Appendix

Automatic Calcification Morphology and Distribution Classification for Breast Mammograms with Multi-task Graph Convolutional Neural Network

Fold 1			
Characteristics	**Group Frequency (percentage %)**		**P-value**
	Train n=152	**Test n=43**	
Age, mean (SD)	54.1 (11.5)	56.2 (13.5)	0.350
Mammographic purpose			0.280
Diagnosis	113	32	
Screening	39	11	
Symptoms			0.597
Mass	71 (46.7)	27 (62.8)	
Breast cancer initiate/subsequent screening	39 (25.7)	9 (20.9)	
Lump	21 (13.8)	3 (7.0)	
Positive cases from prior/outside mammography	14 (9.2)	2 (4.7)	
Bloody discharge	2 (1.3)		
Positive cases from prior/outside mammography and abnormal ultrasound finding	1 (0.7)	1 (2.3)	
From abnormal ultrasound	1 (0.7)	1 (2.3)	
Mass + bloody discharge	1 (0.7)		
Lump + bloody discharge	1 (0.7)		
Discharge	1 (0.7)		
Breast density composition			0.845
Fatty breast	4 (2.6)	1 (2.3)	
Scattered fibroglandular density	28 (18.4)	6 (14.0)	
Heterogeneously dense	106 (69.7)	33 (76.7)	
Extremely dense	14 (9.2)	3 (7.0)	
BI-RADS Category assessment			0.234
4A	6 (3.9)	2 (4.7)	
4B	30 (19.7)	3 (7.0)	
4C	30 (19.7)	8 (18.6)	
5	86 (56.6)	30 (69.8)	
Histopathology			0.892
IDC	88 (57.9)	29 (67.4)	
DCIS	11 (25.6)	52 (34.3)	
IDC+DCIS	6 (4.0)	1 (2.3)	
ILC	1 (0.7)	1 (2.3)	
Cellular changes conclusive for malignancy	1 (0.7)		
Adenocarcinoma	1 (0.7)	1 (2.3)	
ILC + IDC + DCIS	1 (0.7)		
Invasive mucinous carcinoma	1 (0.7)		
Distribution descriptors			0.685
Cluster	74 (48.7)	23 (53.5)	
Regional	45 (29.6)	11 (25.6)	
Segmental	23 (15.1)	4 (9.3)	
Diffuse	8 (5.3)	4 (9.3)	
Linear	2 (1.3)	1 (2.3)	
Morphology descriptors			0.842
Pleomorphic	70 (46.1)	18 (41.9)	
Heterogeneous	26 (17.1)	9 (20.9)	
Linear	11 (7.2)	5 (11.6)	
Pleomorphic + amorphous	11 (7.2)	3 (7.0)	
Pleomorphic + linear	10 (6.6)	2 (4.7)	
Amorphous + linear	6 (3.9)	2 (4.7)	
Amorphous + heterogeneous	5 (3.3)	3 (7.0)	
Amorphous	7 (4.6)		
Heterogeneous + linear	5 (3.3)	1 (2.3)	
Heterogeneous + amorphous + linear	1 (0.7)		

Table 1: Basic characteristics of training and testing sets for fold 1 in 5-fold cross validation

Fold 2			
Characteristics	**Group Frequency (percentage %)**		**P-value**
	Train n=154	**Test n=41**	
Age, mean (SD)	55.2 (11.8)	51.9 (12.3)	0.126
Mammographic purpose			0.898
Diagnosis	113 (73.3)	32 (78.1)	
Screening	41 (26.6)	9 (22.0)	
Symptoms			0.786
Mass	79 (51.3)	19 (46.3)	
Breast cancer initiate/subsequent screening	39 (25.3)	9 (20.9)	
Lump	16 (10.4)	8 (19.5)	
Positive cases from prior/outside mammography	12 (7.8)	4 (9.8)	
Bloody discharge	1 (0.6)	1 (2.4)	
Positive cases from prior/outside mammography and abnormal ultrasound finding	2 (1.3)		
From abnormal ultrasound	2 (1.3)		
Mass + bloody discharge	1 (0.6)		
Lump + bloody discharge	1 (0.6)		
Discharge	1 (0.6)		
Breast density composition			0.417
Fatty breast	5 (3.2)		
Scattered fibroglandular density	25 (16.2)	9 (22.0)	
Heterogeneously dense	109 (70.8)	30 (73.2)	
Extremely dense	15 (9.7)	2 (4.9)	
BI-RADS Category assessment			0.864
4A	6 (3.9)	2 (4.9)	
4B	25 (16.2)	8 (19.5)	
4C	29 (18.8)	9 (22.0)	
5	94 (61.0)	22 (53.7)	
Histopathology			0.535
IDC	92 (59.7)	25 (60.9)	
DCIS	49 (31.8)	14 (34.1)	
IDC+DCIS	5 (3.2)	2 (4.8)	
ILC	2 (1.3)		
Cellular changes conclusive for malignancy	1 (0.6)		
Adenocarcinoma	2 (1.3)		
ILC + IDC + DCIS	1 (0.6)		
Invasive mucinous carcinoma	1 (0.6)		
Distribution descriptors			0.794
Cluster	74 (48.1)	23 (56.1)	
Regional	47 (30.5)	9 (22.0)	
Segmental	21 (13.6)	6 (14.6)	
Diffuse	10 (6.5)	2 (4.9)	
Linear	2 (1.3)	1 (2.4)	
Morphology descriptors			0.578
Pleomorphic	68 (44.2)	20 (48.8)	
Heterogeneous	29 (18.8)	6 (14.6)	
Linear	12 (7.8)	4 (9.8)	
Pleomorphic + amorphous	10 (6.5)	4 (9.8)	
Pleomorphic + linear	9 (5.8)	3 (7.3)	
Amorphous + linear	7 (4.5)	1 (2.4)	
Amorphous + heterogeneous	8 (5.2)		
Amorphous	4 (2.6)	3 (7.3)	
Heterogeneous + linear	6 (3.9)		
Heterogeneous + amorphous + linear	1 (0.6)		

Table 2: Basic characteristics of training and testing sets for fold 2 in 5-fold cross validation

Fold 3			
Characteristics	**Group Frequency (percentage %)**		**P-value**
	Train n=152	**Test n=43**	
Age, mean (SD)	54.3 (12.0)	55.2 (11.9)	0.694
Mammographic purpose			0.441
Diagnosis	110 (71.9)	35 (83.3)	
Screening	43 (28.2)	7 (16.7)	
Symptoms			0.231
Mass	76 (49.7)	22 (52.4)	
Breast cancer initiate/subsequent screening	41 (26.8)	7 (16.7)	
Lump	18 (11.8)	6 (14.3)	
Positive cases from prior/outside mammography	11 (7.2)	5 (11.9)	
Bloody discharge	2 (1.3)		
Positive cases from prior/outside mammography and abnormal ultrasound finding	2 (1.3)		
From abnormal ultrasound	2 (1.3)		
Mass + bloody discharge	1 (0.7)		
Lump + bloody discharge		1 (2.4)	
Discharge		1 (2.4)	
Breast density composition			0.073
Fatty breast	2 (1.3)	3 (7.1)	
Scattered fibroglandular density	28 (18.3)	6 (14.3)	
Heterogeneously dense	112 (73.2)	27 (64.3)	
Extremely dense	11 (7.2)	6 (14.3)	
BI-RADS Category assessment			0.846
4A	7 (4.6)	1 (2.4)	
4B	26 (17.0)	7 (16.7)	
4C	31 (20.3)	7 (16.7)	
5	89 (58.2)	27 (64.3)	
Histopathology			0.921
IDC	89 (58.2)	28 (66.7)	
DCIS	52 (34.0)	11 (26.2)	
IDC+DCIS	5 (3.4)	2 (4.8)	
ILC	1 (0.7)	1 (2.4)	
Cellular changes conclusive for malignancy	1 (0.7)		
Adenocarcinoma	2 (1.3)		
ILC + IDC + DCIS	1 (0.7)		
Invasive mucinous carcinoma	1 (0.7)		
Distribution descriptors			0.870
Cluster	79 (51.6)	18 (42.9)	
Regional	42 (27.5)	14 (33.3)	
Segmental	21 (13.7)	6 (14.3)	
Diffuse	9 (5.9)	3 (7.1)	
Linear	2 (1.3)	1 (2.4)	
Morphology descriptors			0.429
Pleomorphic	64 (41.8)	24 (57.1)	
Heterogeneous	29 (19.0)	6 (14.3)	
Linear	13 (8.5)	3 (7.1)	
Pleomorphic + amorphous	13 (8.5)	1 (2.4)	
Pleomorphic + linear	11 (7.2)	1 (2.4)	
Amorphous + linear	6 (3.9)	2 (4.8)	
Amorphous + heterogeneous	5 (3.3)	3 (7.1)	
Amorphous	7 (4.6)		
Heterogeneous + linear	4 (2.6)	2 (4.8)	
Heterogeneous + amorphous + linear	1 (0.7)		

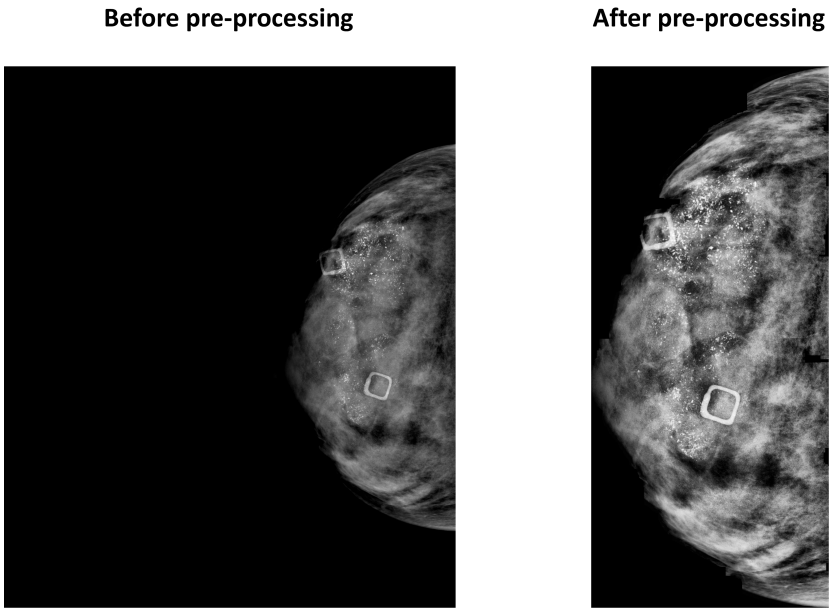
Table 3: Basic characteristics of training and testing sets for fold 3 in 5-fold cross validation

Fold 4			
Characteristics	**Group**		**P-value**
	Frequency (percentage %)		
	Train n=153	**Test n=42**	
Age, mean (SD)	54.3 (11.8)	55.2 (12.6)	0.707
Mammographic purpose			0.878
Diagnosis	115 (75.2)	30 (71.4)	
Screening	38 (24.9)	12 (28.6)	
Symptoms			0.585
Mass	75 (49.0)	23 (54.8)	
Breast cancer initiate/subsequent screening	36 (23.5)	12 (28.6)	
Lump	20 (13.1)	4 (9.5)	
Positive cases from prior/outside mammography	14 (9.2)	2 (4.8)	
Bloody discharge	2 (1.3)		
Positive cases from prior/outside mammography and abnormal ultrasound finding	2 (1.3)		
From abnormal ultrasound	2 (1.3)		
Mass + bloody discharge		1 (2.4)	
Lump + bloody discharge	1 (0.7)		
Discharge	1 (0.7)		
Breast density composition			0.323
Fatty breast	4 (2.6)	1 (2.4)	
Scattered fibroglandular density	28 (18.3)	6 (14.3)	
Heterogeneously dense	105 (68.6)	34 (81.0)	
Extremely dense	16 (10.5)	1 (2.4)	
BI-RADS Category assessment			0.573
4A	7 (4.6)	1 (2.4)	
4B	24 (15.7)	9 (21.4)	
4C	28 (18.3)	10 (23.8)	
5	94 (61.4)	22 (52.4)	
Histopathology			0.764
IDC	89 (58.2)	28 (66.7)	
DCIS	52 (34.0)	11 (26.2)	
IDC+DCIS	6 (4.0)	1 (2.4)	
ILC	2 (1.3)		
Cellular changes conclusive for malignancy	1 (0.7)		
Adenocarcinoma	1 (0.7)	1 (2.4)	
ILC + IDC + DCIS	1 (0.7)		
Invasive mucinous carcinoma		1 (2.4)	
Distribution descriptors			0.178
Cluster	78 (51.0)	19 (45.2)	
Regional	41 (26.8)	15 (35.7)	
Segmental	19 (12.4)	8 (19.0)	
Diffuse	12 (7.8)		
Linear	3 (2.0)		
Morphology descriptors			0.922
Pleomorphic	71 (46.4)	17 (40.5)	
Heterogeneous	25 (16.3)	10 (23.8)	
Linear	12 (7.8)	4 (9.5)	
Pleomorphic + amorphous	12 (7.8)	2 (4.8)	
Pleomorphic + linear	8 (5.2)	4 (9.5)	
Amorphous + linear	6 (3.9)	2 (4.8)	
Amorphous + heterogeneous	7 (4.6)	1 (2.4)	
Amorphous	6 (3.9)	1 (2.4)	
Heterogeneous + linear	5 (3.3)	1 (2.4)	
Heterogeneous + amorphous + linear	1 (0.7)		

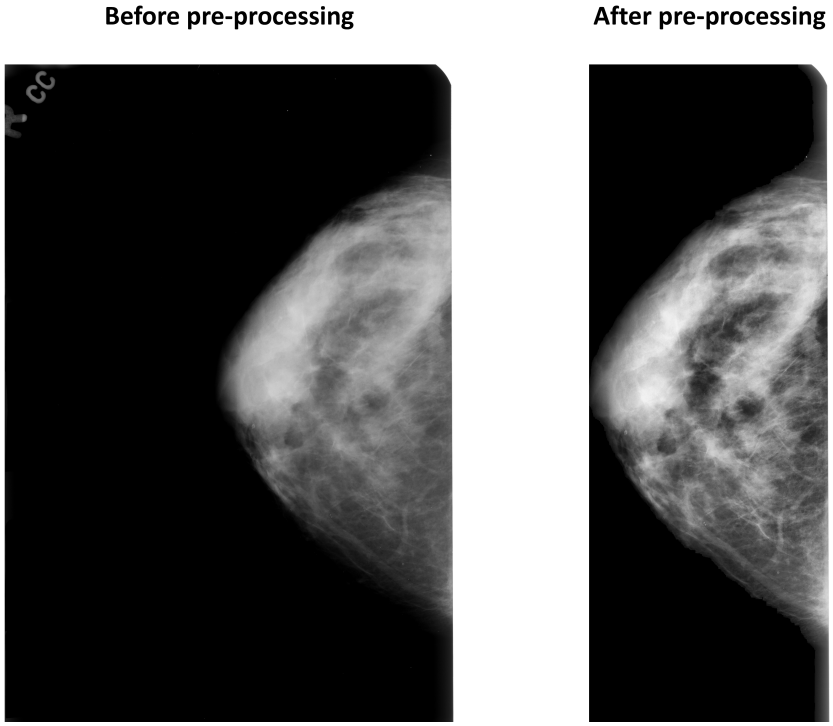
Table 4: Basic characteristics of training and testing sets for fold 4 in 5-fold cross validation

Fold 5			
Characteristics	**Group**		**P-value**
	Frequency (percentage %)		
	Train n=154	**Test n=41**	
Age, mean (SD)	55.1 (11.8)	52.4 (12.4)	0.215
Mammographic purpose			0.411
Diagnosis	115 (74.7)	30 (73.1)	
Screening	39 (25.3)	11 (26.8)	
Symptoms			0.765
Mass	81 (52.6)	17 (41.5)	
Breast cancer initiate/subsequent screening	37 (24.0)	11 (26.8)	
Lump	17 (11.0)	7 (17.1)	
Positive cases from prior/outside mammography	13 (8.4)	3 (7.3)	
Bloody discharge	1 (0.6)	1 (2.4)	
Positive cases from prior/outside mammography and abnormal ultrasound finding	1 (0.6)	1 (2.4)	
From abnormal ultrasound	1 (0.6)	1 (2.4)	
Mass + bloody discharge	1 (0.6)		
Lump + bloody discharge	1 (0.6)		
Discharge	1 (0.6)		
Breast density composition			0.523
Fatty breast	5 (3.2)		
Scattered fibroglandular density	26 (16.9)	8 (19.5)	
Heterogeneously dense	111 (72.1)	28 (68.3)	
Extremely dense	12 (7.8)	5 (12.2)	
BI-RADS Category assessment			0.157
4A	6 (3.9)	2 (4.9)	
4B	27 (17.5)	6 (14.6)	
4C	25 (16.2)	13 (31.7)	
5	96 (62.3)	20 (48.8)	
Histopathology			0.061
IDC	96 (62.3)	21 (51.2)	
DCIS	47 (30.5)	16 (39.0)	
IDC+DCIS	6 (3.8)	1 (2.4)	
ILC	2 (1.3)		
Cellular changes conclusive for malignancy		1 (2.4)	
Adenocarcinoma	2 (1.3)		
ILC + IDC + DCIS		1 (2.4)	
Invasive mucinous carcinoma	1 (0.6)		
Distribution descriptors			0.288
Cluster	74 (48.1)	23 (56.1)	
Regional	49 (31.8)	7 (17.1)	
Segmental	19 (12.3)	8 (19.5)	
Diffuse	9 (5.8)	3 (7.3)	
Linear	3 (1.9)		
Morphology descriptors			0.371
Pleomorphic	71 (46.1)	17 (41.5)	
Heterogeneous	30 (19.5)	5 (12.2)	
Linear	11 (7.1)	5 (12.2)	
Pleomorphic + amorphous	10 (6.5)	4 (9.8)	
Pleomorphic + linear	10 (6.5)	2 (4.9)	
Amorphous + linear	7 (4.5)	1 (2.4)	
Amorphous + heterogeneous	7 (4.5)	1 (2.4)	
Amorphous	4 (2.6)	3 (7.3)	
Heterogeneous + linear	4 (2.6)	2 (4.9)	
Heterogeneous + amorphous + linear		1 (2.4)	

Table 5: Basic characteristics of training and testing sets for fold 5 in 5-fold cross validation



(a) TMU dataset



(b) CBIS-DDSM dataset

Figure 1: Examples of mammograms from the TMU and CBIS-DDSM dataset before and after preprocessing

Type	Methods	ROC-AUC performance on each type of dsitribution descriptor class ±std [95% CI]				
		Diffuse	**Regional**	**Cluster**	**Linear**	**Segmental**
Baselines	ResNet	0.766±0.107 [0.688, 0.866]	0.672±0.094 [0.590, 0.746]	0.640±0.077 [0.568, 0.706]	0.638±0.109 [0.534, 0.720]	0.582±0.096 [0.510, 0.680]
	DenseNet	0.638±0.109 [0.560, 0.746]	0.648±0.091 [0.572, 0.730]	0.628±0.082 [0.548, 0.694]	0.476±0.375 [0.142, 0.788]	0.482±0.060 [0.428, 0.530]
	MobileNet	0.660±0.147 [0.524, 0.782]	0.634±0.074 [0.566, 0.694]	0.624±0.033 [0.596, 0.652]	0.864±0.099 [0.782, 0.940]	0.654±0.109 [0.562, 0.746]
	EfficientNet	0.670±0.148 [0.526, 0.796]	0.720±0.094 [0.624, 0.786]	0.624±0.110 [0.522, 0.710]	0.802±0.191 [0.628, 0.930]	0.556±0.079 [0.504, 0.638]
	GCN	0.574±0.140 [0.458, 0.698]	0.596±0.064 [0.542, 0.654]	0.540±0.061 [0.484, 0.592]	0.876±0.022 [0.858, 0.896]	0.528±0.083 [0.460, 0.596]
	GAT	0.596±0.136 [0.480, 0.712]	0.604±0.055 [0.560, 0.650]	0.518±0.069 [0.460, 0.580]	0.814±0.109 [0.720, 0.908]	0.588±0.117 [0.486, 0.682]
Proposed	**Multi-task, multi-graph, 8-layer GCN**	**0.944±0.059 [0.884, 0.984]**	**0.754±0.058 [0.706, 0.802]**	**0.798±0.055 [0.750, 0.842]**	**0.544±0.060 [0.496, 0.598]**	**0.680±0.078 [0.612, 0.750]**

(a) ROC-AUC performance on each type of dsitribution descriptor class

Type	Methods	ROC-AUC performance on each type of morphology descriptor class ±std [95% CI]			
		Coarse heterogeneous	**Amorphous**	**Fine pleomorphic**	**Fine linear (fine-linear branching)**
Baselines	ResNet	0.528±0.045 [0.496, 0.570]	0.578±0.025 [0.556, 0.598]	0.548±0.027 [0.524, 0.568]	0.554±0.074 [0.506, 0.630]
	DenseNet	0.546±0.073 [0.482, 0.610]	0.634±0.077 [0.578, 0.708]	0.592±0.040 [0.560, 0.628]	0.548±0.079 [0.486, 0.624]
	MobileNet	0.572±0.034 [0.540, 0.600]	0.564±0.036 [0.538, 0.600]	0.556±0.005 [0.552, 0.560]	0.546±0.057 [0.514, 0.604]
	EfficientNet	0.548±0.016 [0.534, 0.562]	0.582±0.036 [0.550, 0.614]	0.558±0.032 [0.530, 0.584]	0.540±0.042 [0.510, 0.580]
	GCN	0.626±0.036 [0.594, 0.656]	0.472±0.171 [0.336, 0.628]	0.566±0.039 [0.528, 0.598]	0.598±0.045 [0.556, 0.636]
	GAT	0.548±0.116 [0.444, 0.644]	0.528±0.139 [0.418, 0.664]	0.564±0.037 [0.534, 0.596]	0.568±0.052 [0.516, 0.606]
Proposed	**Multi-task, multi-graph, 8-layer GCN**	**0.758±0.053 [0.708, 0.802]**	**0.620±0.088 [0.548, 0.702]**	**0.614±0.033 [0.582, 0.636]**	**0.616±0.090 [0.542, 0.690]**

(b) ROC-AUC performance on each type of morphology descriptor class

Table 6: The ROC-AUC performance comparison on TMU dataset between baseline models and proposed model on distribution and morphology classification. ROC-AUC is evaluated on each type of distribution and morphology descriptor.

Type	Methods	ROC-AUC performance on each type of dsitribution descriptor class ±std [95% CI]				
		Diffuse	**Regional**	**Cluster**	**Linear**	**Segmental**
Baselines	ResNet	0.900±0.052 [0.850, 0.946]	0.700±0.148 [0.574, 0.830]	0.546±0.057 [0.496, 0.596]	0.460±0.075 [0.388, 0.516]	0.554±0.066 [0.502, 0.614]
	DenseNet	0.958±0.031 [0.928, 0.980]	0.596±0.183 [0.444, 0.754]	0.638±0.073 [0.572, 0.694]	0.614±0.052 [0.568, 0.658]	0.528±0.111 [0.432, 0.624]
	MobileNet	0.784±0.143 [0.644, 0.896]	0.792±0.055 [0.746, 0.848]	0.640±0.042 [0.602, 0.676]	0.490±0.046 [0.454, 0.532]	0.600±0.030 [0.576, 0.626]
	EfficientNet	0.884±0.152 [0.730, 0.968]	0.640±0.092 [0.560, 0.718]	0.580±0.079 [0.516, 0.656]	0.628±0.091 [0.542, 0.702]	0.620±0.036 [0.586, 0.644]
	GCN	0.698±0.155 [0.566, 0.830]	0.608±0.066 [0.556, 0.674]	0.552±0.057 [0.514, 0.608]	0.558±0.131 [0.438, 0.682]	0.604±0.081 [0.534, 0.674]
	GAT	0.778±0.061 [0.734, 0.836]	0.586±0.042 [0.546, 0.618]	0.586±0.059 [0.532, 0.630]	0.542±0.102 [0.452, 0.628]	0.624±0.087 [0.552, 0.698]
Proposed	**Multi-task, multi-graph, 8-layer GCN**	**0.984±0.017 [0.966, 0.996]**	**0.904±0.053 [0.852, 0.944]**	**0.870±0.018 [0.856, 0.884]**	**0.712±0.070 [0.650, 0.774]**	**0.896±0.036 [0.864, 0.928]**

(a) ROC-AUC performance on each type of dsitribution descriptor class

Type	Methods	ROC-AUC performance on each type of morphology descriptor class ±std [95% CI]				
		Amorphous	**Fine pleomorphic**	**Fine linear (fine-linear branching)**	**Punctate**	**Round and regular**
Baselines	ResNet	0.568±0.031 [0.544, 0.594]	0.538±0.050 [0.496, 0.580]	0.566±0.054 [0.516, 0.614]	0.572±0.029 [0.546, 0.596]	0.776±0.042 [0.744, 0.814]
	DenseNet	0.544±0.037 [0.516, 0.580]	0.522±0.065 [0.474, 0.582]	0.556±0.050 [0.512, 0.596]	0.600±0.087 [0.530, 0.684]	0.832±0.081 [0.756, 0.888]
	MobileNet	0.572±0.056 [0.528, 0.626]	0.550±0.039 [0.516, 0.584]	0.506±0.072 [0.450, 0.570]	0.554±0.049 [0.512, 0.596]	0.816±0.075 [0.750, 0.880]
	EfficientNet	0.564±0.037 [0.536, 0.600]	0.546±0.048 [0.504, 0.592]	0.576±0.080 [0.508, 0.648]	0.556±0.043 [0.518, 0.590]	0.822±0.101 [0.734, 0.910]
	GCN	0.540±0.068 [0.472, 0.596]	0.522±0.048 [0.480, 0.562]	0.548±0.052 [0.508, 0.600]	0.614±0.061 [0.562, 0.666]	0.680±0.223 [0.480, 0.870]
	GAT	0.584±0.036 [0.552, 0.614]	0.550±0.041 [0.508, 0.582]	0.524±0.038 [0.494, 0.560]	0.584±0.092 [0.504, 0.660]	0.518±0.092 [0.442, 0.590]
Proposed	**Multi-task, multi-graph, 8-layer GCN**	**0.586±0.125 [0.494, 0.702]**	**0.610±0.116 [0.500, 0.708]**	**0.664±0.032 [0.636, 0.692]**	**0.740±0.070 [0.678, 0.798]**	**0.906±0.122 [0.784, 0.988]**

(b) ROC-AUC performance on each type of morphology descriptor class

Table 7: The ROC-AUC performance comparison on CBIS-DDSM dataset between baseline models and proposed model on distribution and morphology classification. ROC-AUC is evaluated on each type of distribution and morphology descriptor.

Type	Methods	Distribution				
		ROC-AUC±std [95% CI]	**Precision±std [95% CI]**	**Recall±std [95% CI]**	**F1-score±std [95% CI]**	**Accuracy±std [95% CI] (%)**
Baselines	ResNet	*0.632±0.039 [0.599, 0.666]	*0.386±0.016 [0.374, 0.400]	*0.622±0.012 [0.611, 0.632]	*0.476±0.016 [0.464, 0.490]	*62.165±1.234 [61.145, 63.203]
	DenseNet	*0.668±0.049 [0.626, 0.710]	*0.412±0.043 [0.381, 0.455]	*0.622±0.012 [0.611, 0.633]	***0.488±0.024 [0.467, 0.509]**	*62.165±1.234 [61.145, 63.335]
	MobileNet	*0.660±0.035 [0.632, 0.694]	*0.479±0.110 [0.391, 0.576]	***0.623±0.012 [0.613, 0.633]**	*0.480±0.016 [0.467, 0.493]	***62.331±1.211 [61.311, 63.351]**
	EfficientNet	***0.672±0.030 [0.648, 0.700]**	***0.503±0.119 [0.398, 0.606]**	*0.622±0.012 [0.611, 0.632]	*0.487±0.023 [0.467, 0.507]	*62.165±1.234 [61.145, 63.241]
	GCN	*0.605±0.038 [0.570, 0.635]	*0.386±0.016 [0.374, 0.401]	*0.622±0.012 [0.612, 0.633]	*0.476±0.016 [0.464, 0.491]	*62.165±1.234 [61.152, 63.335]
	GAT	*0.622±0.039 [0.584, 0.651]	*0.387±0.017 [0.374, 0.404]	*0.622±0.012 [0.611, 0.633]	*0.477±0.016 [0.464, 0.492]	*62.165±1.234 [61.145, 63.335]
Proposed	**Multi-task, multi-graph, 8-layer GCN**	**0.873±0.019 [0.859, 0.891]**	**0.742±0.045 [0.705, 0.782]**	**0.758±0.011 [0.749, 0.768]**	**0.727±0.023 [0.708, 0.746]**	**75.845±1.121 [74.870, 76.820]**

(a) Classification performance on distribution descriptors

Type	Methods	Morphology				
		ROC-AUC±std [95% CI]	**Precision±std [95% CI]**	**Recall±std [95% CI]**	**F1-score±std [95% CI]**	**Accuracy±std [95% CI] (%)**
Baselines	ResNet	*0.603±0.020 [0.588, 0.621]	0.624±0.088 [0.545, 0.694]	0.753±0.121 [0.631, 0.833]	0.653±0.155 [0.496, 0.757]	75.293±12.094 [63.053, 83.289]
	DenseNet	*0.611±0.032 [0.578, 0.636]	0.604±0.121 [0.486, 0.694]	0.753±0.121 [0.631, 0.833]	0.654±0.154 [0.499, 0.757]	75.298±12.083 [63.069, 83.289]
	MobileNet	*0.599±0.027 [0.578, 0.624]	0.613±0.178 [0.440, 0.732]	0.753±0.121 [0.630, 0.833]	0.653±0.156 [0.496, 0.758]	75.299±12.110 [63.037, 83.304]
	EfficientNet	***0.613±0.014 [0.602, 0.627]**	0.594±0.146 [0.452, 0.697]	0.753±0.121 [0.631, 0.833]	0.654±0.154 [0.499, 0.757]	75.302±12.085 [63.077, 83.293]
	GCN	*0.580±0.041 [0.547, 0.616]	0.591±0.169 [0.427, 0.706]	0.753±0.121 [0.631, 0.833]	0.653±0.156 [0.497, 0.758]	75.309±12.111 [63.079, 83.310]
	GAT	*0.552±0.022 [0.531, 0.570]	**0.646±0.077 [0.580, 0.712]**	**0.762±0.103 [0.659, 0.833]**	**0.675±0.114 [0.564, 0.758]**	**76.240±10.330 [65.914, 83.323]**
Proposed	**Multi-task, multi-graph, 8-layer GCN**	**0.700±0.044 [0.661, 0.735]**	**0.681±0.076 [0.615, 0.746]**	**0.792±0.050 [0.750, 0.833]**	**0.712±0.055 [0.666, 0.760]**	**79.189±4.989 [74.961, 83.289]**

(b) Classification performance on morphology descriptors

† Statistical tests were performed between baseline and proposed models on all evaluation metrics of both tasks. Statistically significant comparisons are highlighted with *. $p < 0.0001$ for these comparisons, which are below the adjusted significance level (0.008).

Table 8: The performance comparison on CBIS-DDSM dataset between baseline models and proposed model on distribution and morphology classification. Best results in baseline models are represented in red bold. Results from the proposed method are highlighted using black bold.

Type	Methods	Distribution				
		ROC-AUC ±std [95% CI]	**Precision ±std [95% CI]**	**Recall ±std [95% CI]**	**F1-score ±std [95% CI]**	**Accuracy±std [95% CI] (%)**
Ablation study	Task-specific (Dis.)	0.851±0.019 [0.835, 0.867]	0.676±0.071 [0.615, 0.740]	0.720±0.040 [0.687, 0.754]	0.656±0.047 [0.617, 0.697]	71.963±3.988 [68.722, 75.375]
	Single-graph (Rad.)	**0.872±0.025 [0.852, 0.893]**	**0.722±0.085 [0.649, 0.795]**	0.745±0.034 [0.716, 0.775]	0.697±0.047 [0.657, 0.737]	**74.497±3.413 [71.631, 77.503]**
	Single-graph (KNN)	*0.795±0.026 [0.770, 0.815]	*0.455±0.090 [0.379, 0.532]	*0.655±0.046 [0.616, 0.695]	*0.534±0.077 [0.469, 0.599]	*65.549±4.664 [61.545, 69.553]
	2-layer GCN	0.866±0.024 [0.848, 0.888]	0.687±0.037 [0.655, 0.719]	0.743±0.015 [0.730, 0.756]	0.699±0.030 [0.674, 0.729]	74.322±1.504 [72.993, 75.557]
	4-layer GCN	0.869±0.019 [0.856, 0.888]	0.686±0.062 [0.630, 0.745]	**0.745±0.021 [0.728, 0.762]**	0.697±0.042 [0.665, 0.738]	74.496±2.068 [72.795, 76.197]
	16-layer GCN	0.867±0.021 [0.849, 0.887]	0.711±0.041 [0.678, 0.748]	0.742±0.014 [0.729, 0.754]	**0.699±0.028 [0.676, 0.726]**	74.162±1.421 [72.945, 75.389]
Proposed	**Multi-task, multi-graph, 8-layer GCN**	**0.873±0.019 [0.859, 0.891]**	**0.742±0.045 [0.705, 0.782]**	**0.758±0.011 [0.749, 0.768]**	**0.727±0.023 [0.708, 0.746]**	**75.845±1.121 [74.870, 76.820]**

(a) Classification performance on distribution descriptors

Type	Methods	Morphology				
		ROC-AUC±std [95% CI]	**Precision ±std [95% CI]**	**Recall ±std [95% CI]**	**F1-score ±std [95% CI]**	**Accuracy±std [95% CI] (%)**
Ablation study	Task-specific (Mor.)	0.638±0.057 [0.586, 0.685]	0.580±0.165 [0.423, 0.699]	0.752±0.121 [0.635, 0.835]	0.652±0.156 [0.502, 0.760]	75.216±12.104 [63.494, 83.488]
	Single-graph (Rad.)	0.655±0.050 [0.614, 0.698]	**0.671±0.075 [0.610, 0.740]**	0.752±0.121 [0.635, 0.835]	**0.659±0.143 [0.523, 0.761]**	75.225±12.109 [63.503, 83.497]
	Single-graph (KNN)	*0.591±0.028 [0.569, 0.616]	0.580±0.165 [0.423, 0.699]	0.752±0.121 [0.635, 0.835]	0.652±0.156 [0.502, 0.760]	75.216±12.104 [63.494, 83.488]
	2-layer GCN	0.675±0.041 [0.640, 0.709]	0.664±0.083 [0.586, 0.727]	0.752±0.121 [0.636, 0.835]	0.656±0.147 [0.515, 0.761]	**75.237±12.063 [63.558, 83.488]**
	4-layer GCN	**0.679±0.033 [0.646, 0.703]**	0.644±0.108 [0.542, 0.723]	0.752±0.121 [0.630, 0.835]	0.657±0.145 [0.512, 0.760]	75.216±12.104 [63.037, 83.488]
	16-layer GCN	0.675±0.054 [0.632, 0.724]	0.617±0.179 [0.449, 0.749]	**0.752±0.121 [0.635, 0.835]**	0.653±0.156 [0.502, 0.762]	75.219±12.105 [63.497, 83.491]
Proposed	**Multi-task, multi-graph, 8-layer GCN**	**0.700±0.044 [0.661, 0.735]**	**0.681±0.076 [0.615, 0.746]**	**0.792±0.050 [0.750, 0.833]**	**0.712±0.055 [0.666, 0.760]**	**79.189±4.989 [74.961, 83.289]**

(b) Classification performance on morphology descriptors

† **Statistical tests were performed between baseline and proposed models on all evaluation metrics of both tasks. Statistically significant comparisons are highlighted with** *. $p < 0.0001$ **for these comparisons, which are below the adjusted significance level (0.008).**

Table 9: Ablation Study: The performance comparison on CBIS-DDSM dataset between the ablation models and the proposed model on distribution and morphology classification. Best results in ablation studies are represented in red bold. Results from the proposed method are highlighted using black bold. Mor.=Morphology, Dis.=Distribution, Rad.=Radius

Type	Methods	Distribution				
		ROC-AUC±std [95% CI]	**Precision±std [95% CI]**	**Recall±std [95% CI]**	**F1-score±std [95% CI]**	**Accuracy±std [95% CI] (%)**
Baselines	ResNet	*0.654±0.014 [0.643, 0.669]	*0.469±0.062 [0.422, 0.531]	*0.562±0.010 [0.553, 0.570]	*0.432±0.005 [0.427, 0.437]	*56.172±1.050 [55.228, 57.021]
	DenseNet	*0.653±0.022 [0.633, 0.670]	*0.490±0.018 [0.475, 0.507]	*0.567±0.017 [0.554, 0.582]	*0.452±0.023 [0.435, 0.476]	*56.696±1.686 [55.353, 58.151]
	MobileNet	***0.690±0.018 [0.671, 0.701]**	*0.466±0.072 [0.407, 0.541]	***0.574±0.020 [0.558, 0.594]**	*0.464±0.034 [0.436, 0.495]	***57.437±2.024 [55.770, 59.369]**
	EfficientNet	*0.680±0.013 [0.670, 0.692]	***0.562±0.045 [0.515, 0.589]**	*0.570±0.015 [0.558, 0.583]	***0.460±0.026 [0.435, 0.481]**	*57.012±1.530 [55.822, 58.316]
	GCN	*0.673±0.015 [0.660, 0.685]	*0.423±0.058 [0.362, 0.465]	*0.569±0.012 [0.557, 0.578]	*0.429±0.016 [0.414, 0.442]	*56.913±1.235 [55.698, 57.824]
	GAT	*0.659±0.011 [0.649, 0.668]	*0.371±0.036 [0.337, 0.401]	*0.562±0.010 [0.552, 0.569]	*0.427±0.026 [0.404, 0.449]	*56.172±1.050 [55.199, 56.946]
Proposed	**Multi-task, multi-graph, 8-layer GCN**	**0.862±0.028 [0.833, 0.885]**	**0.691±0.021 [0.674, 0.708]**	**0.690±0.021 [0.672, 0.708]**	**0.661±0.012 [0.650, 0.673]**	**68.963±2.071 [67.155, 70.771]**

(a) Classification performance on distribution descriptors

Type	Methods	Morphology				
		ROC-AUC±std [95% CI]	**Precision±std [95% CI]**	**Recall±std [95% CI]**	**F1-score±std [95% CI]**	**Accuracy±std [95% CI] (%)**
Baselines	ResNet	*0.623±0.036 [0.599, 0.659]	0.577±0.081 [0.506, 0.647]	0.680±0.043 [0.641, 0.716]	0.551±0.055 [0.502, 0.598]	67.979±4.275 [64.100, 71.561]
	DenseNet	*0.582±0.016 [0.570, 0.597]	**0.603±0.105 [0.518, 0.695]**	**0.682±0.042 [0.647, 0.722]**	0.555±0.054 [0.509, 0.606]	**68.252±4.198 [64.702, 72.195]**
	MobileNet	*0.672±0.022 [0.658, 0.694]	0.526±0.082 [0.475, 0.608]	0.680±0.043 [0.641, 0.714]	0.551±0.055 [0.501, 0.596]	67.979±4.275 [64.058, 71.429]
	EfficientNet	*0.664±0.034 [0.632, 0.696]	0.580±0.086 [0.506, 0.654]	0.682±0.042 [0.644, 0.718]	0.555±0.054 [0.505, 0.601]	68.252±4.198 [64.440, 71.802]
	GCN	***0.677±0.018 [0.658, 0.689]**	0.466±0.055 [0.419, 0.512]	0.680±0.043 [0.644, 0.716]	0.551±0.054 [0.506, 0.597]	67.979±4.275 [64.396, 71.561]
	GAT	*0.671±0.029 [0.645, 0.694]	0.532±0.108 [0.445, 0.624]	0.681±0.044 [0.641, 0.717]	**0.560±0.061 [0.506, 0.612]**	68.081±4.345 [64.058, 71.689]
Proposed	**Multi-task, multi-graph, 8-layer GCN**	**0.725±0.028 [0.702, 0.748]**	**0.619±0.078 [0.553, 0.682]**	**0.689±0.043 [0.654, 0.724]**	**0.596±0.056 [0.548, 0.643]**	**68.951±4.281 [65.369, 72.395]**

(b) Classification performance on morphology descriptors

† Statistical tests were performed between baseline and proposed models on all evaluation metrics of both tasks. Statistically significant comparisons are highlighted with *. $p < 0.0001$ for these comparisons, which are below the adjusted significance level (0.008).

Table 10: The performance comparison on mixed dataset (TMU + CBIS-DDSM) between baseline models and proposed model on distribution and morphology classification. Best results in baseline models are represented in red bold. Results from the proposed method are highlighted using black bold.