

Comparison of Convolutional Neural Networks Architectures for Screening Mandibular-Plane-to-Hyoid Distance as a Risk Indicator of Obstructive Sleep Apnea on Lateral Cephalograms

Nichakorn Songsang* Udom Thongudomporn** Witsarut Upalananda***

Abstract

Background: Lateral cephalometric radiographs can aid in screening obstructive sleep apnea (OSA) through the mandibular-plane-to-hyoid (MP-H) distance. An automated tool based on this measure could support OSA screening. However, the accuracy of such tools depends on the performance of AI-assisted models, making it essential to evaluate and compare their effectiveness in detecting MP-H distance for reliable clinical application. **Objective:** To evaluate the performance of different convolutional neural network (CNN) architectures in classifying patients into short and long MP-H groups. **Materials and methods:** A total of 304 pre-orthodontic lateral cephalometric radiographs from adults (age 18-56 years) were classified into short MP-H (< 18 mm) and long MP-H ($\geq$ 18 mm) groups. Four CNN architectures (DenseNet-121, ResNet-50, EfficientNet-B0, and MobileNetV3) were trained to classify short and long MP-H groups. To address class imbalance, weighted binary cross entropy loss functions (weights ranging from 1 to 5) were applied, assigning greater penalties to misclassification of the minority class. **Results:** In the scenario without application of weighted cross entropy loss, DenseNet-121 achieved the overall high screening performance, with sensitivity = 0.87, specificity = 0.95, precision = 0.79, F1-score = 0.82, accuracy = 0.93, AUROC = 0.91, and AUPRC = 0.83. MobileNetV3 consistently demonstrated the lowest performance. Weighted loss functions provided inconsistent benefits across architectures. DenseNet-121 showed consistent performance among all weights. **Conclusion:** DenseNet-121 shows potential of screening long MP-H distance from lateral cephalograms. Weighted loss functions may provide improvements, but model selection remains the more critical factor.

Keywords: Cephalometry, Convolutional neural networks, Hyoid bone, Mandible, Sleep apnea

Received: 21-Oct-2025 **Revised:** 29-Dec-2025 **Accepted:** 13-Jan-2026

Corresponding author: Witsarut Upalananda

E-mail: witsarut.u@psu.ac.th

* Dentist, Private Dental Clinic, Bangkok, Thailand

** Professor, Faculty of Dentistry, Prince of Songkla University, Hatyai, Songkhla, Thailand

*** Assistant Professor, Faculty of Dentistry, Prince of Songkla University, Hatyai, Songkhla, Thailand

Introduction

Obstructive sleep apnea (OSA), a multifactorial disorder, is characterized by recurrent episodes of partial (hypopnea) or complete (apnea) upper airway obstruction during sleep.¹ The prevalence of OSA in the Thai population has been reported to be 11.40 % with gender variation of 15.40 % in men and 6.30 % in women.² The standard diagnostic procedure for OSA is polysomnography (PSG), a comprehensive sleep study, which aids in classifying OSA severity based on the Apnea-Hypopnea Index (AHI). It quantifies the average number of apnea and hypopnea events per hour of sleep. The severity of OSA is classified into three categories: mild ($5 \leq \text{AHI} < 15$) moderate ($15 \leq \text{AHI} < 30$), and severe ($\text{AHI} \geq 30$).¹

Orthodontists are involved in the early screening of OSA, as they often see patients who have risk factors and craniofacial features linked to the condition. However, it is not practical to refer all patients without symptoms for a full PSG. Also, PSG availability is often restricted to large hospitals equipped with specialized sleep centers, and the associated costs can be considerable. Normally, dentists are positioned to screen and detect potential patients with an examination and screening questionnaire before referral.³

Lateral cephalometric radiographs offer a non-invasive method for identifying anatomical risk factors related to OSA. By analyzing specific anatomical structures, orthodontists can gain valuable insights into a patient's airway morphology, potentially serving as a cost-effective and efficient initial screening tool.⁴

Several cephalometric anatomical parameters have been reported in association with OSA, including reduction in posterior airway space, enlargement and elongation of uvula, increased tongue size, alteration in cranial base length and angulation, maxillomandibular size and anteroposterior position, increase in anterior lower facial height (ALFH), and hyoid bone position.^{5,6} Nevertheless, evidence from a recent systematic review

indicates that, among these variables, only an increased ALFH, a reduced pharyngeal airway space, and an inferiorly displacement of hyoid bone demonstrated consistent associations with OSA in adult subjects, with negligible heterogeneity across studies.⁶

Hyoid bone is a crucial structure for maintaining airway patency as it anchors several muscles.⁷ The maintenance of pharyngeal airway patency involves at least four muscle groups: the hyoid muscles (the mylohyoid and geniohyoid), which facilitate anterior movement of the hyoid; the muscles of the soft palate, such as the tensor palatini; the pharyngeal constrictor muscles; and the muscles of the tongue, all of which collectively contribute to the dilation or narrowing of the pharynx.⁸

A lower position of the hyoid bone was related to greater mandibular-plane-to-hyoid (MP-H) distance which was reported to be reliability parameter associated in determining severity of OSA.⁹⁻¹¹ Significant variations in MP-H have been observed among patients with different degrees of OSA severity.⁹ The MP-H cutoff values were varied depending on the studied population. In the Thai population, an MP-H cutoff value of ≥ 18 mm is associated with an increased risk of an $\text{AHI} \geq 15$, indicating moderate-to-severe OSA (AUROC = 0.65-0.86).⁹ A study examined the association between hyoid bone position and OSA severity in 133 Korean patients with an $\text{AHI} \geq 5$. A threshold value of 19.45 mm (AUROC = 0.62) was identified as predictive of increased OSA severity, including higher frequency of respiratory events, reduced oxygen saturation, and interrupted sleep patterns.¹⁰ Another study reported a MP-H distance of more than 15.40 mm could indicate OSA in American patients.¹² In Iranian population, a MP-H distance of 14 mm correlated with moderate OSA (AUROC = 0.98).¹³ While in Swedish population, the MP-H distance of greater than 22.80 mm was found in patients with OSA.¹⁴ Although, the cutoff value varied among different population, it was confirmed that the longer the MP-H distance or the more inferior the hyoid

bone placed was associated with OSA severity regardless of ethnicities.⁹⁻¹¹ While multiple craniofacial features may contribute to OSA pathophysiology, inferiorly displaced hyoid represented the most robust and reproducible cephalometric landmark. To clarify, soft tissue landmarks used to indicate airways are often obscured in radiographs due to superimposed structures, making them less reliable. Hence, bony landmarks provide orthodontists with more dependable reference points.¹⁵

Computer vision, employing convolutional neural networks (CNN), enables machines to efficiently process and interpret visual data. In image classification, algorithms identify the objects presented within an image by analysing features such as color, shape, and edges to distinguish between different object classes.¹⁶

Developing an automated diagnostic tool to differentiate patients based on MP-H distance in lateral cephalometric radiographs could be beneficial for screening, early detection and diagnosis and then referred for further investigation. Our study used MP-H as OSA screening parameter as the MP-H distance may serve as an additional diagnostic tool for the screening of severe OSA. The objective of this study was to evaluate the image classification performance of different CNN models in classifying patients by the 18-mm MP-H cutoff using adult pretreatment lateral cephalometric radiographs.

Materials and methods

A total of 304 pre-orthodontic treatment lateral cephalometric radiographs of adults were collected from a private dental clinic in Uttaradit, Thailand. The digital formats of the radiographs were gathered from Orthophos XG 5 DS Ceph (Dentsply, Sirona) radiographic machines and exported as JPEG format (.jpg) with a resolution of 1804 × 2136 pixels. The inclusion criteria included pretreatment lateral cephalograms from orthodontic patients with

permanent dentition (aged 18 years or older). All patients were asymptomatic for OSA and undiagnosed. Patients' information, including gender, date of birth, and date of radiograph being taken, were collected. Lateral cephalograms with improper image quality, poor visual image quality, including brightness, contrast, and noise, poor geometric image quality, including distortion, artifacts, and blur were excluded. Also, patients with a history of head and neck trauma who had been diagnosed with head and neck tumors or cancers and had undergone head and neck surgery were also excluded. All images were manually labeled with Labelme, an open-source image labeling tool.¹⁷ Main examiner (Examiner A) was standardized with an experienced radiologist (Examiner B), and an experienced orthodontist (Examiner C). The annotated landmarks and related measurements were presented in Figure 1.

An interexaminer reliability among three examiners was performed. After one month, the intraexaminer reliability between the first and second trials of the main examiner was performed one month apart to check reproducibility for human ground teeth MP-H distance using intraclass correlation coefficient.

In this study, the short and long MP-H groups were defined based on an 18 mm cutoff value, which was proposed by Banhiran et al. as being correlate with moderate-to-severe OSA in Thai population.⁹ Therefore, the samples with MP-H less than 18 mm were labelled as short MP-H group, those over 18 mm as long MP-H group.

Kolmogorov-Smirnov and Shapiro-Wilk tests were performed to check for normal distribution of data characteristics (age, gender, skeletal classification, divergent pattern, and MP-H distance) in training, validating and test group. All collected data were partitioned using stratified random sampling based on gender and MP-H groups into training, validation, and test sets, with ratios of 0.70, 0.15, and 0.15, respectively. The training set was used to train the models. The validation set was used to optimize

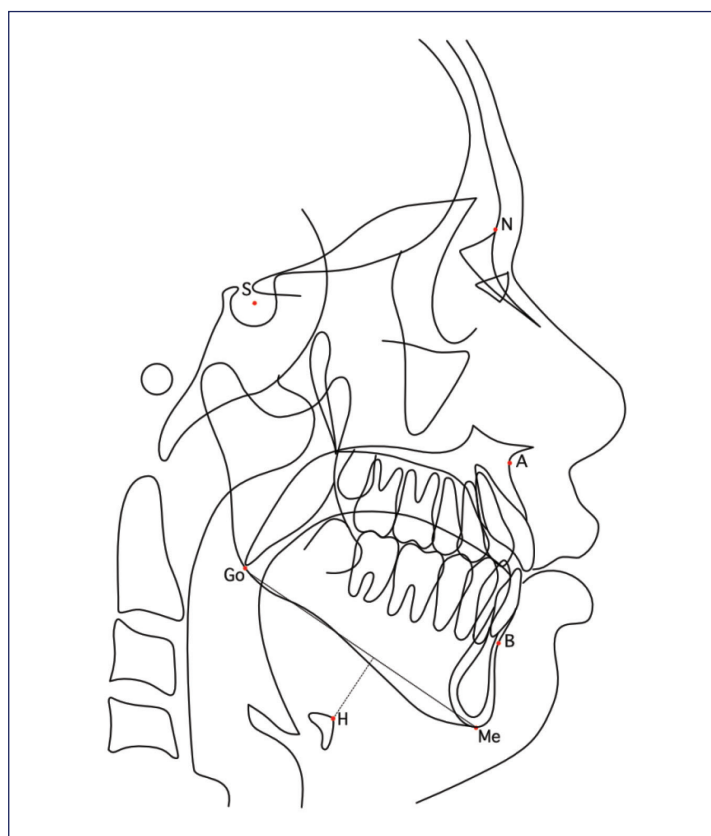


Figure 1 The annotated landmarks included S (Sella), N (Nasion), A (Point A), B (Point B), Go (Gonion), Me (Menton), and H (Hyoidal - the most anterosuperior point of hyoid bone) The SNA, SNB, SN-GoMe angles, and MP-H distance were measured. The ANB and SN-GoMe angle were used to analyze the skeletal classification and divergent pattern of the patients consecutively. The mandibular plane (MP) was constructed using line passing through Go-Me. The MP-H was measured in millimetres by constructing a line from H perpendicular to MP line.

the hyperparameters of a classifier during the training process. Finally, the test set was used to evaluate the performance of the models. For the image preprocessing, the collected radiographs were pre-processed using an automated segmentation pipeline to eliminate non-informative white borders. Using the OpenCV library, a dynamic threshold was established for each image based on the 60th percentile of its pixel intensity distribution to remove peripheral white borders. To standardize the dataset for downstream analysis, images were centrally zero-padded to a uniform dimension of 2037×2037 pixels. This approach preserved the original aspect ratio and spatial integrity of the samples while ensuring a consistent input architecture. Finally, these standardized images

were resized to 224×224 pixels before subsequent processing through the deep learning models (Figure 2). In addition, during the training process, image transformations were applied to the training data as data augmentation, enhancing the model's ability to generalize to unseen data. Augmentation techniques included adding Gaussian noise and rotating images $\pm 10\%$. For the training parameters, model training was performed using the AdamW optimizer with an initial learning rate of 1×10^{-3} and a weight decay of 10×10^{-4} . A OneCycle learning rate scheduling strategy was applied, with the maximum learning rate set to 1×10^{-3} . Thirty percent of the total training steps were allocated to the warm-up phase, followed by cosine annealing. The learning rate was

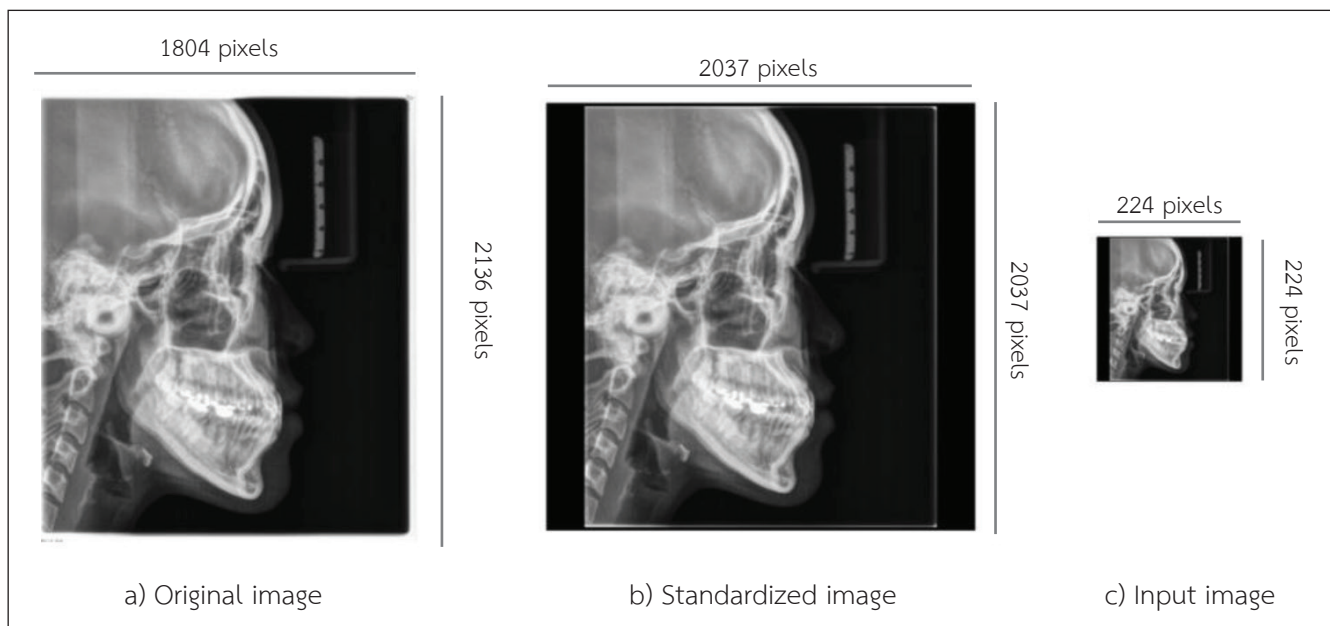


Figure 2 Preprocessing workflow of lateral cephalometric images for CNN input

initially reduced using a division factor of 25 and further decreased with a final division factor of 1×10^4 . All models were trained for a maximum of 100 epochs with a batch size of 64. Early stopping with a patience of 10 epochs was employed to terminate training when no further performance improvement was observed.

Four CNN models, including DenseNet-121, ResNet-50, EfficientNet-B0 and MobileNetV3, were utilized to classify images into risk assessment as either short MP-H or long MP-H group following Thai MP-H cutoff value. Due to the existing constraints in knowledge concerning the most suitable deep learning architecture for airway classification problem, a variety of models were selected. These models represent different methodological approaches designed to improve image classification accuracy, enabling a comprehensive assessment of model performance within this framework. DenseNet, ResNet, EfficientNet architectures were used by previous researchers in CNN-aided airway image classification and landmark detection.^{15,18,19} However, the inclusion of MobileNetV3, a lightweight model, was intended to explore potential use in clinical environments with limited computational resources. Each of these models

uses a unique approach to feature extraction and efficiency. DenseNet-121 (approximately 8 million parameters), invented by Cornell University & Facebook AI Research in 2017, improves feature propagation and reuse by densely connecting each layer to every other layer within a block. It connects each layer to all subsequent layers, promoting feature reuse and improving parameter efficiency.²⁰ ResNet-50 (approximately 25.60 million parameters), developed by Microsoft research in 2016, utilizes residual learning through shortcut connections to bypass layers, solving the vanishing gradient problem, and enabling deeper networks training with high accuracy. Its deep architecture excels at extracting rich, multi-scale features.²¹ EfficientNetB0 (approximately 5 million parameters), proposed by Google brain in 2019, employs a compound scaling method that uniformly scales the network's depth, width, and resolution, achieving a superior balance between accuracy and computational efficiency.²² Finally, MobileNetV3 (approximately 5 million parameters), introduced by Google AI in 2019, is a lightweight and efficient model designed for mobile devices, making it suitable for applications with limited computational resources.²³ Multiple studies reported

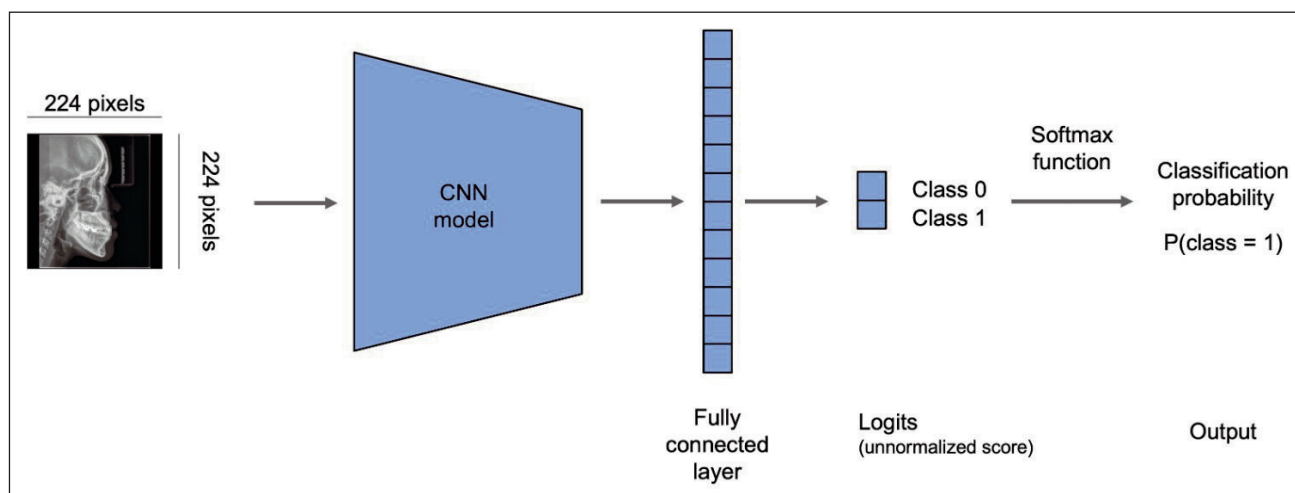


Figure 3 Schematic overview of the CNN-based binary classification workflow

using CNN models to screen patients with OSA or with potential airway obstruction from lateral cephalometric radiographs through image classification with distinct accuracy.^{18,19,24} The overall workflow of CNN models are demonstrated in Figure 3.

The binary cross-entropy loss function was used for the training process. However, as samples classified as long MP-H had relatively low number compared to group with short MP-H, contributed to an imbalanced dataset. Therefore, we further investigated the strategy to deal with image classification task with imbalanced dataset. In this study, we labelled the short MP-H group as class 0 and the long MP-H group as class 1, with class 1 being the minority class in our dataset. Normally, binary cross entropy (CE) loss can be calculated following formula:

$$\mathcal{L}_{CE} = -[y \log(p) + (1-y) \log(1-p)]$$

Where;

y is defined as the true label class which is 0 or 1 representing class 0 or 1, respectively.

p is defined as the predicted probability for class 1, which value ranging from 0 to 1 ($0 < P < 1$).

As we are dealing with the class imbalanced data due to low prevalence of OSA; therefore, we employed the weighted binary cross entropy (WCE) loss in the experiment which can be described as follow:²⁵

$$\mathcal{L}_{WCE} = -[w_1 y \log(p) + w_0 (1-y) \log(1-p)]$$

Where;

w_1 = weight for class 1

w_0 = weight for class 0

As the weight for class 1 and class 0 could be determined differently, we assigned more weight to the minority class. This resulted in more penalties in the training process, when the models predicted the minority class incorrectly. To mitigate class imbalance commonly observed in disorders with low prevalence, we explore the ratio for w_0 and w_1 for training the model with class imbalanced dataset. The class-weight ratios applied for w_0 and w_1 were [1, 5], [1, 4], [1, 3], [1, 2], and [1, 1]. This approach aimed to determine appropriate penalization for incorrectly screening long MP-H cases.

The results from the classification models can be defined as follows:

True Positive (TP): Cases that truly belonged to the long MP-H group and were correctly classified by the model as long MP-H.

True Negative (TN): Cases that truly belonged to the short MP-H group and were correctly classified by the model as short MP-H.

False Positive (FP): Cases that truly belonged to the short MP-H group but are incorrectly classified by the model as long MP-H.

False Negative (FN): Cases that truly belong to the long MP-H group but are incorrectly classified by the model as short MP-H.

A confusion matrix was generated for each model to visualize class-level performance, showing the distribution of true and false classifications between the short and long MP-H groups.

Performance metrics included accuracy, precision, sensitivity (recall), specificity, F1-score, area under the receiver operating characteristic curve (AUROC) and area under the precision-recall curve (AUPRC). To clarify, accuracy is the proportion of correct predictions (both true positives and true negatives) among the total number of cases. Precision is the proportion of positive predictions that are correct. Sensitivity is the proportion of actual positives that are correctly identified. All metrics were calculated based on TP, TN, FP and FN. Moreover, F1 Score is the harmonic mean of precision and recall, providing a single metric that balances the two. Furthermore, AUROC is a measurement for classification models at various threshold levels. The curves were plotted using Matplotlib library, a visualization utility in Python. AUROC represented the overall ability of the model to differentiate between positive and negative classes. The model was interpreted as follows:²⁶

AUROC = 0.50: The model performance is equivalent to random choosing.

$0.50 < \text{AUROC} < 0.70$: The model had a poor performance.

$0.70 \leq \text{AUROC} < 0.80$ is the model with acceptable performance.

$0.80 \leq \text{AUROC} < 0.90$ is the model with excellent performance.

$\text{AUROC} \geq 0.90$ is the model with distinguished performance.

AUPRC serves as a key metric for assessing the performance of classification models, particularly in situations involving imbalanced datasets. The precision-recall curve represents precision on the y-axis and recall on the x-axis across various decision thresholds.

AUPRC provided comprehensive summary of the model's ability to sustain both high precision and recall over the range of thresholds.²⁷ Each model's performance was assessed under all five loss weighting scenarios.

To evaluate the robustness and reliability of our performance metrics, the stratified bootstrap was used to estimate 95 % confidence intervals (CIs) for each model by 1000 iterations. Conduction of bootstrap iterations by randomly sampling with replacement from the set was done. For each resampled dataset, all evaluation metrics were calculated. The 95 % confidence intervals were determined to offer a quantitative assessment of the uncertainty.

Results

The human reliability test of intraexaminer reliability (ICC = 0.97) and interexaminer reliability (ICC = 0.99) of MP-H distance were found to be high. The intraclass correlation coefficient (ICC) value above 0.90 indicated an excellent reliability.²⁸

Of the 304 patients, 210 (69.08 %) were female, and 94 (30.92 %) were male. The mean age was 26.67 ± 7.80 (Min = 18, Max = 56) years. Patients had an average ANB angle of $3.45^\circ \pm 3.17^\circ$ (Min = -6.44° , Max = 12.42°) and SN-GoMe angle of $34.31^\circ \pm 6.31^\circ$ (Min = 20.83° , Max = 59.41°). The mean distance of MP-H was 12.63 ± 5.36 mm (Min = 0.61, Max = 27.62).

Kolmogorov-Smirnov and Shapiro-Wilk tests revealed a normal distribution in SNA, SNB, ANB angle, SN-GoMe angle and MP-H distance among training, validating, and test set ($P > 0.05$). Hence, the data were properly stratified allocation among three datasets. (Table 1).

Confusion matrices of each model under all weighting scenarios are shown in Figure 4, illustrating the class-level prediction outcomes between the short and long MP-H groups. Figure 5 summarize the classification performance of DenseNet-121, ResNet-50, EfficientNet-B0, and MobileNetV3 under

Table 1 Demographic data of all dataset

	Overall dataset (n = 304)	Training dataset (n = 212)	Validating dataset (n = 46)	Test dataset (n = 46)
Short MP-H (n, %)	253 (83.22 %)	177 (83.49 %)	38 (82.61 %)	38 (82.61 %)
Long MP-H (n, %)	51 (16.78 %)	35 (16.51 %)	8 (17.39 %)	8 (17.39 %)
Age (mean $\pm$ SD, years)	26.67 $\pm$ 7.80	26.22 $\pm$ 7.27	28.65 $\pm$ 9.10	26.76 $\pm$ 8.60
Gender				
Female (n, %)	210 (69.08 %)	149 (70.28 %)	31 (67.39 %)	30 (65.22 %)
Male (n, %)	94 (30.92 %)	63 (29.72 %)	15 (32.61 %)	16 (34.78 %)
SNA (mean $^{\circ}$ $\pm$ SD)	85.21 $\pm$ 3.57	85.20 $\pm$ 3.42	85.43 $\pm$ 3.89	85.09 $\pm$ 3.96
SNB (mean $^{\circ}$ $\pm$ SD)	81.76 $\pm$ 4.32	81.89 $\pm$ 4.29	81.63 $\pm$ 4.44	81.34 $\pm$ 4.42
ANB (mean $^{\circ}$ $\pm$ SD)	3.45 $\pm$ 3.17	3.31 $\pm$ 3.38	3.81 $\pm$ 2.55	3.75 $\pm$ 2.67
SN-GoMe (mean $^{\circ}$ $\pm$ SD)	34.31 $\pm$ 6.31	34.04 $\pm$ 6.36	34.58 $\pm$ 6.84	35.27 $\pm$ 5.54
MP-H Distance (mean $\pm$ SD, mm)	12.63 $\pm$ 5.36	12.68 $\pm$ 5.38	12.50 $\pm$ 4.96	12.51 $\pm$ 5.77

MP-H: Mandibular-plane-to-hyoid bone

SD: Standard deviation

Min: Minimum

Max: Maximum

five class-weight configurations, based on bootstrap-estimated evaluation metrics. When comparing model performance, AUROC served as the primary metric, as it provides a measure of overall discriminative ability. Across all architectures, AUROC values ranged from 0.41 to 0.96. The highest AUROC was obtained with ResNet-50 (weight 4; AUROC = 0.96, 95 % CI 0.88 – 1.00), which also achieved perfect specificity (1.00) and precision (1.00). DenseNet-121 maintained consistently high performance across weight settings, with AUROC values $\geq$ 0.89. EfficientNet-B0 achieved its highest AUROC with weight 4 (0.93, 95 % CI 0.83 – 0.99). MobileNetV3 exhibited substantial variability, with its highest AUROC at weight 5 (0.94, 95 % CI 0.86 – 0.99). Based on the AUROC metric, DenseNet-121 and ResNet-50 achieved excellent to distinguished performance, whereas EfficientNet-B0 showed acceptable to excellent performance. MobileNetV3,

however, exhibited performance ranging from poor to distinguished across weight settings.

However, because the study sample was imbalanced, with comparatively fewer cases of long MP-H, AUPRC was additionally considered, as it better reflects discriminative performance under class imbalance and is less sensitive to the predominance of negative cases. Among all models, AUPRC values ranged from 0.19 to 0.90. With ResNet-50, weight = 4 had the highest AUPRC of 0.90 (95 % CI 0.73 – 1.00) followed by Densenet-121, weight = 3 with AUPRC of 0.88 (95 % CI 0.67 – 1.00). Both maintained an excellent performance. Notably, DenseNet-121 demonstrated consistently stable performance across all weighting scenarios. The model achieved AUROC values ranging from 0.89 to 0.92 and AUPRC values from 0.83 to 0.88.

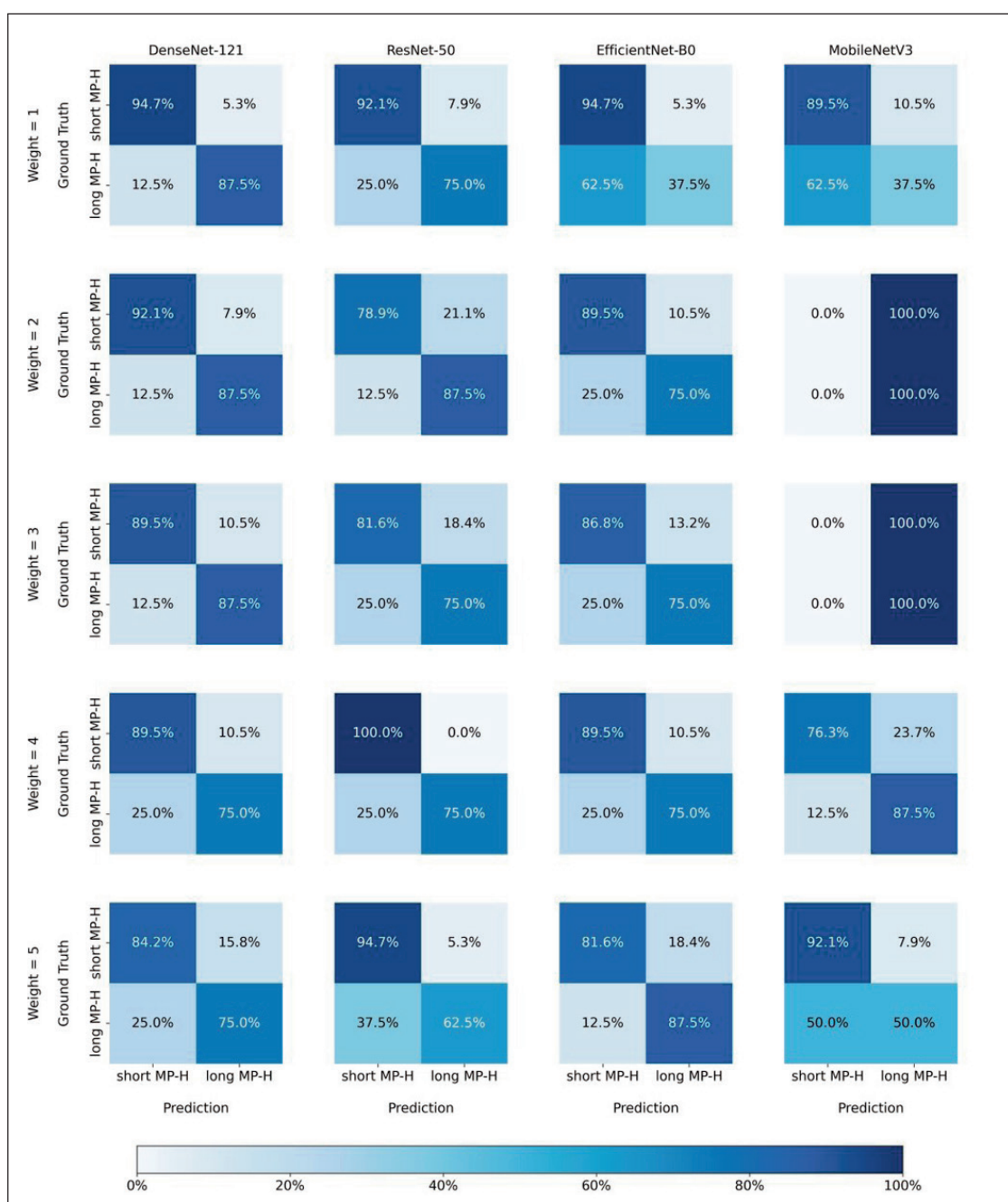


Figure 4 Confusion matrix of DenseNet-121, ResNet-50, EfficientNet-B0, and MobileNetV3 with varying weights (from 1 to 5). ChatGPT 5.5 (OpenAI) was used to assist with the development of the Python code for development of the figure.

Discussion

Assessment of the risk for OSA is considered essential for formulating an effective treatment plan, as undiagnosed or underestimated cases may lead to inadequate management of the condition.³ The study could lead to a development of a beneficial screening tool for dentists to detect possible risk. This study demonstrated the feasibility of employing CNN

models to classify patients into short or long MP-H based on the measurement of the MP-H distance in lateral cephalometric radiographs. The proposed approach demonstrates the potential of AI-assisted imaging analysis as an automated screening tool for OSA risk assessment adjunct to questionnaire, offering advantages over conventional manual measurements, which are often time-consuming and susceptible to

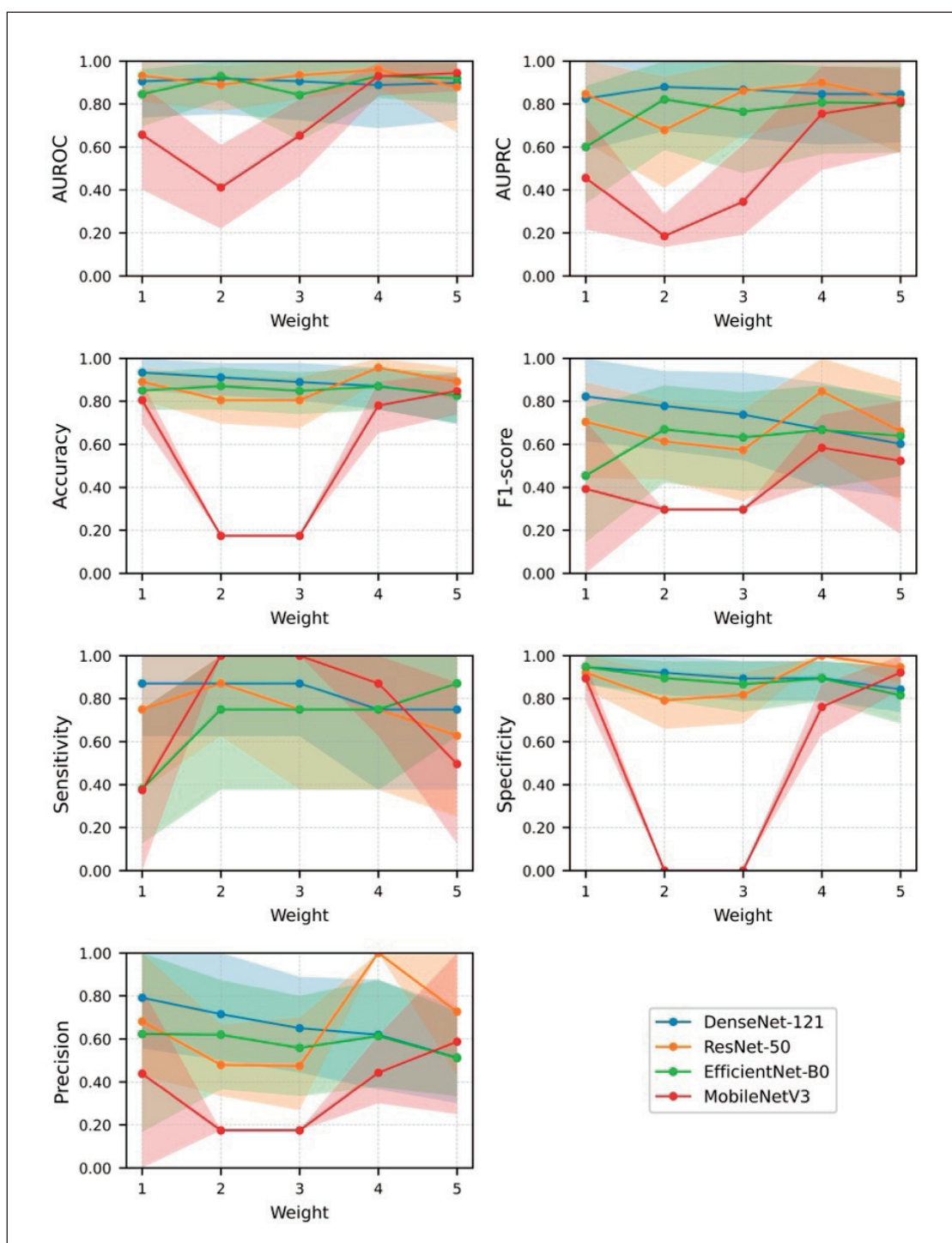


Figure 5 Model performance of DenseNet-121, ResNet-50, EfficientNet-B0, and MobileNetV3 across weight settings (w = 1 – 5). Mean metric values are shown with shaded 95 % confidence intervals for AUROC, AUPRC, accuracy, F1-score, sensitivity, specificity, and precision. ChatGPT 5.5 (OpenAI) was used to assist with the development of the Python code for development of the figure.

inter-examiner variability.

From our study, DenseNet-121 recorded high F1 score, sensitivity, specificity, accuracy, AUROC, and AUPRC, followed by ResNet-50, EfficientNet-B0, and

MobileNetV3. To clarify, DenseNet-121(weight = 1) achieving high F1 score (0.82), sensitivity (0.87), and specificity (0.95), along with acceptable precision (0.79). To develop a radiographic screening tool for identifying

long MP-H patients, minimizing false negatives is important to avoid mis-screened long MP-H groups. DenseNet-121 achieved high sensitivity (0.87) and high specificity (0.95). This suited a medical screening tool, as it minimized missed long MP-H group as most true cases were accurately detected, while also keeping false positives low to avoid unnecessary referral. It also maintained a high F1 score, reflecting a balanced trade-off between sensitivity and precision, which was essential to avoid alarming healthy individuals unnecessarily. Also, ResNet-50 also showed strong performance but had slightly lower sensitivity (0.75). In contrast, EfficientNet-B0, and MobileNetV3 had lower F1 scores and precision, making them less reliable for clinical use. Conversely, MobileNetV3 models demonstrated suboptimal performance in clinical screening contexts, exhibiting low precision and specificity. When the penalty weight was increased from 1 to 3, the model achieved a sensitivity of 1 but a specificity of 0. Increasing the loss function weight for the minority class (long MP-H) caused the model to prioritise minimising false negatives, ultimately resulting in all cases being classified as long MP-H. In terms of AUPRC, DenseNet-121 achieved superior overall performance in classifying MP-H within an imbalanced dataset, where short MP-H outnumbered long MP-H cases. AUPRC and AUROC of DenseNet-121, ResNet-50, EfficientNet-B0, and MobileNetV3 were shown (Figure 2, and 3). With imbalanced nature of data, AUPRC was assessed.²⁷ All DenseNet-121 revealed AUPRC of more than 0.80 which indicated a strong performance. Although, ResNet-50 ($w = 4$) showed highest AUPRC (0.90). However, the model demonstrated a sensitivity of 0.75, indicating that around 25 % of long MP-H cases were not correctly identified during screening. Overall, DenseNet-121 ($weight = 1$) offered the best trade-off between sensitivity and precision, aligning well with the clinical goals of early detection of long MP-H. As the clinical goal is to prioritize screening of long MP-H cases, the high sensitivity is considered. Clinically, DenseNet-121

may be useful for large-scale screening in orthodontic settings, extending beyond traditional cephalometric analysis. The extensive cephalometric databases can be rapidly assessed to identify individuals with long MP-H, facilitating population-based risk assessment, epidemiological investigations, and early detection of patients who may require additional airway evaluation. However, it was never intended to be used as an independent tool, since human validation, clinical examination, and consideration of other risk factors remain necessary.

Previously, CNN model based on ResNet backbone was developed to automatically detect upper airway obstruction in lateral cephalograms based on upper pharyngeal width and lower pharyngeal width cut off value from McNamara's analysis.¹⁸ The developed model possessed a high performance with a sensitivity of 0.86, specificity of 0.89, precision of 0.90, negative predictive value (NPV) of 0.85, and F1-score of 0.88, respectively.¹⁸ Also, researchers utilised DenseNet-121 models to classify patients with OSA and normal patients using lateral cephalometric radiographs labelled according to polysomnography results. The model achieved an accuracy of 0.74. Sensitivity, specificity, F1 score, and AUROC were recorded as 0.77, 0.73, 0.87, and 0.82, respectively. The precision was 0.71, while the NPV was 0.79.¹⁹ In terms of performance, our study found that DenseNet-121 (AUPRC = 0.83-0.88, sensitivity = 0.75-0.87) demonstrated higher performance than ResNet (AUPRC = 0.68-0.90, sensitivity = 0.63-0.87); however, both could be utilized for image classification to short and long MP-H. Since, the polysomnography was not incorporated into the study, we suggested using as screening tool.

Commonly, medical data is imbalanced as the normal population outweighed the number of diseased populations. Multiple strategies are proposed to deal with class imbalanced data including image augmentation and Generative adversarial

networks (GANs) and weighted cross entropy loss. Image augmentation made a slight change (rotation, translation, and flipping) to the original data to allow the models to learn from more diverse data.²⁹ GANs are an advanced machine learning approach used to create new synthetic data that mimics real data, making it difficult to differentiate them apart.³⁰ While, weighted cross entropy loss is a variation of the standard loss function used in model training, where additional weight is given to the minority class to address class imbalance.²⁵ However, the results showed that weights varied across different models. It was therefore important to identify appropriate weights for specific problems. There was no universally applicable solution.

The limitations of this study include the relatively small sample size of lateral cephalometric radiographs, all of which were obtained from a single source, as well as the retrospective nature of the research design. The information on BMI, additional health conditions, and smoking status was unavailable. Nevertheless, all patients were asymptomatic for OSA and undiagnosed. These constraint affects the generalizability of the findings. Future research should include larger datasets collected from multiple centers. Hence, prospective cohort evaluations are recommended to further assess model performance. Inclusion of OSA patients with multiple risks ranging from mild to severe OSA with incorporation of polysomnography for validation and comparison of how models predicted OSA risk from MP-H distance and their performance are also advised. Also, a development of a multimodal, multifactorial screening tools could also be advantageous for screening of other risk factors.

Conclusion

The proposed deep learning-based system shows potential for future clinical implementation in assessing long MP-H distance from lateral cephalometric

radiographs. Among the evaluated configurations, DenseNet-121 with weight = 1 provided the most favorable balance between sensitivity and precision, aligning well with the clinical objectives of screening patients with a long MP-H distance. The effects of weighted loss functions varied across architectures, whereas DenseNet-121 consistently demonstrated high performance across all weighting scenarios, indicating that backbone selection plays a more influential role than class weighting in this task.

Author contributions

Nichakorn Songsang: Software, Formal analysis, Data curation, Visualization, Writing- Original draft preparation, Investigation, Resources, Validation, Writing- Reviewing and Editing, Project administration, Funding acquisition; **Udom Thongudomporn:** Conceptualization, Methodology, Supervision, Investigation, Resources, Validation, Writing- Reviewing and Editing, Project administration, Funding acquisition; **Witsarut Upalananda:** Conceptualization, Methodology, Supervision, Software, Formal analysis, Data curation, Visualization.

Ethical statement

This study was approved by the Human Research Ethics Committee, Faculty of dentistry, Prince of Songkla University (EC6712-053). All data were anonymized to protect participants confidentiality.

Disclosure statement

The authors have no conflicts of interest.

Funding

This research was supported by Graduate school, Faculty of Dentistry, Prince of Songkla university.

AI-assisted writing declaration

The authors used Microsoft Copilot for language editing and ChatGPT (OpenAI, GPT-5.5) for assistance with Python code development for Figures 4 and 5. All content was reviewed by the authors, who take full responsibility for the final manuscript.

Acknowledgments

None.

References

- Spicuzza L, Caruso D, Di Maria G. Obstructive sleep apnoea syndrome and its management. *Ther Adv Chronic Dis* 2015;6(5):273-85.
- Neruntarat C, Chantapant S. Prevalence of sleep apnea in HRH Princess Maha Chakri Srinthorn Medical Center, Thailand. *Sleep Breath* 2011;15(4):641-8.
- Alansari RA. The role of orthodontics in management of obstructive sleep apnea. *Saudi Dent J* 2022;34(3):194-201.
- Ryu HH, Kim CH, Cheon SM, Bae WY, Kim SH, Koo SK, et al. The usefulness of cephalometric measurement as a diagnostic tool for obstructive sleep apnea syndrome: a retrospective study. *Oral Surg Oral Med Oral Pathol Oral Radiol* 2015;119(1):20-31.
- deBerry-Borowiecki B, Kukwa A, Blanks RH. Cephalometric analysis for diagnosis and treatment of obstructive sleep apnea. *Laryngoscope* 1988;98(2):226-34.
- Neelapu BC, Kharbanda OP, Sardana HK, Balachandran R, Sardana V, Kapoor P, et al. Craniofacial and upper airway morphology in adult obstructive sleep apnea patients: a systematic review and meta-analysis of cephalometric studies. *Sleep Med Rev* 2017;31:79-90.
- Auvenshine Rc Dds P, Pettit Nj Dmd MSD. The hyoid bone: an overview. *Cranio* 2020;38(1):6-14.
- Bailey EF. Activities of human genioglossus motor units. *Respir Physiol Neurobiol* 2011;179(1):14-22.
- Banhiran W, Wanichakorntrakul P, Metheetrairut C, Chiewwit P, Planuphap W. Lateral cephalometric analysis and the risks of moderate to severe obstructive sleep-disordered breathing in Thai patients. *Sleep Breath* 2013;17(4):1249-55.
- Jo JH, Park JW, Jang JH, Chung JW. Hyoid bone position as an indicator of severe obstructive sleep apnea. *BMC Pulm Med* 2022;22(1):349.
- Kazmouz S, Calzadilla N, Choudhary A, McGinn LS, Seaman A, Purnell CA. Radiographic findings predictive of obstructive sleep apnea in adults: a systematic review and meta-analysis. *J Craniomaxillofac Surg* 2025;53(2):162-80.
- Riley R, Guilleminault C, Herran J, Powell N. Cephalometric analyses and flow-volume loops in obstructive sleep apnea patients. *Sleep* 1983;6(4):303-11.
- Ahmadi K, Amali A, Saedi B, Dasdar S, Rashedi S, Kianfar N, et al. Evaluation of cephalometric indices in patients with obstructive sleep apnea in comparison with healthy individuals. *Adv Oral Maxillofac Surg* 2022;5:100250.
- Andersson L, Brattström V. Cephalometric analysis of permanently snoring patients with and without obstructive sleep apnea syndrome. *Int J Oral Maxillofac Surg* 1991; 20(3):159-62.
- Yoon HJ, Kim DR, Gwon E, Kim N, Baek SH, Ahn HW, et al. Fully automated identification of cephalometric landmarks for upper airway assessment using cascaded convolutional neural networks. *Eur J Orthod* 2022;44(1): 66-77.
- Manakitsa N, Maraslidis GS, Moysis L, Fragulis GF. A review of machine learning and deep learning for object detection, semantic segmentation, and human action recognition in machine and robotic vision. *Technologies* 2024;12(2):15.
- Russell BC, Torralba A, Murphy KP, Freeman WT. LabelMe: a database and web-based tool for image annotation. *Int J Comput Vis* 2008;77(1):157-73.
- Jeong Y, Nang Y, Zhao Z. Automated evaluation of upper airway obstruction based on deep learning. *Biomed Res Int* 2023;2023:8231425.
- Jeong HG, Kim T, Hong JE, Kim HJ, Yun SY, Kim S, et al. Automated deep neural network analysis of lateral cephalogram data can aid in detecting obstructive sleep apnea. *J Clin Sleep Med* 2023;19(2):327-37.
- Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; 2017 Jul 21-26; Honolulu, HI, USA. p. 4700-8.
- He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; 2016 Jun 27-30; Las Vegas, NV, USA. p. 770-8.
- Tan M, Le QV. EfficientNet: rethinking model scaling for convolutional neural networks [Internet]. *arXiv*; 2019

- [cited 2025 Aug 7]. Available from: <https://arxiv.org/abs/1905.11946>.
23. Howard A, Sandler M, Chu G, Chen L-C, Chen B, Tan M, et al. Searching for MobileNetV3. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV); 2019 Oct 27 - Nov 2; Seoul, South Korea. p. 1314-24.
 24. Tsuiji S, Nagaoka T, Fukuda T, Sakamoto Y, Almeida FR, Nakayama H, et al. Machine learning for image-based detection of patients with obstructive sleep apnea: an exploratory study. *Sleep Breath* 2021;25(4):2297-305.
 25. Özdemir Ö, Sönmez EB. Weighted cross-entropy for unbalanced data with application on COVID X-ray images. In: 2020 Innovations in Intelligent Systems and Applications Conference (ASYU); 2020 Oct 15-17; Istanbul, Turkey. p. 1-5.
 26. de Hond AAH, Steyerberg EW, van Calster B. Interpreting area under the receiver operating characteristic curve. *Lancet Digit Health* 2022;4(12):e853-5.
 27. Ozenne B, Subtil F, Maucourt-Boulch D. The precision-recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. *J Clin Epidemiol* 2015;68(8):855-9.
 28. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med* 2016;15(2):155-63.
 29. Goceri E. Medical image data augmentation: techniques, comparisons and interpretations. *Artif Intell Rev* 2023;1-45.
 30. Goyal M, Mahmoud QH. A systematic review of synthetic data generation techniques using generative AI. *Electronics* 2024;13(17):3509.