scientific reports

Check for updates

OPEN

A deep learning approach to remotely assessing essential tremor with handwritten images

Yumeng Peng[1,4,5,7], Songliang Han[1,7], Di Wu[2], Chenbin Ma[2], Zijing Zeng[5], Ping He[6], Tian Yuan[1,5], Ying Shi[1,5], Lixuan Li[1], Wenjing Yang[1], Longsheng Pan[3]✉ & Zhengbo Zhang[1]✉

Essential tremor (ET) is the most prevalent movement disorder, with its incidence increasing with age, significantly impacting motor functions and quality of life. Traditional methods for assessing ET severity are often time-consuming, subjective, and require in-person visits to medical facilities. This study introduces a novel deep learning-based approach for remotely assessing ET severity using handwriting images, which improves both efficiency and accessibility. We collected approximately 1000 high-quality Archimedean spiral handwriting images from patients in both medical institutions and home settings, creating a robust and diverse dataset. A transfer learning-based model, ETSD-Net, was developed and trained to evaluate ET severity. The model achieved an accuracy of 88.44%, demonstrating superior performance over baseline models. Our approach offers a cost-effective, scalable, and reliable solution for ET assessment, particularly in remote or resource-limited settings, and provides a valuable contribution to the development of more accessible diagnostic tools for movement disorders.

Keywords Essential tremor, Deep learning, Handwriting, Archimedean spiral, Transfer learning, Remotely assessing

Essential tremor (ET) is the most common movement disorder, characterized by rhythmic tremors in the hands, head, or other parts of the body during voluntary movements[1]. It affects an estimated 60 million people worldwide and is expected to rise as the population ages[2]. Regular ET assessments can help monitor disease progression, detect deterioration or treatment response, and adjust management plans accordingly.

Currently, assessing ET is challenging due to the lack of specific tests or biomarkers[3,4], typically relying on clinical observation during face-to-face interactions[5]. Clinicians evaluate tremors based on the clinical rating scale for tremor (CRST)[6] through tasks such as hand-to-nose movements, drinking water, or drawing[3]. Spiral drawing is a commonly used clinical method for assessing ET[7]. The patient is provided with a paper containing a pre-drawn Archimedean spiral (with guideline templates). Two points are marked at the center and the outer edge of the spiral. The patient is instructed to connect the two points without crossing the guideline. Clinicians evaluate the severity of tremor visually based on the patient's drawing (as shown in Table 1). This method is widely used for its simplicity and practicality in clinical settings[8,9].

However, traditional assessment methods require patients to visit medical institutions, schedule appointments with neurologists, and undergo in-person evaluations. Many ET patients are elderly with limited mobility, making the process cumbersome and time-consuming for both patients and clinicians[10]. Additionally, doctor-based subjective assessments are prone to significant bias, and the data are often difficult to retain, preventing continuous monitoring of the patient's condition. Remote intelligent assessment offers a promising solution by reducing costs and improving convenience and accessibility, enabling patients in remote areas or with mobility challenges to receive professional evaluations without geographical constraints[11].

The technology for assessing ET has rapidly evolved in recent decades. Devices such as inertial sensors[12], EMG[13], video equipment[14], and electronic handwriting boards[15,16] have significantly enhanced the objectivity, quantification, and consistency of tremor detection[7]. The application of machine learning and deep learning algorithms in ET assessment has gained increasing attention[7,17]. Ali et al.[18] recorded accelerometer signals

[1]Center for Artificial Intelligence in Medicine, Medical Innovation Research Department, PLA General Hospital, Beijing, China. [2]School of Biological Science and Medical Engineering, Beihang University, Beijing, China. [3]Department of Neurosurgery, Chinese PLA General Hospital, Beijing, China. [4]Department of Neurology, 923Th Hospital of the Joint Logistics Support Force of PLA, Nanning, China. [5]Chinese PLA Medical School, Beijing, China. [6]School of Medicine, Nan Kai University, Tianjin, China. [7]Yumeng Peng and Songliang Han have contributed equally to this work. ✉email: panls301@163.com; zhengbozhang@126.com

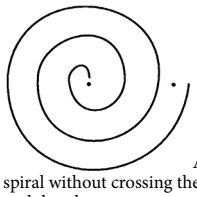
Tasks	Score	Guide
 Asked the patient to join both points of the Archimedean spiral without crossing the lines and without resting your hand or arm on the table. Test each hand	0	Normal
	1	Slightly tremulous, may across lines occasionally
	2	Moderately tremulous or crosses lines frequently
	3	Accomplishes the task very hard, or any errors
	4	Cannot complete the task

Table 1. Rules for tremor severity scoring by Archimedean spiral drawing task in the CRST.

from 35 subjects while they drew Archimedean spiral diagrams, distinguishing ET from healthy controls. Sole-Casals et al.[19] used information from electronic handwriting boards as input to an SVM model to differentiate Parkinson's Disease (PD) and ET. Wang et al.[15] combined convolutional neural networks with electronic handwriting boards to diagnose ET.

Although previous research has highlighted the advantages of objectively assessing and diagnosing ET, these methods also have limitations. Most studies have focused on ET identification, such as distinguishing ET from healthy populations. This binary classification task primarily involves identifying differences in tremor patterns, which are often visually apparent. However, fine-grained assessment of ET tremor severity remains a significant clinical challenge. To our knowledge, no published research has utilized standard CRST-guided handwritten images for grading ET tremor severity. Furthermore, remote home monitoring and self-assessment for ET, essential for tracking disease progression, remain unexplored. To address these gaps, we present the following contributions in this paper.

Proposing a remote ET diagnosis framework

This study introduces an innovative framework for remotely assessing ET severity using handwritten images, overcoming the limitations of traditional clinical settings. By utilizing paper and pen, this approach provides a low-cost, accessible, and scalable solution for ET evaluation, particularly beneficial for elderly patients or those in resource-limited environments.

Creating a high-quality handwritten ET grading dataset

We collected about 1000 high-quality CRST Archimedean spiral handwriting images from more than 300 patients, both from medical institutions and remote home environments, forming a robust dataset with expert ratings. This dataset addresses the shortage of high-quality annotated data for ET severity grading and offers a valuable resource for further research.

Demonstrating the model's strengths

The proposed ETSD-Net model achieved 88.44% accuracy, surpassing traditional physician assessments and baseline models such as ConvNeXt-Tiny, DenseNet, MobileNet-V2, and ResNet50. By using deep learning for image recognition, our model provides an objective, consistent, and quantifiable method for ET severity assessment, especially when clinical evaluations are constrained by time, resources, or expertise. Furthermore, our model's use of Grad-CAM visualizations enhances interpretability, ensuring that the focus remains on relevant features of the handwriting images, thereby improving diagnostic reliability.

Related work

We conducted a search in the Web of Science database using the query "**Title = essential tremor AND (Topic = drawing OR Topic = writing)**", yielding 151 results as of January 9, 2025. After reviewing the titles and abstracts, we excluded 138 studies that focused on genetics, epidemiology, surgery, or pharmacological treatments. Finally, 13 relevant articles were selected for full-text review, and 9 articles focused on ET severity assessment and potentially containing handwriting datasets were summarized in our analysis, as shown in Table 2. Ali et al.[20] recruited 17 ET patients and 18 healthy controls. Participants performed guided Archimedean spiral drawing tasks while wearing an Inertial Measurement Unit (IMU) on the forearm. Classification was performed using a Support Vector Machine, achieving only 68.57% accuracy in estimating the severity of ET. Ma et al.[21], using a digital writing tablet and pen, collected multi-modal data from 147 ET patients. By utilizing transfer learning and an attention mechanism, their system achieved an accuracy ranging from 97.33% to 97.39% for five-category tremor severity classification. Although this accuracy is quite high, the data collection required supervision from researchers and took place in a laboratory setting, making it unsuitable for remote home assessments.

After summarizing the findings, we observed that current studies are predominantly based on IMUs or electronic handwriting boards. At present, no research has focused solely on using handwritten images to assess the severity of ET, nor has a dedicated handwriting image dataset for ET patients been established. Moreover,

Authors (Year)	Sensor	Objectives	datasets	Accuracy
Holly et al.[22]	IMU	Estimating disability	Drawing without guide lines	
Ali et al.[20]	IMU	ET diagnosis/assessing ET severity	Data unavailable	91.42% (diagnosis) 68.57% (Severity assessment)
Ma et al. [21]	Electronic tablet	Assessing ET severity	Based on electronic writing tablet	97.39% (Best)
Adran et al.[16]	Electronic tablet	ET diagnosis	Drawing without guide lines	93.00% (Best)
McGurrin et al.[23]	IMU	Relating sensor-measured tremor to clinical ratings	Data unavailable	
Ali et al.[18]	IMU	Assessing ET severity	Data unavailable	85.71% (binary classification)
Lopez-de-Ipina et al.[24]	Electronic tablet	Analysis of fine motor skills	Drawing without guide lines	
Motin et al.[25]	Electronic tablet	Assessing ET severity	Data unavailable	87.20%
Yu et al.[26]	Electronic tablet	Explore graphomotor function characterization	Drawing circle	

Table 2. Overview of handwriting studies that assess ET.

Characteristic	Grades					p-value
	Overall, N = 315[1]	1, N = 106[1]	2, N = 105[1]	3, N = 70[1]	4, N = 34[1]	
Age	63 ± 10	62 ± 10	63 ± 11	64 ± 11	66 ± 8	0.169[2]
BMI	25.0 ± 3.4	25.4 ± 3.7	24.3 ± 2.8	25.3 ± 3.6	24.8 ± 3.5	0.089[2]
Gender						
Female	85 (27.0%)	34 (32.1%)	27 (25.7%)	14 (20.0%)	10 (29.4%)	0.346[3]
Male	230 (73.0%)	72 (67.9%)	78 (74.3%)	56 (80.0%)	24 (70.6%)	
Duration(years)	21 ± 11	22 ± 11	20 ± 9	20 ± 12	19 ± 11	0.499[2]
Family History	1.03 ± 0.54	1.06 ± 0.61	0.98 ± 0.48	1.09 ± 0.50	0.97 ± 0.52	0.515[2]
CRST Write_R	5.0 ± 4.5	1.5 ± 1.7	4.1 ± 2.7	7.5 ± 2.9	13.9 ± 2.1	<0.001[2]
CRST Write_L	6.8 ± 3.6	3.2 ± 1.7	6.8 ± 2.4	10.0 ± 2.3	10.9 ± 1.9	<0.001[2]
CRST TOTAL	32 ± 21	13 ± 6	29 ± 10	48 ± 10	70 ± 11	<0.001[2]

Table 3. Patient demographics and baseline characteristics. This table presents the patient demographics and baseline characteristics, including age, BMI, gender, disease duration, family history (whether family members have a history of ET), and CRST scores. The CRST is used to assess tremor severity, with higher scores indicating greater severity. CRST Write_R and CRST Write_L refer to the CRST scores for writing tasks performed with the right and left hands, respectively. The CRST TOTAL is the sum of CRST Write_R and CRST Write_L. Duration (years) represents the number of years since the initial ET diagnosis for each patient. [1]Mean ± standard deviation (SD); n (%). [2]One-way ANOVA. [3]Pearson's Chi-squared test.

existing studies primarily rely on data collected in laboratory settings, without addressing methods for remote assessment. The spiral diagram provided by digital handwriting systems lacks a reference template, which increases the time required for physicians to guide patients during the drawing process. Visual diagnosis by experts often involves observing the number of crossings between the handwriting and the template. Furthermore, compared to electronic handwriting boards, pen and paper offer a more natural writing environment, making the diagnosis and assessment closer to real-life conditions. This approach also avoids the costs associated with purchasing and maintaining electronic handwriting boards or IMU devices, making the assessment more affordable and convenient, with the potential for remote evaluation.

Methods

Subjects and protocol

This study aims to develop a convenient and highly accurate method for assessing the severity of ET using handwritten images. It is based on the clinical trial titled "Efficacy and Safety Study of ExAblate Transcranial MRgFUS Thalamic Disruption for Drug-Refractory Idiopathic Tremor" (trial protocol code ET002J) at PLA General Hospital. The trial was approved by the Ethics Committee of the Chinese PLA General Hospital (S2018-021–00/01). The ET002J clinical trial is part of a prospective, single-arm, multi-center clinical trial sponsored by InSightec (ClinicalTrials.gov Identifier: NCT03253991). Patients with symptoms of tremor were remotely recruited through web-based questionnaires (designed for the ET002J clinical trial), telephone interviews, and video consultations. Multiple neurologists screened patients with typical symptoms of ET. Data collection for this experiment began on September 4, 2020. As of March 5, 2024, 315 ET patients at different stages—who had completed the necessary tests and were confirmed not to have other conditions such as PD, hyperthyroidism, hepatolenticular degeneration, or drug-induced tremors—were included in the study. The participants' ages ranged from 30 to 78 years. To provide a clearer understanding of the patient profiles, we categorized the patients into four groups based on their CRST Write scores: Group 1 (0–7 points), Group 2 (8–14 points), Group 3 (15–21 points), and Group 4 (22–28 points), as shown in Table 3.

Research on digital images has found that drawing Archimedean spirals or straight lines offers greater discriminative power than writing[5], likely due to its requirement for continuous movements across multiple planar directions, unlike the predominantly vertical motions associated with writing[27]. Therefore, we chose to have patients use paper and pen to complete the CRST scale drawing A (large Archimedean spiral) as the method for data collection. Patients were randomly divided into two groups: one undergoing evaluation in medical institutions and the other undergoing remote home-based evaluation, at an 8:2 ratio. After a 24-h medication withdrawal, the medical institutions group consisted of 252 patients who completed the task under the guidance of neurologists. The handwritten images were then scanned using the scanning function of the HP LaserJet Pro MFP M226dw printer (Hewlett-Packard, USA) and saved as JPEG format images with a preset resolution of 300 dpi. The 63 patients in the remote home-based group were instructed to print a PDF template of the drawing task on A4 paper at home, complete the drawing as instructed, and then take photos using smartphones or cameras at the highest resolution and upload the images in full resolution JPEG format.

This resulted in a total of 798 high-definition scanned images and 199 photos of patient-produced drawings from the remote group. An expert panel of neurologists assessed the tremor severity of the patients using the CRST scale based on the handwritten images. Each image was independently scored by three neurologists. In cases of scoring discrepancies, the neurologists discussed and re-evaluated the images to reach a consensus score. This process resulted in a high-quality dataset of 997 images. These consensus scores were then used as labels for training the model. The experimental process is illustrated in Fig. 1.

Dataset preparation

In this study, we utilized transfer learning techniques for the classification of handwritten spiral images. Transfer learning is a deep learning method where a pre-trained model, originally developed for a specific task, is repurposed as the starting point for a model on a new task[28–31]. The pre-trained model leverages the prior knowledge gained by the model on the extensive ImageNet dataset to enhance performance on a specialized, private dataset.

Preprocessing is a critical initial step that ensures the input data is suitable and of high quality for training deep learning models. To make our handwritten images compatible with pre-trained deep learning models, we applied several preprocessing steps, including resizing images, normalizing pixel values, and data augmentation[28–31]. Many deep learning models, particularly convolutional neural networks (CNNs) like VGG16 and ResNet, are trained with fixed-size input images, typically 224 × 224 pixels[29]. To ensure compatibility with these pretrained models, we resize our images to 224 × 224 pixels. This size strikes a balance between computational efficiency and image detail. Larger images (e.g., 512 × 512) increase computational costs and training time, while smaller sizes (e.g., 64 × 64) may lose detail, reducing performance. Thus, 224 × 224 pixels provides an optimal compromise[32].

This resizing ensures that the model can effectively process the images without distortion or loss of critical information. The images were normalized using the mean and standard deviation values derived from the ImageNet dataset ([0.485, 0.456, 0.406] for the means and [0.229, 0.224, 0.225] for the standard deviations, respectively). Normalization and resizing as part of preprocessing ensure that the model focuses on learning relevant features from the images, rather than being influenced by variations in color, brightness, or size. This process helps in stabilizing the training process and improves model convergence.

Beyond preprocessing, data augmentation plays a pivotal role in enhancing the model's ability to generalize from the training data to unseen data. It artificially expands the training dataset by applying a series of random transformations that produce plausible variations of the input images. Our data groups were naturally imbalanced, with a higher number of patients with mild tremors, leading to an unequal distribution of samples across each group. To address this imbalance, we applied data augmentation techniques before training the model to balance the groups and ensure a more even distribution of samples. Specifically, we used rotation (±25 degrees), horizontal flipping, vertical flipping, scaling (0.8 to 1.2), additive Gaussian noise (scale 0 to 0.05*255), Gaussian blur (sigma 0 to 3.0), linear contrast adjustment (0.75 to 1.5), brightness multiplication (0.8 to 1.2), and random cropping (up to 10%). By systematically applying these preprocessing and augmentation techniques, we expanded each data group to 800 samples, ensuring that our dataset was robust and capable of training an effective deep learning model for the classification of handwritten spiral images.

Model evaluation

To comprehensively evaluate the applicability of using hand-drawn spiral lines to assess the severity of ET, we incorporated four high-performance deep learning networks as baseline models.

(1) ResNet50[29]: The residual structure alleviates the gradient vanishing problem, making it suitable for extracting complex features. Its mature and stable architecture is a widely used benchmark model in visual tasks.
(2) DenseNet[30]: The dense connectivity mechanism enables efficient feature reuse, achieving good performance with fewer parameters. Its characteristics make it particularly suitable for tasks involving small datasets or requiring deep feature fusion.
(3) ConvNeXt-Tiny[31]: It widely adopts the design principles of modern lightweight convolutional networks, enabling it to capture multi-scale local and global features while balancing performance and efficiency.
(4) MobileNet-V2[28]: By utilizing depthwise separable convolutions and an inverted residual structure, it achieves a balance between computational efficiency and predictive performance. Its lightweight design is highly suitable for deployment on mobile or portable devices.

Considering that the remote diagnosis of ET severity relies on the accuracy and efficiency of the model, we propose ETSD-Net, an improved model based on MobileNetV2. The input size of the ETSD-Net is 224×224 pixels, which is consistent with the preprocessing step where all images were resized to this dimension. It comprises 2

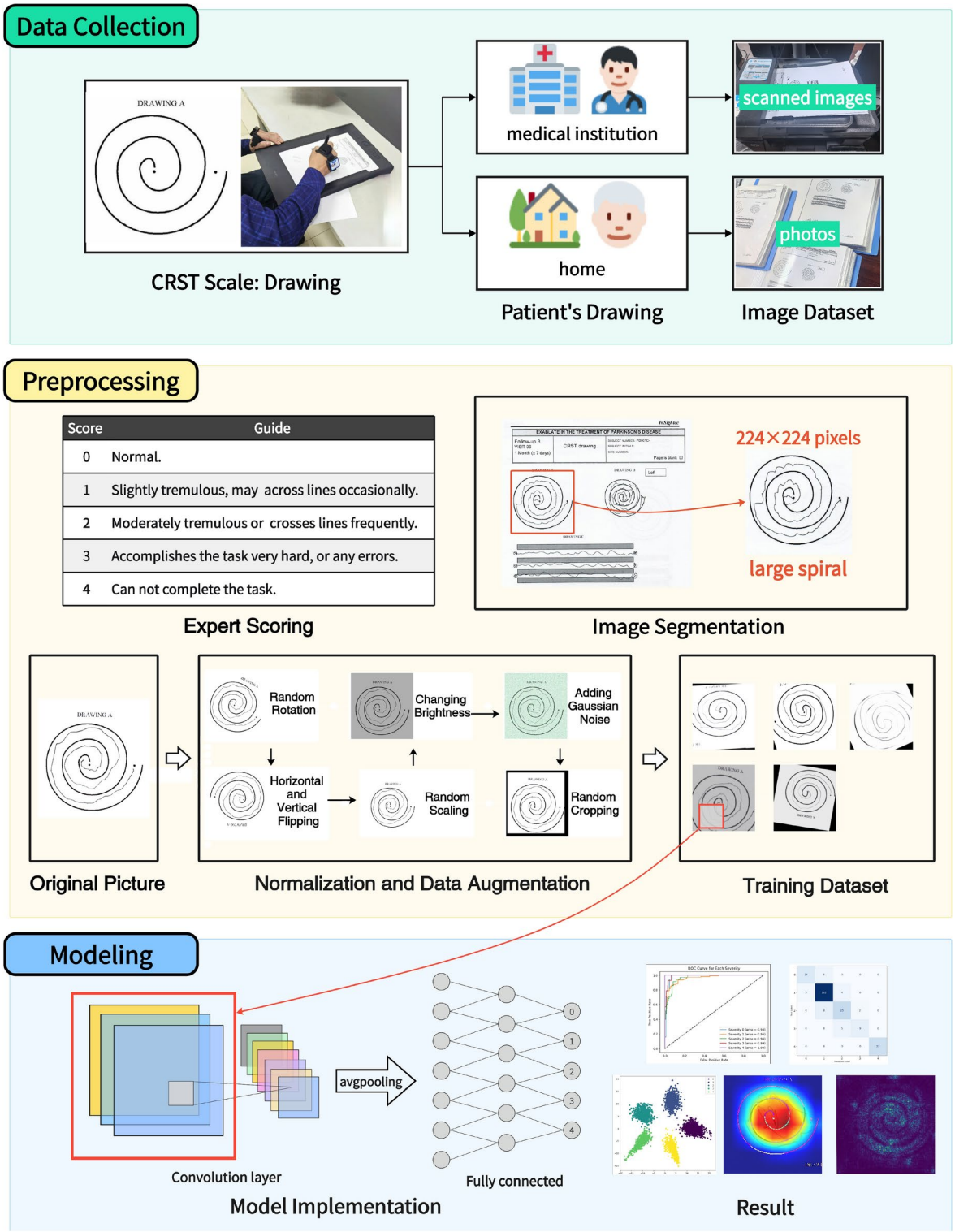


Fig. 1. Schematic diagram of the proposed system for a deep learning approach to remotely assessing ET with handwritten images. This mainly includes: (1) Patients completing the CRST handwriting tasks at home or in medical institutions and collecting handwritten images; (2) Experts scoring and processing images to form a dataset; and (3) modeling and evaluation.

convolution blocks, N inverted residual blocks and 1 fully connected layers, with specific batch normalization layer and activation functions. Specifically, The input of ETSD-Net is first passthrough a convolution block which contains convolution layer, batch normalization layer, and relu layer to extract shallow features. The the shallow features are sent to several stacked inverted residual blocks to capture spatial details and semantic information in

hand-drawn spiral line images, which is benefited from the introduce of a channel-spatio attention mechanism in the inverted residual modules of the network. Then the semantic features are refined by another convolution blocks with average pooling layers. Finally, the classes related features are flattened and sent to a fully connected layers to get the outputs.

Experimental setups

To prevent data leakage and ensure an objective evaluation of both the baseline model and ETSD-Net proposed in this paper, we employed a subject-independent data split strategy, ensuring that the data from the same subject only appears in one of the training, validation, or test sets. Based on this, we divided the data collected in Section. Dataset Preparation in a 6:2:2 ratio.

To reduce the cost of training the model and accelerate the convergence process, both the baseline model and ETSD-Net are trained using transfer learning in this paper. The reason we adopted this approach is because the initial layers in transfer learning models capture generalizable features, while the latter layers are more task-specific.

So, the models underwent a two-phase training process: the first phase involved adding a new classifier, and the second phase focused on fine-tuning the model. In the first phase, for ResNet50, DenseNet, ConvNeXt-Tiny and MobileNet-V2, a fully connected layer corresponding to the five severity levels defined by the CRST was added to the end of each of the four pre-trained models, with the weights of the other layers frozen. For ETSD-Net, since its underlying architecture is MobileNet-V2, we used the pre-trained weights of MobileNet-V2 when loading the model. We froze the weights of the modules that could be matched, and the fine-tuning mainly focused on the spatiotemporal attention module and the final classification layer. In this way, the initial layers of a neural network capture universal features like edges and textures. By keeping these layers unchanged, the model can utilize these learned features without the need for retraining.

During the fine-tuning process, we use a batch size of 64 for iterative training and employed a learning rate schedule with hierarchical decay. The base learning rate was set to 10^{-3}. For the feature extraction layers, the learning rate is decayed by a factor of 0.5 every three layers, while the classification layer is set to the base learning rate without decay. This is because the feature layers contain pre-trained weights and do not require a large learning rate to find optimal weights, whereas the classification layer has not loaded pre-trained weights. The Adam optimizer is used for parameter optimization, and multi-class cross-entropy is used as the loss function for backpropagation. The entire fine-tuning process lasted for 10 epochs. Here, we use a small number of epochs because the model is initialized with pretrained weights from ImageNet. Experimental results show that this initialization method enabled the model to converge within 10 epochs. Therefore, we do not use a larger number of training epochs to avoid overfitting on the small-scale dataset. The model is trained on the training set, and the model with the highest accuracy on the validation set is saved as the best model. The performance metrics used to evaluate the model in this paper are accuracy, F1-score, precision, and recall.

The process is exemplified by ETSD-Net and illustrated in Fig. 2.

Results

The performance of four state-of-the-art transfer learning models—ResNet50, DenseNet, MobileNet-V2, and ConvNeXt-Tiny—against our proposed model using four evaluation metrics: Accuracy, Precision, Recall, and F1-score were compared. The results are summarized in Table 4.

Our ETSD-Net outperforms the baseline models in several key metrics. Specifically, it achieves the highest Accuracy (88.44%), Recall (88.44%), and F1-score (88.45%), demonstrating its superior ability to correctly

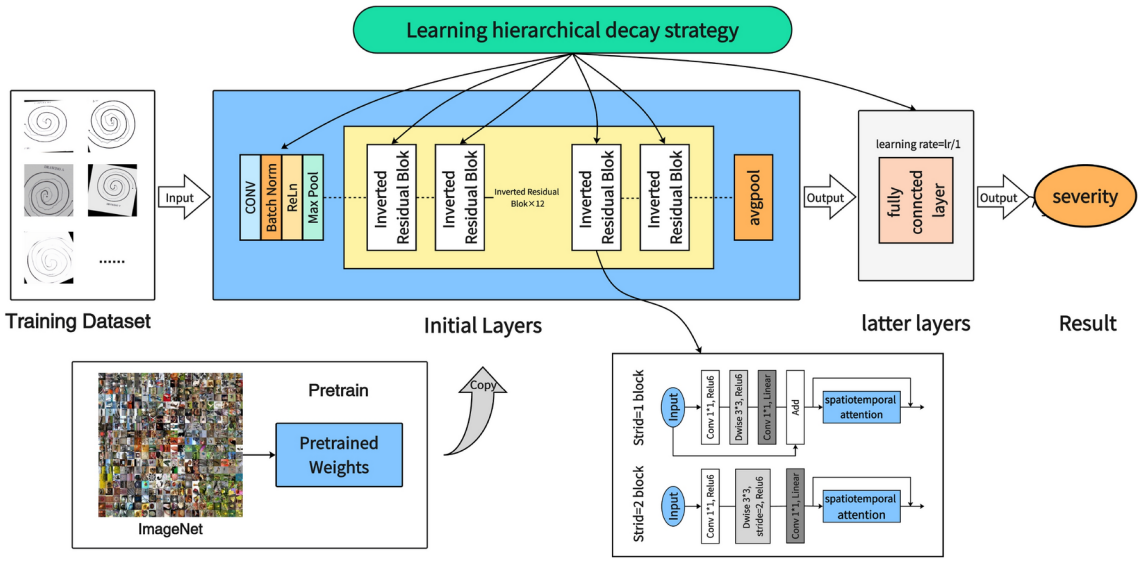


Fig. 2. Integration of ETSD-Net in severity prediction.

classify and balance precision with recall. While ConvNeXt-Tiny shows slightly higher Precision (88.87%), our model achieves a comparable result (88.64%) while maintaining better overall performance across the other metrics.

To evaluate the impact of data augmentation on the performance of our model, we compared the results of ETSD-Net trained on the original dataset and the augmented dataset. The model trained on the augmented dataset achieved an accuracy of 88.44%, compared to 86.43% on the original dataset. The precision, recall and F1-score also showed significant improvements, with the precision increasing from 86.39% to 88.64% and the F1-score from 86.18 to 88.45%.

The performance of our model is demonstrated through the ROC curve and confusion matrix. The ROC curves for each severity level show high areas under the curve (AUC) values, ranging from 0.98 to 0.99, indicating excellent classification capability across all categories. The confusion matrix further highlights the model's robust performance, with most predictions falling along the diagonal, signifying correct classifications (as shown in Fig. 3). These results indicate that our proposed model successfully leverages transfer learning techniques and outperforms existing architectures in multiple evaluation criteria, making it more robust and reliable for the task at hand.

ETSD-Net, built upon the MobileNet-V2 architecture, maintains a lightweight design with comparable computational efficiency to MobileNet-V2. As shown in Table 5, both models exhibit similar FLOPs (0.33 GFLOPs for ETSD-Net vs. 0.30 GFLOPs for MobileNet-V2) and parameter sizes (2.26M for ETSD-Net vs. 2.20M for MobileNet-V2). This similarity reflects the architectural choices aimed at retaining MobileNet-V2's computational efficiency while incorporating improvements to enhance performance.

However, our model's inference time (20.17 ± 10.56 ms) is notably higher than that of MobileNet-V2 (7.21 ± 0.93 ms). This increase in latency can be attributed to the additional modifications designed to improve feature extraction and overall performance. In contrast, other baseline models, such as ResNet50 and ConvNeXt-Tiny, achieve faster inference times (e.g., 5.34 ± 1.28 ms for ConvNeXt-Tiny), but at the cost of significantly higher computational demands (4.45 GFLOPs and 27.80M parameters for ConvNeXt-Tiny).

Figure 4 and Fig. 5 illustrate the Grad-CAM[33] and saliency map[34] visualizations, respectively, for the different models across the CRST input images (0–4). These visualizations collectively reveal how each model allocates attention to specific regions of the input images during classification and provide complementary insights into their focus mechanisms. The Grad-CAM results highlight the broader attention distribution, while the saliency maps emphasize sensitivity to task-relevant regions at a more granular level.

ResNet50 and DenseNet show limited and localized attention distributions in both Grad-CAM and saliency maps, primarily focusing on isolated segments of the spiral. This incomplete coverage suggests that these models struggle to capture the global geometric pattern of the spiral, which is critical for accurate classification. ConvNeXt-Tiny exhibits attention that is heavily concentrated on disconnected, small regions, as seen in both visualizations, indicating a tendency to overfit to local details rather than understanding the spiral's overall structure. MobileNet-V2 demonstrates a more balanced attention pattern compared to the earlier models, with Grad-CAM and saliency maps showing broader coverage of the spiral. However, its focus is still insufficiently global, leaving portions of the structure underrepresented.

In contrast, ETSD-Net achieves the most comprehensive and consistent attention distribution, as evidenced by both Grad-CAM and saliency maps. The model effectively learns the full geometric structure of the spiral, with attention covering both central and peripheral regions. This global focus ensures robust feature extraction and better generalization across varying input complexities. The combined results from Grad-CAM and saliency maps strongly support ETSD-Net's superior performance, as its ability to capture both local and global features is unmatched by the baseline models.

Discussion

Clinical value of this framework

This study innovatively uses handwritten images to enable remote assessment of ET, breaking the current limitation where ET evaluation is confined to clinical settings. We collected and processed 997 high-quality handwriting images from medical institutions and remote home environments, creating a high-quality dataset through rigorous evaluation by a panel of neurological experts. Using a transfer learning approach, we proposed the ETSD-Net model for ET severity assessment, which achieved the best performance with an accuracy of 88.44%, surpassing the baseline models ConvNeXt-Tiny, DenseNet, MobileNet-V2, and ResNet50.

Studies show that scoring tremor severity with scales like the CRST requires trained raters to achieve reliable results[35]. Grimaldi et al.[36] reported an intraclass correlation coefficient of 85.9% for CRST total scores, with variability across subcomponents (e.g., 88.2% for tremor amplitude and 67.1% for daily activity). Similarly, Elble

Method	Accuracy	Precision	Recall	F1-score
ResNet50	86.43%	88.48%	86.43%	86.73%
DenseNet	85.93%	85.92%	85.93%	85.51%
ConvNeXt-Tiny	86.93%	**88.87%**	86.93%	87.21%
MobileNet-V2	87.44%	87.63%	87.44%	87.44%
ETSD-Net (original datasets)	86.43%	86.39%	86.43%	86.18%
ETSD-Net (Ours)	**88.44%**	88.64%	**88.44%**	**88.45%**

Table 4. Performance comparison of baseline models and ETSD-Net. Significant values are given in bold.

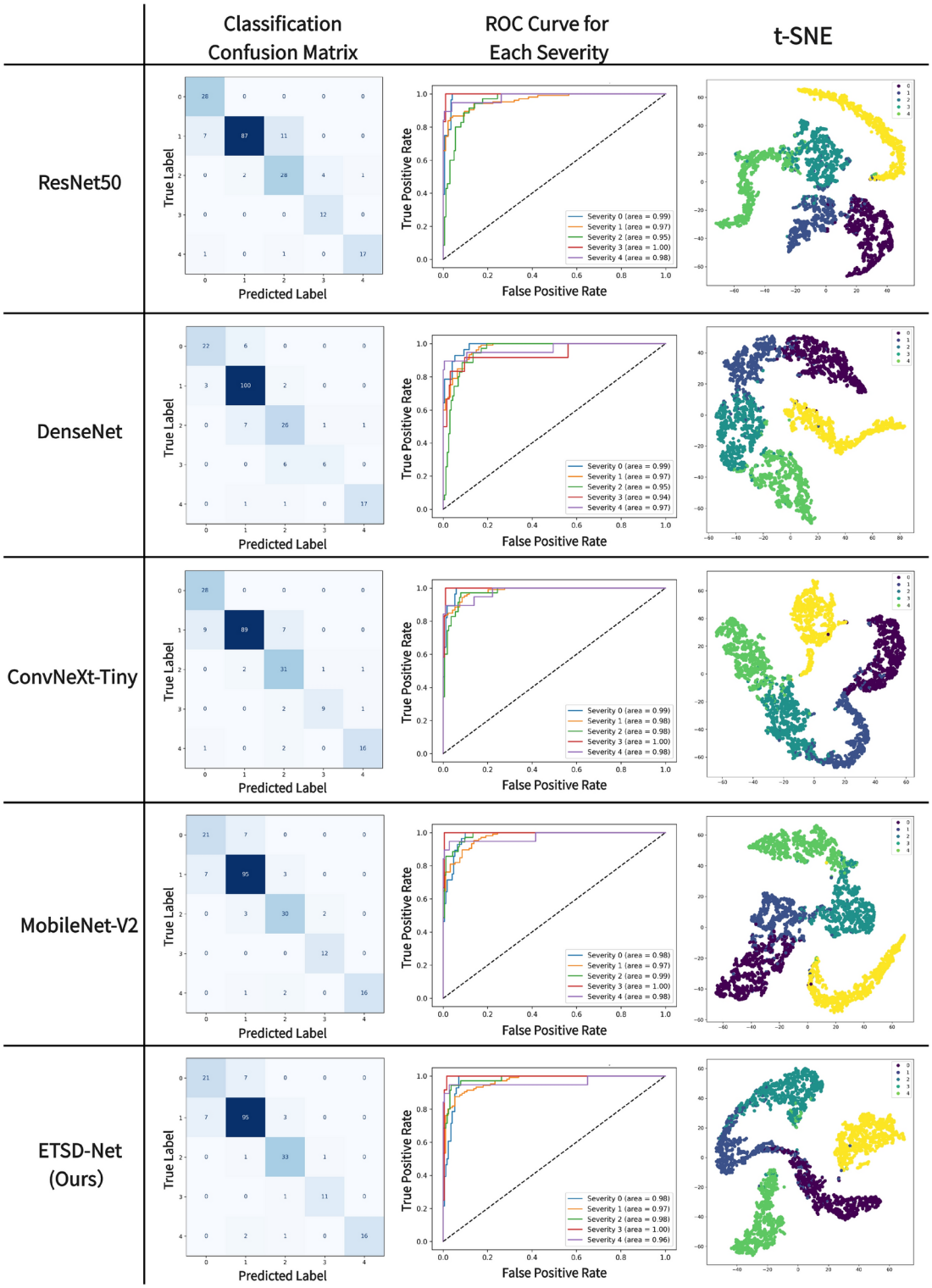


Fig. 3. Performance evaluation of baseline models and ETSD-Net: confusion matrices, ROC curves, and t-SNE visualization for ET severity prediction.

et al.[37] reported validity testing with spiral ratings (80.4%) and handwriting ratings (76.2%) in ET patients. In comparison, our model provides a consistent, objective, and quantifiable method for tremor severity assessment, which is especially useful when clinical evaluations are limited by time, resources, or expertise. Standardizing assessments through algorithms can complement clinical evaluations and improve diagnostic reliability, particularly in remote or resource-limited settings.

Models	Flops	Params	Inference time (ms)
ResNet50	4.13 G	23.52 M	7.17 ± 4.15
DenseNet	2.83 G	6.88 M	18.27 ± 1.64
ConvNeXt-Tiny	4.45 G	27.80 M	5.34 ± 1.28
MobileNet-V2	0.30 G	2.20 M	7.21 ± 0.93
ETSD-Net (Ours)	0.33 G	2.26 M	20.17 ± 10.56

Table 5. Computational efficiency and inference time of baseline models and ETSD-Net. FLOPs (Floating Point Operations) indicate the model's computational complexity. Params (Parameters) affect memory usage and the model's capacity. Inference time is the time taken for the model to process an input and generate an output, measured in milliseconds (ms).

Given that current ET assessment research is primarily based on electronic handwriting boards and IMU devices, our study demonstrates the feasibility of using only paper and pen for remote ET evaluation. This approach is low-cost, preserves the natural writing habits of patients, and is particularly beneficial for elderly ET patients. We believe our study is clinically meaningful and represents a significant step toward integrating remote, automated methods into routine ET assessment.

Model performance and interpretability

In the previous Sect "Related work", we conducted a detailed review of research on handwriting-based ET severity assessment. We found that current studies using handwriting to assess ET severity are primarily based on IMU or electronic tablets. The lowest accuracy reported was 68.57% by Ali et al.[20], who used an IMU-based approach, while the highest accuracy was 97.39% from Ma et al.[21] from our research group, which used an electronic tablet. McGurrin et al.[24], also using an IMU, achieved an accuracy of 93.00%. However, it is important to note that these studies using IMU or electronic tablets involve more complex paradigms, with accelerometer, electronic trajectory, or pressure signals that differ significantly from our handwriting image-based approach. In contrast, our model achieved 88.44% accuracy based solely on handwriting images, far surpassing the clinical accuracy of 76.20%-80.40% reported in handwriting-based studies.

We use a deep learning model for image recognition that processes the input handwriting images directly and extracts relevant features through convolutional layers, without requiring additional input characteristics. The model is designed to autonomously learn important patterns and representations from the input images. However, to enhance interpretability and understand what the model focuses on during its decision-making process, we incorporated Grad-CAM visualizations. These visualizations highlight the specific regions of the handwriting images that the model attends to, providing insights into how the model interprets the handwriting. This approach ensures that the focus remains on the image data. We also used t-SNE for visualization, which shows that the model effectively classifies tremor severity based on the handwriting images.

Limitations and future directions

However, we acknowledge several limitations in this study. Our research is based on handwriting images, and the quality of the images affects the performance of our model. Ink flow issues during writing could have influenced the results. Most patients carefully completed the drawings under the supervision of neurologists or at home. However, some patients, due to urgency for medical attention or other distractions, did not complete the drawings as carefully. This may have led to less accurate drawings compared to those done more attentively. Using a deep learning model to analyze 2D static images, while practical, also limits the study of clinical physiology and the interpretability of the model. In the future, we plan to expand the collection of handwritten images to include those from PD patients, which will help facilitate the early differential diagnosis between PD and ET.

Conclusion

This study presents an effective approach to remotely assessing the severity of ET using handwriting images, offering a practical and accessible method for evaluation. We collected about 1000 high-quality CRST Archimedean spiral handwriting images from more than 300 patients, establishing a robust dataset with expert ratings. Using a transfer learning approach, we developed the ETSD-Net model for ET severity assessment. Our model not only achieves objective and accurate evaluations but also enables remote, low-cost assessments. With an accuracy of 88.44%, ETSD-Net outperforms existing methods and shows great potential for integrating into clinical practice, especially in remote or resource-limited settings. This study represents a meaningful contribution to improving the accessibility and reliability of ET assessment, particularly for elderly patients who may benefit most from home-based evaluations.

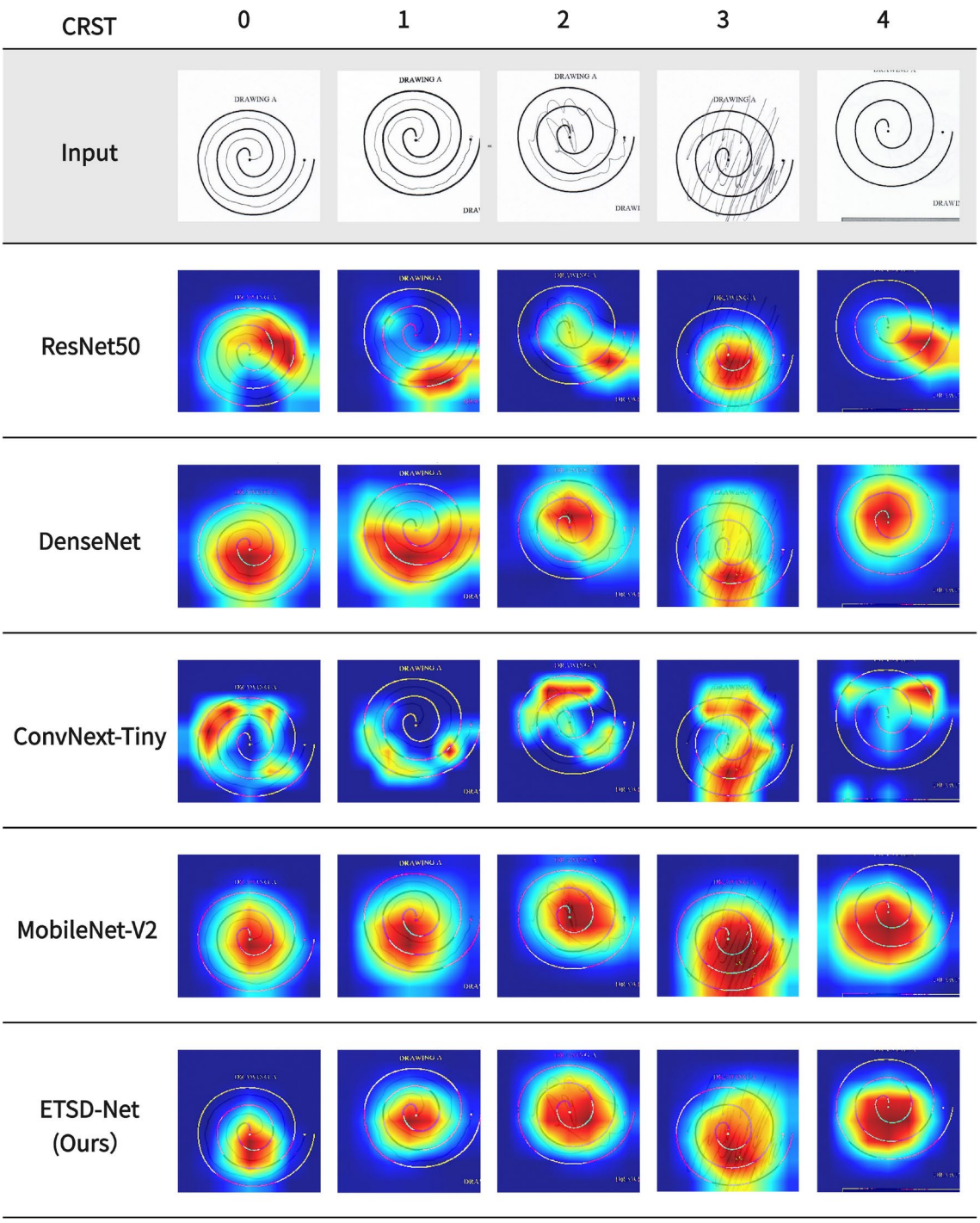


Fig. 4. Grad-CAM heatmap visualizations for baseline models and ETSD-Net.

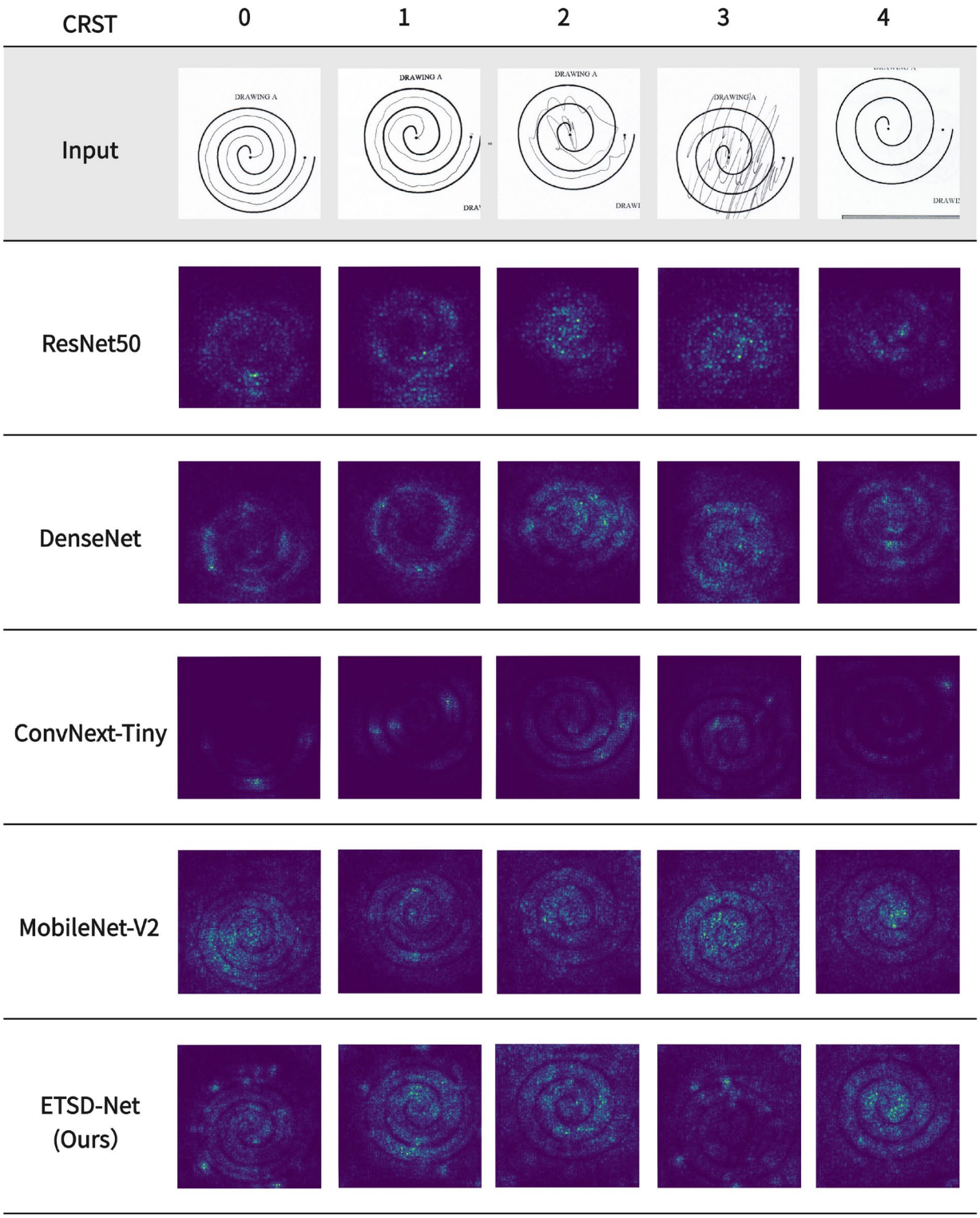


Fig. 5. Saliency map visualizations for baseline models and ETSD-Net.

Data availability

Upon reasonable request, the corresponding author can provide access to the data supporting the findings of this study.

Received: 19 October 2024; Accepted: 17 March 2025
Published online: 28 March 2025

References

1. Welton, T. et al. Essential tremor. *Nat. Rev. Dis. Primers* **7**, 83. https://doi.org/10.1038/s41572-021-00314-w (2021).
2. Louis, E. D. the roles of age and aging in essential tremor: An epidemiological perspective. *Neuroepidemiology* **52**, 111–118. https://doi.org/10.1159/000492831 (2019).
3. Haubenberger, D. & Hallett, M. Essential tremor. *N. Engl. J. Med.* **378**, 1802–1810. https://doi.org/10.1056/NEJMcp1707928 (2018).
4. Bhatia, K. P. et al. Consensus statement on the classification of tremors. from the task force on tremor of the International Parkinson and Movement Disorder Society: IPMDS task force on tremor consensus statement. *Mov. Disord.* **33**, 75–87. https://doi.org/10.1002/mds.27121 (2018).
5. Peters, J. et al. Computerised analysis of writing and drawing by essential tremor phenotype. *BMJ Neurol. Open* **3**, e000212. https://doi.org/10.1136/bmjno-2021-000212 (2021).
6. Fahn S, Tolosa E, Marin C. Clinical rating scale for tremor. *Parkinson's Disease and Movement Disorders.* 225–234 (1988).
7. Peng, Y. et al. Intelligent devices for assessing essential tremor: a comprehensive review. *J. Neurol.* **271**(8), 4733–4750. https://doi.org/10.1007/s00415-024-12354-9 (2024).
8. Alty, J., Cosgrove, J., Thorpe, D. & Kempster, P. How to use pen and paper tasks to aid tremor diagnosis in the clinic. *Pract. Neurol.* **17**, 456–463. https://doi.org/10.1136/practneurol-2017-001719 (2017).
9. Elble, R. J. The essential tremor rating assessment scale. *J. Neurol. Neuromed.* **29**, 507–512 (2016).
10. David-Olawade, A. C. et al. Nursing in the digital age: Harnessing telemedicine for enhanced patient care. *Inform. Health* **1**, 100–110 (2024).
11. Tan, S. Y., Sumner, J., Wang, Y. & Wenjun Yip, A. A systematic review of the impacts of remote patient monitoring (RPM) interventions on safety, adherence, quality-of-life and cost-related outcomes. *npj Digit. Med.* **7**, 1–16. https://doi.org/10.1038/s41746-024-01182-w (2024).
12. Ma, C. et al. Quantitative assessment of essential tremor based on machine learning methods using wearable device. *Biomed. Signal Process. Control* **71**, 103244. https://doi.org/10.1016/j.bspc.2021.103244 (2022).
13. Ruonala V et al. EMG Signal morphology in essential tremor and Parkinson's disease. In *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC).* 5765–5768. https://doi.org/10.1109/EMBC.2013.6610861. (2013).
14. Kovalenko, E. et al. Distinguishing between Parkinson's disease and essential tremor through video analytics using machine learning: A pilot study. *IEEE Sens. J.* **21**, 11916–11925 (2021).
15. Wang, Y. et al. Application of optimized convolutional neural networks for early aided diagnosis of essential tremor: Automatic handwriting recognition and feature analysis. *Med. Eng. Phys.* **113**, 103962 (2023).
16. Adran Otero, J. F. et al. EMD-based data augmentation method applied to handwriting data for the diagnosis of ESSENTIAL TREMOR using LSTM networks. *Sci. Rep.* https://doi.org/10.1038/s41598-022-16741-y (2022).
17. Singh P, Singh SP, Singh DS. An introduction and review on machine learning applications in medicine and healthcare. In *2019 IEEE Conference on Information and Communication Technology (CICT).* https://doi.org/10.1109/CICT48419.2019.9066250. (2019).
18. Ali, S. M. et al. Wearable sensors during drawing tasks to measure the severity of essential tremor. *Sci. Rep.* **12**, 5242. https://doi.org/10.1038/s41598-022-08922-6 (2022).
19. Sole-Casals, J. et al. Discrete cosine transform for the analysis of essential tremor. *Front. Physiol.* **9**, 1947. https://doi.org/10.3389/fphys.2018.01947 (2019).
20. Ali, S. M. et al. Wearable accelerometer and gyroscope sensors for estimating the severity of essential tremor. *IEEE J. Transl. Eng. Health Med.* **12**, 194–203. https://doi.org/10.1109/JTEHM.2023.3329344 (2024).
21. Ma, C. et al. Automatic diagnosis of multi-task in essential tremor: Dynamic handwriting analysis using multi-modal fusion neural network. *Future Gener. Comput. Syst. Int. J. Escience* **145**, 429–441. https://doi.org/10.1016/j.future.2023.03.033 (2023).
22. Holly, P. et al. Estimating disability in patients with essential tremor: Comparison of tremor rating scale, spiral drawing, and accelerometric tremor power. *Mov. Disord. Clin. Pract.* https://doi.org/10.1002/mdc3.14160 (2024).
23. McGurrin, P., McNames, J., Haubenberger, D. & Hallett, M. Continuous monitoring of essential tremor: Standards and challenges. *Mov. Disord. Clin. Pract.* **9**, 1094–1098. https://doi.org/10.1002/mdc3.13558 (2022).
24. Lopez-de-Ipina, K. et al. Analysis of fine motor skills in essential tremor: Combining neuroimaging and handwriting biomarkers for early management. *Front. Hum. Neurosci.* **15**, 648573. https://doi.org/10.3389/fnhum.2021.648573 (2021).
25. Motin, M. A. et al. Computerized screening of essential tremor and level of severity using consumer tablet. *IEEE Access* **9**, 15404–15412. https://doi.org/10.1109/ACCESS.2021.3052186 (2021).
26. Yu, N.-Y., Van Gemmert, A. W. A. & Chang, S.-H. Characterization of graphomotor functions in individuals with Parkinson's disease and essential tremor. *Behav. Res. Methods* **49**, 913–922. https://doi.org/10.3758/s13428-016-0815-0 (2017).
27. Impedovo, D. Velocity-based signal features for the assessment of Parkinsonian handwriting. *IEEE Signal Process. Lett.* **26**, 632–636 (2019).
28. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. & Chen, L. C. MobileNetV2: Inverted residuals and linear bottlenecks. *IEEE* https://doi.org/10.1109/CVPR.2018.00474 (2018).
29. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. *IEEE* https://doi.org/10.1109/CVPR.2016.90 (2016).
30. Huang, G., Liu, Z., Maaten, L. V. D. & Weinberger, K. Q. Densely connected convolutional. *Networks* https://doi.org/10.1109/CVPR.2017.243 (2017).
31. Liu Z et al. A ConvNet for the 2020s. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 11966–11976. https://doi.org/10.1109/CVPR52688.2022.01167. (2022).
32. Russakovsky, O. et al. ImageNet large scale visual recognition challenge. *Int. J. Comput. Vision* **115**, 211–252. https://doi.org/10.1007/s11263-015-0816-y (2015).
33. Selvaraju, R. R. et al. Grad-CAM: Visual explanations from deep networks via gradient-based localization. *IEEE Int. Conf. Comput. Vis.* https://doi.org/10.1109/ICCV.2017.74 (2017).
34. Simonyan, K., Vedaldi, A. & Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. *Computer* https://doi.org/10.48550/arXiv.1312.6034 (2013).
35. Bain, P. G. et al. Assessing tremor severity. *J. Neurol. Neurosurg. Psychiatry* **56**, 868–873 (1993).
36. Grimaldi, G. & Manto, M. Assessment of tremor: Clinical and functional scales. In *Mechanisms and Emerging Therapies in Tremor Disorders* (eds Grimaldi, G. & Manto, M.) 325–340 (Springer, 2013).
37. Elble, R. et al. Task force report: Scales for screening and evaluating tremor: Critique and recommendations: Tremor Scales. *Mov. Disord.* **28**, 1793–1800. https://doi.org/10.1002/mds.25648 (2013).

Acknowledgements

We express our great gratitude to all the participants in the present study.

Author contributions

Yumeng Peng was responsible for the experimental design, data processing, and manuscript writing. Songliang

Han and Di Wu contributed to manuscript writing and the development of the algorithm models. Zhengbo Zhang provided overall project planning and reviewed the manuscript to ensure its accuracy and coherence. Longsheng Pan offered clinical guidance and managed patient interactions throughout the study. Zijin Zeng, Chenbin Ma, Wenjing Yang, Ping He, Tian Yuan, Ying Shi, and Lixuan Li were in charge of data collection and organization. Yumeng Peng, Wenjing Yang, Ping He, Zijin Zeng, and Longsheng Pan were responsible for rating the handwritten images. All authors gave final approval and agreed to be accountable for all aspects of the work.

Funding
This study is supported by National Natural Science Foundation of China (62171471).

Declarations

Competing interests
The authors declare no competing interests.

Ethics declarations
This study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Ethics Committee of Chinese PLA General Hospital (S2018-021–00/01). All methods were carried out in accordance with relevant guidelines and regulations. Informed consent has been obtained from the participants.

Additional information
Correspondence and requests for materials should be addressed to L.P. or Z.Z.