# A. Multi-Task Network Architecture

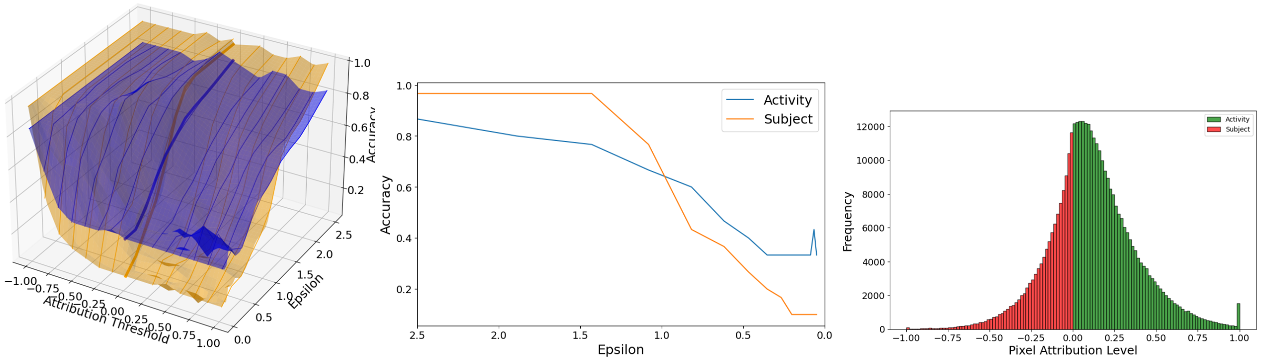
A multi-task network that is able to perform both HAR and subject identification has been constructed. Its architecture is summarised in Table A.1. The backbone ResNet18 convolutional layers was used to extract features from each sample. Subsequently, these features are fed into two branches, allowing for a sample's subject and activity classification.

| *Layer Name* | *Output Size* | *Description* |
|---|---|---|
| **ResNet Features Extraction** | | |
| Conv2d | 64 | 7x7 stride 2 |
| MaxPool2d | 64 | 3x3 stride 1 |
| 4 x Conv2d | 64 | 3x3 stride 1 |
| 4 x Conv2d | 128 | 3x3 stride 2 |
| 4 x Conv2d | 256 | 3x3 stride 1 |
| **1. Subject Branch** | | |
| 4 x Conv2d | 512 | 3x3 stride 1 |
| AvgPool2d | 512 | Adaptive average pooling |
| Flatten | 512 | Convert to a vector |
| Linear | 5/10 | Fully connected layer |
| **2. Activity Branch** | | |
| 4 x Conv2d | 512 | 3x3 stride 1 |
| AvgPool2d | 512 | Adaptive average pooling |
| Flatten | 512 | Convert to a vector |
| Linear | 3 | Fully connected layer |

Table A.1. **Multi-Task Network Architecture Summary.**

# B. Effects of Attribution Threshold and Epsilon for Sal, GradC, IG, IIG and ISG

Figure B.1, Figure B.2, Figure B.3, Figure B.4, Figure B.5 show the impact of various $\epsilon$ and attribution threshold values on the performance of user and activity recognition using the multi-task model. DP noise is driven by the attribution methods Sal, GradC, IG, IIG, ISG, respectively.

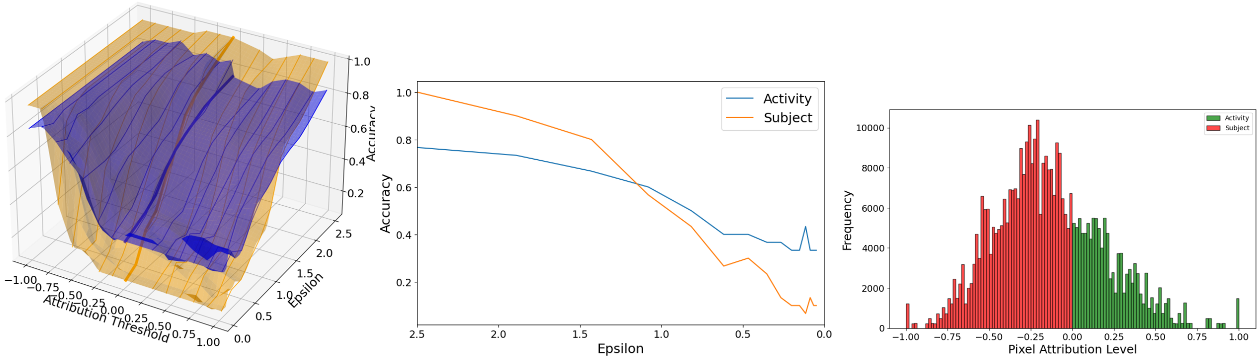


(a) $\epsilon$ and Attribution Thresholds performance

(b) Attribution Threshold 0.00025 against $\epsilon$

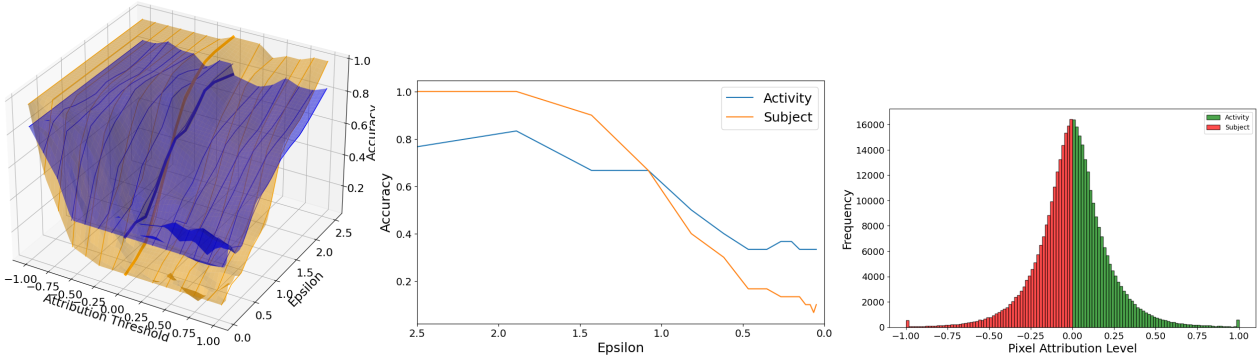
(c) Pixel Attribution for Subject and Activity

Figure B.1. **Impact of Different Saliency (Sal) Attribution Thresholds and $\epsilon$ Levels on the performance of HAR and subject recognition.** a) Multi-task models performance across various $epsilon$ and attribution threshold values, b) Multi-task models performance across various $epsilon$ with attribution threshold equal to 0.00025, c) Histogram of pixel attributions for Subject and Activity recognition.



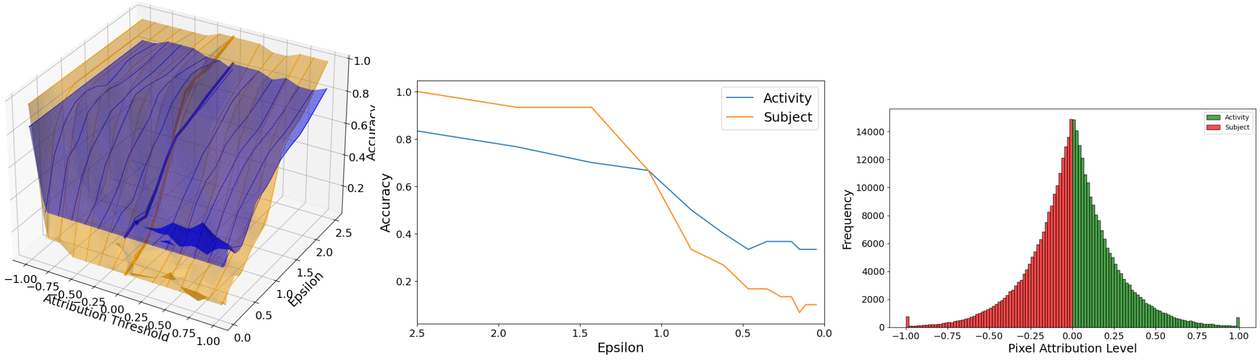
(a) $\epsilon$ and Attribution Thresholds performance (b) Attribution Threshold 0.00025 against $\epsilon$ (c) Pixel Attribution for Subject and Activity

Figure B.2. **Impact of Different Gradcam (GradC) Attribution Thresholds and $\epsilon$ Levels on the performance of HAR and subject recognition.** a) Multi-task models performance across various $epsilon$ and attribution threshold values, b) Multi-task models performance across various $epsilon$ with attribution threshold equal to 0.00025, c) Histogram of pixel attributions for Subject and Activity recognition.



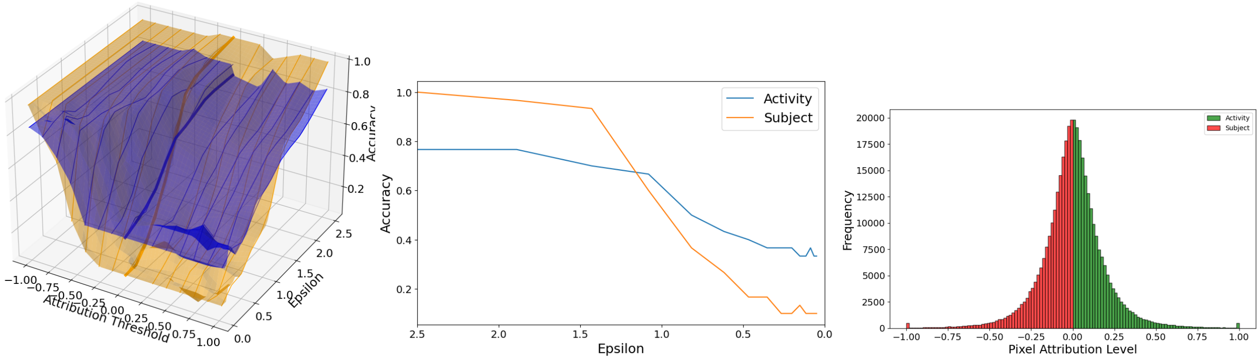
(a) $\epsilon$ and Attribution Thresholds performance (b) Attribution Threshold 0.00025 against $\epsilon$ (c) Pixel Attribution for Subject and Activity

Figure B.3. **Impact of Different Integrated Gradient (IG) Attribution Thresholds and $\epsilon$ Levels on the performance of HAR and subject recognition.** a) Multi-task models performance across various $epsilon$ and attribution threshold values, b) Multi-task models performance across various $epsilon$ with attribution threshold equal to 0.00025, c) Histogram of pixel attributions for Subject and Activity recognition.



(a) $\epsilon$ and Attribution Thresholds performance (b) Attribution Threshold 0.00025 against $\epsilon$ (c) Pixel Attribution for Subject and Activity

Figure B.4. **Impact of Different Integrated InputX Gradient (IIG) Attribution Thresholds and $\epsilon$ Levels on the performance of HAR and subject recognition.** a) Multi-task models performance across various $epsilon$ and attribution threshold values, b) Multi-task models performance across various $epsilon$ with attribution threshold equal to 0.00025, c) Histogram of pixel attributions for Subject and Activity recognition



(a) $\epsilon$ and Attribution Thresholds performance (b) Attribution Threshold 0.00025 against $\epsilon$ (c) Pixel Attribution for Subject and Activity

Figure B.5. **Impact of Different Integrated Integrated Smooth Gradient (ISG) Attribution Thresholds and $\epsilon$ Levels on the performance of HAR and subject recognition.** a) Multi-task models performance across various $epsilon$ and attribution threshold values, b) Multi-task models performance across various $epsilon$ with attribution threshold equal to 0.00025, c) Histogram of pixel attributions for Subject and Activity recognition.

## C. Implementation Details

For a fair comparison, the multi-task/HAR baseline model, Optics, IDG-DP and other tested DP-based attribution methods, were trained with 32 batch size, 400 epochs using cross-entropy loss and a Stochastic Gradient Descent (SGD) optimizer with a learning rate of 0.001 and a momentum value of 0.9. An early stopping mechanism with a patience of 8 epochs and a minimum delta of 0.01 was implemented to ensure better model convergence. The Captum library [45] was used for generating attributions, except for the IDG where the saliency function in [2] was utilized to generate attributions.

For the optic masking, the convolutional 2D layer was used to generate masks, the noise strength value used is 0.5, this mask is added to the original input to create a noisy input. The optic mask model were trained using 600 epochs with the same optimizer, momentum, learning rate and loss function with the baseline model.

Table C.2 describes the number of samples used for each evaluation procedure. The 60 blueand 25 samples used in the shadow model black-box MIA are for generating member and non-member data to evaluate the success rate of the MIA attack. Additionally, Label-only-10 has 10 samples for evaluation, while Label-only-20 consists of 20 samples for evaluating the success rate of the attack.

Table C.2. Distribution of training and testing samples for each implementation procedure.

| Procedure | Training | Testing |
|---|---|---|
| HAR | 60 | 30 |
| Blackbox MIA | 25 | 25 |
| Rule-based MIA | 25 | 25 |
| Blackbox MIA with 3 shadow model | 60 | 30 |
| Blackbox MIA with 10 shadow model | 25 | 25 |
| Label-Only 25 | 25 | 25 |
| Label-Only 10 | 20 | 20 |
| Label-Only 20 | 40 | 30 |

*Evaluation samples for Label-Only 10 = 10, and 20 for Label-Only 20.

## D. Performance Comparison with the inclusion of Baseline DP (Base-DP) against HAR and tested attacks

Laplace noise was introduced to the baseline multi-task model to create the Base-DP model (without attributions guidance) for comparative analysis with the tested methods. An attribution threshold of 0.00025 and an $\epsilon$ value of 1.20 were applied in constructing the Base-DP model.

[2] https://github.com/chasewalker26/Integrated-Decision-Gradients

For completeness, we added the HAR performance of Base-DP along with the performance of various attacks on the Base-DP model on all the results presented in the Results section.

The investigation results in Table D.3 demonstrate IDG's effective utility and resistance to black-box MIA attacks based on shadow models. IDG demonstrate better performance compared with tested methods, including the baseline with DP especially in terms of utility, attack accuracy and precision.

Table D.3. Performance evaluation comparison for the baseline and various multi-task DP based activity models against black-box MIA using shadow models. Shadow dataset size = 60, $epsilon = 1.20$ attribution threshold = 0.00025 For Base-DP, IG-DP, Sal-DP, IIG-DP, ISG-DP and IDG-DP. noise mask for Optics = 0.50

| Model | Clean Test Accuracy (%)↑ | Total Attack Accuracy (%)↓ | Total Attack Precision (%)↓ | Total Attack Recall (%)↓ |
|---|---|---|---|---|
| Baseline | 83.33 | 56.67 | 54.17 | 86.67 |
| Base-DP | 90.00 | 56.67 | 55.00 | 73.33 |
| Optics | 63.33 | 63.33 | 64.29 | 60.00 |
| IG-DP | 93.33 | 56.67 | 55.00 | 73.33 |
| GradC-DP | 80.00 | 56.67 | 54.17 | 86.67 |
| Sal-DP | 80.00 | 56.67 | 55.00 | 73.33 |
| IIG-DP | 90.00 | **36.67** | **37.50** | **40.00** |
| ISG-DP | 90.00 | 46.47 | 45.45 | **33.33** |
| IDG-DP | **96.70** | **36.67** | 37.50 | 40.00 |

Figure D.6 (a) illustrates the results of a Label-Only attack [9] evaluated using the data distribution for Label-Only 25 as reported in Table C.2. IDG-DP consistently outperforms all tested methods in terms of clean test accuracy. In attack accuracy, IDG-DP and Base-DP exhibit similar performance, yet IDG-DP demonstrates superior utility. Figure D.6 further demonstrates the utility and attack performance of IDG-DP and Base-DP for Label-Only 20, with additional data records provided in Table C.2.

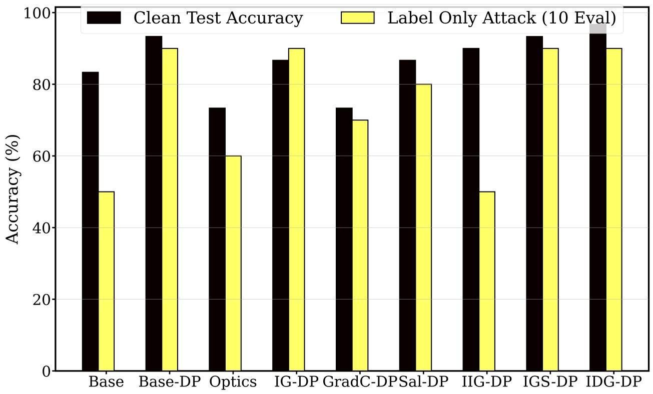


Figure D.6. Models performance comparison against Label-Only MIA: Label Only MIA, attack training size = 20, attack test size = 20, evaluation size = 10, $epsilon = 1.20$ attribution threshold = 0.00025 For Base-DP, IG-DP, Sal-DP, IIG-DP, ISG-DP and IDG-DP. noise mask for Optics = 0.50