Solving the WiFi Sensing Dilemma in Reality Leveraging Conformal Prediction

Kailong Wang
WINLAB, Rutgers University
kailong.wang@rutgers.edu

Cong Shi
WINLAB, Rutgers University
cs1421@scarletmail.rutgers.edu

Jerry Cheng
New York Institute of Technology
jcheng18@nyit.edu

Yan Wang
Temple University
y.wang@temple.edu

Minge Xie
Rutgers University
mxie@stat.rutgers.edu

Yingying Chen
WINLAB, Rutgers University
yingche@scarletmail.rutgers.edu

ABSTRACT

With the wide deployment of smart environments and IoT devices, WiFi sensing has demonstrated its great convenience and contactless sensing capabilities in supporting a broad array of applications. However, designing a ubiquitous WiFi sensing system for heterogeneous scenarios in practice is still a big dilemma as the system performs poorly when the testing data is significantly different from the training data caused by domain variations. To address this dilemma, existing studies involve extra efforts to develop new features or even to retrain the original model under environmental variations. However, none of them can resolve the dilemma completely. In this work, we conduct a comprehensive study on the domain variation problem to make WiFi sensing robust and accurate in reality. Our definition of domains is comprehensive and includes environments, surrounding settings, user differences, user's facing directions, user's positions relative to WiFi sensors, and user participating time frames. Our innovation is to achieve reliable WiFi sensing across all the domains based on the conformal prediction framework. Our approach quantifies the conformity (i.e., similarity) between the testing WiFi samples and the training samples, then labels the testing samples with the most probable class(es). We develop a novel cross-domain transformal prediction scheme based on the multivariate kernel density estimation to effectively assess and learn the conformity of each domain in the training data. To meet various application-specific requirements, we further develop two approaches to fuse the knowledge of conformity derived from the training domains to perform predictions. Extensive experiments with both self-collected and public datasets show that our framework can improve prediction accuracies from 30% to 74% improvements in three most representative WiFi-based applications across six types of domain variations.

CCS CONCEPTS

• **Human-centered computing → Human computer interaction (HCI)**.

SenSys '22, November 6–9, 2022, Boston, MA, USA

ACM ISBN 978-1-4503-9886-2/22/11…$15.00
https://doi.org/10.1145/3560905.3568529

KEYWORDS

WiFi Sensing, Domain Variations, Conformal Prediction

ACM Reference Format:
Kailong Wang, Cong Shi, Jerry Cheng, Yan Wang, Minge Xie, and Yingying Chen. 2022. Solving the WiFi Sensing Dilemma in Reality Leveraging Conformal Prediction. In *ACM Conference on Embedded Networked Sensor Systems (SenSys '22), November 6–9, 2022, Boston, MA, USA.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3560905.3568529

1 INTRODUCTION

Mobile sensing is the core enabler of the wider deployment of smart environments and IoT devices. Among various sensing modalities in mobile devices, wireless signals, and the especially pervasive WiFi signals, have demonstrated their great convenience and sensing capability in practice due to their widely deployed infrastructure and non-intrusive characteristics. Existing studies have designed many WiFi sensing systems for various applications, including activity/gesture recognition [21, 35], vital sign monitoring [16], and user identification [27, 31], etc. While these systems can provide promising results under specific conditions, extending such WiFi sensing systems as ubiquitous solutions for heterogeneous practical scenarios, such as different environments and devices, various participants and participating time frames, etc., is still a challenging and difficult task, when there are differences, known as *domain variations*, between training and testing samples. This challenge is known as the *domain variation problem*, which is one of the most critical research problems in WiFi sensing.

It is common that the same WiFi sensing system will be deployed across different environments (e.g., different rooms and buildings). Even when the WiFi sensing system only needs to operate in a single environment, it still suffers from many problems caused by domain variations such as furniture movements, users' facing directions and positions changes. Note that it is hard for a user to keep the same facing direction and position in every scenario. Such minute differences are also considered as domain variations in reality. Some domain variation problems in WiFi sensing systems have been studied. Features with low correlations to the environments have been developed to achieve high gesture recognition accuracy across different rooms, facing directions, and positions [35]. EI [11] is an activity recognition framework that can work across different environments using domain adaptation. Data augmentation methods have been proposed to generate synthetic or virtual training samples [4] or to reuse knowledge [8] from different tasks to improve the generalization ability of a WiFi-based activity recognition

across devices and subjects. Adversarial learning technique has also been used to remove unpredictable environment-specific factors to perform user authentication across different furniture placements and users' positions in a room [19]. Overall, these studies only focus on a limited number of domains and still have dilemma when facing various domains in reality. They also require extra effort to develop new features or retrain the model when environment changes. It remains difficult to widely deploy such systems.

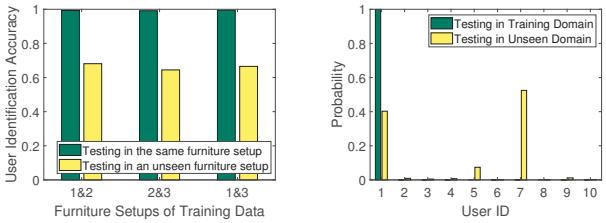
In contrast to the existing work, we conduct a comprehensive study on the domain variation problem and design a robust and accurate WiFi sensing framework under heterogeneous domain variations. Specifically, we define the following domain variation categories that most WiFi sensing systems are subject to in reality. *Environment:* WiFi sensing systems usually exploit pervasive WiFi infrastructures in indoor environments, when deploying these systems in reality, the same WiFi sensing system is likely to be used across different environments. These environments have a variety of physical characteristics, such as room sizes, layouts, and building structures. *Setting:* Furthermore, the settings within the same indoor environment are subject to change from time to time. For instance, different placements of furniture and sensing devices may cause different patterns in reflections and dispersion of WiFi signals. *User:* Most of the WiFi sensing applications involve human subjects. Therefore, variations due to human subjects with different different physiological and behavioral traits are common in WiFi sensing applications. *User's Facing Direction:* In reality, user's facing directions are changing dynamically in WiFi sensing applications. Although minor in scale, such domain variations are especially challenging because they are unpredictable and cannot be well addressed without extensive retraining efforts [5]. *User's Position:* A user can be at different locations or proximate positions from the trained location/position in a room, resulting in complicated domain variations in the relative positions to WiFi sensing devices. *Timelines:* Dynamic varying temperatures, humidity, and hardware states may also render wireless channel conditions unstable across different timelines [9, 17, 18]. Overall, the domain variations can result in changing multi-path effects and bringing noises into wireless signals. They will also lead to fluctuating patterns in fine-grained WiFi signal measurements (e.g., channel state information (CSI)), thereby causing signal profile mismatches and degraded sensing performance in various WiFi sensing applications. In this study, we focus on three most critical WiFi sensing applications: user identification, activity recognition, and gesture recognition, across different environments or within the same environment. These applications are the essential components of a broad spectrum of mobile applications in practice, including mobile healthcare, smart home, and Internet of Things.

To address the domain variation problem, we develop a low-effort framework to achieve reliable and accurate WiFi sensing across multiple domains in practical deployment. Our system resorts to conformal prediction [25] to determine the conformity (i.e., similarity) between the source data (training data) and target data for predictions based on a quantification metric derived from the training data. The basic idea is to leverage WiFi signals from a few domains (i.e., two or more) to assess the conformity of the testing WiFi signals, which may be from an unseen domain. Compared to existing machine-learning-based approaches, conformal prediction is a non-parametric approach to handling shifts in WiFi signals. It can achieve classification results without the need for generating new features or retraining under domain variations.

However, realizing such a practical WiFi sensing framework is challenging. The success of conformal prediction is built upon an effective metric to quantify conformity. Existing learning-based measurement algorithms assume the training and testing data to be identically and independently distributed (i.i.d.). This assumption is no longer valid under domain variations. Developing a quantification metric resilient to domain variations is necessary to realize conformal prediction. In addition, different from traditional classification techniques that output a single class label given an input, conformal prediction resolves the prediction dilemma in reality through performing prediction on each class and producing a set of class labels as output. The prediction accuracy and the size of the predicted set should be balanced based on applications.

Specifically, we design a scheme based on kernel density estimation (KDE) to assess and learn the conformity across domains in the training data. Compared to other learning-based algorithms, which rely on a well-defined mapping relationship between training and testing data (i.e., i.i.d assumption), our scheme leverages the conformity learned from the training data to quantify domain variations in the testing inputs. The cross-domain relationships derived from the training data relieve our framework from the i.i.d assumption and are statistically more reliable under different domain variations. To meet the application-specific requirements on the accuracy and size of the class set, we develop two approaches to fuse the knowledge of conformity derived from training domains, with priorities on maximizing the accuracy and minimizing the number of classes. We summarize the contributions of our work as follows:

- We conduct a comprehensive study of the domain variation problems in various WiFi sensing applications and show the feasibility of achieving high WiFi sensing performance across typical domains in real deployment without requiring extra efforts for collecting new data, generating new features, or retraining prediction models.
- In contrast to existing studies, we develop a holistic WiFi sensing framework using conformal prediction that can ensure high prediction accuracy when facing different domain variations in reality. We develop novel kernel density-based nonconformity measure and cross-domain conformal prediction with two fusion approaches that can more accurately determine the most possible class(es) of the input data.
- We realize the proposed framework for typical WiFi sensing applications (i.e., user identification, activity classification, and gesture recognition) to perform a thorough study on the effectiveness of our framework with domain variations of six categories: environments, settings, user, user's facing directions, user's positions, and timelines.
- We conduct comprehensive experiments with both self-collected and public WiFi sensing datasets. The results validate that our conformal prediction-based framework can effectively mitigate cross-domain errors and improve the prediction accuracies from 30% to 74% in three WiFi sensing applications with domain variations in six categories.



(a) User identification accuracy for the training and unseen domains

(b) Probabilities for user 1's CSI data in the training and the unseen domain

Figure 1: Impacts of domain variations to deep-learning-based user identification with the CSI data collected in three furniture settings (i.e., setups 1, 2, 3).

2 BACKGROUND AND PRELIMINARY

2.1 Deep-learning-enabled WiFi Sensing

WiFi techniques have been used in a multitude of mobile and IoT devices, such as voice assistants, smart refrigerators, and laptops, to connect the devices and exchange data. Specifically, the channel state information (CSI) of WiFi describes how the wireless signals propagate over multiple orthogonal frequency division multiplexing (OFDM) subcarriers between a pair of devices. The CSI is captured during WiFi signal propagation with the combined effects of scattering, fading, and multi-path. As a result, it contains information of human bodies, motions, and surrounding environment (e.g., furnitures, walls, etc.) As CSI is readily available on most current WiFi systems (e.g., 802.11n and its successors), significant research efforts have been devoted to investigating using CSI for sensing applications. Among them, activity recognition [30], gesture recognition [2], and user identification [20] are the three most critical applications. The key idea of WiFi sensing is to extract discriminative features from the CSI measurements in order to capture characteristics of the involved activities and human subjects. The features are then fed to deep learning models to train with a set of target classes. With the strong capabilities of modeling both linear and non-linear mapping relationships, deep learning models often significantly outperform traditional machine learning models and human-craft analytical methods [15].

2.2 Problem Scope

Domain Variations in WiFi Sensing. Despite the promising results of deep learning, current studies have found that the wireless sensing approaches are susceptible to domain variations [11, 19]. In the context of WiFi sensing, a domain is defined as an impacting factor of the signal patterns of CSI. Due to the omnidirectional signal propagation, the CSI captures substantial information specific to these impacting factors. Changes in any of them, which we refer to as *domain variations*, will result in the *data distribution drifts*. In this paper, we aim to explore using statistical assessment metrics to quantify such data distribution drifts and improve the robustness of WiFi sensing. Instead of extracting new features and updating the model's weights, our approach uses these metrics to leverage intermediate results of the deep learning models (representations) for the quantification of domain variations. We focus on the following domain variations in six categories:

Environment. Users may use wireless sensing based applications (e.g., smart home control, user authentication) across rooms, such as different rooms of a house or different offices in a company building. As the multi-path of WiFi is dominated by wall reflections, the change of room layout will greatly alter the multi-path. Such changes make the CSI intensities and fluctuation patterns vary greatly, even with same activities/gestures from one participant.

Setting. The placements of furniture and appliances can vary from day to day in practical scenarios. These room objects reflect and diffract WiFi signals, and thus the multi-path of WiFi signals will be impacted by the placement changes of these objects. Similar to environment variations, the multi-path changes will alter the intensities and fluctuation patterns of CSI.

User. Many sensing applications (e.g., smart home control, health care monitoring) are required to have reliable performance among different users. However, the physiological (e.g., heights, length of arm and leg) and behavioral characteristics (e.g., gesture preferences) are varying from person to person. Such differences will result in changes in CSI patterns, making it difficult to apply a trained model for wireless sensing to a new user.

User's Facing Direction. A user may also perform activities/gestures with different facing directions. The changes in facing direction will alter the angle of signals reflected and diffracted by the user's body, thereby impacting the signal patterns.

User's Position. The user may also perform the activities/gestures at slightly different positions every time, which results in different relative distances/angles to the WiFi devices. As CSI is very sensitive to distance/angle changes, the variations of the user's position can change the intensities and fluctuation patterns of CSI.

Timeline. Existing studies [9, 17, 18] reveal that CSI may change over time (e.g., different time on one day or different days) due to different temperatures, humidity, and hardware imperfections. Thus, even when the wireless signal propagation conditions are static (e.g., no furniture settings changes and human movements), the intensities of CSI can be greatly different across timelines.

Impacts of Domain Variations. Without losing generality, we illustrate the impacts of domain variations by performing the user identification with different setting variations (i.e., furniture placements). We consider using the feature extraction methods and the deep learning models developed in an existing study [19]. The CSI data and classes are collected from 10 participants who are asked to perform the same set of pre-defined activities (e.g., walking, sitting down). We separately collect three sets of CSI data from three furniture settings, consisting of one sofa, one microwave oven, three cabinets, and five chairs. Figure 1(a) shows the user identification accuracy by using the CSI data under two furniture settings for training. We can find that the models achieve close to 100% accuracies in the training setup, but the accuracies decrease by over 30% if the testing CSI data is from a third (i.e., unseen) furniture setting. These preliminary results show the significant impacts of domain variations in traditional deep learning methods.

The key reason for the degraded performance of deep learning under domain variations is attributed to the data drifts. The classifier in deep learning models (e.g., the last fully-connected layer with SoftMax activation) assumes the training and testing data follow the same distribution (i.e., i.i.d assumption) [24]. However,

this assumption does not hold under the domain variations, where the data drifts alter the distribution of testing inputs. As a result, there may not have a class with an overwhelmingly high predicted probability. Since a traditional classifier is designed to predict a class with the maximum probability, misclassifications might occur. Figure 1(b) illustrates this problem by showing the probability distributions of predicting a target user (i.e., User 1) based on CSI data collected in a training environment (i.e., Env 1). When testing with an unseen domain (i.e., Env 3), the classifier produces prediction probabilities for all the users, among which User 1, 5 and 7 have relatively higher values. In this example, instead of User 1 (ground truth), User 7 is selected because of the related probability is slightly higher than that of User 1. With this, we observe a typical case of performance degradation in traditional deep learning models under domain variations due to the violation of i.i.d. assumption.

3 SYSTEM DESIGN

3.1 Cross-Domain Conformal Prediction

Mathematically, conformal prediction uses a *nonconformity measurement function* to examine how nonconformal (dissimilar) a WiFi sample, s, is compared to a *calibration set* derived from the training data. Similar to a validation set for cross-validation, the calibration set, whose name is from a prior work of conformal prediction [24], is a small proportion of the dataset that is not overlapping with the training data. Conformal prediction needs to first build a nonconformity measurement function using the training dataset, and then have a calibration dataset as the nonconformity profile. In the inference phase, the nonconformity scores of the testing data are compared with the profile to quantify the conformity.

We denote the calibration subset of the k^{th} class as $C_k, k \in \{1, \ldots, K\}$, where K is the total number of classes. Thus, the calibration set can be denoted as $C = \cup_{k=1}^{K} C_k$. We denote the nonconformity measurement function for the k^{th} class as $\hat{f}_k(\cdot)$. It is learned from the feature representations of a subset of training data of class k, which is non-overlapped with C_k. The nonconformity score of s is then calculated via: $a_k^{(s)} = \hat{f}_k(s)$. The smaller $a_k^{(s)}$ is, the less likely the testing input fits into the profile of class k. Conformal prediction determines whether s belongs to class k by quantifying the *degree of conformity*, denoted as $d_k^{(s)}$, which is the proportion of feature representations that s is conformal within C_k. The quantification process is essentially a comparison between the nonconformity score of s and samples in the calibration set. Given a significance level ϵ and the degrees of conformity of all K classes, conformal prediction produces a predictive set $\{k : d_k^{(s)} \geq 1 - \epsilon\} \in \{1, \ldots, K\}$.

To achieve a valid conformity quantification, conformal prediction requires s to be *exchangeable* [25] with the feature representations in C_k. Different from the i.i.d assumption, the exchangeability allows s and the representation in C_k drawn from a similar data distribution. However, this assumption could still be violated in practice when domain variations occur. To enable robust predictions under domain variations, we design a cross-domain conformal prediction framework that meets the requirement of exchangeability, even when domain variations occur. Our idea is to include at least two calibration subsets of CSI data from different domains. Instead of using the calibration subset of a single domain, such an approach quantifies the impacts of domain variations in terms of nonconformity scores across two domains. It then leverages the cross-domain nonconformity to score the degree of conformity on s, which may be from a third (unseen) domain. We design our own nonconformity measurement function (e.g., $\hat{f}_k(\cdot)$) and calibration set (e.g., C_k) to realize such a cross-domain approach. The detailed designs of our nonconformity measurement function and calibration set are elaborated in Section 4.

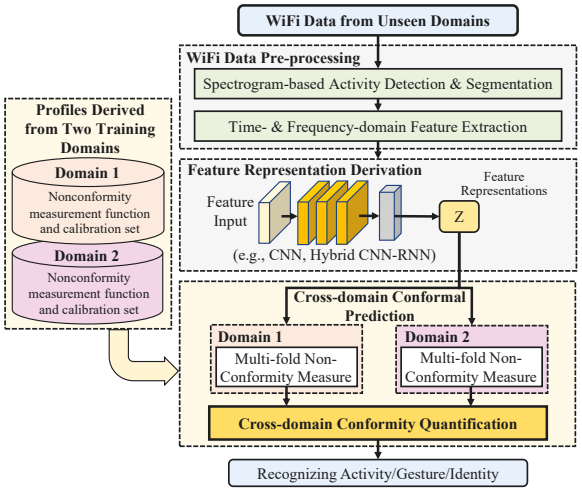


Figure 2: Overview of the designed cross-domain conformal prediction framework.

Our cross-domain conformal prediction framework exhibits several fundamental differences compared to prior solutions to domain variations [4, 11, 19, 23, 32, 33]. First, it does not require collecting any new CSI data and labels (e.g., from an unseen/unknown domain) for retraining or adapting the deep learning model. In contrast, prior domain-adaptation-based methods [11, 19] need to use a considerable amount of new data (e.g., CSI data without labels) collected from a target domain for retraining, so as to align the data drifts between the training and testing data. RISE [32] designs an anomaly detector with incremental learning based on the intermediate features of conformal prediction to enhance the robustness of machine learning models. That approach still requires moderate human effort to label the misclassified data to retrain the model. Second, our framework does not assume any prior knowledge of the impacts of target domains, while prior studies relying on data augmentation techniques [4, 23] need to synthesize the distortions of domain variations on training data, which are difficult to generalize for unseen domains. Third, our framework works as an orthogonal solution to domain variations compared to prior approaches based on feature engineering and domain adaptation. It does not require any new features or model updates, and thus it can be easily integrated into current WiFi sensing systems.

3.2 Challenges

We face several key challenges to realize the proposed cross-domain conformal prediction framework that enhances the robustness of deep learning models under domain variations:

Quantifying the Nonconformity of Testing CSI Data. The success of conformal prediction is built upon the effectiveness of

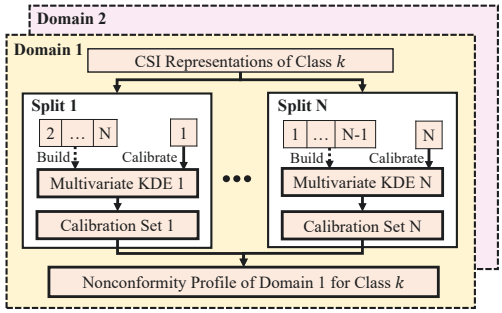


Figure 3: Illustration of the process to construct the kernel-density-based nonconformity measure and calibration subset for class k. For each domain in the training data, we split the data into N folds. We take turns to use one fold as the calibration subset and the reminding fold to build the kernel density estimator (KDE). The same process is applied to all other classes for conformal prediction.

the nonconformity measurement function $\hat{f}_k(\cdot)$, which quantifies how dissimilar a testing CSI input is compared to a calibration set. Existing parametric measurement algorithms (e.g., linear regression and deep learning) rely on mathematically learning a well-defined or predictable mapping relationship between the CSI testing data and the target objectives. However, such a relationship can be easily distorted under practical scenarios with data drifts caused by domain variations. Thus, it is essential to design a nonconformity measurement function that is robust to the data drifts.

Maximizing Training Data Utilization for Calibrating Conformal Prediction. Conformal prediction needs to build a nonconformity measurement function and a calibration set to quantify the conformity. To enable effective conformity quantification, the data used to build the nonconformity measurement function cannot be reused in the calibration set. This practice limits the utilization of the training data for calibration, resulting in suboptimal performance. Therefore, it is desirable to have solutions fully utilize all training data in the calibration set.

Satisfying Application-specific Requirements on Accuracy and Size of Predicted Class Set. Different from traditional classifiers that output a single class label given a testing CSI input, conformal prediction examines the degree of conformity on individual classes and produces a class set. Consequently, there is a trade-off between the prediction accuracy (i.e., the class set includes the correct class label) and the number of predicted classes. In reality, different applications have different requirements on the accuracy and the number of predicted classes. For example, for some personalized applications exploiting wireless sensing, such as recommending TV content and adjusting room temperature, it is desirable to minimize the number of predicted identities. This capability is necessary to enable the applications to provide more appropriate personalized services. We need to adapt our conformal prediction framework based on such requirements.

3.3 Framework Overview

To address the aforementioned challenges, we design a cross-domain conformal prediction framework as illustrated in Figure 2. We consider a scenario where the labeled training CSI data are collected under at least two training domains. The trained model based on both deep learning and conformal prediction can directly operate on new CSI data collected under various types of domain variations (e.g., the changes of room and furniture placement) without re-training/adaptation. It does not require explicitly extracting new features regarding specific domains. Instead, it quantifies the conformity of the testing input without assuming the types of domain variations. To showcase the effectiveness of our framework, we apply our framework to three representative CSI-based applications: gesture recognition, activity recognition, and user identification.

CSI Data Pre-processing. Our framework takes the time-series CSI measurements from WiFi-enabled devices (e.g., voice assistants, smart refrigerators) as input. It first performs *Spectrogram-based Activity Detection & Segmentation* that determines the CSI segments of user activity/gesture through time-frequency analyses based on CSI amplitudes. A set of time- and frequency-domain features are extracted to characterize the user's activity, identity, and gesture uniqueness (e.g., speeds of motions, gesture preferences).

Feature Representation Derivation. Our framework then employs deep learning models to further compute feature representations, which are the outputs of the last layer prior to the classifier. Existing studies [19, 20, 35] show that leveraging the feature representations from deep learning models for classification is computationally efficient and is robust to small-scale input CSI variations (e.g., minor activity differences across repeats). Also it can be applied to different models (e.g., CNN, Hybrid CNN-RNN) without any modifications to the model architecture.

Cross-domain Conformal Prediction. The core component of our framework is a conformity quantification process based on domain variations in the training dataset. Compared to traditional conformal prediction, which relies on the data of a single domain, our cross-domain framework leverages the nonconformity in data across two (or more) training domains to quantify the conformity of testing CSI data. Particularly, we build pairs of nonconformity measurement function and calibration set for each individual domain in the training dataset. Such a quantification process meets the requirement of exchangeability [25] even when domain variations occur. In addition, we design the *Multi-fold Nonconformity Measure* that takes turns to use part of the data for calibration and the rest to build the nonconformity measurement function. It utilizes all training data in the calibration set and thus significantly enhances the performance of nonconformity measure. To meet the application-specific requirements on the prediction accuracy and the number of classes, we develop two different approaches to perform *Cross-domain Conformity Quantification* and determine the class set, with priorities on maximizing the prediction accuracy and minimizing the size of the class set, respectively.

4 CROSS-DOMAIN CONFORMAL PREDICTION FRAMEWORK

4.1 Density Based Nonconformity Measure

Kernel Density Estimator. To achieve effective conformal prediction, we first need to design a nonconformity measurement function $\hat{f}_k(\cdot)$ to quantify how nonconformal (dissimilar) the feature representations of a testing CSI input, s, is compared to those of a calibration subset C_k. The intermediate outputs of deep learning

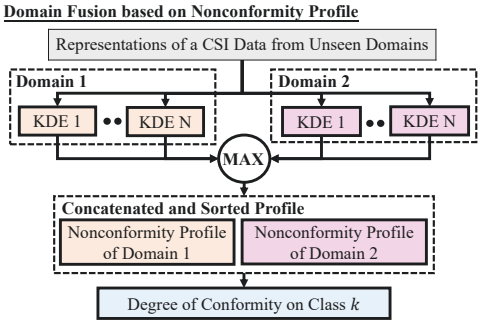


Figure 4: Illustration of conformity assessment through jointly considering all nonconformity profiles from both domains. Such a joint nonconformity profile helps to reduce the number of classes in the class set.

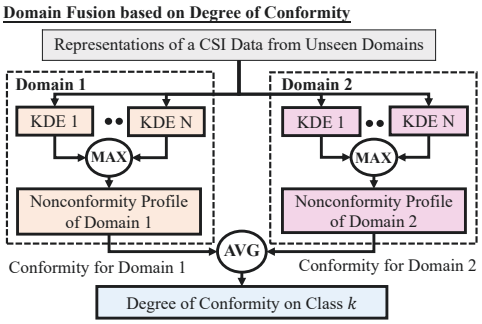


Figure 5: Illustration of conformity assessment through separately considering the nonconformity profile from each domain. Such a scheme helps to enlarge the prediction region of conformal prediction.

model from the last dense layer prior to the classifier are used as the feature representations. On appearance, existing classification algorithms (e.g., SoftMax layer, linear regression, deep learning) may also be used to quantify the nonconformity. For example, the probability mapping relationship $P(k|s)$ from SoftMax layer characterizes the probability of the input s belongs to a class k. However, based on Bayes' theorem, we have $P(k|s) = \frac{P(s|k)P(k)}{P(s)}$, where the distribution of the CSI data, $P(s)$, can be significantly impacted by domain variations. Thus, these classification algorithms are not reliable under domain variations. To address this challenge, instead of $P(k|s)$, we exploit the probability mapping relationship $P(s|k)$ to quantify the nonconformity of the CSI input against a class k, in order to achieve robustness under domain variations.

In our framework, we adopt the multivariate kernel density estimator (KDE) as the nonconformity measurement function [6] to quantify whether a new input point s belongs to the k^{th} class or not (i.e., estimating $P(s|k)$). The KDE approach has shown to have optimal performance under weak assumptions on the testing data [13], where the data drifts may occur. Multivariate KDE is a non-parametric estimator based on the distribution of a data set. In our study, the data is from a training set of class k: $B_k = \{x_{k,1}, \ldots, x_{k,m_k}\}$, where m_k is the size of the set. The multivariate KDE is then specified as follows:

$$\hat{f}_k(x) = \frac{1}{m_k} \sum_{i=1}^{m_k} K_H(x - x_{k,i}), \tag{1}$$

where $K_H(\cdot)$ is a multivariate kernel, a symmetric probability density function; H is a bandwidth matrix, which is a diagonal matrix with 1 in all the diagonal elements; $K_H(x) = |H|^{-1/2} K(H^{-1/2}x)$, where $K(\cdot)$ is a Gaussian Kernel. When a KDE is fitted, we can use it as our nonconformity scores for any data point, either from a test CSI input (e.g., s), or a calibration set (e.g., C_k), by replacing x with the values of the particular feature representations.

Multi-fold Nonconformity Measure. Existing approaches of conformal prediction need to build a nonconformity measurement function using the training dataset and a nonconformity profile using the calibration dataset to quantify the conformity. As discussed in Section 3.1, the training data is not overlapping with the calibration dataset, which is similar to the structure of the train-validation setup in the cross-validation. Such a practice limits the utilization of the dataset because only a small proportion of dataset will be used to construct the nonconformity profile, resulting in suboptimal performance. To address this challenge, we explore a multi-fold approach to fully utilize all training data. We illustrate our multi-split nonconformity measurement function in Figure 3. For the data of each domain, we partition the data into N folds of equal size. We leave one fold as the calibration set (C_k) and use the remaining $N-1$ folds (B_k) to build the multivariate KDE. We repeat this process N times for each domain until all folds have been used as the calibration set once. For each testing input, our framework learns N nonconformity measurement functions regarding each domain. We refer to the nonconformity scores from all the N calibration subsets as the nonconformity profile.

4.2 Cross-domain Conformity Assessment

Class Set Prediction based on Nonconformity Profile. For each pair of nonconformity measurement functions $\hat{f}_k(\cdot)$ and calibration set C_k, conformal prediction compares the of the calibration set of class k and computes the *degree of conformity*, which quantifies the level of uncertainty on feature representations s of a testing input. The smaller the degree of conformity, the more similar s compared to the feature representations in C_k. We denote the calibration set as, $C_k = \{s_{k,1}, ..., s_{k,n_k}\}$, where n_k is the size of C_k. We use $\hat{f}_k(\cdot)$ to get the nonconformity scores for each set of feature representations in the calibration set: $\{a_{k,1}, a_{k,2}, ..., a_{k,n_k}\}$. For a calibration set and the nonconformity score of the testing CSI input $a_k^{(s)}$, the degree of conformity can be computed via:

$$d_k^{(s)} = \frac{COUNT\{i \in \{1, ..., n_k\} : a_{k,i} \geq a_k^{(s)}\}}{n_k}, \tag{2}$$

where $COUNT(\cdot)$ denotes the operation of counting the number of n_k instances that meet the criteria (i.e., $a_{k,i} \geq a_k^{(s)}$). The degree of conformity is essentially the proportion of feature representations in the calibration set with nonconformity scores greater than the testing CSI input. If the degree of conformity is large, the testing CSI input is conformitive (similar) to the calibration set. It is likely that the input s belongs to class k. Otherwise, the testing input is non-conformitive (dissimilar) to the calibration set. Based on $d_k^{(s)}$, our framework performs a test to determine whether to include

class k in the class set. Given a significance level ϵ and the degree of conformity of all K classes, conformal prediction produces a predicted set $\{k : d_k^{(s)} \geq 1-\epsilon\} \in \{1,\ldots,K\}$. The larger the ϵ, the smaller the prediction region, meaning that fewer number of classes would be included in the class set.

Equation 2 formulates the computation of degree of conformity given a single pair of nonconformity measurement function $\hat{f}_k(\cdot)$ and calibration set. In our framework, we consider calculating the degree of conformity based on multiple nonconformity measurement functions and calibration sets from all D domains. This will enable effective conformal prediction under domain variations. Based on different requirements for WiFi sensing, we develop two schemes that fuse the nonconformity measurement functions and calibration sets. Specifically, we design *Domain Fusion Based on Nonconformity Profile* to minimize the size of the predicted class set, and we also develop *Domain Fusion based on Degree of Conformity* to maximize the probability the predicted set includes the correct label.

Domain Fusion based on Nonconformity Profile. To minimize the size of the predicted set, our idea is to reject classes with feature representations nonconformal to the *majority* of the data across all training domains. To realize such a rejection process, as illustrated in Figure 4, we select the maximum nonconformity score from the scores derived with all $D \times N$ KDEs as $a_k^{(s)}$ for computing the degree of conformity. It quantifies the maximum dissimilarity between s and all calibration subsets. We concatenate and sort the nonconformity scores of calibration subsets from all D domains and all N folds to generate C_k, which includes nonconformity scores of all folds (i.e., all training data). Under such a setting, $a_k^{(s)}$ will lay on the tail of the nonconformity scores distribution if s is dissimilar to the majority of the domains (i.e., calibration subsets of all training domains). To take into considerations of the scale differences among different calibration subsets, we normalize the nonconformity scores of each subset to zero mean and unit variance before the concatenation. Our framework then computes the degree of conformity, which is the proportion of nonconformity scores in C_k greater than $a_k^{(s)}$ as we formulated in Equation 2. Based on the degree of conformity, our framework then determines whether s belongs to class k. We repeat such a conformity quantification for all K classes to obtain the final predicted set.

Domain Fusion based on Degree of Conformity. To enhance the prediction accuracy, we propose to compute *degree of conformity* based on individual domains and leverage the average degree of conformity for class set prediction. By separately considering the degree of conformity of different training domains, the scheme leads to larger prediction regions, making the generated class set to have a higher probability to include the correct class. We illustrate the flow of our scheme in Figure 5. For each training domain, we concatenate the calibration subsets of N folds to generate C_k. We perform the same normalization process as described above to remove the scale differences among different calibration subsets. The maximum nonconformity score from the N KDEs is used to compare with C_k to compute the degree of conformity. Our framework then uses the averaged degree of conformity of all training domains to determine whether s is belonging to predicted class k. We apply this process for all K classes to produce the final class set. Such a fusion scheme extends the prediction regions for achieving higher prediction accuracy.

5 PERFORMANCE EVALUATION

5.1 WiFi Datasets

Public Dataset for Gesture Recognition. We evaluate the gesture recognition performance of our framework on Widar3.0 dataset [35]. It is a public gesture dataset involving 4 types of domain variations, including *environment*, *user*, *user's facing direction*, and *user's position* variations. CSI data of 6 gestures (i.e., pushing & pulling, sweeping, clapping, sliding, drawing a circle, and drawing a zigzag) conducted by 9 persons are collected. The WiFi packet transmission rate is 1000 packets per second. To examine the impacts of environmental changes, we use the CSI data collected in three rooms with different layouts, including an empty classroom, a spacious hall, and an office room, furnished with desks and chairs with a $2m \times 2m$ square sensing area. The dataset contains gesture data from 5 locations (i.e., northwest, northeast, southwest, southeast and the center) and 5 orientations (i.e., facing northwest, north, northeast, east and southeast) in the sensing area. The CSI data collected from 5 positions and 5 orientations are used to evaluate our framework under position and orientation variations, respectively.

Self-collected Dataset for Activity Recognition/User Identification. To further evaluate our approach in real environment settings, we collect our own dataset under *setting*, *user*, *user's position*, and *timeline* variations. Two laptops equipped with Intel 5300 NICs are used to collect CSI data. We collect CSI data of 6 activities (i.e., picking up a remote control, sitting, exercising, using a stove, and walking in two different trajectories) performed by 10 users in a residential apartment. The residential apartment has the size of 33ft × 17ft and the office has the size of 21ft × 12ft. The residential apartment includes common room objects, such as sofas, home appliances, chairs, and desks. To emulate setting variations, we collect the data under three different furniture settings, with one sofa, one microwave oven, three cabinets, and 5 chairs moved at least 3ft for each setting. To study user's position variations, we collect CSI data of 3 activities (i.e., sitting, stretching the body, and typing on a keyboard) in an office. The office environment has different types of furniture, such as desks, chairs, and cabinets. For each activity, the user is asked to perform the activity at 5 different proximate positions at least one foot away from each other. The experiment is repeated in the morning, afternoon, and night of a same day for evaluating our framework's robustness across different timeline.

5.2 Deep Learning Models

While our cross-domain conformal prediction framework should work for all deep learning models, we particularly focus on the following models designed for the three applications.

Gesture Recognition Model. We implement a hybrid CNN-RNN model based on the method of Widar3.0 [35]. In particular, we use Doppler Frequency Shift (DFS) [29] extracted from the spectrogram of CSI amplitude as the feature for recognizing gestures under different domain variations. The hybrid CNN-RNN model leverages a 2D constitutional layer, a pooling layer, two fully-connected layers, and a single layer of gated recurrent units to learn the temporal

patterns in CSI. A fully connected layer with SoftMax activation function is used as the classifier for gesture recognition.

Activity Recognition Model. We implement a CNN-based model based on existing work [19] for activity recognition. The model takes normalized amplitude and spectrogram of CSI as inputs under different variation scenarios. The CNN model adopts 3 convolutional layers to learn the time and frequency features in CSI, respectively. Two fully-connected layers followed by a SoftMax activation function are used to predict the class of activities. Note that we did not implement the domain discriminator, which needs to be trained with CSI data from a target domain.

User Identification Models. We adopt the same deep learning model architecture for user identification as the one developed for activity recognition. Unlike activity recognition, we use the labels of users' identities with normalized amplitudes and spectrograms of CSI to train the model for user identification.

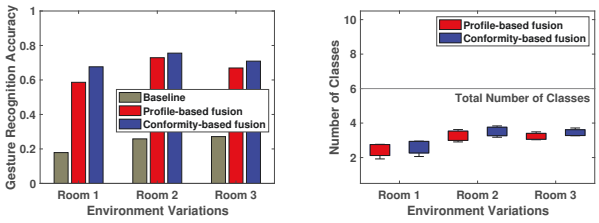
5.3 Evaluation Setup and Methodology

Baseline Methods. We use the models introduced in Section 5.2 to derive feature representations. To perform conformal prediction, we replace the classifier with our framework to calculate the degree of conformity. We compare the performance of our framework with the original deep learning models relying on the classifiers for gesture recognition/activity recognition/user identification.

Evaluation Metrics. We focus on using the following two metrics to evaluate our framework. *1) Gesture Recognition/Activity Recognition/User Identification Accuracy:* this is the percentage of the testing CSI data being included in the class sets predicted by our framework. Note that the class set produced by conformal prediction may contain no labels, a single label, or multiple labels. This is because conformal prediction quantifies the conformity (uncertainty) of the testing data and dynamically determine a class set. It is essentially different from traditional classification approaches that output the probabilities of a single or a set of top-k labels. Thus, we use classification accuracy to examine the performance based on the original paper on conformal prediction [24], instead of top-k accuracies or precision/record or F1 scores, which normally assumes a singleton or a fixed number of labels in the prediction. *2) Average Number of Classes:* this is the mean of the size of the predicted sets predicted by our framework. It is desirable to have a small number of classes in the predicted set. As there is a trade-off between the accuracy and the number of classes, we show all the results with both metrics. We repeat our experiments three times and produce barplots of prediction accuracies and boxplots of number of labels in the predicted sets under each scenario.

5.4 Performance Across Environments

Firstly, we show the performance of our framework on across-environment gesture recognition using the Widar3.0 dataset. The gesture data is collected in three rooms of different sizes. We first use data from any two rooms to construct nonconformity measure and calibration sets, then make predictions for the users in the third (unseen) room. Figure 6(a) displays the gesture recognition performance of our framework compared with the baseline. We find that the baseline model has 27.2% accuracy due to the significant impacts from room variations. In contrast, our cross-domain framework has much higher gesture recognition accuracy. The final results are 65.5% and 72.8% for the two fusion approaches, which improve significantly over the baseline results. Even under large scale domain variations like room changes, our framework can still keep the system accuracy at about 70%. Although the results are still lower than 80%, our framework shows significantly improvement on the classification performance, improving the accuracy to about 40% from 27% of the baseline without applying the conformal prediction method. The results demonstrate our conformity assessment via conformal prediction is more reliable compared to the traditional classifiers. The numbers of classes for the two approaches are shown in Figure 6(b) and the average of both are 2.9 and 3.0. The profile-based approach has a smaller average number of classes. This is because the aggregated calibration set excludes classes dissimilar to the testing input.

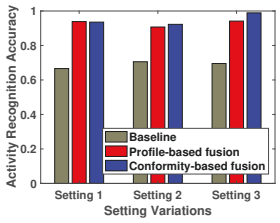


(a) Gesture recognition accuracy under environment variations (b) Boxplots of number of classes under environment variations

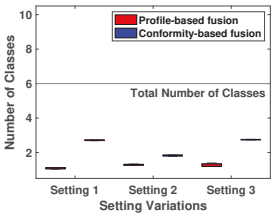
Figure 6: Training-Testing Combinations Across Environments, Gesture Recognition: Gesture recognition accuracy (a) and the boxplots of number of classes (b) based on the Widar3.0 dataset. The gesture data is collected in three different rooms. For each combination (e.g., Room 1), we use data of two rooms (i.e., other than Room 1) for constructing nonconformity measure and calibration sets and the data of Room 1 for testing. ϵ is chosen to be 0.35 and 0.28 for profile-based and conformity-based fusion approaches, respectively.

5.5 Performance Across Settings

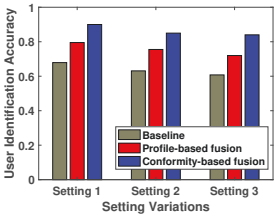
Activity Recognition under Furniture Setting Variations. Secondly, we evaluate activity recognition task under across-setting using our self-collected apartment datasets. The data is collected in one apartment room with three furniture settings. We use the data from two of the settings for training the deep learning model and constructing nonconformity measure and calibration sets. We apply our conformal prediction framework for activity recognition in a third (unseen) setting. Figure 7(a) gives the activity recognition accuracies of our framework and the baseline. We find that the baseline model has an average 68.1% accuracy due to setting variations. In contrast, our framework has average net accuracy improvements of 27.3 and 29.2% respectively under the two fusion approaches. Our system accuracy is over 90% in general and can be as high as 98.8% in some settings. The average numbers of classes of the two approaches are 1.2 and 2.5 respectively as shown in Figure 7(b). Based on these experiments, profiled-based approach is better to handle tasks with strict requirement of number of classes, while the conformity-based approach is needed if the application would like to relax the tolerance of number of classes.



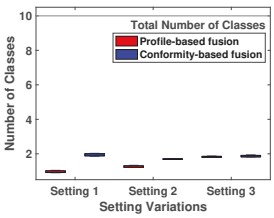
(a) Activity recognition under setting variations



(b) Boxplots of number of classes under setting variations



(c) User identification under setting variations



(d) Boxplots of number of classes under setting variations

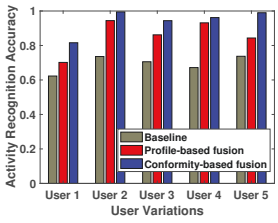
Figure 7: Training-Testing Combinations Across Settings, Activity Recognition and User Identification: Activity recognition accuracy (a), user identification accuracy (c) and the boxplots of number of classes (b)(d) based on the self-collected apartment dataset. The data is from three furniture settings. For each combination (e.g., Setting 1), we use data of two settings (i.e., other than Setting 1) to build nonconformity measure and calibration sets while the data of Setting 1 for testing. ϵ is 0.02 for both profile-based and conformity-based fusion approach for activity recognition. For user identification, ϵ is selected to be 0.03 and 0.04 for two fusion approaches.

User Identification under Furniture Setting Variations. We also show the user identification performance under the same apartment dataset and experiment setup. The results are in Figure 7(c). The average accuracy improvements of the two approaches are 12.4% and 23.6% which bring the system performance up to 79.8% and 91.2% compared to the baseline. Figure 7(d) is the boxplot of numbers of predicted classes with average 1.2 and 1.8. We have similar observations on the trade-off between the accuracy and the number of classes for the two approaches.

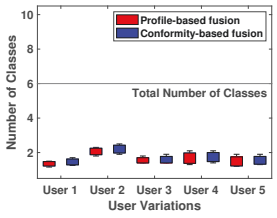
5.6 Performance Across Users

Activity Recognition under User Variations. We show the performance of our framework on self-collected apartment dataset under user variations. The data is the same as those used for user identification/activity recognition. We use the data of four users for training and a fifth (unseen) user for testing. Figure 8(a) gives the activity recognition performance of our framework and the baseline, which is average at 63.0%. Our proposed conformal prediction is able to provide 21.4% and 30.9% accuracy improvement and bring the system performance to 84.4% and 93.9% with average number of classes 1.6 and 1.9 as shown in Figure 8(b).

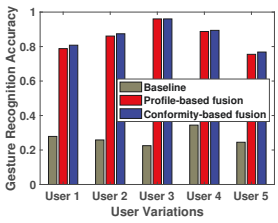
Gesture Recognition under User Variations. We also show the performance of gesture recognition on Widar3.0 dataset under



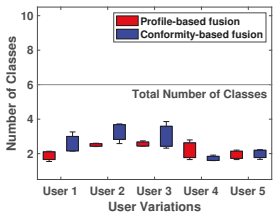
(a) Activity recognition under user variations



(b) Boxplots of number of classes under user variations



(c) Gesture recognition under user variations



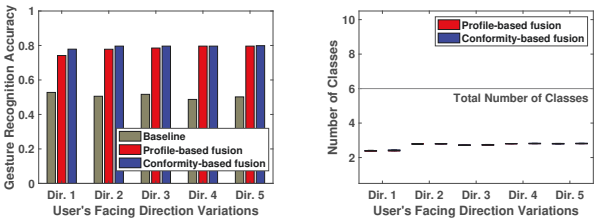
(d) Boxplots of number of classes under user variations

Figure 8: Training-Testing Combinations Across Users, Activity Recognition and Gesture Recognition: Activity recognition accuracy (a), gesture recognition accuracy (c) and the boxplots of number of classes (b)(d) based on the self-collected apartment dataset and public Widar3.0 dataset respectively. The activity data is collected from ten different users, and Widar3.0 dataset contains data of nine different users. For each combination (e.g., User1), we use data of random four users(i.e., other than User1) to build nonconformity measure and calibration sets and the data of User1 for testing. ϵ is 0.01 for both profiled-based and conformity-based fusion approaches for activity recognition. For gesture recognition, ϵ is chosen to be 0.15 and 0.3 respectively.

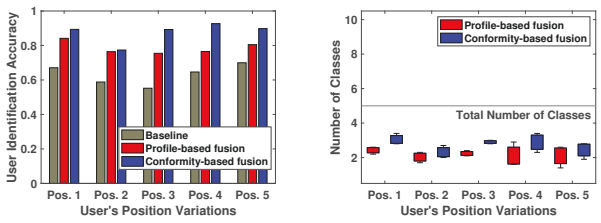
user variations. We use data from four users to construct nonconformity measures and calibration sets and test the result on one of the unseen user. The baseline of gesture recognition under user variations is as low as 27.1%. With our proposed method, the performance can be improved to over 80% and can be as high as 96.7%. The average numbers of classes are 2.2 and 2.5 for the profiled-based and the conformity-based fusion approaches respectively. The results demonstrate that our conformal prediction can retain model performance even if the performance of the deep learning model drops to a very low level.

5.7 Performance Across User's Facing Directions

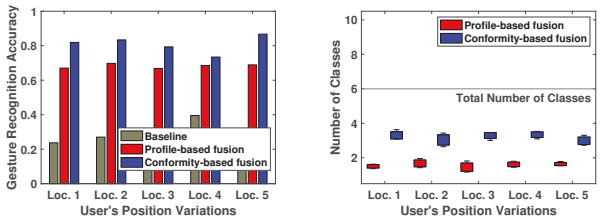
Gesture Recognition under User's Facing Direction Variations. We then study the performance of our framework under user's facing direction variations for gesture recognition on the Widar3.0 dataset. The data is collected in a position with five different facing directions of the user. We use the data of two directions for training. We then test the deep learning model and cross-domain conformal prediction in a third (unseen) direction. Figure 9(a) presents the user identification performance of our framework and the baseline. We find that the average accuracy of



(a) Gesture recognition under user's facing direction variations (b) Boxplots of number of classes under user's facing direction variations



(a) User identification under user's position variations (b) Boxplots of number of classes under user's position variations



(c) Gesture recognition under user's position variations (d) Boxplots of number of classes under user's position variations

Figure 9: Training-Testing Combinations Across User's Facing Directions, Gesture Recognition: Gesture recognition accuracy (a) and the boxplots of number of classes (b) based on the public Widar3.0 dataset. The data is collected at one spot with five facing directions. For each combination (e.g., (Dir)ection 1), we use data of two facing directions (i.e., other than Dir. 1) for constructing nonconformity measure and calibration sets and the data of Dir. 1 for testing. ϵ is selected to be 0.17 and 0.06 with the two fusion approaches.

Figure 10: Training-Testing Combinations Across User's Positions, User Identification : User Identification accuracy (a), Gesture Recognition accuracy (c) and the boxplots of number of classes (b)(d) based on the self-collected office dataset and public Widar3.0 dataset. The user identification data is collected from five positions. For each combination (e.g., (Pos)ition 1), we use data of two positions (i.e., other than Pos. 1) for constructing nonconformity measure and calibration sets and the data of Pos. 1 for testing. ϵ is 0.02 for both profile-based and conformity-based fusion approaches. Similarly, gesture recognition data is collected from five different torso locations. The experiment setup is the same as user identification task. ϵ is chosen to be 0.15 and 0.3 respectively.

the baseline model is 50.3%. Our proposed method turns out to be robust to such variations with the average gesture recognition accuracy of 78.8% and 80.2% for the profiled-based fusion approach and the conformity-based fusion approach respectively. The average numbers of classes are 2.2 and 2.4 as shown in Figure 9(b). Though addressing the impacts of facing directions is challenging for conformal prediction because the high similarity between domains makes uncertainty assessment hard to benefit from the cross-domain information, our fusion methods are still able to provide close to 30% accuracy improvement.

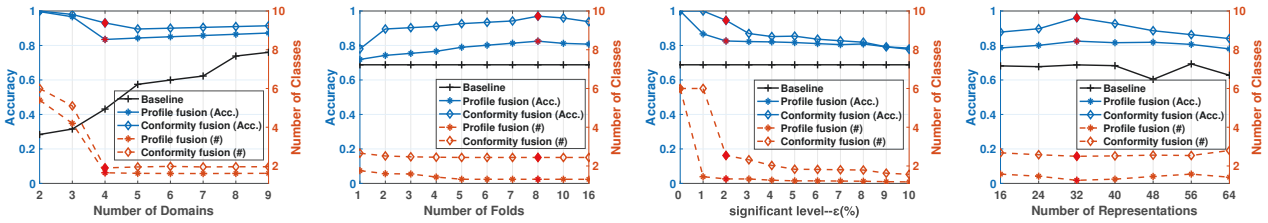
5.8 Performance Across User's Positions

User Identification under Proximate Position Variations. Here we present the performance of user identification with proximate position variation using our self-collected office dataset. The dataset is collected in five different proximate positions in an office environment. We use the data of two positions for constructing the nonconformity measure and calibration sets. We then test the DNN model with our proposed schemes in a third (unseen) position. Figure 10(a) gives the user identification accuracies of our framework and the baseline. We find that the baseline model has an average accuracy of 63.3%. Our cross-domain framework is robust to such variation and bring the average accuracy to 78.2% and 87.5% for profile-based schemes and conformity-based schemes respectively. The average numbers of predicted classes of the two schemes are 1.8 and 2.6 as shown in Figure 10(b). The results demonstrate the effectiveness of our scheme on improving the model's robustness.

Gesture Recognition under Torso Location Variations. At the same time Figure 10(c) shows our proposed method also works for gesture recognition under torso location variation using the Widar3.0 dataset. The baseline of DNN model has accuracy of 23.7%. We utilize the data from two torso locations for training and the data from a third (unseen) location for testing. Our proposed method is able to bring the system accuracy to average 68.2% and 80.7% with average number of classes 1.5 and 3.0 (as shown in Figure 10(d)). This demonstrates that our proposed method not only works for large variations such as room variation or furniture setting variation, it also works for relatively small variation.

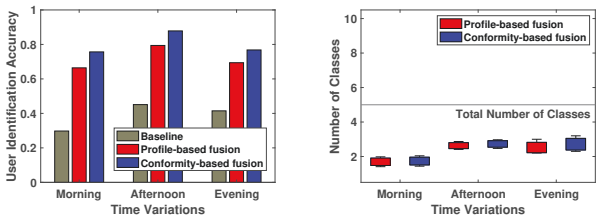
5.9 Performance Across Timelines

Finally, we show the performance of user identification across timelines using the self-collected office dataset. The data is collected in three time frames: morning, afternoon and evening. We use the data from two of three time frames for training and calibration of our cross-domain conformal prediction system. Then we use the data of remaining time frame for testing. Figure 12(a) shows the user identification performance of our framework and the baseline. We find that the accuracy of baseline model can be as low as 29.8% due to the significant variations in the wireless channel conditions across the day. In comparison, our framework is robust to such variations. The average user identification accuracy improvements are 33.3% and 41.2% and bring the system performance to as high as 79.1% and 87.0% for the profiled-based fusion approach and the conformity-based fusion approach respectively. The average predicted numbers of classes are 2.2 and 2.4 as shown in Figure 12(b). The results demonstrate that our proposed method can handle domain variations of both spatial and temporal changes.



(a) Number of Domains (b) Number of Folds (c) Significant Level–ϵ(%) (d) Number of Feature Representations

Figure 11: Performance of conformal prediction under different parameter settings.



(a) User identification under time variations (b) Boxplot of number of classes under time variations

Figure 12: Training-Testing Combinations Across Time, User Identification: User identification accuracy (a) and the boxplots of number of classes (b) based on the self-collected office dataset. The data are collected from three time of the day. For each combination (e.g., Morning), we use data of two time of the day (i.e., other than Morning) for training and the data of Morning for testing.

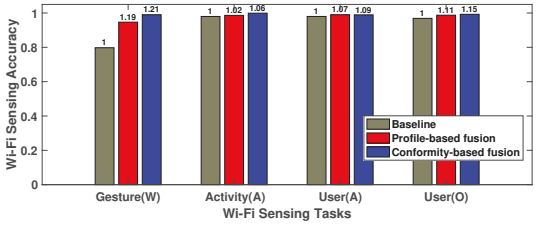


Figure 13: WiFi sensing tasks without domain variation. Each of Gesture Recognition, Activity Recognition or User Identification task is performed with one of (W)idar3.0, (A)partment or (O)ffice dataset(e.g., Gesture(W) is Gesture Recognition task with Widar3.0 dataset). The average number of classes is labeled on top of the bar plot.

5.10 Impacts of Framework Parameters

Impact of the Number of Training Domains. We showcase the impacts of number of training domain using user variations. In our previous experiments, we use 4 users for training, which is necessary to train the feature extractor of our DNN model. We illustrate this point and study the impact of the number of training domains in this section. We perform activity recognition task on self-collected apartment dataset, which has 10 users and 6 activities. Following the same experiment setup as subject variation, we compare the system performance using different numbers of users as training data. As shown in Figure 11(a), the DNN model has accuracy as low as 28% if there are only 2 users as training data. The DNN model performance will gradually improve with the increasing number of users in training set. The conformal prediction performs differently compared to DNN model. With as few as 4 users, the conformal prediction has accuracy above 84.4% and 93.9% with average number of predicted classes 1.6 and 1.9 for profile-based and conformity-based fusion schemes, respectively. This illustrates that conformal prediction can be effective without involving too many domains. This is a big advantage compared to traditional deep learning method in real-world applications.

Impact of the Number of Folds. Our proposed method has a key component which is the N-fold stratified splitting. We study the impact of number of folds on activity recognition with across furniture placement setup. Figure 11(b) displays the results with fold number $1 \sim 10$, and 16. The average number of predicted classes becomes stable after 5 folds and the performance accuracy reach the peak at 8 folds. The results show that the a relatively larger number of fold can boost the performance of conformal prediction.

Impact of Significant Level–ϵ(%). In this study, we choose different ϵ to balance the competing requirements of high accuracy and low number of predicted classes. In the previous experiments, we select the ϵ to achieve the highest possible prediction accuracy with the average number of classes under 3 when all tasks have at least 6 classes. In this section, we study the impact of ϵ on the performance of our proposed framework. We conduct this evaluation using the activity recognition task with across furniture placement setup. As shown in Figure 11(c), the average number of predicted classes and prediction accuracy will decrease when ϵ is increasing. Note that $1 - \epsilon$ is the usual confidence level. When the confidence level is decreasing (i.e., the significant level is increasing), the model will reduce the size of predicted set. This also leads to performance drop because the true class is more likely to be missing from the predictions. It is also worth noting that when the confidence level is 100% (i.e., significant level is 0), the prediction accuracy is 100% and the average number of classes is 6 (out of 6). This is because only when including all classes, we are 100% in confidence to claim that the true class is included in the outputs. Since all classes are in the prediction set, the accuracy is 100% trivially. From our experiments, the average number of predicted classes decreases sharply at the significant level of 0.01 for the profile-based method and 0.02 for the conformity-based method. As a result, we choose 0.02 for ϵ in our experiments by using the elbow method [22].

Impact of Feature Representation Dimensions. Similar with traditional deep learning method, the model performance is directly related to the quality of features. In this study, our features

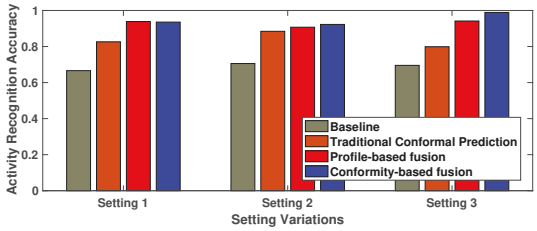


Figure 14: Training-Testing Combinations Across Settings, Activity Recognition: Activity recognition accuracy based on the self-collected apartment dataset. The data is from three furniture settings. For each combination (e.g., Setting 1), we use data of two settings (i.e., other than Setting 1) to build nonconformity measure and calibration sets while the data of Setting 1 for testing. ϵ is 0.02 for both profile-based and conformity-based fusion approach for activity recognition.

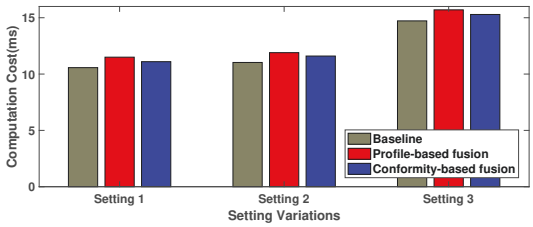


Figure 15: Computation cost of Across Settings, Activity Recognition: Activity Recognition computation cost on the self-collected apartment dataset. The data are collected from three furniture settings. For each combination (e.g., other than Setting 1) to build nonconformity measure and calibration sets while the data of Setting 1 for testing. ϵ is 0.02 for both profile-based and conformity-based fusion approach for activity recognition.

are the feature representations from the last dense layer of DNN models. Choosing the right number of feature representations is critical for conformal prediction to produce reliable results. We study this parameter on activity recognition with across furniture placement setup by obtaining the prediction results with the numbers of representations to be 16, 24, 32, 40, 48, 56, and 64. As shown in Figure 11(d), the DNN model becomes unstable when the number is greater than 40, while conformal prediction still maintains high prediction accuracy. When the representation number is 32, conformal prediction reaches the highest prediction accuracy and the smallest average number of classes. Besides, DNN also reaches its highest accuracy with 32 as the number of neurons of its last dense layer before SoftMax layer. This study indicates that though conformal prediction is robust to the number of feature representations, it still benefits from a carefully tuned DNN parameter. The fact that conformal prediction is not sensitive to such parameter changes makes it easier to be deployed in real life scenarios.

5.11 In-Domain Accuracy

We show that conformal prediction is effective under cross domain setup, it is still unknown how it performs without domain variation. In this section, we show the results of Gesture Recognition on Widar3.0 dataset, Activity Recognition on Apartment dataset and User Identification on both Apartment and Office dataset. As shown in Figure 13, four tasks has the baseline accuracies of 79.77%, 97.98%, 97.99%, 96.88%. The profile-based conformal prediction has accuracies of 94.66%, 98.62%, 99.01%, 98.72% with average number of classes of 1.19, 1.02, 1.07, 1.09 and conformity-based conformal prediction accuracies of 98.99%, 99.89%, 98.91%, 99.24% with average number of classes of 1.21, 1.06, 1.13, 1.15 respectively. Conformal prediction is able to provide marginal improvement on top of the high baseline accuracy under in-domain setup without involving many multiclass predictions. The results confirm the effectiveness of conformal prediction on testing data without domain variations.

5.12 Computation Cost

To evaluate the computation cost of our framework as the step of kernel density estimation might be computationally intensive, we measure the average inference time of CNN baseline, profile-based conformal prediction, and conformity-based conformal prediction using activity recognition across different furniture placement setups. As shown in figure 15, three settings have the baseline computation cost of 10.57ms, 11.03ms and 14.72ms. In comparison, the profile-based conformal prediction adds an extra cost of 0.93ms, 0.87ms, 0.98ms, and extra costs of 0.53ms, 0.57ms and 0.57ms for conformity-based conformal prediction. The results show that our framework based on kernel density estimation only incur less than 1ms computational costs per sample.

5.13 Comparing with Standard Conformal Prediction

As explained in the section 3.1, the proposed design is to meet the requirement of exchangeability. In figure 14, we further demonstrate the necessity of such design over the standard setup. We find that though the traditional conformal prediction setup improves the performance over the standard setup. The traditional conformal prediction method improves by average 12.7% over the deep learning baseline. Our proposed method is much better with 27.4% and 29.2% improvements under the two fusion mechanisms respectively. Therefore, our system is more robust in real life scenario.

6 DISCUSSION

Reducing the Number of Predicted Classes. Our work is the first attempt to exploit conformal prediction to address domain variations in WiFi sensing. We demonstrate through extensive experiments that our framework can significantly boost the performance of WiFi sensing systems under six types of representative domain variations, while maintaining a reasonable number of class labels in the prediction results. Such a capability allows wireless sensing to support many real-world applications without any stringent requirements on the number of predicted classes. Those applications include health care monitoring, smart home control, and personalized service (e.g., suggesting TV viewing contents, adjusting temperature/lighting), etc. We are aware that the current framework is not ready for applications requiring a singleton prediction (e.g., user authentication). More sophisticated algorithms to fuse the nonconformity scores and calibration sets of multiple domains

could be designed to improve the quantification of the degree of conformity, leading to more precise results.

Combining with Other Approaches. Our framework provides a solution orthogonal to existing approaches to improve the robustness of Wi-Fi sensing systems against domain variations. It can be easily combined with existing machine-learning-based approaches to enhance prediction performance. For instance, as our framework leverages deep learning to generate feature representations, domain adaptation techniques [10, 19] can be utilized to learn more robust representations that can improve the conformity quantification performance. In addition, we may apply data augmentation [33] to expand the dataset and construct more effective nonconformity measurement functions and larger calibration sets.

Leveraging Data from More Domains. In Section 5.10, our framework's performance is studied when different numbers of users were used as domains for activity recognition. As shown in Figure 11(a), the number of labels is reduced as the number of the training domain increases. It means that by leveraging data from more domains (i.e., more number of users) will have less uncertainty in generating the class set. In the future, we plan to explore algorithms that can fuse the nonconformity scores and calibration sets of multiple domains. This will improve the quantification of conformity measurements for more precise results.

Generalizing to other implementations. Our proposed framework grounded on conformal prediction is a versatile solution to the Wi-Fi sensing dilemma in reality. As shown in section 5, we demonstrate through extensive experiments that our framework is feasible for the three most representative Wi-Fi sensing applications. Only requiring the representations, our framework can be applied to any deep learning models, including classification models to quantify the conformity between source and target data, where domain variations may occur in the inference phase. For example, our framework can be applied to localization systems by examining the conformity of CSI data regarding Wi-Fi fingerprints.

7 RELATED WORK

WiFi Sensing based on Deep Learning. With the strong capabilities to model complex mapping relationships, deep learning has been widely used to support various WiFi sensing tasks, including but not limited to gesture recognition [2, 14], activity recognition [26, 28], user identification/authentication [12, 20], localization [1], and emotion detection [34]. These approaches features extracted from WiFi measurements into an output, with classification as the most prominent learning task. For example, WiSDAR [26] combines convolutional neural network with long short term memory units to classify WiFi signals of multiple antennas into a set of human activities. Shi *et al.* [20] design a system that extracts statistical features from CSI measurements associated with human daily activities and leverages an encoder-decoder-based network to identify users. These approaches and systems have shown the feasibility and initial success of WiFi sensing. However, they all face the challenge of domain variation problem, where the classification performance degrades significantly when reality factors change.

Existing Approaches to Mitigate Domain Variations. Efforts have been made to investigate the domain variation problems in the context of WiFi sensing [4, 11, 19, 23, 35]. For example, data augmentation [4, 8] has been exploited to generate synthetic or virtual training WiFi data to improve the robustness of deep learning models under domain variations. Domain adaptation techniques [11, 19] are used to transfer the knowledge learned from one WiFi environment to a target environment, by leveraging unlabeled WiFi data collected in the target environment. However, these approaches require to generate new features or retrain/adapt the model, unlikely to cover all possible domain variations in reality.

Conformal Prediction. Different from learning-based approaches relying on deriving a mapping relationship, conformal prediction performs statistical assessment based on training data to perform predictions. It quantifies the conformity between the testing data and a calibration set to determine a set of class labels as the output. Conformal prediction has been used for online learning [24], drug development/recovery [7], and image classification [3]. RISE [32] first applies conformal prediction on wireless sensing to detect data that is likely to be misclassified. Incremental learning with extra labeling is leveraged to improve the robustness of sensing systems. Compared to RISE, our cross-domain conformal prediction framework explores the relationships across training domains to quantify the conformity of testing CSI data. It relieves our framework of the i.i.d. assumption and thus enables applying conformal prediction in an unseen domain without any extra labeling efforts.

8 CONCLUSION

This work aims to understand the domain variation problem in WiFi sensing and develop a low-effort WiFi sensing framework that meets the sensing reality requirements of deployment in various real-world scenarios. Towards this end, we comprehensively investigate the impact of six typical domain variations (i.e., environments, settings, users, users' facing directions, users' positions, and timelines) in the three critical WiFi sensing applications (i.e., user identification, activity recognition, and gesture recognition). We propose a holistic WiFi sensing framework based on conformal prediction that can ensure robust cross-domain sensing performance without extra effort for data collection, feature modification, or model retraining/adaptation. The unique cross-domain conformal prediction scheme leverages multivariate kernel density estimation and nonconformity measurement functions derived from a few training domains to effectively assess the conformity of the testing WiFi data even if the data is from an unseen domain. We further design two fusion approaches to combine the nonconformity scores derived from the training domains to quantify the degree of conformity, with priorities on maximizing the prediction accuracy and minimizing the number of classes. Extensive experiments show that our framework has great performance of the three WiFi sensing applications across the six categories of domain variations.

9 ACKNOWLEDGMENT

This work was partially supported by the National Science Foundation Grants CCF1909963, CCF2000480, CCF2028858, CCF2028873, CNS2120276, CNS2120396, CNS1801630, CNS2120350, DMS2015373, DMS2027855.

REFERENCES

[1] Moustafa Abbas, Moustafa Elhamshary, Hamada Rizk, Marwan Torki, and Moustafa Youssef. 2019. WiDeep: WiFi-based accurate and robust indoor localization system using deep learning. In *2019 IEEE International Conference on Pervasive Computing and Communications (PerCom*. IEEE, 1–10.

[2] Heba Abdelnasser, Moustafa Youssef, and Khaled A Harras. 2015. Wigest: A ubiquitous wifi-based gesture recognition system. In *2015 IEEE conference on computer communications (INFOCOM)*. 1472–1480.

[3] Anastasios Angelopoulos, Stephen Bates, Jitendra Malik, and Michael I. Jordan. 2020. Uncertainty Sets for Image Classifiers using Conformal Prediction. *arXiv* (Sept. 2020). https://doi.org/10.48550/arXiv.2009.14193 arXiv:2009.14193

[4] Hong Cai, Belal Korany, Chitra R Karanam, and Yasamin Mostofi. 2020. Teaching rf to sense without rf training measurements. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 4 (2020), 1–22.

[5] Zhi-An Deng, Zhiyu Qu, Changbo Hou, Weijian Si, and Chunjie Zhang. 2018. WiFi Positioning Based on User Orientation Estimation and Smartphone Carrying Position Recognition. *Wireless Commun. Mobile Comput.* 2018 (Sept. 2018), 5243893. https://doi.org/10.1155/2018/5243893

[6] Tarn Duong and Martin L Hazelton. 2005. Cross-validation bandwidth matrices for multivariate kernel density estimation. *Scandinavian Journal of Statistics* 32, 3 (2005), 485–506.

[7] Martin Eklund, Ulf Norinder, Scott Boyer, and Lars Carlsson. 2015. The application of conformal prediction to the drug discovery process. *Ann. Math. Artif. Intell.* 74, 1 (June 2015), 117–132. https://doi.org/10.1007/s10472-013-9378-2

[8] Taesik Gong, Yeonsu Kim, Jinwoo Shin, and Sung-Ju Lee. 2019. MetaSense: few-shot adaptation to untrained conditions in deep mobile sensing. In *SenSys '19: Proceedings of the 17th Conference on Embedded Networked Sensor Systems*. Association for Computing Machinery, New York, NY, USA, 110–123. https://doi.org/10.1145/3356250.3360020

[9] Jinyang Huang, Bin Liu, Pengfei Liu, Chao Chen, Ning Xiao, Yu Wu, Chi Zhang, and Nenghai Yu. 2020. Towards anti-interference wifi-based activity recognition system using interference-independent phase component. In *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*. IEEE, 576–585.

[10] Wenjun Jiang, Chenglin Miao, Fenglong Ma, Shuochao Yao, Yaqing Wang, Ye Yuan, Hongfei Xue, Chen Song, Xin Ma, Dimitrios Koutsonikolas, et al. 2018. Towards environment independent device free human activity recognition. In *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking (ACM MobiCom)*. 289–304.

[11] Wenjun Jiang, Chenglin Miao, Fenglong Ma, Shuochao Yao, Yaqing Wang, Ye Yuan, Hongfei Xue, Chen Song, Xin Ma, Dimitrios Koutsonikolas, Wenyao Xu, and Lu Su. 2018. Towards Environment Independent Device Free Human Activity Recognition. In *MobiCom '18: Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*. Association for Computing Machinery, New York, NY, USA, 289–304. https://doi.org/10.1145/3241539.3241548

[12] Hao Kong, Li Lu, Jiadi Yu, Yingying Chen, Xiangyu Xu, Feilong Tang, and Yi-Chao Chen. 2021. MultiAuth: Enable Multi-User Authentication with Single Commodity WiFi Device. In *Proceedings of the Twenty-second International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing*. 31–40.

[13] Jing Lei, James Robins, and Larry Wasserman. 2013. Distribution-free prediction sets. *J. Amer. Statist. Assoc.* 108, 501 (2013), 278–287.

[14] Chenning Li, Manni Liu, and Zhichao Cao. 2020. WiHF: enable user identified gesture recognition with WiFi. In *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*. IEEE, 586–595.

[15] Jian Liu, Hongbo Liu, Yingying Chen, Yan Wang, and Chen Wang. 2019. Wireless sensing for human activity: A survey. *IEEE Communications Surveys & Tutorials* 22, 3 (2019), 1629–1645.

[16] Jian Liu, Yan Wang, Yingying Chen, Jie Yang, Xu Chen, and Jerry Cheng. 2015. Tracking Vital Signs During Sleep Leveraging Off-the-shelf WiFi. In *Proceedings of the 16th ACM International Symposium on Mobile Ad Hoc Networking and Computing (ACM MobiHoc)*. 267–276.

[17] Pengfei Liu, Panlong Yang, Wen-Zhan Song, Yubo Yan, and Xiang-Yang Li. 2019. Real-time identification of rogue WiFi connections using environment-independent physical features. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*. IEEE, 190–198.

[18] Xinping Rao, Zhi Li, Yanbo Yang, and Shengyang Wang. 2020. DFPhaseFL: a robust device-free passive fingerprinting wireless localization system using CSI phase information. *Neural Computing and Applications* 32, 18 (2020), 14909–14927.

[19] Cong Shi, Jian Liu, Nick Borodinov, Bruno Leao, and Yingying Chen. 2020. Towards environment-independent behavior-based user authentication using wifi. In *2020 IEEE 17th International Conference on Mobile Ad Hoc and Sensor Systems (MASS)*. IEEE, 666–674.

[20] Cong Shi, Jian Liu, Hongbo Liu, and Yingying Chen. 2017. Smart user authentication through actuation of daily activities leveraging WiFi-enabled IoT. In *Proceedings of the 18th ACM International Symposium on Mobile Ad Hoc Networking and Computing*. 1–10.

[21] Deepika Singh, Erinc Merdivan, Ismini Psychoula, Johannes Kropf, Sten Hanke, Matthieu Geist, and Andreas Holzinger. 2017. Human Activity Recognition Using Recurrent Neural Networks. In *Machine Learning and Knowledge Extraction*. Springer, Cham, Switzerland, 267–274. https://doi.org/10.1007/978-3-319-66808-6_18

[22] Robert L. Thorndike. 1953. Who belongs in the family? *Psychometrika* 18, 4 (Dec. 1953), 267–276. https://doi.org/10.1007/BF02289263

[23] Aditya Virmani and Muhammad Shahzad. 2017. Position and orientation agnostic gesture recognition using wifi. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*. 252–264.

[24] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. 2005. *Algorithmic learning in a random world*. Springer Science & Business Media.

[25] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. 2005. Conformal prediction. *Algorithmic learning in a random world* (2005), 17–51.

[26] Fangxin Wang, Wei Gong, and Jiangchuan Liu. 2018. On spatial diversity in WiFi-based human activity recognition: A deep learning-based approach. *IEEE Internet of Things Journal* 6, 2 (2018), 2035–2047.

[27] Wei Wang, Alex X Liu, and Muhammad Shahzad. 2016. Gait recognition using wifi signals. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (ACM Ubicomp)*. 363–373.

[28] Wei Wang, Alex X Liu, Muhammad Shahzad, Kang Ling, and Sanglu Lu. 2015. Understanding and modeling of wifi signal based human activity recognition. In *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking (ACM MobiCom)*. 65–76.

[29] Wei Wang, Alex X Liu, Muhammad Shahzad, Kang Ling, and Sanglu Lu. 2017. Device-free human activity recognition using commercial WiFi devices. *IEEE Journal on Selected Areas in Communications* 35, 5 (2017), 1118–1131.

[30] Yan Wang, Jian Liu, Yingying Chen, Marco Gruteser, Jie Yang, and Hongbo Liu. 2014. E-eyes: device-free location-oriented activity identification using fine-grained WiFi signatures. In *Proceedings of the 20th annual international conference on Mobile computing and networking (ACM MobiCom)*. 617–628.

[31] Yunze Zeng, Parth H Pathak, and Prasant Mohapatra. 2016. WiWho: WiFi-based Person Identification in Smart Spaces. In *Proceedings of 15th ACM/IEEE International Conference on Information Processing in Sensor Networks (IEEE IPSN)*. 1–12.

[32] Shuangjiao Zhai, Zhanyong Tang, Petteri Nurmi, Dingyi Fang, Xiaojiang Chen, and Zheng Wang. 2021. RISE: robust wireless sensing using probabilistic and statistical assessments. In *MobiCom '21: Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*. Association for Computing Machinery, New York, NY, USA, 309–322. https://doi.org/10.1145/3447993.3483253

[33] Jie Zhang, Zhanyong Tang, Meng Li, Dingyi Fang, Petteri Nurmi, and Zheng Wang. 2018. CrossSense: Towards cross-site and large-scale WiFi sensing. In *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*. 305–320.

[34] Mingmin Zhao, Fadel Adib, and Dina Katabi. 2016. Emotion recognition using wireless signals. In *Proceedings of the 22nd annual international conference on mobile computing and networking*. 95–108.

[35] Yue Zheng, Yi Zhang, Kun Qian, Guidong Zhang, Yunhao Liu, Chenshu Wu, and Zheng Yang. 2019. Zero-Effort Cross-Domain Gesture Recognition with Wi-Fi. In *MobiSys '19: Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services*. Association for Computing Machinery, New York, NY, USA, 313–325. https://doi.org/10.1145/3307334.3326081