Article

Surgically Oriented Ultrasound-Based AI of Median Nerve Morphology as a Decision Support for Carpal Tunnel Release: A Calibrated ConvNeXt-CBAM Framework With Supervised Contrastive Warm-Up

Ann. Ital. Chir., 2026 97, 1: 74–83
https://doi.org/10.62713/aic.4332

Keyue Yang[1], Xinkun He[1], Pingping Ye[2], Yao Qian[1], Xiaofeng Teng[1], Xueyuan Li[1], Xin Wang[1], Hong Chen[1]

[1]Department of Hand Microsurgery and Plastic Reconstructive Surgery, The Sixth Hospital of Ningbo, 315010 Ningbo, Zhejiang, China
[2]Department of Special Examinations, The Sixth Hospital of Ningbo, 315010 Ningbo, Zhejiang, China

AIM: Carpal tunnel syndrome (CTS) is a compressive neuropathy commonly encountered in hand surgery, and decisions regarding whether operative decompression should be conducted rely on accuracy and clinical interpretability of imaging findings. Ultrasound is commonly used to directly visualize the median nerve in this condition, but remains operator-dependent and is limited by inconsistent diagnostic thresholds. This study aimed to develop a surgically oriented ultrasound-based artificial intelligence (AI) model that provides strong discriminative power, reliable probability calibration for preoperative counseling, and anatomy-aware explainability aligned with the nerve targeted in carpal tunnel release.

METHODS: In this retrospective study, adults with suspected CTS, who had ultrasound image of adequate quality and met criteria for the clinical standard, were included; cases with prior carpal tunnel surgery or non-diagnostic images were excluded. We analyzed 2900 wrist ultrasound examinations, reserving an a priori 20% test set (n = 580 images) at the patient level. DenseNet-121 was utilized as the reference baseline. The proposed model used ConvNeXt-T augmented with Convolutional Block Attention Modules (CBAM), optimized with a supervised contrastive warm-up before standard fine-tuning; probabilities were post-hoc calibrated by temperature scaling. Images underwent de-identification, normalization, and ultrasound-appropriate augmentation. Primary outcome was discrimination (receiver operating characteristic (ROC), area under the curve (AUC), average precision (AP)) with bootstrap bands; secondary outcomes included accuracy, precision, recall, F1 score (harmonic mean of precision and recall), confusion matrices, probability distributions, class-conditional score separation, and Grad-CAM++ agreement with expert-defined regions of interest. Test set labels used a fixed 0.5 threshold.

RESULTS: The proposed model outperformed the baseline in terms of discrimination (AUC 0.904 vs 0.821; AP 0.907 vs 0.831). Aggregate metrics also favored the proposed approach (accuracy 0.91 vs 0.83; precision 0.89 vs 0.81; recall 0.88 vs 0.82; F1 score 0.88 vs 0.81). Confusion matrices showed concurrent reductions in false positives (58→33, –43%) and false negatives (52→35, –33%): baseline true-negatives (TN)/false-positives (FP)/false-negatives (FN)/true-positives (TP) = 232/58/52/238; proposed TN/FP/FN/TP = 257/33/35/255. Predicted-probability histograms and class-conditional densities indicated more confident, better-separated outputs with calibration. Grad-CAM++ overlays were more compact and nerve-concordant relative to expert contours, supporting anatomy-aligned interpretability for surgical planning.

CONCLUSIONS: A calibrated, explainable ConvNeXt-CBAM ultrasound classifier delivers reliable probabilities and anatomically faithful saliency that are directly actionable for surgical triage, timing of decompression, and preoperative counseling in CTS. These findings support ultrasound-based AI as a practical adjunct to clinical assessment and nerve conduction studies in hand surgery, warranting prospective, multicenter validation and workflow integration.

Keywords: carpal tunnel syndrome; hand surgery; median nerve; artificial intelligence; deep learning; ConvNeXt

Introduction

Carpal tunnel syndrome (CTS) is one of the most common compressive neuropathies encountered in surgical practice, with the rising global incidence rates highlighting its burden on hand surgery workflows [1]. This condition is the primary indication for operative decompression via open or endoscopic carpal tunnel release [2]. Decisions to operate, the timing of intervention, and patient counseling all hinge on an accurate and clinically interpretable assessment of disease probability and severity, ideally supported by metrics with confidence intervals to quantify uncertainty. Although nerve conduction study is frequently used, such a detection approach is costly, time-consuming, uncomfortable for patients, and limited in terms of anatomic context

Submitted: 2 September 2025 Revised: 9 October 2025 Accepted: 15 October 2025 Published: 10 January 2026
Correspondence to: Hong Chen, Department of Hand Microsurgery and Plastic Reconstructive Surgery, The Sixth Hospital of Ningbo, 315010 Ningbo, Zhejiang, China (e-mail: chenhong_6612@126.com).
Editor: Hui Lu

for surgical planning [3]. High-frequency ultrasound, by contrast, is a bedside modality that directly depicts the median nerve and flexor retinaculum [4], supports documentation of cross-sectional area (CSA), flattening ratio, wrist-to-forearm CSA ratio, echotexture, and bowing of the retinaculum, and is increasingly integrated into preoperative surgical workflows. However, interpretation of its findings remains operator-dependent; thresholds vary across devices and institutions; and surgically relevant image features (e.g., focal enlargement and its relation to the retinaculum) are not consistently quantified, limiting reproducibility and reducing confidence in sonography as a sole basis for surgical decision-making [5].

Artificial intelligence (AI) offers a means to standardization of ultrasound interpretation for surgical decision-making [6], but its clinical utility depends on two major requirements, beyond raw discriminative performance. First, probability calibration must be reliable so that predicted risks correspond to observed outcomes. Surgeons depend on well-calibrated probabilities to determine operative candidacy, prioritize higher-risk cases, and communicate expected benefits, potentially reducing unnecessary procedures as demonstrated through decision-curve analysis [7]. Second, interpretation should be anatomically faithful, allowing the surgeon to confirm that the algorithm's attention is centered on the median nerve rather than on the peri-lesional tissues that do not require decompression [8]. Much of the existing literature about imaging emphasizes classification accuracy using conventional backbones and standard saliency methods, with limited attention to calibration and to rigorous alignment between model attention and expert-defined nerve regions—two gaps that compromise the clinical utility of AI for preoperative planning [9].

To address these gaps, we designed a surgically oriented ultrasound-based AI framework that couples modern convolutional representation learning with explicit calibration and anatomy-aware explanability, tailored to guiding decision-making for carpal tunnel surgery [10]. We established DenseNet-121 as a transparent baseline and developed a proposed model based on ConvNeXt-T augmented with Convolutional Block Attention Modules (CBAM) [11]. Training involved a supervised contrastive warm-up to promote within-class clustering and between-class separation, followed by fine-tuning of a binary classifier [12]. To deliver probabilities that are usable for guiding surgical decision-making, we applied temperature scaling as a post-hoc calibration step [13]. To ensure anatomical interpretability, we employed Grad-CAM++ and compared saliency against expert-defined regions of interest (ROIs) for the median nerve, emphasizing concordance that can be visually audited by surgeons prior to surgical intervention [14].

The study is designed around surgically relevant endpoints, with potential for multicenter validation to enhance generalizability. Discriminative performance was assessed by using receiver operating characteristic (ROC) analysis and in terms of average precision with 95% confidence intervals; threshold-dependent behavior was evaluated using accuracy, precision, recall, F1 score (harmonic mean of precision and recall), and confusion matrices to explicitly illustrate the trade-off between missed cases and unnecessary operations; calibration was assessed using reliability analysis (including expected calibration error and Brier score) and probability distributions; and anatomy-aware explainability was evaluated in terms of the spatial agreement between Grad-CAM++ and expert-defined ROIs. By aligning model development and evaluation with surgical use cases—triage, timing of decompression, and preoperative counseling—this work aims to advance the function of ultrasound-based AI from generic classification toward surgical decision support in CTS, paving the way for National Institutes of Health (NIH)-funded prospective trials linking predictions to postoperative outcomes.

Methods

We conducted a retrospective ultrasound imaging study of adults evaluated for suspected CTS under approval of the Ethics Committee of the Sixth Hospital of Ningbo (approval number is 2024-20(K)) and in accordance with the Declaration of Helsinki. Subjects were deemed eligible based on a definitive clinical reference standard for CTS, determined by experienced clinicians incorporating patient-reported histories (e.g., nocturnal paresthesia persisting >3 months), physical examination findings (e.g., positive Tinel or Phalen signs), and electrodiagnostic findings (e.g., prolonged distal motor latency >4.5 ms or reduced sensory nerve conduction velocity). Discrepancies were adjudicated by two neurologists or hand surgeons, each with over 10 years of experience, until consensus was reached. Exclusion criteria included a history of prior carpal tunnel surgery, traumatic wrist deformity, or artifacts that compromise visualization of the median nerve. To align with standard surgical practice, images were acquired on high-frequency linear transducers using a preset optimized for superficial nerve imaging, with transducer frequency, depth, dynamic range, and overall gain systematically recorded. The ultrasound devices employed in this study encompassed a range of systems commonly used in clinical settings, with their brands, probe frequencies, and percentage of usage summarized in **Supplementary Table 1**. This distribution reflects institutional variability, which was mitigated through preprocessing techniques such as intensity normalization and speckle noise augmentation to ensure model robustness across devices. Scans were performed at the carpal tunnel inlet (pisiform–scaphoid level) with the forearm supinated and the wrist in neutral or slight extension; where available, a forearm reference level was additionally obtained to enable wrist-to-forearm comparisons. All Digital Imaging and Commu-

Table 1. Baseline demographic and clinical characteristics of the study population.

Characteristic	Value (n = 580)
Total patients	580
Age (years)	Mean: 54.3 ± 7.2
Gender	
Male	157 (27.0%)
Female	423 (73.0%)
Affected side	
Right	290 (50.0%)
Left	218 (37.5%)
Bilateral	72 (12.5%)
Prognosis	
Complete recovery	400 (69.0%)
Partial recovery	151 (26.0%)
No improvement	29 (5.0%)
Electrodiagnostic grade	
Mild	232 (40.0%)
Moderate	261 (45.0%)
Severe	87 (15.0%)

nications in Medicine (DICOM) headers and frames were de-identified prior to analysis.

The dataset comprised ultrasound examinations from 580 unique patients evaluated for suspected CTS, with an average of two original static images per patient (range 1–4, totaling 1160 originals manually selected as transverse views at the carpal tunnel inlet and forearm), expanded via augmentation to 2900 images for robust training. Key demographic and clinical characteristics of the study population, extrapolated from available data, are summarized in Table 1 to enhance transparency on sample composition.The 'prognosis' entry in Table 1 captures post-diagnostic clinical outcomes during follow-up, classified as: complete recovery, marked by full symptom resolution (e.g., no paresthesia, baseline grip strength, negative provocation tests) and return to prior functional status; partial recovery, with mild residual symptoms (e.g., occasional paresthesia) but enhanced function enabling daily tasks without constraints; and no improvement, where symptoms persist or worsen despite conservative measures.

An independent test set comprising 20% of the cohort (n = 580 images) was reserved a priori at the patient level to prevent cross-subject leakage; when both wrists were available for a given patient, both sides were assigned to the same split. The remaining 2320 studies formed the development partition (training/validation) for model building and hyperparameter selection, with validation subsets sampled at the patient level using class stratification. To enhance robustness in this resource-limited setting, online augmentations including random rotations (±15°), translations (±10% of dimensions), and elastic deformations (sigma = 5 pixels) were applied during training. These preserved key diagnostic features, such as median nerve cross-sectional area, while simulating clinical variations (e.g., probe angulation), resulted in an effective ~1.2-fold increase in training diversity, as validated by narrower 95% bootstrap confidence intervals on area under the curve (AUC) compared to non-augmented runs. Two fellowship-trained musculoskeletal sonographers delineated the median nerve on a subset of images using polygonal or elliptical contours to create ROIs for quality assurance and for subsequent spatial agreement analyses; disagreements were resolved internally until a consensus was reached. These ROIs were not used in training supervision for the classifier (to avoid label leakage) but served as independent ground truth for interpretability evaluation. For model inputs, images were preprocessed to remove embedded rulers and user interface (UI) overlays, converted to grayscale if necessary, resized to a square canvas with aspect-ratio–preserving padding, and intensity-normalized. Ultrasound-appropriate data augmentation included ±10–15° rotations, small translations and random crops, mild brightness/contrast/gamma adjustments, limited elastic deformations (sigma = 5 pixels), and low-amplitude speckle injection to mimic device variability, which were applied online during training to expand the effective dataset from 1160 original images to 2900. All thresholds and operating points reported for confusion-matrix analyses were selected without reference to the held-out test set, unless explicitly stated.

A DenseNet-121 classifier with a global average pooling head and a single sigmoid output was trained as a reference baseline. Our proposed model was built on ConvNeXt-T and augmented with CBAM. ConvNeXt-T follows a modernized hierarchical convolutional design with four stages of progressively downsampled feature maps and blocks that combine large-kernel depthwise convolutions (7 × 7), pointwise (1 × 1) channel mixing, gaussian error linear unit (GELU) activations, LayerNorm, residual connections, and stochastic depth; channels generally scale across stages (e.g., 96→192→384→768), with stage-wise depths approximately [3, 3, 9, 3]. To enhance the representation specifically for median nerve morphology, we inserted CBAM in a sequential attention manner after the residual computation of each stage (post-activation, pre-residual addition for the next stage) to keep computational overhead modest while maximizing coverage of multiscale features involved in nerve boundary, fascicular pattern, and retinacular bowing. CBAM applies channel attention followed by spatial attention, as encapsulated in the following expression:

$$M_c(F) = \sigma(\mathrm{MLP}(\mathrm{GAP}(F)) + \mathrm{MLP}(\mathrm{GMP}(F)))$$

where global average pooling (GAP)/global max pooling (GMP) denotes global average/max pooling over spatial dimensions, the two-layer multi-layer perceptron (MLP) shares weights across the two pooled descriptors. Spatial attention forms a two-channel spatial descriptor by concate-

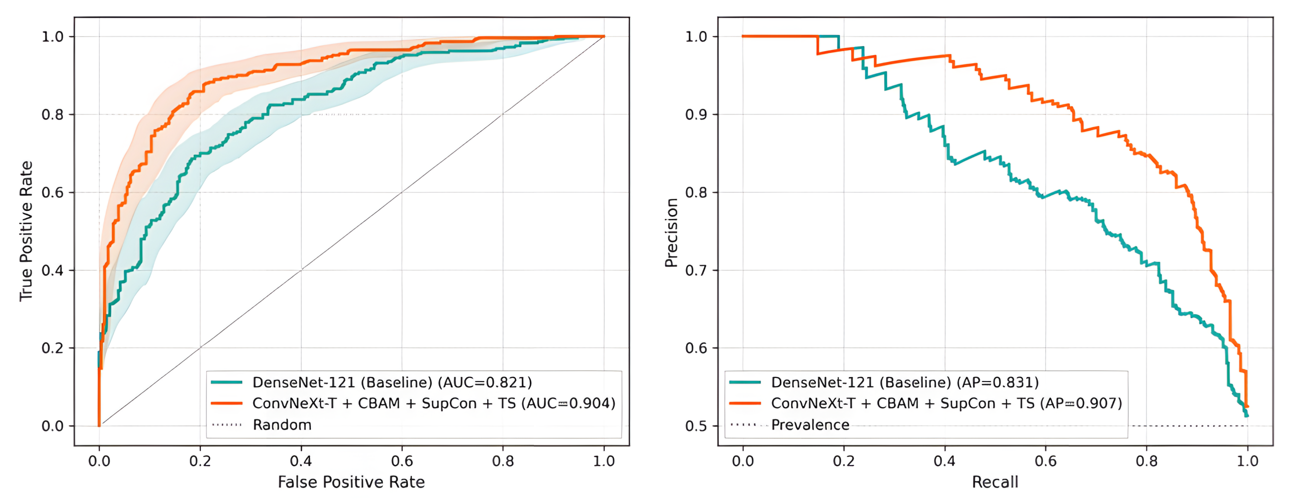

Fig. 1. Discriminative performance on the held-out test set. ROC curves with bootstrap 95% confidence bands (left) and precision–recall curves with prevalence reference (right) for DenseNet-121 (baseline) and ConvNeXt-T + CBAM + SupCon + Temperature scaling. Abbreviations: AUC, area under the curve; ROC, receiver operating characteristic; CBAM, Convolutional Block Attention Modules; AP, average precision; TS, test set.

nating average- and max-pooled maps across channels and filtering with a 7×7 convolution, i.e.,

$$M_s\left(F_c\right)=\sigma\left(\mathrm{Conv}_{7\times 7}\left(\left[\mathrm{AvgPool}_c\left(F_c\right);\mathrm{MaxPool}_c\left(F_c\right)\right]\right)\right)$$

This channel-then-spatial ordering allows the network to first emphasize diagnostically salient feature planes (e.g., fascicular texture channels) and then focus attention onto their relevant loci (e.g., the intraneural contour at the tunnel inlet), which is advantageous for the fine-grained, shape-dependent cues that guide surgical decision-making.

To improve class separation before supervised classification and to reduce reliance on threshold-sensitive training early on, optimization began with a supervised contrastive warm-up. From each input, two stochastic "views" were generated using independent augmentation draws. The supervised contrastive objective is expressed as follows:

$$\mathcal{L}_{\mathrm{SupCon}}=\sum_i \frac{-1}{|P(i)|}\sum_{p\in P(i)}\log\frac{\exp\left(z_i\cdot z_p/\tau\right)}{\sum_{a\neq i}\exp\left(z_i\cdot z_a/\tau\right)}$$

with temperature τ set empirically (e.g., 0.07). This warm-up ran for a fixed number of epochs with the classifier head either detached or lightly supervised. Subsequently, we switched to standard binary cross-entropy on the sigmoid output, removed the projection head, and fine-tuned the entire network end-to-end. Intuitively, the warm-up encourages CTS images to form tight clusters in representation space (despite acquisition/device variability) and pushes non-CTS away. This process stabilizes optimization and reduces overlap around mid-range probabilities during the subsequent discriminative training phase.

Given the centrality of probability calibration to surgical triage, we applied temperature scaling as a post-hoc calibration step on the validation set. This procedure adjusts confidence without changing the ranking, thereby preserving ROC and precision-recall (PR) performance while enhancing reliability for preoperative counseling and shared decision-making.

Both models were trained with AdamW, cosine-annealed learning rates with warm starts, label smoothing during the cross-entropy phase, and weight decay tuned via grid search. Mini-batches were balanced by class, and training employed early stopping based on validation loss. Mixed-precision training was used where available. Unless otherwise noted, class labels on the test set were derived at a fixed threshold of 0.5 to support confusion-matrix analyses; alternative operating points (e.g., high-sensitivity screening vs high-precision preoperative confirmation) were explored by sweeping the threshold on validation curves, and were reported with the corresponding metrics. For interpretability, Grad-CAM++ heatmaps were computed with respect to the last convolutional stage of each network, normalized to [0,1], and overlaid on the source images; spatial agreement with expert-defined ROIs was quantified using intersection-over-union (IoU) without any morphological post-processing.

Results

All analyses were performed on the independent test set, with class labels derived from predicted probabilities using a fixed decision threshold of 0.5. As shown in Fig. 1, the proposed ConvNeXt-T + CBAM model—trained with a supervised contrastive warm-up and calibrated via temperature scaling—demonstrated superior discriminative per-

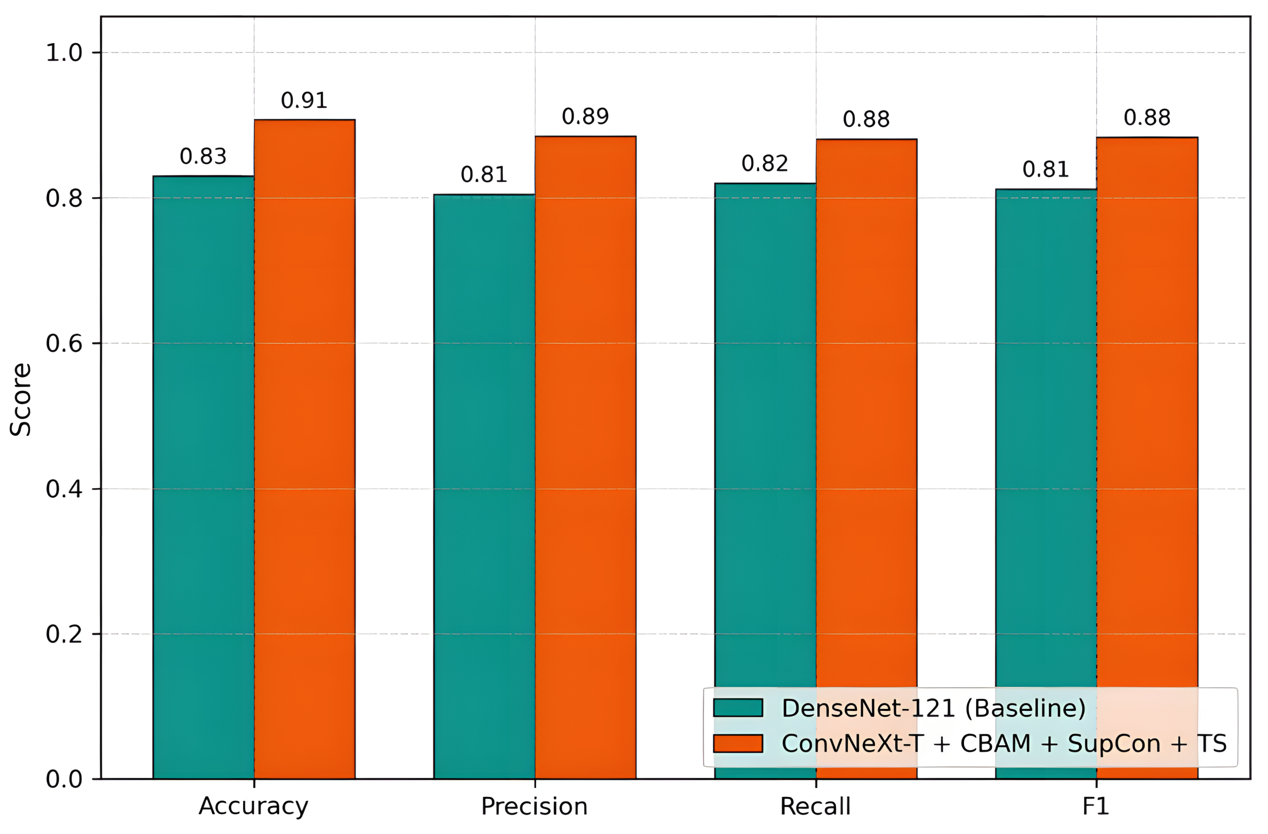

Fig. 2. Comparison of aggregate classification metrics (accuracy, precision, recall, and F1 score (harmonic mean of precision and recall)) between DenseNet-121 (baseline) and the proposed ConvNeXt-T + CBAM + SupCon + Temperature scaling model on the test set.

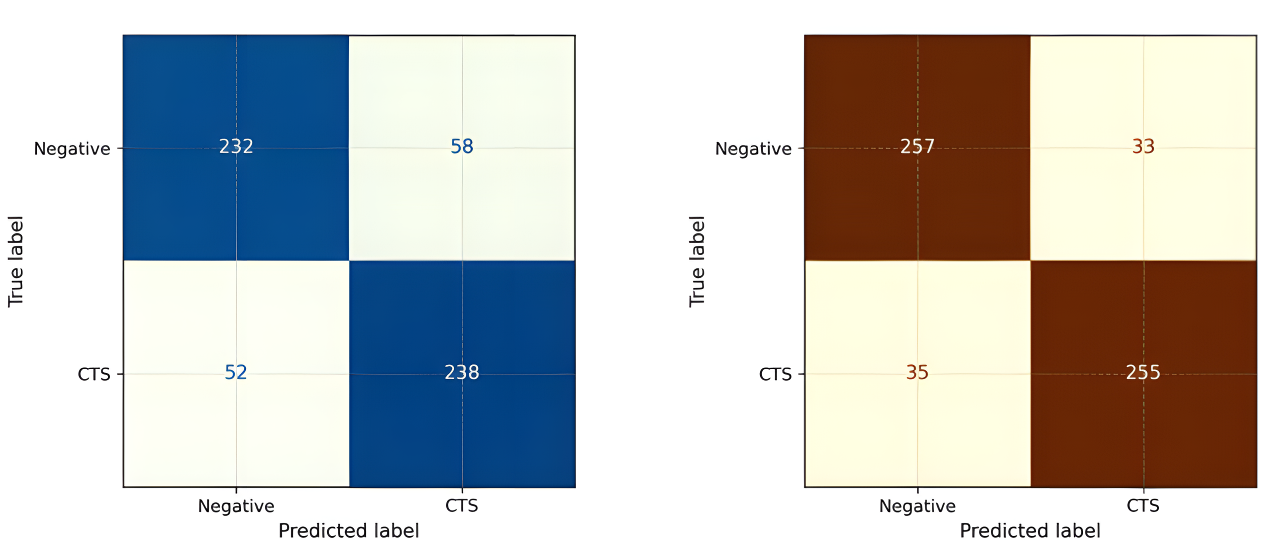

Fig. 3. Confusion matrices at a fixed decision threshold of 0.5 on the test set. (Left) baseline—DenseNet-121; (right) proposed model—ConvNeXt-T + CBAM + SupCon + Temperature scaling. Values in the cells display the observed counts for true-negatives (TN), false-positives (FP), false-negatives (FN), and true-positives (TP). CTS, carpal tunnel syndrome.

formance relative to the reference baseline, DenseNet-121. The ROC curve for the proposed model had an AUC of 0.904, exceeding the baseline AUC of 0.821. Bootstrap 95% confidence bands were uniformly narrower and exhibited minimal overlap with the baseline across a wide range of false-positive rates, indicating greater statistical stability. Precision–recall analysis similarly favored the proposed approach (average precision (AP) 0.907 vs 0.831 for the baseline); the dashed reference line in the PR panel denotes class prevalence in this balanced test cohort. To ensure fair threshold-dependent comparisons, temperature scaling was fitted separately for both models on the validation set, and model-specific decision thresholds selected to maximize F1 score on validation were fixed prior to testing. Discrimination metrics (AUC/AP) were reported with 95% bootstrap CIs and AUC differences were tested using DeLong's method. Calibration was quantified by expected calibration erro (ECE) and Brier score with bootstrap CIs; anatomy-aware explainability was evaluated by computing the mean IoU between Grad-CAM++ overlays and expert-defined ROIs on the annotated subset. Performance metrics across training, validation, and test sets are presented in **Supplementary Table 2**.

Aggregate classification metrics are summarized in Fig. 2. Compared with the baseline, the proposed model achieved higher accuracy (0.91 vs 0.83), precision (0.89 vs 0.81), recall (0.88 vs 0.82), and F1 score (0.88 vs 0.81), consistent with the gains observed in AUC and AP.

Error profiles derived from confusion matrices are presented in Fig. 3. For the baseline, true-negatives (TN) = 232, false-positives (FP) = 58, false-negatives (FN) = 52, true-positives (TP) = 238; for the proposed model, TN = 257, FP = 33, FN = 35, TP = 255. These translate to improvements in sensitivity (0.821→0.879) and specificity (0.800→0.886). In absolute terms, FP decreased from 58 to 33 (–43%) and FN from 52 to 35 (–33%), representing clinically meaningful reductions in both rates and missed CTS diagnoses.

The distribution of predicted probabilities for all test instances shifted toward more confident outputs with the proposed model, accompanied by a modest reduction in variance (Fig. 4), consistent with the observed reductions in both FP and FN. Class-conditional score densities further demonstrate improved separability (Fig. 5): the proposed model shows less overlap between negative and CTS distributions in the mid-probability band (~0.45–0.65), whereas the baseline exhibits broader tails and greater interclass mixing—patterns consistent with its lower precision and recall.

Finally, qualitative explainability results illustrate Grad-CAM++ overlays (color) alongside expert-defined ROIs (white dashed contours) (Fig. 6). The baseline tended to highlight broader peri-lesional regions that frequently extended beyond the expert-annotated median nerve, whereas the proposed model produced compact, well-localized saliency predominantly confined within the expert boundary. This closer spatial concordance between model attention and expert annotations is concordant with the above-mentioned improvements in quantitative performance.

Discussion

This study proposes and evaluates a surgically oriented ultrasound-based AI framework for CTS that integrates a modern backbone (ConvNeXt-T) with CBAM, a supervised contrastive warm-up, and post-hoc temperature scaling [15]. Interpreted through a surgical lens, three features are most critical for guiding surgical decision-making: (i) sufficiently robust discriminative power to support triage; (ii) well-calibrated probabilities suitable for preoperative counseling and threshold selection; and (iii) anatomy-aware explainability that reassures that the model is attending specifically to the median nerve rather than to the non-actionable peri-lesional tissue [16]. With these features, the proposed model outperformed the DenseNet-121 (baseline) and produced Grad-CAM++ overlays that aligned more closely with expert contours, thereby addressing common barriers to deploying AI for preoperative planning and consent [14].

Methodological Innovations and Model Design

From a methodological perspective, the ConvNeXt-T architecture augmented with CBAM is well-suited for analyzing ultrasound images of the median nerve. ConvNeXt's large-kernel depth-wise convolutions and stage-wise hierarchy are effective for capturing multiscale edge, texture, and shape cues, while CBAM's sequential channel-then-spatial gating first accentuates fascicular texture channels and then sharpens attention along the intraneural contour at the tunnel inlet [17]. The supervised contrastive warm-up further improves representation geometry before conventional supervised fine-tuning: by explicitly pulling together embeddings of CTS cases and pushing apart non-CTS, the network enters the classification phase with reduced class overlap, which is reflected in improved precision–recall behavior and a tighter mid-probability band [18]. Temperature scaling, applied on a held-out validation set, then adjusts confidence without altering ranking, yielding probabilities that map more faithfully to observed risk—precisely the quantity surgeons need when articulating the benefits and trade-offs of decompression. Together, these design choices move the system beyond raw accuracy toward surgery-ready outputs: calibrated risk estimates, transparent localization, and thresholdable operating points tailored to clinical intent (e.g., a high-sensitivity threshold for ruling out CTS in primary screening vs a high-precision threshold when confirming surgical candidacy) [19]. Compared to previous AI studies on CTS ultrasound, our method, incorporated with built-in calibration tailored for supporting surgical decision-making, achieves an AUC of 0.904, an accuracy of 0.91, and a false negative rate of 0.121, surpassing Moser *et al*. (2024)'s [4] segmentation performance (Dice 0.76) and Waki *et al*. (2025)'s [5] video-based diagnostics (accuracy 0.75).

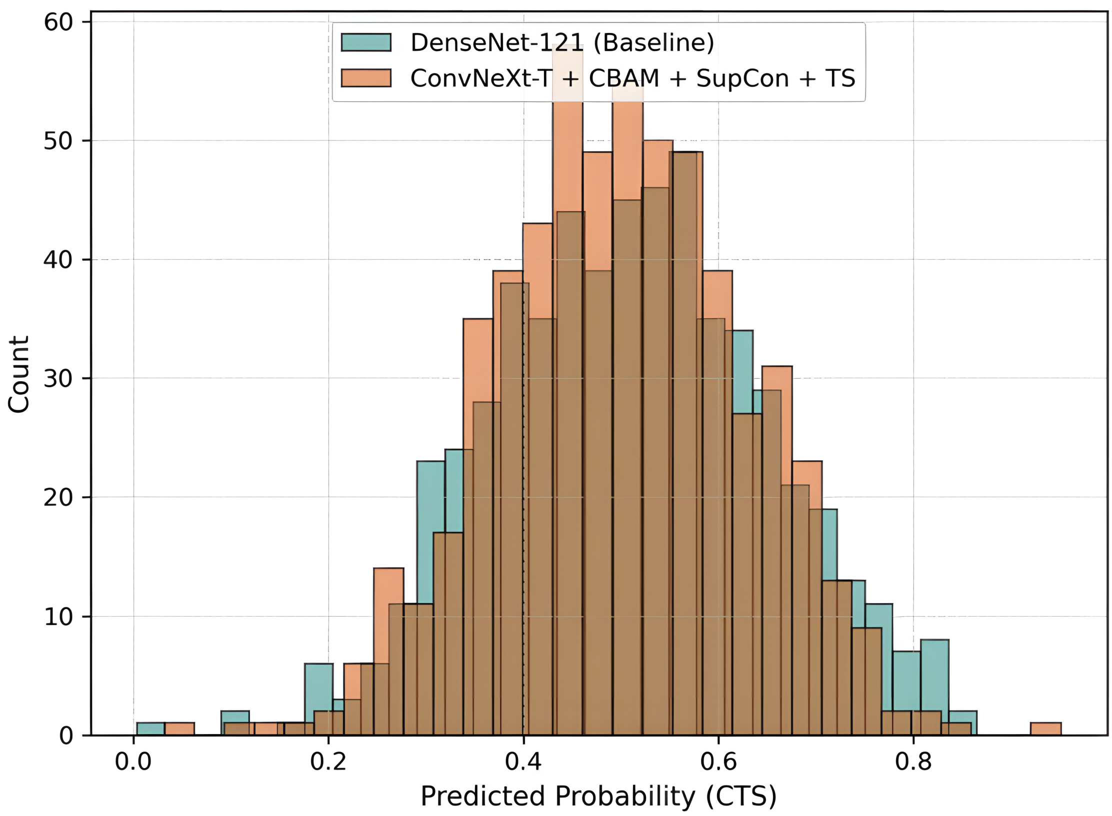

Fig. 4. Distribution of predicted probabilities of all test cases for DenseNet-121 (baseline) and the proposed ConvNeXt-T + CBAM + SupCon + Temperature scaling model. The results illustrate a shift toward more confident outputs with the calibrated ConvNeXt-T + CBAM + SupCon approach.

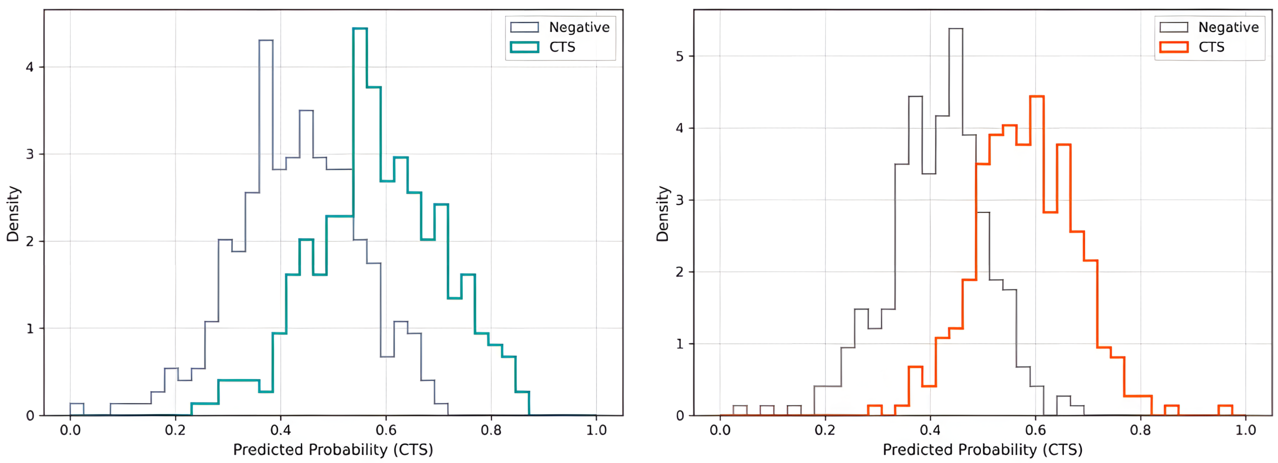

Fig. 5. Class-conditional score distributions (density) for negative vs CTS cases. (Left) baseline— DenseNet-121; (right) proposed model—ConvNeXt-T + CBAM + SupCon + Temperature scaling. The proposed model exhibits reduced inter-class overlap and stronger separability.

Clinical Implications for Surgical Decision-Making

From a clinical standpoint, the observed reduction in both FN and FP has practical implications along the perioperative pathway [20]. A lower FN rate mitigates delays to necessary decompression and the progression of denervation-related morbidity, whereas a lower FP rate decreases referrals and minimizes patient exposure to surgical risks without a clear indication [21]. The probability histograms and class-conditional score densities indicate a shift toward more confident and better-separated outputs,

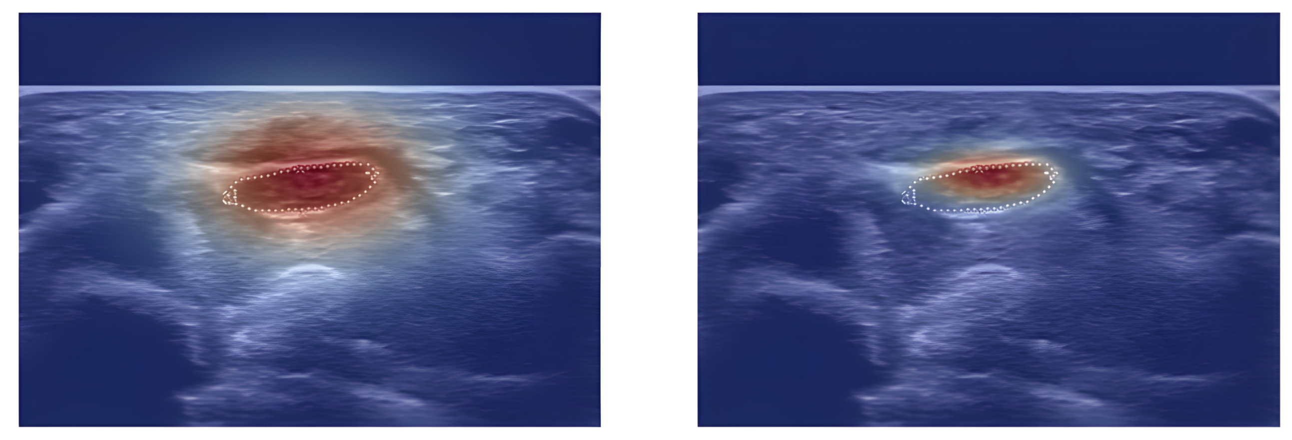

Fig. 6. Visual explainability and localization: Grad-CAM++ overlays (color) vs expert-defined regions of interest (white dashed contours) on representative ultrasound images. (Left) the reference baseline DenseNet-121 shows broader peri-lesional activation extending beyond the median nerve; (right) the proposed ConvNeXt-T + CBAM + SupCon + Temperature scaling model exhibits compact, well-localized attention largely confined within the expert boundary.

which in practice translate into clearer "gray zone" boundaries for surgeons and patients [22]. Equally important, the Grad-CAM++ overlays demonstrated more compact, nerve-concordant attention in the proposed model than in the baseline. Such anatomically faithful visualizations provide surgeons with evidence they can examine in conferences or clinical settings, fostering trust and facilitating focused discussion of the nerve segment targeted for decompression [23].

Limitations

Our work should be interpreted in light of several limitations. First, this is a retrospective, single-center study, inherently carrying biases such as ultrasound equipment variability, operator dependence, and limited patient demographics that constrain generalizability; therefore, external, prospective, multicenter validation investigation involving instruments from different vendors and varying acquisition presets is warranted to ensure that the findings are generalizable in other settings and demographics. Second, despite the implementation of standardized protocols in image acquisition, variations in the real-world clinical settings, such as those inherent to ultrasound equipment, operators, and imaging parameters, can significantly hamper median nerve visualization and diminish performance of the tested model. To counter this shortcoming, we propose post-hoc temperature recalibration for equipment adaptation, operator training aligned with recent protocols [18], domain adaptation approaches like unified generative models for handling synthetic-to-real shifts, or federated learning to mitigate this in future iterations [24]. Third, while temperature scaling improved calibration on our data, real-world shifts in prevalence and scanner heterogeneity can induce calibration drift. Therefore, lightweight site-specific recalibration (e.g., recalculating the temperature parameter on a small local validation set) is recommended before deployment. Fourth, saliency methods—including Grad-CAM++—are post-hoc approximations and can be brittle. Although our overlays demonstrated strong qualitative concordance with expert-defined ROIs, implementing prospective quality-control safeguards (e.g., automatic flags when attention extends beyond a learned nerve prior) is recommended. Fifth, due to retrospective constraints, the current study failed to conduct quantitative net benefit analyses, including cost-effectiveness analysis or decision curves, for the proposed model. Thus, future research should include pragmatic trials integrating these analyses with postoperative outcomes such as Disabilities of the Arm, Shoulder and Hand (DASH) scores. Finally, although the proposed model in this study is framed as surgery-guided, the reference standard of the model relied primarily on clinical diagnosis and electromyography data, without incorporating postoperative follow-up data such as symptom improvement or functional recovery, thereby impeding direct surgical validation. Future prospective trials should incorporate these outcomes, such as DASH scores for symptom relief and complication tracking, to strengthen clinical relevance.

Future Directions

The aforementioned limitations suggest clear directions for future investigations. A prospective, multicenter study incorporating device-specific domain adaptation and periodic calibration checks is needed to confirm the generalizability of our results. Incorporating cine ultrasound and simple patient-level aggregators (e.g., exam-level pooling across frames and levels) may further stabilize predictions. Automated or semi-automated nerve segmentation could be integrated upstream to standardize ROIs and enable attention-based guardrails. Finally, embedding the tool in the surgical workflow, including Picture Archiving and Communication System (PACS), real-time inference at the point of care, and standardized reporting that couples calibrated probabilities with attention maps, would facilitate consistent triage, support shared decision-making, and optimize surgical planning. Future validation could include five centers (three academic, two community), covering mainstream scanner brands such as GE, Philips, and Siemens. Per-site sample size of 200 (total n = 1000) determined based on power calculations to detect a 5% change in AUC with 80% power (alpha = 0.05), as well as incorporation of domain adaptation strategies to mitigate inter-equipment variability, are recommended.

In summary, by fusing a modern convolutional backbone and attention mechanism with representation-level contrastive learning, post-hoc calibration, and anatomy-aligned explainability, we developed a model that is specifically actionable for guiding surgical decision-making. The framework complements existing clinical assessment and nerve conduction studies by providing calibrated probabilities and transparent, anatomically grounded localizations that surgeons can rely on when selecting candidates for carpal tunnel release, determining timing of intervention, and counseling patients regarding expected outcomes. Further prospective, multicenter validation, coupled with outcome-linked evaluation, is warranted to confirm the clinical utility of the proposed ultrasound-based AI model and realize its potential as a routine tool in hand surgery practice.

Conclusions

In this paper, we present an ultrasound-based AI framework for CTS that integrates a modern ConvNeXt-T backbone with CBAM attention, a supervised contrastive warm-up, and post-hoc temperature scaling to generate anatomically faithful outputs that are both well-calibrated and highly discriminative. By aligning model design with surgical priorities, such as triage, timing of decompression, and preoperative counseling, the system provides probability estimates that correspond to observed risk and Grad-CAM++ localizations that can be visually audited against the median nerve, thereby supporting confident selection of surgical candidates while reducing both missed cases and unnecessary procedures. These findings position ultrasound-based AI as a practical adjunct to clinical assessment and nerve conduction studies in hand surgery. Future work should prioritize prospective, multicenter validation, device-specific recalibration, integration with cine ultrasound and patient-level data aggregation, and seamless workflow deployment, ultimately linking predictions to postoperative outcomes to demonstrate net benefits for surgical care.

Availability of Data and Materials

The data analyzed are available from the corresponding author upon reasonable request.

Author Contributions

KYY and HC designed the study and carried it out. KYY, HC, XKH, PPY, YQ and XFT supervised the data collection. KYY, HC, XKH, PPY, XYL and XW analyzed the data. KYY, HC, XKH, PPY, YQ and XFT interpreted the data. KYY and HC drafted and reviewed the manuscript. All authors have been involved in revising it critically for important intellectual content. All authors gave final approval of the version to be published. All authors have participated sufficiently in the work to take public responsibility for appropriate portions of the content and agreed to be accountable for all aspects of the work in ensuring that questions related to its accuracy or integrity are addressed.

Ethics Approval and Consent to Participate

Ethical approval was obtained from the Ethics Committee of the Sixth Hospital of Ningbo for the biomedical research involving humans titled “Clinical Study on Absorbable External Nerve Repair Materials”. The approval number is 2024-20(K). Written informed consent was obtained from a legally authorized representative for anonymized patient information to be published in this article. This study adheres to the relevant principles of the Declaration of Helsinki.

Acknowledgment

Not applicable.

Funding

This study was supported by the Medicine, Health, Science and Technology Program of Zhejiang Province (2024KY386) to KYY, the Science and Technology Program of Ningbo (2023Z187) to HC, and Ningbo Top Medical and Health Research Program (No.2022020506).

Conflict of Interest

The authors declare no conflict of interest.

Supplementary Material

Supplementary material associated with this article can be found, in the online version, at https://doi.org/10.62713/aic.4332.

References

[1] Gebrye T, Jeans E, Yeowell G, Mbada C, Fatoye F. Global and regional prevalence of carpal tunnel syndrome: A meta-analysis based on a systematic review. Musculoskeletal Care. 2024; 22: e70024. https://doi.org/10.1002/msc.70024.

[2] Chen Q, Zou X, Xia Y, Hu Y, Chen C, Zheng P. B-Mode ultrasound imaging in diagnosing carpal tunnel syndrome: an auxiliary diagnostic tool for hand surgeons. Frontiers in Neurology. 2024; 15: 1325464. https://doi.org/10.3389/fneur.2024.1325464.

[3] Braham S, Moussa A, Bouhoula M, Ben Meriem N, Annen I, Sakly G, *et al*. Exploring ultrasound and electromyography for carpal tunnel syndrome diagnosis: a comprehensive comparative study and implications for occupational medicine. Frontiers in Neurology. 2024; 15: 1490873. https://doi.org/10.3389/fneur.2024.1490873.

[4] Moser F, Muller S, Lie T, Langø T, Hoff M. Automated segmentation of the median nerve in patients with carpal tunnel syndrome. Scientific Reports. 2024; 14: 16757. https://doi.org/10.1038/s41598-024-65840-5.

[5] Waki T, Sato Y, Tsukamoto K, Yamada E, Yamamoto A, Ibara T, *et al*. Effectiveness of Comprehensive Video Datasets: Toward the Development of an Artificial Intelligence Model for Ultrasonography-Based Severity Diagnosis of Carpal Tunnel Syndrome. Journal of Ultrasound in Medicine: Official Journal of the American Institute of Ultrasound in Medicine. 2025; 44: 557–566. https://doi.org/10.1002/jum.16619.

[6] Kim SW, Kim S, Shin D, Choi JH, Sim JS, Baek S, *et al*. Feasibility of artificial intelligence assisted quantitative muscle ultrasound in carpal tunnel syndrome. BMC Musculoskeletal Disorders. 2023; 24: 524. https://doi.org/10.1186/s12891-023-06623-3.

[7] Loos NL, Hoogendam L, Souer JS, van Uchelen JH, Slijper HP, Wouters RM, *et al*. Algorithm versus expert: machine learning versus surgeon-predicted symptom improvement after carpal tunnel release. Neurosurgery. 2024; 95: 110–117. https://doi.org/10.1227/neu.0000000000002848.

[8] Roy S, Koehler G, Ulrich C, Baumgartner M, Petersen J, Isensee F, *et al*. MedNeXt: Transformer-Driven Scaling of ConvNets for Medical Image Segmentation. Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. Lecture Notes in Computer Science, vol 14223. Springer: Cham. 2023.

[9] Nissar I, Alam S, Masood S, Kashif M. MOB-CBAM: A dual-channel attention-based deep learning generalizable model for breast cancer molecular subtypes prediction using mammograms. Computer Methods and Programs in Biomedicine. 2024; 248: 108121. https://doi.org/10.1016/j.cmpb.2024.108121.

[10] Huang SC, Pareek A, Jensen M, Lungren MP, Yeung S, Chaudhari AS. Self-supervised learning for medical image classification: a systematic review and implementation guidelines. NPJ Digital Medicine. 2023; 6: 74. https://doi.org/10.1038/s41746-023-00811-0.

[11] Woo S, Park J, Lee JY, Kweon IS. Cbam: Convolutional block attention module. In Proceedings of the European conference on computer vision (ECCV) (pp. 3–19). 2018.

[12] Latha M, Kumar PS, Chandrika RR, Mahesh TR, Kumar VV, Guluwadi S. Revolutionizing breast ultrasound diagnostics with EfficientNet-B7 and Explainable AI. BMC Medical Imaging. 2024; 24: 230. https://doi.org/10.1186/s12880-024-01404-3.

[13] Lilli L, Giarnieri E, Scardapane S. A calibrated multiexit neural network for detecting urothelial cancer cells. Computational and Mathematical Methods in Medicine. 2021; 2021: 5569458. https://doi.org/10.1155/2021/5569458.

[14] Mekki YM, Rhim HC, Daneshvar D, Pouliopoulos AN, Curtin C, Hagert E. Applications of artificial intelligence in ultrasound imaging for carpal-tunnel syndrome diagnosis: a scoping review. International Orthopaedics. 2025; 49: 965–973. https://doi.org/10.1007/s00264-025-06497-1.

[15] Mehmood Y, Bajwa UI. Brain tumor grade classification using the ConvNext architecture. Digital Health. 2024; 10: 20552076241284920. https://doi.org/10.1177/20552076241284920.

[16] Li Z, Li H, Ralescu AL, Dillman JR, Altaye M, Cecil KM, *et al*. Joint self-supervised and supervised contrastive learning for multimodal MRI data: Towards predicting abnormal neurodevelopment. Artificial Intelligence in Medicine. 2024; 157: 102993. https://doi.org/10.1016/j.artmed.2024.102993.

[17] Singh R, Gupta S, Mohamed HG, Bharany S, Rehman AU, Ghadi YY, *et al*. Advancing prenatal healthcare by explainable AI enhanced fetal ultrasound image segmentation using U-Net++ with attention mechanisms. Scientific Reports. 2025; 15: 19612. https://doi.org/10.1038/s41598-025-04631-y.

[18] Battaglia F, Troisi L, Cigna E, Stagno d'Alcontres F, Rizzo V, Delia G. Standardized High-Resolution Ultrasound Protocol for the Diagnosis and Monitoring of Carpal Tunnel Syndrome: A Mixed-Design Observational Study. Diagnostics (Basel, Switzerland). 2025; 15: 1593. https://doi.org/10.3390/diagnostics15131593.

[19] Peng J, Zeng J, Lai M, Huang R, Ni D, Li Z. One-Stop Automated Diagnostic System for Carpal Tunnel Syndrome in Ultrasound Images Using Deep Learning. Ultrasound in Medicine and Biology. 2024; 50: 304–314. https://doi.org/10.1016/j.ultrasmedbio.2023.10.009.

[20] Rani V, Kumar M, Gupta A, Sachdeva M, Mittal A, Kumar K. Self-supervised learning for medical image analysis: a comprehensive review. Evolving Systems. 2024; 15: 1607–1633.

[21] Thirumal P, Narayanan R, Chakravarty P, Yareeda S, Ravula P. Diagnostic potential of high-resolution ultrasound of median nerve in assessing the severity of carpal tunnel syndrome. The Journal of Global Radiology. 2025; 10: 752. https://doi.org/10.7191/jgr.808.

[22] Zhang X, Xu A, Ouyang G, Xu Z, Shen S, Chen W, *et al*. Wavelet-Guided Multi-Scale ConvNeXt for Unsupervised Medical Image Registration. Bioengineering (Basel, Switzerland). 2025; 12: 406. https://doi.org/10.3390/bioengineering12040406.

[23] Dahan F, Shah JH, Saleem R, Hasnain M, Afzal M, Alfakih TM. A hybrid XAI-driven deep learning framework for robust GI tract disease diagnosis. Scientific Reports. 2025; 15: 21139. https://doi.org/10.1038/s41598-025-07690-3.

[24] Chaddad A, Hu Y, Wu Y, Wen B, Kateb R. Generalizable and explainable deep learning for medical image computing: An overview. Current Opinion in Biomedical Engineering. 2025; 33: 100567.