A APPENDIX

A.1 Detailed Data Set Descriptions

A summary of the characteristics of each data set is provided in Table 5. In detail, we use the following data sets:

- Hospital is a benchmark data set used in the data cleaning literature (Rekatsinas et al., 2017). In our experiments, we inject missing errors instead.
- NYPD [1] contains violation crimes reported to the New York City Police Department (NYPD). We focus on the latest snapshot of the data set (June 2019) and remove some of the attributes relating only to time. In addition, since we cannot obtain ground truth in a semi-automated fashion for existing missing values, we simply remove all the rows containing missing values and later inject missing errors.
- Chicago [2] consists of data on Chicago taxi rides. Chicago contains naturally-occurring missing values. As far as we know, no continuous errors exist beyond the case where the entire tuple is missing. The primary attributes with missing values are the "Census Tracts" and "Community Area" columns which denote the US census tracts and the Chicago-defined community areas the taxi ride took place, respectively. In order to judge the performance of MDI methods on the naturally-occurring errors, ground truth for census tracts may be queried by latitude-longitude from an FCC API[3] and community areas can be subsequently matched via the census tract with a join table for community areas [4]. Since this data set is the tens of gigabytes in size we sample 40000 rows from each of the top 10 taxi companies by number of rides.
- Phase is a data set on three-phase current traces collected by a third-party company. Due to privacy policies we cannot share the source in this paper.
- We use 10 data sets from the UCI repository [5] (Dua & Graff, 2017). All of them are used as-is in the experiments except the Eye EEG data set which contains outlier values we subsequently dropped (if they are beyond three standard deviations from the mean).

[1] https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Current-Year-To-Date-/5uac-w243

[2] https://data.cityofchicago.org/Transportation/Taxi-Trips/wrvz-psewurl/

[3] https://www.fcc.gov/census-block-conversions-api

[4] http://robparal.blogspot.com/2012/04/census-tracts-in-chicago-community.html

[5] http://archive.ics.uci.edu/ml/datasets/

Table 5: Data sets used in experiments sorted by the proportion of discrete to continuous attributes.

Data Set	$\|r\|$	# Continuous Attributes	# Discrete Attributes
Tic-Tac-Toe	958	0	10
Hospital	1000	2	14
Mammogram	831	1	5
Thoracic	470	3	14
Contraceptive	1473	2	8
Solar Flare	1066	3	10
NYPD	32399	4	13
Credit	653	6	10
Australian	691	6	9
Chicago	400k	11	7
Balance	625	4	1
Eye EEG	14976	14	1
Phase	9628	4	0
CASP	45730	10	0

We present functional relationships in the above data sets in Table 7. The functional relationships fall on the spectrum of the synthetic experiments (see Section 5.5): `Trip Total` from Chicago and `D` from Phase both correspond to $k = 4$ and $k = 3$ for $f(\mathcal{X}_Y) \in$ **linear**, respectively; `class` from Balance corresponds to $k = 4$ for $f(\mathcal{X}_Y) \in$ **interact**; all other functional relationships where $\mathcal{X}_Y$ is some variant of 2-D coordinates correspond to $k = 2$ and $f(\mathcal{X}_Y) \in$ **kernel**.

A.2 Hyperparameter Tuning

Table 6: Set of hyperparameters for each model over which we perform grid search cross-validation.

Method	Hyperparameter	Search Space
AimNet	dropout %	$[0, 0.25, 0.5]$
HCQ	weight decay	$[0, 0.01, 0.1]$
XGB	gamma	$[0, 0.1, 1]$
MIDAS	keep %	$[0.8, 0.65, 0.5]$
GAIN	alpha	$[0.1, 1, 10]$
MF	# trees	$[50, 100, 300]$
MICE	# iterations	$[1, 3, 5]$

Given the large number of hyperparameters in each of the baseline methods and the numerous data sets we wish to benchmark against, it is intractable to perform a thorough hyperparameter search for every baseline method. For all baselines, we begin with their default parameters as described in their corresponding papers or specified in their open-source implementations. We choose the most influential hyperparameter for each method, as shown in Table 6, to perform grid search cross-validation across.

Table 7: Functional relationships with real-valued domains for each data set where $Y = f(\mathcal{X}_Y)$. $*$: since Y and $\mathcal{X}_Y$ are both continuous and f is linear, all permutations are also valid functional relationships.

Data Set	Y	$\mathcal{X}_Y$ (continuous)
Chicago	Pickup Census Tract	{Pickup Centroid Latitude, Pickup Centroid Longitude}
	Dropoff Census Tract	{Dropoff Centroid Latitude, Dropoff Centroid Longitude}
	Pickup Community Area	{Pickup Centroid Latitude, Pickup Centroid Longitude}
	Dropoff Community Area	{Dropoff Centroid Latitude, Dropoff Centroid Longitude}
	Trip Total*	Fare + Tips + Tolls + Extras
NYPD	ADDR_PCT_CD	{Latitude, Longitude} {X_COORD_CD, Y_COORD_CD}
	PATROL_BORO	{Latitude, Longitude} {X_COORD_CD, Y_COORD_CD}
	BORO_NM	{Latitude, Longitude} {X_COORD_CD, Y_COORD_CD}
Phase	D*	A + B + C
Balance	class	left_distance × left_weight − right_distance × right_weight

A.3 Training

For all experiments a default embedding size $k = 64$ is used with a maximum pruned domain size of $D = 50$ for AimNet. We always train AimNet with 20 epochs (although we observe in almost all data sets AimNet converges in fewer than 3), and each mini-batch consists of 1 sample if $|\mathcal{D}| \leq 2000$ or 32 samples if $|\mathcal{D}| > 2000$. Since AimNet uses a mini-batch approach, samples that have a smaller domain than D have negative values randomly sampled into its softmax loss since the softmax arguments are padded anyways to size D for mini-batch training.

A.4 Chicago Deep-Dive

In Figure 6 we plot the latitude-longitude coordinates for a region of Chicago from the Chicago data set and all samples in that region. We label each point on the plane with its `Pickup Census Tract` label. Note that since these are centroid `Latitudes` and `Longitudes` there are multiple samples per coordinate point. We display both the coordinates of the observed samples (which a model can train on) and the missing samples (which have coordinates but have missing census tracts). There is an apparent gap between missing and observed samples for a particular census tract. We demarcate the true boundaries of the census tracts in Figure 7 and notice that the observed and missing samples do indeed within their corresponding census tract boundaries. Upon further inspection, the systematic separation between observed and missing samples arise from different taxi companies utilizing different centroid coordinates for census tracts while having inconsistent census tract reporting standards.

A.5 MAR/MNAR Injection on Real Data Sets

Suppose the error percentage is x and e.g. $x = 20\%$. In either cases, let A be the target attribute in which we inject missing values. We uniform-randomly choose one if its continuous dependent attribute B. We then inject missing values into A according to its type:

Continuous Target (MAR injection)

1. Sort all tuples in ascending or descending order (randomly chosen) with respect to values of $t[B]$.
2. Uniform-randomly choose a contiguous interval $I_{A|B}$ in $t[B]$ equivalent to x of all tuples. $I_{A|B}$ cannot start nor end at the endpoints of $t[B]$.
3. Inject missing into all $t[A]$ cells whose co-occurring values in $t[B]$ is within $I_{A|B}$.

Discrete Target (MNAR injection)

1. Group tuples by the values of $t[A]$
2. For a given group g_i (for some value $v_i \in \text{dom}(A)$), randomly choose a cut-off value c. c will either be the x-th largest or smallest (randomly chosen) among all the co-occurring values in B (i.e. $g_i[B]$).
3. Inject missing into all $g_i[A] = v_i$ cells whose co-occurring values in attribute B is beyon the cut-off c. That is if c is the x-th largest then inject missing if $g_i[B] > c$ or if c is the x-th smallest then inject missing if $g_i[B] < c$.

A.6 Detailed Ablation Study Results

We perform an ablation study by removing the attention layer while imputing on both synthetic and MAR and MNAR injected data sets. We plot the results of removing the attention layer on the *kernel* data sets from Section 5.5 in Figure 8. For $|Y| = 1$ (each synthetic data set consists of only one set of Y and $\mathcal{X}$) in the top row the attention mechanism contributes nothing to the performance. However, once we introduce $|Y| = 5$ sets of $\mathcal{X}, Y$ we observe that as the number of classes increases, the attention mechanism eventually accounts for $> 50\%$ of the prediction accuracy. For sufficiently difficult imputation problems where the number of classes is large, and where there are other irrelevant

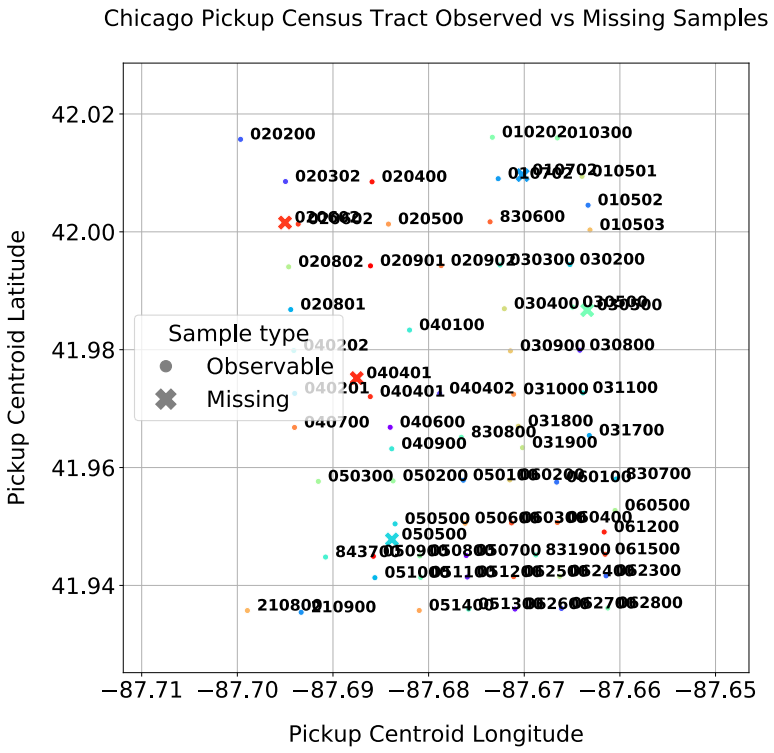

Figure 6: Observed vs missing `Pickup Census Tract` samples in the Chicago data set.

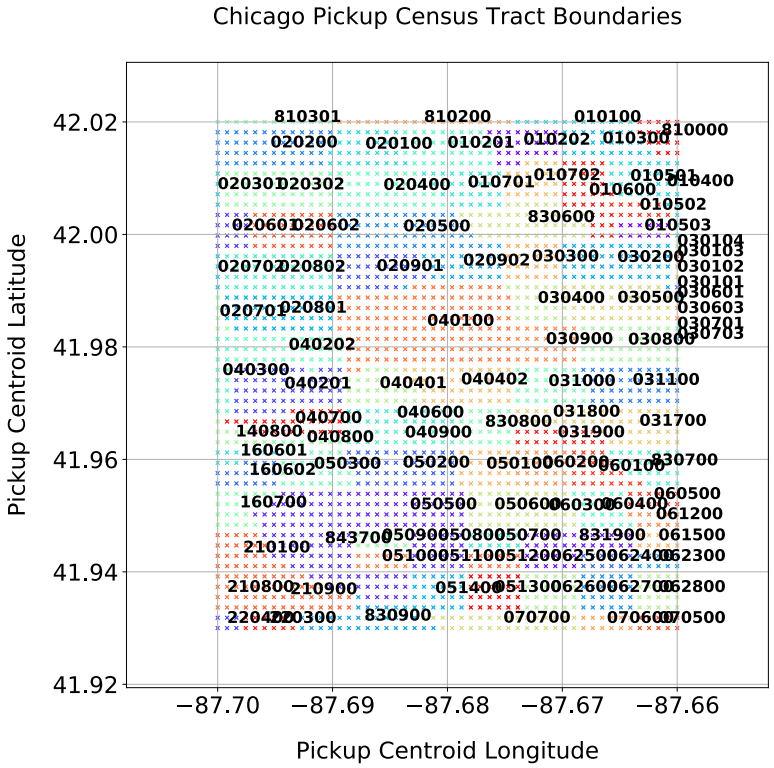

Figure 7: Approximate bounding polygons of attribute `Pickup Census Tract` for the Chicago data set.

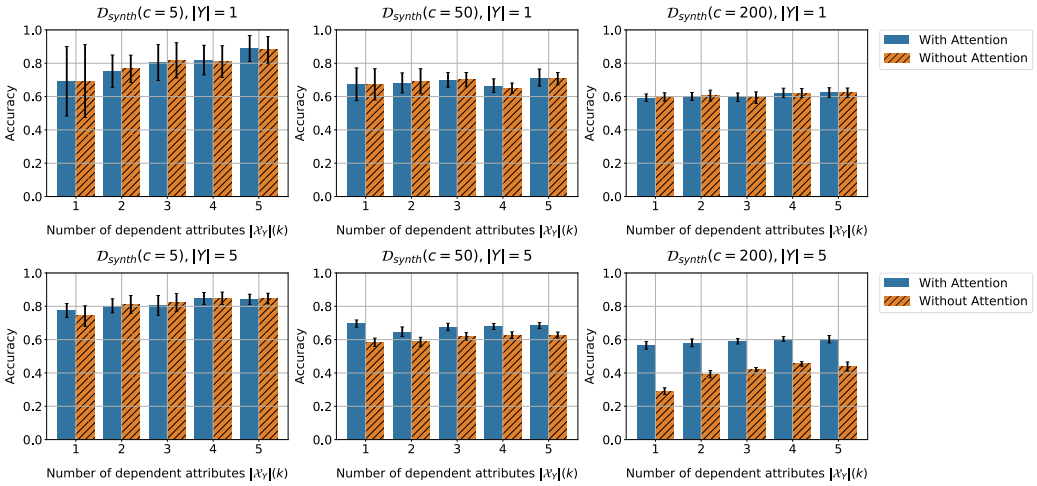

Figure 8: Accuracy on synthetic data sets of AimNet with/without attention layer with $|\mathcal{D}_{synth}| = 5000$, $|\mathcal{Y}| = 1, 5$ varying number of dependent attributes $|\mathcal{X}_Y|$ over different class sizes $c = 5, 50, 200$.

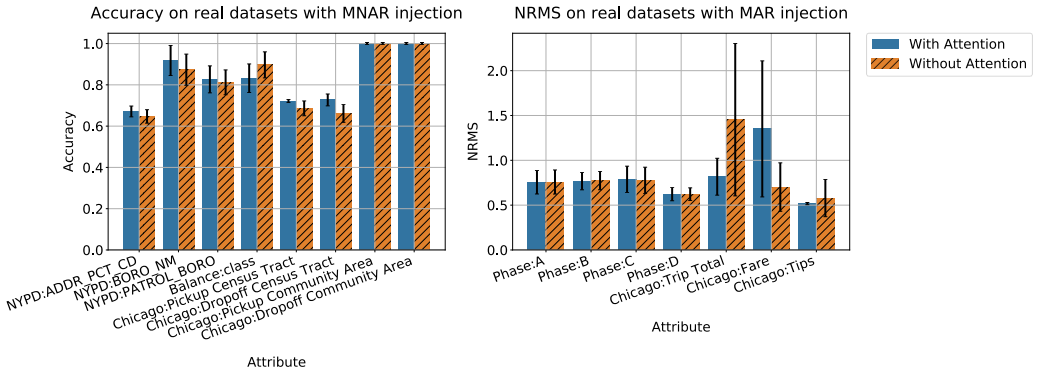

Figure 9: Imputation Accuracy and NRMS error on real data sets under MNAR/MAR injections of AimNet with and without attention layer.

attributes to the functional relationship in the data set, the attention mechanism helps identify the correct inputs $\mathcal{X}_Y$ for a given Y attribute.

We perform the same ablation study on the real-world MAR and MNAR injected data sets and plot the results in Figure 9. We find that for discrete attributes, the attention mechanism accounts for $5 - 10\%$ of the imputation accuracy on the NYPD and Chicago attributes, all of which exhibit the *kernel* functional relationship for $k = 2$. Interestingly for the `class` attribute in the Balance data set the attention layer has a non-trivial negative impact on imputation accuracy. In fact, without the attention layer AimNet would outperform all other baselines on Balance in Table 4 after accounting for noise. For continuous attributes, the attention layer has no effect except on the Chicago attributes with the functional relationships in Equation 5.

A.7 Sensitivity Analysis

We vary the hyperparameters of AimNet to assess the sensitivity of the model to hyperparameter perturbations, specifically dropout rate, max domain size and the embedding size k. The sensitivity analysis is performed on the NYPD data set with MCAR-injected missing values where $p = 0.2$ and the results are shown in Table 8. We observe that for discrete predictions and the corresponding imputation accuracy, the model can perform about the same regardless of the dropout rate, max domain size, and embedding size. In fact a smaller domain size and embedding size would improve run time and make AimNet even more competitive in a practical setting. We do observe that for continuous target attributes, dropout has a negative impact on NRMS. On the other hand, the embedding size has a negligible effect. It is therefore recommended to not use dropout for imputing continuous attributes: one may choose to train a separate AimNet with just the continuous target attributes.

Table 8: Sensitivity analysis of AimNet's hyperparameters.

Accuracy on discrete attributes for the NYPD data set				
	dropout	max domain size	embedding size	AimNet
base	0.25	50	64	0.921
(dropout rate)	0.0			0.918
	0.5			0.920
(max domain size)		10		0.917
		100		0.921
(embedding size)			16	0.920
			32	0.921
			128	**0.922**
NRMS on continuous attributes for the NYPD data set				
	dropout rate	max domain size	embedding size	AimNet
base	0.0	50	64	**0.150**
(dropout rate)	0.25			0.281
	0.5			0.509
(embedding size)			16	0.159
			32	**0.150**
			128	0.153

A.8 Error Percentage Analysis

We additionally vary the percentage of missing values for $p = 0.4, 0.6$ for MCAR-injected errors (in addition to $p = 0.2$ in the main experiments). The results for $p = 0.4$ and $p = 0.6$ are tabulated in Table 9 and Table 10, respectively. Unsurprisingly the accuracy and NRMS of all methods deteriorate with more missing values. We observe however that AimNet maintains its lead compared to the other baselines and in fact marginally outperforms XGB and MF on continuous imputation on the CASP data set when $p = 0.6$. This suggests that not only is AimNet competitive regardless of the missingness proportion but it in fact outpaces the baselines empirically when missingness increases.

Table 9: Imputation accuracy and NRMS error on the test set under **MCAR** error injections with $p = 0.4$ missingness across our AimNet model and the other baselines.The means of 10 trials per data set/method with different pseudo-random seeds are reported. Results for each method are after cross-validation on a holdout set. The best results for each data set are **bolded** as well as any results that overlap within the confidence interval ($\pm$ 1 standard deviation).

Data Set	Accuracy on discrete attributes (ACC $\pm$ std)						
	AimNet	HCQ	XGB	MIDAS	GAIN	MF	MICE
Tic-Tac-Toc	$\mathbf{0.53 \pm 0.01}$	0.5 ± 0.01	0.52 ± 0.01	0.44 ± 0.02	0.35 ± 0.01	0.5 ± 0.01	0.46 ± 0.01
Hospital	$\mathbf{0.95 \pm 0.0}$	$\mathbf{0.95 \pm 0.0}$	0.91 ± 0.01	0.24 ± 0.01	0.14 ± 0.02	0.94 ± 0.01	0.7 ± 0.01
Mammogram	$\mathbf{0.73 \pm 0.02}$	$\mathbf{0.72 \pm 0.01}$	$\mathbf{0.72 \pm 0.02}$	0.71 ± 0.02	0.35 ± 0.01	0.66 ± 0.02	0.63 ± 0.02
Thoracic	$\mathbf{0.85 \pm 0.01}$	0.84 ± 0.01	0.84 ± 0.01	0.83 ± 0.02	0.52 ± 0.15	$\mathbf{0.85 \pm 0.01}$	0.75 ± 0.03
Contraceptive	$\mathbf{0.63 \pm 0.01}$	$\mathbf{0.63 \pm 0.01}$	0.62 ± 0.01	0.62 ± 0.01	0.43 ± 0.01	0.62 ± 0.01	0.55 ± 0.01
Solar Flare	$\mathbf{0.76 \pm 0.01}$	0.75 ± 0.01	0.75 ± 0.01	0.66 ± 0.01	0.46 ± 0.02	0.74 ± 0.01	0.65 ± 0.01
NYPD	0.87 ± 0.0	0.85 ± 0.0	$\mathbf{0.88 \pm 0.0}$	0.75 ± 0.0	0.15 ± 0.01	$\mathbf{0.88 \pm 0.0}$	0.58 ± 0.0
Credit	$\mathbf{0.73 \pm 0.01}$	0.7 ± 0.01	$\mathbf{0.73 \pm 0.01}$	0.6 ± 0.01	0.39 ± 0.01	$\mathbf{0.73 \pm 0.01}$	0.63 ± 0.01
Australian	$\mathbf{0.7 \pm 0.01}$	0.66 ± 0.01	0.68 ± 0.01	0.6 ± 0.01	0.46 ± 0.01	0.69 ± 0.01	0.59 ± 0.01
Balance	$\mathbf{0.73 \pm 0.03}$	$\mathbf{0.72 \pm 0.03}$	$\mathbf{0.71 \pm 0.03}$	0.64 ± 0.03	0.45 ± 0.05	0.63 ± 0.05	0.64 ± 0.04
Eye EEG	0.67 ± 0.01	0.62 ± 0.01	0.73 ± 0.01	0.55 ± 0.01	0.52 ± 0.03	$\mathbf{0.78 \pm 0.01}$	0.53 ± 0.01
Data Set	NRMS on continuous attributes (NRMS $\pm$ std)						
	AimNet	HCQ	XGB	MIDAS	GAIN	MF	MICE
Hospital	$\mathbf{0.81 \pm 0.04}$	1.1 ± 0.08	0.92 ± 0.07	440.63 ± 61.35	2.26 ± 1.18	0.89 ± 0.04	1.23 ± 0.09
Mammogram	$\mathbf{0.92 \pm 0.02}$	1.02 ± 0.04	0.98 ± 0.05	1.12 ± 0.08	1.05 ± 0.06	1.01 ± 0.03	1.25 ± 0.07
Thoracic	$\mathbf{0.94 \pm 0.01}$	1.09 ± 0.05	1.03 ± 0.11	5.64 ± 7.16	1.23 ± 0.22	0.99 ± 0.06	1.32 ± 0.12
Contraceptive	$\mathbf{0.9 \pm 0.02}$	1.12 ± 0.04	0.94 ± 0.02	1.11 ± 0.02	1.17 ± 0.05	0.99 ± 0.02	1.23 ± 0.06
Solar Flare	$\mathbf{0.93 \pm 0.09}$	$\mathbf{0.98 \pm 0.09}$	$\mathbf{1.0 \pm 0.1}$	10772.7 ± 4057.37	$\mathbf{1.0 \pm 0.09}$	1.04 ± 0.11	1.16 ± 0.07
NYPD	0.32 ± 0.01	0.44 ± 0.13	0.28 ± 0.0	0.69 ± 0.03	3.63 ± 0.17	$\mathbf{0.22 \pm 0.01}$	0.62 ± 0.01
Credit	$\mathbf{0.97 \pm 0.03}$	1.24 ± 0.03	1.26 ± 0.41	1.15 ± 0.07	1.2 ± 0.08	1.12 ± 0.18	1.34 ± 0.11
Australian	$\mathbf{0.96 \pm 0.02}$	1.23 ± 0.03	1.19 ± 0.2	1.14 ± 0.12	1.27 ± 0.16	1.07 ± 0.13	1.6 ± 0.7
Eye EEG	0.48 ± 0.0	0.71 ± 0.03	0.47 ± 0.0	0.91 ± 0.01	1.0 ± 0.28	$\mathbf{0.44 \pm 0.0}$	0.67 ± 0.01
Phase	$\mathbf{0.52 \pm 0.01}$	0.58 ± 0.0	0.53 ± 0.01	0.97 ± 0.01	1.14 ± 0.26	0.58 ± 0.01	0.73 ± 0.01
CASP	0.5 ± 0.01	1.5 ± 0.26	$\mathbf{0.49 \pm 0.01}$	0.88 ± 0.01	0.83 ± 0.09	$\mathbf{0.48 \pm 0.01}$	0.73 ± 0.03

Table 10: Imputation accuracy and NRMS error on the test set under **MCAR** error injections with $p = 0.6$ missingness across our AimNet model and the other baselines. The means of 10 trials per data set/method with different pseudo-random seeds are reported. Results for each method are after cross-validation on a holdout set. The best results for each data set are **bolded** as well as any results that overlap within the confidence interval ($\pm$ 1 standard deviation).

Data Set	Accuracy on discrete attributes (ACC $\pm$ std)						
	AimNet	HCQ	XGB	MIDAS	GAIN	MF	MICE
Tic-Tac-Toc	$\mathbf{0.48 \pm 0.01}$	$\mathbf{0.48 \pm 0.0}$	0.47 ± 0.01	0.43 ± 0.01	0.39 ± 0.01	0.44 ± 0.01	0.4 ± 0.01
Hospital	$\mathbf{0.86 \pm 0.01}$	$\mathbf{0.86 \pm 0.01}$	0.68 ± 0.01	0.24 ± 0.0	0.11 ± 0.01	0.79 ± 0.01	0.37 ± 0.01
Mammogram	$\mathbf{0.69 \pm 0.01}$	$\mathbf{0.69 \pm 0.01}$	$\mathbf{0.68 \pm 0.01}$	$\mathbf{0.68 \pm 0.01}$	0.34 ± 0.02	0.62 ± 0.02	0.58 ± 0.02
Thoracic	$\mathbf{0.85 \pm 0.01}$	0.84 ± 0.01	0.83 ± 0.0	0.84 ± 0.01	0.51 ± 0.13	0.84 ± 0.01	0.72 ± 0.04
Contraceptive	$\mathbf{0.62 \pm 0.01}$	$\mathbf{0.62 \pm 0.01}$	0.61 ± 0.01	0.61 ± 0.01	0.42 ± 0.02	0.6 ± 0.01	0.53 ± 0.01
Solar Flare	$\mathbf{0.72 \pm 0.01}$	$\mathbf{0.72 \pm 0.01}$	0.71 ± 0.01	0.66 ± 0.01	0.45 ± 0.03	0.7 ± 0.01	0.61 ± 0.01
NYPD	0.77 ± 0.0	0.76 ± 0.0	$\mathbf{0.79 \pm 0.0}$	0.67 ± 0.0	0.15 ± 0.0	0.78 ± 0.0	0.45 ± 0.0
Credit	$\mathbf{0.69 \pm 0.01}$	0.66 ± 0.01	0.68 ± 0.01	0.6 ± 0.01	0.38 ± 0.02	0.68 ± 0.01	0.57 ± 0.01
Australian	$\mathbf{0.67 \pm 0.01}$	0.65 ± 0.01	0.65 ± 0.01	0.59 ± 0.01	0.45 ± 0.02	0.65 ± 0.01	0.54 ± 0.02
Balance	$\mathbf{0.67 \pm 0.03}$	$\mathbf{0.64 \pm 0.04}$	0.63 ± 0.03	0.51 ± 0.06	0.46 ± 0.04	0.54 ± 0.03	0.55 ± 0.04
Eye EEG	0.63 ± 0.01	0.6 ± 0.01	0.66 ± 0.01	0.54 ± 0.01	0.52 ± 0.03	$\mathbf{0.67 \pm 0.01}$	0.52 ± 0.01
Data Set	NRMS on continuous attributes (NRMS $\pm$ std)						
	AimNet	HCQ	XGB	MIDAS	GAIN	MF	MICE
Hospital	$\mathbf{0.9 \pm 0.04}$	1.16 ± 0.12	1.01 ± 0.08	139.97 ± 24.15	3.79 ± 0.36	0.95 ± 0.05	1.27 ± 0.09
Mammogram	$\mathbf{0.94 \pm 0.02}$	1.05 ± 0.06	1.0 ± 0.05	1.13 ± 0.06	1.1 ± 0.11	1.01 ± 0.04	1.28 ± 0.08
Thoracic	$\mathbf{0.99 \pm 0.02}$	1.13 ± 0.05	1.17 ± 0.11	3.54 ± 5.28	1.18 ± 0.09	1.07 ± 0.04	1.43 ± 0.2
Contraceptive	$\mathbf{0.94 \pm 0.01}$	1.15 ± 0.04	1.01 ± 0.02	1.12 ± 0.02	1.31 ± 0.14	1.14 ± 0.04	1.29 ± 0.05
Solar Flare	$\mathbf{0.98 \pm 0.06}$	$\mathbf{1.01 \pm 0.06}$	1.05 ± 0.09	4454.64 ± 1610.59	1.07 ± 0.17	1.13 ± 0.14	1.24 ± 0.19
NYPD	0.56 ± 0.0	0.58 ± 0.04	0.45 ± 0.0	0.8 ± 0.01	3.51 ± 0.14	$\mathbf{0.42 \pm 0.01}$	0.98 ± 0.0
Credit	$\mathbf{0.99 \pm 0.01}$	1.33 ± 0.19	1.23 ± 0.25	1.14 ± 0.11	1.24 ± 0.11	1.13 ± 0.13	1.43 ± 0.26
Australian	$\mathbf{0.98 \pm 0.01}$	1.37 ± 0.26	1.29 ± 0.42	1.1 ± 0.04	1.25 ± 0.12	1.13 ± 0.19	1.38 ± 0.09
Eye EEG	0.59 ± 0.0	0.82 ± 0.04	$\mathbf{0.57 \pm 0.0}$	0.97 ± 0.01	1.61 ± 0.24	$\mathbf{0.57 \pm 0.0}$	0.79 ± 0.0
Phase	$\mathbf{0.64 \pm 0.0}$	0.71 ± 0.01	0.65 ± 0.0	1.0 ± 0.0	1.45 ± 0.51	0.71 ± 0.01	0.91 ± 0.01
CASP	$\mathbf{0.58 \pm 0.01}$	2.06 ± 0.51	0.59 ± 0.01	0.94 ± 0.01	1.2 ± 0.22	0.62 ± 0.0	0.88 ± 0.03