Automated Skeletal Classification with Lateral Cephalometry Based on Artificial Intelligence

H.J. Yu, S.R. Cho, M.J. Kim, W.H. Kim, J.W. Kim, and J. Choi

Appendix

METHODS DETAILS

Model Architecture and Training details

The diagnostic model was structured with custom layers connected to the base model, where the features extracted from the cephalogram through the layer were concatenated with gender information. The structure of custom layers and hyperparameters were optimized to efficiently combine and analyze the features from the cephalogram and gender information. Keras was used as the framework and was trained using an optimizer, Adam, with standard parameters (beta_1 = 0.9, beta_2 = 0.999 and learning rate = 0.0001). X-rays sized 2,510 x 2,000 require very large computational loads and complex models. Thus, the images were downsampled to 375 x 300 through cropping and resizing while preserving as much diagnostic information as possible. The diagnostic model architecture is illustrated with further detail in Figure 1-b.

Statistical Analysis

Confusion matrices, diagnostic accuracy, sensitivity, specificity, the receiver operating characteristic (ROC) curves, and the area under the curve (AUC) with 95% confidence intervals were used to test the system performance. Confidence interval (95%) was obtained using the Clopper-Pearson method (Newcombe 1998). The following algorithms were used for the calculation of accuracy, sensitivity, and specificity (Memar and Faradji 2017):

Accuracy (AC) = (TP+TN)/(TP+FN+TN+FP) (%)

Sensitivity (SN) = TP/(TP+FN) (%)

Specificity (SP) = TN/(TN+FP) (%)

where TP denotes true positives, TN denotes true negatives, FP denotes false positives, and FN denotes false negatives. P values less than 0.001 were considered significant.

The ROC curves and the AUCs were calculated for each skeletal class. Two types of averages: micro-average and macro-average were used in this work (Yang 1999). AUC is an effective and comprehensive measure of sensitivity and specificity for assessing the inherent validity of a diagnostic test and the overall performance of the ROC curve. Additionally, high AUC values confirm the accuracy by which the model can distinguish patients with or without diseases from a wide range of operating points.

Appendix Table 1. Descriptive statistics: The numbers of patient data distributed in each skeletal class of Models I-III prior to undersampling.

	Model I	Model II	Model III
Sagittal			
Class I	4,326	3,738	3,398
Class II	779	504	412
Class III	785	581	511
Total	5,890	4,823	4,321
Vertical			
Normal	4,115	3,511	3,120
Hyperdivergent	835	598	491
Hypodivergent	940	650	548
Total	5,890	4,759	4,159

Appendix Table 2. Numerical relationship of data between each class of the two diagnostics.

	Normal	Hyperdivergent	Hypodivergent
Model I			
Class I	3,098	541	687
Class II	457	228	94
Class III	560	66	159
Model II			
Class I	2,292	314	389
Class II	230	134	43
Class III	355	33	91
Model III			
Class I	1,874	214	294
Class II	158	113	26
Class III	274	27	69

Appendix Table 3. The area under the curve (AUC) and CI of cephalometric analysis for Models I, II, and III.

	AUC [95% CI]
Sagittal	
Model I	
Class I	0.889 [0.851, 0.927]
Class II	0.950 [0.924, 0.977]
Class III	0.967 [0.945, 0.988]
Micro-average	0.939 [0.910, 0.968]
Macro-average	0.938 [0.909, 0.967]
Model II	
Class I	0.944 [0.909, 0.978]
Class II	0.978 [0.957, 1.0]
Class III	0.981 [0.961, 1.0]
Micro-average	0.970 [0.944, 0.995]
Macro-average	0.970 [0.944, 0.996]
Model III	
Class I	0.965 [0.935, 0.996]
Class II	0.973 [0.946, 1.0]
Class III	0.991 [0.975, 1.0]
Micro-average	0.974 [0.947, 1.0]
Macro-average	0.978 [0.954, 1.0]
Vertical	
Model I	
Normal	0.892 [0.856, 0.928]
Hyperdivergent	0.957 [0.933, 0.980]
Hypodivergent	0.959 [0.935, 0.982]
Micro-average	0.939 [0.911, 0.967]
Macro-average	0.937 [0.909, 0.966]
Model II	
Normal	0.969 [0.945, 0.993]
Hyperdivergent	0.988 [0.973, 1.0]
Hypodivergent	0.992 [0.980, 1.0]
Micro-average	0.976 [0.955, 0.997]
Macro-average	0.984 [0.967, 1.0]
Model III	
Normal	0.971 [0.945, 0.996]
Hyperdivergent	0.989 [0.974, 1.0]
Hypodivergent	0.988 [0.974, 1.0]
Micro-average	0.984 [0.965, 1.0]
Macro-average	0.984 [0.965, 1.0]

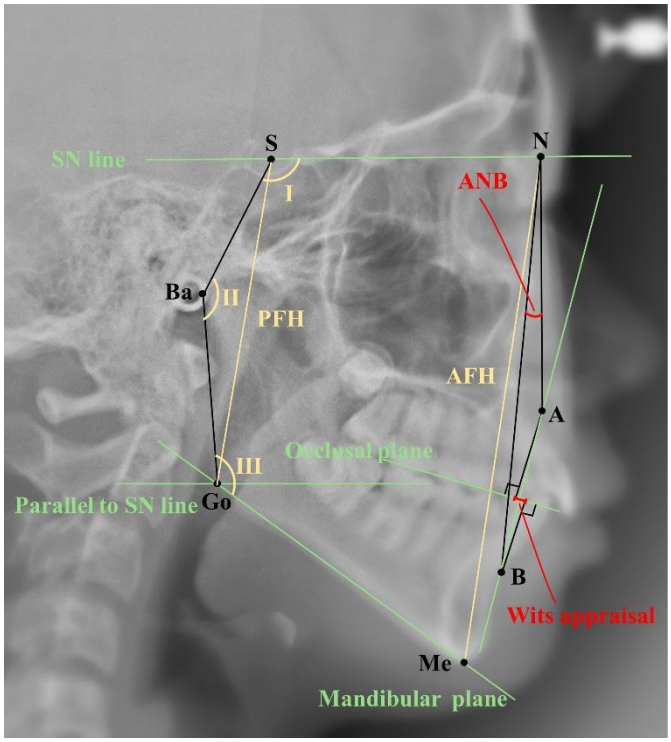
Appendix Table 4. The numbers of patient data assigned to training, validation and test set by for skeletal components in skeletal discrepancies and its severity.

	Training	Validation	Test
Discrepancies			
Maxilla	1,914	396	396
Mandible	2,007	423	423
Severity			
Class II	563	117	117
Class III	569	117	117

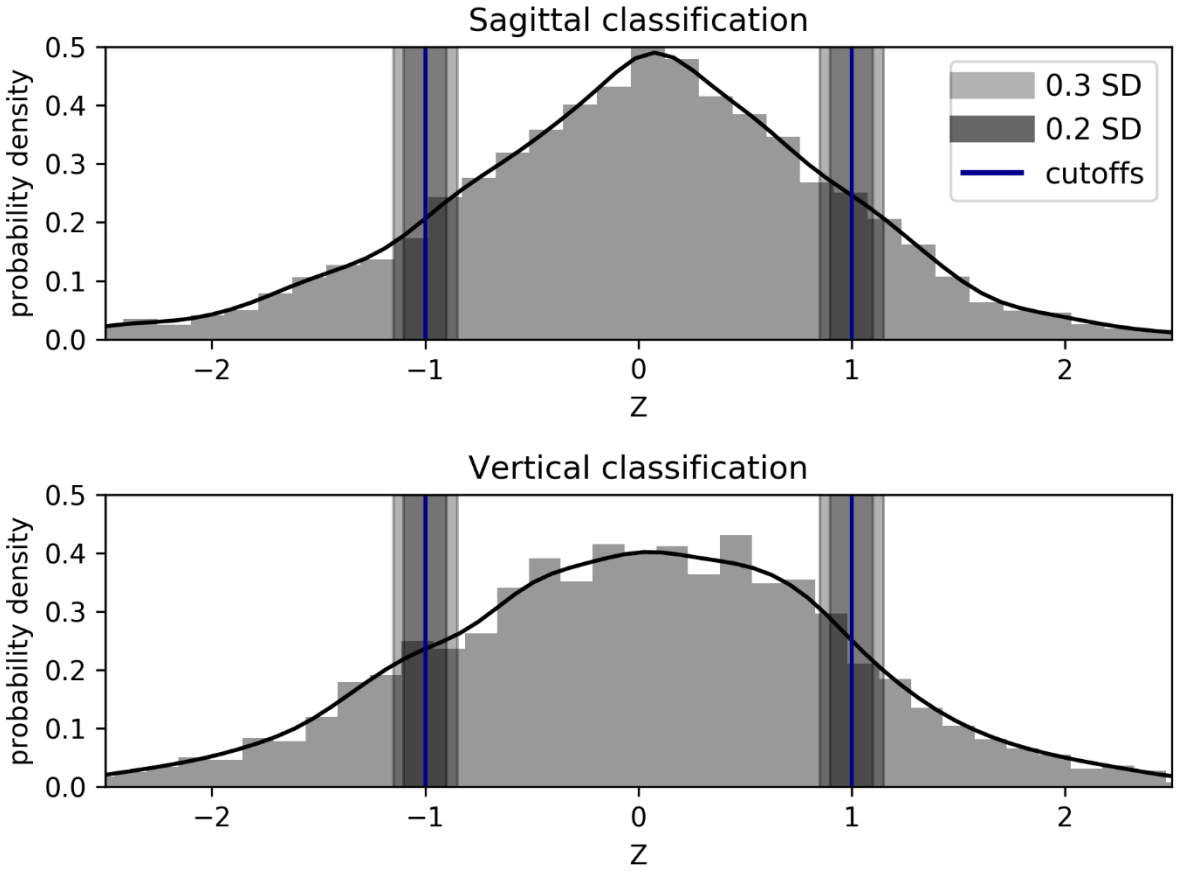
Appendix Table 5. Performance of cephalometric analysis for skeletal components in skeletal discrepancies and its severity.

	SN [95% CI]	SP [95% CI]	AC [95% CI]
Discrepancies			
Maxilla			
Normal	68.93 [60.30, 76.70]	83.71 [78.69, 89.75]	78.79 [74.43, 82.71]
Protrusion	81.81 [74.17, 87.99]	89.77 [85.47, 93.15]	87.12 [83.42, 90.26]
Retrusion	84.09 [76.72, 89.87]	93.94 [90.34, 96.50]	90.66 [87.35, 93.34]
Mean	78.28 [70.40, 84.86]	89.14 [84.84, 92.53]	85.52 [81.73, 88.77]
Mandible			
Normal	65.25 [56.78, 73.06]	89.72 [85.57, 93.00]	81.56 [77.53, 85.14]
Protrusion	88.65 [82.27, 93.37]	92.55 [88.84, 95.33]	91.25 [88.14, 93.77]
Retrusion	89.36 [83.06, 93.92]	89.36 [85.16, 92.71]	89.36 [86.02, 92.13]
Mean	81.09 [74.02, 86.79]	90.54 [86.52, 93.68]	87.39 [83.90, 90.34]
Severity			
Class II			
Moderate	56.41 [39.62, 72.19]	82.05 [71.72, 89.83]	73.50 [64.55, 81.23]
Severe	48.72 [32.42, 65.22]	83.33 [73.19, 90.82]	71.79 [62.73, 79.72]
Mild	82.05 [66.47, 92.46]	78.21 [67.41, 86.76]	79.49 [71.03, 86.39]
Mean	62.39 [46.17, 76.62]	81.20 [70.77, 89.13]	74.93 [66.10, 82.45]
Class III			
Moderate	64.10 [47.18, 78.80]	82.05 [71.72, 89.83]	76.07 [67.30, 83.47]
Severe	51.28 [34.78, 67.58]	79.49 [68.84, 87.80]	70.09 [60.93, 78.20]
Mild	82.05 [66.47, 92.46]	87.18 [77.68, 93.68]	85.47 [77.76, 91.30]
Mean	65.81 [49.47, 79.61]	82.91 [72.75, 90.43]	77.21 [68.66, 84.32]

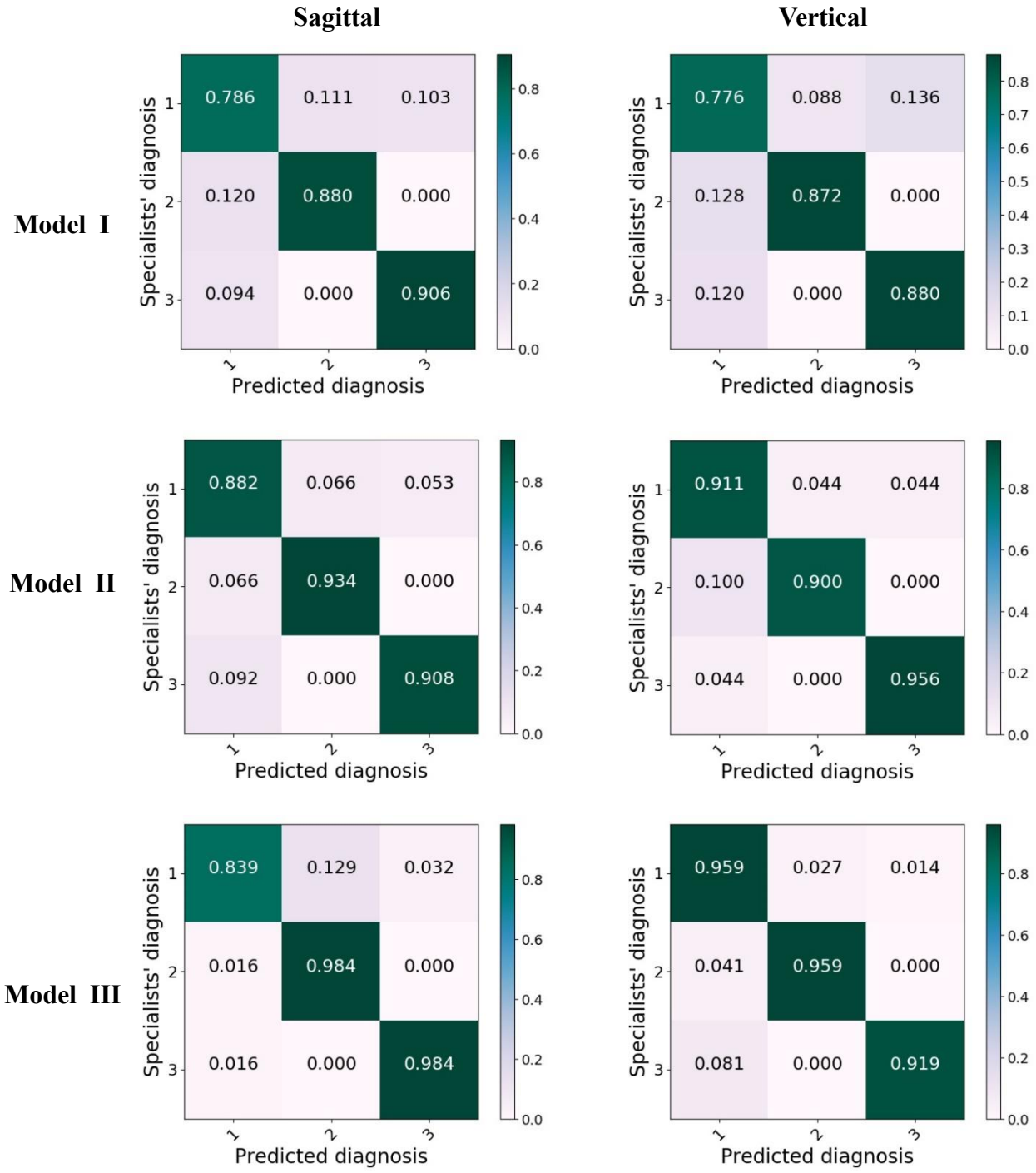
AC, accuracy
SN, sensitivity
SP, specificity



Appendix Figure 1. Cephalometric analysis used in this study. *Landmarks*. S, sella; N, nasion; A, subspinale; B, supramentale; Me, menton; Go, gonion; Ba, basion. PFH: posterior facial height. AFH: anterior facial height. ANB and Wits appraisal are described in the figure. Jarabak's ratio is PFH/AFH. Björk'sum is Saddle angle (I) + Articular angle (II) + Gonial angle (III).



Appendix Figure 2. Normalization graphs for sagittal and vertical classifications illustrating Model II excluding data in the interval of 0.2SD from the cutoff and Modell III excluding data in the interval of 0.3SD from the classification cutoff.



Appendix Figure 3. Confusion matrices. Row: Models (Model I, II, III). Column: cephalometric analysis (Sagittal, Vertical).

References

Memar P, Faradji F. 2017. A novel multi-class eeg-based sleep stage classification system. J IEEE Transactions on Neural Systems Rehabilitation Engineering. 26(1):84-95.

Newcombe RG. 1998. Two-sided confidence intervals for the single proportion: Comparison of seven methods. J Statistics in medicine. 17(8):857-872.

Yang Y. 1999. An evaluation of statistical approaches to text categorization. J Information retrieval. 1(1-2):69-90.