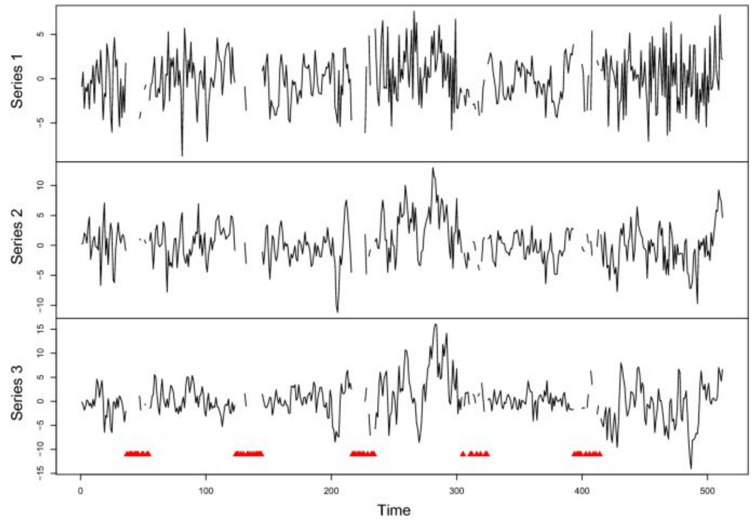
Mitigating missing data for Human Activity Recognition

Team:
Rushi Bhatt
Ronak Kaoshik
Jinru Cai

Overall Project Goals

- Building robust models for missing data imputation and downstream task of HAR classification



Specific Aims

1. Investigate and artificially create different types of missingness for HAR：
    a. MCAR
    b. MAR
    c. MNAR
2. Train and fine-tuning models for Imputation task
    a. Definition
    b. Methods and models
    c. Evaluation
3. Train and fine-tuning models for Classification task
    a. Definition
    b. Methods and models
    c. Evaluation

Current State of the Art and Limitations

- Only baseline approaches like mean/median/KNN based methods have been tested on UCI HAR and PAMAPS2 dataset for imputation+classification related tasks.
- BRITS is the current SOTA approach that has explored the data imputation problem in human activity recognition domain particularly for localization (spatial coordinates data).
- Additionally, BRITS has many limitations like lengthy training cycle, lack of support for training on data with missingness in the raw signal, compounding error due to being auto-regressive.

Novelty and Importance of our approach

- Our approach is inspired by SAITS, a self-attention based network that derives its training objectives from the likes of masked language modeling.
- Our approach is the first of its kind that effectively deals with the tasks of imputation and multi-class human activity recognition. We have extensively tested it on both UCI HAR and PAMAPS2 dataset.
- Our approach has faster training and inference time cycles as compared to BRITS.
- Our approach performs consistently better than BRITS when tested across variable missingness settings.
- Our approach has direct usage potential on datasets where the raw data itself contains missing information.

Technical Approach: Dataset

- UCI HAR [1]
  - 6 activities (WALKING, WALKING_UPSTAIRS, WALKING_DOWNSTAIRS, SITTING, STANDING, LAYING)
  - 128 readings/ window (~10k time series)
  - 3-axis accelerometer and 3-axis gyroscope
- PAMAP 2 [2, 3]
  - 24 activities
  - 52 features
    - Heart Beat
    - IMU hand (2 x 3-axis accelo, 3-axis gyro, 3-axis magneto, orientation)
    - IMU chest
    - IMU ankle
  - We found window size of 40, as best for imputation and classification performance

[1] Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra and Jorge L. Reyes-Ortiz. A Public Domain Dataset for Human Activity Recognition Using Smartphones. 21th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, ESANN 2013. Bruges, Belgium 24-26 April 2013

[2] A. Reiss and D. Stricker. Introducing a New Benchmarked Dataset for Activity Monitoring. The 16th IEEE International Symposium on Wearable Computers (ISWC), 2012.

[3] A. Reiss and D. Stricker. Creating and Benchmarking a New Dataset for Physical Activity Monitoring. The 5th Workshop on Affect and Behaviour Related Assistance (ABRA), 2012.

Data Missingness

Types of data missingness:

- Missing completely at random (MCAR), missing values are independent of any other values
- Missing at random (MAR), missing values depend only on observed values
    - E.g., missingness rate for walking is higher than that for standing?
    - E.g., missingness rate for sensor X is higher than sensor Y?
- Missing not at random (MNAR), missing values depend on both observed and unobserved values
    - E.g., if you run very fast (more acceleration), missingness is likely to be higher than normal run

Synthetically introducing MCAR, MAR, and MNAR

- **MCAR**: Generates missingness completely at random based on a given missingness rate ‘p’
- **MAR** (with logistic masking model): First, a subset of variables with no missing values is randomly selected. The remaining variables have missing values according to a logistic model with random weights, re-scaled so as to attain the desired proportion of missing values on those variables.
- **MNAR** (with logistic masking model[4]): It is implemented by selecting missingness probabilities with a logistic model, taking all variables as inputs. Hence, values that are inputs can also be missing.

[4] Yoon, J., Jordon, J. & Schaar, M.. (2018). GAIN: Missing Data Imputation using Generative Adversarial Nets. Proceedings of the 35th International Conference on Machine Learning, in Proceedings of Machine Learning Research 80:5689-5698 Available from https://proceedings.mlr.press/v80/yoon18a.html.

Baseline Imputation methods

- **Zero imputation -** Filling the missing data directly with 0.
- **Mean imputation -** Filling the missing data with mean of the values in the window.
- **Median imputation -** Filling the missing data with median of the values in the window.

BRITS[5]

Unidirectional Uncorrelated Recurrent Imputation RITS-I

time series $\mathbf{X}$

31	/	/	32	27	22
6	17	/	/	/	13
/	107	/	87	66	90

$\mathbf{x}_1$ $\mathbf{x}_2$ ...... $\mathbf{x}_6$

masking vectors

1	0	0	1	1	1
1	1	0	0	0	1
0	1	0	1	1	1

$\mathbf{m}_1$ $\mathbf{m}_2$ ...... $\mathbf{m}_6$

$$\frac{1}{T}\sum_{t=1}^{T}\ell_t + \mathcal{L}_{out}(\mathbf{y},\hat{\mathbf{y}})$$

$$\ell_t = \langle \mathbf{m}_t, \mathcal{L}_e(\mathbf{x}_t, \hat{\mathbf{x}}_t) \rangle .$$

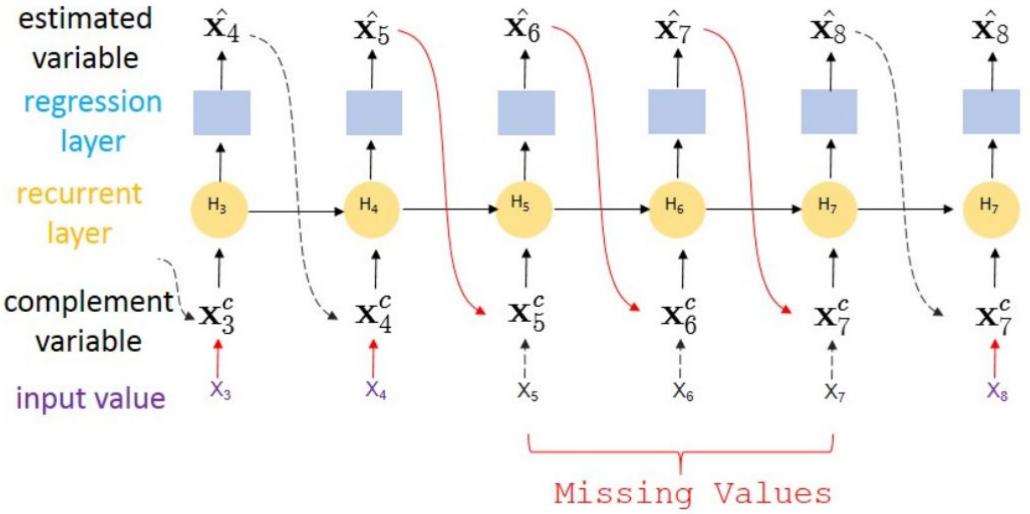


Figure 2: Imputation with unidirectional dynamics.

[5] W. Cao, D. Wang, J. Li, H. Zhou, L. Li, και Y. Li, 'BRITS: Bidirectional Recurrent Imputation for Time Series', στο Advances in Neural Information Processing Systems, 2018, τ. 31.

Our approach - SAITS[6]

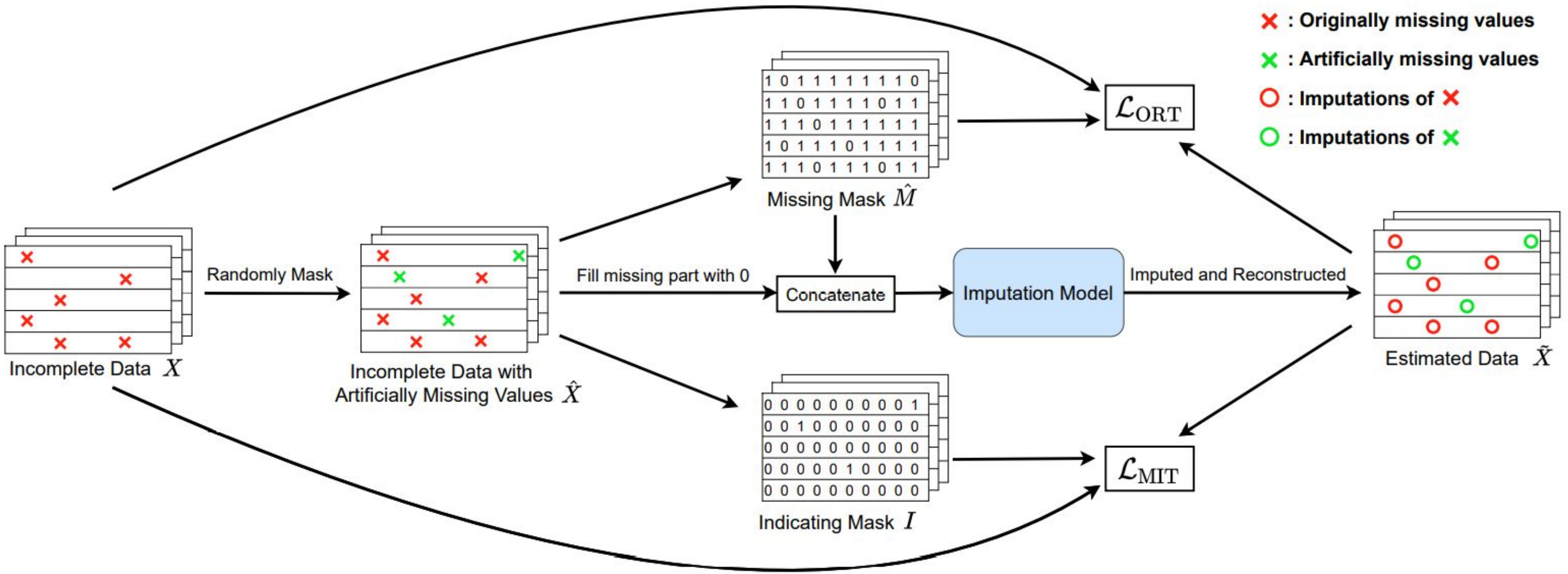
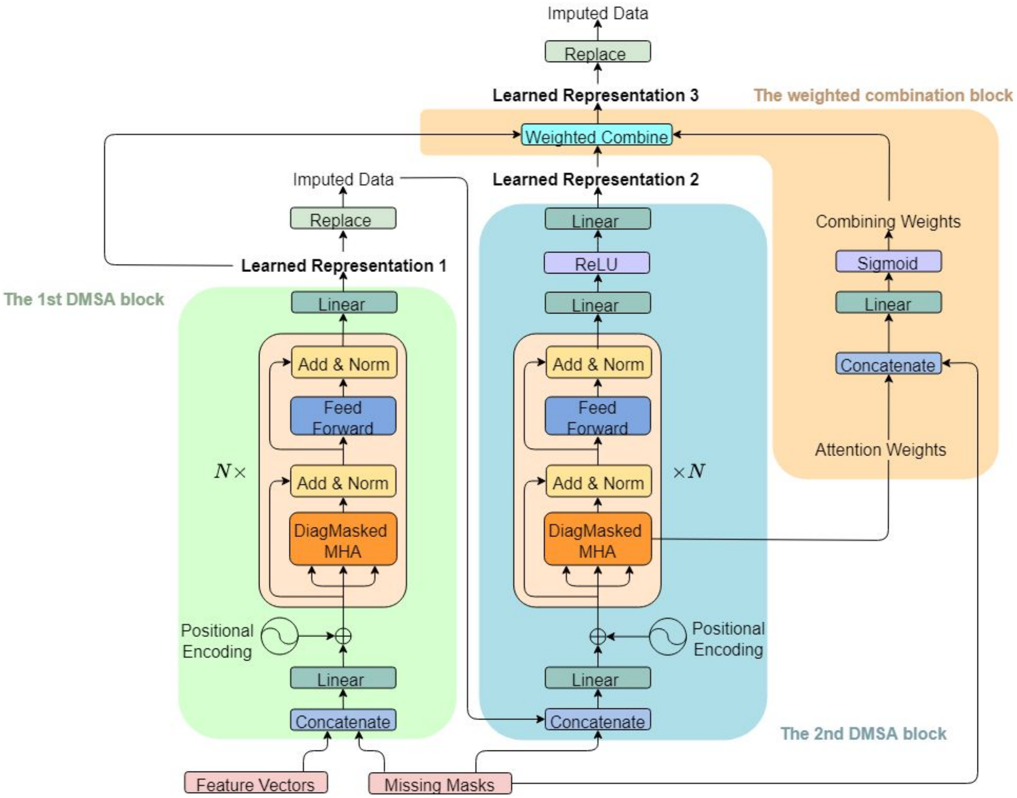


Figure 1: A graphical overview of the joint-optimization training approach.

[5] W. Du, D. Côté, και Y. Liu, 'SAITS: Self-Attention-based Imputation for Time Series'. arXiv, 2022.

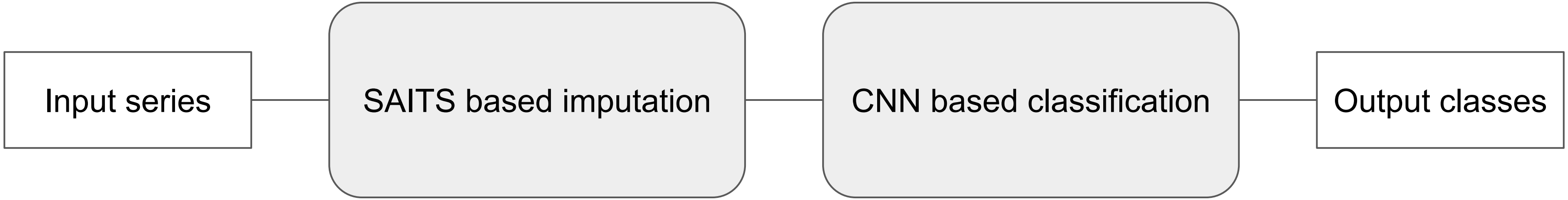
Our approach - SAITS[6]



$$\mathcal{L} = \mathcal{L}_{\mathrm{ORT}} + \lambda\, \mathcal{L}_{\mathrm{MIT}}$$

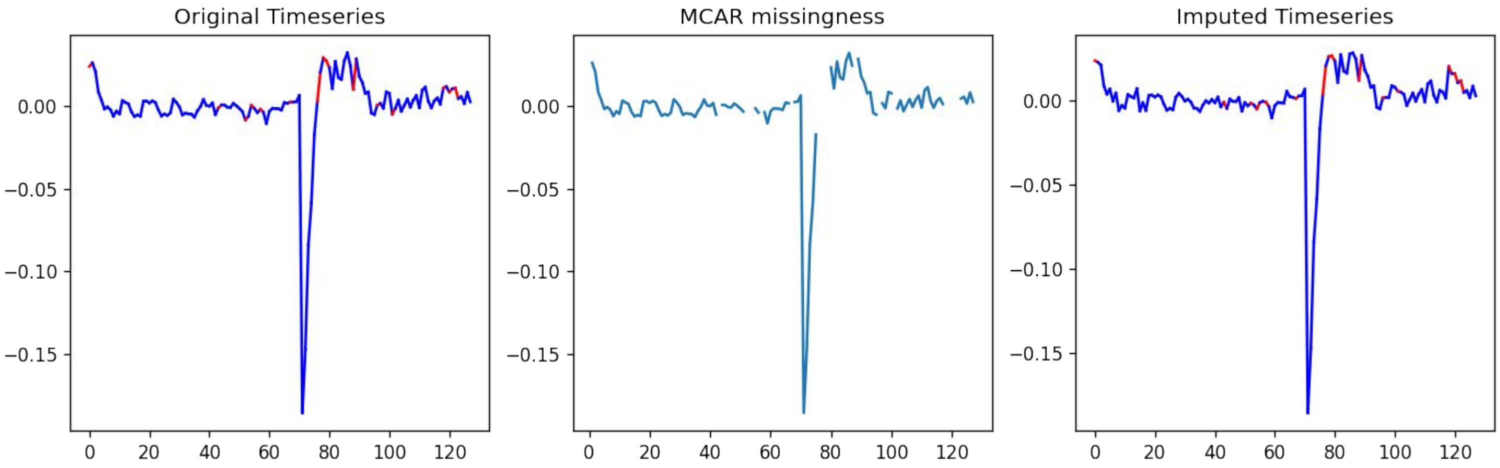
Our approach - Classification

- Both BRITS and SAITS have made use of RNN based classifiers for their downstream recognition tasks.
- We observed that RNNs tend to take a long time for training on our HAR datasets.
- Hence, we made use of a CNN based classifier.
- For reference, our CNN based classifier took about **15 seconds/epoch** whereas the compex RNN classifiers were taking upto **8-9 minutes/epoch**.



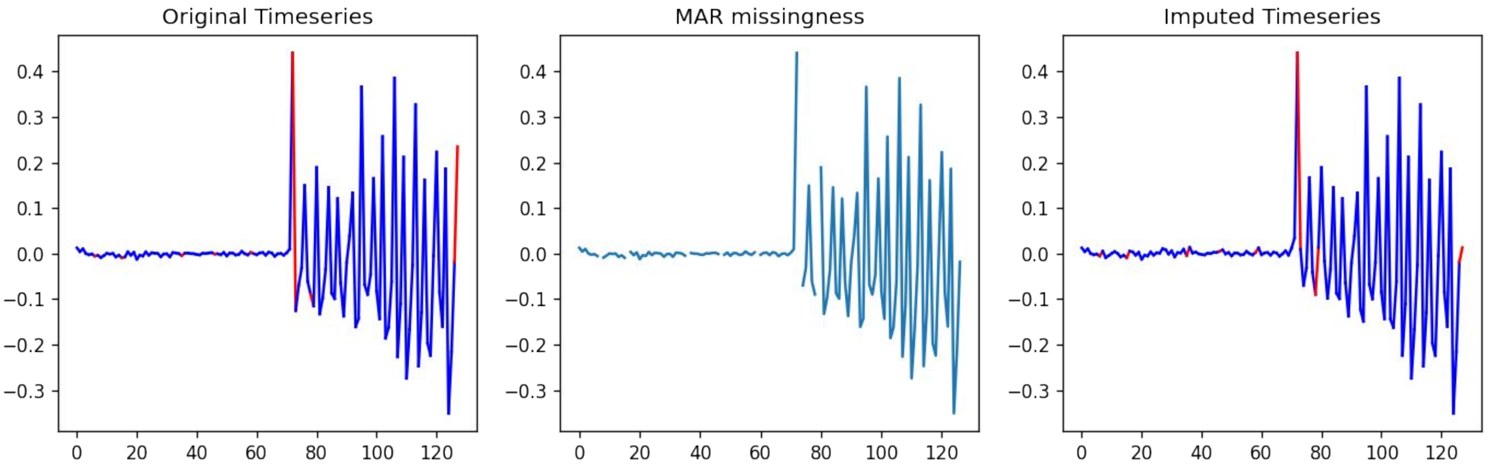
Imputation Performance

Performance of SAITS on UCI HAR data with MCAR



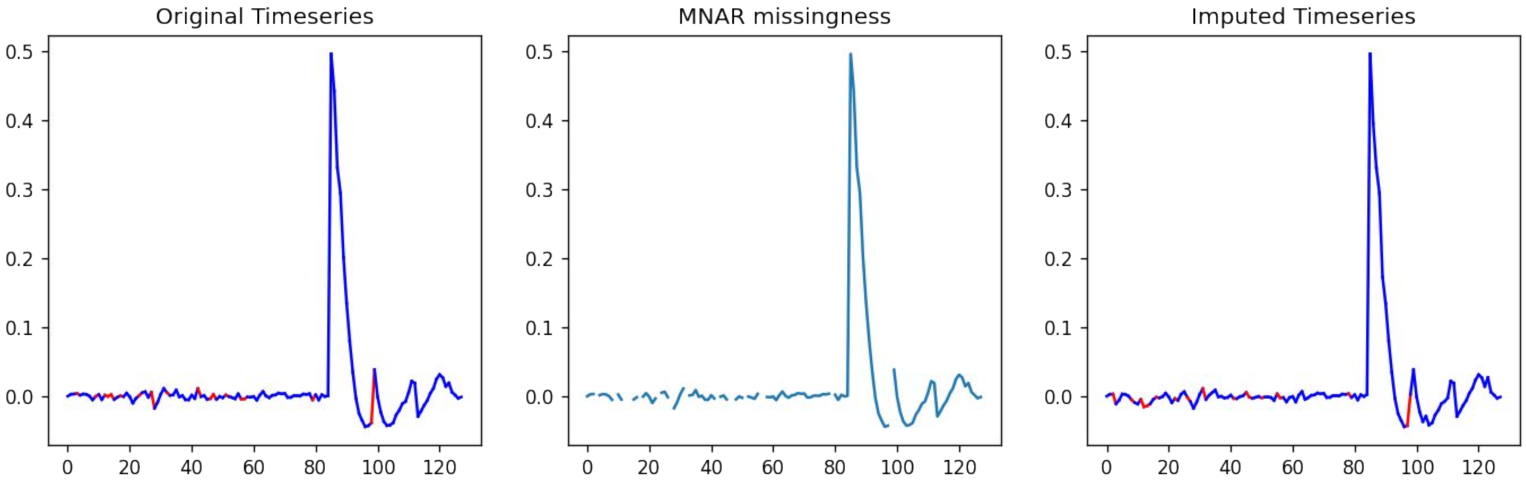
Imputation Performance

Performance of SAITS on UCI HAR data with MAR



Imputation Performance

Performance of SAITS on UCI HAR data with MNAR



Evaluation metrics

Metrics for imputation

$$\mathrm{MAE}\left(estimation, target, mask\right) = \frac{\sum_d^D \sum_t^T |(estimation - target) \odot mask|_t^d}{\sum_d^D \sum_t^T mask_t^d}$$

$$\mathrm{RMSE}\left(estimation, target, mask\right) = \sqrt{\frac{\sum_d^D \sum_t^T \left(((estimation - target) \odot mask)^2\right)_t^d}{\sum_d^D \sum_t^T mask_t^d}}$$

$$\mathrm{MRE}\left(estimation, target, mask\right) = \frac{\sum_d^D \sum_t^T |(estimation - target) \odot mask|_t^d}{\sum_d^D \sum_t^T |target \odot mask|_t^d}$$

Metrics for Classification

Accuracy, and F1-Score

Results

	MCAR			MAR			MNAR		
	MAE	RMSE	MRE	MAE	RMSE	MRE	MAE	RMSE	MRE
Zero Imputation	0.242178	0.413517	2.86e+14	0.241268	0.412009	2.86e+14	0.241828	0.413479	2.87e+14
Mean Imputation	0.777366	0.898209	9.49e-01	0.777086	0.897553	9.48e-01	0.775993	0.897156	9.48e-01
Median Imputation	0.756448	0.875055	9.51e-01	0.756012	0.874249	9.51e-01	0.755060	0.873975	9.50e-01
BRITS Imputation	0.123852	0.221070	5.11e-01	0.128857	0.225485	5.34e-01	0.123321	0.217309	5.09e-01
SAITS Imputation	0.032241	0.076779	1.33e-01	0.036456	0.083063	1.50e-01	0.035134	0.079133	1.45e-01

Table 1:Imputation performance for MCAR, MAR, and MNAR missingness type for UCI HAR dataset (10 % missingness rate)

Results

	MCAR		MAR		MNAR	
	Acc.	F1-Score	Acc.	F1-Score	Acc.	F1-Score
Zero Imputation	25.22%	0.22	27.84%	0.23	18.92%	0.17
Mean Imputation	44.23%	0.46	38.66%	0.40	42.25%	0.45
Median Imputation	45.64%	0.48	33.27%	0.35	40.42%	0.42
BRITS Imputation	91.17%	0.94	86.62%	0.88	89.85%	0.90
SAITS Imputation	99.16%	0.99	99.16%	0.99	97.16%	0.97

Table 2: Classification performance for MCAR, MAR, and MNAR missingness type for UCI HAR dataset (10 % missingness rate)

Results

	Zero Imputation			Mean Imputation			Median Imputation			BRITS Imputation			SAITS Imputation		
	MAE	RMSE	MRE	MAE	RMSE	MRE	MAE	RMSE	MRE	MAE	RMSE	MRE	MAE	RMSE	MRE
10 %	0.241908	0.412919	2.86e+14	0.776911	0.897924	0.94e+00	0.755901	0.874692	0.95e+00	0.071421	0.189044	1.01e+00	0.032241	0.076779	1.33e-01
20 %	0.265971	0.454341	6.31e+14	0.854312	0.987369	1.04e+00	0.831223	0.961850	1.04e+00	0.103246	0.184325	9.57e-01	0.039748	0.088735	1.64e-01
30 %	0.290546	0.495989	1.03e+15	0.932390	1.077443	1.33e+00	0.907377	1.049769	1.14e+00	0.095643	0.186893	9.97e-01	0.040181	0.098522	1.66e-01
40 %	0.314458	0.536854	1.49e+15	1.009885	1.167183	1.23e+00	0.982692	1.137118	1.23e+00	0.113654	0.189329	9.99e-01	0.039748	0.088735	1.64e-01
50 %	0.338482	0.578073	2.00e+15	1.088094	1.257354	1.33e+00	1.058985	1.225220	1.33e+00	0.113654	0.19139	1.10e00	0.046216	0.111676	1.91e-01
60 %	0.362732	0.619305	2.58e+15	1.165613	1.347124	1.423687	1.134844	1.313155	1.42e+00	0.115624	0.198277	1.13e+01	0.063642	0.136941	2.63e-01
70 %	0.386891	0.660723	3.21e+15	1.243150	1.436976	1.51e+00	1.211076	1.401609	1.52e+00	0.115988	0.199835	1.23e+00	0.063642	0.136941	2.636e-01
80 %	0.411090	0.702014	3.90e+15	1.320939	1.527140	1.61e+0	1.288818	1.491931	1.61e+00	0.118903	0.21341	1.83e+00	0.063642	0.136941	2.63e-01
90 %	0.435221	0.743202	4.67e+15	1.400586	1.709149	1.37e+00	1.370342	1.713084	2.97e+00	0.123216	0.23548	2.13e+00	0.063642	0.136941	2.63e-01

Table 3: Imputation performance over varying missingness rate for UCI HAR dataset (MCAR)

Results

	BRITS Imputation		SAITS Imputation	
	Acc.	F1-Score	Acc.	F1-Score
10%	91.17%	0.94	99.16%	0.99
20%	90.73%	0.92	96.67%	0.96
30%	82.62%	0.81	91.33%	0.90
40%	76.86%	0.74	82.97%	0.81
50%	68.43%	0.66	75.97%	0.76
60%	59.14%	0.56	63.74%	0.62
70%	48.22%	0.46	56.22%	0.56
80%	43.86%	0.39	50.12%	0.51
90%	39.45%	0.37	42.36%	0.43

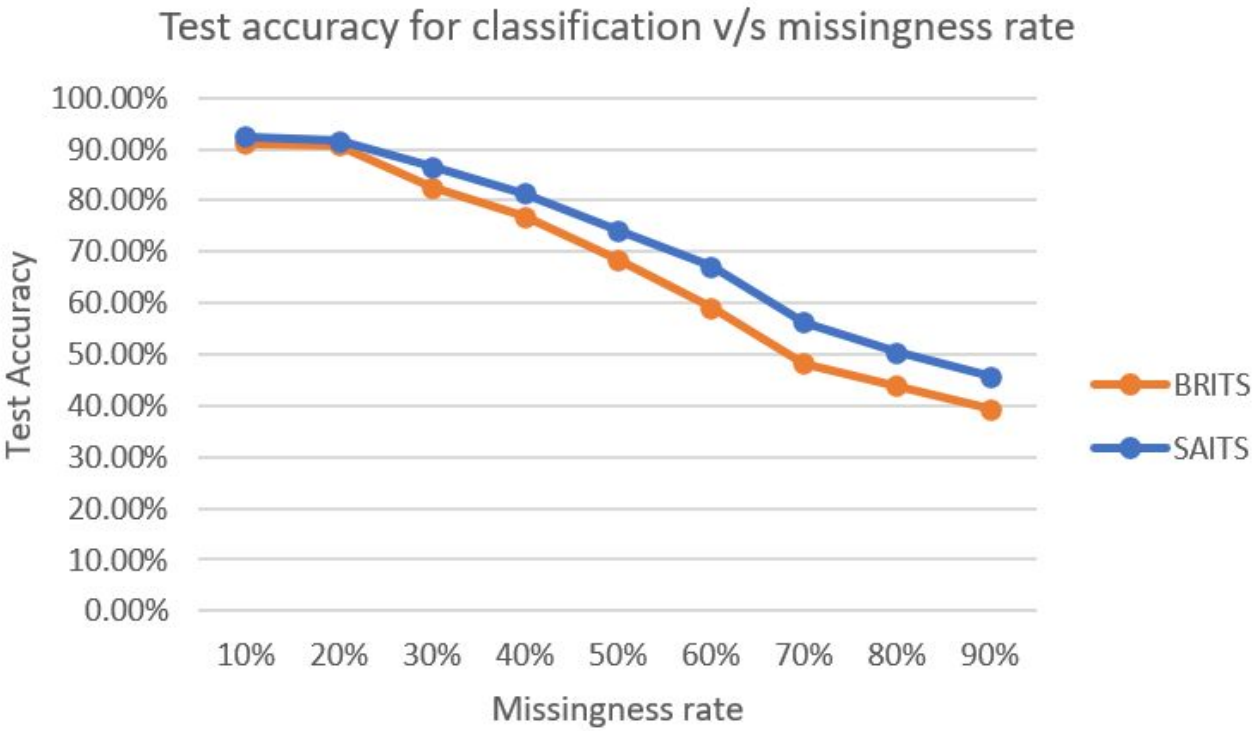
Table 4: Classification performance over varying missingness rate for UCI HAR dataset (MCAR)

Results

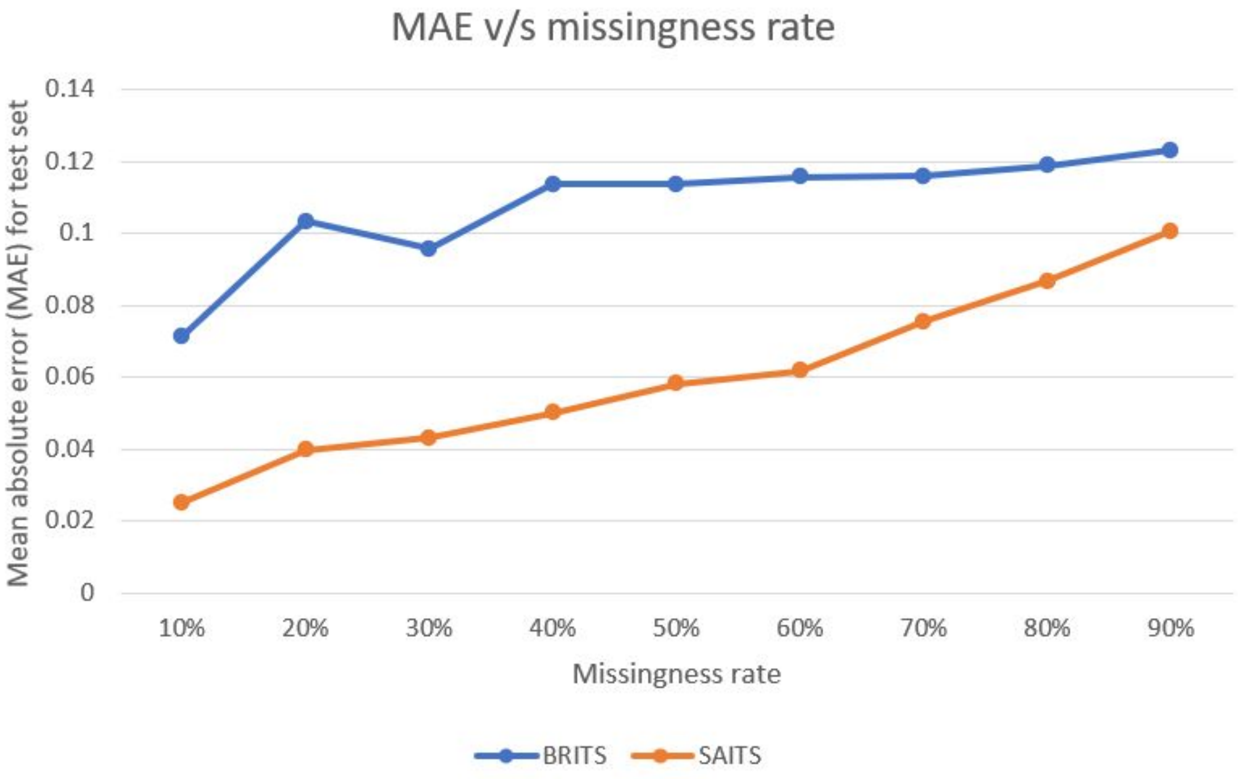
	MCAR			MAR			MNAR		
	MAE	RMSE	MRE	MAE	RMSE	MRE	MAE	RMSE	MRE
Zero Imputation	43.001326	296.209078	1.27e+18	44.795440	302.214721	1.32e+18	43.000029	296.078032	1.26e+18
Mean Imputation	1827.534261	2135.402027	1.00e+00	1843.997157	2148.611512	1.00e+00	1824.700344	2132.157724	1.00e+00
Median Imputation	1827.516164	2135.420835	1.00e+00	1843.981177	2148.633080	1.00e+00	1824.687548	2132.176473	1.00e+00
BRITS Imputation	30.418686	147.324833	4.02e-01	31.762073	151.236273	3.68e-01	36.282232	150.060937	4.00e-01
SAITS Imputation	18.222896	61.980596	4.2e-01	17.388033	53.826534	2.33e-01	26.891733	70.600211	3.65e-01

Table 5:Imputation performance for MCAR, MAR, and MNAR missingness type for PAMAPS 2 dataset (10% missingness rate)

Results (UCI HAR): Plots for varying missingness



Results (UCI HAR): Plots for varying missingness



Discussion & Conclusion

From our evaluations, we conclude the following:

- SAITS **outperforms** BRITS by ~ 31% in MAE and achieves 2 ~ 3 times faster training speed. Furthermore, SAITS outperforms baseline imputation approaches by ~ 400% in MAE.
- SAITS is **robust** to different types of missingness - namely tested on MCAR, MAR, and MNAR with varying missingness rate
- SAITS is **multiutility** - it not only outperforms other methods on imputation metrics but also does well on downstream task of HAR classification.
- SAITS **generalizes** well as we have tested it on two popular - UCI HAR and PAMAPS2 dataset.

Future Work

- Our best classifier (CNN-based) on SAIT imputed data achieved ~92.5 % for UCI HAR, and ~91 % for PAMAPS2 (with 10% MCAR); however their existing SOTA models have achieved ~99+ % test accuracy. Therefore we see new direction of future research in on experimenting with SOTA/building better classifiers.
- IMU signals from smartphones and smartwatches are also used for various other applications like gait classification, step counting, and gesture control; therefore, the reconstructed signals from our imputation method can be utilized to improve the performance of the above mentioned tasks.

Team member contribution

1. Rushi
   a. Implemented baseline imputation methods
   b. Implemented simulation script for artificially introducing MCAR, MAR, and MNAR
   c. Implemented SAITS model and performed initial experiments
2. Ronak
   a. Implemented BRITS model
   b. Conducted experiments of varying % of missingness
   c. Implemented downstream classifier
3. Jerry
   a. Performed literature review to compare different models with current models we use.
   b. Existing code review
   c. Code for future improvement searching