Medical Engineering &
Physics

IPEM
Institute of Physics and Engineering in Medicine

PAPER • OPEN ACCESS

Optical hand kinematics for machine learning-based screening of carpal tunnel syndrome

To cite this article: Şerife Çelikbaş *et al* 2026 *Med. Eng. Phys.* **147** 075014

View the article online for updates and enhancements.

You may also like

- Leveraging wearable sensors and machine learning for posture-based detection of carpal tunnel syndrome
Nikita Gautam, Sunil Kumar Gupta and Amit Shrivastava
- A Review on Risk Factors and Diagnosis of Carpal Tunnel Syndrome
Nikita Gautam and Amit Shrivastava
- Long exposure convolutional memory network for accurate estimation of finger kinematics from surface electromyographic signals
Weiyu Guo, Chenfei Ma, Zheng Wang et al.

This content was downloaded from IP address 161.9.142.52 on 18/08/2026 at 14:12

Medical Engineering & Physics

IPEM Institute of Physics and Engineering in Medicine

Check for updates

PAPER

OPEN ACCESS

RECEIVED
17 March 2026

REVISED
3 July 2026

ACCEPTED FOR PUBLICATION
21 July 2026

PUBLISHED
29 July 2026

Original content from this work may be used under the terms of the Creative Commons Attribution 4.0 licence.

Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

Optical hand kinematics for machine learning-based screening of carpal tunnel syndrome

Şerife Çelikbaş[1,4,*], **Abdurrahim Akgündoğdu**[2] and **Ayşegül Gündüz**[3]

[1] Biomedical Device Technology, Namık Kemal University, Tekirdağ, Türkiye
[2] Department of Electrical and Electronics Engineering, İstanbul University-Cerrahpaşa, İstanbul, Türkiye
[3] Cerrahpaşa Faculty of Medicine, Department of Neurology, İstanbul University-Cerrahpaşa, İstanbul, Türkiye
[4] Current address: Department of Biomedical Engineering, Institute of Graduate Studies, İstanbul University-Cerrahpaşa, İstanbul, Türkiye.
[*] Author to whom any correspondence should be addressed.

E-mail: serifecelikbas@nku.edu.tr

Keywords: carpal tunnel syndrome, hand kinematics, machine learning, clinical decision support, optical camera-based assessment

Supplementary material for this article is available online

Abstract

Carpal tunnel syndrome is usually assessed by clinical examination and electrodiagnostic testing, but these methods may not always be practical for rapid screening. In this work, it was explored whether hand motion recorded with a low-cost optical camera could add useful information to routine clinical data. Demographic variables, symptom-related findings obtained after the Phalen test, and five thumb-to-finger opposition tasks were collected. Hand landmarks were extracted with MediaPipe and used to derive kinematic descriptors, which were then analyzed together with clinical variables by machine-learning and hybrid deep-learning models. To reduce the risk of overly optimistic estimates, the main evaluation was carried out at hand level with leave-one-group-out cross-validation. In the primary leakage-safe comparison, the best Dataset 4 multimodal model yielded the strongest binary screening profile. Compared with the clinical Dataset 1 baseline, accuracy increased from 82.6% (95% CI 75.0%–90.2%) to 87.0% (79.3%–93.5%), sensitivity from 70.0% (54.4%–85.1%) to 80.0% (64.7%–93.4%), specificity from 88.7% (80.3%–95.6%) to 90.3% (83.1%–96.8%), and ROC-AUC from 0.845 (0.745–0.934) to 0.855 (0.752–0.945). In the same-architecture tests, the kinematic variables did not behave uniformly across model families. The clearest gain among the classical models was seen with light gradient boosting machine. Feature-level results were also consistent with the model findings, as selected thumb-opposition coordinates differed across stages and appeared to capture a functional component not represented by symptom scores. In the four-class severity task, the improvement was small: accuracy increased from 81.5% (72.8%–89.1%) to 82.6% (73.9%–90.2%), and macro-$F1$ increased from 0.706 (0.544–0.823) to 0.748 (0.588–0.850). In this cohort, the optical hand-kinematic measures mainly strengthened the screening analysis. Their use for finer severity staging will need confirmation in larger external datasets and in participant-level analyses.

1. Introduction

Carpal tunnel syndrome (CTS) is often diagnosed as a nerve condition causing tingling, numbness, and pain in the hand and fingers along the median nerve (Rotaru-Zavaleanu *et al* 2024). In routine clinical practice, diagnosis is initially based on bedside provocation tests such as Phalen's, Reverse Phalen's, and Durkan's maneuvers. However, in some patients, provocative tests may still leave uncertainty about the diagnosis. Tinel's sign reflects this problem well, because its diagnostic accuracy has been reported differently across studies. Therefore, its result should be evaluated together with the other clinical findings rather than used alone for screening (Gopal *et al* 2024). When clinical findings are not sufficient, nerve conduction studies are generally used to support the diagnosis, while imaging may be helpful

© 2026 The Author(s). Published on behalf of Institute of Physics and Engineering in Medicine by IOP Publishing Ltd.

when structural information is needed (Sonoo *et al* 2018). Ultrasonography is commonly based on the median nerve cross-sectional area measured at the carpal tunnel inlet, but the diagnostic cut-off values are not identical across populations (Luo *et al* 2025). MRI, particularly MR neurography, can show median nerve morphology and multiparametric changes in more detail; however, its cost and accessibility make it less practical for screening-based clinical use (Wang *et al* 2025). No single diagnostic method used in routine practice fully captures both the structural and functional changes associated with CTS. There are still several limitations in the current diagnosis of CTS. Imaging techniques depend on specialized equipment and expert evaluation, while electrophysiological tests are invasive and not always ideal for early diagnosis or repeated assessment. Delayed diagnosis can therefore allow CTS to progress from sensory complaints to muscle weakness, thenar atrophy, deformity, and reduced grip strength. This is an important functional issue, and (Silva *et al* 2024) also emphasized the need for alternative approaches that may support earlier diagnosis and follow-up after treatment in CTS cases with thenar atrophy. In more advanced cases, surgery may be needed, although recovery may differ between patients. Therefore, simple and non-invasive approaches are still needed to support earlier screening and severity assessment of CTS beyond specialized clinical settings.

Current diagnostic approaches differ not only in diagnostic performance but also in patient comfort, cost, and accessibility. Although nerve conduction studies are still widely accepted as a reference method, they are invasive and require considerable time and clinical resources (Wipperman and Goerl 2016, Sasaki *et al* 2022, Tüfekçi and Boz 2022, Öten *et al* 2026). Although ultrasound and MR neurography can add structural and quantitative imaging information, their dependence on imaging systems and expert interpretation may limit their use as rapid screening tools in all clinical settings. In recent years, AI has attracted growing interest in supporting CTS diagnosis and management. Recent deep-learning studies have also attempted to improve CTS diagnosis by modeling complex clinical and electrophysiological patterns; however, these approaches generally remain dependent on specialized data sources rather than simple functional recordings (Elseddik *et al* 2023). Previous studies have shown that combined diagnostic approaches may help with treatment decisions in mild to moderate CTS and with postoperative evaluation in more severe cases (Wipperman and Goerl 2016). Still, many of the AI-based methods proposed so far rely on specialized equipment, invasive data acquisition, or highly controlled settings, limiting their usefulness in everyday clinical practice.

Numerous studies have investigated automated and data-driven strategies for CTS diagnosis. Electrophysiological analyses have shown that a sensory conduction velocity threshold of 40.5 m/s can achieve high diagnostic sensitivity and specificity (Tüfekçi and Boz 2022). Other work has questioned whether commonly used clinical grading systems adequately reflect disease severity, highlighting the need for more objective and reproducible measures (Sasaki *et al* 2022). Machine learning studies based on nerve conduction and clinical data have reported diagnostic accuracies above 90% in selected cohorts, and analyses of larger datasets have suggested that CTS severity can be differentiated with reasonable reliability.

More recent research has moved toward non-invasive motion- and kinematics-based approaches. Alterations in wrist and finger motion during task-based assessments have been documented in CTS populations relative to healthy controls. On the basis of these observations, wearable sensor-based systems have been used to measure grip strength and joint movement, mainly in rehabilitation settings. Gyroscope-based tools have also been employed to identify reductions in thumb rotation and palmar abduction in CTS patients (Kuroiwa *et al* 2018). In parallel, mobile and tablet-based applications have been introduced to assess thumb movement speed, pen pressure, and drawing patterns, with reported sensitivities ranging from 71% to 92% and specificities of up to 99% (Fujita *et al* 2019, Koyama *et al* 2021, Watanabe *et al* 2021). Video-based approaches using computer vision have further suggested that brief, task-specific hand motion recordings can provide clinically meaningful diagnostic information (Tsukamoto *et al* 2024).

Despite these advances, many existing systems still depend on specialized sensors, wearable devices, or laboratory-based equipment, which limits their scalability and routine clinical use. Low-cost optical cameras make it easier to record video. With computer vision and machine learning, these videos can be used to obtain detailed information about hand movements without using specialized sensors.

In this study, low-cost optical camera recordings were combined with machine learning and hybrid deep-learning approaches to investigate whether camera-derived hand kinematics could support CTS screening and severity assessment. Demographic and symptom-related variables were evaluated together with kinematic features obtained from five thumb-finger opposition tasks. Particular attention was given to a clinically realistic evaluation setting by analysing the data at hand level and by reducing the risk of

information leakage during model assessment. In this way, the additional value of kinematic information beyond routine clinical variables could be examined more reliably.

The main contributions of this study are as follows:

- A machine learning-based framework for CTS screening and severity assessment is presented using video recordings acquired with a standard optical camera.
- Hand kinematic patterns are extracted from camera recordings without the use of dedicated sensors or wearable devices.
- The contribution of demographic, symptom-related, and motion-based feature groups to CTS classification is systematically examined.
- The proposed approach is evaluated in a hand-level, leakage-aware setting intended to provide a more clinically realistic estimate of performance.

2. Materials and methods

2.1. Dataset and data acquisition

Volunteers presenting to the Electromyography Laboratory of the Neurology Department at İstanbul University–Cerrahpaşa were enrolled. Each participant was informed in advance and provided written informed consent. Ethics approval for the study was obtained from the İstanbul University–Cerrahpaşa Ethics Committee before participant recruitment (No. 2024/174). The study included 46 participants: 27 women and 19 men, aged 21–76 years, with a mean age of 51.9 ± 17.4 years. Of the 46 participants, 45 were right-hand dominant, and 1 was left-hand dominant. Data from 92 hands were included in the analysis. A desktop application developed in Python (Kivy) and a 1080p Logitech USB webcam were used during acquisition. Recordings were collected under routine clinical lighting conditions, including natural daylight in outpatient rooms, and were securely transferred to a Firebase database during acquisition (figure 1).

Previous studies have shown that age, sex, and hand dominance may influence both the likelihood of developing CTS and the way symptoms are experienced or reported (Stevens *et al* 1992). For this reason, age, sex, birthplace, and dominant hand were recorded for each participant. After the preliminary phase, each participant completed a one-minute Phalen's test. During this test, pain, numbness, and paresthesia were scored in 22 predefined hand regions using a four-level ordinal scale (none, mild, moderate, severe) (figure 2).

Subjective symptom information was complemented by camera-based motion recording. After completion of the Phalen test, five thumb-finger opposition tasks were recorded: *oppositio digitorum*, *oppositio pollicis et indicis*, *oppositio pollicis et medii*, *oppositio pollicis et annularis*, and *oppositio pollicis et digiti minimi*. Each task lasted approximately five seconds. These movements were selected because CTS may affect thumb opposition, fine coordination, and hand function (Wiesman *et al* 2003, Afshar *et al* 2025). Hand landmarks were extracted from the videos with the Google MediaPipe Hands framework. Before recording, each thumb–finger opposition task was demonstrated to the participant. For each trial, the target thumb–finger opposition posture was reached and held briefly during the approximately five-second recording period; repeated cyclic tapping was not performed. Thumb-finger contact pressure was not measured. Participants were instructed to keep the hand visible to the camera throughout the task. However, the recordings were obtained under routine outpatient-room lighting conditions in order to reflect a practical screening environment.

2.1.1. Hand tracking and joint detection algorithm

A number of methods have been proposed for hand tracking and joint localization. Some methods reconstruct 3D hand pose from conventional RGB images using convolutional neural networks, whereas others address occlusion using multi-view geometry or weakly supervised keypoint estimation (Simon *et al* 2017, Zimmermann and Brox 2017). In the present study, the open-source MediaPipe Hands framework was adopted because it provides real-time 2.5D hand landmark estimation from a single RGB image (Zhang *et al* 2020). The framework first detects the palm to define a region of interest and then predicts 21 hand landmarks together with relative depth information.

In the present dataset, 21 landmarks were extracted for each of the five opposition tasks (figure 3). In some recordings, individual landmark values were missing because of reflections, occlusion, or rapid hand motion. The analysis was therefore designed to reduce the impact of frame-level landmark

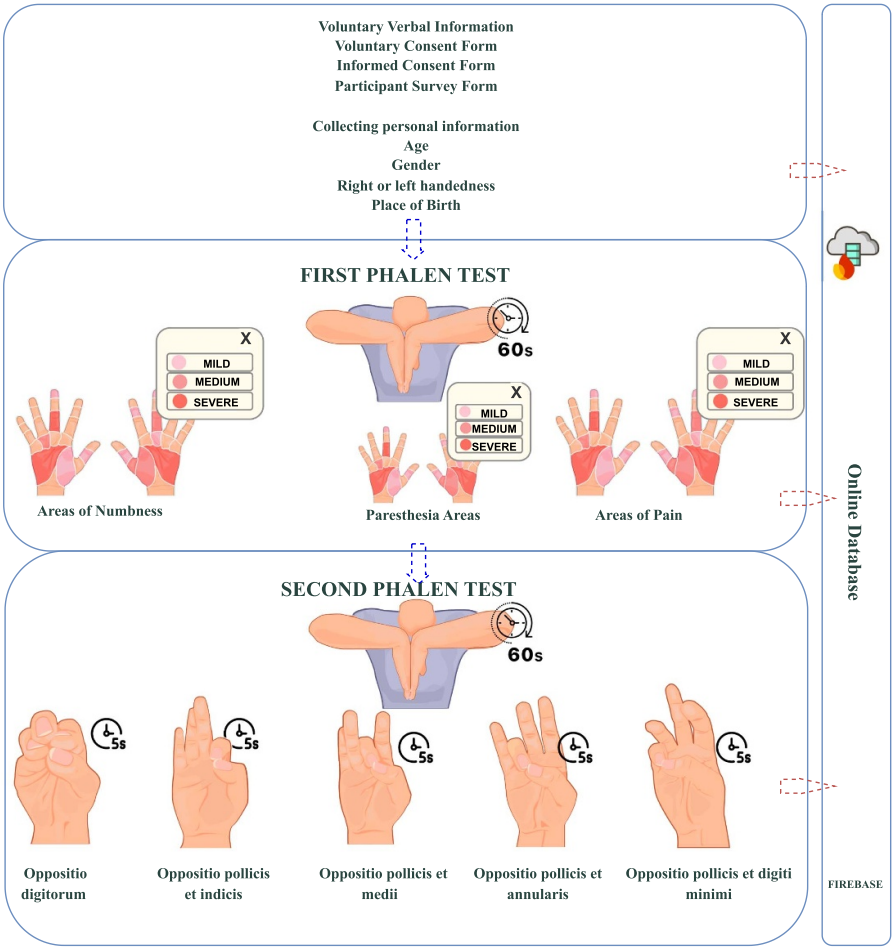


Figure 1. Workflow diagram for data acquisition.

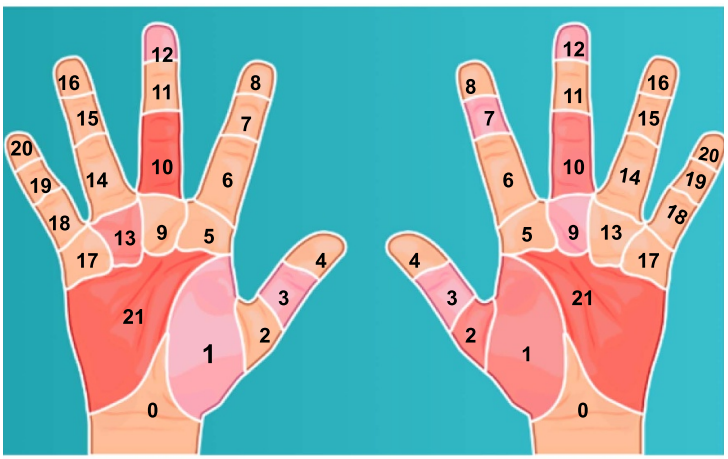


Figure 2. Numbered hand regions used for pain, numbness, and paresthesia scoring.

variability. Instead of relying on all raw coordinates, we used scale-normalized distances, selected anchor coordinates, fold-safe imputation, and hand-level aggregation of repeated observations. For the tabular analyses, missing values were set to 0 before hand-level aggregation. For the deep-learning analyses, missing values were imputed with a median imputer fitted on the training partition of each outer fold and then applied to the corresponding validation and test partitions.

Rather than using the full set of raw landmark coordinates directly, a compact kinematic representation was constructed. For each gesture, scale-normalized distances between clinically relevant landmarks

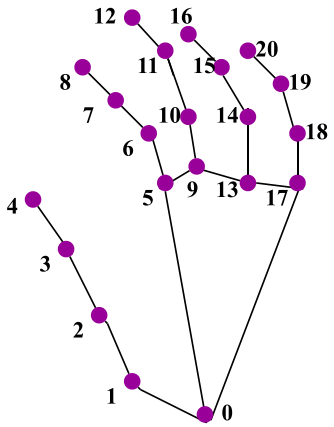


Figure 3. MediaPipe hand landmarks used in the study (Zhang *et al* 2020).

Table 1. Engineered kinematic feature groups used in Datasets 3 and 4.

Feature group	No. of features
Scale-normalized distances	45
Joint angles	70
Sequence descriptors	63
Cross-gesture summaries	20
Total	198

and finger joint angles were calculated. In addition, the ordered set of five opposition tasks was treated as a short motion sequence, and descriptors related to speed, acceleration, smoothness, and cross-gesture variation were derived. These engineered features were combined with a small set of raw anchor coordinates from the wrist and fingertips. In this way, coarse geometric information was retained while the dimensionality and noise of the full coordinate set were reduced.

2.1.2. Dataset construction

After data collection, four dataset configurations were prepared.

- **Dataset 1** contained demographic variables and symptom scores. Specifically, it included age, sex, birthplace, dominant hand, and the regional pain, numbness, and paresthesia scores obtained after the Phalen test.
- **Dataset 2** extended Dataset 1 by adding selected landmark coordinates. Six anchor landmarks were retained for each gesture, namely the wrist and fingertip landmarks (0, 4, 8, 12, 16, and 20), resulting in 90 additional coordinate features.
- **Dataset 3** contained only camera-derived motion information. It consisted of 198 engineered kinematic features derived from distances, joint angles, sequence-based descriptors, and cross-gesture summary statistics, together with 90 raw anchor coordinates.
- **Dataset 4** combined the full clinical block of Dataset 1 with the compact kinematic block used in Dataset 3. This dataset was used as the main multimodal fusion configuration.

The 198 engineered kinematic features in Dataset 3 and Dataset 4 were grouped as shown in table 1.

A schematic overview of the selected landmarks used as anchor coordinates is shown in figure 4.

The resulting datasets contained 70, 160, 288, and 358 input features for Datasets 1–4, respectively. A summary is provided in table 2. These feature counts refer to the row-level dataset definitions before hand-level aggregation. At the raw recording level, the dataset contained 460 rows. Because repeated recordings were available for each hand, the primary model evaluation was later carried out in a leakage-aware manner by aggregating repeated observations to the hand level. This yielded 92 hand-level groups. Class labels were assigned by a neurologist using clinical findings together with electrophysiological assessment. At the hand level, 62 hands were classified as stage 0 (non-CTS), whereas 30 hands showed CTS (6 stage 1, 15 stage 2, and 9 stage 3). For the binary task, stages 1–3 were grouped as CTS. Staging

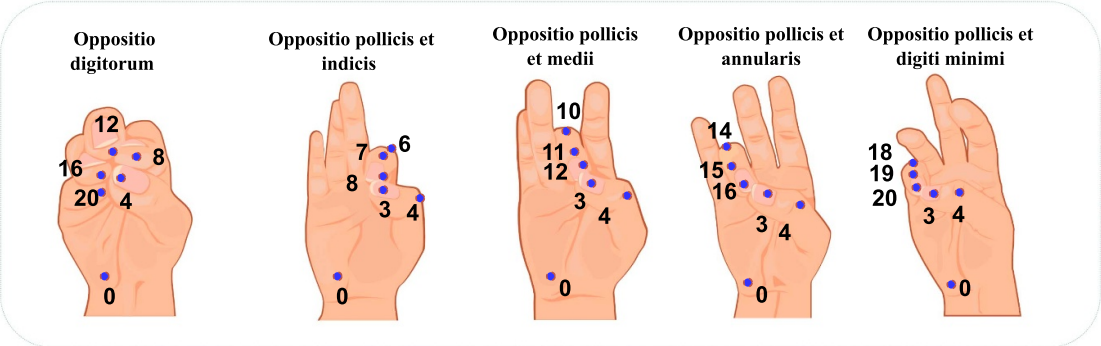


Figure 4. Selected MediaPipe landmarks used as anchor coordinates in the compact kinematic representation.

Table 2. Dataset configurations used in the study.

Dataset	Content	No. of features
Dataset 1	Demographic variables and symptom scores from 22 hand regions (pain, numbness, paresthesia)	70
Dataset 2	Dataset 1 plus selected landmark coordinates from six anchor landmarks across five gestures	160
Dataset 3	Engineered kinematic features derived from distances, joint angles, sequence descriptors, and cross-gesture summaries, together with raw anchor coordinates	288
Dataset 4	Dataset 1 combined with the compact camera-derived kinematic block used in Dataset 3	358

was supported by electrophysiological parameters such as median sensory nerve conduction velocity, with stage 0 corresponding to values above 50 m s^{-1}, stage 1 to approximately 45–50 m s^{-1}, stage 2 to approximately 40–45 m s^{-1}, and stage 3 to values below 40 m s^{-1} (Tüfekçi and Boz 2022).

2.2. Machine learning and hybrid deep-learning approaches

For the final tabular comparison, four tree-based models were included: extreme gradient boosting (XGB), light gradient boosting machine (LGBM), an XGB+LGBM soft-voting ensemble, and an XGB+RF soft-voting ensemble. These models were preferred because they are well suited to tabular biomedical data and can model nonlinear relationships between clinical and kinematic variables (Quinlan 1996, Bühlmann and Yu 2010, Kamiński *et al* 2018, Bentéjac *et al* 2021, Ali *et al* 2023, Lokker *et al* 2024). The Random Forest model was configured with 300 trees, a maximum depth of 12, a minimum split size of 5, and a minimum leaf size of 2. XGB used 600 trees, a maximum depth of 4, a learning rate of 0.05, subsampling of 0.9, and column subsampling of 0.9. LGBM used 800 trees, a learning rate of 0.03, and 31 leaves. In the soft-voting ensembles, class probabilities were averaged with weights of 2:1 in favor of XGB. All randomized components were initialized with random state 42.

In addition to the tabular models, deep-learning and hybrid models were also evaluated for the datasets that contained camera-derived motion information. The nested gesture encoder was designed as a two-level structure. At the first level, each of the five opposition tasks was processed as a separate gesture block using a shared gesture-level multilayer perceptron. Each gesture block contained 18 coordinate inputs from the six selected anchor landmarks, followed by two fully connected layers with 128 units, ReLU activation, and dropout. At the second level, the five gesture-level representations were combined using attention pooling and then concatenated with the auxiliary clinical/symptom feature block before classification.

In the FrozenNestedEWC variant, partial freezing refers to keeping the first linear layer of the shared gesture encoder fixed during stage-B fine-tuning. This was used to preserve the lower-level gesture representation learned during the first stage. Elastic weight consolidation (EWC) was used as a regularization penalty to stabilize fine-tuning in the small-sample setting, rather than as a continual-learning benchmark Kirkpatrick *et al* (2017).

Two deep architectures were considered: a nested gesture encoder and a frozen nested encoder regularized with EWC and replay-based fine-tuning. Hybrid variants were obtained by combining the learned deep embeddings with classical machine-learning heads. In this way, purely tabular models, purely deep models, and multimodal fusion models based on learned motion representations could be compared within the same framework. For the deep-learning and hybrid analyses, the available `coord_` variables were reshaped as gesture-wise coordinate blocks corresponding to five gestures and six anchor landmarks

Table 3. Hyperparameter summary for the evaluated classical, deep-learning, and hybrid models.

Model	Hyperparameters
Random forest	300 trees; `max_depth` = 12; `min_samples_split` = 5; `min_samples_leaf` = 2; `random_state` = 42
XGB	600 trees; `max_depth` = 4; learning rate = 0.05; subsample = 0.9; `colsample_bytree` = 0.9; `random_state` = 42
LGBM	800 trees; learning rate = 0.03; `num_leaves` = 31; `random_state` = 42
Soft voting	XGB+LGBM and XGB+RF; 2:1 weighting toward XGB
NestedGestureNet	Batch size = 32; AdamW learning rate = 10^{-3}; `weight_decay` = 10^{-4}; dropout = 0.20; `max_epochs` = 100; patience = 10
FrozenNestedEWC	Stage-A = 40 epochs; Stage-B = 80 epochs; EWC $\lambda = 5.0$; replay = 0.20; label smoothing = 0.05
Hybrid heads	XGB head: 220 trees, depth = 4, learning rate = 0.05; LGBM head: 350 trees, 31 leaves, learning rate = 0.03

(wrist and fingertips). The remaining variables were treated as an auxiliary feature block. Within each outer training fold, missing values were imputed using a median imputer fitted on the training partition, followed by standardization with training-fold statistics. A group-wise inner validation split was then created using `GroupShuffleSplit` with a validation proportion of 0.20 and random state 42. Models were trained with AdamW using a batch size of 32, a learning rate of 10^{-3}, a weight decay of 10^{-4}, dropout of 0.20, a maximum of 100 epochs, and early stopping with a patience of 10 epochs. For the FrozenNestedEWC model, stage-A pretraining was run for 40 epochs and stage-B fine-tuning for 80 epochs, with partial freezing of the first gesture-encoder linear layer, an EWC penalty of $\lambda = 5.0$, a replay ratio of 0.20, and label smoothing of 0.05. In the hybrid models, deep embeddings were classified with XGB and/or LGBM heads trained within each outer fold; the hybrid XGB head used 220 trees, depth 4, and a learning rate of 0.05, whereas the hybrid LGBM head used 350 trees, 31 leaves, and a learning rate of 0.03. For binary classification, the decision threshold was tuned on the inner validation split by maximizing MCC over thresholds from 0.05 to 0.95 in steps of 0.05. The fixed hyperparameter settings used in the evaluated classical, deep-learning, and hybrid models are summarized in table 3.

This table documents the fixed model settings used in the final comparisons and clarifies that the reported results were obtained under a predefined modeling protocol.

Before the final outer-fold evaluation, the model settings were kept fixed, and no choice was made by looking at the held-out test folds. Since the dataset included only 92 hands, relatively conservative settings were used to reduce overfitting. These settings were based on common default choices, preliminary internal trials, and the size of the dataset. The inner validation split was used for early stopping and binary-threshold selection only. It was not used to tune the models repeatedly according to test-fold results. Therefore, this analysis should be considered a fixed-protocol model comparison rather than a full hyperparameter-search study.

To reduce information leakage, the primary evaluation was carried out with Leave-One-Group-Out cross-validation at the hand level. Groups were defined from the hand identifier, and repeated observations from the same hand were never split across training and test folds. Before model fitting, repeated rows belonging to the same hand were aggregated by feature-wise mean and standard deviation. This yielded 92 hand-level groups from 460 raw recording rows. For comparison purposes, an additional optimistic protocol based on non-independent row-level stratified splitting was also evaluated. All 46 participants contributed both hands. Therefore, although hand-level LOGO prevents repeated recordings from the same hand from crossing training and test folds, it does not fully remove possible participant-level dependence between the two hands of the same participant. To address this issue, a bilateral label-concordance audit and a participant-level leave-one-participant-out sensitivity analysis were additionally performed.

All preprocessing steps were fitted on the training data only and then applied to the held-out fold. Continuous variables were standardized using training-fold statistics. Class imbalance was addressed only in the training partition. Since the dataset included fewer CTS hands than non-CTS hands, this step was used to limit the tendency of the model to favor the majority class, a common problem in machine-learning studies with imbalanced data (Altalhan *et al* 2025). In the final tree-based baseline models, synthetic oversampling was disabled and balanced class/sample weights were used instead. When auxiliary synthetic minority oversampling technique (SMOTE) experiments were performed, SMOTE ($k = 3$, random state 42) was applied only to the training fold after scaling and never to validation or test data. In the tabular analysis of Dataset 4, the full clinical block was retained, raw coord_ features were excluded from tabular feature selection, and the engineered kinematic block was reduced within each training fold using mutual-information ranking (mutual_info_classif, random state 42). The top 25 kinematic features

selected in the training fold were then carried forward to the corresponding held-out fold. In the protocol comparison experiments, the clinical block and the motion block were also modeled separately and combined at the decision level in a late-fusion setting. Model performance was assessed using sensitivity, accuracy, specificity, $F1$-score, macro-$F1$, balanced accuracy, Matthews correlation coefficient (MCC), and ROC-AUC. For the primary Dataset 1 versus Dataset 4 comparisons, 95% confidence intervals were estimated using non-parametric bootstrap resampling of the out-of-fold hand-level predictions. Resampling was performed at the hand level with 500 bootstrap replicates. For each metric, the central 95% interval was obtained from the empirical bootstrap distribution by using its lower and upper 2.5% tails. For the binary task, CTS was defined as the union of stages 1–3 versus stage 0. For the multi-class task, all four stages were retained. Complete model-level performance results for all evaluated classical, deep-learning, and hybrid models are provided in Supplementary Table S.

2.3. Statistical analysis

Before analysis, the data were examined to see if they satisfied the expected patterns of normality and equal variance. The Shapiro–Wilk test (Shapiro and Wilk 1965) was used to evaluate distribution patterns, and Levene's test (Brown and Forsythe 1974) to assess variance consistency between groups. Because the data were not normally distributed, comparisons across groups were carried out with the Kruskal–Wallis test (Kruskal and Wallis 1952, Gastwirth *et al* 2009). When this test showed a significant result, pairwise differences were explored using Dunn's post-hoc procedure (George and Mallery 2010). All statistical decisions were based on a threshold of $p < 0.05$. In addition to formal hypothesis testing, the statistically retained variables were also ranked using a practical separation score to facilitate interpretation of the strongest between-stage contrasts.

For each variable, a separation score was calculated for every pairwise stage comparison as follows:

$$S_{ab} = \frac{|\tilde{x}_a - \tilde{x}_b|}{\mathrm{IQR}_{\mathrm{all}} + \epsilon}.$$

Here, $\tilde{x}_a$ and $\tilde{x}_b$ denote the median values of the two stages being compared. $\mathrm{IQR}_{\mathrm{all}}$ represents the interquartile range of the same variable across all hand-level observations, and ϵ was included only to prevent division by zero.

This score was used only to prioritize already significant findings and not to define statistical significance. For concise reporting in the main text, only representative coordinates from the strongest pairwise contrast (stage 0 vs stage 3) with separation $\geqslant 0.20$ are summarized in table 5.

2.4. Proposed methodology

A multistage workflow was designed for CTS screening and severity classification using demographic variables, regional post-Phalen symptom scores, and camera-derived hand kinematics (figure 5). The workflow comprised four main stages: data acquisition, landmark extraction and feature construction, leakage-aware model development, and performance evaluation.

The recordings described in section 2 were first processed into four dataset configurations with increasing feature complexity. Dataset 1 contained demographic variables and regional symptom scores obtained after the Phalen test. Dataset 2 extended this configuration by adding selected landmark coordinates from six anchor landmarks across five gestures. Dataset 3 included only camera-derived motion information and consisted of engineered kinematic features together with anchor coordinates. Dataset 4 combined the complete clinical block of Dataset 1 with the compact kinematic block of Dataset 3 and served as the main multimodal fusion dataset.

The raw video recordings were processed with MediaPipe Hands, and 21 landmarks were identified for each opposition task. Rather than using all raw coordinates directly, a more compact motion representation was created. This representation included scale-normalized landmark distances, joint angles, sequence-based descriptors, and summary measures across gestures. Wrist and fingertip anchor coordinates were also added to preserve basic geometric information while limiting redundancy and noise.

To obtain clinically more realistic performance estimates, the primary analysis was carried out in a leakage-aware setting. Repeated recordings belonging to the same hand were aggregated before model fitting by computing summary statistics at hand level, thereby yielding 92 hand-level groups from 460 raw recording rows. Preprocessing was performed using only the training data and then applied to the held-out fold. Continuous features were standardized using training-fold statistics. Class imbalance was handled only in the training data. In the final tree-based baselines, balanced class/sample weights were used. SMOTE ($k = 3$, random state 42) was reserved for auxiliary analyses and, when enabled, was applied only to the training fold after scaling.

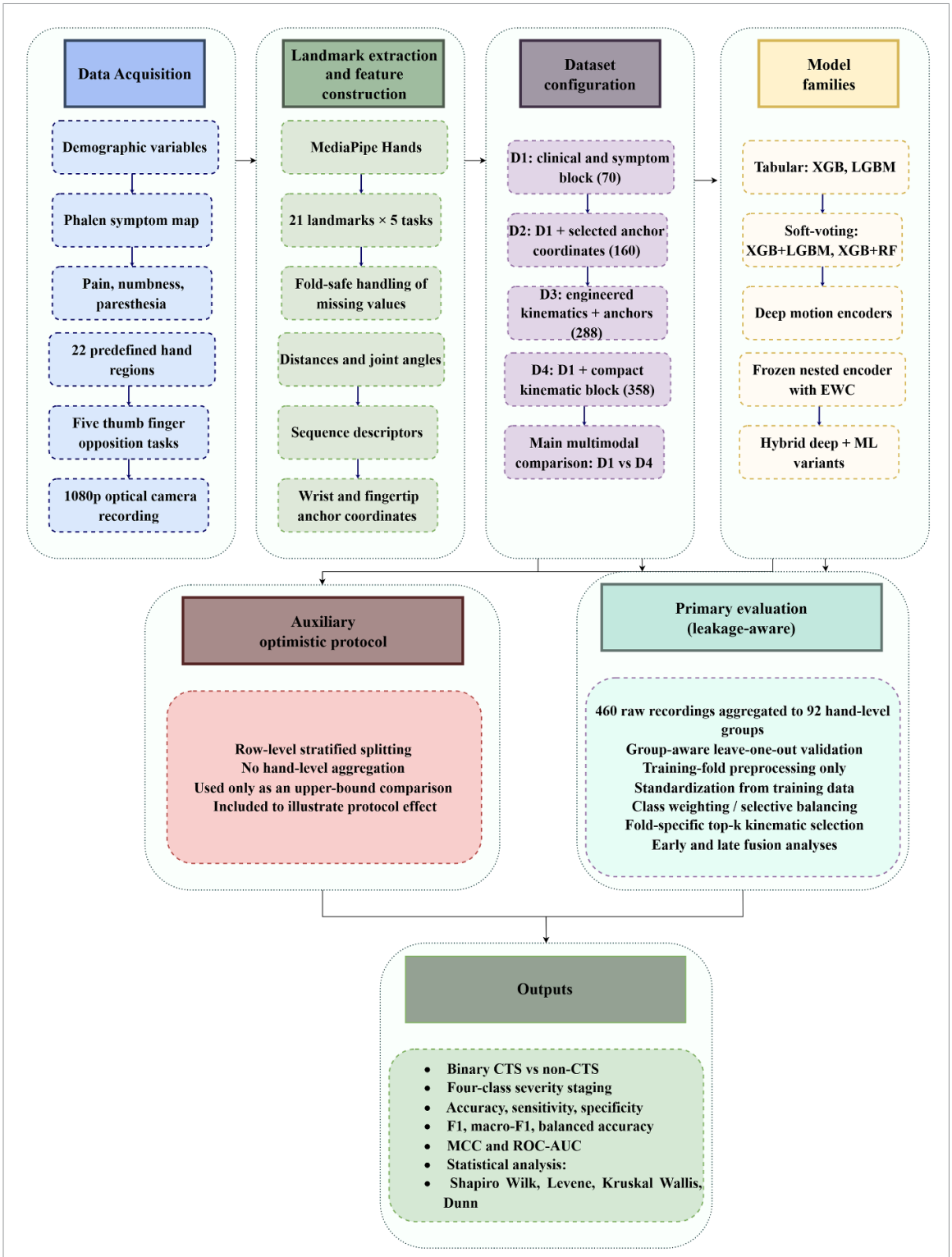


Figure 5. Workflow of the proposed CTS screening and severity assessment framework. SMOTE denotes Synthetic Minority Oversampling Technique, and EWC denotes elastic weight consolidation.

For tabular modeling, XGB, LGBM, XGB+LGBM, and XGB+RF were evaluated across the four dataset configurations. In addition, deep-learning and hybrid models were used for the camera-derived motion datasets. In the multimodal experiments, the clinical and kinematic blocks were examined both in an early-fusion setting and in a late-fusion setting in which the two branches were modeled separately and combined at decision level. In the Dataset 4 tabular analyses, the clinical block was retained in full, whereas the kinematic block was reduced within each training fold using fold-specific top-k feature selection to limit overfitting and avoid feature-selection leakage.

Model development was performed for both binary CTS detection and four-class severity classification. In the primary analysis, generalization was assessed using group-aware leave-one-out validation at the hand level. For comparison, a second protocol based on non-independent row-level splitting was also evaluated as a more optimistic setting. Model performance was reported using sensitivity, specificity, accuracy, *F*1-score, macro-*F*1, balanced accuracy, MCC, and ROC-AUC. The selected features from Dataset 4 were also examined statistically, as described in section 2.3.

3. Results

3.1. Primary leakage-safe classification results

The primary analysis was performed under the leakage-safe, hand-level evaluation protocol. Figure 7 summarizes both the primary Dataset 1 versus Dataset 4 comparisons and the fixed-LGBM ablation across all four predefined dataset configurations, while figure 6 shows the corresponding binary confusion matrices for the primary models. In the binary task, the best model obtained from Dataset 1 was the XGB+LGBM ensemble, whereas the best model obtained from Dataset 4 was the Hybrid_FrozenNestedEWC+ML model. As shown in figure 7(a), the point estimates favored the multimodal Dataset 4 model across the main binary screening metrics. Accuracy increased from 82.6% (95% CI 75.0%–90.2%) to 87.0% (79.3%–93.5%), sensitivity from 70.0% (54.4%–85.1%) to 80.0% (64.7%–93.4%), specificity from 88.7% (80.3%–95.6%) to 90.3% (83.1%–96.8%), and ROC-AUC from 0.845 (0.745–0.934) to 0.855 (0.752–0.945). These intervals show the uncertainty expected from a modest 92-hand dataset, but the direction of change was favorable for Dataset 4, particularly for sensitivity, MCC, and the reduction in false-negative classifications. The corresponding confusion matrices in figure 6 show that the Dataset 4 model reduced the number of false negatives from 9 to 6 and the number of false positives from 7 to 6. Overall, the total number of misclassified hands decreased from 16 to 12. To address the attribution of this improvement, Dataset 1 and Dataset 4 were also compared under identical classical architectures. This analysis showed that the added kinematic block did not affect all model families in the same way. The clearest directional gain was observed with LGBM, where accuracy increased from 0.815 to 0.837 and MCC increased by +0.056 after adding the kinematic block to the clinical/symptom variables. XGB and XGB+RF showed small positive MCC changes, whereas XGB+LGBM showed a small decrease. The exact McNemar tests did not show a clear architecture-independent difference in this restricted set of classical models. Therefore, the Dataset 4 finding should be read cautiously: camera-derived kinematic descriptors may improve symptom-based learning when they are used in a suitable multimodal model, but their effect is not the same across model families (figure 9).

3.1.1. Ablation across the four predefined dataset configurations

The four predefined feature configurations were also tested with the same LGBM model to separate the effect of the feature set from the effect of the classifier. In the binary task, accuracy was 0.815 for Dataset 1, 0.707 for Dataset 2, 0.652 for Dataset 3, and 0.837 for Dataset 4. Sensitivity was 0.700, 0.500, 0.400, and 0.767, respectively, and MCC was 0.576, 0.315, 0.181, and 0.632. These results show that the engineered kinematic block alone did not exceed the clinical/symptom feature block. The strongest binary result under fixed LGBM was obtained when the clinical and kinematic variables were used together in Dataset 4.

In the four-class task, the fixed-LGBM comparison gave accuracies of 0.815 for Dataset 1, 0.630 for Dataset 2, 0.663 for Dataset 3, and 0.815 for Dataset 4. Dataset 4 matched Dataset 1 in accuracy and balanced accuracy, while giving slightly higher macro-*F*1 and ROC–AUC. Thus, in this cohort, the camera-derived features should not be viewed as a stand-alone replacement for symptom-based variables. Their main value appeared when they were added to the clinical/symptom block in the multimodal Dataset 4 representation.

Although the ROC–AUC difference between the best leakage-safe binary models was modest, the Dataset 4 model showed higher sensitivity and slightly higher specificity than Dataset 1. The confusion matrices also showed fewer false negatives and fewer false positives for Dataset 4 than for Dataset 1.

In the four-class severity analysis, the best Dataset 1 model was LGBM, whereas the best Dataset 4 model was Hybrid_NestedGestureNet+LGBM. As shown in figure 7(b), the point estimates again favored Dataset 4, although the magnitude of change was smaller than in the binary task. Accuracy increased from 0.815 (95% CI 0.728–0.891) to 0.826 (0.739–0.902), balanced accuracy from 0.706 (0.546–0.846) to 0.748 (0.597–0.871), macro-*F*1 from 0.706 (0.544–0.823) to 0.748 (0.588–0.850), MCC from 0.634 (0.471–0.785) to 0.656 (0.500–0.800), and ROC–AUC from 0.849 (0.752–0.923) to 0.851

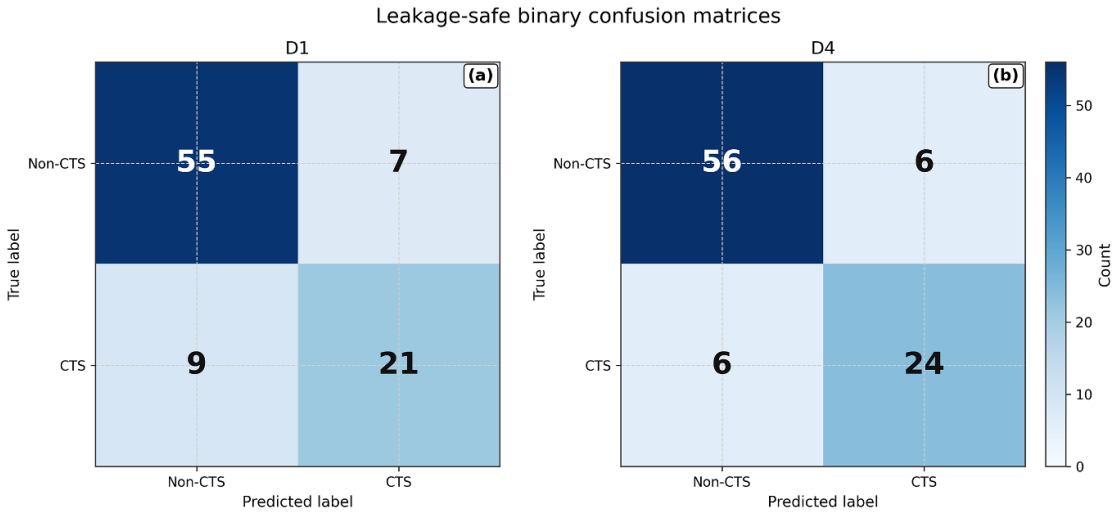


Figure 6. Leakage-safe binary confusion matrices for the best-performing Dataset 1 and Dataset 4 models. Panel (a) shows the Dataset 1 XGB+LGBM ensemble, and panel (b) shows the Dataset 4 Hybrid_FrozenNestedEWC+ML model. Rows correspond to the true classes and columns to the predicted classes. The binary task included 62 non-CTS hands and 30 CTS hands. The Dataset 4 model produced fewer false negatives and fewer false positives than the Dataset 1 baseline.

(0.745–0.940). The improvement was larger in the binary task than in the four-class task, although Dataset 4 retained a modest advantage in the four-class setting. The wider intervals and the smaller point-estimate changes indicate that four-class severity staging remains more uncertain than binary CTS screening in the present sample. To make the uncertainty around the primary performance estimates explicit, bootstrap 95% confidence intervals for the main Dataset 1 and Dataset 4 comparisons are summarized in table 4. These intervals were calculated from the out-of-fold hand-level predictions and therefore reflect the variability expected under the leakage-safe evaluation protocol. The binary point estimates consistently favored Dataset 4, with the clearest changes observed for sensitivity and MCC. The confidence intervals also show the uncertainty expected from the modest 92-hand dataset. For the four-class severity task, the differences between Dataset 1 and Dataset 4 were smaller, so these results are interpreted as supportive and exploratory rather than as definitive evidence of improved severity staging.

3.2. Effect of the evaluation protocol

To examine the effect of protocol choice, Dataset 1 and a fixed late-fusion Dataset 4 baseline were compared under leakage-safe and optimistic evaluation protocols (figure 8(a)).

Under the leakage-safe protocol, the Dataset 1 clinical-only model achieved 79.3% accuracy, 73.3% sensitivity, 82.3% specificity, an MCC of 0.543, and a ROC–AUC of 0.786. The corresponding Dataset 4 late-fusion baseline achieved 80.4% accuracy, 73.3% sensitivity, 83.9% specificity, an MCC of 0.563, and a ROC–AUC of 0.801.

When the same comparison was repeated under the optimistic row-level protocol, the apparent performance increased substantially. The Dataset 1 model increased to 82.6% accuracy, 78.7% sensitivity, 84.5% specificity, an MCC of 0.617, and a ROC–AUC of 0.911, whereas the Dataset 4 late-fusion baseline increased to 84.3% accuracy, 78.0% sensitivity, 87.4% specificity, an MCC of 0.648, and a ROC–AUC of 0.913. As shown in figure 8(b), the ROC curve under the optimistic protocol was displaced upward relative to the leakage-safe protocol, indicating higher apparent performance under non-independent splitting.

3.3. Secondary analyses

The classical machine-learning block was examined separately to determine whether the gains obtained with Dataset 4 were uniform across the tabular models. The observed changes were not uniform across models. In the binary setting, the largest MCC improvement was observed for LGBM (ΔMCC $= +0.056$), whereas XGB and XGB+RF showed only marginal gains (ΔMCC $= +0.006$ for both). For XGB+LGBM, a decrease in MCC was observed (ΔMCC $= -0.022$). ROC–AUC decreased slightly for all four classical binary models. Overall, the same-architecture comparison indicates that the kinematic block contributed differently across classical learners. The LGBM result supports a directional benefit from adding kinematic descriptors to the symptom-based feature set, whereas the smaller changes in the other classical models suggest that this information is most useful when the model architecture can

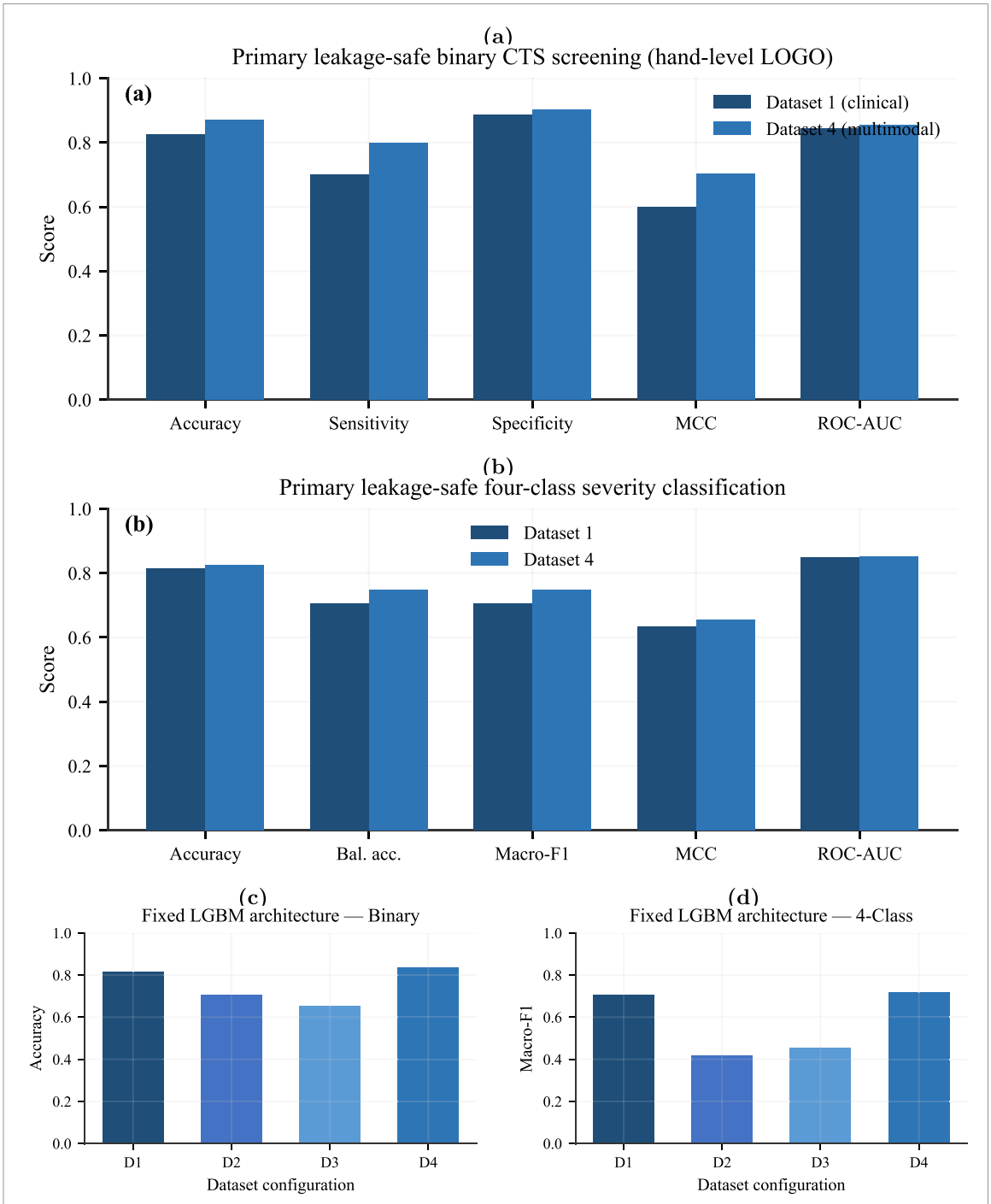


Figure 7. Primary and ablation-based leakage-safe classification results. Panels (a) and (b) show the best Dataset 1 and Dataset 4 models for the binary CTS screening and four-class severity tasks, respectively. Panels (c) and (d) show all four predefined dataset configurations under the same fixed LGBM architecture. Bars show point estimates; bootstrap 95% confidence intervals for the primary Dataset 1 and Dataset 4 comparisons are reported in the Results text and table 4. The primary Dataset 4 model showed the strongest binary screening profile, whereas the fixed-architecture ablation shows that the added kinematic information was most useful when combined with the clinical/symptom block rather than used alone.

integrate the clinical and motion-derived patterns effectively. The same-architecture MCC comparison is shown in figure 9.

A discordant-pair comparison was also performed for the binary classical models. The number of discordant cases was small in all comparisons, and the exact McNemar p-values ranged from 0.50 to 1.00 across the tested models. More specifically, the comparison yielded $b = 2$, $c = 2$, and $p = 1.00$ for XGB; $b = 0$, $c = 2$, and $p = 0.50$ for LGBM; $b = 2$, $c = 2$, and $p = 1.00$ for XGB+RF; and $b = 2$, $c = 1$, and $p = 1.00$ for XGB+LGBM. The paired tests therefore suggest a cautious interpretation. The classical same-architecture analysis helps to show the direction of change and the possible role of the added kinematic variables, but it does not prove that Dataset 4 performed better under all models. This is also in

Table 4. Bootstrap 95% confidence intervals for the primary leakage-safe Dataset 1 and Dataset 4 comparisons.

Task	Metric	Dataset 1	Dataset 4
Binary	Accuracy	0.826 (0.750–0.902)	0.870 (0.793–0.935)
Binary	Sensitivity	0.700 (0.544–0.851)	0.800 (0.647–0.934)
Binary	Specificity	0.887 (0.803–0.956)	0.903 (0.831–0.968)
Binary	MCC	0.598 (0.413–0.784)	0.703 (0.547–0.856)
Binary	ROC–AUC	0.845 (0.745–0.934)	0.855 (0.752–0.945)
Four-class	Accuracy	0.815 (0.728–0.891)	0.826 (0.739–0.902)
Four-class	Balanced accuracy	0.706 (0.546–0.846)	0.748 (0.597–0.871)
Four-class	Macro-*F*1	0.706 (0.544–0.823)	0.748 (0.588–0.850)
Four-class	MCC	0.634 (0.471–0.785)	0.656 (0.500–0.800)
Four-class	ROC–AUC	0.849 (0.752–0.923)	0.851 (0.745–0.940)

Values are point estimates with 95% bootstrap confidence intervals in parentheses. Confidence intervals were obtained by hand-level bootstrap resampling of out-of-fold predictions.

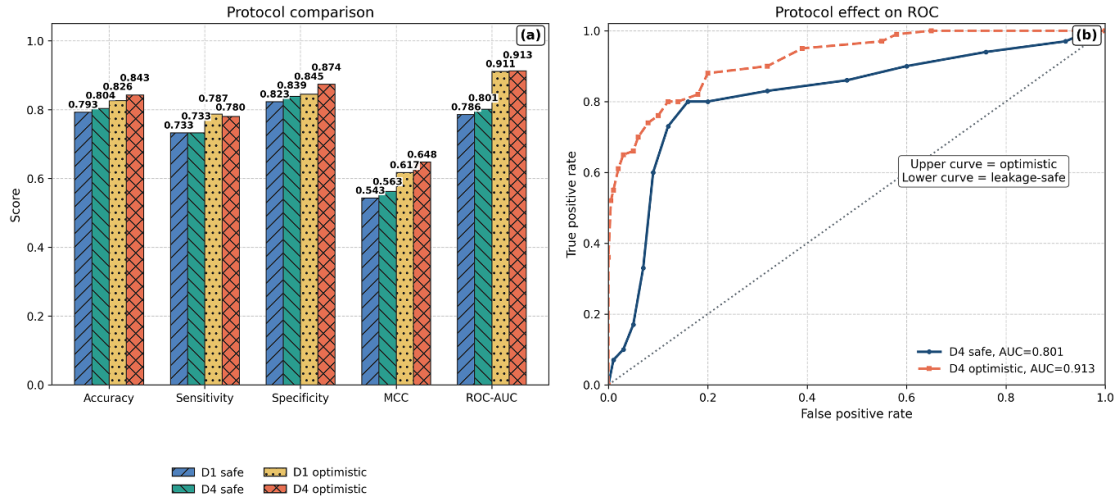


Figure 8. Effect of evaluation protocol on binary performance. Panel (a) compares Dataset 1 and Dataset 4 under leakage-safe and optimistic evaluation protocols using a fixed late-fusion baseline. Panel (b) shows the ROC curves for the Dataset 4 late-fusion baseline under the same two protocols. The optimistic protocol yields consistently higher performance estimates than the grouped hand-level evaluation.

line with the primary analysis, where the best result was obtained when clinical/symptom variables and motion-derived descriptors were used together in the strongest multimodal setting. Because both hands were available for each participant, bilateral label concordance was also examined. Binary labels were found to be concordant in 40 of 46 participants (87.0%), and four-class severity labels were found to be concordant in 38 of 46 participants (82.6%). This supports the possibility that the two hands of the same participant may not be statistically independent.

A participant-level leave-one-participant-out sensitivity analysis was therefore performed using LGBM. More conservative estimates than the primary hand-level protocol were produced by this analysis. In the binary task, 0.630 accuracy, 0.633 sensitivity, 0.629 specificity, and 0.628 ROC–AUC were achieved by Dataset 1, whereas 0.467 accuracy, 0.233 sensitivity, 0.581 specificity, and 0.380 ROC–AUC were achieved by Dataset 4. These results are reported as a sensitivity analysis rather than as the primary performance estimate, because the clinical label and screening target in the present study were defined at the hand level.

A complete model-level comparison was also added to show whether the headline hybrid result reflected a single isolated model variant. In the binary task, several Dataset 4 hybrid configurations reached the same highest accuracy of 0.870, including Hybrid_NestedGestureNet+ML, Hybrid_FrozenNestedEWC+ML, and Hybrid_NestedGestureNet+LGBM. The fixed Dataset 4 LGBM model also remained competitive, reaching 0.837 accuracy. In the four-class task, the best result was

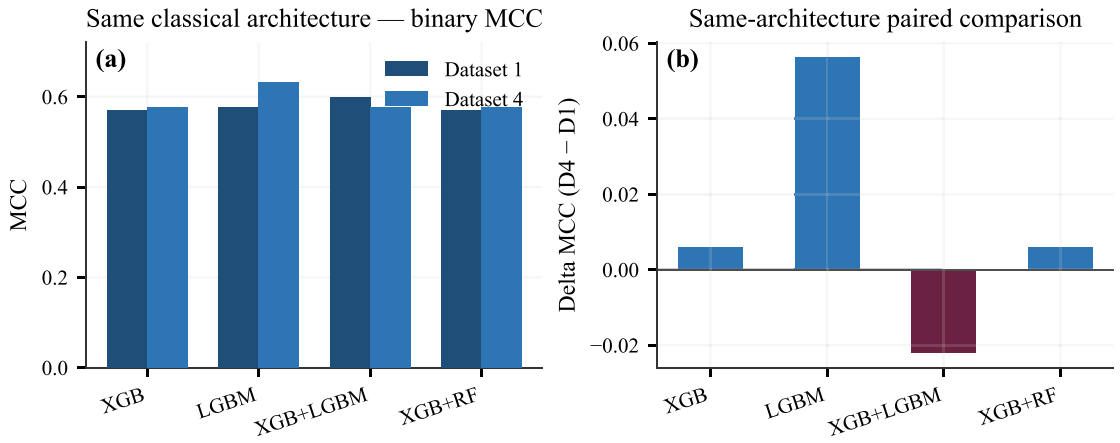


Figure 9. Same-architecture binary comparison between Dataset 1 and Dataset 4 using classical machine-learning models. Panel (a) shows MCC values for both datasets under identical model families. Panel (b) shows ΔMCC calculated as Dataset 4 minus Dataset 1. Positive values indicate higher MCC after adding the kinematic block to the clinical/symptom variables. The largest directional improvement was observed for LGBM, supporting an architecture-dependent contribution of the kinematic descriptors.

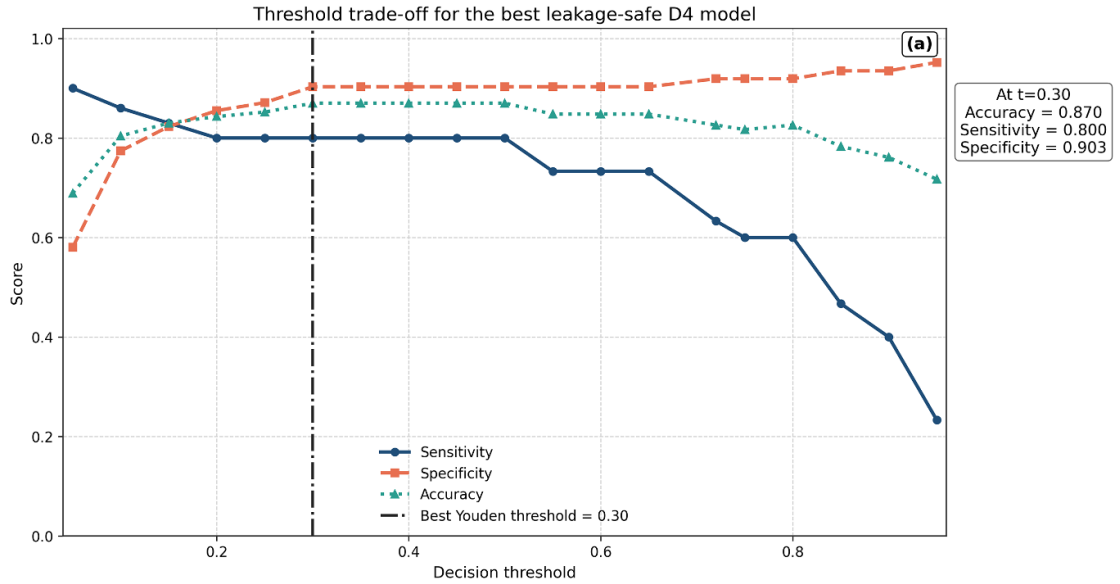


Figure 10. Threshold trade-off analysis for the best leakage-safe Dataset 4 binary model. The vertical line marks the Youden-optimal threshold at 0.30. Around this operating point, sensitivity and specificity remain well balanced, whereas higher thresholds improve specificity at the cost of a substantial decline in sensitivity.

obtained with Hybrid_NestedGestureNet+LGBM, with 0.826 accuracy and 0.748 macro-*F*1, while several other Dataset 4 models remained close to the Dataset 1 baseline. The complete per-model results are provided in Supplementary table S.

Finally, threshold analysis was performed for the best leakage-safe Dataset 4 binary model (figure 10). A relatively stable operating region was observed between thresholds of approximately 0.30 and 0.40. At the Youden-optimal threshold of 0.30, accuracy remained 87.0%, while sensitivity and specificity were balanced at 80.0% and 90.3%, respectively.

At higher thresholds, specificity increased further, but this occurred at the expense of a marked decline in sensitivity. These results show that thresholds below 0.50 yielded a more balanced trade-off between sensitivity and specificity in binary CTS screening.

3.4. Statistical feature-level findings

To complement the classification results, stage-related differences in selected clinical and kinematic variables were examined using the non-parametric framework described in section 2.3. Regional symptom scores obtained after the Phalen test showed a clear stage-dependent pattern, with the most consistent

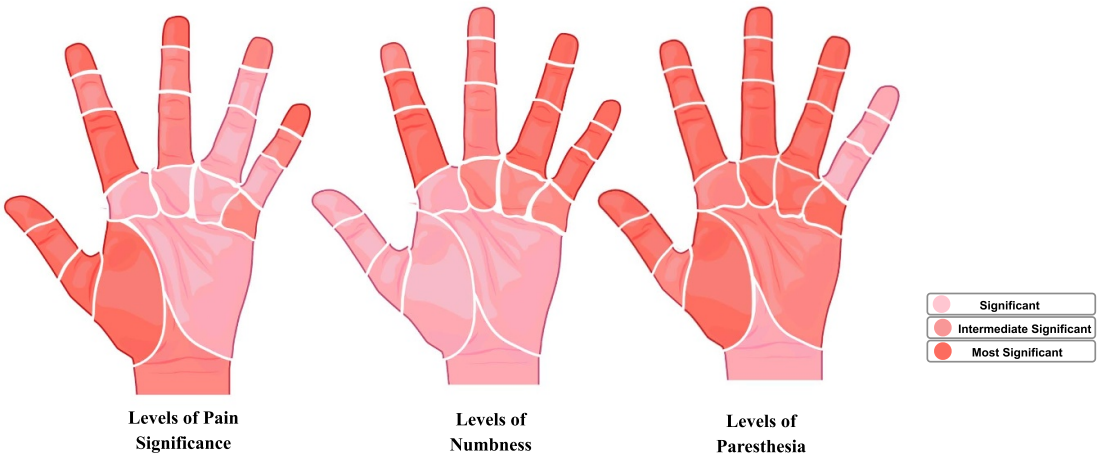


Figure 11. Regional symptom maps after the Phalen test showing the relative strength of stage-related differences for pain, numbness, and paresthesia across the predefined hand areas. Shading reflects the four-stage Kruskal–Wallis screening followed by Dunn post-hoc comparisons across all available hand-level observations. Darker shading indicates stronger evidence of between-stage separation. The representative coordinates summarized in table 5 are a separate stage 0 versus stage 3 summary of the strongest retained kinematic contrasts.

Table 5. Representative coordinates from the strongest stage 0 vs stage 3 contrasts. Only coordinates retained after the Kruskal–Wallis/Dunn screening and exceeding the descriptive separation threshold of 0.20 are shown. The separation values are reported as generated by the analysis pipeline and are included only to rank the relative strength of the retained contrasts.

Finger	Coordinate	Group comparison	Separation
Index	Index-thumb landmark 14 (*z*-axis)	0 vs 3	0.6511
	Index-thumb landmark 13 (*y*-axis)	0 vs 3	0.6460
	Index-thumb landmark 12 (*z*-axis)	0 vs 3	0.6000
Middle	Middle-thumb landmark 0 (*x*-axis)	0 vs 3	0.3445
	Middle-thumb landmark 12 (*y*-axis)	0 vs 3	0.2018
Little	**Little-thumb landmark 14** (*z*-axis)	**0 vs 3**	**1.0000**
	Little-thumb landmark 17 (*z*-axis)	0 vs 3	0.8187
	Little-thumb landmark 11 (*z*-axis)	0 vs 3	0.5569
	Little-thumb landmark 1 (*x*-axis)	0 vs 3	0.5149
	Little-thumb landmark 0 (*z*-axis)	0 vs 3	0.2540
	Little-thumb landmark 17 (*y*-axis)	0 vs 3	0.2110

differences appearing in the median-nerve-dominant thumb region. As shown in figure 11, pain, numbness, and paresthesia were not distributed uniformly across the hand; instead, the strongest between-stage contrasts were concentrated in clinically meaningful sensory regions.

The kinematic variables showed a similar tendency. The largest pairwise contrasts were observed between stage 0 and stage 3, particularly for index-thumb and little-thumb opposition coordinates, most often along the *z*-axis. Additional differences were also observed for selected middle-thumb coordinates, suggesting that both symptom distribution and thumb-related movement characteristics contribute to stage separation. Representative coordinates from the strongest contrasts are summarized in table 5.

The feature-level results help explain the improvement seen in the best Dataset 4 model. The main retained differences were found in thumb-opposition coordinates, particularly little-thumb and index-thumb movements, which are functionally related to median nerve involvement in CTS. Therefore, the kinematic variables appear to add stage-related functional information to the regional symptom scores, rather than only increasing the number of model inputs.

4. Discussion

This study evaluated whether hand kinematics extracted from low-cost optical camera recordings can provide clinically meaningful information for CTS screening and severity assessment when combined with conventional clinical variables. The main result was that the multimodal approach showed its clearest benefit in the binary classification setting. Under leakage-safe hand-level evaluation, the best Dataset 4 model showed higher point estimates for accuracy, sensitivity, specificity, MCC, and ROC–AUC than the best Dataset 1 baseline, with bootstrap confidence intervals reported to show the uncertainty around these estimates. The confidence intervals and same-architecture analyses indicate that the magnitude of this gain should be interpreted cautiously in this modest single-center dataset. In contrast, the improvement in the four-class severity task was smaller. Taken together, these findings suggest that camera-derived motion features are most informative for distinguishing CTS from non-CTS hands, whereas finer stage separation remains more difficult.

The stronger effect in the binary setting is clinically plausible. The recorded opposition gestures directly involve motor functions that depend on median nerve integrity and thenar muscle performance. CTS is known to affect thumb opposition, palmar abduction, and coordinated fine motor control, even when such deficits are not striking on routine visual inspection (Fernandes *et al* 2013, Kuroiwa *et al* 2018, Sonoo *et al* 2018, Saggar *et al* 2024). In the present study, these motion characteristics were not represented only by raw landmark coordinates, but also by derived descriptors based on distances, joint relationships, temporal variation, and cross-gesture summaries. The improvement observed in the best multimodal binary model, together with the feature-level separation of selected thumb-opposition coordinates, suggests that the optical recordings added functional information that was not fully captured by symptom reports alone.

From a practical perspective, the binary results are the most encouraging. The best leakage-safe Dataset 4 model increased sensitivity from 70.0% to 80.0% while also slightly improving specificity. For a screening-oriented application, this is an important pattern, because a useful prediagnostic tool should identify more affected hands without causing a disproportionate increase in false positives. In the present study, this balance was reflected not only in the summary metrics but also in the confusion matrices, where the multimodal model reduced both false negatives and false positives relative to the clinical baseline. This suggests that the added motion information was not merely shifting the decision boundary in one direction, but improving overall discrimination.

The more modest gains in the four-class analysis should be interpreted carefully. Severity grading in CTS is inherently more difficult than binary discrimination because stage boundaries are less distinct and may reflect overlapping clinical and electrophysiological profiles rather than sharply separated motor phenotypes (Park *et al* 2021, Sasaki *et al* 2022). The fact that Dataset 4 still showed consistent, though limited, improvements across accuracy, balanced accuracy, macro-$F1$, MCC, and ROC–AUC indicates that the kinematic features retained some severity-related information. However, the current data do not support a strong claim that optical motion analysis alone can resolve detailed stage classification with high precision. In its present form, the approach appears better suited to screening and broad stratification than to replacing established methods for fine-grained grading.

These results are best interpreted in the context of the existing literature. Established CTS pathways rely heavily on clinical examination, nerve conduction studies, and, in selected cases, imaging (Wipperman and Goerl 2016, Tüfekçi and Boz 2022, Bakalis *et al* 2024). Other studies have explored more portable strategies, including tablet-based screening, smartphone-based anomaly detection, and video-based movement analysis (Fujita *et al* 2019, Koyama *et al* 2021, Watanabe *et al* 2021, Tsukamoto *et al* 2024). As summarized in table 6, the present study differs from many earlier approaches by combining symptom-based clinical information with low-cost, non-contact optical kinematics obtained from multiple opposition gestures, and by evaluating the resulting models under a leakage-safe hand-level protocol. The present results do not suggest that this framework replaces electrodiagnostic assessment or specialist examination. Rather, they support its role as a practical adjunct for screening and broad functional stratification, especially when a low-cost and non-invasive workflow is desirable.

A particularly important observation in this study was the magnitude of the protocol effect. When the same late-fusion baseline was evaluated under an optimistic row-level split, the apparent performance increased substantially and the ROC curve shifted upward compared with the leakage-safe grouped evaluation. This finding has methodological importance beyond the present dataset. In repeated-measure biomedical datasets, samples originating from the same hand or participant can easily inflate model estimates if grouping is ignored. For this reason, the leakage-safe hand-level protocol used here should be considered the more credible basis for interpretation. The lower but more realistic

Table 6. Compact comparative overview of CTS diagnostic approaches.

Study	Approach / device	Key result	Main value / limitation
(Wipperman and Goerl 2016)	Clinical tests; EMG/US/MRI support	Flick: 93% sens., 96% spec.	Broad clinical overview; requires equipment and expert interpretation.
(Tüfekçi and Boz 2022)	EMG thresholding	94.2% sens., 90% spec.	Useful for progression assessment; invasive and clinic-based.
(Sasaki *et al* 2022)	Review of EMG grading	$r = -0.786$	Shows inconsistency in current grading; no new method proposed.
(Öten *et al* 2026)	ML on EMG signals	High accuracy	Automates EMG-based diagnosis; still EMG-dependent.
(Tsamis *et al* 2021)	ML electrodiagnosis with EMG	95% neurophys.; 89% clinical	May reduce observer error; remains invasive and sensor-dependent.
(Kuroiwa *et al* 2018)	Thumb-angle analysis with gyroscope	Reduced motion in CTS	Objective thumb-motion metric; needs extra hardware.
(Fujita *et al* 2019)	tablet-based SVM	93% sens., 73% spec.	Portable screening tool; moderate specificity.
(Watanabe *et al* 2021)	Spiral drawing + SVM	82% sens., 71% spec.	Simple early screening; depends on drawing skill.
(Koyama *et al* 2021)	Smartphone anomaly detection	94% sens., 67% spec.; AUC 0.86; specific AUC 0.99	Accessible mobile screening; low specificity in the general model.
(Park *et al* 2021)	Severity ML using clinical + EMG	76%–91% acc.	Highlights pain and BMI; no fine-motor motion data.
(Tsukamoto *et al* 2024)	Grip video via smartphone app	89% sens., 83% spec.; $r = 0.68$	No extra hardware; grip-only input and no staging.
This study	**Multimodal ML with joint kinematics + symptoms; USB/web camera + app**	**2-class: 87.0% acc., 80.0% sens., 90.3% spec., AUC 0.855; 4-class: 82.6% acc., bal. acc. 0.748, macro-*F*1 0.748, AUC 0.851**	**Low-cost, non-contact CTS screening and broad severity assessment from multiple opposition gestures; limited by single-center data, modest sample size, and need for external validation.**

Abbreviations: sens., sensitivity; spec., specificity; acc., accuracy; bal. acc., balanced accuracy; AUC, area under the curve.

values reported under this protocol are more informative for translation than inflated estimates derived from non-independent splitting.

The threshold analysis adds another clinically relevant dimension to the main findings. The best leakage-safe Dataset 4 binary model showed a relatively stable operating region between thresholds of approximately 0.30–0.40, with the Youden-optimal point at 0.30. At this threshold, sensitivity and specificity remained well balanced. This is relevant because screening tools are rarely judged by accuracy alone. In practical use, the operating threshold needs to reflect the relative cost of missed cases and unnecessary referrals. The present results suggest that the default 0.50 threshold is not necessarily the most appropriate choice for a screening-oriented CTS application and that a lower threshold may offer a more useful balance when sensitivity is clinically prioritized. The statistical follow-up analyses were broadly consistent with the model-level findings. Stage-related differences were concentrated in the median-nerve sensory territory and in thumb-opposition kinematics, especially for index-thumb and little-thumb movements. This pattern suggests that the added value of the optical recordings was not merely numerical; rather, the motion-derived variables appear to capture functionally meaningful abnormalities that are aligned with the clinical phenotype of CTS. The same-architecture analyses put this result into a more cautious context. They indicate that the kinematic variables were not superior in every classical model; their value depended on how they were combined with the symptom variables. This is reasonable, because the symptom scores describe the sensory distribution after the Phalen test, whereas the optical measures describe thumb-opposition function. In the best multimodal model, using both sources increased sensitivity and reduced false negatives, suggesting that the screening representation gained clinically useful information.

Several limitations should be considered when interpreting these results. First, the study was conducted at a single center and included a relatively limited number of participants and hand-level groups. A further limitation is the possible dependence between the two hands of the same participant. Hand-level grouping was used in the primary analysis because the clinical label and the screening target were defined at the hand level. However, similar binary or severity labels were observed in both hands of many participants in the bilateral concordance audit, and more conservative estimates were obtained in the participant-level sensitivity analysis. For this reason, validation at the participant level in a larger cohort is needed before clinical use. The findings therefore require confirmation in larger and more heterogeneous populations. A further limitation is related to the potential dependence between the two hands of the same participant. In the primary analysis, hand-level grouping was used because the clinical label and the screening target were defined at the hand level. However, the bilateral concordance audit indicated that the same binary or severity label was present in both hands for many participants. In addition, more conservative estimates were obtained in the participant-level sensitivity analysis. Therefore, participant-level validation in a larger cohort is required before clinical deployment can be considered. In addition, contact pressure during opposition was not quantified, and the recordings were not obtained under a fully standardized motion-capture setup. Future studies should test whether pressure sensing, stricter camera positioning, or multi-view recordings improve robustness. Second, the optical representation was based on 2.5D landmark estimates rather than fully calibrated three-dimensional motion capture. Although this choice supports simplicity and low cost, landmark quality may still be influenced by occlusion, lighting conditions, hand orientation, and camera placement (Simon *et al* 2017, Zimmermann and Brox 2017, Zhang *et al* 2020). It should also be noted that thumb–finger contact pressure was not quantified, and the recordings were acquired without a fully standardized motion-capture setup. In future work, whether robustness can be improved through pressure sensing, stricter camera positioning, or multi-view recording should be examined. Third, the four-class problem remains difficult, and the modest improvement in that setting indicates that broader datasets or alternative feature representations may be necessary for more reliable staging. Finally, because the present analysis was cross-sectional, it cannot show whether the kinematic markers remain stable over time or change after treatment. This point should be examined in follow-up studies, especially since electrophysiological findings in untreated CTS may also vary during the course of the disease (Benli Küçük 2023).

Future work should focus on external validation across different centers, recording conditions, and patient groups. Longitudinal recordings would be particularly valuable for determining whether camera-derived kinematic features track progression, recovery, or response to treatment. It would also be useful to test whether the approach can be simplified into a brief and robust workflow suitable for outpatient triage, telemedicine-supported assessment, or settings with limited access to electrodiagnostic testing (Alenoghena *et al* 2023, Shaik *et al* 2023). Technology-supported follow-up has also been studied in CTS. For example, Sidek and Mohamed (2022) proposed an IoT-based rehabilitation monitoring approach, showing that such systems may be useful not only for diagnosis but also for rehabilitation follow-up; however, they may still depend on additional sensors or dedicated infrastructure. With further validation, a low-cost optical assessment of this kind could become a practical adjunct to existing CTS pathways by providing rapid, non-invasive, and functionally relevant information before confirmatory testing is considered.

5. Conclusion

In conclusion, this study shows that hand kinematics obtained from low-cost optical camera recordings can provide useful complementary information for machine learning-based screening of CTS when combined with routine clinical variables. The clearest contribution of the multimodal approach was seen in binary CTS detection under leakage-safe hand-level evaluation. By comparison, only a limited improvement was observed for four-class severity classification. In addition, it was revealed that the evaluation protocol itself has a considerable impact on the resulting performance estimates, which highlights the necessity of group-aware validation in repeated-measure biomedical datasets. These findings suggest that non-contact optical kinematic analysis could be useful as a supportive tool for CTS screening and general functional classification. However, it should not be seen as a substitute for conventional electrodiagnostic assessment. Before routine clinical use can be considered, this approach should be evaluated in larger multicenter cohorts, with longitudinal follow-up and participant-level validation.

Data availability statement

The data cannot be made publicly available upon publication because they contain sensitive personal information. The data that support the findings of this study are available upon reasonable request from the authors.

Supplementary model results available at: https://doi.org/10.1088/1873-4030/ae8dce/data1.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Ethical approval and consent to participate

The study protocol was approved by the İstanbul University—Cerrahpaşa Non-Interventional Clinical Research Ethics Committee (Approval No. 2024/174; 17 November 2024). All procedures were conducted in accordance with the Declaration of Helsinki. Written informed consent was obtained from all participants.

CRediT authorship contribution statement

Şerife Çelikbaş: Conceptualization, Methodology, Data curation, Software, Writing—original draft, Visualization. **Abdurrahim Akgündoğdu**: Formal analysis, Investigation, Supervision, Validation, Writing—review & editing. **Ayşegül Gündüz**: Supervision, Validation, Writing—review & editing.

ORCID iDs

Şerife Çelikbaş 0000-0001-6118-9335
Abdurrahim Akgündoğdu 0000-0001-8113-0277
Ayşegül Gündüz 0000-0003-2365-0850

References

Afshar A, Tabrizi A and Shariyate M J 2025 Autonomic component of carpal tunnel syndrome *J. Hand Surg. Am.* **50** 491–6
Alenoghena C O, Ohize H O and Adejo A O 2023 Telemedicine: a survey of telecommunication technologies, developments and challenges *J. Sens. Actuator Netw.* **12** 20
Ali Z A, Abduljabbar Z H, Taher H A, Sallow A B and Almufti S M 2023 Exploring the power of extreme gradient boosting algorithm in machine learning: a review *Acad. J. Nawroz Univ.* **12** 320–34
Altalhan M, Algarni A and Alouane M T-H 2025 Imbalanced data problem in machine learning: a review *IEEE Access* **13** 13686–99
Bakalis D, Kontogiannis P, Ntais E, Simos Y V, Tsamis K I and Manis G 2024 Carpal tunnel syndrome automated diagnosis: a motor vs. sensory nerve conduction-based approach *Bioengineering* **11** 175
Benli Küçük E 2023 Temporal changes in electrophysiological parameters in untreated patients with carpal tunnel syndrome *Cureus* **15** e46039
Bentéjac C, Csörgő A and Martínez-Muñoz G 2021 A comparative analysis of gradient boosting algorithms *Artif. Intell. Rev.* **54** 1937–67
Brown M B and Forsythe A B 1974 Robust tests for the equality of variances *J. Am. Stat. Assoc.* **69** 364–7
Bühlmann P and Yu B 2010 Boosting *WIREs Comput. Stat.* **2** 69–74
Elseddik M *et al* 2023 Deep learning-based approaches for enhanced diagnosis and comprehensive understanding of carpal tunnel syndrome *Diagnostics* **13** 3211
Fernandes C H, Meirelles L M, Neto J R, Nakachima L R, Santos J B D and Faloppa F 2013 Carpal tunnel syndrome with thenar atrophy: evaluation of the pinch and grip strength in patients undergoing surgical treatment *Hand* **8** 60–63
Fujita K, Watanabe T, Kuroiwa T, Sasaki T, Nimura A and Sugiura Y 2019 A tablet-based app for carpal tunnel syndrome screening: diagnostic case-control study *JMIR mhealth uhealth* **7** e14172

Gastwirth J L, Gel Y R and Miao W 2009 The impact of Levene's test of equality of variances on statistical theory and practice *Stat. Sci.* **24** 343–60
George D and Mallery P 2010 *SPSS for Windows Step by Step: a Simple Guide and Reference, 17.0 Update* 10th edn Allyn & Bacon)
Gopal S, Minns C J and Kumaran B 2024 Diagnostic test accuracy of Tinel's test in adults with carpal tunnel syndrome—a systematic review *Phys. Ther. Rev.* **29** 222–33
Kamiński B, Jakubczyk M and Szufel P 2018 A framework for sensitivity analysis of decision trees *Cent. Eur. J. Oper. Res.* **26** 135–59
Kirkpatrick J *et al* 2017 Overcoming catastrophic forgetting in neural networks *Proc. Natl Acad. Sci. USA* **114** 3521–6
Koyama T, Sato S, Toriumi M, Watanabe T, Nimura A, Okawa A, Sugiura Y and Fujita K 2021 A screening method using anomaly detection on a smartphone for patients with carpal tunnel syndrome: diagnostic case-control study *JMIR mhealth uhealth* **9** e26320
Kruskal W H and Wallis W A 1952 Use of ranks in one-criterion variance analysis *J. Am. Stat. Assoc.* **47** 583–621
Kuroiwa T, Fujita K, Nimura A, Miyamoto T, Sasaki T and Okawa A 2018 A new method of measuring the thumb pronation and palmar abduction angles during opposition movement using a three-axis gyroscope *J. Orthop. Surg. Res.* **13** 288
Lokker C *et al* 2024 Boosting efficiency in a clinical literature surveillance system with LGBM *PLoS Digit. Health* **3** e0000299
Luo Y T, Huang Y T, Chiu V, Chang Y W and Horng Y S 2025 Diagnostic meta-analysis of the efficacy of ultrasonography for diagnosing carpal tunnel syndrome: a comparison between Asian and non-Asian populations *J. Formos. Med. Assoc.* **124** 666–74
Öten E, Aygün Bilecik N and Uğur L 2026 Use of machine learning methods in diagnosis of carpal tunnel syndrome *Comput. Methods Biomech. Biomed. Eng.* **29** 838–48
Park D *et al* 2021 Machine learning-based approach for disease severity classification of carpal tunnel syndrome *Sci. Rep.* **11** 17464
Quinlan J R 1996 Bagging, boosting and C4.5 *Proc. National Conf. Artificial Intelligence (AAAI)* (available at: https://cdn.aaai.org/AAAI/1996/AAAI96-108.pdf)
Rotaru-Zavaleanu A D, Lungulescu C V, Bunescu M G, Vasile R C, Gheorman V and Gresita A 2024 Occupational carpal tunnel syndrome: a scoping review of causes, mechanisms, diagnosis and intervention strategies *Front. Public Health* **12** 1407302
Saggar S K, Thaman R G, Mohan G and Kumar D 2024 Mapping neurophysiological patterns in carpal tunnel syndrome: correlations with Tinel's and Phalen's signs *Cureus* **16** e58168
Sasaki T, Koyama T, Kuroiwa T, Nimura A, Okawa A, Wakabayashi Y and Fujita K 2022 Evaluation of the existing electrophysiological severity classifications in carpal tunnel syndrome *J. Clin. Med.* **11** 1685
Shaik T, Tao X, Higgins N, Li L, Gururajan R, Zhou X and Acharya U R 2023 Remote patient monitoring using artificial intelligence: current state, applications and challenges *WIREs Data Min. Knowl. Discov.* **13** e1485
Shapiro S S and Wilk M B 1965 An analysis of variance test for normality (complete samples) *Biometrika* **52** 591–611
Sidek M A A and Mohamed M, 2022 IoT-based rehabilitation monitoring system for carpal tunnel syndrome *Evol. Electr. Electron. Eng.* **3** 537–46 (available at: https://publisher.uthm.edu.my/periodicals/index.php/eeee/article/view/8619)
Silva A R, Adao T and Macedo M P 2024 Carpal tunnel syndrome with thenar atrophy: alternative methods for early diagnosis and surgical treatment follow-up *Proc. Comput. Sci.* **239** 816–23
Simon T, Joo H, Matthews I and Sheikh Y 2017 Hand keypoint detection in single images using multiview bootstrapping (arXiv:1704.07809)
Sonoo M, Menkes D L, Bland J D P and Burke D 2018 Nerve conduction studies and EMG in carpal tunnel syndrome: do they add value? *Clin. Neurophysiol. Pract.* **3** 78–88
Stevens J C, Beard C M, O'Fallon W M and Kurland L T 1992 Conditions associated with carpal tunnel syndrome *Mayo Clin. Proc.* **67** 541–8
Tsamis K I, Kontogiannis P, Gourgiotis I, Ntabos S, Sarmas I and Manis G 2021 Automatic electrodiagnosis of carpal tunnel syndrome using machine learning *Bioengineering* **8** 181
Tsukamoto K, Matsui R, Sugiura Y and Fujita K 2024 Diagnosis of carpal tunnel syndrome using a 10-s grip-and-release test with video and machine learning analysis *J. Hand Surg. Eur* **49** 634–6
Tüfekçi A and Boz C 2022 Electrophysiological evaluation of carpal tunnel syndrome *Arch. Med. Sci.* (https://doi.org/10.5114/aoms/153468)
Wang Y *et al* 2025 Combination of morphological and multiparametric MR neurography enhances carpal tunnel syndrome diagnosis and evaluation *Sci. Rep.* **15** 184
Watanabe T, Koyama T, Yamada E, Nimura A, Fujita K and Sugiura Y 2021 The accuracy of a screening system for carpal tunnel syndrome using hand drawing *J. Clin. Med.* **10** 4437
Wiesman I M, Novak C B, Mackinnon S E and Winograd J M 2003 Sensitivity and specificity of clinical testing for carpal tunnel syndrome *Can. J. Plast. Surg.* **11** 70–72
Wipperman J and Goerl K 2016 Carpal tunnel syndrome: diagnosis and management *Am. Fam. Phys.* **94** 993–9 (available at: www.aafp.org/pubs/afp/issues/2016/1215/p993)
Zhang F *et al* 2020 Mediapipe hands: on-device real-time hand tracking (arXiv:2006.10214)
Zimmermann C and Brox T 2017 Learning to estimate 3D hand pose from single RGB images (arXiv:1705.01389)