Optimizing Stroke Detection Using Evidential Networks and Uncertainty-Based Refinement

Faranak Akbarifar, Sean P. Dukelow, Albert Jin, Parvin Mousavi, and Stephen H. Scott

***Abstract*— Evaluating neurological impairments post-stroke is essential for assessing treatment efficacy and managing subsequent disabilities. Conventional clinical assessment methods depend largely on clinicians' visual and physical evaluations, resulting in coarse rating systems that frequently miss subtle impairments or improvements. Interactive robotic devices, like the Kinarm Exoskeleton system, are transforming the assessment of motor impairments by offering precise and objective movement measurements. In this study, we analyzed kinematic data from 337 stroke patients and 368 healthy controls performing three Kinarm tasks. Using deep learning methods, particularly an evidential network, we distinguished impaired participants from healthy controls while generating measures of prediction uncertainty. By retraining the network with the least uncertain samples and refining the test set by excluding the top 10% most uncertain samples, we improved the sensitivity of detecting subtle impairments in minimally impaired stroke patients (those scoring normal on the CMSA) from 0.55 to 0.75. We further extended the model to detect impairments associated with transient ischemic attack (TIA), resulting in an increased detection accuracy from 0.86 to 0.92. The model's ability to identify subtle motor deficits, even in TIA patients who show no observable symptoms on standard clinical exams, highlights its significant clinical utility. Detecting TIA is critical, as individuals who experience a TIA have a substantially higher risk of recurrent stroke. This work highlights the immense potential of integrating deep learning with uncertainty estimation to enhance the detection of stroke-related impairments, potentially paving the way for personalized post-stroke rehabilitation.**

***Index Terms*— Stroke assessment, deep learning, Kinarm, transient ischemic attack, uncertainty estimation.**

I. Introduction

CLINICAL assessment plays a pivotal role in the healthcare system, guiding patient care and evaluating the effectiveness of novel therapies. Whereas advancements have improved various clinical assessment tools, such as improved imaging techniques and novel blood-based biomarkers, evaluating brain function still heavily relies on visual and physical examinations by clinicians [1]. This is notably apparent for quantifying impairments associated with stroke. For example, the Fugl-Meyer Assessment (FMA) [2] and Chedoke-McMaster Stroke Assessment (CMSA) [3] quantify motor abilities. However, these assessment tools have a ceiling effect, where individuals receive the highest score suggesting no measurable impairments and yet still display visible impairments in motor skills [4], [5].

Recent advancements in robotics and motion analysis technologies have facilitated the development of innovative tools for the evaluation of motor function in stroke survivors [1], [6], [7], [8], [9], [10], [11], [12]. One such system is the Kinarm Exoskeleton Lab (Kinarm, Kingston, ON, Canada) [13] that quantifies upper limb sensory, motor and cognition functions using a suite of behavioural tasks called Kinarm Standard Tests (KST)^{TM} [13], [14]. During each task, this system records precise upper limb motion data, extracts spatiotemporal parameters, and leverages statistical models to compare a participant's performance against healthy individuals, considering factors such as age, sex, and handedness [1], [10], [15].

Machine learning techniques combined with kinematic measurements acquired by robotic systems have also demonstrated potential for relating participants' performance to clinical measures. Previous studies [16], [17] have used these kinematic features to predict a number of clinical scales using a simple neural network. The findings from these studies imply that these features not only capture a similar amount of information as clinical scores but potentially exceed them in informational depth. Notably in [18], a number of

Received 30 July 2024; revised 3 December 2024; accepted 10 January 2025. Date of publication 20 January 2025; date of current version 28 January 2025. This work was supported in part by the Canadian Institutes of Health Research (CIHR) Operating Grant MOP 106662; in part by the Heart and Stroke Foundation of Canada Grant-in-Aid G-13-0003029; in part by the Alberta Innovates Health Solutions Team Grant 201500788; in part by Ontario Research Fund—Research Excellence Grants ORF-RE 04-47 and RE-09-112; and in part by the Vector AI Institute, a Canada Canadian Institute for Advanced Research - Artificial Intelligence (CIFAR AI) Chair, Natural Sciences and Engineering Research Council of Canada (NSERC), and Queen's University. *(Corresponding author: Faranak Akbarifar.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Queen's University Health Sciences and Affiliated Teaching Hospitals Research Ethics Board (HSREB) and the University of Calgary Conjoint Health Research Ethics Board.

Faranak Akbarifar and Parvin Mousavi are with the School of Computing, Queen's University, Kingston, ON K7L 2N8, Canada (e-mail: f.akbarifar@queensu.ca; mousavi@queensu.ca).

Sean P. Dukelow is with the Hotchkiss Brain Institute, University of Calgary, Calgary, AB T2N 4N1, Canada (e-mail: sean.dukelow@albertahealthservices.ca).

Albert Jin is with the Department of Medicine, Queen's University, Kingston, ON K7L 3N6, Canada (e-mail: ayj@queensu.ca).

Stephen H. Scott is with the Centre for Neuroscience Studies, Queen's University, Kingston, ON K7L 3N6, Canada (e-mail: steve.scott@queensu.ca).

This article has supplementary downloadable material available at https://doi.org/10.1109/2025.3531768, provided by the authors.

Digital Object Identifier 10.1109/TNSRE.2025.3531768

classical machine learning techniques, as well as an artificial neural network, have shown promise in differentiating stroke participants from control groups based on the kinematic features acquired from a robotic reaching task. A more recent study [19] utilized an autoencoder-based classifier to distinguish least impaired (CMSA = 7, scored as normal) stroke participants from healthy controls, with an AUC equal to 0.84, yet room for improvement in classification sensitivity remains.

The current research extends this work by introducing a novel dimension of analysis - the estimation of confidence or uncertainty [20], [21] in the predictions generated by a deep learning model. An evidential network [22] is utilized to obtain these measures of confidence, which prove to be invaluable in refining both training and testing datasets. It is posited that instances for which the network expresses uncertainty represent regions of the data distribution where the model has encountered an insufficient number of samples. By systematically refining the datasets based on these uncertainty measures, the research aims to enhance the sensitivity and specificity of detection of stroke-related impairments.

Moreover, we expanded the model's application to a distinct yet related clinical challenge: the detection of impairments associated with transient ischemic attack (TIA) [23]. TIA, or "mini-stroke", is defined as a temporary neurological dysfunction resulting from reduced blood flow to specific areas of the brain without permanent brain damage as indicated by the absence of infarction on imaging studies [24].Symptoms like weakness, paralysis, or speech difficulties usually resolve within 24 hours, often within an hour. However, our recent work using KST highlights that many of these are identified as impaired compared to healthy controls 7 to 14 days after the TIA event [25]. Detecting TIA is of paramount importance because individuals who experience a TIA have a significantly higher risk of having a recurrent stroke. Studies have shown that the risk of stroke is particularly high in the days and weeks following a TIA, making early detection and intervention crucial to prevent subsequent strokes [26], [27], [28]. The hypothesis underlying extension of the current work to TIA is that the deep learning model, initially trained on stroke patient data, has, in essence, learned to detect impairment. Applying it to TIA patients explores its efficacy in distinguishing TIA individuals from healthy controls based on subtle, but measurable motor deficits.

This research presents a comprehensive investigation into the potential of deep learning, uncertainty estimation, and Kinarm Exoskeleton data in the context of stroke and TIA assessment of motor impairments. The study not only contributes to the broader discourse on neurological assessment but also addresses a critical clinical need by offering an innovative method for identifying impairments in individuals with mild impairments including individuals with stroke or TIA.

II. Methods

A. Participants

Participants with stroke and healthy volunteers who performed three Kinarm tasks—Visually Guided Reaching (VGR) [29], Object-Hit (OH) [30], and Object-Hit-and-Avoid (OHA) [31]—were recruited from Kingston, Ontario, and Calgary, Alberta. Clinical assessments validated stroke status using the CMSA and neuroimaging. The CMSA evaluates motor impairment post-stroke, focusing on the arm and hand, with scores ranging from 1 (severe impairment) to 7 (normal function). For the present study, the CMSA assessed the impaired arm, providing a standardized measure for motor recovery [3].

Demographics and clinical features of participants are summarized in Table I. Stroke participants were recruited from inpatient acute stroke or rehabilitation units at Foothills Medical Centre, Dr. Vernon Fanning Care Centre (Calgary, Alberta), and Providence Care, St. Mary's of the Lake Hospital (Kingston, Ontario). Inclusion criteria for stroke participants included a first clinical stroke within the last 35 days and being at least 18 years old. Exclusion criteria included other neurological disorders, upper limb orthopedic impairments, or difficulties understanding task instructions [32]. Furthermore, participants with multiple or bilateral strokes were systematically excluded from the dataset. Subsequent assessments were removed during data cleaning. Healthy controls met the same criteria, had no stroke history or neurological conditions, had normal or corrected vision, and no upper limb musculoskeletal injuries.

In addition, participants qualified for inclusion in this study if they had experienced a TIA, as defined by modern criteria [23], with no prior history of stroke or other symptomatic neurological conditions, and no history of orthopedic injuries to the upper extremities. Additionally, they needed to be asymptomatic and have received a National Institutes of Health Stroke Scale (NIHSS) score of 0 within the first 24 hours after presentation. However, our findings revealed that some participants initially scored as 0 on the NIHSS, indicating no observable neurological deficit, but later had scores of less than 7 on the CMSA. This discrepancy suggests that the NIHSS, used for rapid emergency assessments, may miss mild impairments detected by the more detailed CMSA, which takes about 15 minutes per limb to administer.

This study was reviewed and approved by the Queen's University Health Sciences and Affiliated Teaching Hospitals Research Ethics Board (HSREB) and the University of Calgary Conjoint Health Research Ethics Board. All participants gave their written informed consent to have their data collected for research purposes before performing the assessment.

B. Apparatus

Data were gathered using the Kinarm Exoskeleton Lab [13], [15], an interactive robotic system with integrated virtual reality for measuring upper limb sensorimotor function. Participants sat in an adjustable-height chair with robotic linkages aligned to their shoulder and elbow joints, allowing flexion and extension while supporting the arm against gravity (Fig. 1 **(a)**). The virtual reality display provided visual feedback aligned with the horizontal workspace, obscuring the direct view of the participants' arms.

TABLE I

DEMOGRAPHIC AND CLINICAL DATA FOR THE STUDY PARTICIPANTS, INCLUDING INDIVIDUALS IN THE HEALTHY CONTROL GROUP, THOSE WITH A HISTORY OF STROKE, AND THOSE WHO EXPERIENCED TRANSIENT ISCHEMIC ATTACKS (TIA)

	Control (n=368)	**Stroke (n=337)**	**TIA (n=30)**
Age	46 (18–93)	62 (18–92)	68 (49-88)
Sex	167 M, 201 F	221 M, 115 F, 1 O	13 M, 17 F
Dominant hand	330 R, 34 L, 4 A	310 R, 26 L, 1 A	28 R, 2 L
Days since stroke	...	16 (1-84)	11 (4-23)
Types of stroke	...	290 I, 45 H, 2 U	30 U
CMSA [1-7]			
Weakest arm	[0 ... 0, 368]	[16, 40, 48, 23, 58, 56, 96]	[0, 0, 0, 0, 0, 5, 25]
Strongest arm	[0 ... 0, 368]	[0, 0, 0, 1, 12, 61, 263]	[0, 0, 0, 0, 0, 6, 24]

Data are presented as the mean (range). Square brackets for CMSA scores indicate the actual number of individuals who obtained a given score on the test.
CMSA: Chedoke-McMaster Stroke Assessment, M: Male, F: Female, O: Other, R: Right, L: Left, A: Ambidextrous, I: Ischemic Stroke, H: Hemorrhagic Stroke, U: Unknown, Weakest arm: Affected/Non-dominant for stroke/control, Strongest arm: Less affected/dominant for stroke/control

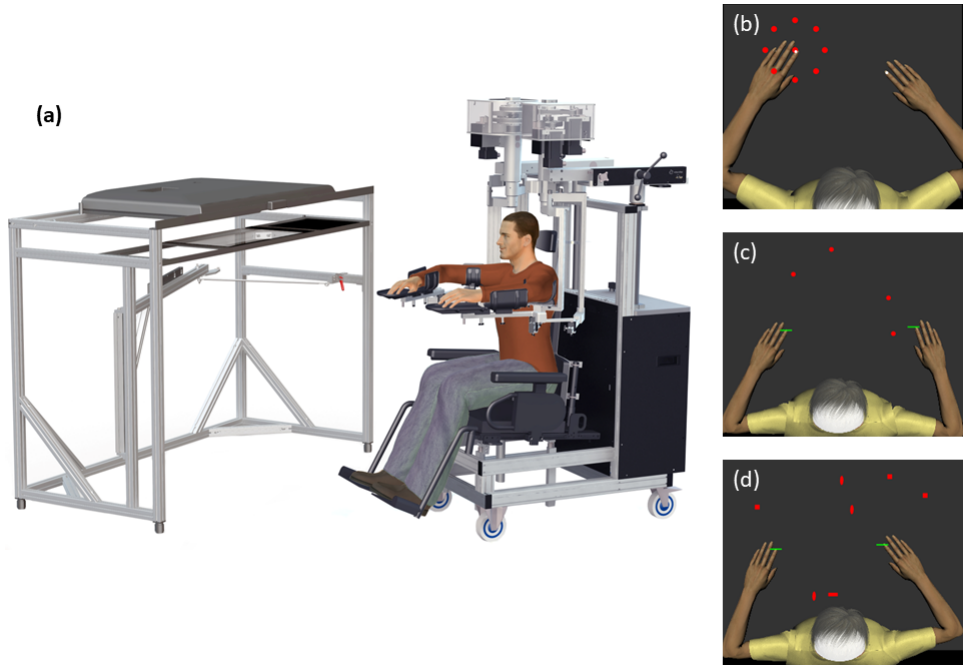

Fig. 1. Apparatus and tasks. **(a)** The Kinarm exoskeleton robot used for all tasks. **(b)** The virtual workspace for VGR. Participants must move their index finger (represented by a white circle) from the central target to one of the peripheral targets that appear randomly and evenly on a circle. **(c)** The virtual workspace for OH. Participants must hit as many red targets as possible using their fingertips (represented by green paddles). **(d)** The virtual workspace for OHA. Participants must hit as many targets as possible avoiding distractors. The targets' shape is communicated to the participant at the beginning of the experiment.

C. Tasks

The Kinarm Standard Tests (KST) [13], [25] are a suite of behavioral tasks designed to assess sensory, motor, and cognitive functions within 3 to 6 minutes each, allowing comprehensive assessment within an hour. The robotic system records hand movements, and automated analysis quantifies performance using task-specific spatiotemporal features. These features are combined into a single Task Score for each task, serving as a comprehensive performance metric [1], [14].

We included three Kinarm tasks, VGR, OH and OHA to broaden our investigation into motor skills. In the VGR task, similar to all KST, participants' view of their hands was obscured. The virtual reality system provided visual feedback on hand position and spatial objectives aligned within the horizontal workspace (Fig. 1 **(b)**). Participants moved a white circle, representing their hand, from a central target to one of eight peripheral targets, repeating each movement up to eight times for 64 trials. Fourteen movement parameters were derived from hand position and motion during each trial, detailed in [29].

In OH, participants used green paddles, representing their hands, to strike 300 red objects falling toward them within a

horizontal virtual plane (Fig. 1 **(c)**). The objects were evenly distributed across ten invisible bins along the horizontal axis, each releasing objects randomly and cycling through the process with increasing rate and speed, as detailed in [30].

In OHA, participants hit specific shapes of falling red objects and avoided others (Fig. 1 **(d)**). They were informed about the target shapes beforehand, with other shapes serving as distractors. Participants had to hit 200 targets while avoiding 100 distractors, as described in [31].

Within each task, control group parameters were standardized using Box-Cox transforms and linear regressions to account for age, sex, and handedness [1], [33]. This standardization ensures that demographic factors, such as age, sex, and handedness, are accounted for, minimizing their influence on the classification outcomes. As a result, demographic details were not directly included in the classifier, allowing the model to focus on motor function features relevant to stroke classification. To create a global Task Score, z-scores of task parameters were transformed: best performance mapped to 0 and poorer performance to positive values (absolute values for two-sided measures, zeta-transforms for one-sided measures). The root-sum-square distance of z-scores for healthy participants was Box-Cox transformed to achieve a standard normal distribution, then converted to zeta-scores. Stroke participant scores were converted using the same transformation parameters as controls. Task Scores for healthy controls followed a Normal distribution, identifying participants as impaired if their Task Scores exceeded 1.96, indicating performance at or beyond the 95^{th} percentile of controls [1], [14].

D. Data Preprocessing

We normalized all features across samples so that 1 denoted the poorest performance and 0 the best. Input feature vectors were constructed from VGR features for the affected and unaffected arms, followed by OH and OHA features adjusted to align with the affected/unaffected labeling. Originally, OH and OHA features were recorded as right/left. For the control group, the feature order was non-dominant/dominant. Each individual had 62-feature vectors and two label sets: stroke/control status and detailed CMSA labels.

E. Uncertainty Estimation Using Evidential Loss

Uncertainty estimation in deep learning is crucial for addressing the reliability and interpretability of model predictions. Techniques like Bayesian Neural Networks quantify uncertainty by integrating over the weights of the network [34]. Monte Carlo Dropout approximates Bayesian inference by applying dropout during inference [35]. These methods enhance the trustworthiness of AI systems, especially in high-stakes domains like healthcare and autonomous driving.

The evidential method introduced by Sensoy et al. [22] has gained popularity in the medical field for its computational efficiency, which is critical for timely patient diagnosis and treatment. Evidential networks assume that class prediction likelihoods follow a Dirichlet distribution, parameterized by $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_K)$, where K is the number of classes. The parameters α_k represent the "evidence" for each class, and the total evidence $\sum_{k=1}^{K} \alpha_k$ reflects the network's confidence in its predictions.

The evidential loss function consists of two key components:

1) **Prediction Loss**: This term ensures accurate class predictions by minimizing the expected negative log-likelihood:

$$\mathcal{L}_{\text{pred}} = \mathbb{E}_{\text{Dir}(\boldsymbol{\alpha})}[-\log p(y)], \tag{1}$$

where $p(y)$ is the predicted probability for the true class y.

2) **Regularization Loss**: This term penalizes overconfident predictions when the evidence is insufficient. The penalty is proportional to the total uncertainty:

$$\mathcal{L}_{\text{reg}} = \lambda \cdot \text{KL}[\text{Dir}(\boldsymbol{\alpha})||\text{Dir}(\mathbf{1})], \tag{2}$$

where λ is a regularization parameter, and $\mathbf{1}$ represents a uniform Dirichlet prior.

The total loss is given by:

$$\mathcal{L} = \mathcal{L}_{\text{pred}} + \mathcal{L}_{\text{reg}}. \tag{3}$$

Here, the evidence adjustment term $\lambda \cdot \text{KL}[\text{Dir}(\boldsymbol{\alpha})||\text{Dir}(\mathbf{1})]$ ensures that the model does not produce overconfident predictions without sufficient evidence, fostering a balance between prediction accuracy and uncertainty.

By using the evidential approach, the network can efficiently provide both accurate predictions and meaningful uncertainty estimates, making it particularly well-suited for medical applications where decisions must be made under uncertainty [22], [36].

Furthermore, the application of evidential networks extends beyond our study. Amini et al. [37] applied these networks in medical imaging, demonstrating their utility in quantifying diagnostic uncertainty. Malinin and Gales [36] explored their potential in predictive uncertainty estimation in deep learning, providing insights into their broader applicability. Moreover, the work of Der Kiureghian and Ditlevsen [38] underlines the importance of evidential networks in risk assessment, crucial for decision-making in uncertain environments.

F. SHAP Values for Feature Importance

Feature importance in machine/deep learning, pertains to methodologies that assign a rating to each feature according to their effectiveness in accomplishing the classification task. We used Shapley values, derived from cooperative game theory [39], to identify the most influential features. To calculate a Shapley value for a particular feature, we evaluate how its presence/absence influences the overall outcome. This is done by examining its contribution within different combinations or groupings of features. The Shapley value represents the average contribution of the feature across various scenarios. We applied the SHAP method [40] to compute these values for each feature, providing a detailed understanding of their importance. SHAP values help us interpret the impact of each feature on predictions when it deviates from a baseline value.

G. Structure of the Network

We chose a simple Multi-Layer Perceptron (MLP) as our deep classifier for the binary stroke/control classification task. We compared our model's performance with shallow methods: Support Vector Machine (SVM), Random Forest (RF), and Principal Component Analysis followed by Linear Discriminant Analysis (PCA-LDA). Our decision to use an MLP was motivated by the limited size of our dataset, as simpler models with fewer parameters are less prone to overfitting in such scenarios. Additionally, MLPs are versatile and computationally efficient, making them well-suited for our classification task. Our network had four dense layers, with the final layer and loss function modified for an evidential network. Each layer used the Leaky ReLU activation function and batch normalization. Networks were trained with the Adam optimizer, using a learning rate of 1×10^{-3} and a weight decay of 0.005 for L2 regularization. Training was performed with a batch size of 8. Instead of training for a fixed number of epochs, an early stopping mechanism was implemented to prevent overfitting by halting training when the validation loss did not improve for 10 consecutive epochs. A step-based learning rate scheduler reduced the learning rate by a factor of 0.1 every 7 epochs.

The model parameters were initialized using PyTorch's default initialization method. To ensure reproducibility, fixed random seeds were used when training a model. The evidential loss function was set to the Expected Cross Entropy loss, which balances prediction accuracy and uncertainty estimation.

These hyperparameter choices were optimized for the limited size of our dataset and the simplicity of the MLP architecture, ensuring computational efficiency and robust performance.

H. Data Stratification and Model Evaluation

Data were stratified so that each subject's data appeared in only one subset (60% train, 20% validation, 20% test). We trained 20 models using different random seeds and reported the mean ($\pm$ SD) of classification metrics, including accuracy, sensitivity, specificity, and AUC, on a test set of stroke and control participants (II, upper rows). The best model, selected based on AUC, is shown in the middle rows of Table II. We refined the train set by removing the top 10% uncertain data and reported the performance of the 10% model in the lower rows of Table II. Additionally, we used the trained 10% model to make inferences on the TIA test set.

III. Results

A. Participants

This study included Kinarm assessments of 337 stroke patients and 368 controls with no prior sensorimotor impairment. Demographics and clinical features are summarized in Table I. Ninety percent of controls and 92% of stroke participants were right-hand dominant, with four controls and one stroke patient having mixed handedness [41]. Additionally, 73% of stroke participants showed impairments based on the CMSA score of their most affected arm (Score < 7), with 86% having ischemic strokes and 14% hemorrhagic strokes.

TIA test set consisted of 30 participants assessed with Kinarm about two weeks after the TIA event [25], with data preprocessing as detailed in Section " Data preprocessing". Percentages of TIA participants performing below the 95th percentile of controls in terms of Kinarm Task Scores were 22.5% in VGR, 2.5% in OH, and 7.5% in OHA.

B. Data Visualization

Fig. 2 illustrates a Uniform Manifold Approximation and Projection (UMAP) visualization of all data utilized in this study. UMAP, effective for dimensionality reduction, captures intricate relationships in high-dimensional datasets, facilitating the exploration of detailed patterns and structures [42]. The depicted data points are color-coded: green for controls, magenta for strokes, and blue for TIAs. Marker sizes in the main plot correspond to CMSA values of the most affected arms of individuals with stroke and TIA (and non-dominant arms of control participants). Subplots only show participants with a specific CMSA value in the same mentioned color-code. Notably, participants with lower CMSA values are farther from control data points. This distinction is primarily evident in the distance between the most-impaired stroke participants, at the extreme left of the plot, and the control participants at the extreme right. While there are discernible overlaps in the areas predominantly associated with certain CMSA values, this might be attributed to the dimension reduction performed for 2D visualization purposes. TIA participants are typically not identified as impaired in the CMSA, with a small number scoring 6, as indicated in Table I. Scattering of these data points among other stroke and control participants underscores the need for a sophisticated method that utilizes all available features to differentiate the impairments related to this group.

C. Binary Classification Results on Stroke/Control Test Set

Upper rows of Table II summarize the binary classification performance in different experiments of this study using SVM, RF, PCA-LDA, and a 4-layer feedforward network adapted for evidence calculation. These methods exhibit strong performance, and their results are quite similar. However, the evidential network not only achieves classification accuracy but also provides insights into the network's uncertainty for individual data samples. Middle rows of Table II present the performance of the top models from the previous table, focusing on the least impaired stroke participants (CMSA Scores $= 7$ for both arms) and the controls. Notably, MLP-evidential method outperforms the others, increasing sensitivity to 0.55 compared to our previous study [19]. In our efforts to enhance this model's performance, we leveraged the network's uncertainty values to systematically eliminate the most uncertain data samples from the training set and then retraining the network. This process was iteratively applied, removing 5%, 10%, 20%, 30%, and 35% of the training data deemed most uncertain. The results highlight that the 10% model (trained after removing 10% of the most

uncertain data,) outperformed the others in sensitivity and AUC, with negligible compromises in specificity, as detailed in lower rows of Table II. Consequently, we adopted this superior-performing model for our subsequent experiments.

D. Uncertainty Analysis

The mean uncertainty for healthy controls was 0.43 (SD = 0.08, range = 0.30–0.74), while for stroke patients, it was 0.42 (SD = 0.09, range = 0.28–0.77). This demonstrates comparable uncertainty distributions between the two groups. Misclassified samples exhibited a higher mean uncertainty (0.53, SD = 0.10) compared to correctly classified samples (0.43, SD = 0.08), indicating that uncertainty reflects prediction confidence rather than solely misclassification.

We found it pertinent to delve deeper into the subset of the training set composed of the 10% "Removed" samples. Table III summarizes their statistical characteristics, showing that this uncertain data mainly includes control and stroke participants with comparable Kinarm Task Scores. We also filtered the test set by excluding the most uncertain samples. As shown in Fig. 3 **(b)**, removing just 10% of the most uncertain samples increased sensitivity from 0.55 to 0.75.

E. Distinguishing Impairments Associated With TIA

We applied our trained 10% model to the TIA test set, achieving an accuracy of 0.86. In other words, this model identified 86% of subjects as "impaired", effectively distinguishing them from healthy controls. This performance, achieved without integrating uncertainty values, surpasses the previously reported outcomes in Section " Participants" by 64% for the most severe impairment rate, which was based solely on VGR features.

Fig. 3 **(c)** illustrates the accuracy enhancement achieved by filtering the TIA set for uncertain samples. In line with the results from the earlier test set, removing 10% of uncertain data leads to an improvement in the accuracy of detecting TIA-related impairments, increasing it from 0.86 to 0.92.

F. Important Features Using SHAP Values

Having obtained the SHAP values for individual participants, we calculated the affecting classification outcomes. Fig. 4 shows the top 20 important features in each test set. Remarkably, 75% of these features are shared between the two data sets, though in different orders. Notably, speed-related features are prominent in the TIA dataset (Fig. 4**(b)**). Across both groups, features such as "no movement end" and "end target not reached," (Refer to [29] for a brief description) which reflect challenges in completing the reach or sustaining the hand at spatial targets after reaching, effectively measuring the number of successful movements in VGR, are among the pivotal factors. Important features also come from both arms of VGR and the number of successfully hit targets in the other two tasks.

Fig. 2. UMAP Visualization of Control (green), Stroke (magenta), and TIA (blue) participants achieved from the features of 3 Kinarm tasks; VGR, OH, and OHA. Marker sizes in the main plot are proportional to CMSA of weakest arms (affected arm of Stroke and non-dominant arm of Control). Subplots have focused on participants with specific CMSA scores.

IV. Discussion

According to our previous research [19], robot-based behavioral assessments provide a wealth of objective information. This information proves highly valuable, as it can be effectively harnessed by both shallow and deep machine learning methods to differentiate between the majority of stroke participants, including those with minimal impairment, and control participants (See Table II). In the current study, we applied an evidential network to discern between these groups, incorporating uncertainty quantification to enhance model reliability and interpretability. This method not only bolsters confidence in predictions but also provides a novel framework for assessing diagnostic certainty, paving the way

TABLE II

COMPREHENSIVE OVERVIEW OF CLASSIFICATION PERFORMANCES: THE UPPER ROWS PRESENT CLASSIFICATION PERFORMANCE RESULTS FOR SVM, RF, PCA-LDA, AND MLP-EVIDENTIAL CLASSIFIERS ON THE ENTIRE TEST DATASET, REPORTED AS THE MEAN (Â± STANDARD DEVIATION). THE MIDDLE ROWS SHOWCASE THE CLASSIFICATION PERFORMANCE OF THE BEST MODELS ON CMSA-7/CONTROL. THE LOWER ROWS ILLUSTRATE THE CLASSIFICATION PERFORMANCE OF THE MLP-EVIDENTIAL CLASSIFIER ON CMSA-7/CONTROL WITH VARIOUS TRAINING DATASETS GENERATED BY REMOVING 5%, 10%, 20%, 30%, AND 35% OF THE TOP UNCERTAIN DATA SAMPLES

Method	Accuracy	Sensitivity	Specificity	AUC
SVM	0.90 (±0.003)	0.83 (±0.006)	0.94 (±0.001)	0.90 (±0.003)
RF	0.90 (±0.003)	0.82 (±0.007)	0.99 (±0.005)	0.90 (±0.003)
PCA-LDA	0.90 (±0.002)	0.82 (±0.006)	0.98 (±0.005)	0.90 (±0.003)
MLP-evidential loss	0.90 (±0.001)	0.85 (±0.001)	0.95 (±0.004)	0.90 (±0.001)
SVM (BM)	0.87	0.5	0.98	0.74
RF (BM)	0.89	0.51	1	0.75
PCA-LDA (BM)	0.87	0.5	0.98	0.74
MLP-evidential loss (BM)	0.88	0.55	0.96	0.85
0% removed train set	0.88	0.55	0.96	0.85
5% removed train set	0.89	0.55	0.99	0.84
10% removed train set	0.88	0.59	0.96	0.86
20% removed train set	0.88	0.59	0.96	0.85
30% removed train set	0.88	0.55	0.97	0.84
35% removed train set	0.88	0.55	0.98	0.80

Note: (BM) stands for the best model. *x% removed train set* represents the original train set with x% of the top uncertain data samples removed.

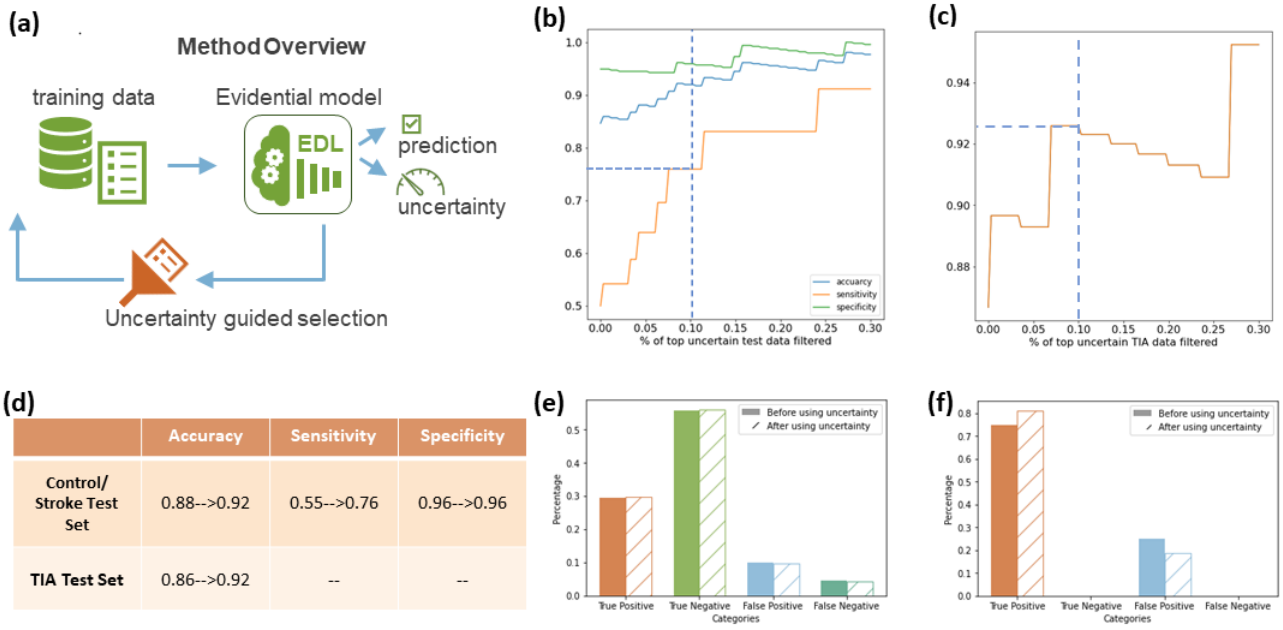

Fig. 3. Uncertainty Method. **(a)** Overview of the method illustrating the initial training of our evidential network on training data. Post-training, the network provides class predictions (stroke/control) along with uncertainty values. Subsequently, these values are employed to identify a subset of training data on which the network expresses higher confidence. This refined subset is then used for retraining the network. **(b)** Alteration in performance metrics following the filtration of the test set for the top x% of uncertain samples. Notably, the sensitivity in distinguishing minimally impaired stroke participants from controls increases from 0.5 to 0.76 by abstaining from making decisions about only 10% of the most uncertain data samples. **(c)** Shift in performance metrics after filtering the TIA set for the top x% of uncertain test samples. Demonstratively, the accuracy in detecting TIA subjects as impaired improves from 0.86 to 0.92 by refraining from decisions regarding 10% of the most uncertain data samples. **(d)** Table of alterations in classification performance metrics in the original and TIA test sets, after filtering 10% of uncertain test subjects. **(e)** Modification in the count of True Positive, True Negative, False Positive, and False Negative predictions before and after incorporating uncertainty values in the initial test set. (Positive: Stroke, Negative: Control). The reductions in all counts indicate instances where the network exhibits uncertainty even with correctly classified samples. **(f)** Revision in the count of True Positive, True Negative, False Positive, and False Negative predictions before and after the application of uncertainty values in the TIA test set.

for more personalized and adaptive rehabilitation strategies. Different uncertainty values are generated due to factors linked to both the data characteristics and the network's learning dynamics [22], [43], [44], [45]. Primary factors include the complexity and novelty of the data, as well as the level of noise within the data. High uncertainty often arises from samples divergent from the training set [46] and complex and ambiguous data types [47]. Considering

TABLE III
SUMMARY OF THE REMOVED 10 PERCENT MOST UNCERTAIN TRAINING DATA: ROWS 2-4 EXHIBIT THE MEAN TASK SCORES FOR THE VGR, OH, AND OHA TASKS OF EXCLUDED PARTICIPANTS, ALONG WITH THEIR RESPECTIVE STANDARD DEVIATIONS. PARTICIPANTS EXCEEDING 1.96 (THE 95TH PERCENTILE CUT-OFF BASED ON CONTROL NORMATIVE VALUES) ARE IDENTIFIED AS IMPAIRED

	Control	Stroke	Total
No. of removed	26 (60%)	17 (40%)	43 (100%)
VGR Task Scores	1.18 (±0.56)	1.77 (±0.74)	1.41 (±0.70)
OH Task Scores	1.21 (±0.82)	1.21 (±0.61)	1.21 (±0.74)
OHA Task Scores	1.62 (±0.80)	1.25 (±0.76)	1.47 (±0.80)
% of impaired in VGR	12	41	53
% of impaired in OH	15	12	27
% of impaired in OHA	23	18	41

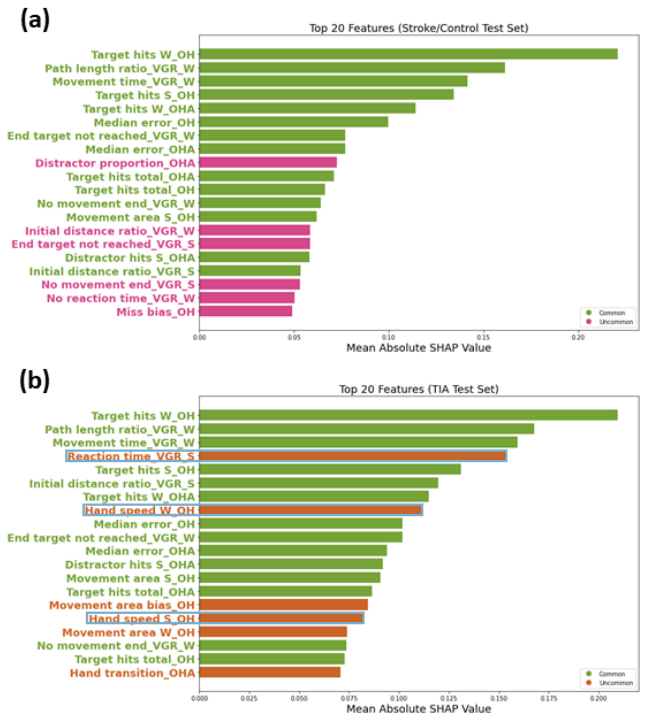

Fig. 4. Identification of top 20 important features for stroke prediction using the SHAP method. **(a)** in the stroke/control test set, **(b)** in the TIA test set. Common features between the two test sets (70%) are shown in green. Speed-related features in TIA are highlighted with blue frames. VGR: VGR feature, OH: OH feature, OHA: OHA feature, S: feature of the stronger arm, W: feature of the weaker arm.

number of participants in our data set and given that the efficacy of deep networks is intricately linked to the volume of training samples [48], [49], uncertain samples likely come from underrepresented distributions, leading to less reliable predictions. As shown in Table III, our observations suggest that the most uncertain samples performed similarly on Kinarm tasks. These samples included both control participants and some minimally impaired stroke participants who were clinically identified as having a stroke but exhibited task scores comparable to those of uncertain control participants, particularly on the OH and OHA tasks. This overlap suggests that the uncertainty arises due to similarities in task performances between these groups. Furthermore, this may indicate the existence of undetected subgroups within the control population, with similar performance to minimally impaired stroke participants. As a result, uncertainty values have a wide range of applications. They are crucial in data curation, helping to identify and flag samples that require additional analysis or inclusion in training datasets [50]. In high-stakes fields like medicine, these uncertainty estimates can guide clinical decision-making, prompting further review or alternative diagnostic approaches are employed when model confidence is low. Such practices not only incrementally refine model accuracy but also boost the safety and dependability of AI applications in critical areas.

As observed in the Results section, the application of an evidential network for refining our training set enhanced the detection of stroke participants with minimal impairments from healthy controls. It is crucial to mention that the notion of increasing accuracy by removing samples with lower certainty is not circular. When we remove the uncertain samples from the training set, we do not consider the classification predictions of either the training or test sets. Instead, we focus solely on the uncertainty values of the training samples. Thus, it is not trivial that the classification performance on the test set would improve by retraining the model with a refined training set and reporting the performance on the test set. It's noteworthy that rather than selecting a fixed uncertainty threshold, we removed the top 10% most uncertain training samples when this improved performance metrics. A similar approach was applied to the test set, where removing the top 10% most uncertain samples improved classification performance.

In real-world scenarios, during inference on the test set, the actual labels are often unknown. The only information available is the network's prediction and the associated uncertainty of that prediction. Consequently, when removing the most uncertain samples from the test set, there is no way to know if we are discarding correctly or incorrectly classified subjects. As depicted in Fig. 3**(e)** and Fig. 3 **(f)**, the changes in the number of true/false positives/negatives in the original and TIA test sets after filtering out 10% of the most uncertain samples indicate that the network shows uncertainty even for correctly classified samples. This reduction across all counts underscores that uncertain samples are not necessarily misclassified but rather those requiring further scrutiny. Furthermore, the mean and standard deviation of Task Scores in Table III indicate that the excluded data comprise control participants and minimally impaired stroke participants. Specifically, their performance on OH and OHA tasks deviated from the assigned binary label of control/stroke. Consequently, our network, while making decisions about these individuals, faces uncertainty due to conflicting signals from diverse feature sets. Notably, the distinctive features related to the number of hit objects and distractors, integral for tasks involving object interaction, closely align with those identified as significant for binary stroke/control classification

based on SHAP values, as depicted in Fig. 4. By setting these uncertain samples aside, we identify subjects who need more thorough investigation to detect their impairments accurately. Therefore, the process of removing uncertain samples serves not only to enhance model performance but also to highlight areas where additional data or analysis is needed, ultimately contributing to the development of more robust and reliable models. Hence, the network's performance can be accurately evaluated without including these samples and leaving those for further examination by the expert. Future studies using this approach could lead to a better identification of patient subgroups, which would be helpful for both research and clinical purposes.

As highlighted in the Results section, the utilization of the SHAP method offered valuable insights into the most significant features within both test sets. In the TIA dataset, the notable contribution of speed-related features to classification as "impaired" suggests that participants in this group perform their tasks at a significantly slower rate (Fig. 4). Features like "no movement end" and "end target not reached" in both test sets aligns with our previous study, which identified such difficulties as a reason for subclustering within various CMSA levels of patients [19]. Furthermore, it is worth noting that features from both arms of VGR are instrumental, justifying the inclusion of features from both arms. In particular, features related to the number of successfully hit targets have been selected. This observation underscores the ability of the network to discern important features, such as the number of hits, as might be expected when assessing OH and OHA tasks through intuitive human observation. This stands in contrast to the statistical approach, which assigns equal weight to all features, potentially overlooking these nuanced but significant distinctions. The SHAP analysis revealed that the top five features had a significant influence on classification outcomes, shedding light on the most clinically relevant motor impairments distinguishing stroke from healthy individuals. While these features are highly informative, our results indicate that relying solely on the top five features would not have achieved the same level of performance. The inclusion of additional features allowed the model to capture more subtle impairments, thereby enhancing the robustness of the classification. Clinically, this suggests that a broader assessment of motor functions provides a more comprehensive understanding of a patient's condition, which is essential for designing effective, personalized interventions.

Our research aims to expand the limited literature on applying machine learning techniques to analyze robotic assessment data in stroke diagnosis. This field of machine learning in robotic assessment remains relatively novel, and our findings contribute to the field by showcasing the potential accuracy of this approach for classification tasks for detection of subtle impairments and uncertainty estimation. While previous works like [16], [17] focused on regression tasks predicting clinical stroke scales within different robotic systems and tasks, our focus on stroke/control classification aligns with their conclusions: that robotic features hold comparable information to clinical scores. Comparatively, while Agrafiotis et al.'s [16] work also harnesses the potential of robotic systems for clinical assessment, our study advances the methodology by integrating task feature weighting, allowing for a more refined analysis of feature importance in stroke detection. Additionally, the researchers in [18] utilized data from the Kinarm and VGR task but with a smaller sample size. Their reported 94% accuracy using a simple artificial neural network exceeds our average but lacks sensitivity and specificity details, nor does it cover detecting impairments in minimally impaired stroke participants. Contrarily, [19] employed deep learning, providing insights into impairments, yet our study, leveraging deep neural networks with evidential losses, excels in detecting subtle impairments and estimating uncertainty. Our approach aids clinicians by identifying samples for further examination, enhancing stroke detection sensitivity. Additionally, our network outperformed the method in [25] by 45% in detecting TIA-related impairments. Our emphasis on task feature integration and weighting, allowing the network to learn feature importance, could account for this improved performance compared to methods giving equal weight to all features.

A. Clinical Implications

Our study carries potential clinical implications. By leveraging machine learning models trained on Kinarm assessment data, clinicians can adopt a data-driven approach to assess stroke-related impairments, thereby fostering the development of patient-centered rehabilitation strategies. The advancements in robotic assessments for monitoring stroke recovery and identifying distinct patient subgroups are likely to transform post-stroke recovery. These innovations will not only pave the way for personalized medicine but also enhance future research into the anatomical and pathophysiological mechanisms behind specific movement abnormalities.

The ability to estimate uncertainty adds an additional layer of information that can guide clinical decision-making. By identifying cases with higher uncertainty, clinicians can prioritize further evaluation, potentially leading to earlier intervention and improved patient outcomes. Additionally, for minimally impaired subjects who score normal on clinical tests such as CMSA, our classifier can detect subtle impairments that may otherwise go unnoticed. For these subjects, rehabilitation plans can be tailored and continued until the classifier identifies them as normal with an acceptable level of uncertainty. This targeted approach ensures that even subtle motor deficits are addressed, facilitating more personalized and effective rehabilitation strategies. Moreover, our application of the model to TIA patients highlights its versatility. Detecting impairments following TIA is important as timely diagnosis can prevent subsequent strokes. Our model's superior performance in differentiating TIA patients from healthy controls, compared to previous statistical methods, underscores its potential clinical relevance and potential utility in a broader range of neurological assessments.

B. Limitations

It is essential to acknowledge the limitations of our study. Firstly, our study relies on a specific set of Kinarm assessment

tasks, and the generalizability of our model to other stroke assessment protocols requires further investigation. These Kinarm tasks identify behavioural impairments, but do not provide any direct information on the pathophysiological mechanisms of these impairments. Furthermore, our research primarily focuses on binary classification between stroke patients and healthy controls. Future studies could explore more granular classifications, such as distinguishing different stroke subtypes or assessing the severity of impairment. While the average age of the healthy subjects in our dataset was lower than that of the stroke subjects, we ensured that this did not influence the classification outcomes. As described in Subsection C (Tasks) of the Methods section, we used z-scores standardized via Box-Cox transforms and linear regressions to account for potential confounders, including age, sex, and handedness. By employing these standardized scores, the classification accuracy reflects the motor function impairments rather than demographic differences, thereby mitigating the impact of age on the results. Finally, the present approach assumes that individuals in our healthy cohort do not have any underlying neurological impairments. Interestingly, many of the most uncertain data samples were from individuals in the healthy cohort and may reflect minor changes in performance related to underlying undiagnosed conditions.

V. Conclusion

In this manuscript, we demonstrated how deep learning techniques can be utilized to analyze data from the Kinarm, a robotic upper extremity motion assessment tool, to distinguish impairments related to strokes or TIAs from healthy control behavior. Our study advances the application of machine learning in stroke assessment through kinematic performance metrics, demonstrating not just competitive accuracy but the added value of uncertainty estimation for clinical support. The capacity to gauge prediction uncertainty enhances identification of out-of-distribution samples, refining data sets that will lead to a more reliable model which will potentially lead to more information about data quality and subgroups of patients. Notably, a significant number of uncertain samples were from healthy controls, indicating a broad behavioral spectrum. By training our network with both stroke and healthy data, we effectively distinguished impairments in TIA participants from controls, showcasing our network's ability to detect subtle impairments. This was particularly evident in the detection of TIA participants, who, despite exhibiting no symptoms on standard clinical exams and no apparent structural damage using imaging techniques, were identified through the sensitivity of the Kinarm tool and kinematic measures. Machine learning's enhanced sensitivity in detecting these impairments underscores the potential of our approach. While acknowledging limitations and the need for further research, our findings hold promise for assessing mild stroke and TIA impairments, highlighting the critical role of machine learning in augmenting neurological assessments.

Acknowledgment

The authors would like to thank Simone Appaqaq, Helen Bretzke, Ethan Heming, Kimberly Moore, and Justin Peterson for their roles in recruiting and assessing participants and in managing the data after collection. They declare that S. H. Scott is the Co-Founder and CSO of BKIN Technologies (dba as Kinarm), the company that commercializes the robotic technology used in the present study. All other authors confirm no conflict of interest.

References

[1] S. H. Scott, C. R. Lowrey, I. E. Brown, and S. P. Dukelow, "Assessment of neurological impairment and recovery using statistical models of neurologically healthy behavior," *Neurorehabilitation Neural Repair*, vol. 37, no. 6, pp. 394–408, Jun. 2023.

[2] A. R. Fugl-Meyer, L. Jääskö, I. Leyman, S. Olsson, and S. Steglind, "The post-stroke hemiplegic patient. 1. A method for evaluation of physical performance," *Scand. J. Rehabil. Med.*, vol. 7, no. 1, pp. 13–31, 1975.

[3] C. Gowland et al., "Measuring physical impairment and disability with the Chedoke–McMaster stroke assessment," *Stroke*, vol. 24, no. 1, pp. 58–63, Jan. 1993.

[4] D. J. Gladstone, C. J. Danells, and S. E. Black, "The Fugl–Meyer assessment of motor recovery after stroke: A critical review of its measurement properties," *Neurorehabilitation Neural Repair*, vol. 16, no. 3, pp. 232–240, Sep. 2002.

[5] C. Camona, K. B. Wilkins, J. Drogos, J. E. Sullivan, J. P. A. Dewald, and J. Yao, "Improving hand function of severely impaired chronic hemiparetic stroke individuals using task-specific training with the ReIn-hand system: A case series," *Frontiers Neurol.*, vol. 9, p. 923, Nov. 2018.

[6] K. Park, B. R. Ritsma, S. P. Dukelow, and S. H. Scott, "A robot-based interception task to quantify upper limb impairments in proprioceptive and visual feedback after stroke," *J. NeuroEngineering Rehabil.*, vol. 20, no. 1, p. 137, Oct. 2023.

[7] B. Lee, I. D. Saragih, and S. O. Batubara, "Robotic arm use for upper limb rehabilitation after stroke: A systematic review and meta-analysis," *Kaohsiung J. Med. Sci.*, vol. 39, no. 5, pp. 435–445, May 2023.

[8] A. Schwarz, C. M. Kanzler, O. Lambercy, A. R. Luft, and J. M. Veerbeek, "Systematic review on kinematic assessments of upper limb movements after stroke," *Stroke*, vol. 50, no. 3, pp. 718–727, 2019.

[9] S. Balasubramanian, R. Colombo, I. Sterpi, V. Sanguineti, and E. Burdet, "Robotic assessment of upper limb motor function after stroke," *Amer. J. Phys. Med. Rehabil.*, vol. 91, no. 11, pp. S255–S269, 2012.

[10] S. H. Scott and S. P. Dukelow, "Potential of robots as next-generation technology for clinical assessment of neurological disorders and upper-limb therapy," *J. Rehabil. Res. Develop.*, vol. 48, no. 4, pp. 335–353, 2011.

[11] H. I. Krebs et al., "Robot-aided neurorehabilitation: A robot for wrist rehabilitation," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 15, no. 3, pp. 327–335, Sep. 2007.

[12] H. I. Krebs et al., "Robot-aided neurorehabilitation: From evidence-based to science-based rehabilitation," *Topics Stroke Rehabil.*, vol. 8, no. 4, pp. 54–70, Jan. 2002.

[13] Kinarm. (2021). *Kinarm Exoskeleton Lab*. [Online]. Available: https://kinarm.com/

[14] L. E. R. Simmatis, S. Early, K. D. Moore, S. Appaqaq, and S. H. Scott, "Statistical measures of motor, sensory and cognitive performance across repeated robot-based testing," *J. NeuroEngineering Rehabil.*, vol. 17, no. 1, p. 86, Jul. 2020.

[15] S. H. Scott, "Apparatus for measuring and perturbing shoulder and elbow joint positions and torques during reaching," *J. Neurosci. Methods*, vol. 89, no. 2, pp. 119–127, Jul. 1999.

[16] D. K. Agrafiotis et al., "Accurate prediction of clinical stroke scales and improved biomarkers of motor impairment from robotic measurements," *PLoS ONE*, vol. 16, no. 1, Jan. 2021, Art. no. e0245874.

[17] H. I. Krebs et al., "Robotic measurement of arm movements after stroke establishes biomarkers of motor recovery," *Stroke*, vol. 45, no. 1, pp. 200–204, Jan. 2014.

[18] N. Chalmers, G. Seaborn, J.-Y. Jung, J. I. Glasgow, and S. H. Scott, "Recombination of common sensory-motor impairment evaluation techniques using a committee of classifiers," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Sep. 2009, pp. 857–860.

[19] F. Akbarifar, S. P. Dukelow, P. Mousavi, and S. H. Scott, "Computer-aided identification of stroke-associated motor impairments using a virtual reality augmented robotic system," *Comput. Methods Biomechanics Biomed. Engineering: Imag. Visualizat.*, vol. 10, no. 3, pp. 252–259, May 2022.

[20] J. Gawlikowski et al., "A survey of uncertainty in deep neural networks," 2021, *arXiv:2107.03342*.

[21] E. Hullermeier and W. Waegeman, "Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods," *Mach. Learn.*, vol. 110, no. 3, pp. 457–506, 2021.

[22] M. Sensoy, L. Kaplan, and M. Kandemir, "Evidential deep learning to quantify classification uncertainty," 2018, *arXiv:1806.01768*.

[23] R. L. Sacco et al., "Aha/Asa expert consensus document an updated definition of stroke for the 21st century a statement for healthcare professionals from the American heart association/American stroke association," *Stroke*, vol. 44, no. 7, pp. 2064–2089, 2013.

[24] Heart Stroke Found. Canada. (2021). *Definitions—Secondary Prevention of Stroke*. [Online]. Available: https://www.strokebestpractices.ca/recommendations/secondary-prevention-of-stroke/definitions

[25] L. Simmatis, J. Krett, S. H. Scott, and A. Y. Jin, "Robotic exoskeleton assessment of transient ischemic attack," *PLoS ONE*, vol. 12, no. 12, Dec. 2017, Art. no. e0188786.

[26] J. D. Easton et al., "Definition and evaluation of transient ischemic attack: A scientific statement for healthcare professionals from the American heart association/American stroke association stroke council; council on cardiovascular surgery and anesthesia; council on cardiovascular radiology and intervention; council on cardiovascular nursing; and the interdisciplinary council on peripheral vascular disease: The American academy of neurology affirms the value of this statement as an educational tool for neurologists," *Stroke*, vol. 40, no. 6, pp. 2276–2293, Jun. 2009.

[27] A. J. Coull, J. K. Lovett, and P. M. Rothwell, "Population based study of early risk of stroke after transient ischaemic attack or minor stroke: Implications for public education and organisation of services," *BMJ*, vol. 328, no. 7435, p. 326, Feb. 2004.

[28] S. C. Johnston, D. R. Gress, W. S. Browner, and S. Sidney, "Short-term prognosis after emergency department diagnosis of TIA," *JAMA*, vol. 284, no. 22, pp. 2901–2906, Dec. 2000.

[29] A. M. Coderre et al., "Assessment of upper-limb sensorimotor function of subacute stroke patients using visually guided reaching," *Neurorehabilitation Neural Repair*, vol. 24, no. 6, pp. 528–541, Jul. 2010.

[30] K. Tyryshkin et al., "A robotic object hitting task to quantify sensorimotor impairments in participants with stroke," *J. NeuroEng. Rehabil.*, vol. 11, no. 1, p. 47, Apr. 2014.

[31] T. C. Bourke, C. R. Lowrey, S. P. Dukelow, S. D. Bagg, K. E. Norman, and S. H. Scott, "A robot-based behavioural task to quantify impairments in rapid motor decisions and actions after stroke," *J. NeuroEng. Rehabil.*, vol. 13, no. 1, pp. 1–13, Dec. 2016.

[32] T. Vanbellingen et al., "A new bedside test of gestures in stroke: The apraxia screen of TULIA (AST)," *J. Neurol., Neurosurg. Psychiatry*, vol. 82, no. 4, pp. 389–392, Apr. 2011.

[33] *KST Summary*. Accessed: Jul. 30, 2024. [Online]. Available: https://kinarm.com/download/kst-summary-analysis-version-3-7/

[34] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. 33rd Int. Conf. Mach. Learn.*, Jun. 2016, pp. 1050–1059.

[35] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[36] A. Malinin and M. Gales, "Predictive uncertainty estimation via prior networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, Dec. 2018, pp. 7047–7058.

[37] A. Amini, W. Schwarting, A. Soleimany, and D. Rus, "Deep evidential regression," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 14927–14937.

[38] A. D. Kiureghian and O. Ditlevsen, "Aleatory or epistemic? Does it matter?" *Struct. Saf.*, vol. 31, no. 2, pp. 105–112, Mar. 2009.

[39] L. S. Shapley, "Notes on the N-person game—II: The value of an N-person game," RAND Corp., Santa Monica, CA, USA, Tech. Rep. RM-670, 1951.

[40] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[41] J. F. Veale, "Edinburgh handedness inventory—Short form: A revised version based on confirmatory factor analysis," *Laterality, Asymmetries Body, Brain Cognition*, vol. 19, no. 2, pp. 164–177, Mar. 2014.

[42] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform manifold approximation and projection for dimension reduction," 2018, *arXiv:1802.03426*.

[43] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, Jan. 2016, pp. 1–16.

[44] A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, Mar. 2017, pp. 1–20.

[45] Y. Gal and Z. Ghahramani, "Bayesian convolutional neural networks with Bernoulli approximate variational inference," 2015, *arXiv:1506.02158*.

[46] I. Osband, C. Blundell, A. Pritzel, and B. Van Roy, "Deep exploration via bootstrapped DQN," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, Feb. 2016, pp. 1–9.

[47] A. G. Wilson and P. Izmailov, "Bayesian deep learning and a probabilistic perspective of generalization," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, Jan. 2020, pp. 4697–4708.

[48] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[49] A. Halevy, P. Norvig, and F. Pereira, "The unreasonable effectiveness of data," *IEEE Intell. Syst.*, vol. 24, no. 2, pp. 8–12, Mar. 2009.

[50] R. S. Olson, R. J. Urbanowicz, P. C. Andrews, N. Lavender, L. C. R. Kidd, and J. H. Moore, "Automating biomedical data science through tree-based pipeline optimization," in *Proc. 19th Eur. Conf. Appl. Evol. Comput.*, Porto, Portugal. Cham, Switzerland: Springer, Jan. 2016, pp. 123–137.