MULTI-FI: Enhancing Wi-Fi Sensing Accuracy via Multi-view and Context Fusion

Iñaki Bravo*†, Claudio Fiandrino*
*IMDEA Networks Institute, Spain, †Universidad Carlos III de Madrid, Spain
Email: {name.surname}@networks.imdea.org

***Abstract*—Wi-Fi sensing enables innovative applications in healthcare and surveillance by providing continuous, contactless monitoring through existing infrastructure. Moreover, exploiting information from different receivers within an environment (i.e., different views) and using it as input to Deep Learning (DL) models facilitates sophisticated use cases. Previous works have demonstrated that this approach, also known as collaborative sensing, increases sensing coverage and significantly boosts accuracy and robustness. However, existing DL models for collaborative Wi-Fi sensing exhibit two critical limitations: *(1)* They do not consider the optimal fusion point for data from different receivers; *(2)* They assign equal importance to all receivers, irrespective of their sensing capacity. In this paper, we address the above gaps by proposing MULTI-FI, a novel model enhancement framework consisting of two modules. The first one benchmarks the appropriate fusion strategy, given the characteristics of the sensing scenario. By leveraging physical context information like location and orientation and the appropriate fusion strategy, the second module instruments a model re-design strategy. Our validation of MULTI-FI spans across several SoTA DL models, three real-world datasets and two applications, and shows improvements over the original model version in every case, with average accuracy gains of 8.2% and up to 29.6%.**

I. INTRODUCTION

Over the past decades, Wi-Fi has established itself as a leading wireless communication technology, with more than 19.5 billion devices currently in use worldwide [1]. Its massive adoption has motivated both academia and industry[1] to explore capabilities beyond communications. Among them, sensing has emerged as a prominent functionality, offering advantages such as robustness to occlusions and preservation of visual privacy compared to camera-based solutions. Moreover, Wi-Fi sensing can be implemented on conventional off-the-shelf (COTS) devices [2], without requiring specialized hardware. These benefits have driven the establishment of a dedicated task group and standardization effort, namely IEEE 802.11bf [3], and enabled a wide range of applications, including gesture recognition [4], people tracking [5], breath monitoring [6], human activity recognition [7], and gait identification [8].

Wi-Fi sensing systems typically leverage Channel State Information (CSI) and learn its association with a target sensing task. Early efforts relied on analytical models for pattern identification and classification, but advances in Deep Learning (DL) have made data-driven approaches the dominant paradigm, as demonstrated in previous work [6]. However, more advanced applications require the use of data from receivers (Rx) deployed at various locations, i.e., a multistatic architecture [5], [9], [10]. This setup increases sensing coverage, improves robustness, and yields additional accuracy gains. Fig. 1 shows how multiple receivers collect CSI from different positions within the environment, providing distinct "views" that are fused and processed by a DL model to enable *collaborative sensing*. Previous work has validated the effectiveness of this paradigm across applications and environments [4], [11], [12], and its adoption is expected to expand further with emerging topologies such as EasyMesh Wi-Fi ™.

Our motivation is that existing collaborative Wi-Fi sensing approaches remain suboptimal. Our analysis shows that State-of-The-Art (SoTA) DL models can improve accuracy by up to 30% through simple yet principled design changes. In particular, we identify two distinct limitations: current models *(L1)* do not explicitly determine where complementary views should be fused within the network, i.e., the optimal fusion depth; *(L2)* treat all views as equally informative, ignoring that the sensing capacity of a receiver strongly depends on the physical context, such as its location and orientation relative to the sensing target and the transmitter (Tx).

In this paper, we propose MULTI-FI, a framework that addresses the above limitations in model design, and increases the accuracy across all scenarios tested. MULTI-FI consists of two modules. The first module addresses *(L1)* by systematically assessing the most suitable multi-view fusion strategy for a given sensing scenario. Our results show that model accuracy can vary by up to 29.1% depending on the fusion strategy adopted. Despite its impact, this design choice has largely been overlooked in Wi-Fi sensing. Most existing multi-view models adopt a single fusion strategy without systematically evaluating its alternatives. In particular, most previous work relies on *input fusion*, where data from all receivers are combined before being fed to the classification model [4], [13]. While this approach is theoretically optimal under the assumption of infinite training data [14], in practical settings it presents drawbacks: *(1)* leads to overfitting; *(2)* assumes that all computation is performed on the edge/cloud; *(3)* models must be retrained whenever the number of available receivers changes. In this work, we perform a systematic analysis of alternative multi-view fusion strategies and demonstrate that even SoTA models can benefit from adopting a different fusion design.

The second module of MULTI-FI addresses *(L2)* by incorporating physical context features to improve *collaborative sensing*. Previous work has shown that the relative position

[1]https://www.originwirelessai.com

between the sensing target and transceivers significantly affects the sensing accuracy [15], [16], particularly in dense 802.11 deployments [17]. Although prior work has already explored view selection, they typically rely on task-specific architectures. Additionally, view selection is less accurate than ours (will be shown in Subsection II-C), since it inhibits the extraction of cross-view patterns. MULTI-FI differentiates itself by providing a model-agnostic enhancement layer; unlike existing solutions that require specialized "view-aware" models, MULTI-FI can be wrapped around any SoTA backbone (e.g., Convolutional Neural Network (CNN), transformers, etc.). Nonetheless, most models overlook view importance altogether. Specifically, models based on *input fusion* implicitly expect the model to prioritize informative views without guidance, while *score fusion* typically assigns equal importance to all views or selects a single view based on maximum confidence. MULTI-FI overcomes this limitation by explicitly identifying and prioritizing privileged views over less informative ones. It first estimates physical context features, e.g., location and orientation for gesture recognition, and trajectory and direction for gait identification. Then, these features are embedded into a multidimensional context vector, from which modulating weights are derived to dynamically adjust the contribution of each view for the final collaborative prediction.

We validate MULTI-FI as a model enhancement framework using five popular models, emphasizing the diversity of architectures. Among those, we include two well-established models tailored to Wi-Fi sensing: a version of Widar 3.0 [4] (CNN + Gated Recurrent Unit (GRU)) and WiFlexFormer [18] (Transformer). Additionally, we select three ubiquitous models from different domains with proven success for Wi-Fi sensing [19]: BiLSTM from time series and ResNet18 [20] and MobileNetV3 [21] from computer vision. Throughout the paper, we leverage three popular real-world datasets targeting different applications. To illustrate the crucial role that the physical context plays, we conduct preliminary experiments in SimWiSense [11] and Widar [4]. Then, we extensively evaluate MULTI-FI on Widar for gesture recognition and on GaitID [8] for gait identification. In addition, we study the effect of the number of views and the robustness of the different fusion strategies to errors in context extraction.

The key contributions ("C") and findings ("F") of our study are summarized as follows:

C1. We develop MULTI-FI, a model enhancement framework tailored to multistatic Wi-Fi sensing. We address two key weaknesses of current SoTA models by leveraging multi-view and context fusion. Results show improvements in every scenario, with an average accuracy gain of 8.2%.

C2. We verify the robustness of MULTI-FI across varying number of receivers and noisy context extraction.

F1. We find that for the majority of Wi-Fi sensing scenarios, the default approach in current models, *input fusion*, is suboptimal. This fusion strategy was inherited from computer vision, where RGB channels are highly correlated pixel-wise; however, this is typically not the case for multistatic wireless sensing. Thus, simply adopting a fusion scheme better-suited to the nature of the problem improves accuracy even in SoTA models.

F2. We find that, irrespective of the fusion strategy employed, all models improve accuracy over baseline when *collaborative sensing* is guided by physical context. Thus, both modules in MULTI-FI can be leveraged independently to improve baseline models.

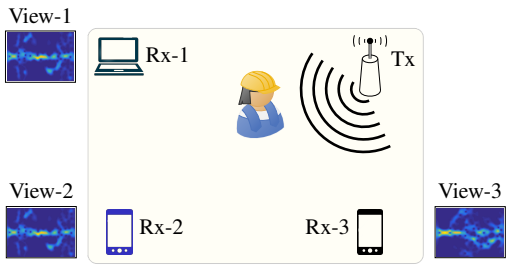


Fig. 1. Exemplary Sensing Scenario

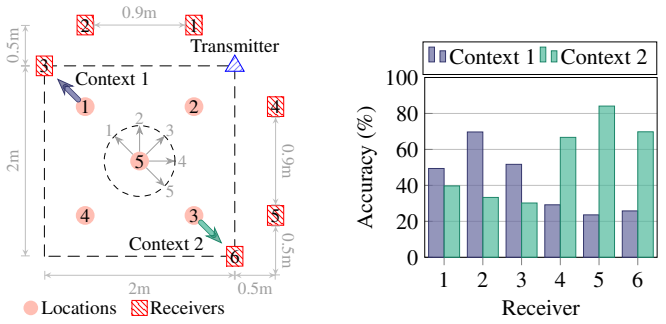
II. BACKGROUND AND MOTIVATION

A. Wi-Fi Sensing Fundamentals

Wi-Fi sensing relies on CSI, Received Signal Strength Indicator (RSSI), or Beamforming Feedback Information (BFI), with CSI being the most widely studied due to its fine-grained characterization of the propagation environment. For sensing, packets are typically transmitted at a fixed rate, forming a time series that captures channel variations over time. Modern Wi-Fi systems employ *Multiple-Input-Multiple-Output* (MIMO) and *Orthogonal Frequency Division Multiplexing* (OFDM), yielding a CSI value for each subcarrier and transceiver antenna pair at every time instant. CSI is acquired by transmitting pre-defined Long Training Symbols (LTFs) symbols in the packet preamble; the receiver estimates the channel by comparing the transmitted and received signals, modeled as $y = Hx + n$, where n denotes white Gaussian noise. Under a scattering model, each element of the CSI matrix H is given by:

$$H(f_c, t) = \sum_{p=1}^{P} \alpha_p(t) e^{-j2\pi f_c \tau_p(t)}, \tag{1}$$

where $\alpha_p(t)$ and $\tau_p(t)$ denote the amplitude attenuation and propagation delay of path p, with f_c the subcarrier frequency and P the number of paths. Then, $H \in \mathbb{C}^{N \times M \times F \times T}$, where N, M, F, and T denote transmit/receive antennas, subcarrier number, and time points per sample correspondingly. As illustrated in Fig. 1, multipath signals are perturbed by static or dynamic obstacles, and their temporal evolution enables environment sensing. Raw CSI suffers from amplitude fluctuations, synchronization errors, and phase instabilities; some works use only $|H|$ [22], others apply calibration or denoising, and many extract features such as Doppler Frequency Shift (DFS) [4], [5], [7], [13] to suppress static background effects. We adopt the DFS approach from [4]; for broader CSI processing methods, see [7], [23].



(a) Widar setup (b) Results

Fig. 2. Results for Widar. (a) The evaluation setup. (b) Shows a cross-receiver accuracy comparison for Context 1 and Context 2.

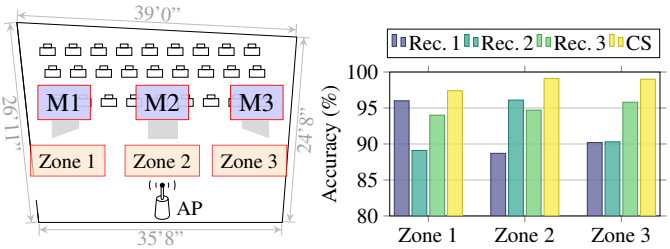
B. Multi-view Classification

Multi-view classification aims to leverage information from multiple domains to improve learning [24], [25]. Data can be homogeneous (e.g., images) or heterogeneous (e.g., text and images); here, we focus on homogeneous data from different perspectives, i.e., views. Success relies on two principles [24]: 1) *consensus*, ensuring consistency across views to extract common features, and 2) *complementarity*, where views provide unique information to boost performance. Fusion strategies vary by depth [25]: *input fusion* concatenates views as input channels; *early* and *late fusion* extract per-view features, merging them either before further feature extraction (early) or directly for prediction (late); *score fusion* treats views independently, combining outputs as an ensemble [26]. Thus, any Wi-Fi sensing scenario with more than one receiver operating simultaneously can be treated as a multi-view problem. While these setups are highly common, to the best of our knowledge, no work has addressed *(L1)*. Hence, we focus on this research question: **How does the multi-view fusion strategy impact the accuracy of a Wi-Fi sensing model?**

C. Motivation: Context Influence on Sensing

The lack of cross-domain generalization limits Wi-Fi sensing adoption, as variations in environment, setup, or user/target position can drastically reduce accuracy [16]. Prior solutions include conformal prediction [16], consistent feature derivation [4], [27], domain adaptation [28], domain generalization [22], and data augmentation [29]. However, regardless of how similar train and test distributions are, sensing performance is influenced by the geometry between transmitter, receiver and target, motivating *collaborative sensing*. MULTI-FI can thus enhance accuracy both in-domain and cross-domain when combined with these approaches.

To show the impact of physical context on sensing, we conduct two experiments. First, we train a ResNet18 [20] on a Widar subset [4] and evaluate receiver accuracy in two distinct contexts: Context 1 comprises samples from location 1, orientation 1 (Fig. 2(a)), and Context 2 from location 3, orientation 5. Fig. 2(b) shows that Context 1 yields higher accuracy for receivers 1–3, while Context 2 favors receivers 4–6, with differences up to 50%, highlighting the role of physical context. Moreover, as Deep Neural Networks (DNNs) are typically overconfident in their predictions [30], they are not a reliable way of assessing view capacity independently.



(a) SimWiSense Setup (b) Results

Fig. 3. Results for SimWiSense. (a) The evaluation setup. (b) shows model accuracy for each receiver in each zone and collaborative sensing (CS), which yields improvements with respect to single receiver in every case.

The second experiment demonstrates the broad relevance of *collaborative sensing* and the impact of physical context on sensing capacity. Using the SimWiSense setup [11], we replicate their proximity experiment with multi-user activity recognition, adding *collaborative sensing*. Fig. 3(a) depicts three users performing simultaneous activities in separate zones, with signals captured by three receivers placed near each zone. Fig. 3(b) shows that (1) the closest receiver achieves the highest accuracy in each zone, highlighting the role of location, and (2) combining all views outperforms any single receiver. This underscores the potential of advanced *collaborative sensing* schemes that account for proper use of contextual information.

Drawing conclusions from these experiments and aiming to solve *(L2)*, we focus on the following research question: **In a *collaborative sensing* scenario, how should we assess and integrate the sensing capacity of different receivers into the DL model to obtain better overall predictions?**

III. MULTI-FI FRAMEWORK

This section details MULTI-FI from a technical perspective. Section III-A covers different multi-view fusion strategies (module ❶, Fig. 4), targeting *L1*; Section III-B describes contextual feature extraction and embedding (module ❷) to enhance *collaborative sensing*, targeting *L2*.

A. Multi-view Fusion

Problem definition. The multi-view classification problem can be mathematically defined as in (2).

$$f : X_V \rightarrow \{1, \ldots, k\} \mid X_V = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(n_V)}\}, \quad (2)$$

where k is the number of classes, n_V is the total number of different views and X_V represents the collection of those available views. However, even when considering the same input and output, there are different ways for the classification model, $f()$, to integrate the complementary views in X_V.

Multi-view strategies. To characterize the different fusion strategies, we show first that any classification model $f()$ can be split in two blocks as in (3).

$$\left.\begin{aligned} z &= \mathcal{C}(x) \\ y &= \mathcal{L}(z) \end{aligned}\right\} \Rightarrow y = f(x) = (\mathcal{L} \circ \mathcal{C})(x), \quad \text{where} \quad y \in \mathbb{R}^{1\times k}. \quad (3)$$

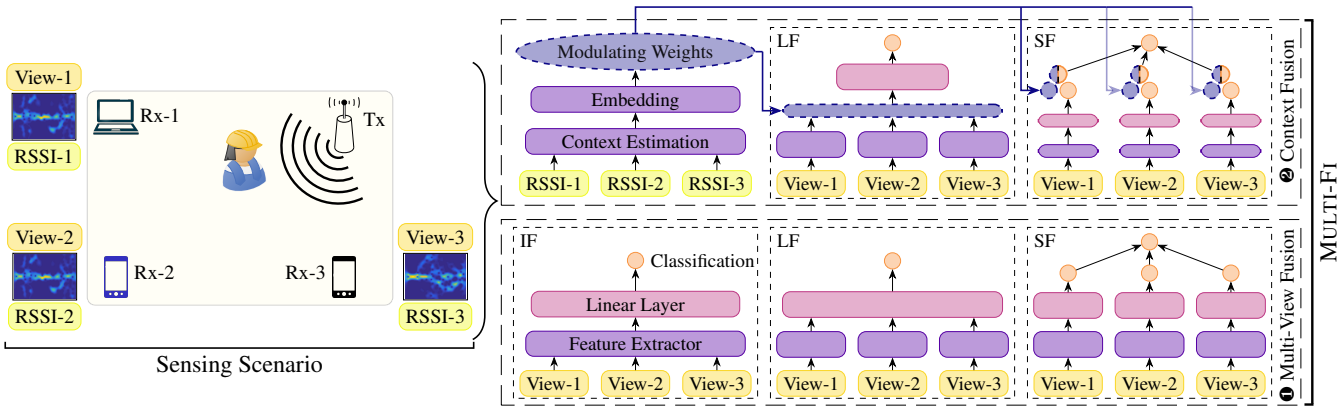


Fig. 4. The architectural design of MULTI-FI

Here, $\mathcal{C}$ represents the "core" of the classification model, which includes all layers until the last linear block. Thus, z represents the convolutional features array for CNNs, the hidden states for GRUs or Transformers, etc. Meanwhile, $\mathcal{L}$ represents the final linear block, which takes the extracted features from the backbone and outputs the prediction array of the classifier, y. With this notation, we can characterize the three tested multi-view fusion strategies as follows:

- **Input Fusion (IF)** merges all the views for the same sample together as different input channels. Then, they are fed to the classification model, $f()$.

$$y_V = f(X_V) = (\mathcal{L} \circ \mathcal{C})(X_V). \tag{4}$$

- **Late Fusion (LF)** extracts features independently for each view. Then, they are concatenated into a single array, and fed to the linear block, $\mathcal{L}$, for a final prediction, y_V:

$$\begin{aligned} z^{(v)} &= \mathcal{C}(x^{(v)}), \quad for \quad v = 1, \ldots, n_V \\ y_V &= \mathcal{L}(z^{(1)} \| z^{(2)} \| \ldots \| z^{(n_V)}). \end{aligned} \tag{5}$$

- **Score Fusion (SF):** processes each view independently until a prediction array is obtained per view, $y^{(v)}$. Then, these individual predictions are combined to obtain a collaborative prediction, y_V. Typically, this step involves cross-view averaging or selecting the class with highest confidence among all views and classes, k^M:

$$\begin{aligned} y^{(v)} &= f(x^{(v)}) \\ y_V &= \begin{cases} \frac{1}{n_V} \sum_v y^{(v)} & \text{opt 1} \\ y^{(v^M)} \mid (v^M, k^M) = \underset{v,k}{\text{argmax}}\, y_k^{(v)} & \text{opt 2.} \end{cases} \end{aligned} \tag{6}$$

The placement of feature extraction (e.g., DFS) modules and classification models is flexible, which is compliant with IEEE 802.11bf standardization guidelines. They can be deployed at client stations, access points (AP), or an external server/cloud, depending on system architecture and use-case requirements. However, in multi-view scenarios, at least some level of centralized computation is typically needed to perform coordination and fusion between views.

B. Context Fusion

This subsection analyzes module ❷ of MULTI-FI (see the top of Figure 4). As discussed in Section II-C, the sensing capacity of a view is influenced by the geometry between transceivers and sensing target, among other variables. Moreover, receivers with poor sensing quality can degrade the overall accuracy by adding noise to the decision process. MULTI-FI overcomes this issue by dynamically modulating view importance, thereby addressing *(L2)*. To assess view importance, it leverages the physical context of the sensing target with respect to the transceivers. Thus, this module consists of two steps: physical context extraction and model fusion.

1) Physical Context Extraction: The first step is to extract the desired physical context features, i.e., location and orientation of the target with respect to the transceivers. For gait identification, we focus on path and direction. Here, we work on a discretized version of space, with limited predefined available locations/orientations in the dataset (see Fig. 2(a)). Thus, instead of leveraging the bulkier CSI data, we opt for RSSI. While CSI yields higher resolution, extracting physical context via RSSI time-series offers two advantages: *(i)* robustness to phase noise, and *(ii)* low computational overhead. However, there exist works proposing alternative solutions [5] for human localization and tracking, with different trade-offs in terms of accuracy and computational overhead. These are orthogonal to MULTI-FI and could be leveraged for context extraction. In fact, any improvements in the context extractor would strengthen MULTI-FI's effect. Without loss of generalization, we use a time series of RSSI values for each receiver, which we feed to a lightweight time series classifier, MiniRocket [31]. Thus:

$$L/O = g(\text{RSSI}), \quad \text{where} \quad \text{RSSI} \in \mathbb{R}^{T \times n_V}, \tag{7}$$

where L/O represents physical context features and $g()$ is the time series classifier. T represents the temporal dimension, defined as the actual duration time of the sample (s), multiplied by the specified rate for sensing (r), i.e., $T = r \cdot s$.

2) Physical Context Fusion: Upon extraction of physical context features, MULTI-FI integrates them into the classification model. The previous step yields a categorical variable without intrinsic meaning (the label numbers do not translate to

any real physical measure); thus, a more meaningful representation is needed. For this, MULTI-FI leverages an embedding layer, whose output is a multi-dimensional representation of the context variables. It is a matrix of trainable weights, $E \in \mathbb{R}^{L/O \times d}$, which takes the categorical context label and returns its d-dimensional representation, where d is a design parameter. Then, these embeddings go through a linear block, $\mathcal{M}()$, to connect them with the modulating weights, w. Finally, those weights are multiplied element-wise with the corresponding elements in the classifier. Next, we describe how this step applies to each multi-view strategy, as depicted in the top part of Fig. 4.

Score fusion. Here, the classifier is left intact. The weights, w, modulate the prediction arrays for each of the receivers before their combination. That is: *1)* Each view gets an independent prediction from the classifier. *2)* These prediction arrays are modulated via element-wise multiplication with the weights, w. *3)* The new arrays are averaged to get a definitive prediction.

$$\left.\begin{aligned} y^{(v)} &= f(x^{(v)}) \\ w &= \mathcal{M}(E(l)) \end{aligned}\right\} y^{(v)*} = w \odot y^{(v)} \Rightarrow y_V = \frac{1}{n_v} \sum_v y^{(v)*}. \tag{8}$$

Where E is the embedding layer and $l \in [1, \ldots, L]$ is the context label of the sample. Meanwhile, $y^{(v)*}$ is the obtained modulated version of the prediction array for a given view, v.

Late fusion. Here, as the different views are fused within the model, we do not leave the original model intact anymore. We get the embeddings of context features and their corresponding modulating weights, w, in parallel to the classification model. Then, these weights adjust the importance of each view by altering their set of extracted features, $z^{(v)} \Rightarrow z^{(v)*}$. Finally, these modulated features are fused together and go through the last block of linear layers, as they would in the original *late fusion* version of the model.

$$\left.\begin{aligned} z^{(v)} &= \mathcal{C}(x^{(v)}) \\ w &= \mathcal{M}(E(l)) \end{aligned}\right\} z^{(v)*} = w \odot z^{(v)} \Rightarrow y_V = \mathcal{L}(z^{(1)*} \| \ldots). \tag{9}$$

Input fusion. While input fusion strategies can integrate contextual data, they introduce systemic inefficiencies that outweigh its potential accuracy benefits. Unlike late or score fusion, which makes it possible parallel execution, input fusion requires a sequential pipeline that halts classification until input-channel modulation is complete. This creates significant overhead, scaling operations from low-dimensional vectors ($\approx 10^0 - 10^1$ elements) to higher-dimensional tensors (e.g., x has size $n_V \times T \times 121$, where $T \approx 10^2$–10^3); thus, compromising real-time inference. Furthermore, input-level weighting suffers from feature dilution; modulating raw signals before the network distinguishes sensing features from environment-specific noise is less effective than prioritizing the high-level semantic embeddings utilized in later stages. Therefore, we exclude context fusion for *input fusion* models.

IV. EVALUATION

A. Implementation Details

Datasets. To test the effectiveness of MULTI-FI, potential data sets must satisfy two requirements: *1)* At least two receivers operating simultaneously to ensure *collaborative sensing* capacity. *2)* Different positions of the sensing target, allowing us to assess the influence of physical context on the sensing capacity. We also aim to test MULTI-FI for different applications. These guidelines led us to: Widar [4] for gesture recognition and GaitID [8] for gait identification. Both datasets employ 1 transmitter and 6 receivers. Widar contains gestures performed by 17 people across 3 rooms in 5 different locations and orientations (see Fig. 2(a) for exact positions). However, to ensure consistency with previous works, we limit the pool of gestures to: *push&pull*, *sweep*, *clap*, *slide*, *draw-O* and *draw-Z*. GaitID includes data collected from 2 rooms and 11 users. Here, users walk a number of times through each of the 4 defined paths in the 2 possible directions. For both datasets, the input data fed to the classification models is DFS, which is provided by the dataset authors.

Models. We experiment with 5 popular Wi-Fi sensing models, with diverse architectures, to validate MULTI-FI as a model-agnostic contribution. We also favor models with open source codes, to prevent implementation mismatches. We include a version of Widar3.0 (Wi) [4], a model tailored to Wi-Fi sensing which incorporates CNNs followed By GRUs. We also choose WiFlexFormer (WF) [18], a transformer-based architecture specifically developed for sensing purposes. Additionally, we borrow three staples from different fields, with a proven record for sensing tasks, as demonstrated in [19]: ResNet18 (RN18) [20], MobileNetV3 (MN) [21] and BiLSTM (Bi)[19].

Experimental design. We train each model version for five runs and report the mean and standard deviation. All models have been trained for 40 epochs with a step decay scheduler and a cross-entropy loss function. The initial learning rate is 0.001 and the optimizer is Stochastic Gradient Descent (SGD). Experiments have been performed with a NVIDIA A100 80GB GPU, using the PyTorch framework. Code using alternative frameworks (e.g., SimWiSense preliminary experiment used TensorFlow) was converted, aiming to preserve the original characteristics. For CSI preprocessing, we follow the instructions and code in [4]. Additionally, the number of MiniRocket kernels used for context extraction is 3×10^4. This value is within the tested range in the original work [31] and chosen via hyperparameter search. Kernel length is fixed to 11, which is the recommended value by the authors. Additionally, due to discrepancies in RSSI sample length, it is fixed to $T = 2000$.

B. Results

Table I shows the results for the five tested models for each fusion strategy, with and without context. MULTI-FI boosts accuracy across all scenarios by leveraging a proper multi-view fusion strategy, enhanced via context fusion. The average accuracy gain is 8.2%, reaching up to 29.6% for Widar on the

TABLE I
ACCURACY RESULTS. UNDERLINE IS ORIGINAL MODEL AND **BOLD** IS AFTER MULTI-FI. MEAN±STD (%)

Model	Dataset	Input Fusion	Late Fusion		Score Fusion		MULTI-FI ΔAcc.
			No Context	Context	No Context	Context	
ResNet18	Widar	88.1±1.2	83.4±1.4	88.9±1.6	89.9±0.4	**90.8**±0.7	2.7
	GaitID	98.5±0.3	98.0±0.7	**98.8**±0.3	94.6±1.3	95.5±1.3	0.3
MobileNet	Widar	93.6±0.3	95.7±0.6	96.9±0.4	96.8±0.3	**97.1**±0.2	3.5
	GaitID	94.7±0.6	96.5±0.8	**98.4**±0.5	88.2±2.1	88.9±2.8	3.7
BiLSTM	Widar	89.5±1.0	91.0±1.1	92.1±1.5	95.6±0.5	**96.2**±0.6	6.7
	GaitID	70.8±4.0	84.8±3.1	**87.2**±3.2	75.0±10.3	81.4±8.2	16.4
Widar	Widar	84.5±8.2	91.3±0.6	91.6±0.3	92.6±0.8	**93.5**±0.5	9.0
	GaitID	60.1±4.4	89.2±0.7	89.5±1.2	85.2±1.4	**89.7**±1.8	29.6
WiFlexFormer	Widar	88.3±0.9	57.7±16.9	61.3±12.3	90.2±2.3	**91.3**±1.6	3.0
	GaitID	84.3±3.3	87.1±10.4	**91.5**±7.2	85.5±6.0	88.7±5.4	7.2

GaitID dataset. We note that MULTI-FI does not just improve on weak baselines, but on very well performing models as well. Next, we study the contribution of each module individually.

1) Multi-view Fusion Results: Here, we compare the results of the three fusion strategies, ignoring context fusion (see *input fusion* column and "No Context" columns for *score* and *late fusion* from Table I). As expected, no single fusion strategy outperforms the others across models and datasets, a result consistent with previous research on different applications [26]. Nonetheless, we are able to extract valuable insights from the data. An initial takeaway is that the choice of fusion strategy greatly influences model accuracy. In fact, the average difference across models and datasets between best and worst performing alternatives amounts to 8.0%, reaching up to 29.1%. Thus, the choice of an appropriate fusion strategy is a crucial system variable which shouldn't be overlooked. As wireless sensing models are typically either inspired (leveraging CNNs and Vision Transformers (ViT)) or adopted from computer vision, they tend to process different views as different input channels. This is the natural format of an image (RGB channels) however, it is not trivial to consider this optimal for other applications, including wireless sensing. In fact, all five models tested here (recall these are widely used models and datasets) originally follow an *input fusion* strategy which, according to our results, fails to be the best alternative in 9 out of 10 scenarios. Anticipating the optimal fusion strategy *a priori* is a daunting task since the interaction between model, sensing task, and dataset peculiarities lead to a coupled, non-convex optimization landscape. To the best of our knowledge, no research work has provided a rigorous way of handling this *a priori* decision. Current SoTA approaches propose Neural Architecture Search (NAS) [32], [33] schemes to avoid an exhaustive search. Despite this, we observe that in Widar *score fusion* performs the best on all five models, while in GaitID *late fusion* is the best choice in four out of five models. We attribute this difference to the difficulty of the task. In Widar there are only 6 different classes and individual views perform reasonably well (low bias), thus *score fusion* excels, as it offers the highest robustness to noise (low variance). Contrary, GaitID presents a harder task, with 11 classes and lower individual accuracy. Here, *score fusion* yields higher bias and the capacity of *late fusion* to extract cross-view patterns proves useful at lowering it, despite a higher propensity to overfitting.

2) Context Fusion Results: We examine the impact of context fusion single-handedly. Firstly, we state that MiniRocket achieves 99% and 98% accuracy for context extraction in Widar and GaitID, respectively. Next, we discuss the results for the context fusion module (❷) which are shown in Table I.

Score Fusion. We present the results in the "Context" column of *score fusion* in Table I. Our approach modulates view importance, in lieu of taking the usual approaches described in (6). Thus, we benchmark our results against the best performing option between those: mean and max. confidence across views ("No Context"). All models benefit from the proposed module in both datasets, although to different extents. The results show an average accuracy improvement of 2.0% across models and datasets.We notice that models with high initial accuracy exhibit more modest improvements. Nevertheless, this is an expected finding since these models already had more correct views per sample and, thus, view weighting becomes less relevant. Additionally, we observe that cross-view mean typically outperforms max. confidence. This finding supports the notion that DNN confidences alone are not a reliable approach for assessing receiver sensing capacity.

Late Fusion. We show these results in the "Context" column of *late fusion* in Table I. Our baseline is the same model with the same fusion strategy but without context modulation ("No Context"). Our results show an average accuracy boost of 2.2% across models and datasets. As in *score fusion*, MULTI-FI's context module is able to further increase the accuracy gains provided by a correct view fusion scheme. We highlight this concept, as improving accuracy becomes harder the better performing the original model is.

C. Micro-benchmarks

Impact of Number of Receivers. We vary the number of available views. Instead of using all six receivers, we limit the views to the ones from receivers 1, 3 and 5 from Figure 2(a). The selection criteria for the remaining receivers focuses on maintaining a balanced setup and preserving sensing coverage; favoring receivers on both sides of the symmetry axis and placed at varying distances from the Tx. We use a *score fusion* strategy on GaitID and compare the effect of context integration on accuracy. Figure 5(a) portrays the complete results. We observe an average improvement of 3.8% against the mean and 7.6% against max. confidence. These results support the improvements provided by context fusion, irrespective of the number of views used for *collaborative sensing*.

Added Complexity of Context Fusion. We estimate the extra computation necessary for context fusion. For this, we use ResNet18 on Widar and evaluate the extra trainable parameters and FLOPS induced by the context module, including both context extraction and fusion. For *score fusion*, we find a 1.5% increase in trainable parameters and a 1.1% increase in extra FLOPS. For *late fusion*, we observe similar numbers: 1.4% and 1.1%, respectively. These numbers are well within the natural runtime noise of batch training, showcasing how little extra complexity MULTI-FI adds.

Context Error Robustness. To assess the robustness of MULTI-FI under conditions where the context extraction may be hindered, we perform the following experiment on GaitID for both *late* and *score fusion*. Firstly, we select a certain ratio of samples from the test set and randomly change their context prediction. Then, we study the effect on the overall accuracy of the system. We test with context error rates ranging from 0(MULTI-FI context predictions) to 1 (totally wrong predictions); for each error rate we average the results across 5 random seeds. We compare each model against its baseline version (dashed lines, "No Cxt"); note that since baseline models ignore context information, they are unaffected by varying error rates. Results are shown in Figure 5. As expected, less accurate context predictions lead to lower system accuracy. Nonetheless, we observe differences in this effect depending on the fusion strategy employed. For *score fusion*, context is beneficial as evidenced by the downward trend, but even when considering wrong context predictions only, accuracy remains above baseline. This is because cross-view mean (baseline) does not allow enough flexibility for the model to differentiate between classes. Thus, for *score fusion*, MULTI-FI improves accuracy via two mechanisms: 1) View weighting via context (subject to accurate context extraction). 2) Class weighting, learning to balance confidence among classes (e.g., class 3 tends to exhibit overconfidence, so lower weights are assigned to their logits). For *late fusion*, results differ; for most models an error rate of 0.2 already degrades accuracy below baseline. This result indicates that *score fusion* versions are more resilient to unstable conditions and/or weak context extractors.

V. RELATED WORK

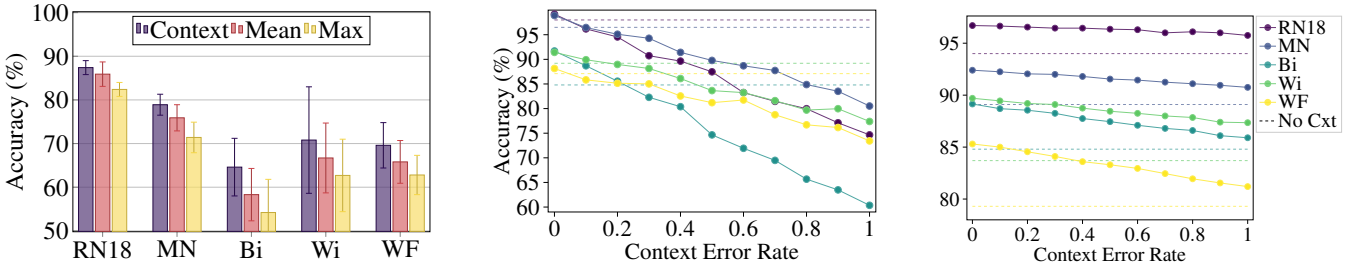
Wi-Fi Sensing Datasets. Next, we outline a brief review on publicly available datasets for Wi-Fi sensing. Besides the renowned Widar dataset [4], there are several options for gesture and activity recognition. We highlight XRF55 [34] due to its large size and combination of data types: RFID, mmWave, infrared, etc. We highlight the work in [35], creating a dataset with the IEEE 802.11ax standard and anonymized video ground truth recordings. For gait identification, options are more limited; we highlight CAUTION [36]. Recent datasets are tailored towards more complex goals: MM-Fi [37] targets multi-modal sensing for 4D human perception, SimWiSense [11] targets simultaneous multi-user sensing and OctoNet [38] captures human motion across several different points in the frequency spectrum and biometrics.

Wi-Fi Sensing Models. Regarding activity and gesture recognition, a plethora of models is available. SenseFi [19] conducts a benchmark for different architectures. SLNet [13] proposes a complex-valued DNN with spectrogram generation. Models in [39] and THAT [40] are based on the transformer architecture. OneSense [12] focus on the one-shot scenario. Meanwhile, WiHF [27] targets gesture recognition and human identification simultaneously. HybridZone [2] merges acoustic and Wi-Fi signals. Less alternatives have been developed for gait identification. AutoFi [41] follows a geometric self-supervised learning approach. XGait [42] uses Inertial Measurement Unit (IMU) signals to simulate RF samples for training enhancement. Meanwhile, Wi-PIGR [43] works on the generalization of gait identification across unknown paths.

Multi-view Learning. There is growing interest in the field, with survey papers covering the area [24], [25]. Different fusion strategies have been explored across various contexts: [26] evaluates ResNet models across three domains, while [14] approaches multi-view binary classification from a mathematical perspective. For CNN-Transformer architectures, [44] compares standard fusion methods against a distillation-based hybrid approach. Furthermore, BM-NAS [33] and DC-NAS [32] employ NAS to efficiently identify optimal multi-view strategies without resorting to brute-force methods.

VI. CONCLUSIONS

In this paper, we propose MULTI-FI, a novel framework to enhance model design specifically tailored to Wi-Fi sensing. Our framework addresses two key limitations present in previous models (*(L1)* and *(L2)*), and consequentially improves sensing accuracy across the board. Firstly, it selects an appropriate multi-view fusion scheme. Then, it leverages the physical context of the sensing target, to modulate the importance of each view for *collaborative sensing*. We validate MULTI-FI across several diverse SoTA models, datasets and applications, exhibiting accuracy improvements in every tested scenario. In fact, MULTI-FI achieves an average improvement of 8.2%, reaching up to 29.6% improvement in the best-case scenario, on already well performing models.



(a) Varying Nº of available receivers

(b) Context Error: *late fusion*

(c) Context Error: *score fusion*

Fig. 5. Robustness micro-benchmarks.

ACKNOWLEDGMENTS

This work has been partially funded by project TUCAN6-CM (TEC-2024/COM-460), funded by the Madrid Regional Government (ORDEN 5696/2024). Claudio Fiandrino is a Ramón y Cajal awardee (RYC2022-036375-I), funded by MCIU/AEI/10.13039/501100011033 and the ESF.

REFERENCES

[1] Wi-Fi Alliance. (2025) Discover wi-fi. [Online]. Available: https://www.wi-fi.org/discover-wi-fi

[2] M. Li *et al.*, "Hybrid Zone: Bridging acoustic and Wi-Fi for enhanced gesture recognition," in *Proc. of IEEE INFOCOM*, 2024, pp. 981–990.

[3] R. Du *et al.*, "An overview on ieee 802.11 bf: Wlan sensing," *IEEE Communications Surveys & Tutorials*, vol. 27, no. 1, pp. 184–217, 2024.

[4] Y. Zheng *et al.*, "Zero-effort cross-domain gesture recognition with Wi-Fi," in *Proc. of ACM MobiSys*, 2019, pp. 313–325.

[5] D. Wu *et al.*, "WiTraj: Robust indoor motion tracking with WiFi signals," *IEEE Trans. on Mobile Computing*, vol. 22, no. 5, 2021.

[6] F. Adib *et al.*, "Smart homes that monitor breathing and heart rate," in *Proc. of ACM CHI*, 2015, pp. 837–846.

[7] F. Meneghello *et al.*, "Sharp: Environment and person independent activity recognition with commodity IEEE 802.11 access points," *IEEE Trans. on Mobile Computing*, 2022.

[8] Y. Zhang *et al.*, "GaitID: Robust Wi-Fi based gait recognition," in *Proc. of WASA*. Springer, 2020, pp. 730–742.

[9] X. Li *et al.*, "Optimal ai model splitting and resource allocation for device-edge co-inference in multi-user wireless sensing systems," *IEEE Trans. on Wireless Communications*, 2024.

[10] G. Yin *et al.*, "FewSense, towards a scalable and cross-domain Wi-Fi sensing system using few-shot learning," *IEEE Trans. on Mobile Computing*, vol. 23, no. 1, pp. 453–468, 2022.

[11] K. F. Haque *et al.*, "SimWiSense: Simultaneous multi-subject activity classification through Wi-Fi signals," in *Proc. of IEEE WoWMoM*, 2023, pp. 46–55.

[12] L. Zhao *et al.*, "One is enough: Enabling one-shot device-free gesture recognition with cots WiFi," in *Proc. of IEEE INFOCOM*, 2024, pp. 1231–1240.

[13] Z. Yang *et al.*, "{SLNet}: A spectrogram learning neural network for deep wireless sensing," in *Proc. USENIX NSDI*, 2023, pp. 1221–1236.

[14] L. M. Pereira *et al.*, "A comparative analysis of early and late fusion for the multimodal two-class problem," *IEEE Access*, vol. 11, 2023.

[15] X. Wang *et al.*, "Placement matters: Understanding the effects of device placement for WiFi sensing," *Proc. of the ACM IMWUT*, 2022.

[16] K. Wang *et al.*, "Solving the WiFi sensing dilemma in reality leveraging conformal prediction," in *Proc. of ACM SenSys*, 2022, pp. 407–420.

[17] A. Blanco *et al.*, "Augmenting mmWave localization accuracy through sub-6 GHz on off-the-shelf devices," in *Proc. of ACM MobiSys*, 2022.

[18] J. Strohmayer *et al.*, "WiFlexFormer: Efficient WiFi-based person-centric sensing," *arXiv preprint arXiv:2411.04224*, 2024.

[19] J. Yang *et al.*, "SenseFi: A library and benchmark on deep-learning-empowered WiFi human sensing," *Patterns*, vol. 4, no. 3, 2023.

[20] K. He *et al.*, "Deep residual learning for image recognition," in *Proc. of IEEE CVPR*, 2016, pp. 770–778.

[21] A. G. Howard *et al.*, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[22] D. Wang *et al.*, "AirFi: Empowering WiFi-based passive human gesture recognition to unseen environment via domain generalization," *IEEE Trans. on Mobile Computing*, vol. 23, no. 2, pp. 1156–1168, 2022.

[23] Y. Ma *et al.*, "WiFi sensing with channel state information: A survey," *ACM Computing Surveys*, vol. 52, no. 3, pp. 1–36, 2019.

[24] Z. Yu *et al.*, "A review on multi-view learning," *Frontiers of Computer Science*, vol. 19, no. 7, p. 197334, 2025.

[25] X. Yan *et al.*, "Deep multi-view learning methods: A review," *Neurocomputing*, vol. 448, pp. 106–129, 2021.

[26] M. Seeland *et al.*, "Multi-view classification with convolutional neural networks," *Plos one*, vol. 16, no. 1, 2021.

[27] C. Li *et al.*, "WiHF: Enable user identified gesture recognition with WiFi," in *Proc. of IEEE INFOCOM*, 2020, pp. 586–595.

[28] J. Chen *et al.*, "LAGER: Label-free domain-adaptive wireless gesture recognition via latent feature alignment and augmentation," *IEEE Internet of Things Journal*, 2024.

[29] G. Chi *et al.*, "RF-diffusion: Radio signal generation via time-frequency diffusion," in *Proc. of ACM MobiCom*, 2024, pp. 77–92.

[30] C. Guo *et al.*, "On calibration of modern neural networks," in *Proc. of ICML*, 2017, pp. 1321–1330.

[31] A. Dempster *et al.*, "Minirocket: A very fast (almost) deterministic transform for time series classification," in *Proc. of ACM SIGKDD*, 2021, pp. 248–257.

[32] X. Liang *et al.*, "Dc-nas: Divide-and-conquer neural architecture search for multi-modal classification," in *Proc. of the AAAI conference on artificial intelligence*, vol. 38, no. 12, 2024, pp. 13 754–13 762.

[33] Y. Yin *et al.*, "Bm-nas: Bilevel multimodal neural architecture search," in *Proc. of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 8, 2022, pp. 8901–8909.

[34] F. Wang *et al.*, "Xrf55: A radio frequency dataset for human indoor action analysis," *Proc. of ACM IMWUT*, vol. 8, pp. 1–34, 2024.

[35] M. Cominelli *et al.*, "Exposing the CSI: A systematic investigation of CSI-based Wi-Fi sensing capabilities and limitations," in *Proc. of IEEE PerCom*, 2023, pp. 81–90.

[36] D. Wang *et al.*, "CAUTION: A robust WiFi-based human authentication system via few-shot open-set recognition," *IEEE Internet of Things Journal*, vol. 9, no. 18, pp. 17 323–17 333, 2022.

[37] J. Yang *et al.*, "Mm-fi: Multi-modal non-intrusive 4d human dataset for versatile wireless sensing," *NeurIPS*, vol. 36, pp. 18 756–18 768, 2023.

[38] D. Yuan *et al.*, "Octonet: A large-scale multi-modal dataset for human activity understanding grounded in motion-captured 3d pose labels," in *NeurIPS Datasets and Benchmarks Track*, 2026.

[39] F. Luo *et al.*, "Vision transformers for human activity recognition using WiFi channel state information," *IEEE Internet of Things Journal*, 2024.

[40] B. Li *et al.*, "Two-stream convolution augmented transformer for human activity recognition," in *Proc. of the AAAI conference on artificial intelligence*, vol. 35, no. 1, 2021, pp. 286–293.

[41] J. Yang *et al.*, "Autofi: Toward automatic Wi-Fi human sensing via geometric self-supervised learning," *IEEE Internet of Things Journal*, vol. 10, no. 8, pp. 7416–7425, 2022.

[42] H. Yang *et al.*, "XGait: cross-modal translation via deep generative sensing for RF-based gait recognition," in *Proc. of ACM SenSys*, 2023.

[43] L. Zhang *et al.*, "Wi-PIGR: Path independent gait recognition with commodity Wi-Fi," *IEEE Trans. on Mobile Computing*, vol. 21, no. 9, pp. 3414–3427, 2021.

[44] S. Black *et al.*, "Multi-view classification using hybrid fusion and mutual distillation," in *Proc. of the IEEE/CVF WACV*, 2024, pp. 270–280.