



OPEN ACCESS

EDITED BY

Junpei Zhong,
Hong Kong Polytechnic University, Hong
Kong SAR, China

REVIEWED BY

Antonio Parziale,
University of Salerno, Italy
Giuseppe De Gregorio,
University of Salerno, Italy

*CORRESPONDENCE

Mohamad Eid,
✉ mohamad.eid@nyu.edu

RECEIVED 24 March 2023

ACCEPTED 29 August 2023

PUBLISHED 13 September 2023

CITATION

Babushkin V, Alsuradi H, Jamil MH,
Al-Khalil MO and Eid M (2023), Assessing
handwriting task difficulty levels through
kinematic features: a deep-learning
approach.

Front. Robot. AI 10:1193388.

doi: 10.3389/frobt.2023.1193388

COPYRIGHT

© 2023 Babushkin, Alsuradi, Jamil,
Al-Khalil and Eid. This is an open-access
article distributed under the terms of the
[Creative Commons Attribution License
\(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or
reproduction in other forums is
permitted, provided the original author(s)
and the copyright owner(s) are credited
and that the original publication in this
journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted
which does not comply with these terms.

Assessing handwriting task difficulty levels through kinematic features: a deep-learning approach

Vahan Babushkin^{1,2}, Haneen Alsuradi¹,
Muhammad Hassan Jamil¹, Muhamed Osman Al-Khalil³ and
Mohamad Eid^{1*}

¹Applied Interactive Multimedia Lab, Engineering Division, New York University Abu Dhabi, Abu Dhabi, United Arab Emirates, ²Tandon School of Engineering, New York University, New York, NY, United States, ³Arabic Studies Program, New York University Abu Dhabi, Abu Dhabi, United Arab Emirates

Introduction: Handwriting is a complex task that requires coordination of motor, sensory, cognitive, memory, and linguistic skills to master. The extent these processes are involved depends on the complexity of the handwriting task. Evaluating the difficulty of a handwriting task is a challenging problem since it relies on subjective judgment of experts.

Methods: In this paper, we propose a machine learning approach for evaluating the difficulty level of handwriting tasks. We propose two convolutional neural network (CNN) models for single- and multilabel classification where single-label classification is based on the mean of expert evaluation while the multilabel classification predicts the distribution of experts' assessment. The models are trained with a dataset containing 117 spatio-temporal features from the stylus and hand kinematics, which are recorded for all letters of the Arabic alphabet.

Results: While single- and multilabel classification models achieve decent accuracy (96% and 88% respectively) using all features, the hand kinematics features do not significantly influence the performance of the models.

Discussion: The proposed models are capable of extracting meaningful features from the handwriting samples and predicting their difficulty levels accurately. The proposed approach has the potential to be used to personalize handwriting learning tools and provide automatic evaluation of the quality of handwriting.

KEYWORDS

artificial neural networks, deep learning, learning from demonstration, machine learning, sensorimotor learning

1 Introduction

Handwriting is a complex sensorimotor skill that combines sensory, cognitive, and motor processes to master. Forming legible handwriting requires a coordinated movement of the hand, arm and fingers while continuously evaluating the visual (formed letter) and the haptic (pressure from the pen) feedback received towards adjusting the trajectory, alignment and force of writing. The ability of children to produce fluent and readable handwriting is important for educational development, achievements in school, and self-esteem (Chang and Yu, 2014). The advances in assistive technologies, along with a better understanding

of the difficulty of handwriting, led to new and innovative methods of handwriting development/evaluation (Maor et al., 2011) and rehabilitation (Fancher et al., 2018). With the availability of smart devices (such as touch devices and hand tracking technologies), the analysis of handwriting has reached new heights, such as handwriting evaluation (Rosenblum et al., 2003), detection/prediction of handwriting difficulties (Drotár and Dobeš, 2020), and novel remediation/intervention methods (Fancher et al., 2018).

The transformation from traditional teaching methods to technology-enhanced learning paved the way for exploring personalized learning paradigms (Jenkins et al., 2016). In fact, research has found that personalizing the learning task for individual learners improves motivation, attitude, and learning (Corbalan et al., 2006). A fully personalized handwriting learning system would involve a mechanism to evaluate the quality of handwriting as a sensorimotor skill and provide a learning task with an appropriate level of difficulty. Giving learners a learning task that is too difficult or too easy can hinder the learning process and cause a downgrade (Curto et al., 2015). Traditionally, the difficulty level of a handwriting task is evaluated through expert assessment using task-specific performance measures. Researchers have developed known metrics for evaluating the difficulty of handwriting tasks, including the size and complexity of the shape of the handwriting task, the number and sequence length of stroke, and the quality of sensorimotor skill of the writer (Danna and Velay, 2015). Understanding the difficulty level of a handwriting task is a challenge due to the high subjectivity/diversity among experts' judgments. Furthermore, manually selecting a handwriting task for an appropriate level of difficulty is inefficient. Besides, automatically evaluating the difficulty level of a handwriting task is ill-defined. By using tablets and hand motion tracking, various handwriting features can be measured and analyzed, e.g., pressure, handwriting speed, in-air movement, stops and lifting of a pen, text shape, and time needed to complete the task, to evaluate the difficulty of a handwriting task. With the availability of data, it is tempting to utilize machine and/or deep learning to automatically classify the difficulty level of a handwriting task.

Machine learning has been used to detect if a child suffers from writing disabilities by using the kinematics of handwriting and questionnaires data (Pagliarini et al., 2017). Recent studies demonstrate the utilization of machine learning methods to diagnose dysgraphia (Mekyska et al., 2016; Asselborn et al., 2018; Asselborn et al., 2020; Dui et al., 2020). For instance, a Random Forest model was developed to identify dysgraphic children; the study included 54 third-grade children and used a 10-item questionnaire for Hebrew handwriting proficiency (HPSQ) (Rosenblum, 2008) to identify poor writing. Another study used commodity tablets and proposed a tool to diagnose dysgraphia (Asselborn et al., 2018); data from 298 children, including 56 with dysgraphia, were labeled using the Concise Assessment Scale for Children's Handwriting (Beknopte beoordelingsmethode voor kinderhand-schriften, BHK) test and utilized it to train a Random Forest classifier to detect dysgraphia. Results demonstrated that the proposed model has comparable accuracy to experts. A subsequent study proposed a serious games tablet application for handwriting skill screening at the preliterate stage (Dui et al., 2020). Instead of a global score based on the BHK test, the authors in

(Asselborn et al., 2020) proposed a detailed profile that incorporates the kinematics, pressure, pen tilt and static features (such as the letter shape) in the evaluation of handwriting skills. Based on this evaluation, personalized remediation tasks that are very specific to the individual child's needs can be developed. To our knowledge, no study examined both stylus and hand kinematics to evaluate the difficulty level of handwriting tasks.

This paper presents a data-driven model to provide an unbiased evaluation of the difficulty of the handwriting task, which in turn facilitates the personalized acquisition of the handwriting skill. The system will also be helpful for evaluating the quality of handwriting and informing the simulation of sensorimotor skills. Using the developed model, we examine prominent sensorimotor features based on kinematics that are found in both the hand (such as hand orientation and grip characteristics) and the stylus (including stylus orientation, pressure, and temporal characteristics). The proposed model is evaluated with data collected from five handwriting experts who completed the handwriting of all Arabic alphabet letters. We made the choice of using Arabic orthography as it captures a wider range of difficulty levels in handwriting (Naz et al., 2014), particularly due to its cursive nature, complex ligatures, and numerosity of topological features (Kacem et al., 2012). The model is expected to learn features that could deem a task difficult/easy. This task could go beyond the 28 letter of the Arabic alphabet such as letters-like shapes as well as word segments.

Two models have been considered—the one learning from the aggregated opinion of all experts about the difficulty of a given handwriting task, and the other learning the distribution of experts' estimates of the difficulty level of a given handwriting task. While the first model can tell in average what difficulty level will be assigned to the handwriting task, the second model can explicitly predict how many experts estimate the handwriting task of a particular difficulty level (in order to capture the subjectivity/diversity of expert assessment of the handwriting difficulty level). Furthermore, Shapley values are applied to identify features that are prominent for the evaluation of the difficulty of handwriting.

2 Proposed approach

2.1 Experimental setup

A snapshot of the experimental setup is shown in Figure 1A. The proposed setup consists of a HUION GT-116 tablet that is equipped with a pen-like stylus, and an Ultraleap Stereo IR 170 hand movement tracker. As shown in Figure 1, the hand tracking device is attached to a rigid stand in a position where the writer's hand movement can be reliably tracked. Note that the setup is portable to facilitate data collection at different sites.

2.2 Experimental task and protocol

The handwriting of all the 28 letters of the Arabic alphabet in standard form are included in the study. The experts are asked to write each letter separately with their own speed, initiating the process of recording by pressing the "Start" button on the screen and then "Stop" to submit the data or "Clear" to cancel the current

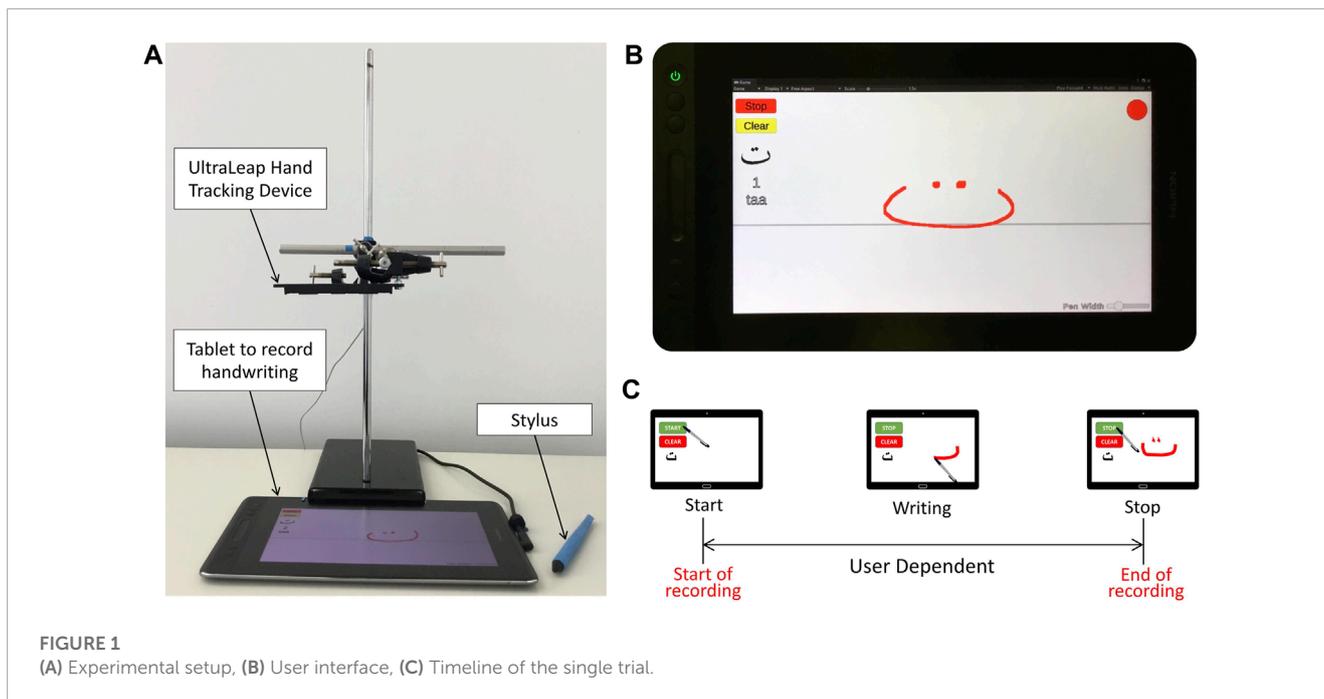


FIGURE 1
(A) Experimental setup, (B) User interface, (C) Timeline of the single trial.

recording and repeat the task for the same letter (Figure 1B). The 28 letters are presented in alphabetical order, starting from the first letter ا (“alif”) to the last letter ي (“yaa”). This is repeated 20 times for every letter so we have a total of 560 samples of handwriting to every expert. Experts are encouraged to take a few minutes break whenever needed.

2.3 Participants

Two groups of experts are recruited for this study. All experts were native Arabic speakers with no known disorders in handwriting. Five handwriting experts (3 female, 2 male, ages 35–55 years) are recruited to complete the handwriting tasks and 20 Arabic teaching experts (10 male, 10 female, ages 35–55 years) are recruited to complete the online survey. Experts were selected based on the following criteria: 1) more than 10 years of experience in teaching Arabic handwriting and 2) currently work in official (statutory work) or extra-official settings (non-statutory work). Additional inclusion criteria are defined for the group who are recruited to complete the handwriting tasks: 1) available for in-person meeting (for recording the handwriting tasks) and 2) no known disorders in handwriting. The 5 handwriting experts did not rate the difficulty level of the handwriting tasks. Note that the study was conducted in compliance with the Declaration of Helsinki, following its norms and regulations, and with an authorized protocol by the New York University Abu Dhabi Institutional Review Board (IRB: #HRPP-2020–12).

2.4 Data collection

A total of 117 spatio-temporal and kinematic features representing the point on the multidimensional skill path are

collected (see Table 1). The recordings from the tablet and the hand tracking device are synchronized to concatenate the two feature sets into one. The data were sampled with the rate of 25 Hz.

2.5 Data preprocessing

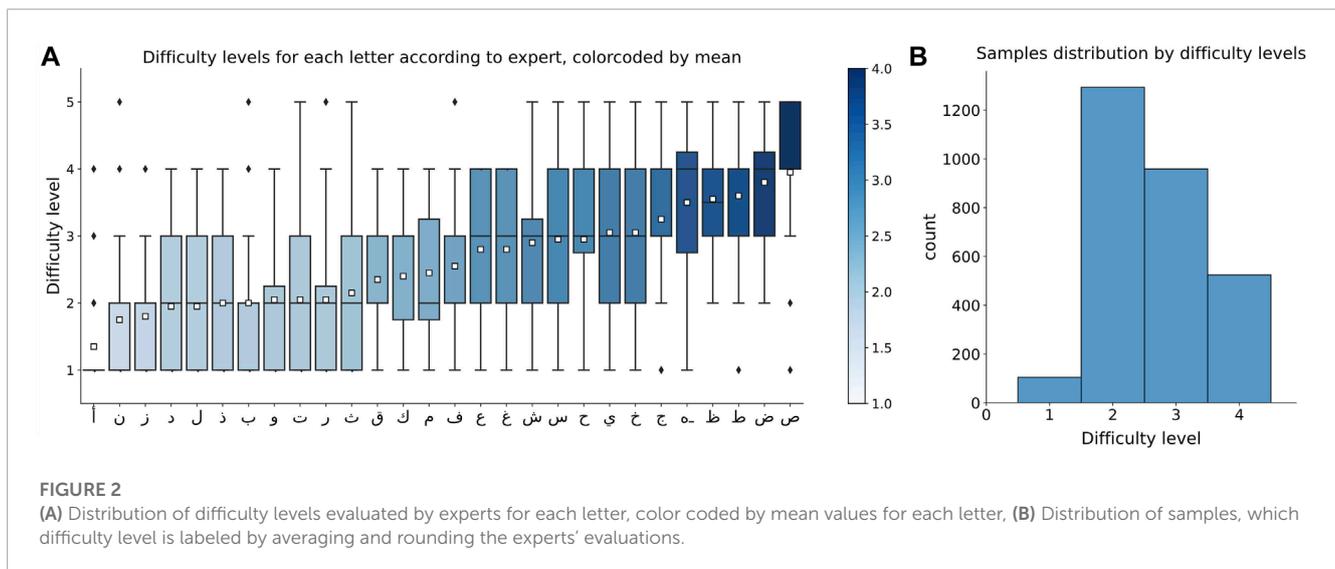
Due to the differences in the experts’ writing speed, the collected raw data represent a multivariate time series of 117 features of different duration, ranging from 15 time points to 234 time points. The data with duration less than 24 time points are excluded from the analysis, because it requires at least 1 s to accomplish the writing task and everything below this threshold represents a partially written letter. Also, the first and last two time points are trimmed to remove any noise during pen’s approaching to and distancing from the tablet surface. To make the data uniform, the features \times time points tensors are padded by zeros to create 117×234 matrices. All the variables are normalized before feeding them to the classification model.

2.5.1 Multiclass single-label classification

The traditional multiclass single-label classification is concerned with training a machine learning model on a set of samples, each having a unique class label from a set of disjoint class labels either binary (two classes) or multiclass (more than two classes) (Sorower, 2010). In this work the labels for multiclass single-label classification were obtained by extracting the mean values from the distribution of the difficulty levels evaluated by experts for each letter and rounding it to the nearest integer (Figure 2, A). It is clear that most of the letters were evaluated as of difficulty level 2 and 3 by the majority of experts. The experts were usually quite reluctant to mark the letter of difficulty level 5, therefore the taken means are always below 5 and therefore difficulty level 5 was excluded from consideration. The apparent imbalance in letters distribution by difficulty levels (Figure 2B) prompted us to seek another labeling approach that

TABLE 1 The 117 features collected during the experiment along with timestamp and hand ID (“Right”/“Left”).

Feature type	Feature name
Stylus kinematics	(x, y, z) coordinates of tooltip’s trajectory ($z = \text{const}$), pressure, azimuth, altitude, proximity
Hand kinematics	Fingers: the distal ends (x, y, z) coordinates of the bone for distal, intermediate, proximal, metacarpal for all 5 fingers and additionally a proximal end of the metacarpal bone for Index, Middle, Ring and Pinky fingers
	(x, y, z) coordinates of palm center
	(x, y, z) coordinates of hand pinch position (thumb and index if they are pinched)
	(x, y, z) coordinates of hand predicted pinch position
	(x, y, z) coordinates of hand wrist position
	(x, y, z) coordinates of elbow position (estimated if not in view)
	(x, y, z) coordinates of hand arm center (midpoint of the bone)
	(v_x, v_y, v_z) palm speed
	(n_x, n_y, n_z) hand palm normal
	(r_x, r_y, r_z, r_w) hand rotation
	Palm width, pitch, yaw, roll
	Hand pinch strength, hand pinch distance
	Hand grab angle
	Hand arm length (length of the bone)
	Hand arm width (average width of flesh around the bone)



reflects each expert’s opinion regarding the difficulty of a given letter while concurrently preserving the overall distribution of difficulty estimates. For this reason, the multiclass-multilabel classification approach has been adopted alongside multiclass single-label classification.

2.5.2 Multiclass-multilabel classification

In contrast to a single-label classification where class labels are mutually exclusive, the multilabel classification (or multi-output classification) is a generalization of multiclass single-label

classification that deals with multiple mutually non-exclusive repeated classes or “labels.” A given training set consists of n samples $S = (\mathbf{x}_i, \mathbf{y}_i)$, $i = 1, \dots, n$, $\mathbf{x}_i = (x_i^1, \dots, x_i^f) \in \mathcal{X}$, $\mathbf{y}_i = (y_i^1, \dots, y_i^l) \in \mathcal{Y}$, where \mathcal{X} is the instance space, \mathcal{Y} is the label vector space, f is the number of features and l is the number of labels in a label vector. The samples are independent, identically distributed and are randomly drawn from an unknown distribution D . The multilabel learning aims to produce a multilabel classifier $h: \mathcal{X} \rightarrow \mathcal{Y}$ that optimizes specific evaluation function (i.e., loss function) (modified from (Zhang and Zhou, 2007)). In this work a vector $\mathbf{y}_i = (y_i^1, \dots, y_i^l)$ of labels is referred to “label vector”.

Emerged from automatic text-categorization for medical diagnosis, the multilabel classification can be accomplished using Artificial Neural Networks that allow to specify the number of outputs in the output layer. The applications of multilabel classification can be found in music/movie categorization, where a single melody/movie can belong to different genres, or in semantic scene classification, where a picture can be categorized under more than one conceptual class at the same time (e.g., sunsets and beaches) (Tsoumakas and Katakis, 2007; Sorower, 2010).

To create the labels vectors for multiclass-multilabel classification, we define a labels vector of size 1×5 for each letter. Each label in the labels vector refer to the number of experts who evaluated the letter as a particular difficulty level. Each label can range from 0 to 20 (since there were 20 experts in overall). For example, for letter | “alif” the target label vector is [16,2,1,1,0], meaning that sixteen experts out of twenty evaluated it as of level 1, two experts as level 2, one expert as level 3 and as level 4 each and no one as level 5.

2.6 Feature selection: Shapley values

Shapley values, first introduced in the context of Game Theory (Shapley, 1953), are becoming more popular for interpreting the contribution of each feature in the prediction process of a trained ML model (Fryer et al., 2021). The idea behind Shapley values is to evaluate the influence of each feature on the outcome of the model by sequentially replacing each of the features, one at a time, with random uniformly-distributed values. A machine learning model is then repeatedly trained on the samples from the training set with the introduced randomly populated feature. Shapley values are obtained by calculating the difference between the outcome of the model considering the random feature and the outcome of the model considering the original feature. This is done on all instances of the validation set. Then, Shapley values are averaged across the samples of the validation set to get the overall influence of each feature. The model is iteratively retrained on the different subsets of the features, replacing a different feature with the random values each time. At the end, the distribution of averaged Shapley values is obtained, allowing to test the hypothesis if the replaced feature is more important than the introduced random feature.

In general, Shapley value of feature with index $f \in F = \{1, \dots, d\}$, where F is a set of all features' indices, can be defined as a weighted average over all marginal contributions $M_f(S)$ of f , where the marginal contribution of f to submodel $S \subset F$ is defined as a difference in evaluation when f is added to the submodel:

$$M_f(S) = C(S \cup \{f\}) - C(S) \quad (1)$$

where C is an evaluation function.

Thus, the Shapley value ϕ_f of feature of index f is

$$\phi_f = \sum_{S \in 2^{F \setminus \{f\}}} \omega(S) M_f(S), \quad (2)$$

and the weights are $\omega(S) = \frac{|S|!(|F|-|S|-1)!}{|F|}$ (Fryer et al., 2021).

3 Multiclass single-label classification

3.1 Architecture

The 1D CNN model for multiclass single-label classification contains two 1D convolution layers for data preprocessing and five fully connected layers for learning (Figure 3). The first convolution layer has 128 channels and accepts a matrix of 117 features and 234 time points as an input. The convolution is performed along the time dimension using 1×5 kernel. The resulting 128×230 matrix is passed to the input of the second 1D convolution layer with the same 1×5 kernel sliding along the time dimension. The output from the second 1D convolution layer is flattened and passed to the dense layers with 256, 128, 64, 32 and finally 4 neurons. To stabilize the learning process and to prevent overfitting, the batch normalization and dropout of 25% were applied after each layer. The Rectified Linear Unit (ReLU) activation was used in all layers except the last one (output layer) that uses Softmax as an activation function. The model was trained on 5000 epochs with Adaptive Moment Estimation (Adam) optimizer and sparse categorical crossentropy loss. The optimal learning rate of 10^{-4} was found empirically. Due to the significant imbalance in the number of samples for average difficulty levels 1 and 4 in comparison to difficulty levels 2 and 3 (see Figure 2) weights are used in the loss function to penalize higher for wrong prediction of levels 1 and 4 during the model training.

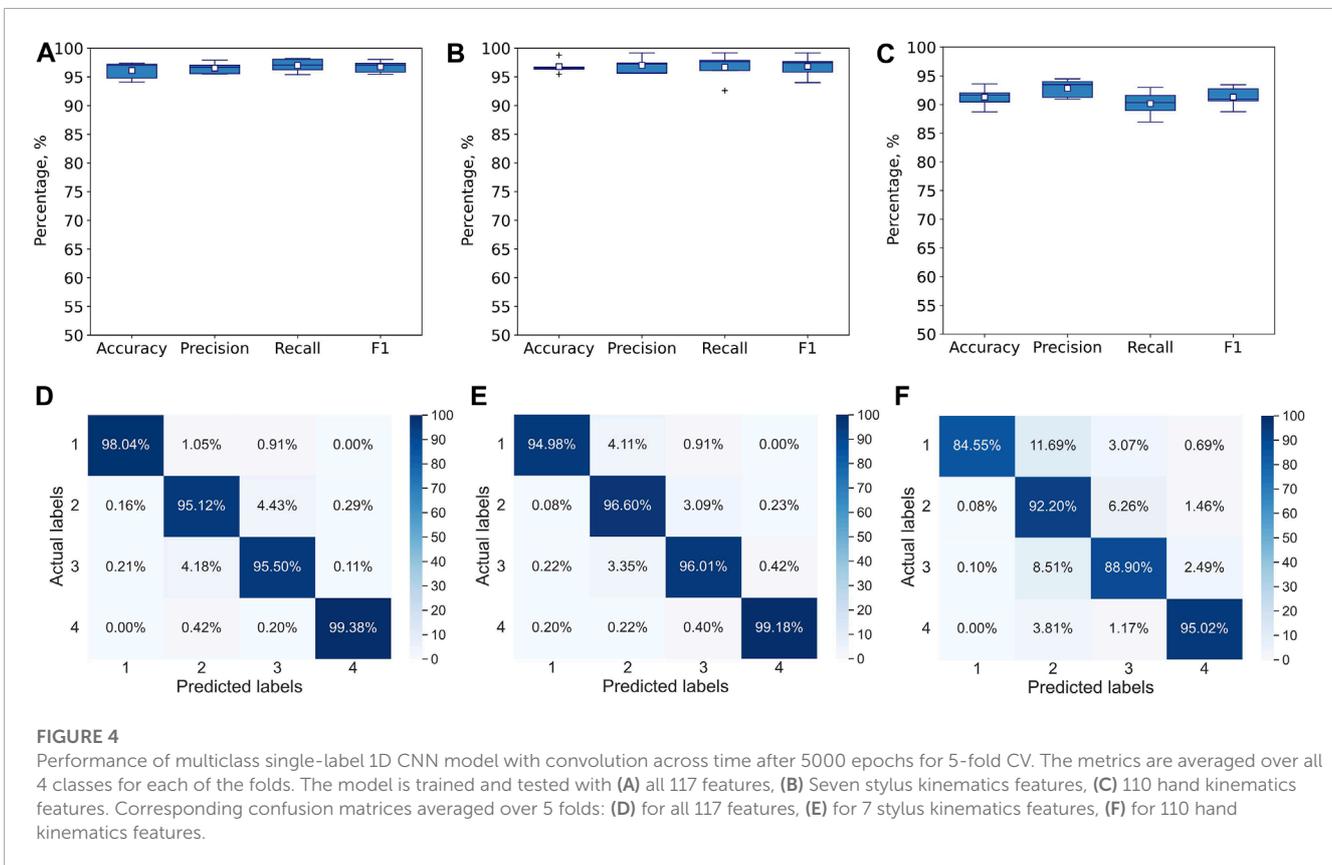
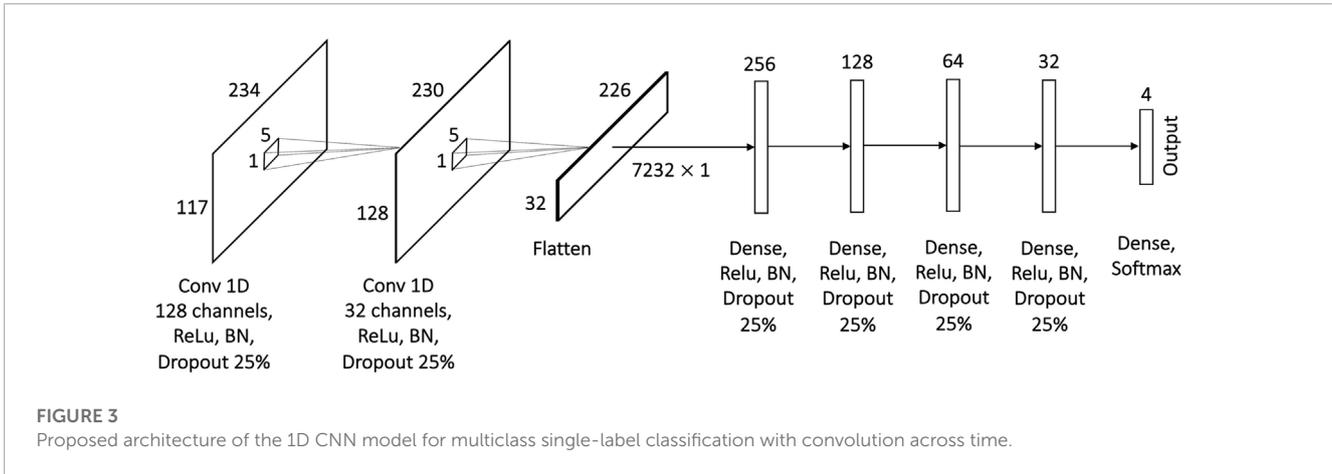
3.2 Results

3.2.1 Model evaluation

The 1D CNN model is evaluated using 5-fold cross-validation (Figure 4) and leave-one-expert-out cross-validation (Figure 5) in terms of accuracy, precision, recall and F1-score. To understand the effect of hand-kinematics related features, the model is trained on three different datasets. One dataset contained all 117 features, and the other two contained only the 7 stylus kinematics features and the 110 hand kinematics features, respectively. Training on all three datasets is run for 5000 epochs and evaluated with 5-fold and leave-one-expert-out cross-validation.

The 1D CNN model for multiclass single-label classification achieves accuracy of 96%, recall of 97% and precision of 97% after 5-fold cross-validation with all 117 features (Figures 4A, D). With the stylus kinematics features, the model achieves highest average accuracy, recall and precision were all 97% (Figures 4B, E). With the hand kinematics features, the average accuracy drops to 91%, the recall to 90%, and precision to 93% (4, C, F).

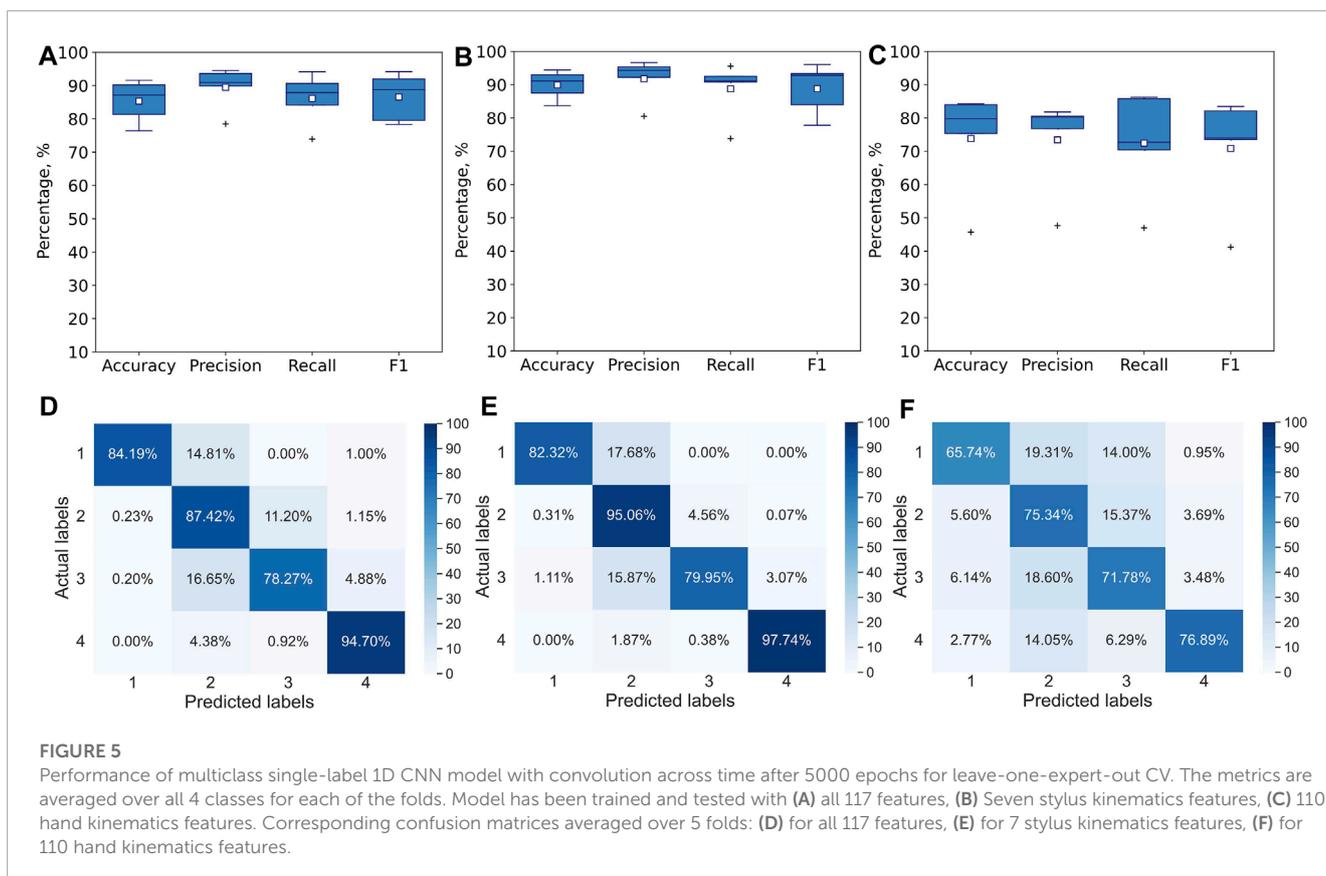
A similar analysis has been performed with leave-one-expert-out cross-validation. As expected, the performance of the model drops. The skill of handwriting is individualized and varies from expert to expert and since during the leave-one-expert-out cross-validation the model is trained with data from any of the four experts and tested on data from the fifth's, it learns handwriting features common to these four experts but fails to capture some of the individualized aspects of the handwriting of the fifth expert. For example, if the fifth expert is left-handed while the remaining four experts are right-handed, the model will learn only right-handed features determining difficulty levels (e.g., the tilt of the



pen) and will not be able to infer the difficulty level from the same features of the left-handed individual, if they differ significantly from right-handed participants. Similar to 5-fold cross validation, the highest accuracy of 90% with recall of 89% and precision of 92% has been achieved on the dataset with stylus kinematics features (Figures 5B, E), while the lowest one of 74% with recall of 72% and precision of 73% is observed on the dataset with the hand tracking features (Figures 5C, F). Despite the noticeable drop in performance, the model is still capable to use some general features to infer the difficulty level of the handwriting task.

3.2.2 Feature analysis

To determine which features are considered as significant for classifying handwriting tasks into four difficulty levels, Shapley values are calculated. The single-run accuracy of the trained model was 98%. Shapley values are calculated using randomly selected 100 instances from the training set as the background distribution and randomly selected 100 instances from the testing set for validation. For each of the 100 test instances, the calculated Shapley values are selected only for the classes that correspond to the true labels. The obtained Shapley values are averaged across 100 validation



instances. The 20 most prominent features for the multiclass single-label classification are shown in Figure 6. These results demonstrate that positional coordinates of the stylus and pressure data as well as roll and hand speed from hand tracking device are significant features for evaluating the difficulty level of the handwriting task.

4 Multiclass-multilabel classification

4.1 Architecture

The developed model for multiclass-multilabel classification has architecture similar to the multiclass single-label model. In other words, it also contains two 1D convolution layers for data preprocessing and five fully connected layers for learning (Figure 7). The input to the first convolution layer of 128 channels is a matrix of 117 features and 234 time points. The convolution is performed along the time dimension using 1×5 kernel. The resulting 128×230 matrix is passed to the input of the second 1D convolution layer with the same 1×5 kernel sliding along the time dimension. The output from the second 1D convolution layer is flattened and passed to the dense layers with 256, 128, 64 and 32 neurons. The only difference from the multiclass single-label model's architecture is that in the multilabel model the output of the penultimate layer with 32 neurons is passed to 5 parallel dense layers with 20 neurons and Softmax activation. Each one of the 5 output layers represents a difficulty level. A single output layer produces a number between

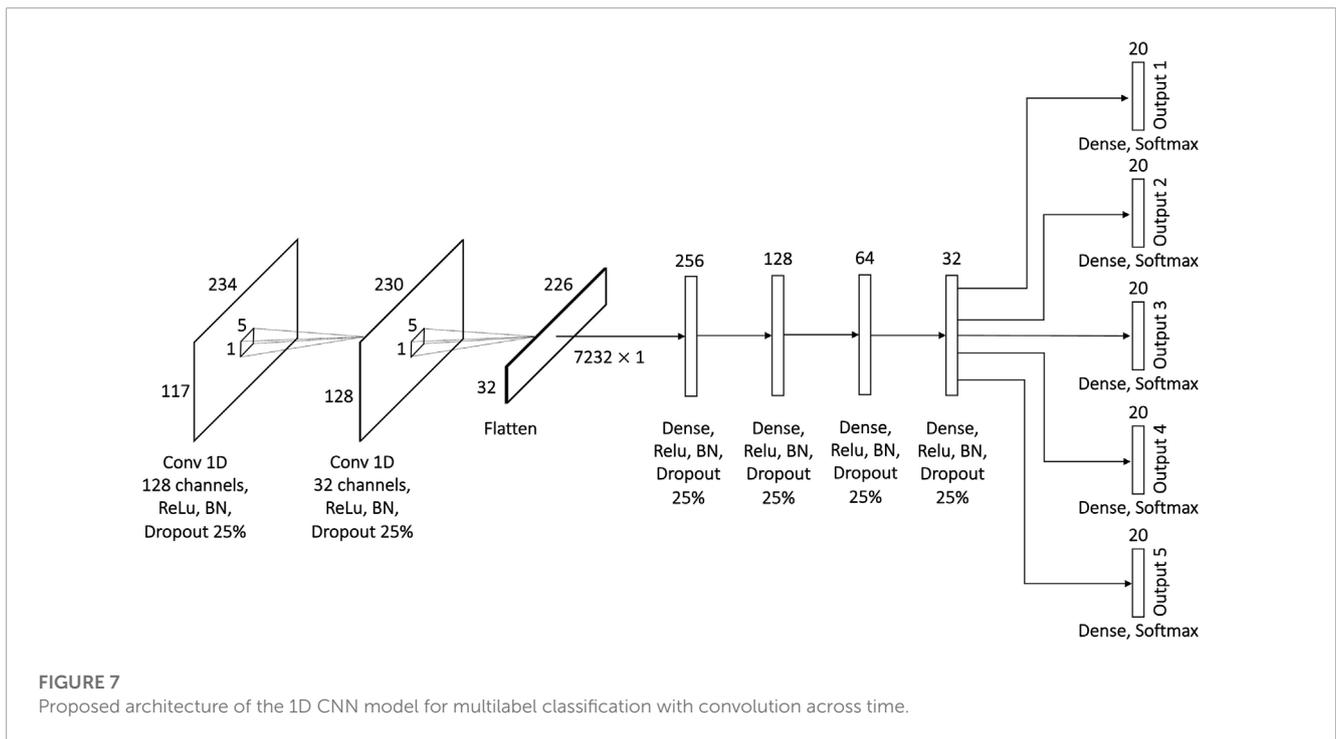
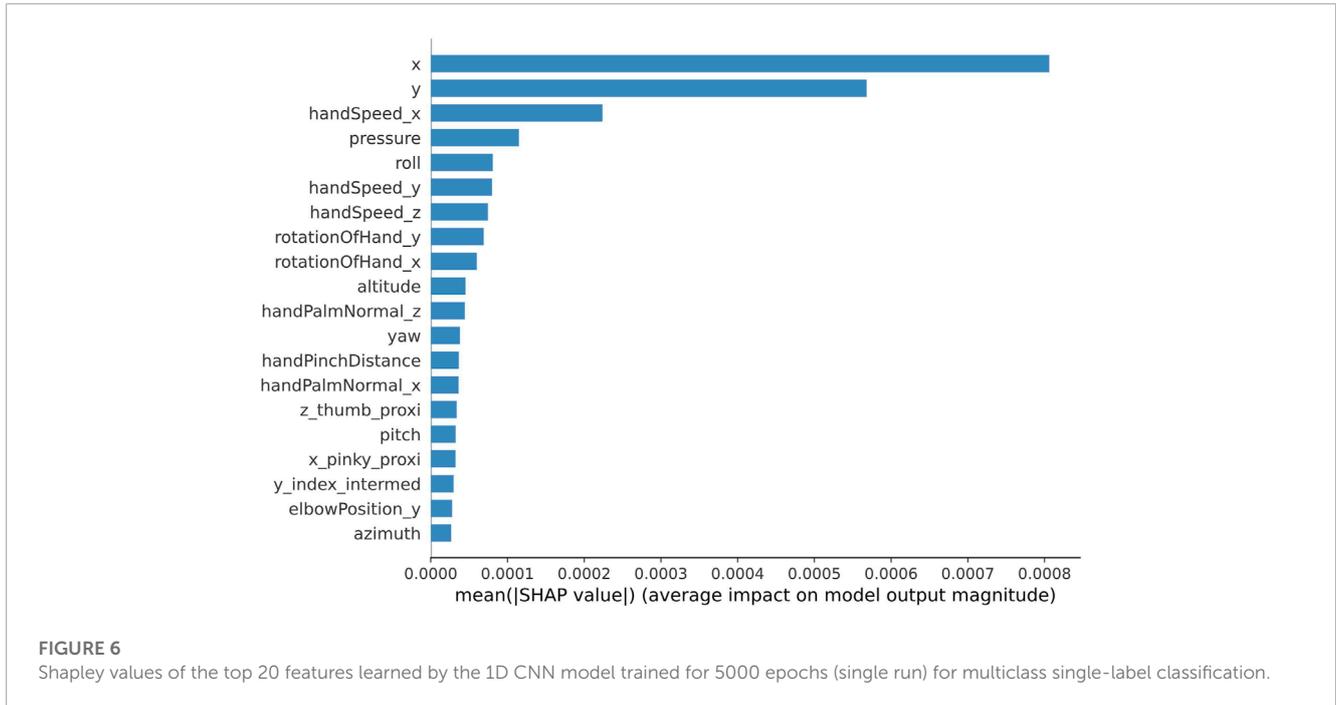
1 and 20 which represents the number of experts who estimated the letter as of a corresponding difficulty level. The outputs from each of these layers later combined to form a label vector of 5 elements. To stabilize the learning process and to prevent overfitting the batch normalization and dropout of 25% were applied after each layer. The Rectified Linear Unit (ReLU) activation used in all layers except the output layers. The model was trained on 5000 epochs with Adaptive Moment Estimation (Adam) optimizer and categorical crossentropy loss. The optimal learning rate of 10^{-4} was found empirically.

The network for multilabel classification outputs a label vector of 5 elements, whose index-position corresponds to the given difficulty level and values corresponds to the number of experts evaluating the letter with a particular difficulty level. We consider the performance metrics such as accuracy, precision, recall and F1-score separately for each of the difficulty level across 5 folds. The accuracy is averaged across 5 difficulty levels and 5 folds.

4.2 Results

4.2.1 Model evaluation

The 1D CNN model was evaluated using 5-fold cross-validation and leave-one-expert-out cross-validation. Apart from common metrics such as accuracy, precision, recall and F1-score Godbole and Sarawagi (2004), averaged over all samples for a given fold/expert and for each difficulty level, the Exact Match Ratio (EMR) (Sorower, 2010) was used to give additional insight into the model's



performance in case of the multilabel classification. The EMR score shows how many elements (labels) in predicted label vector equal to the labels in actual label vector (both labels' values and positions must coincide). The requirements were to predict both the label value and its position in the label vector. Formally speaking, if we denote

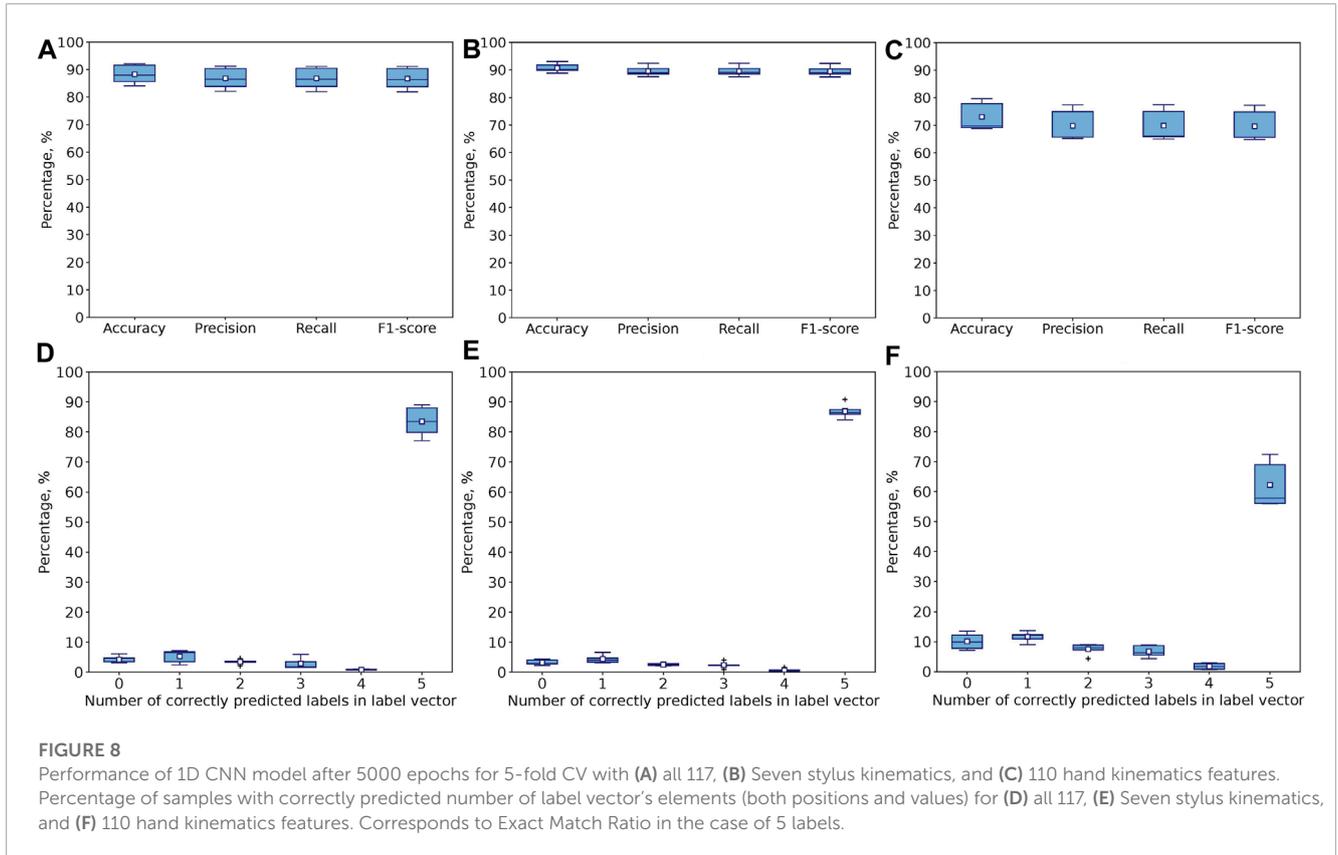
$$y_i = (y_i^1, \dots, y_i^l) \in \mathcal{Y}$$

–an actual label vector, where \mathcal{Y} is the label vector space, $i = 1, \dots, n$, n –number of samples, l is the number of labels in a label vector, and

$$\hat{y}_i = (\hat{y}_i^1, \dots, \hat{y}_i^l) \in \hat{\mathcal{Y}}$$

–a predicted label vector of n th sample, where $\hat{\mathcal{Y}}$ is the predicted label vector space, then the exact match ratio is defined as

$$EMR = \frac{1}{n} \sum_{i=1}^n \mathcal{I}(y_i = \hat{y}_i), \tag{3}$$



Where \mathcal{I} is the indicator function. formula 3 can be extended if we are interested in exact match of vectors with any number of correctly predicted labels. For example, in case of $l-1$ correctly predicted labels:

$$EMR^{l-1} = \frac{1}{n} \sum_{i=1}^n \mathcal{I} \left(y_i^j = \hat{y}_i^j \Big|_{j=2}^l \wedge y_i^1 \neq \hat{y}_i^1 \right) + \dots + \frac{1}{n} \sum_{i=1}^n \mathcal{I} \left(y_i^j = \hat{y}_i^j \Big|_{j=1}^{l-1} \wedge y_i^l \neq \hat{y}_i^l \right), \quad (4)$$

The accuracy was calculated as a ratio of the number of correctly predicted labels to the total number of labels for a given instance:

$$A = \frac{1}{n} \sum_{i=1}^n \frac{|y_i \cap \hat{y}_i|}{|y_i \cup \hat{y}_i|} \quad (5)$$

The precision was calculated as a ratio of the number of correctly predicted labels to the total number of actual labels averaged over all instances:

$$P = \frac{1}{n} \sum_{i=1}^n \frac{|y_i \cap \hat{y}_i|}{|\hat{y}_i|} \quad (6)$$

The recall was calculated as a ratio of the number of correctly predicted labels to the total number of predicted labels averaged over all instances

$$R = \frac{1}{n} \sum_{i=1}^n \frac{|y_i \cap \hat{y}_i|}{|\hat{y}_i|} \quad (7)$$

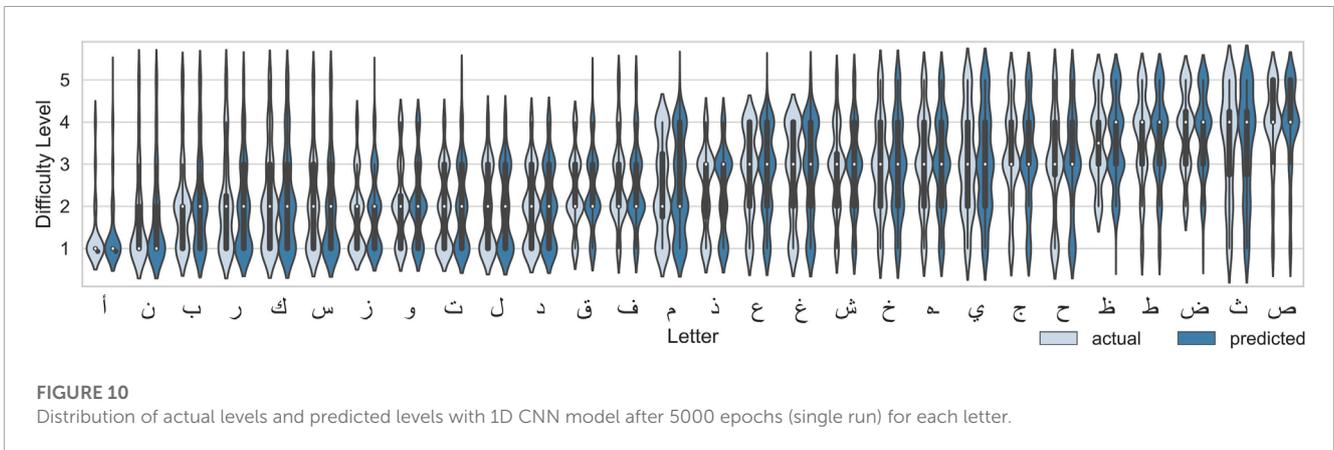
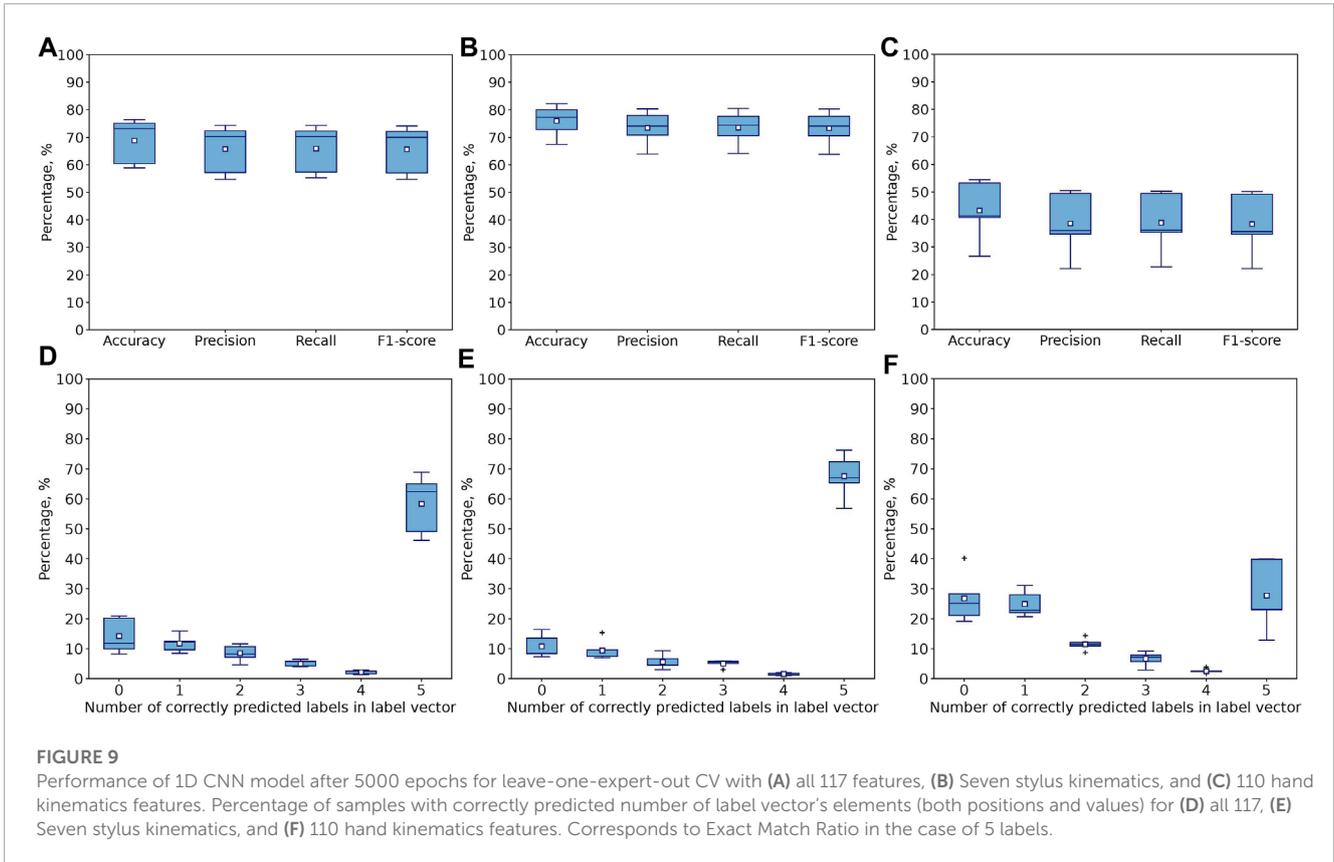
The F1 score was calculated as a harmonic mean of the precision and recall:

$$F_1 = \frac{1}{n} \sum_{i=1}^n \frac{2|y_i \cap \hat{y}_i|}{|y_i| + |\hat{y}_i|} \quad (8)$$

The metrics were averaged over all folds. For example, Figure 8D. Shows the performance of the model across 5 folds with all 117 features by comparing how many label vectors have at least one common element (both value and position are the same), two common elements, and so on. For 5 common elements, depending on the fold, the Exact Match Ratio varies between 77.1% and 89.1%. Notice that 3.1%–6.1% of all samples were predicted incorrectly (none of the elements of predicted and actual label vectors coincide).

The results of model's evaluation are demonstrated in Figure 8 (5-fold cross-validation) and Figure 9 (leave-one-expert-out cross-validation) with accuracy, precision, recall and F1 averaged over all samples for each difficulty level on the upper row (A., B., C) and the accuracy of predicting the exact value and position of the element corresponding to the difficulty level in the label vector on the lower row (D., E., F.). The model trained for 5000 epochs performs better for 5-fold cross-validation rather than leave-one-expert-out cross-validation, because the latter excludes the unique features, attributable to the experts from training, while the 5-fold cross-validation ensures that samples of all experts are equally present both in training and testing sets.

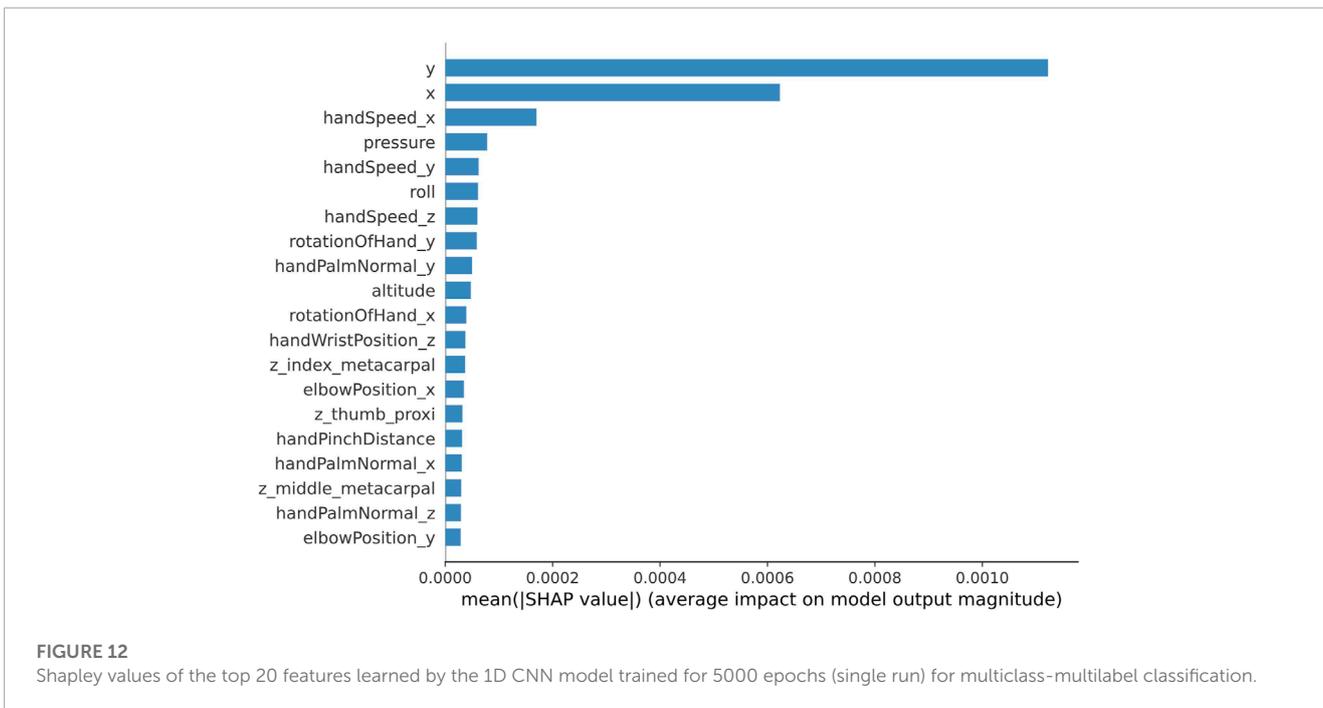
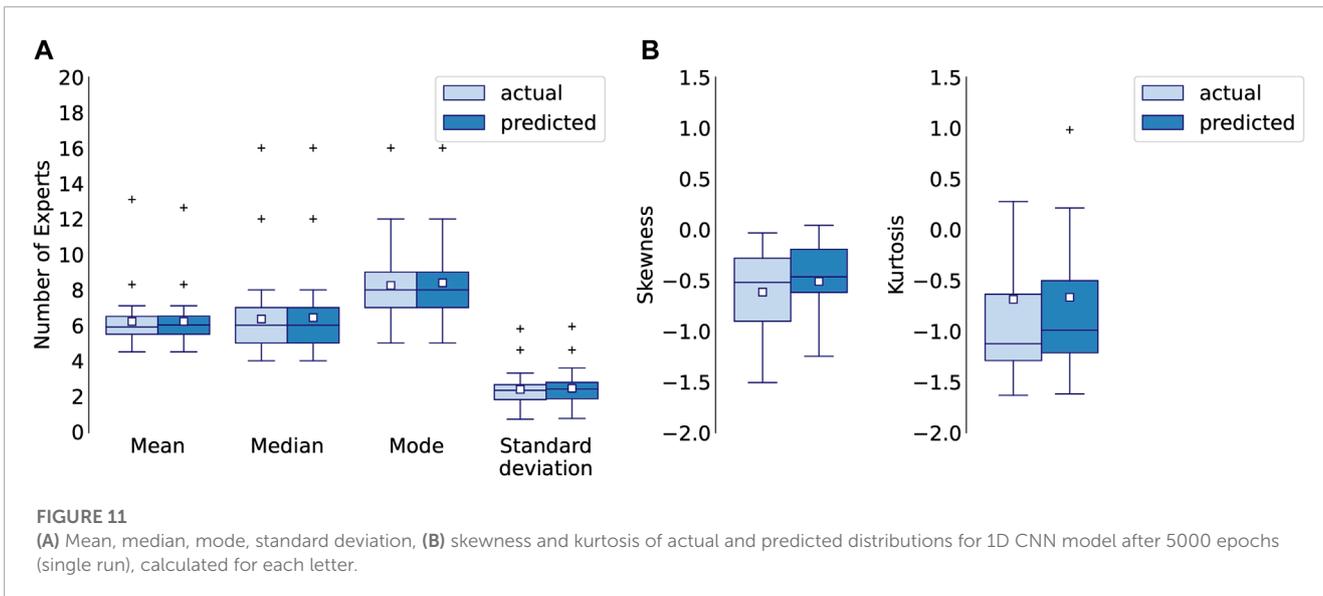
The prevalence of metrics obtained for the model trained and evaluated on the stylus kinematics features over hand kinematics features in the multiclass single-label classification is also observed for the multilabel case. It is more prominent for the leave-one-expert-out cross-validation (Figure 9C) but also true for the 5-fold cross-validation (Figure 8), where the use of 110 hand-tracking features results in increased variation in metrics across folds. For



example, the average accuracy score for 5-fold cross validation drops from 91% to 73% when switching from stylus kinematics features to hand kinematics features. For leave-one-expert-out cross validation this gap is more significant (the average accuracy drops from 76% to 43%).

It is also interesting to evaluate whether the 1D CNN model trained for 5000 epochs on the dataset containing all 117 features is capable of capturing the distribution of the number of experts (the values of each element) in the label vector. Figure 10 shows the violin plots for each letter comparing the actual and predicted distributions of number of experts who assigned the corresponding difficulty level to that letter. At the first glance, it is easy to notice that

most of the distribution shapes are predicted correctly. Moreover, some letters have a similar distribution, for example, most of the experts evaluated ا “alif” and ن “nuun” as belonging to level 1, while the letters ز (“zaay”), and و (“waaw”) were evaluated by the majority of experts as level 1 and 2 and as level 4 by only a few experts. In general the distributions are reconstructed correctly, with noticeable similarities in distributions of different letters (ب (“baa”) and ر (“raa”), or ح (“Haa”), خ (“khaa”), and ي (“yaa”). In some cases as ا “alif” most of the experts agree on the difficulty level, while for other cases like م (“miim”) there is almost uniform distribution of expert evaluations across the five levels; meaning that the experts cannot agree on how challenging it is to write this letter.



The 1D CNN model for multiclass-multilabel classification accurately predicts the parameters of the distribution of the number of experts across difficulty levels. Figure 11 A. compares mean, median and mode of the predicted and actual distributions and 11 B. shows the standard deviation, kurtosis and skewness of the predicted and actual distributions for each letter.

4.2.2 Feature analysis

Similar analysis with Shapley values was conducted for the multiclass-multilabel classification. The average accuracy of the model across all classes and all labels was 89.8%, the average

precision 90.2%, the average recall and F1 score were 90% and 89.8% correspondingly, and the Exact Match Ratio was 85.7%. Shapley values were calculated for every element of the label vector in the same way as shown previously with the single-label classification. Finally, the obtained Shapley values were averaged across 100 validation instances and then average along 5 elements of the label vector. Figure 12B. Shows the most prominent 20 features. These results mostly coincide with the results obtained for the multiclass single-label classification. It appears that the handwriting stylus positional coordinates, hand speed and roll from the hand tracking device and also the pressure data are the most prominent for evaluating the difficulty level of the handwriting task.

5 Discussion

In general, the 1D CNN model demonstrates better performance for 5-fold cross-validation in comparison to the leave-one-expert-out cross-validation, both for single-label and multilabel classification. The leave-one-expert-out cross-validation shows the ability of the model to rate the task included in the training set but executed differently by a left-out expert in comparison to the similar task in the training set. The model will perform well if it does not use the individual features of expert's sensorimotor skill for ranking the difficulty level of the handwriting task. Therefore, the lower performance of the model in leave-one-expert-out cross validation can be indicative of model's ability to focus on the features describing the individuality of expert's sensorimotor skill in the evaluation of the handwriting tasks. Similarly, during the 5-fold cross-validation for every fold the model is trained on 80% of the tasks coming from all the experts and therefore, it learns common features defining the difficulty level of the task. In this case, the high performance of the model might indicate its ability to pick up the features associated with the experts' generic skills.

Another common observation between the single-label and multilabel models is that both models perform the best when trained based on the stylus kinematics. On the other hand, relying on the hand kinematics (i.e. 110 features obtained from the handtracking system) results in a relatively compromised performance, due to the high dimensionality of feature space that leads to the model overfitting. Nonetheless, it is interesting to observe that it is possible to predict the difficulty level of a handwriting task with an accuracy that is considerably above the chance level by relying on hand kinematics only without any direct information about the trajectory of the task.

Explainability analysis with Shapley values provides further insight on which features contribute the most towards the classification of the difficulty level. The previously mentioned conclusion on the prevalence of handwriting trajectory features is consolidated by the results of Shapley values, where it has been demonstrated that the importance of the handwriting trajectory features, collected from the stylus, prevail over the hand kinematics features. The top 3 handwriting trajectory features are geometric coordinates of the stylus-tip (x and y) and the pressure while the top three hand kinematics features include hand velocity components and the roll, for single-label classification. For multilabel classification the most prominent are handwriting trajectory features as stylus tip coordinates, pressure, height (altitude) of the stylus tilt. These features prevail over the top most prominent hand kinematics features including all three components of hand velocity, and roll.

As mentioned before, the motivation for multiclass-multilabel classification is twofold: 1) The majority of the alphabets are ranked as difficulty levels 2 and 3, leading to highly unbalanced ratings as shown in Figure 2. 2) Alphabets with similar means (i.e., they have the same labels in multiclass single-label classification) could have very different underlying distributions. Both motivations are linked to the fact that the difficulty level of a handwriting task is subjective; and thus, alphabets are more characterized with distributions of difficulty levels rather than a single level. Figure 10 gives an insight on how well the multilabel model performed in classifying the

distribution of difficulty level compared to the actual one. It can be observed that different letters have different distribution patterns. For example, some letters have uni-modal distribution where experts somewhat agree on a difficulty level (easy, intermediate, or hard) while other letters are bimodal where experts have split opinions on the difficulty level of the letter (mode1: easy, mode2: hard). Other letters show a uniform distribution where experts have uniformly split over the different difficulty levels. These major differences in the distributions demonstrate the need of multilabel classification.

6 Conclusion

The main purpose of this work is to offer a system for automatic evaluation of the difficulty level of handwriting task. The traditional Machine Learning approach for a single-label classification relies on the difficulty level defined as an average of the distribution of difficulty levels evaluated by experts, rounded to the nearest integer. While this approach can give an idea of the overall difficulty of the handwriting task, the knowledge of the distribution of experts' evaluations allow to understand the extend of subjectivity of experts' judgment. A highly biased distribution with smaller standard deviations will indicate that most of the experts are agree on the difficulty level of the task, while the low values of standard deviation of multimodal distribution indicates the disagreement among the experts. The proposed multilabel model automatically determines the distribution of experts' opinions about the difficulty level of the task from the spatio-temporal features of the hand without asking the experts to evaluate it. This information will help to select the task that corresponds to the skill level of the writer. The model can be used in adaptive handwriting teaching systems that adjust the difficulty level of handwriting task based on the evaluation of the handwriting quality of a learner.

As per future work, we plan to adapt the model to evaluate the difficulty level of other handwriting tasks than Arabic letters. For instance, the model can be re-trained/modified to evaluate the difficulty level of handwriting in other languages or even abstract shapes. Furthermore, the collected data was acquired on a digit tablet. It would be interesting to evaluate the model with another acquisition system that allows writing on a physical paper (e.g., Wacom Bamboo) or on devices that simulate paper (e.g., reMarkable).

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: <https://osf.io/kq86d/>.

Author contributions

VB and HA contributed to the CNN model development and evaluation, VB, MJ, MA-K, and ME developed the data collection

protocol, VB and MJ designed and developed the experimental setup, VB collected the data. VB, HA, MA-K, and ME contributed to the conceptualization of the solution. All authors contributed to the article and approved the submitted version.

Funding

This research is supported by the ASPIRE Award for Research Excellence (AARE-2019) program (project number: AARE19-159). This work is also supported by the NYUAD Center for Artificial Intelligence and Robotics (CAIR), funded by Tamkeen under the NYUAD Research Institute Award, CG010.

References

- Asselborn, T., Dillenbourg, P., and Chapatte, M. (2020). Extending the spectrum of dysgraphia: A data driven strategy to estimate handwriting quality. *Sci. Rep.* 10, 3140. doi:10.1038/s41598-020-60011-8
- Asselborn, T., Gargot, T., Kidziński, L., Johal, W., Cohen, D., Jolly, C., et al. (2018). Automated human-level diagnosis of dysgraphia using a consumer tablet. *npj Digit. Med.* 1, 42. doi:10.1038/s41746-018-0049-x
- Chang, S.-H., and Yu, N.-Y. (2014). The effect of computer-assisted therapeutic practice for children with handwriting deficit: A comparison with the effect of the traditional sensorimotor approach. *Res. Dev. Disabil.* 35, 1648–1657. doi:10.1016/j.ridd.2014.03.024
- Corbalan, G., Kester, L., and Van Merriënboer, J. J. (2006). Towards a personalized task selection model with shared instructional control. *Instr. Sci.* 34, 399–422. doi:10.1007/s11251-005-5774-2
- Curto, P., Mamede, N., and Baptista, J. (2015). “Automatic text difficulty classifier,” in Proceedings of the 7th International Conference on Computer Supported Education, Lisbon Portugal, May 23 - 25, 2015, 36–44.
- Danna, J., and Velay, J.-L. (2015). Basic and supplementary sensory feedback in handwriting. *Front. Psychol.* 6, 169. doi:10.3389/fpsyg.2015.00169
- Drotár, P., and Dobeš, M. (2020). Dysgraphia detection through machine learning. *Sci. Rep.* 10, 21541. doi:10.1038/s41598-020-78611-9
- Dui, L. G., Lunardini, F., Termine, C., Matteucci, M., Stucchi, N., Borghese, N., et al. (2020). A tablet app for handwriting skill screening at the preliteracy stage: instrument validation study. *JMIR Serious Games* 8, e20126. doi:10.2196/20126
- Fancher, L. A., Priestley-Hopkins, D. A., and Jeffries, L. M. (2018). Handwriting acquisition and intervention: A systematic review. *J. Occup. Ther. Sch. Early Intervention* 11, 454–473. doi:10.1080/19411243.2018.1534634
- Fryer, D., Strümke, I., and Nguyen, H. (2021). *Shapley values for feature selection: The good, the bad, and the axioms*. arXiv. doi:10.48550/ARXIV.2102.10936
- Godbole, S., and Sarawagi, S. (2004). “Discriminative methods for multi-labeled classification,” in *Advances in knowledge discovery and data mining*. Editors H. Dai, R. Srikant, and C. Zhang (Berlin, Heidelberg: Springer Berlin Heidelberg), 22–30.
- Jenkins, S., Williams, M., Moyer, J., George, M., and Foster, E. (2016). *The shifting paradigm of teaching: Personalized learning according to teachers*. Retrieved from

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Knowledge Works: http://knowledgeworks.org/sites/default/files/ul/teacher-condition_s.pdf.

Kacem, A., Aouiti, N., and Belaïd, A. (2012). “Structural features extraction for handwritten arabic personal names recognition,” in 2012 international conference on frontiers in handwriting recognition (IEEE), Bari, Italy, September 18-20, 2012, 268–273.

Maor, D., Currie, J., and Drewry, R. (2011). The effectiveness of assistive technologies for children with special needs: A review of research-based studies. *Eur. J. Special Needs Educ.* 26, 283–298. doi:10.1080/08856257.2011.593821

Mekyska, J., Faundez-Zanuy, M., Mzourek, Z., Galáz, Z., Smekal, Z., and Rosenblum, S. (2016). Identification and rating of developmental dysgraphia by handwriting analysis. *IEEE Trans. Human-Machine Syst.* 47, 235–248. doi:10.1109/THMS.2016.2586605

Naz, S., Razzak, M. I., Hayat, K., Anwar, M. W., and Khan, S. Z. (2014). *Challenges in baseline detection of Arabic script based languages*. Cham: Springer International Publishing, 181–196.

Pagliarini, E., Scocchia, L., Vernice, M., Zoppello, M., Balottin, U., Bouamama, S., et al. (2017). Children's first handwriting productions show a rhythmic structure. *Sci. Rep.* 7, 5516. doi:10.1038/s41598-017-05105-6

Rosenblum, S. (2008). Development, reliability, and validity of the handwriting proficiency screening questionnaire (HPSQ). *Am. J. Occup. Ther.* 62, 298–307. doi:10.5014/ajot.62.3.298

Rosenblum, S., Parush, S., and Weiss, P. L. (2003). The in air phenomenon: temporal and spatial correlates of the handwriting process. *Percept. Mot. Ski.* 96, 933–954. doi:10.2466/pms.2003.96.3.933

Shapley, L. (1953). A value for n-persons games. *Ann. Math. Stud.* 28, 307–318.

Sorower, M. S. (2010). *A literature survey on algorithms for multi-label learning*, 25. arXiv.

Tsoumakas, G., and Katakis, I. (2007). Multi-label classification: an overview. *Int. J. Data Warehous. Min. (IJDWM)* 3, 1–13. doi:10.4018/jdwm.2007070101

Zhang, M.-L., and Zhou, Z.-H. (2007). ML-KNN: A lazy learning approach to multi-label learning. *Pattern Recognit.* 40, 2038–2048. doi:10.1016/j.patcog.2006.12.019