RSSI with a scaling factor of $\frac{1}{|X|}$.

$$|H(k, n)| = \frac{|Y(k, n)|}{|X(k, n)|}. \quad (27)$$

The most advanced solution is to utilize the Doppler spectrum to separate dynamic information from static. However, the Doppler spectrum provides a velocity component of $v \cos \alpha$,

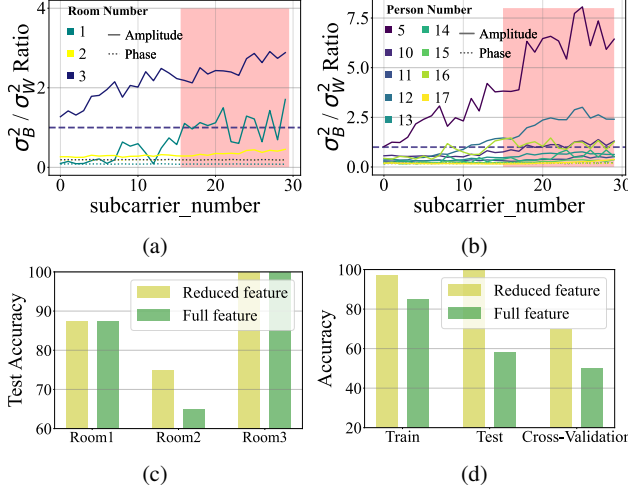


Fig. 24: Evaluation of the SignFi model on the Widar3 dataset. Between-class to within-class ratios for (a) different rooms (Subject 3, Antenna 1), and (b) different subjects (Room 1, Antenna 3). (c) Room classification accuracy using full and reduced features (Amplitude of Antenna 1, Subcarriers 15–30). (d) Human-independency evaluation, selecting features and subjects with ratios greater than 1 (Amplitude of Subcarriers 0–15 of Antenna 2 and 15–30 of Antenna 3 for Subjects 5, 10, and 16 as training/testing, with Subject 12 as unseen for cross-validation).

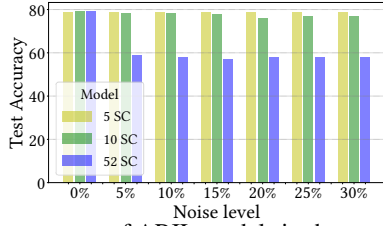


Fig. 25: The accuracy of ARIL models in the presence of noise on the most important subcarrier (i.e., 28). The results show that the models trained on less subcarriers are more resilient.

where α is the motion orientation. Consequently, the Doppler spectrum is heavily influenced by the position and motion orientation of the individual to the sender and receiver [55]. Additionally, it is crucial to have a clean dataset since the spectrum is highly impacted by the phase offsets. Based on the above intuition, two points should be taken into account. First, find a phase-offset-removing algorithm that can increase the ratio of between-class to within-class variance of the CSI's phase to be in the range of the amplitude. Second, develop the Doppler spectrum which is position and motion orientation independent.

E. Investigating the model reliability and trustworthiness

eXSense and, more broadly, XAI, not only verify black-box ML models but also help analyze their behavior. As shown in Section VIII-C, a key strength of DL is handling large input spaces and automatically extracting features through layers such as convolution. This enables models to process

extensive datasets and detect hidden patterns. However, identifying which features are truly meaningful remains essential for trustworthiness. High accuracy alone is insufficient if decisions rely on spurious correlations [56]. While XAI reveals feature contributions, in WiFi sensing with CSI inputs, it cannot fully verify their reliability. Unlike images, CSI features are not self-evident. Here, eXSense is valuable for interpreting results and explaining feature selection. It demonstrates the reliability of ARIL and SignFi. In summary, DL finds meaningful features in large datasets, while eXSense helps analyze and verify them, enabling more reliable models.

IX. CONCLUSION

We proposed eXSense, a workflow to utilize the potential of XAI techniques for understanding blackbox ML-based WiFi sensing solutions. Specifically, we adapted two complementary state-of-the-art XAI techniques to obtain explainability both in the presence as well as the absence of access to the parameters of the blackbox model. We performed an extensive evaluation of two WiFi sensing case studies using eXSense. Based on the insights from the explanations, we optimized their models, which enhanced the overall performance with no/minimal impact on the accuracy of the model. eXSense sheds light on the parts of data which carry more weight in decision making and explains the logic behind that. Having an XAI method that provides the signal parameters for these parts (e.g., AoA or Doppler) can be helpful in verifying the operation of WiFi sensing proposals that use raw CSI samples.

REFERENCES

- [1] W. Jiang, C. Miao, F. Ma, S. Yao, Y. Wang, Y. Yuan, H. Xue, C. Song, X. Ma, D. Koutsoukolas *et al.*, "Towards environment independent device free human activity recognition," in *Proceedings of ACM MobiCom*, 2018, pp. 289–304.
- [2] Y. Xu, W. Yang, J. Wang, X. Zhou, H. Li, and L. Huang, "Wistep: Device-free step counting with wifi signals," *Proceedings of ACM IMWUT*, vol. 1, no. 4, pp. 1–23, 2018.
- [3] G. Wang, Y. Zou, Z. Zhou, K. Wu, and L. M. Ni, "We can hear you with wi-fi!" *IEEE Transactions on Mobile Computing*, vol. 15, no. 11, pp. 2907–2920, 2016.
- [4] H. Li, W. Yang, J. Wang, Y. Xu, and L. Huang, "Wifinger: Talk to your smart devices with finger-grained gesture," in *Proceedings of ACM UbiComp*, 2016, pp. 250–261.
- [5] S. M. Hernandez and E. Bulut, "Adversarial occupancy monitoring using one-sided through-wall wifi sensing," in *Proceedings of ICC*, 2021, pp. 1–6.
- [6] S. Palipana, D. Rojas, P. Agrawal, and D. Pesch, "Falldefi: Ubiquitous fall detection using commodity wi-fi devices," *Proceedings of ACM IMWUT*, pp. 1–25, 2018.
- [7] J. C. Soto, I. Galdino, E. Caballero, V. Ferreira, D. Muchaluat-Saade, and C. Albuquerque, "A survey on vital signs monitoring based on wi-fi csi data," *Computer Communications*, vol. 195, pp. 99–110, 2022.
- [8] Y. Gu, T. Liu, J. Li, F. Ren, Z. Liu, X. Wang, and P. Li, "Emosense: Data-driven emotion sensing via off-the-shelf wifi devices," in *Proceedings of IEEE ICC*, 2018, pp. 1–6.
- [9] R. Du, H. Hua, H. Xie, X. Song, Z. Lyu, M. Hu, Y. Xin, S. McCann, M. Montemurro, T. X. Han *et al.*, "An overview on ieee 802.11 bf: Wlan sensing," *IEEE Communications Surveys & Tutorials*, vol. 27, no. 1, pp. 184–217, 2024.
- [10] F. Wang, J. Feng, Y. Zhao, X. Zhang, S. Zhang, and J. Han, "Joint activity recognition and indoor localization with wifi fingerprints," *IEEE Access*, vol. 7, pp. 80058–80068, 2019.
- [11] Y. Ma, G. Zhou, S. Wang, H. Zhao, and W. Jung, "Signfi: Sign language recognition using wifi," *Proceedings of ACM IMWUT*, pp. 1–21, 2018.
- [12] Y. Zheng, Y. Zhang, K. Qian, G. Zhang, Y. Liu, C. Wu, and Z. Yang, "Zero-effort cross-domain gesture recognition with wi-fi," in *Proceedings of ACM MobiSys*. Association for Computing Machinery, 2019, p. 313–325.

- [13] F. Meneghello, D. Garlisi, N. Dal Fabbro, I. Tinnirello, and M. Rossi, "Sharp: Environment and person independent activity recognition with commodity iee 802.11 access points," *IEEE Transactions on Mobile Computing*, 2022.
- [14] D. Adesina, C.-C. Hsieh, Y. E. Sagduyu, and L. Qian, "Adversarial machine learning in wireless communications using rf data: A review," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 1, pp. 77–100, 2022.
- [15] H. Cai, B. Korany, C. R. Karanam, and Y. Mostofi, "Teaching rf to sense without rf training measurements," *Proceedings of ACM IMWUT*, vol. 4, no. 4, pp. 1–22, 2020.
- [16] T. Gong, Y. Kim, J. Shin, and S.-J. Lee, "Metasense: few-shot adaptation to untrained conditions in deep mobile sensing," in *Proceedings of ACM SenSys*, 2019, pp. 110–123.
- [17] R. Hesse, S. Schaub-Meyer, and S. Roth, "Fast axiomatic attribution for neural networks," in *NeurIPS*, vol. 34, 2021, pp. 19 513–19 524.
- [18] M. Danilevsky, K. Qian, R. Aharonov, Y. Katsis, B. Kavas, and P. Sen, "A survey of the state of explainable ai for natural language processing," in *Proceedings of the 1st conference of the Asia-Pacific chapter of the association for computational linguistics and the 10th international joint conference on natural language processing*, 2020, pp. 447–459.
- [19] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *International conference on machine learning*, 2017, pp. 3319–3328.
- [20] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the gdpr," *Harv. JL & Tech.*, vol. 31, p. 841, 2017.
- [21] Y. Ma, S. Arshad, S. Muniraju, E. Torkildson, E. Rantala, K. Doppler, and G. Zhou, "Location-and person-independent activity recognition with wifi, deep neural networks, and reinforcement learning," *ACM Transactions on Internet of Things*, vol. 2, no. 1, pp. 1–25, 2021.
- [22] E. Brophy, J. J. D. Veiga, Z. Wang, A. F. Smeaton, and T. E. Ward, "An interpretable machine vision approach to human activity recognition using photoplethysmograph sensor data," *arXiv preprint arXiv:1812.00668*, 2018.
- [23] Z. Shi, F. Chang, Y. Jia, J. Li, Y. Qiu, J. Miao, W. Jiang, X. Guo, X. Han, and W. Tang, "Classifying and understanding of dairy cattle health using wearable inertial sensors with random forest and explainable artificial intelligence," *IEEE Sensors Letters*, 2024.
- [24] A. Dubey, N. Lyons, A. Santra, and A. Pandey, "Xai-bayeshar: A novel framework for human activity recognition with integrated uncertainty and shapely values," in *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2022, pp. 1281–1288.
- [25] D. Das, Y. Nishimura, R. P. Vivek, N. Takeda, S. T. Fish, T. Ploetz, and S. Chernova, "Explainable activity recognition for smart home systems," *ACM Transactions on Interactive Intelligent Systems*, vol. 13, no. 2, pp. 1–39, 2023.
- [26] L. Yuan, J. Andrews, H. Mu, A. Vakil, R. Ewing, E. Blasch, and J. Li, "Interpretable passive multi-modal sensor fusion for human identification and activity recognition," *Sensors*, vol. 22, no. 15, p. 5787, 2022.
- [27] S. Wang, M. A. Qureshi, L. Miralles-Pechuán, T. Huynh-The, T. R. Gadekallu, and M. Liyanage, "Explainable ai for 6g use cases: Technical aspects and research challenges," *IEEE Open Journal of the Communications Society*, vol. 5, pp. 2490–2540, 2024.
- [28] A. Das and P. Rad, "Opportunities and challenges in explainable artificial intelligence (xai): A survey," *arXiv preprint arXiv:2006.11371*, 2020.
- [29] T. Zheng, Z. Chen, S. Ding, and J. Luo, "Enhancing rf sensing with deep learning: A layered approach," *IEEE Communications Magazine*, vol. 59, no. 2, pp. 70–76, 2021.
- [30] J. Rožman, H. Hagrass, J. A. Perez, D. Clarke, B. Müller, and S. F. Data, "Privacy-preserving gesture recognition with explainable type-2 fuzzy logic based systems," in *2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. IEEE, 2020, pp. 1–8.
- [31] Y. Ren, Y. Wang, S. Tan, Y. Chen, and J. Yang, "Person re-identification in 3d space: A {WiFi} vision-based approach," in *32nd USENIX Security Symposium (USENIX Security 23)*, 2023, pp. 5217–5234.
- [32] S. A. Siddiqui, D. Mercier, M. Munir, A. Dengel, and S. Ahmed, "Tsviz: Demystification of deep learning models for time-series analysis," *IEEE Access*, vol. 7, pp. 67 027–67 040, 2019.
- [33] H. Strobelt, S. Gehrmann, H. Pfister, and A. M. Rush, "Lstmvis: A tool for visual analysis of hidden state dynamics in recurrent neural networks," *IEEE transactions on visualization and computer graphics*, vol. 24, no. 1, pp. 667–676, 2017.
- [34] H. Strobelt, S. Gehrmann, M. Behrisch, A. Perer, H. Pfister, and A. M. Rush, "Seq2seq-vis: A visual debugging tool for sequence-to-sequence models," *IEEE transactions on visualization and computer graphics*, vol. 25, no. 1, pp. 353–363, 2018.
- [35] N. Tadayon, M. T. Rahman, S. Han, S. Valaee, and W. Yu, "Decimeter ranging with channel state information," *IEEE Transactions on Wireless Communications*, vol. 18, no. 7, pp. 3453–3468, 2019.
- [36] J. Huang, B. Liu, C. Chen, H. Jin, Z. Liu, C. Zhang, and N. Yu, "Towards anti-interference human activity recognition based on wifi subcarrier correlation selection," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 6, pp. 6739–6754, 2020.
- [37] K. Beckh, S. Müller, M. Jakobs, V. Toborek, H. Tan, R. Fischer, P. Welke, S. Houben, and L. von Rueden, "Explainable machine learning with prior knowledge: An overview," *arXiv preprint arXiv:2105.10172*, 2021.
- [38] G. Erion, J. D. Janizek, P. Sturmfels, S. M. Lundberg, and S.-I. Lee, "Improving performance of deep learning models with axiomatic attribution priors and expected gradients," *Nature machine intelligence*, vol. 3, no. 7, pp. 620–631, 2021.
- [39] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you?" explaining the predictions of any classifier," in *Proceedings of SIGKDD*, 2016, pp. 1135–1144.
- [40] "AI Explainability Whitepaper," Google, Tech. Rep., 2020.
- [41] L. S. Shapley, *A Value for N-Person Games*. Santa Monica, CA: RAND Corporation, 1952.
- [42] T. P. Michalak, K. V. Aadithya, P. L. Szczepanski, B. Ravindran, and N. R. Jennings, "Efficient computation of the shapley value for game-theoretic network centrality," *Journal of Artificial Intelligence Research*, vol. 46, pp. 607–650, 2013.
- [43] C. Molnar, *Interpretable Machine Learning*, 2nd ed., 2022. [Online]. Available: <https://christophm.github.io/interpretable-ml-book>
- [44] A. Barredo Arrieta, N. Diaz-Rodriguez, J. Del Ser, A. Bannetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information Fusion*, vol. 58, pp. 82–115, 2020.
- [45] F. Gringoli, M. Schulz, J. Link, and M. Hollick, "Free your csi: A channel state information extraction platform for modern wi-fi chipsets," in *Proceedings of WINTeCH*, 2019, pp. 21–28.
- [46] P. Sturmfels, S. Lundberg, and S.-I. Lee, "Visualizing the impact of feature attribution baselines," *Distill*, vol. 5, no. 1, p. e22, 2020.
- [47] J. Klaise, A. V. Loooveren, G. Vacanti, and A. Coca, "Alibi explain: Algorithms for explaining machine learning models," *Journal of Machine Learning Research*, vol. 22, no. 181, pp. 1–7, 2021.
- [48] R. Kommiya Mothilal, D. Mahajan, C. Tan, and A. Sharma, "Towards unifying feature attribution and counterfactual explanations: Different means to the same end," in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 2021, pp. 652–663.
- [49] A. Goldsmith, *Wireless communications*. Cambridge university press, 2005.
- [50] T. S. Rappaport, *Wireless communications: principles and practice*. Cambridge University Press, 2024.
- [51] X. Li and K. Pahlavan, "Super-resolution toa estimation with diversity for indoor geolocation," *IEEE transactions on wireless communications*, vol. 3, no. 1, pp. 224–234, 2004.
- [52] C. J. Anders, D. Neumann, W. Samek, K.-R. Müller, and S. Lapuschkin, "Software for dataset-wide xai: From local explanations to global insights with zennit, corelay, and virelay," *PloS one*, vol. 21, no. 1, p. e0336683, 2026.
- [53] S. Qaadan, A. AlZamer, M. Mohamed, A. Alshare, and B. Schmuelling, "Introducing feature importance consistency (fic) in xai models: A comprehensive evaluation using rank correlation coefficients," in *2025 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA)*. IEEE, 2025, pp. 1–6.
- [54] Z. Yang, Y. Zhang, K. Qian, and C. Wu, "{SLNet}: A spectrogram learning neural network for deep wireless sensing," in *20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23)*, 2023, pp. 1221–1236.
- [55] K. Niu, X. Wang, F. Zhang, R. Zheng, Z. Yao, and D. Zhang, "Rethinking doppler effect for accurate velocity estimation with commodity wifi devices," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 7, pp. 2164–2178, 2022.
- [56] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, and K. Müller, "Explainable ai: interpreting, explaining and visualizing deep learning," *Springer LNCS*, vol. 11700, 2019.