Supplementary Materials

Structural Quantization of Multimodal Neural Networks for Skin Cancer Classification to Optimize Performance in Edge Mobile Devices

Table 1. Distribution of dermatological images by categories and groups of pigmented skin lesions

№	Group	Category	Abbreviation	Quantity
1.	Benign	vascular lesions	*vl*	253
2.		nevus	*nv*	27878
3.		solar lentigo	*sl*	270
4.		dermatofibroma	*df*	246
5.		seborrheic keratosis	*sk*	1464
6.		benign keratosis	*bk*	1099
7.	Malignant	actinic keratosis	*ak*	869
8.		basal cell carcinoma	*bcc*	3393
9.		squamous cell carcinoma	*scc*	656
10.		melanoma	*ml*	5597

Table 2. Distribution of dermatological metadata by age and standardized groups

№	Age	Quantity	Group	№	Age	Quantity	Group
1.	0	55	young ($\leq 44\ years$)	10.	45	3559	middle ($45 \leq 59\ years$)
2.	5	114		11.	50	3933	
3.	10	1668		12.	55	3228	
4.	15	7966		13.	60	2676	elderly ($60 \leq 74$ years)
5.	20	439		14.	65	2583	
6.	25	875		15.	70	2529	
7.	30	1588		16.	75	2124	senile (≥ 75 years)
8.	35	2233		17.	80	1629	
9.	40	2989		18.	85	1537	

Table 3. Distribution of dermatological metadata by gender and localization of pigmented lesions on the patient's body

№	Group	Meaning	Quantity
1.	Localization on the body	posterior torso	17326
2.		anterior torso	7440
3.		lower extremity	7182
4.		head/neck	5375
5.		upper extremity	3844
6.		palms/soles	411
7.		lateral torso	83
8.		oral/genital	64
9.	Gender	male	23294
10.		female	18431

Table 4. Distribution of images by diagnostic categories in training, validation and test sets

Set	*vl, (%)*	*nv, (%)*	*sl, (%)*	*df, (%)*	*sk, (%)*	*bk, (%)*	*ak, (%)*	*bcc, (%)*	*scc, (%)*	*ml, (%)*	**Total, *(%)***
Training	151 (0.36)	16730 (40.07)	162 (0.40)	146 (0.37)	878 (2.21)	659 (1.58)	521 (1.31)	2035 (5.10)	394 (0.89)	3359 (8.03)	**25035** (60.00)
Validation	51 (0.12)	5574 (13.36)	54 (0.13)	50 (0.12)	293 (0.70)	220 (0.50)	174 (0.42)	679 (1.63)	131 (0.31)	1119 (2.68)	**8345** (20.00)
Test	51 (0.12)	5574 (13.36)	54 (0.13)	50 (0.12)	293 (0.70)	220 (0.50)	174 (0.42)	679 (1.63)	131 (0.31)	1119 (2.68)	**8345** (20.00)

Table 5. Performance and compression metrics of quantized multimodal models based on CNN AlexNet tested on a personal computer

Quantization type AlexNet	Inference time per image, ms	Speed-up, x	Model size, MB	Compression ratio, x
FP32	1.5485±0.0098	-	27.71	-
PTQ_CNN	1.1337±0.0090	1.3467±0.0079	20.67	1.34
PTQ_MLP	1.5122±0.0116	1.0057±0.0045	27.70	1.01
PTQ_classif	1.4785±0.0096	0.9862±0.0254	14.02	1.98
PTQ_full	**1.1002±0.0090**	**1.3796±0.0293**	**6.98**	**3.97**
QAT_CNN	1.1583±0.0063	1.3666±0.0131	20.67	1.34
QAT_MLP	1.5323±0.0101	0.9868±0.0180	27.70	1.01
QAT_classif	1.5260±0.0125	0.9866±0.0180	14.02	1.98
QAT_full	**1.1365±0.0060**	**1.3685±0.0085**	**6.98**	**3.97**

Inference time per image - average processing time for a single image; Speed-up - ratio relative to the FP32 baseline; Model size - total storage of the trained model; Compression ratio - model size reduction after quantization relative to the FP32 baseline.

Table 6. Performance and compression metrics of quantized multimodal models based on CNN Shufflenet_v2 tested on a personal computer

Quantization type Shufflenet_v2	Inference time per image, ms	Speed-up, x	Model size, MB	Compression ratio, x
FP32	3.4672±0.0062	-	24.91	-
PTQ_CNN	6.3617±0.0049	0.5449±0.0010	10.08	2.47
PTQ_MLP	**3.0344±0.0040**	**1.1426±0.0020**	24.91	1.00
PTQ_classif	3.1250±0.0041	1.1095±0.0020	21.73	1.15
PTQ_full	6.4625±0.0040	0.5365±0.0010	**6.88**	**3.62**
QAT_CNN	6.3633±0.0041	0.5448±0.0010	10.08	2.47
QAT_MLP	**3.0406±0.0040**	**1.1403±0.0020**	24.91	1.00
QAT_classif	3.0688±0.0040	1.1297±0.0020	21.73	1.15
QAT_full	6.2966±0.0039	0.5506±0.0010	**6.88**	**3.62**

Table 7. Performance and compression metrics of quantized multimodal models based on CNN VGG_16 tested on a personal computer

Quantization type VGG_16	Inference time per image, ms	Speed-up, x	Model size, MB	Compression ratio, x
FP32	32.4718±0.6258	-	105.42	-
PTQ_CNN	16.4410±0.5527	1.9745±0.0784	63.42	1.66
PTQ_MLP	32.0857±0.4694	1.0120±0.0254	105.42	1.00
PTQ_classif	31.7819±0.8323	1.0217±0.0317	68.49	1.54
PTQ_full	**16.4222±0.6397**	**1.9768±0.0886**	**26.48**	**3.98**
QAT_CNN	16.3665±0.5863	1.9839±0.0831	63.42	1.66
QAT_MLP	31.7670±0.6449	1.0222±0.0278	105.42	1.00
QAT_classif	31.6527±0.7290	1.0259±0.0303	68.49	1.54
QAT_full	**16.2604±0.4695**	**1.9973±0.0710**	**26.48**	**3.98**

Table 8. Characteristics of peripheral devices used to evaluate the performance of the proposed MNNs based on various CNN FP32 baseline models and quantized models by PTQ and QAT methods

Parameter	Samsung Galaxy A56 (SM-A566B)	ASUS Zenfone 2 (Z00UD)
SoC / Platform	Snapdragon 7 Gen 1	Intel Atom Z3580
CPU Architecture	1×Cortex-A720 @ 2.9 GHz + 3×Cortex-A720 @ 2.6 GHz +	4× Moorefield @ 2.3 GHz

	4×Cortex-A520 @ 1.95 GHz	
CPU Cores	8 (1+3+4)	8
Process Technology	4 nm	22 nm
RAM	8 GB	4 GB
Display Resolution	1080 × 2113 px	1080 × 1920 px
GPU	Xclipse 540	PowerVR G6430
Android Version	16 (API 36)	6.0.1 (API 23)

Table 9. Results of testing the performance of the proposed MNN models based on CNN AlexNet on the Samsung A56 mobile edge device

Quantization type AlexNet	Inference time per image, ms	FPS	CPU utilization, %	Power consumption, W	Energy per image, J	Estimated temperature, °C
FP_32	50.56±10.26	49.83±2.10	0.20±0.01	0.3269±0.05	0.0165±0.05	25.83±0.12
PTQ_CNN	21.10±3.78	44.63±1.12	0.12±0.04	0.3163±0.01	0.0098±0.06	25.85±0.06
PTQ_MLP	28.58±1.32	39.96±0.75	0.15±0.13	0.3170±0.02	0.0081±0.01	25.82±0.01
PTQ_classif	23.05±4.63	43.48±1.03	0.11±0.06	0.3154±0.05	0.0082±0.02	25.83±0.03
PTQ_full	17.61±1.20	46.58±0.36	0.10±0.08	0.3149±0.01	0.0077±0.09	25.81±0.02
QAT_CNN	27.98±1.66	45.74±1.83	0.13±0.02	0.3143±0.02	0.0087±0.01	25.82±0.08
QAT_MLP	26.56±3.18	37.61±1.53	0.14±0.08	0.3161±0.04	0.0083±0.02	25.83±0.04
QAT_classif	23.22±2.87	43.05±1.32	0.16±0.02	0.3147±0.01	0.0072±0.04	25.84±0.06
QAT_full	21.88±2.83	50.41±1.86	0.11±0.04	0.3118±0.02	0.0071±0.03	25.81±0.04

Table 10. Results of testing the performance of the proposed MNN models based on CNN AlexNet on the Asus ZenFone 2 mobile edge device

Quantization type AlexNet	Inference time per image, ms	FPS	CPU utilization, %	Power consumption, W	Energy per image, J	Estimated temperature, °C
FP_32	232.18±8.04	4.31±1.09	0.72±0.04	0.4179±0.04	0.0969±0.03	26.02±0.03
PTQ_CNN	154.09±5.84	6.57±1.53	0.49±0.07	0.3781±0.03	0.0582±0.02	25.96±0.02
PTQ_MLP	223.86±6.56	4.53±2.01	0.70±0.03	0.4138±0.03	0.0925±0.01	26.01±0.02
PTQ_classif	194.19±8.99	5.91±2.14	0.69±0.05	0.3734±0.06	0.0538±0.05	25.98±0.03
PTQ_full	78.28±12.11	12.72±3.89	0.36±0.04	0.3409±0.06	0.0268±0.03	25.84±0.04
QAT_CNN	158.38±5.98	6.39±1.83	0.48±0.04	0.3801±0.03	0.0601±0.02	25.97±0.02
QAT_MLP	226.24±6.58	4.49±1.57	0.71±0.02	0.4149±0.03	0.0937±0.03	26.01±0.02
QAT_classif	145.00±12.42	6.91±2.25	0.68±0.05	0.3746±0.05	0.0542±0.06	25.93±0.04
QAT_full	79.06±8.14	12.85±3.47	0.39±0.06	0.3402±0.03	0.0265±0.05	25.88±0.03

Table 11. Results of testing the performance of the proposed MNN models based on CNN Shufflenet v2 on the Samsung A56 mobile edge device

Quantization type Shufflenet_v2	Inference time per image, ms	FPS	CPU utilization, %	Power consumption, W	Energy per image, J	Estimated temperature, °C
FP_32	51.98±4.61	19.21±1.46	0.25±0.07	0.3263±0.02	0.0169±0.06	25.83±0.02
PTQ_CNN	36.39±5.29	27.54±1.38	0.19±0.03	0.3182±0.03	0.0115±0.01	25.82±0.01
PTQ_MLP	50.38±4.51	21.72±1.86	0.21±0.05	0.3276±0.01	0.0174±0.02	25.82±0.03
PTQ_classif	45.86±3.76	17.91±2.63	0.22±0.02	0.3284±0.02	0.0183±0.01	25.81±0.02
PTQ_full	35.01±4.88	28.61±1.37	0.18±0.04	0.3175±0.01	0.0111±0.02	25.81±0.01
QAT_CNN	38.29±6.07	26.11±1.34	0.18±0.04	0.3199±0.03	0.0122±0.02	25.82±0.02
QAT_MLP	50.57±5.11	19.8±1.41	0.23±0.03	0.3256±0.01	0.0164±0.01	25.83±0.01
QAT_classif	47.20±4.76	17.52±1.69	0.21±0.01	0.3297±0.03	0.0188±0.02	25.82±0.03
QAT_full	34.91±4.56	28.69±1.27	0.17±0.02	0.3171±0.01	0.0110±0.03	25.81±0.02

Table 12. Results of testing the performance of the proposed MNN models based on CNN Shufflenet_v2 on the Asus ZenFone 2 mobile edge device

Quantization type Shufflenet_v2	Inference time per image, ms	FPS	CPU utilization, %	Power consumption, W	Energy per image, J	Estimated temperature, °C
FP_32	496.33±26.45	2.19±0.03	1.97±0.12	0.5521±0.03	0.2742±0.02	26.41±0.02
PTQ_CNN	187.39±10.43	5.37±0.08	0.74±0.07	0.3956±0.05	0.0740±0.01	26.03±0.01
PTQ_MLP	492.11±23.71	2.06±0.07	1.99±0.08	0.5503±0.02	0.2708±0.03	26.42±0.01
PTQ_classif	473.87±22.12	2.13±0.04	1.95±0.03	0.5418±0.08	0.2564±0.02	26.45±0.03
PTQ_full	160.69±14.23	6.28±0.05	0.67±0.03	0.3819±0.03	0.0614±0.01	26.02±0.02
QAT_CNN	187.12±11.64	5.38±0.03	0.75±0.04	0.3961±0.06	0.0739±0.01	26.07±0.03
QAT_MLP	540.35±55.35	1.98±0.11	2.13±0.06	0.5749±0.03	0.3118±0.04	26.49±0.05
QAT_classif	516.67±53.72	1.94±0.04	2.18±0.05	0.5627±0.03	0.2920±0.05	26.42±0.03
QAT_full	165.12±15.66	6.19±0.02	0.72±0.03	0.3841±0.02	0.0634±0.01	26.01±0.02

Table 13. Results of testing the performance of the proposed MNN models based on CNN VGG_16 on the Samsung A56 mobile edge device

Quantization type VGG_16	Inference time per image, ms	FPS	CPU utilization, %	Power consumption, W	Energy per image, J	Estimated temperature, °C
FP_32	533.11±6.55	4.32±1.12	1.92±0.11	0.5189±0.04	0.2974±0.03	26.11±0.01
PTQ_CNN	317.78±9.18	4.42±1.84	1.62±0.31	0.3932±0.03	0.1739±0.03	26.12±0.02
PTQ_MLP	493.08±3.65	4.12±1.82	1.76±0.18	0.3983±0.02	0.2867±0.01	26.20±0.03
PTQ_classif	383.92±5.90	3.04±1.55	1.27±0.33	0.5124±0.01	0.2823±0.02	26.12±0.01
PTQ_full	286.87±7.28	5.06±1.12	1.19±0.09	0.3967±0.02	0.1719±0.05	26.18±0.02
QAT_CNN	311.39±10.45	4.42±1.98	1.62±0.18	0.5221±0.06	0.2092±0.07	26.31±0.03
QAT_MLP	454.58±3.60	4.11±1.27	1.72±0.13	0.5981±0.02	0.2775±0.01	26.28±0.02
QAT_classif	386.04±7.69	4.41±1.83	1.68±0.09	0.5941±0.04	0.2733±0.03	26.29±0.01
QAT_full	289.98±10.10	4.61±1.59	1.61±0.11	0.4989±0.05	0.1942±0.06	26.24±0.02

Table 14. Results of testing the performance of the proposed MNN models based on CNN VGG_16 on the Asus ZenFone 2 mobile edge device

Quantization type VGG_16	Inference time per image, ms	FPS	CPU utilization, %	Power consumption, W	Energy per image, J	Estimated temperature, °C
FP_32	2170.37±285.68	0.51±0.09	7.96±1.13	1.4014±0.14	3.0816±0.77	28.56±0.42
PTQ_CNN	1997.64±272.57	0.56±0.13	7.29±0.11	1.3139±0.13	2.6609±0.65	28.37±0.31
PTQ_MLP	2135.03±177.57	0.53±0.07	7.88±0.62	1.3834±0.09	2.9687±0.44	28.58±0.26
PTQ_classif	2051.50±135.01	0.58±0.11	8.24±0.51	1.3426±0.06	2.7623±0.31	28.46±0.22
PTQ_full	**1703.96±143.24**	**0.66±0.14**	**6.83±0.48**	**1.1657±0.07**	**1.9963±0.27**	**27.94±0.21**
QAT_CNN	1974.85±125.41	0.55±0.12	7.43±0.52	1.3028±0.06	2.5783±0.26	28.36±0.29
QAT_MLP	2848.48±344.18	0.45±0.15	10.58±1.32	1.7458±0.17	5.0310±1.03	29.47±0.43
QAT_classif	2287.35±252.83	0.47±0.10	9.14±1.83	1.4629±0.23	3.4477±1.14	28.79±0.67
QAT_full	**1854.69±230.60**	**0.58±0.11**	**7.41±0.76**	**1.2428±0.13**	**2.4466±0.86**	**28.13±0.36**

Table 15. Modality contribution analysis in PTQ-quantized MNN models based on the CNN AlexNet

Quantization type AlexNet	Experimental Condition	Accuracy, %	F1-Score	MCC	Confidence	Entropy	Modality Contribution, %	Agreement with Full, %	Correct Predictions
PTQ_CNN	**Full Multimodal**	85.00 ± 0.4776	0.8310 ± 0.0025	0.7066 ± 0.0098	0.8373 ± 0.0030	0.4292 ± 0.0090	100.00	-	7,122
	Only Images	78.69 ± 0.5200	0.7891 ± 0.0028	0.6295 ± 0.0042	0.7671 ± 0.0033	0.6124 ± 0.0100	95.27	88.82	6,567
	Only Metadata	13.41 ± 0.5000	0.0317 ± 0.0118	0.0000 ± 0.0079	0.7603 ± 0.0050	0.8885 ± 0.0180	4.73	18.03	1,119
	Fusion Gain	**+6.31 ± 0.7050**	**+0.0419 ± 0.0030**	**+0.0771 ± 0.0107**	**+0.0702 ± 0.0043**	**-0.1829 ± 0.0150**	**-**	**-**	**+555**
PTQ_MLP	**Full Multimodal**	85.22 ± 0.4124	0.8328 ± 0.0025	0.7095 ± 0.0084	0.8358 ± 0.0030	0.4324 ± 0.0090	100.00	-	7,138
	Only Images	79.01 ± 0.4500	0.7916 ± 0.0028	0.6342 ± 0.0042	0.7664 ± 0.0033	0.6132 ± 0.0100	95.29	89.08	6,593
	Only Metadata	13.41 ± 0.5000	0.0317 ± 0.0118	0.0000 ± 0.0079	0.7450 ± 0.0050	0.9151 ± 0.0180	4.71	17.97	1,119
	Fusion Gain	**+6.21 ± 0.6100**	**+0.0412 ± 0.0030**	**+0.0753 ± 0.0095**	**+0.0694 ± 0.0043**	**-0.1819 ± 0.0150**	**-**	**-**	**+545**
PTQ_classif	**Full Multimodal**	85.15 ± 0.4775	0.8358 ± 0.0025	0.7092 ± 0.0098	0.8353 ± 0.0030	0.4328 ± 0.0090	100.00	-	7,146
	Only Images	79.53 ± 0.5200	0.7976 ± 0.0028	0.6388 ± 0.0042	0.7655 ± 0.0033	0.6154 ± 0.0100	95.33	88.75	6,637
	Only Metadata	13.41 ± 0.5000	0.0317 ± 0.0118	0.0000 ± 0.0079	0.7421 ± 0.0050	0.9217 ± 0.0180	4.67	16.12	1,119
	Fusion Gain	**+5.62 ± 0.7050**	**+0.0382 ± 0.0030**	**+0.0704 ± 0.0107**	**+0.0707 ± 0.0043**	**-0.1863 ± 0.0150**	**-**	**-**	**+509**
PTQ_full	**Full Multimodal**	85.08 ± 0.5042	0.8349 ± 0.0025	0.7067 ± 0.0105	0.8374 ± 0.0030	0.4286 ± 0.0090	100.00	-	7,137
	Only Images	79.16 ± 0.5500	0.7944 ± 0.0028	0.6329 ± 0.0042	0.7678 ± 0.0033	0.6105 ± 0.0100	95.30	88.53	6,606
	Only Metadata	13.41 ± 0.5000	0.0317 ± 0.0118	0.0000 ± 0.0079	0.7115 ± 0.0050	1.0413 ± 0.0180	4.70	16.43	1,119
	Fusion Gain	**+5.92 ± 0.7450**	**+0.0405 ± 0.0030**	**+0.0738 ± 0.0112**	**+0.0696 ± 0.0043**	**-0.1819 ± 0.0150**	**-**	**-**	**+531**

Table 16. Modality contribution analysis in QAT-quantized MNN models based on the CNN AlexNet

Quantization type AlexNet	Experimental Condition	Accuracy, %	F1-Score	MCC	Confidence	Entropy	Modality Contribution, %	Agreement with Full, %	Correct Predictions
QAT_CNN	**Full Multimodal**	85.25 ± 0.4131	0.8354 ± 0.0025	0.7125 ± 0.0085	0.8406 ± 0.0030	0.4223 ± 0.0090	100.00	-	7,145
	Only Images	78.74 ± 0.4500	0.7911 ± 0.0028	0.6339 ± 0.0042	0.7661 ± 0.0033	0.6181 ± 0.0100	95.28	88.46	6,571
	Only Metadata	13.41 ± 0.5000	0.0317 ± 0.0118	0.0000 ± 0.0079	0.7007 ± 0.0050	1.0692 ± 0.0180	4.73	17.40	1,119
	Fusion Gain	**+6.51 ± 0.6100**	**+0.0443 ± 0.0030**	**+0.0786 ± 0.0094**	**+0.0739 ± 0.0043**	**-0.1958 ± 0.0150**	**-**	**-**	**+574**
QAT_MLP	**Full Multimodal**	85.51 ± 0.4748	0.8385 ± 0.0025	0.7172 ± 0.0097	0.8434 ± 0.0030	0.4110 ± 0.0090	100.00	-	7,156
	Only Images	79.01 ± 0.5200	0.7951 ± 0.0028	0.6350 ± 0.0042	0.7737 ± 0.0033	0.5905 ± 0.0100	95.29	88.44	6,593
	Only Metadata	13.41 ± 0.5000	0.0317 ± 0.0118	0.0000 ± 0.0079	0.6677 ± 0.0050	1.0861 ± 0.0180	4.71	15.94	1,119
	Fusion Gain	**+6.50 ± 0.7000**	**+0.0434 ± 0.0030**	**+0.0822 ± 0.0106**	**+0.0697 ± 0.0043**	**-0.1816 ± 0.0150**	**-**	**-**	**+563**
QAT_classif	**Full Multimodal**	85.36 ± 0.5278	0.8381 ± 0.0025	0.7132 ± 0.0109	0.8401 ± 0.0030	0.4226 ± 0.0090	100.00	-	7,160
	Only Images	79.19 ± 0.5800	0.7945 ± 0.0028	0.6347 ± 0.0042	0.7689 ± 0.0033	0.6065 ± 0.0100	95.30	88.54	6,608
	Only Metadata	13.41 ± 0.5000	0.0317 ± 0.0118	0.0000 ± 0.0079	0.6799 ± 0.0050	1.0275 ± 0.0180	4.70	15.77	1,119
	Fusion Gain	**+6.17 ± 0.7800**	**+0.0436 ± 0.0030**	**+0.0785 ± 0.0118**	**+0.0712 ± 0.0043**	**-0.1839 ± 0.0150**	**-**	**-**	**+552**
QAT_full	**Full Multimodal**	85.23 ± 0.2895	0.8315 ± 0.0025	0.7105 ± 0.0059	0.8409 ± 0.0030	0.4238 ± 0.0090	100.00	-	7,123
	Only Images	78.98 ± 0.3200	0.7916 ± 0.0028	0.6330 ± 0.0042	0.7642 ± 0.0033	0.6303 ± 0.0100	93.45	88.52	6,591
	Only Metadata	14.84 ± 0.5500	0.0638 ± 0.0118	0.0233 ± 0.0079	0.6551 ± 0.0050	1.0307 ± 0.0180	6.55	18.12	1,238
	Fusion Gain	**+6.25 ± 0.4300**	**+0.0399 ± 0.0030**	**+0.0775 ± 0.0070**	**+0.0767 ± 0.0043**	**-0.1930 ± 0.0150**	**-**	**-**	**+532**

Table 17. Modality contribution analysis in PTQ-quantized MNN models based on the CNN ShuffleNet_V2

Quantization type ShuffleNet_V2	Experimental Condition	Accuracy, %	F1-Score	MCC	Confidence	Entropy	Modality Contribution, %	Agreement with Full, %	Correct Predictions
PTQ_CNN	**Full Multimodal**	76.14 ± 0.1611	0.6937 ± 0.0025	0.5080 ± 0.0047	0.8084 ± 0.0030	0.6046 ± 0.0090	100.00	-	6,350
	Only Images	74.79 ± 0.1780	0.6845 ± 0.0028	0.4922 ± 0.0042	0.7354 ± 0.0033	0.8155 ± 0.0100	53.28	96.99	6,241
	Only Metadata	66.81 ± 0.2514	0.5351 ± 0.0118	0.0000 ± 0.0079	0.9468 ± 0.0050	0.2760 ± 0.0180	46.72	74.73	5,575
	Fusion Gain	**+1.35 ± 0.2400**	**+0.0092 ± 0.0030**	**+0.0158 ± 0.0063**	**-0.1384 ± 0.0043**	**+0.3286 ± 0.0150**	**-**	**-**	**+109**
PTQ_MLP	**Full Multimodal**	82.05 ± 0.2359	0.7835 ± 0.0025	0.6368 ± 0.0041	0.8095 ± 0.0030	0.5014 ± 0.0090	100.00	-	6,839
	Only Images	78.99 ± 0.2600	0.7615 ± 0.0028	0.5850 ± 0.0042	0.7659 ± 0.0033	0.6047 ± 0.0100	54.84	90.95	6,592
	Only Metadata	66.81 ± 0.2514	0.5351 ± 0.0118	0.0000 ± 0.0079	0.7186 ± 0.0050	1.0213 ± 0.0180	45.16	71.7	5,575
	Fusion Gain	**+3.06 ± 0.3500**	**+0.0220 ± 0.0030**	**+0.0518 ± 0.0058**	**+0.0436 ± 0.0043**	**-0.1236 ± 0.0150**	**-**	**-**	**+247**
PTQ_classif	**Full Multimodal**	81.38 ± 0.2479	0.7698 ± 0.0025	0.6185 ± 0.0043	0.8147 ± 0.0030	0.4863 ± 0.0090	100.00	-	6,764
	Only Images	78.31 ± 0.2720	0.7428 ± 0.0028	0.5666 ± 0.0042	0.7735 ± 0.0033	0.5817 ± 0.0100	54.60	91.43	6,535
	Only Metadata	66.81 ± 0.2514	0.5351 ± 0.0118	0.0000 ± 0.0079	0.5979 ± 0.0050	1.3121 ± 0.0180	45.40	73.43	5,575
	Fusion Gain	**+3.07 ± 0.3670**	**+0.0270 ± 0.0030**	**+0.0519 ± 0.0060**	**+0.2168 ± 0.0043**	**-0.8258 ± 0.0150**	**-**	**-**	**+229**
PTQ_full	**Full Multimodal**	76.49 ± 0.1906	0.7043 ± 0.0025	0.5013 ± 0.0034	0.7986 ± 0.0030	0.6355 ± 0.0090	100.00	-	6,363
	Only Images	73.98 ± 0.2100	0.6894 ± 0.0028	0.4570 ± 0.0042	0.7188 ± 0.0033	0.8586 ± 0.0100	52.97	94.63	6,174
	Only Metadata	66.81 ± 0.2514	0.5351 ± 0.0118	0.0000 ± 0.0079	0.8301 ± 0.0050	0.7004 ± 0.0180	47.03	76.91	5,575
	Fusion Gain	**+2.51 ± 0.2850**	**+0.0149 ± 0.0030**	**+0.0443 ± 0.0053**	**-0.0315 ± 0.0043**	**-0.2231 ± 0.0150**	**-**	**-**	**+189**

Table 18. Modality contribution analysis in QAT-quantized MNN models based on the CNN ShuffleNet_V2

Quantization type ShuffleNet_V2	Experimental Condition	Accuracy, %	F1-Score	MCC	Confidence	Entropy	Modality Contribution, %	Agreement with Full, %	Correct Predictions
QAT_CNN	**Full Multimodal**	77.79 ± 0.1508	0.7383 ± 0.0025	0.5497 ± 0.0036	0.7385 ± 0.0030	0.8025 ± 0.0090	100.00	-	6,486
	Only Images	71.07 ± 0.1650	0.6772 ± 0.0028	0.4462 ± 0.0042	0.5944 ± 0.0033	1.2182 ± 0.0100	55.15	84.95	5,931
	Only Metadata	59.68 ± 0.2800	0.5782 ± 0.0118	0.2581 ± 0.0079	0.5733 ± 0.0050	1.3360 ± 0.0180	44.86	70.57	4,980
	Fusion Gain	**+6.72 ± 0.2230**	**+0.0611 ± 0.0030**	**+0.1035 ± 0.0054**	**+0.1441 ± 0.0043**	**-0.4157 ± 0.0150**	**-**	**-**	**+555**
QAT_MLP	**Full Multimodal**	83.02 ± 0.2209	0.8004 ± 0.0025	0.6630 ± 0.0037	0.8307 ± 0.0030	0.4347 ± 0.0090	100.00	-	6,920
	Only Images	79.11 ± 0.2420	0.7706 ± 0.0028	0.6003 ± 0.0042	0.7941 ± 0.0033	0.5181 ± 0.0100	54.89	89.43	6,602
	Only Metadata	66.81 ± 0.2514	0.5351 ± 0.0118	0.0000 ± 0.0079	0.8978 ± 0.0050	0.4135 ± 0.0180	45.11	69.95	5,575
	Fusion Gain	**+3.91 ± 0.3270**	**+0.0298 ± 0.0030**	**+0.0627 ± 0.0056**	**-0.0671 ± 0.0043**	**+0.0212 ± 0.0150**	**-**	**-**	**+318**
QAT_classif	**Full Multimodal**	82.51 ± 0.1545	0.7927 ± 0.0025	0.6512 ± 0.0024	0.8317 ± 0.0030	0.4294 ± 0.0090	100.00	-	6,861
	Only Images	79.11 ± 0.1700	0.7664 ± 0.0028	0.5956 ± 0.0042	0.7978 ± 0.0033	0.5059 ± 0.0100	55.58	90.87	6,602
	Only Metadata	65.24 ± 0.2800	0.6084 ± 0.0118	0.2940 ± 0.0079	0.5634 ± 0.0050	1.3112 ± 0.0180	44.42	64.88	5,444
	Fusion Gain	**+3.40 ± 0.2290**	**+0.0263 ± 0.0030**	**+0.0556 ± 0.0046**	**+0.0339 ± 0.0043**	**-0.0765 ± 0.0150**	**-**	**-**	**+259**
QAT_full	**Full Multimodal**	72.60 ± 0.4664	0.7053 ± 0.0025	0.4759 ± 0.0085	0.6574 ± 0.0030	1.0341 ± 0.0090	100.00	-	6,014
	Only Images	55.00 ± 0.5100	0.5756 ± 0.0028	0.3090 ± 0.0042	0.3628 ± 0.0033	1.8171 ± 0.0100	44.20	65.58	4,590
	Only Metadata	66.81 ± 0.2514	0.5351 ± 0.0118	0.0000 ± 0.0079	0.7215 ± 0.0050	0.9091 ± 0.0180	55.80	66.02	5,575
	Fusion Gain	**+5.79 ± 0.5320**	**+0.1297 ± 0.0030**	**+0.1669 ± 0.0094**	**-0.0641 ± 0.0043**	**+0.1250 ± 0.0150**	**-**	**-**	**+439**

Table 19. Modality contribution analysis in PTQ-quantized MNN models based on the CNN VGG_16

Quantization type VGG_16	Experimental Condition	Accuracy, %	F1-Score	MCC	Confidence	Entropy	Modality Contribution, %	Agreement with Full, %	Correct Predictions
PTQ_CNN	**Full Multimodal**	87.15 ± 0.1631	0.8566 ± 0.0019	0.7478 ± 0.0045	0.8546 ± 0.0027	0.3817 ± 0.0078	100.00	-	7,277
	Only Images	82.42 ± 0.1750	0.8312 ± 0.0021	0.6941 ± 0.0042	0.8077 ± 0.0031	0.4931 ± 0.0091	56.04	90.02	6,878
	Only Metadata	66.81 ± 0.2514	0.5351 ± 0.0118	0.0000 ± 0.0079	0.6538 ± 0.0049	1.2630 ± 0.0225	43.96	68.98	5,575
	Fusion Gain	**+4.73 ± 0.1920**	**+0.0254 ± 0.0023**	**+0.0537 ± 0.0051**	**+0.0469 ± 0.0041**	**-0.1079 ± 0.0162**	**-**	**-**	**+399**
PTQ_MLP	**Full Multimodal**	87.38 ± 0.1256	0.8592 ± 0.0016	0.7518 ± 0.0034	0.8580 ± 0.0026	0.3712 ± 0.0071	100.00	-	7,302
	Only Images	82.79 ± 0.1470	0.8342 ± 0.0018	0.6990 ± 0.0037	0.8132 ± 0.0030	0.4767 ± 0.0089	56.17	90.32	6,909
	Only Metadata	66.81 ± 0.2514	0.5351 ± 0.0118	0.0000 ± 0.0079	0.8285 ± 0.0048	0.7430 ± 0.0178	43.83	69.26	5,575
	Fusion Gain	**+4.59 ± 0.1810**	**+0.0250 ± 0.0020**	**+0.0528 ± 0.0048**	**+0.0448 ± 0.0039**	**-0.1055 ± 0.0142**	**-**	**-**	**+393**
PTQ_classif	**Full Multimodal**	87.05 ± 0.0824	0.8554 ± 0.0016	0.7435 ± 0.0022	0.8587 ± 0.0026	0.3680 ± 0.0071	100.00	-	7,269
	Only Images	83.21 ± 0.0920	0.8358 ± 0.0018	0.6989 ± 0.0037	0.8139 ± 0.0030	0.4743 ± 0.0089	56.31	90.69	6,944
	Only Metadata	66.81 ± 0.2514	0.5351 ± 0.0118	0.0000 ± 0.0079	0.8332 ± 0.0048	0.7221 ± 0.0178	43.69	69.96	5,575
	Fusion Gain	**+3.84 ± 0.1240**	**+0.0196 ± 0.0020**	**+0.0446 ± 0.0042**	**+0.0448 ± 0.0039**	**-0.1063 ± 0.0142**	**-**	**-**	**+325**
PTQ_full	**Full Multimodal**	86.00 ± 0.0927	0.8456 ± 0.0016	0.7239 ± 0.0032	0.8488 ± 0.0026	0.3968 ± 0.0071	100.00	-	7,181
	Only Images	81.38 ± 0.1030	0.8209 ± 0.0018	0.6726 ± 0.0037	0.8013 ± 0.0030	0.5105 ± 0.0089	55.68	89.78	6,791
	Only Metadata	66.81 ± 0.2514	0.5351 ± 0.0118	0.0000 ± 0.0079	0.6119 ± 0.0048	1.2924 ± 0.0178	44.32	69.41	5,575
	Fusion Gain	**+4.62 ± 0.1390**	**+0.0247 ± 0.0020**	**+0.0513 ± 0.0048**	**+0.0475 ± 0.0039**	**-0.1047 ± 0.0142**	**-**	**-**	**+390**

Table 20. Modality contribution analysis in QAT-quantized MNN models based on the CNN VGG_16

Quantization type VGG_16	Experimental Condition	Accuracy, %	F1-Score	MCC	Confidence	Entropy	Modality Contribution, %	Agreement with Full, %	Correct Predictions
QAT_CNN	**Full Multimodal**	88.26 ± 0.1103	0.8761 ± 0.0016	0.7753 ± 0.0031	0.8733 ± 0.0026	0.3397 ± 0.0071	100.00	-	7,385
	Only Images	82.43 ± 0.1220	0.8383 ± 0.0018	0.7058 ± 0.0037	0.8434 ± 0.0030	0.4208 ± 0.0089	56.04	90.71	6,879
	Only Metadata	66.81 ± 0.2514	0.5351 ± 0.0118	0.0000 ± 0.0079	0.8098 ± 0.0048	0.7794 ± 0.0178	43.95	66.99	5,575
	Fusion Gain	**+5.83 ± 0.1660**	**+0.0378 ± 0.0020**	**+0.0695 ± 0.0048**	**+0.0299 ± 0.0039**	**-0.0811 ± 0.0142**	**-**	**-**	**+506**
QAT_MLP	**Full Multimodal**	88.11 ± 0.2863	0.8745 ± 0.0016	0.7680 ± 0.0064	0.8715 ± 0.0026	0.3451 ± 0.0071	100.00	-	7,404
	Only Images	83.08 ± 0.3170	0.8402 ± 0.0018	0.7091 ± 0.0037	0.8285 ± 0.0030	0.4554 ± 0.0089	56.26	90.29	6,933
	Only Metadata	66.81 ± 0.2514	0.5351 ± 0.0118	0.0000 ± 0.0079	0.8655 ± 0.0048	0.5952 ± 0.0178	43.74	68.54	5,575
	Fusion Gain	**+5.03 ± 0.4280**	**+0.0343 ± 0.0020**	**+0.0589 ± 0.0072**	**+0.0060 ± 0.0039**	**-0.1103 ± 0.0142**	**-**	**-**	**+471**
QAT_classif	**Full Multimodal**	88.03 ± 0.0778	0.8705 ± 0.0016	0.7686 ± 0.0023	0.8724 ± 0.0026	0.3440 ± 0.0071	100.00	-	7,350
	Only Images	82.23 ± 0.0850	0.8343 ± 0.0018	0.6975 ± 0.0037	0.8378 ± 0.0030	0.4395 ± 0.0089	55.98	90.50	6,862
	Only Metadata	66.81 ± 0.2514	0.5351 ± 0.0118	0.0000 ± 0.0079	0.7860 ± 0.0048	0.8622 ± 0.0178	44.02	67.83	5,575
	Fusion Gain	**+5.80 ± 0.1150**	**+0.0362 ± 0.0020**	**+0.0711 ± 0.0042**	**+0.0346 ± 0.0039**	**-0.0950 ± 0.0142**	**-**	**-**	**+488**
QAT_full	**Full Multimodal**	87.56 ± 0.0571	0.8631 ± 0.0016	0.7572 ± 0.0013	0.8520 ± 0.0026	0.3939 ± 0.0071	100.00	-	7,312
	Only Images	81.17 ± 0.0620	0.8253 ± 0.0018	0.6816 ± 0.0037	0.8078 ± 0.0030	0.5083 ± 0.0089	55.61	88.71	6,774
	Only Metadata	66.81 ± 0.2514	0.5351 ± 0.0118	0.0000 ± 0.0079	0.6557 ± 0.0048	1.2265 ± 0.0178	44.39	68.66	5,575
	Fusion Gain	**+6.39 ± 0.0640**	**+0.0378 ± 0.0020**	**+0.0756 ± 0.0039**	**+0.0442 ± 0.0039**	**-0.1144 ± 0.0142**	**-**	**-**	**+538**

Table 21. Comparative table of structurally quantized MNNs based on various CNNs and well-known intelligent systems for classifying pigmented skin lesions with their implementation in edge mobile devices

№	Article	Year	Features	Modality with CNN architecture	Optimization Technique	Acc, %	Speed-up on mobile Device
1.	[41]	2023	A method for optimizing convolutional neural network architecture by using image cropping, weight clustering, and PTQ quantization is proposed	Image (Unimodal) ResNet	PTQ_full	79.76 (-1,64)	Simulation on Raspberry PI Edge without measuring processing speed
2.	[42]	2023	A quantization method QAT is proposed to pre-trained models for classification of the Nailmelonma dataset	Image (Unimodal) *VGG19*	QAT_full	81.00 (-11.00)	GPU GeForce GTX 1650 Super 9.5 ms (1.85x)
				Image (Unimodal) *MobileNet*	QAT_full	84.00 (-8.00)	GPU GeForce GTX 1650 Super 3.5 ms (3.29x)
				Image (Unimodal) *ResNet152_V2*	QAT_full	91.00 (-1.00)	GPU GeForce GTX 1650 Super 7.3 ms (2.67x)
3.	[43]	2025	A distilled model optimized by standard post-training quantization PTQ is proposed	Image (Unimodal) *EfficientNetB0*	PTQ_full	78,13 (-14,09)	*GPU P100* 41 ms (2.95 x)
4.	The proposed method		An adaptive structural quantization method for multimodal models for efficient deployment on mobile edge devices while preserving efficiency	Image+ Metadata (Multimodal) *AlexNet*	Structural quantization PTQ_full Compress 3.97x	85.08 (-0,06)	*Samsung A56* 17,61 ms (2,87x) *Asus ZenFone 2* 78,28 ms (2.97x)
				Image+ Metadata (Multimodal) *ShuffleNet_v2*	Structural quantization QAT_MLP Compress 1.01x	83.02 (+0.97)	*Samsung A56* 50.57 ms (1.03x) *Asus ZenFone 2* 540.35 ms (0.92x)
				Image+ Metadata (Multimodal) *VGG_16*	Structural quantization QAT_CNN Compress 1.66x	88.26 (+0.89)	*Samsung A56* 311.39 ms (1.71x) *Asus ZenFone 2* 1974.85 ms (1.10x)

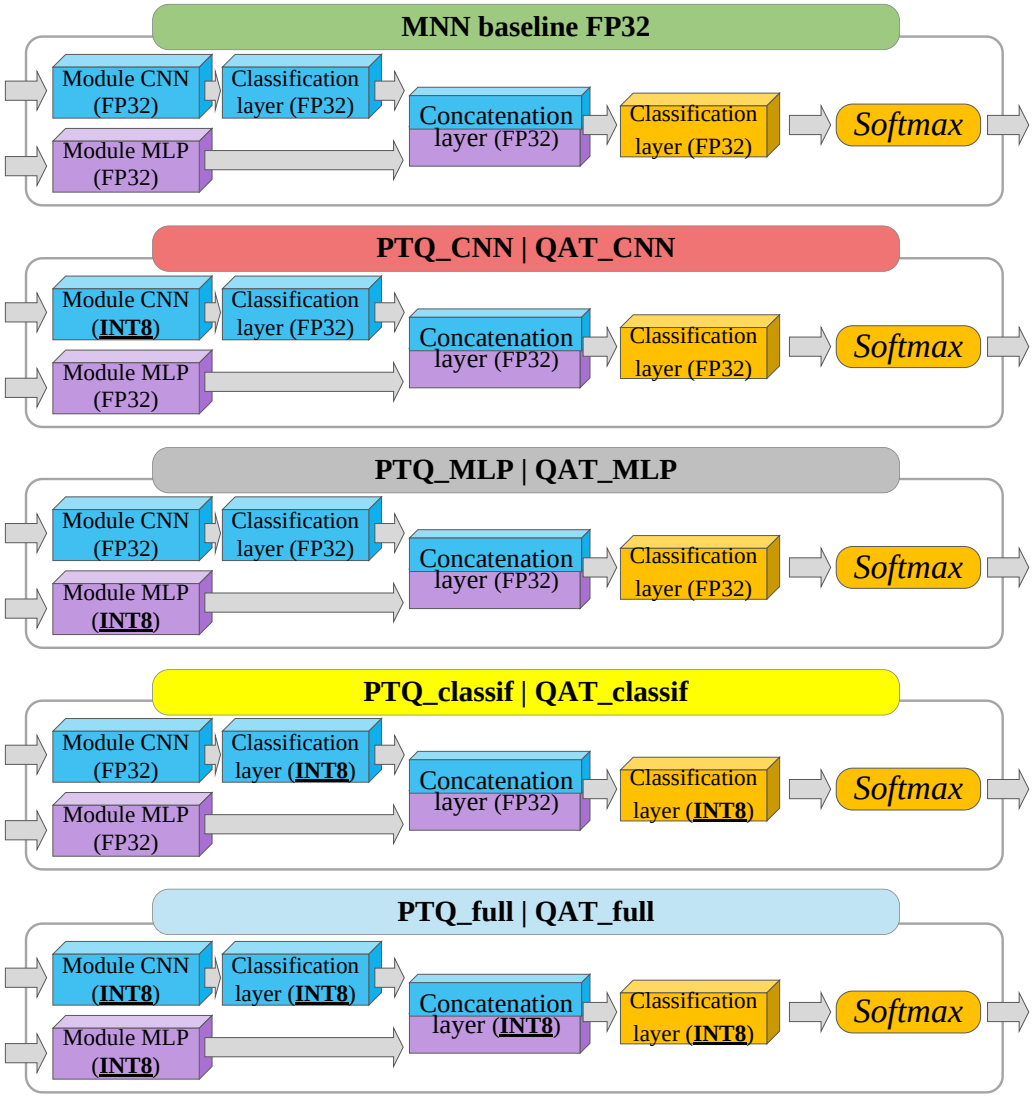


Figure 1. Schematic representation of the FP32 baseline MNN and its INT8-quantized variants, including fully quantized models and models with selectively quantized structural modules

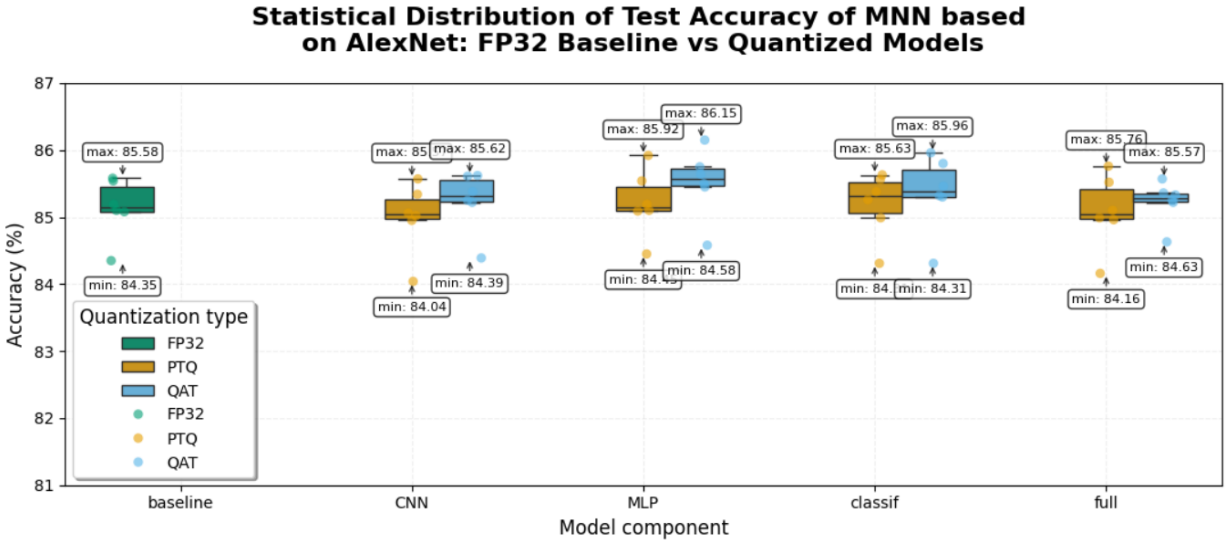


Figure 2. Statistical distribution of test accuracy for the MNN based on the AlexNet CNN for the Baseline FP32 model and quantized models using the PTQ and QAT methods

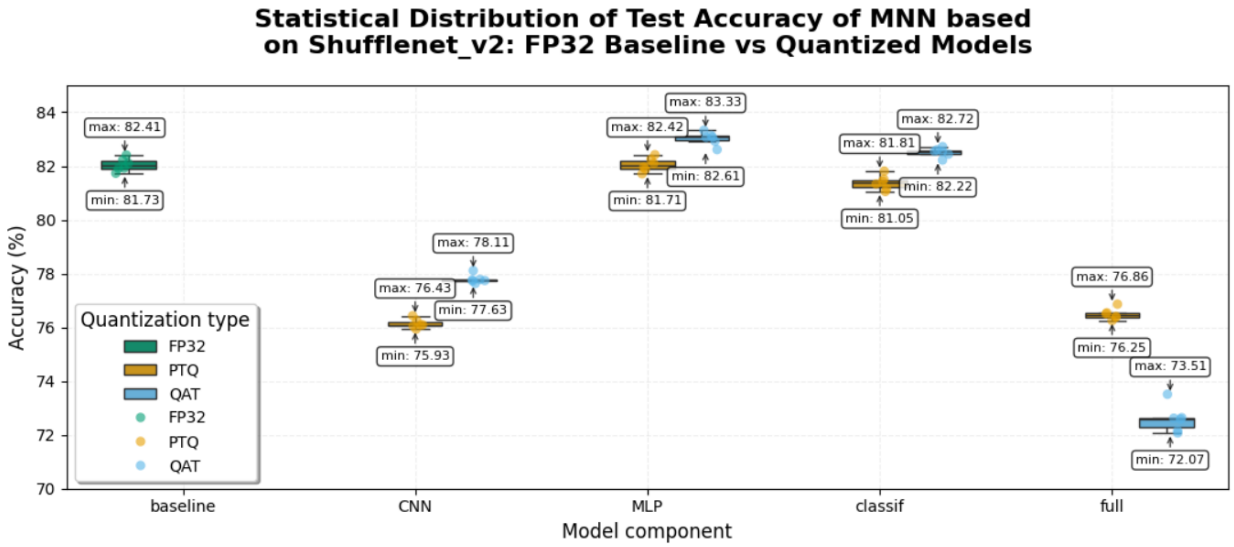


Figure 3. Statistical distribution of test accuracy for the MNN based on the Shufflenet_v2 CNN for the Baseline FP32 model and quantized models using the PTQ and QAT methods

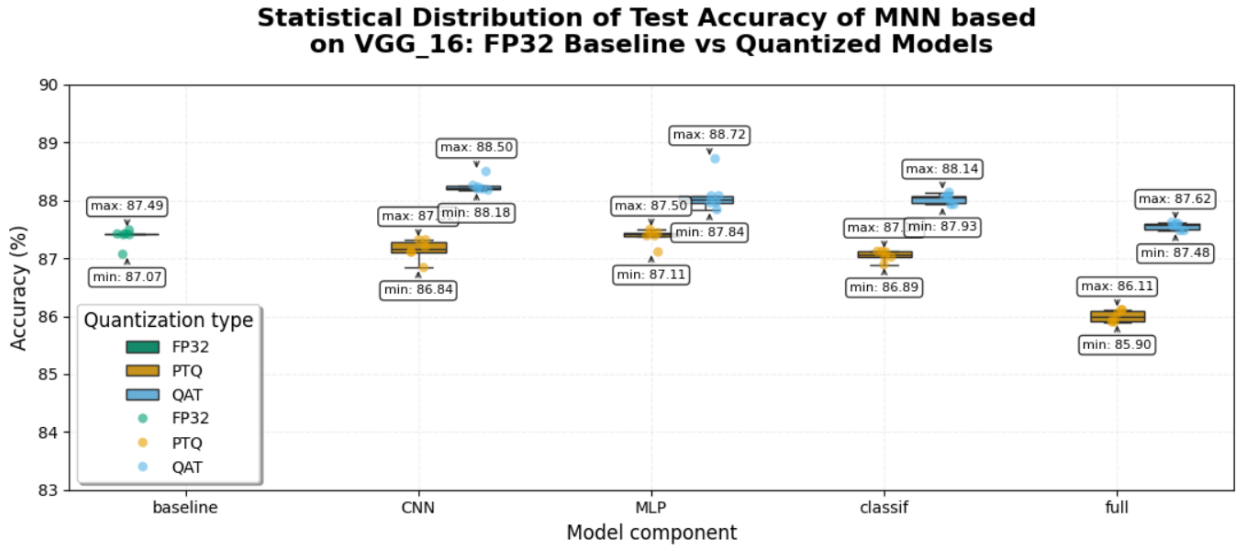
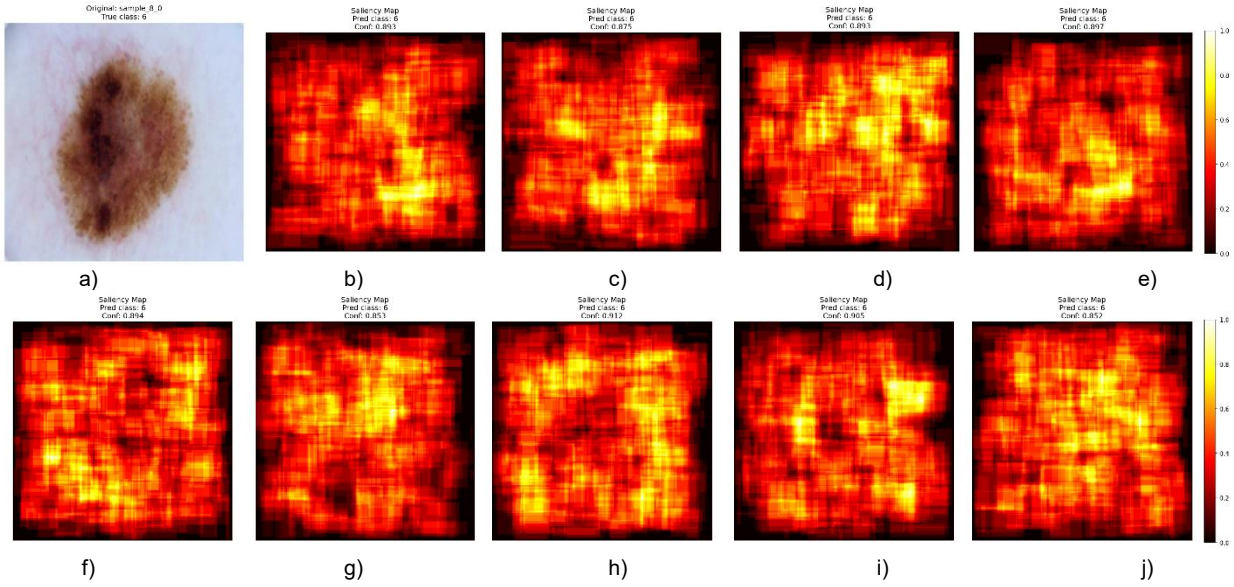


Figure 4. Statistical distribution of test accuracy for the MNN based on the VGG_16 CNN for the Baseline FP32 model and quantized models using the PTQ and QAT methods



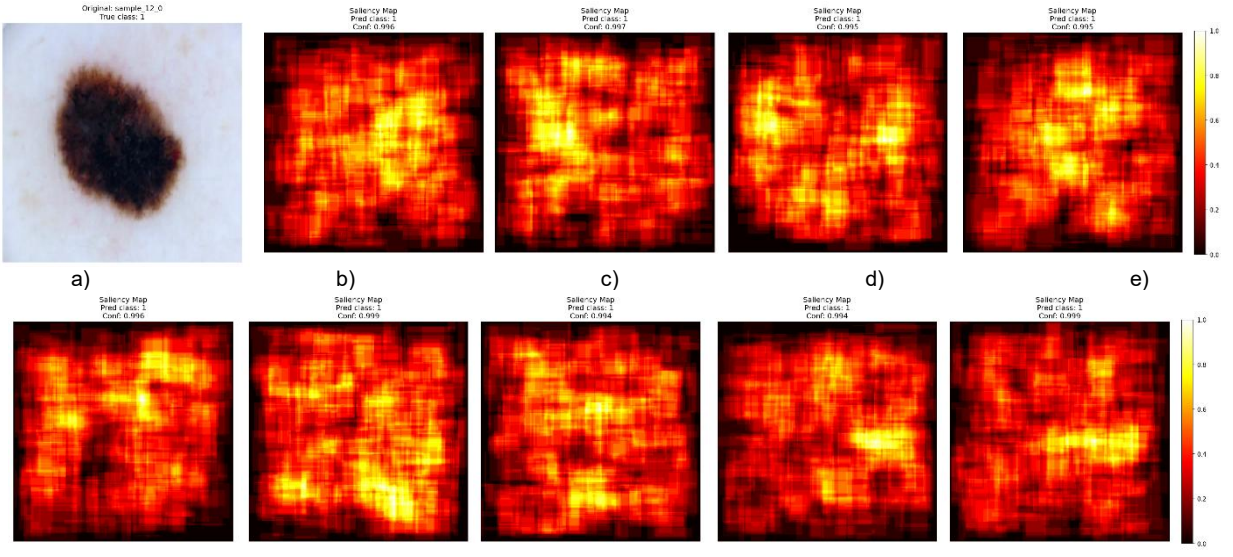


Figure 5. Original dermatological images (a) and attention maps obtained as a result of testing MNN based on CNN AlexNet: b) FP32 baseline; c) PTQ_CNN; d) PTQ_MLP; e) PTQ_classif; f) PTQ_full; g) QAT_CNN; h) QAT_MLP; i) QAT_classif; j) QAT_full

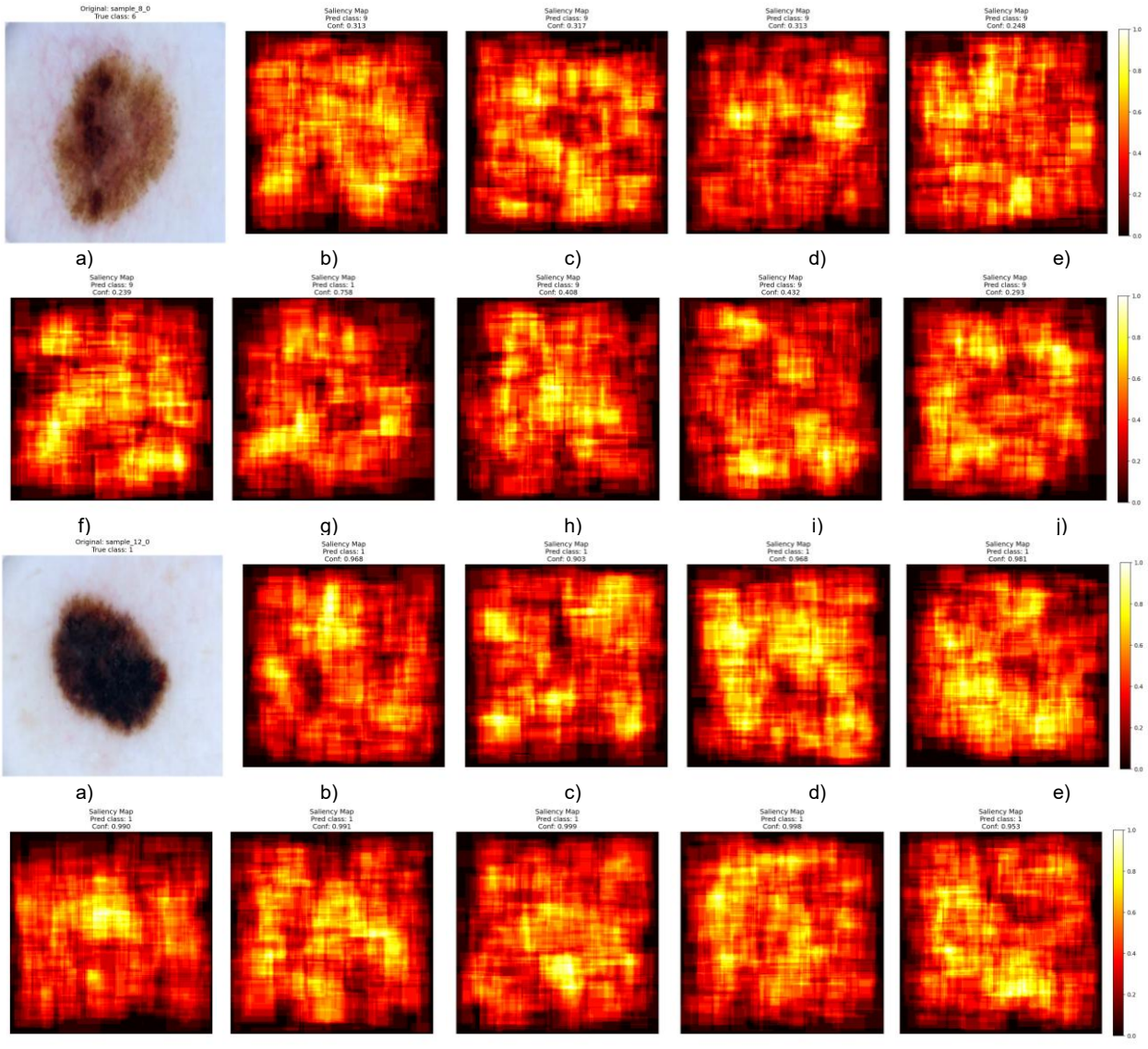
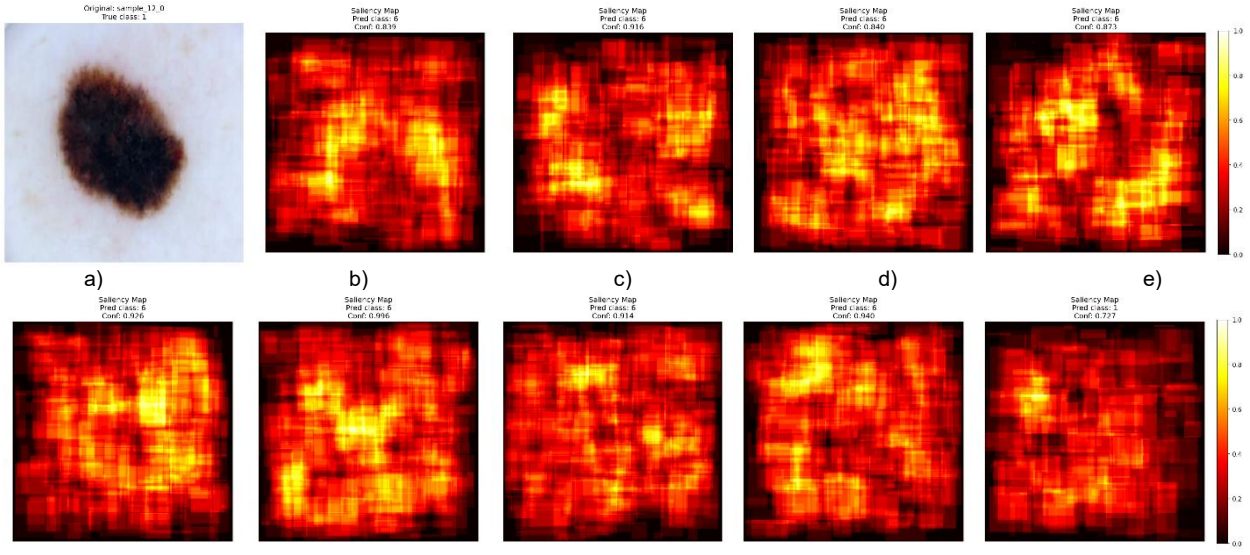


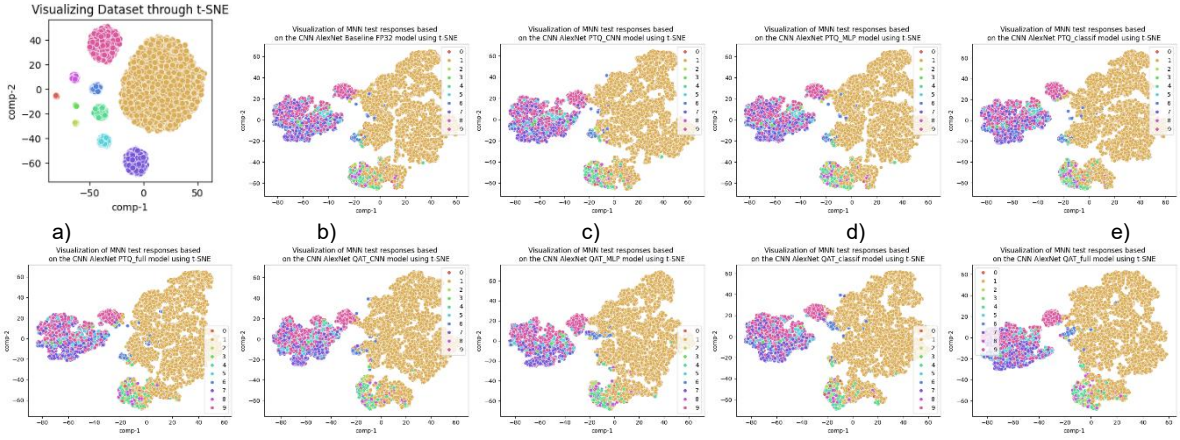
Figure 6. Original dermatological images (a) and attention maps obtained as a result of testing MNN based on CNN Shufflenet_v2: b) FP32 baseline; c) PTQ_CNN; d) PTQ_MLP; e) PTQ_classif; f)
PTQ_full; g) QAT_CNN; h) QAT_MLP; i) QAT_classif; j) QAT_full



a) b) c) d) e)

f) g) h) i) j)

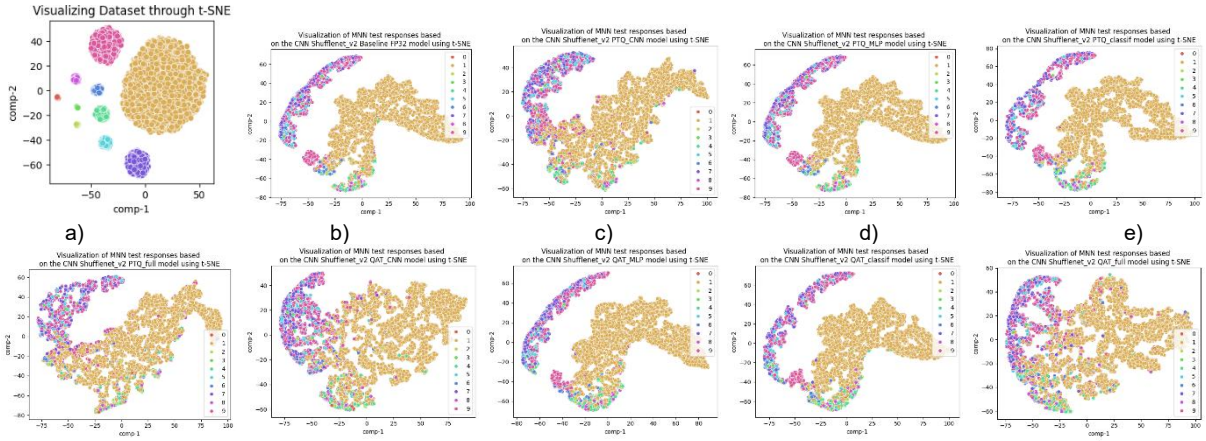
Figure 7. Original dermatological images (a) and attention maps obtained as a result of testing MNN based on CNN VGG_16: b) FP32 baseline; c) PTQ_CNN; d) PTQ_MLP; e) PTQ_classif; f) PTQ_full; g) QAT_CNN; h) QAT_MLP; i) QAT_classif; j) QAT_full



a) b) c) d) e)

f) g) h) i) j)

Figure 8. Visualization of the original distribution of test images (a) and the responses of MNN models based on AlexNet in 2D space using the t-SNE method: b) FP32 baseline; c) PTQ_CNN; d) PTQ_MLP; e) PTQ_classif; f) PTQ_full; g) QAT_CNN; h) QAT_MLP; i) QAT_classif; j) QAT_full



a) b) c) d) e)

f) g) h) i) j)

Figure 9. Visualization of the original distribution of test images (a) and the responses of MNN models based on Shufflenet_v2 in 2D space using the t-SNE method: b) FP32 baseline; c) PTQ_CNN; d) PTQ_MLP; e) PTQ_classif; f) PTQ_full; g) QAT_CNN; h) QAT_MLP; i) QAT_classif; j) QAT_full

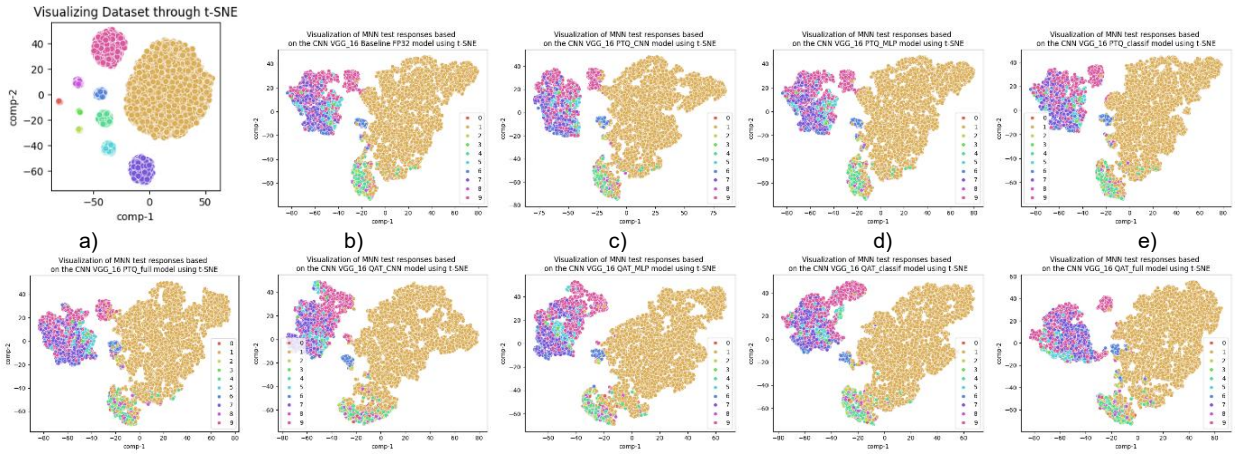


Figure 10. Visualization of the original distribution of test images (a) and the responses of MNN models based on VGG_16 in 2D space using the t-SNE method: b) FP32 baseline; c) PTQ_CNN; d) PTQ_MLP; e) PTQ_classif; f) PTQ_full; g) QAT_CNN; h) QAT_MLP; i) QAT_classif; j) QAT_full

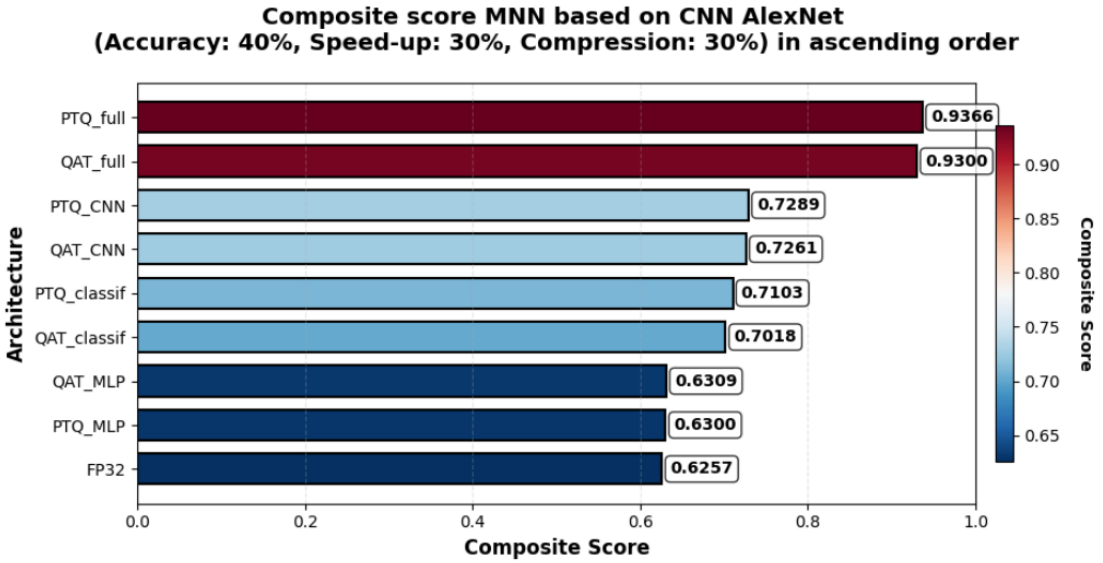


Figure 11. Composite evaluation plot of MNN based on CNN AlexNet (Accuracy: 40%, Speedup: 30%, Compression: 30%) in ascending order

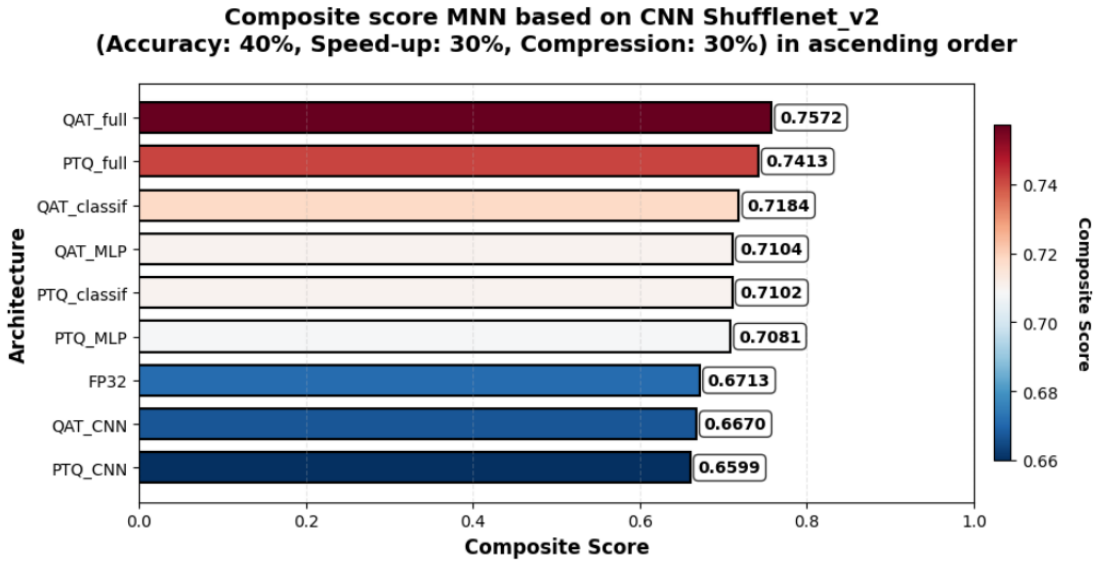


Figure 12. Composite evaluation plot of MNN based on CNN Shufflenet_v2 (Accuracy: 40%, Speedup: 30%, Compression: 30%) in ascending order

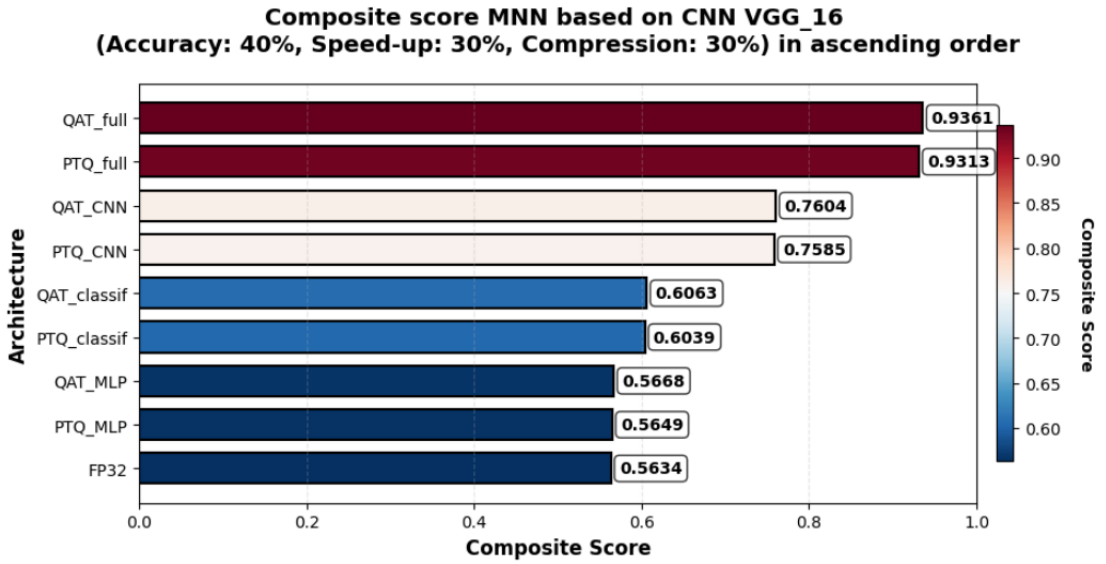


Figure 13. Composite evaluation plot of MNN based on CNN VGG_16 (Accuracy: 40%, Speedup: 30%, Compression: 30%) in ascending order

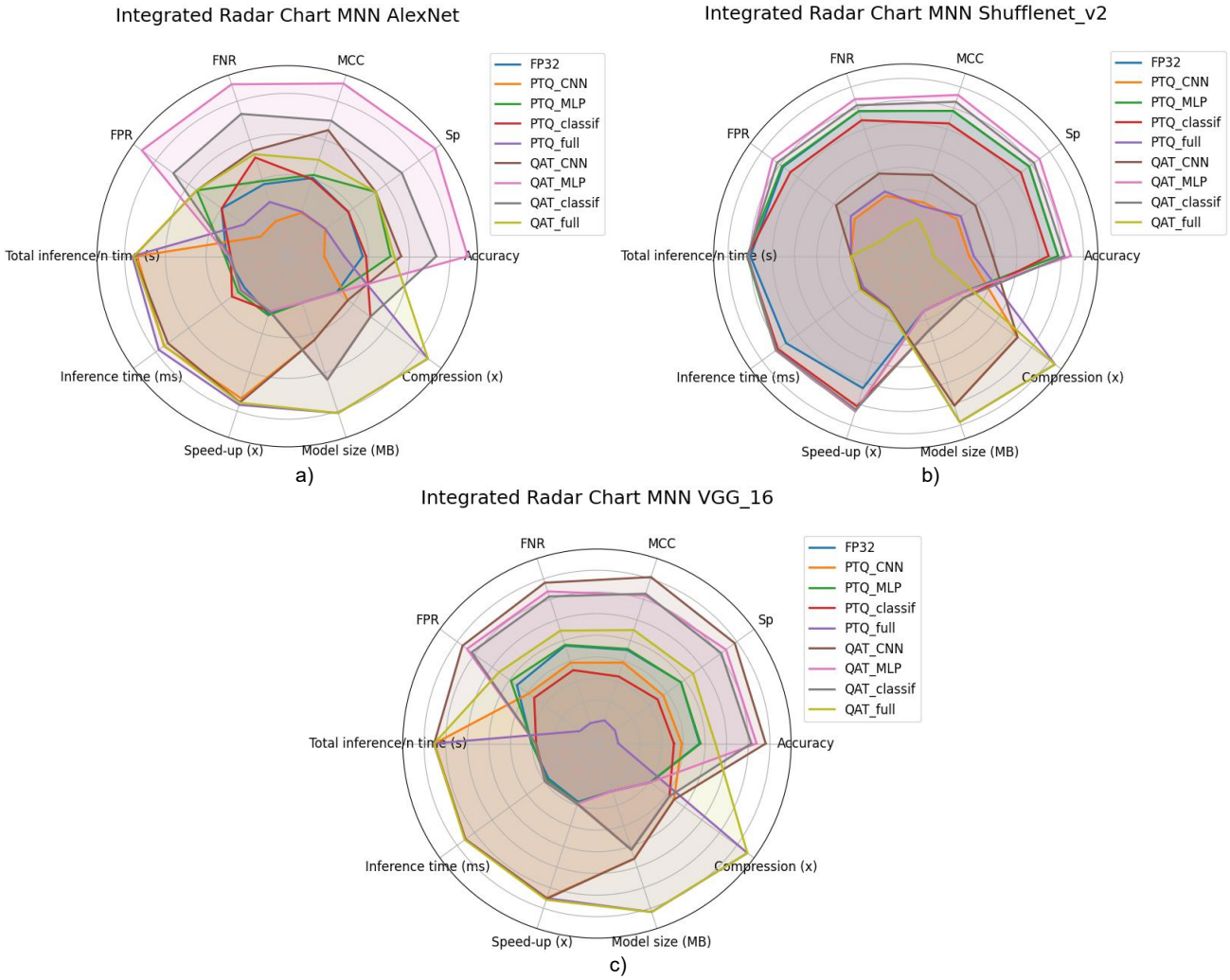
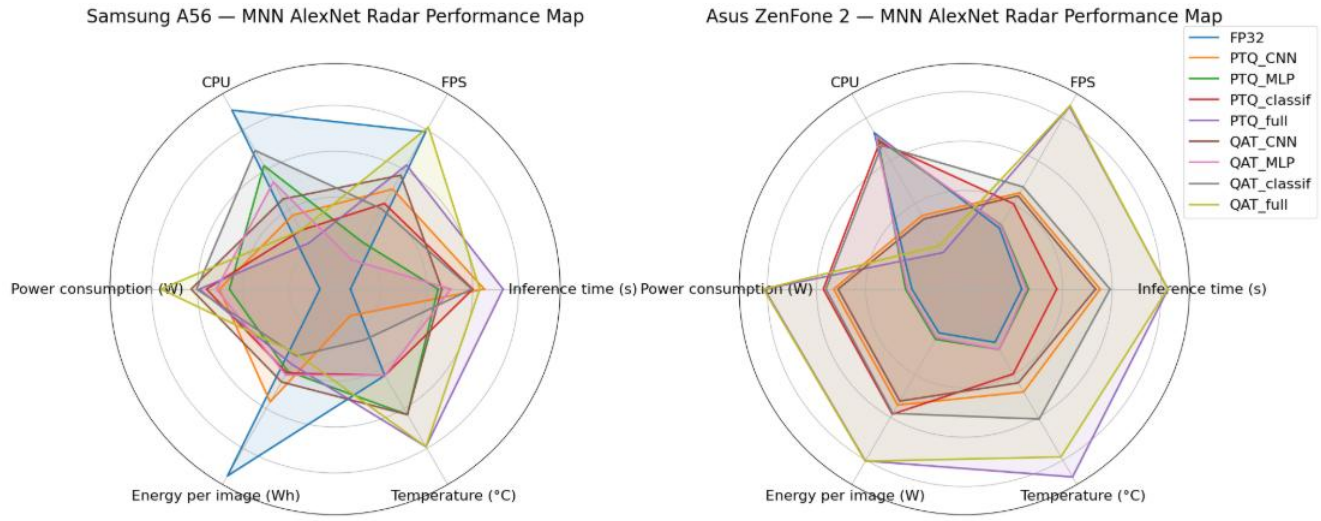
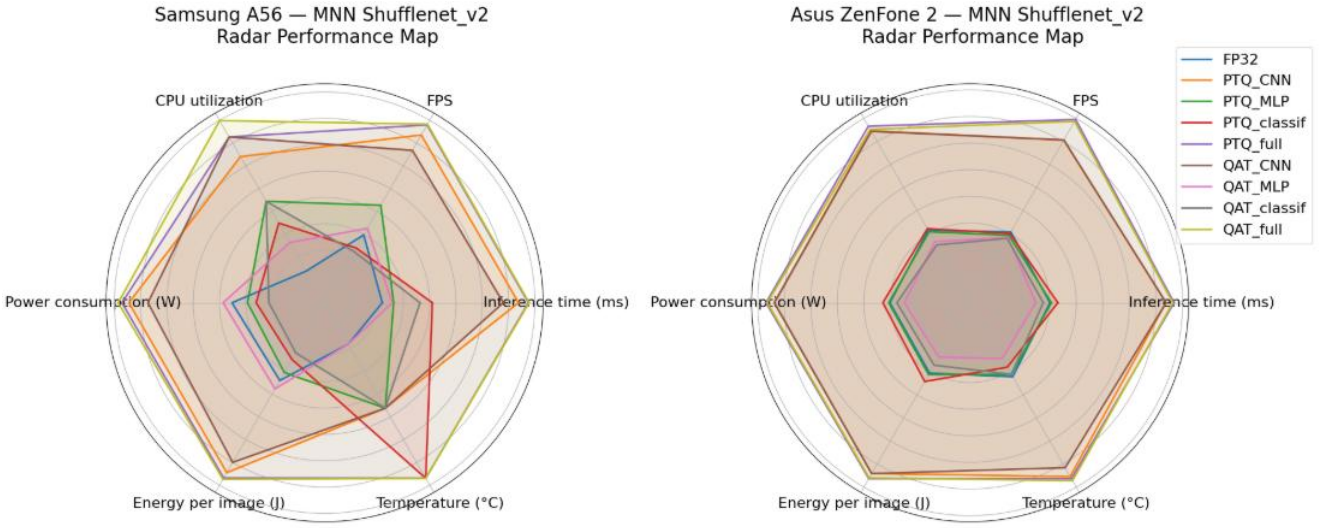


Figure 14. Integrated radar map for the proposed MNN FP32 baseline models and quantized models using PTQ and QAT methods based on CNN architectures: a) AlexNet; b) Shufflenet_v2; c) VGG_16



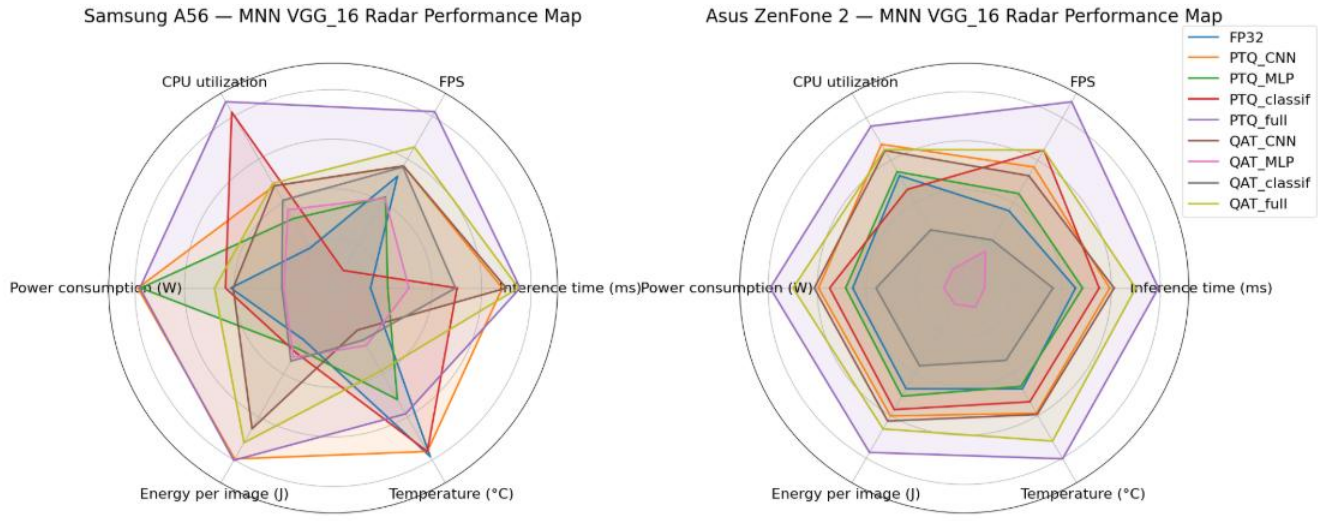
a) b)

Figure 15. Integrated radar performance map for the proposed MNN FP32 baseline models and quantized models by PTQ and QAT methods based on AlexNet CNN architecture on mobile edge devices: a) Samsung Galaxy A56; b) ASUS Zenfone 2



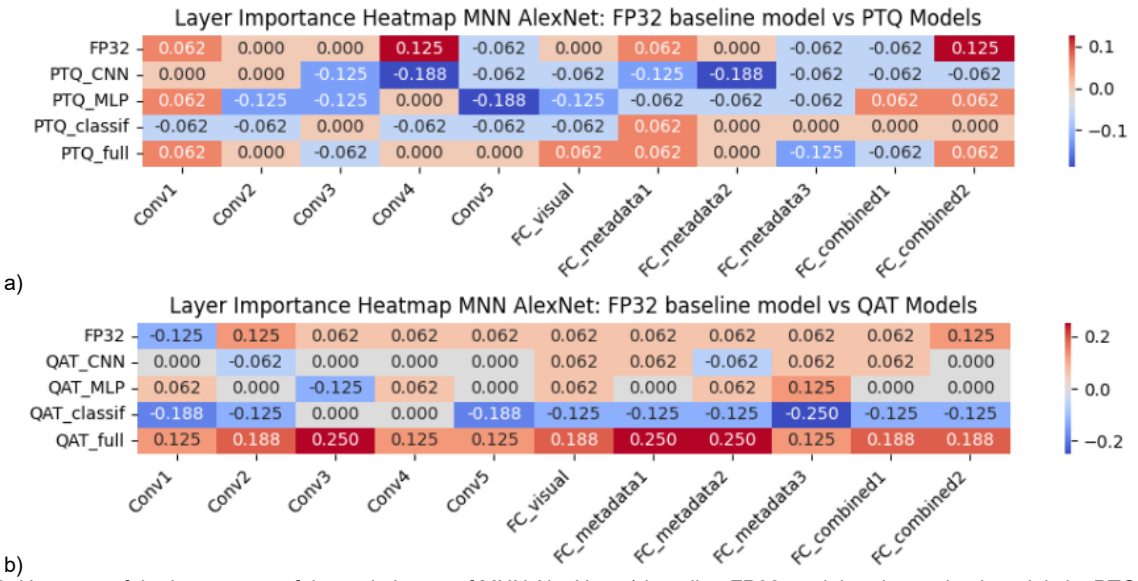
a) b)

Figure 16. Integrated radar performance map for the proposed MNN FP32 baseline models and quantized models by PTQ and QAT methods based on Shufflenet_v2 CNN architecture on mobile edge devices: a) Samsung Galaxy A56; b) ASUS Zenfone 2



a) b)

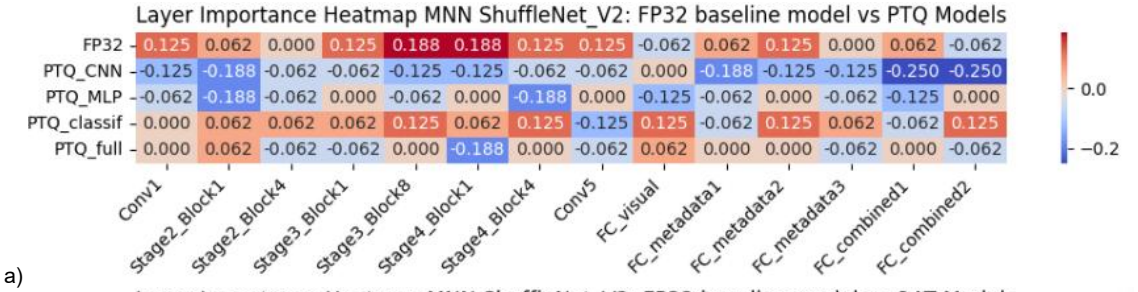
Figure 17. Integrated radar performance map for the proposed MNN FP32 baseline models and quantized models by PTQ and QAT methods based on VGG_16 CNN architecture on mobile edge devices: a) Samsung Galaxy A56; b) ASUS Zenfone 2



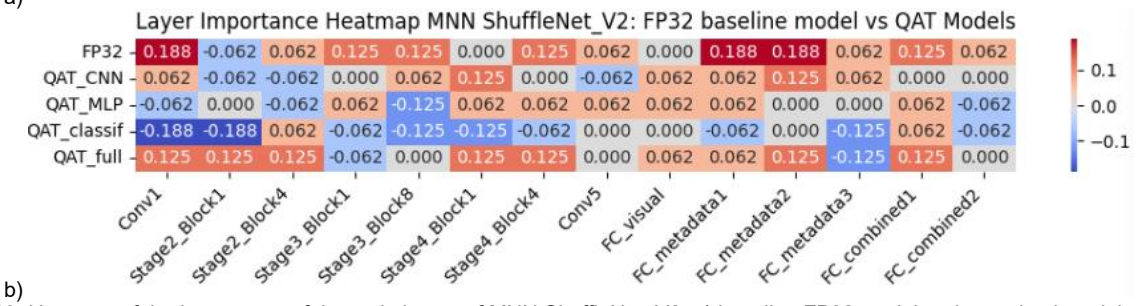
a)

b)

Figure 18. Heatmap of the importance of the main layers of MNN AlexNet: a) baseline FP32 model and quantized models by PTQ method; b) baseline FP32 model and quantized models by QAT method

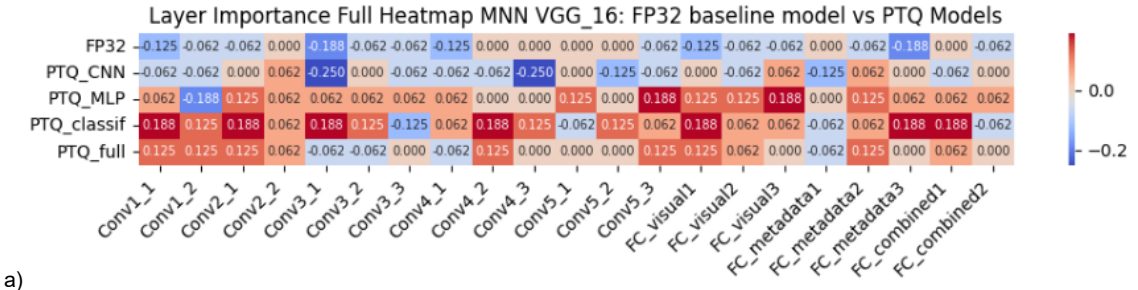


a)

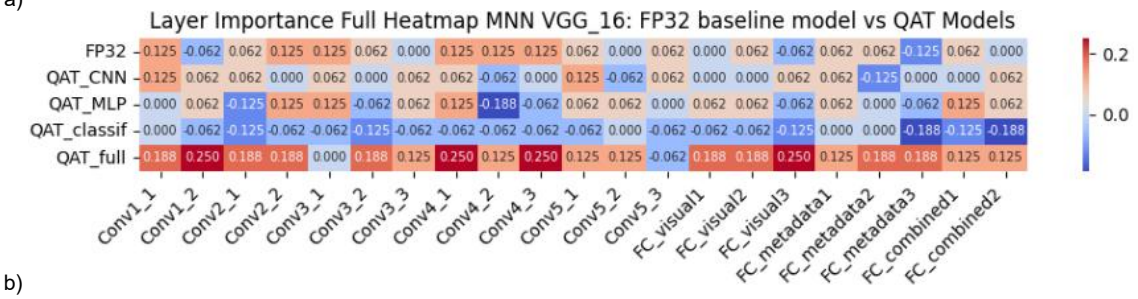


b)

Figure 19. Heatmap of the importance of the main layers of MNN ShuffleNet_V2: a) baseline FP32 model and quantized models by PTQ method; b) baseline FP32 model and quantized models by QAT method



a)



b)

Figure 20. Heatmap of the importance of the main layers of MNN VGG_16: a) baseline FP32 model and quantized models by PTQ method; b) baseline FP32 model and quantized models by QAT method