Image and graph convolution networks improve microbiome based machine learning accuracy - Supplementary Material

August 19, 2022

1 Links to microbiome movies

Movie of DiGiulio case-control study: https://drive.google.com/file/d/ 1qLZNOHqVolUe-OwySmTf0mNg8egMAFJR3/view?usp=sharing
Movie of Diabimmune case-control study:
https://drive.google.com/file/d/1MC4KwfiIe-1ab-05GE0yNATIXsVmg0xR/view?usp=sharing

Hyper parameter	Search space
L1 loss coefficient	[0,1]
Weight decay	[0,0.5]
Learning rate	[0.001,0.01,0.05]
Batch size	[32, 64, 218, 256]
Activation function	[RelU, ElU, tanh]
Dropout	[0,0.05,0.1,0.2,0.3,0.4,0.5]
Linear dimension 1 division factor	[1,11]
Linear dimension 2 division factor	[1,6]
GCN layer	[2,10]
Kernel size 1 of first Conv	[1,8]
Kernel size 2 of first Conv	[1,20]
Stride first Conv	[1,9]
Stride second Conv	[1,9]
Padding first Conv	[0,4]
Padding second Conv	[0,4]
Channels	[1,16]
Channel 2	[1,16]

Table 1: Hyperparameters search space table, where the weight decay controls the L2-regularization, the dropout is similar to all the layers, the number of neurons in the FCNs is determined according to the variables "Linear dimension 1 division factor" and "Linear dimension 2 division factor", GCN layer means the size of the GCN layer, the kernel sizes are determined by "Kernel size 1 of first Conv" (row-size) and "Kernel size 2 of first Conv" columns there were also different hyperparameters for the kernel's size of the second convolution layer.

Table 2: Default parameters of the neural networks (iMic, gMic, gMic+v and FCN)

	Default value used
Optimizer	Adam
Max-pooling	if number of model's parameters < 5000 -no pooling if number of model's parameters > 5000 - $pooling = 2$
Batch-normalization	No batch-normalization was used

Dataset	Model	L1 loss coefficient	Weight decay	Learning rate	Batch size	Activation function	Dropout	Gcn dimension	Linear dimension 1	Linear dimension 2
IBD	FCN sub pca	0.017	0.079	0.01	128	tanh	0.1	-	222	38
	iMic CNN1	0.048	0.010	0.001	32	elU	0.4	-	Total / 6	Total / 5
	iMic CNN2	0.142	0.039	0.001	256	tanh	0.2	-	Total / 1	Total / 8
	FCN	-	0.276	0.166	30	Relu	0.1	-	78	91
	gMic	-	0.021	0.008	10	Relu	0.1	5	124	121
	gMic+v	-	0.806	0.002	70	elu	0.15	9	67	23
CD	iMic CNN1	0.078	0.094	0.01	128	RelU	0.2	-	130	284
	iMic CNN2	0.154	0.107	0.001	64	tanh	0.1	-	Total / 1	Total / 1
	FCN	0.180	0.072	0.001	256	tanh	0.2	-	Total / 10	Total / 11
	gMic	-	0.006	0.366	5	Relu	0.06	-	64	141
	gMic	-	0.011	0.001	10	elu	0.05	5	67	128
	gMic+v	-	0.002	0.155	5	elu	0.2	7	58	42

Dataset	Model	L1 loss coefficient	Weight decay	Learning rate	Batch size	Activation function	Dropout	Gcn dimension	Linear dimension 1	Linear dimension 2
Nugent	iMic CNN1	0.109	0.04	0.001	128	tanh	0.2	-	35	145
	iMic CNN2	0.225	0.011	0.001	64	tanh	0.4	-	Total / 7	Total / 1
	FCN	0.397	0.008	0.001	64	tanh	0	-	5	6
	gMic	-	1.7e-4	0.058	10	Relu	0.35	-	69	27
	gMic	-	0.043	0.257	5	tanh	0.33	2	51	145
	gMic+v	-	0.023	0.069	5	tanh	0.27	7	165	90
Cirrhosis	iMic CNN1	0.230	0.002	0.001	32	elU	0.5	-	280	80
	iMic CNN2	0.121	0.375	0.001	64	tanh	0.2	-	Total / 7	Total / 1
	FCN	0.003	0.092	0.001	256	elU	0.3	-	Total / 9	Total / 1
	gMic	-	0.008	0.004	70	tanh	0.4	-	86	50
	gMic	-	0.031	5e-4	5	elu	0.2	6	53	200
	gMic+v	-	0.015	1e-4	100	tanh	0.3	9	76	74
Milk allergy	iMic CNN1	0.475	0.026	0.001	64	tanh	0.2	-	147	42
	iMic CNN2	0.065	0.102	0.001	256	elU	0.3	-	Total / 7	Total / 4
	FCN	0.319	0.025	0.001	256	elU	0	-	Total / 9	Total / 1
	gMic	-	0.005	0.233	10	Relu	0.12	-	169	128
	gMic	-	0.017	0.122	5	elu	0.1	5	26	136
	gMic+v	-	0.012	0.051	5	elu	0.3	6	164	200
Nuts allergy	iMic CNN1	0.277	0.092	0.001	128	elU	0.4	-	268	138
	iMic CNN2	0.294	0.035	0.001	128	tanh	0.5	-	Total / 8	Total / 5
	FCN	0.398	0.008	0.001	64	tanh	0	-	Total / 5	Total / 6
	gMic	-	0.086	0.498	100	elu	0.1	-	113	81
	gMic	-	0.004	7e-4	5	Relu	0.2	5	101	109
	gMic+v	-	0.014	0.001	5	tanh	0.3	8	170	31
Peanuts allergy	iMic CNN1	0.399	0.095	0.01	64	tanh	0.5	-	201	136
	iMic CNN2	0.376	0.0007	0.001	256	RelU	0.05	-	Total / 4	Total / 4
	FCN	0.187	0.025	0.001	256	elu	0.5	-	Total / 9	Total / 3
	gMic	-	0.003	0.015	50	tanh	0.13	-	29	18
	gMic	-	0.032	2e-4	50	tanh	0.18	10	111	92
	gMic+v	-	0.005	0.157	10	elu	0.15	6	42	129
CA	iMic CNN1	0.017	0.091	0.001	256	RelU	0.05	-	354	78
	iMic CNN2	0.069	0.084	0.001	64	tanh	0.1	-	Total / 5	Total / 4
	FCN	0.003	0.092	0.001	256	elu	0.3	-	Total / 9	Total / 1
	gMic	-	0.001	0.014	5	Relu	0.3	-	168	122
	gMic	-	0.010	0.369	10	tanh	0.35	3	143	141
	gMic+v	-	0.002	0.003	5	elu	0.22	2	83	131
MF	iMic CNN1	0.235	0.002	0.001	64	RelU	0.1	-	309	57
	iMic CNN2	0.377	0.012	0.001	128	RelU	0.5	-	Total / 9	Total / 2
	FCN	0.029	0.019	0.001	128	elu	0.3	-	Total / 7	Total / 11
	gMic	-	1.5e-4	0.031	30	elu	0.04	-	61	180
	gMic	-	0.052	0.004	50	tanh	0.19	6	148	118
	gMic+v	-	0.050	0.001	5	elu	0.36	10	191	146

Table 3: **Table of hyperparameters used**

Dataset	Model	Kernel size 1 of first Conv	Kernel size 2 of first Conv	Kernel size 1 of second Conv	Kernel size 2 of second Conv	Stride first Conv	Stride second Conv	Padding first Conv	Padding second Conv	Channels	Channel 2
IBD	iMic CNN1	5	17	-	-	3	-			14	-
	iMic CNN2	2	5	2	4	2	3	1	3	9	14
CD	iMic CNN1	5	10	-	-	3	-	-	-	14	-
	iMic CNN2	3	6	2	5	1	2	3	2	4	16
Nugent	iMic CNN1	6	6	-	-	8	-	-	-	15	-
	iMic CNN2	3	6	1	4	3	3	2	0	9	13
Cirrhosis	iMic CNN1	2	12	-	-	3	-	-	-	15	-
	iMic CNN2	2	8	1	7	2	3	2	2	5	15
Milk allergy	iMic CNN1	4	13	-	-	3	-	-	-	8	-
	iMic CNN2	4	5	1	9	1	1	2	0	9	8
Nuts allergy	iMic CNN1	3	11	-	-	5	-	-	-	24	3
	iMic CNN2	3	6	3	6	3	-	2	0	9	13
Peanuts allergy	iMic CNN1	7	15	-	-	5	-	-	-	12	-
	iMic CNN2	3	7	2	9	1	1	2	1	8	14

Dataset	Model	Kernel size 1 of first Conv	Kernel size 2 of first Conv	Kernel size 1 of second Conv	Kernel size 2 of second Conv	Stride first Conv	Stride second Conv	Padding first Conv	Padding second Conv	Channels	Channel 2
CA	iMic CNN1	5	17	-	-	5	-	-	-	9	-
	iMic CNN2	2	8	1	7	2	3	2	2	5	15
MF	iMic CNN1	6	16	-	-	2	-	-	-	13	-
	iMic CNN2	4	6	2	4	1	4	2	3	8	6

Table 4: **Special hyperparameters of iMic**

	RF hyperparameters		SVC hyperparameters	
	# of trees	Split criterion	Regularization coefficient	Kernel
IBD	100	gini	1.0	rbf
CD	100	gini	1.0	rbf
Ravel	100	gini	1.0	rbf
Cirrhosis	200	gini	1.0	rbf
Milk allergy	100	gini	0.5	rbf
Nut allergy	10	entropy	0.2	rbf
Peanut allergy	200	gini	0.5	poly
MF	100	gini	1.0	rbf
CA	10	gini	1.0	rbf

Table 5: All simple models' hyperparameters (RF, SVC)

Table 6: Datasets abbrevations

Data full name	Abbrevation
Inflammatory Bowel Disease	IBD
Crohn's disease	CD
Ulcerative colitis	UC
Male vs female	MF
Caucasian vs Afro-Americans	CA

	iMic's running times (seconds)
IBD	1.45
CD	1.49
Ravel	0.88
Cirrhosis	1.85
Milk allergy	23.15
Nut allergy	2.45
Peanut allergy	11.37
MF	4.25
CA	0.47

Table 7: iMic's running times (training) on Intel (R) Core (TM) i9-9900 CPU 3.10GHz