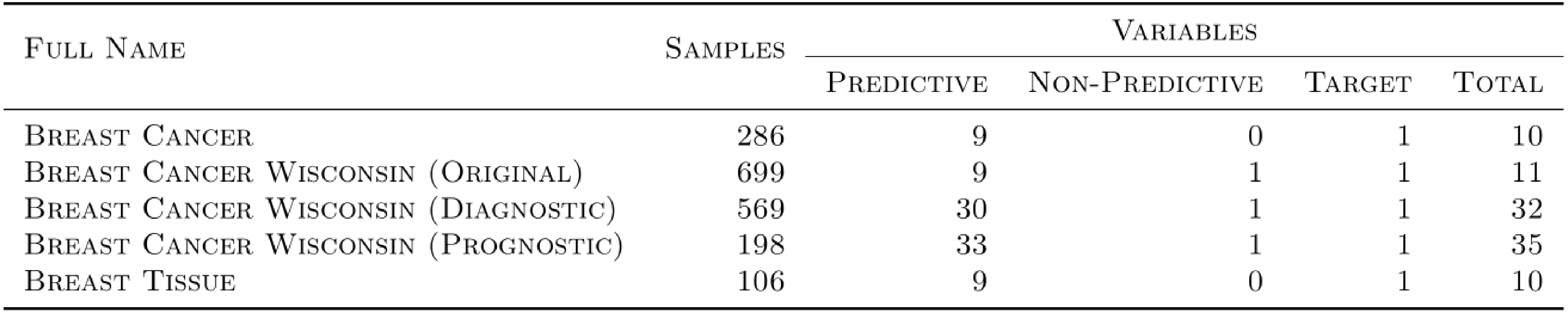
Working with Deep Generative Models and Tabular Data Imputation

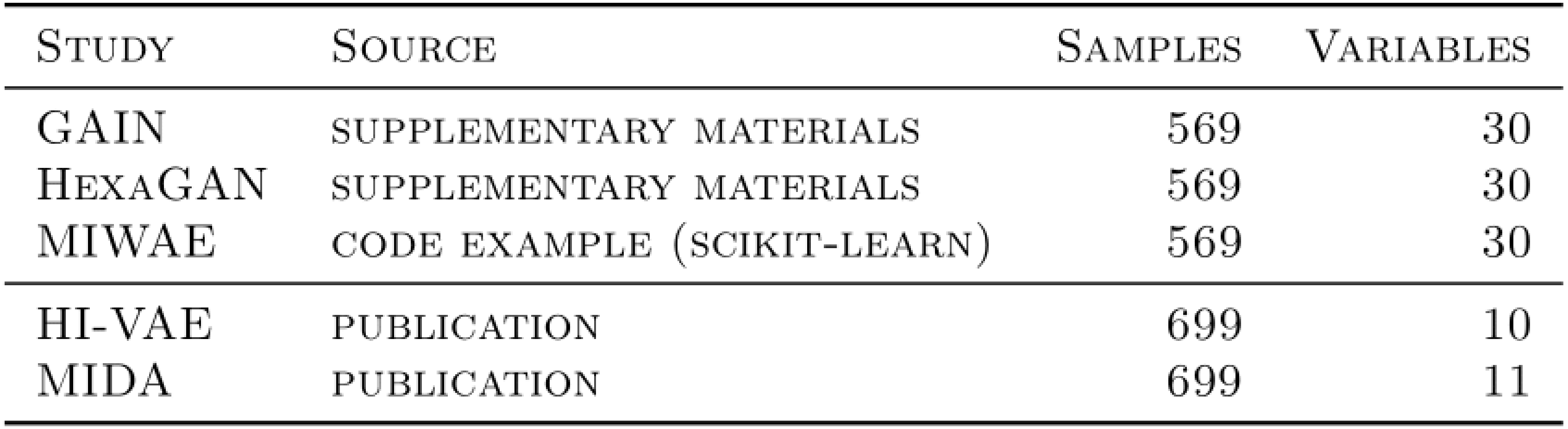
Ramiro Camino <ramiro.camino@uni.lu>

Issues	Examples	Possible Solutions
Unclear data acquisition	Ambiguous UCI Repository references, private data, unclear synthetic dataset generation	Publish code to download data, publish data in public repository, and/or establish common benchmarks
Underspecified pre-processing	Feature scaling, feature selection and categorical variable encoding	Publish pre-processing code
Mismatching metrics	MSE, RMSE, normalized RMSE, custom versions of MSE	Establish common benchmarks
Undefined hyperparameters	Not all values mentioned in the publication, values as arguments to define in the code	Publish hyperparameter search code or ALL the selected values explicitly
Undocumented “training tricks”	Simplifications taken in practice not documented in the publication, GAN training tricks or heuristics not cited or justified	Explain and justify implementation details in publication or document them in the code
Unfair model comparison	No hyperparameter search for other models, model capacities not compared	Publish hyperparameter search code for all models and compare all model capacities

Example: “Breast” UCI Dataset

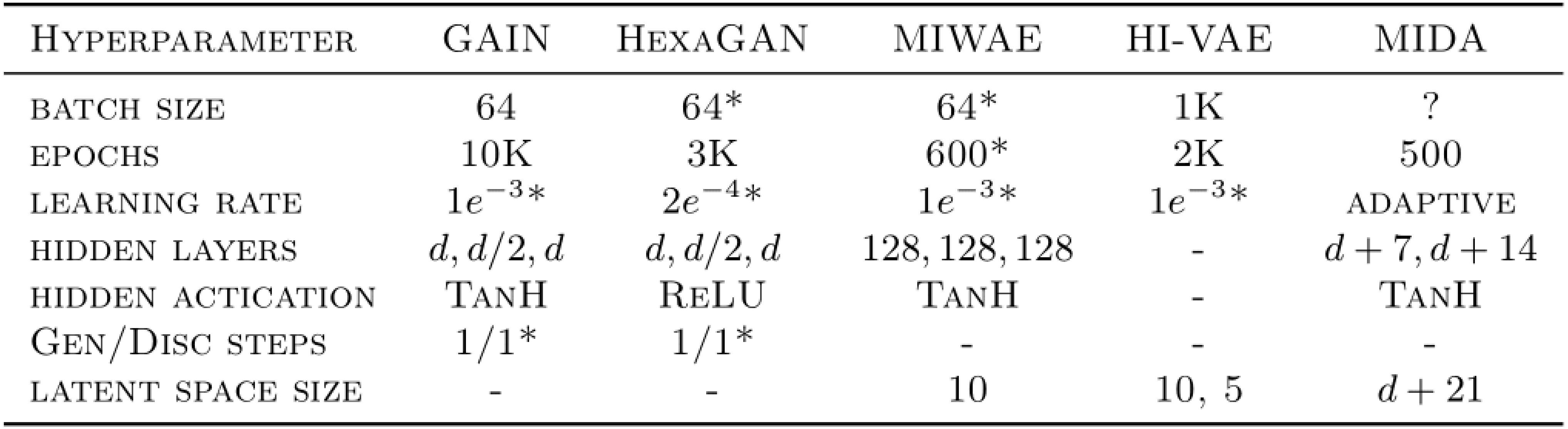


Full Name	Samples	Variables			
		Predictive	Non-Predictive	Target	Total
Breast Cancer	286	9	0	1	10
Breast Cancer Wisconsin (Original)	699	9	1	1	11
Breast Cancer Wisconsin (Diagnostic)	569	30	1	1	32
Breast Cancer Wisconsin (Prognostic)	198	33	1	1	35
Breast Tissue	106	9	0	1	10



Study	Source	Samples	Variables
GAIN	supplementary materials	569	30
HexaGAN	supplementary materials	569	30
MIWAE	code example (scikit-learn)	569	30
HI-VAE	publication	699	10
MIDA	publication	699	11

Common Hyperparameters



Hyperparameter	GAIN	HexaGAN	MIWAE	HI-VAE	MIDA
batch size	64	64*	64*	1K	?
epochs	10K	3K	600*	2K	500
learning rate	$1e^{-3}*$	$2e^{-4}*$	$1e^{-3}*$	$1e^{-3}*$	adaptive
hidden layers	$d, d/2, d$	$d, d/2, d$	$128, 128, 128$	-	$d+7, d+14$
hidden actication	TanH	ReLU	TanH	-	TanH
Gen/Disc steps	1/1*	1/1*	-	-	-
latent space size	-	-	10	10, 5	$d+21$

(?) Not found (-) Not applicable (*) Inferred from source code or examples