RESEARCH

Robotics-Assisted Stroke Rehabilitation with Machine Learning-Based Residual Severity Classification

Raymond M. Greenfield[1†], Russell Jeter[3†], Stephen N. Housley[2,3] and Igor Belykh[1,4*]

Abstract

Background: Stroke therapy is essential to reduce impairments and improve motor movements by engaging autogenous neuroplasticity. This study uses supervised learning methods to address a clinician's autonomous classification of stroke residual severity labeled data towards improving in-home robotics-assisted stroke rehabilitation.

Methods: Thirty-three stroke patients participate in in-home therapy sessions using the Motus Nova robotics rehabilitation technology to capture upper and lower body motion. The therapy session summary data is based on high-resolution movement and assistance data and clinician-informed discrete stroke residual severity labels. This arises from a final processed dataset of $32,902$ patient sessions based on the maximum score per patient per session. Four machine learning algorithms are used to classify stroke residual severity: light gradient boosting, extra trees, deep neural networks, and classical logistic regression. Their performance measures are evaluated to identify which method maximizes stroke residual severity classification accuracy.

Results: We demonstrated that the light gradient boosting method provides the most reliable autonomous detection of stroke severity. This method achieved an average of 94% accuracy, measured using the F1-score performance measure with 10-fold cross-validation. All variables collected from each patient session impact the model's classification accuracy and contribute to 95% of the explained variance.

Conclusion: We showed how objectively measured rehabilitation training paired with machine learning methods can be used to identify the residual stroke severity class with efforts to enhance in-home self-guided, individualized stroke rehabilitation. As data from rehabilitation practices are often of comparable size and nature to the data collected in our study, this suggests that the light gradient boosting method should be considered a standard, more efficient tool for this analysis.

Keywords: Stroke; Rehabilitation Robotics; Machine Learning; Artificial Intelligence; Physical Therapy; Neuroplasticity

Background

Stroke is a leading cause of mortality and disability worldwide, and the economic costs of treatment and post-stroke care are substantial [1]. In 2019, there were 12.2 million incident cases of stroke, 101 million prevalent stroke cases, and 6.55 million deaths from stroke [2]. The severity of a stroke can range from mild to severe, with severe strokes often leading to long-term disability or even death. Stroke rehabilitation typically involves a team of healthcare professionals, including doctors, nurses, therapists, and other specialists. The specific goals and interventions of stroke rehabilitation vary depending on the individual's needs and abilities. They may include physical therapy to improve mobility, occupational therapy to improve the ability to perform daily activities, speech therapy to improve communication skills, and cognitive therapy to improve memory, problem-solving, and other cognitive abilities. This study takes steps to make this goal of stroke patient recovery in-home and autonomous via robotics-assisted stroke rehabilitation and classifying stroke residual severity via machine learning methods.

Machine learning in healthcare and stroke rehabilitation is not a new concept (see [3, 4, 5, 6] as notable examples of this vast research field and [7] for a systematic review of machine learning methods for post-stroke rehabilitation recovery prediction). In particular, multiple studies have been performed to predict outcomes in patient survival, locoregional recurrences, and long-term outcomes in ischemic stroke patients [8, 9]. Similarly, studies focused on motor function have leveraged retrospective healthcare data and targeted predicting the short- and long-term functional ability [10, 11, 12]. Such studies represent an exciting step forward in stroke rehabilitation but have some limitations. These limitations include the use of healthcare data that is infrequently measured (sometimes entirely limited to admission data), which can hamper the performance of models that rely on large datasets for generalizability. Similarly, most studies limit their scope to predicting short- and long-term outcomes and may fail to capture some of the day-to-day changes stroke survivors experience.

This study aims to overcome these limitations by quantifying the progress of patient improvement via day-to-day in-home therapy sessions using the Motus Nova robotics rehabilitation technology [13] that captures upper and lower body motion. The Motus Hand and Motus Foot are robotic therapeutic devices designed to be used by stroke survivors with residual upper and lower extremity impairments at home without needing help from a clinician or caregiver. The Motus Hand and Motus Foot engage the affected wrist or ankle of the user, guiding them through various therapeutic exercises targeting various functional tasks (e.g., gross motor control, fine motor control, and precision tracking). Earlier versions of the technology have been shown to have clinically significant improvements in depressive symptoms, functional independence, upper extremity use in functional tasks, distance walking, and gait speed [14, 15, 13].

Utilizing machine learning promises to have a central role in rehabilitation decision-making in determining if patient therapy is improving. Machine learning is the methodology that allows computers to learn from experience. By constructing and training supervised classifiers to learn decision rules from data, automatic solutions can be exploited to make predictions on new data [16, 17]. Like in many healthcare, disease, or machine learning research applied in a clinical setting, labeling of patient data by a clinician is necessary [3]. This study applies the same heuristic methodologies. Our goal is to use supervised machine learning methods to address a clinician's autonomous classification of stroke residual severity labeled data towards improving in-home robotics-assisted individualized stroke rehabilitation.

*Correspondence: ibelykh@gsu.edu
[1]Department of Mathematics and Statistics, Georgia State University, Atlanta, Georgia, USA
Full list of author information is available at the end of the article
†Equal contributor

Methods

Study Design

The Motus Hand and the Motus Foot each consist of two major components: a peripheral (see the bottom panel of Fig. 1 for a close-up of the Motus Hand peripheral) that the patient attaches to their affected limb and an interactive console that guides their therapy routine and assessment using a video game interface. The peripherals have a pneumatic actuator that can dynamically provide assistance/resistance by filling an air muscle in the peripheral that moves the wrist/ankle joint. The wrist/ankle joint of the peripheral has an embedded angle and pressure sensor that transmits live angle and pressure data to the console. This allows the console to give the user immediate visual feedback of their movement through avatars in a video game on the screen. The therapeutic video game activities can provide a dynamic feedback loop consisting of in-game goals (ships to shoot or coins to collect, for example) that drive user movements, which correspond to movement onscreen, which allow the console to react and set new goals/obstacles. This feedback loop is designed to promote sensory-motor function.

A therapy session with the Motus Hand or Foot consists of stretching, gross motor control, fine motor control, and endurance exercises, depending on the patient's needs. This process is depicted in Fig. 1 where a Motus Hand user is playing "Cosmic Tennis," a gross motor control exercise that plays like the classic arcade game Pong [18]. The user's wrist/ankle movement corresponds to the movement of the paddle on the right-hand side of the screen, and the goal is to hit the ball back and forth to score on the Artificial Intelligence (AI)–controlled opponent.

The Motus Hand and Foot collect high-resolution angle and pressure data from sensors embedded in the wrist/ankle joints and the pressure management system. This high-resolution data is collected at a frequency of $30-50$ Hz, and stored in a time series database. Other information collected during a therapy session includes score, peripheral type (Motus Hand or Motus Foot), current game, etc. (see Table 1 for a full list of data collected). For this study, data was collected from 33 stroke patients, with a total of $32,902$ therapy sessions.

Figure 1 Patients do therapy sessions with the Motus Hand or Motus Foot using a pneumatically-driven exogenous robotic device worn on the affected hand/arm/foot (the Motus Hand is depicted in the bottom panel). The peripheral acts as a game controller (through an angle sensor embedded in the wrist joint) that allows users to play therapeutic video games that dynamically adapt to their needs and provide the requisite assistance/resistance (computer screen in the bottom panel).

To use the data collected during a therapy session to classify a patient's stroke residual severity autonomously, each patient was given a guided assessment with a clinician using the Motus Hand or Foot to classify them as having a High Range of Motion (ROM), Low ROM, or No ROM.

To find an ideal classifier, we use to consider the training and performance of four machine learning algorithms: a light gradient boosting [19], extra trees classifier [20], deep neural network [21], and multi-class logistic regression [22]. A practical model is then constructed using the most common data measured in each session based on the maximum score per session per patient. Unsupervised learning methods are then applied to the final dataset, such as the correlation matrix and principal component analysis, to show that all variables collected are relevant to the study. This includes a 10-fold cross-validation on the final dataset with the mean and standard deviation of accuracy from each computational experiment. From here, the following metrics determine the model's performance, including the accuracy, precision, and recall from the confusion matrix. The macro average F1-score was used to judge the efficacy of the models, as this is a multi-classification problem [23], and as such, accuracy would be an insufficient measure. Figure 2 provides a high-level overview of the data collection, analysis, processing, and modeling that produce the final classification results.

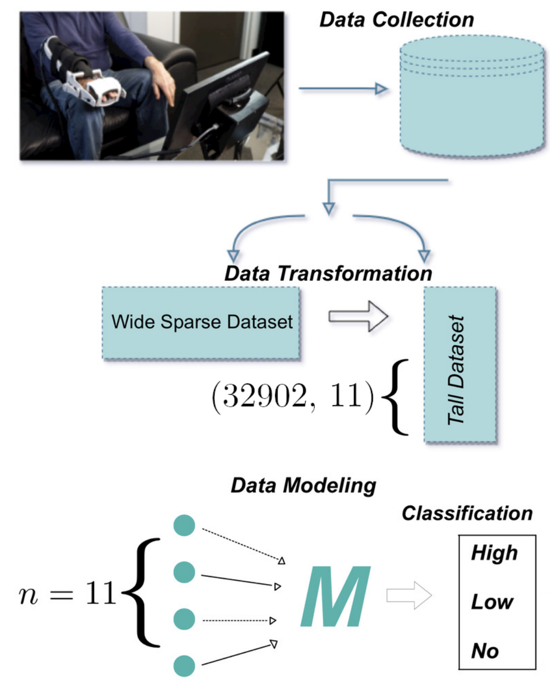

Figure 2 Concept diagram of the overall data analysis and modeling. Data is gathered from $32,902$ therapy sessions using the Motus Hand and Motus Foot. The data is processed and used in a supervised machine-learning model to classify the stroke severity of the patient. A total of 11 predictor variables contribute to classifying stroke severity.

Details of Data Collection

Throughout a therapy session using the Motus Hand or Motus Foot, live angle data (measured in degrees from a natural midpoint in wrist/ankle placement) is collected from the sensor embedded in the wrist or ankle joint at 30-50 readings per second. This "raw" angle sensor data is then stored in a high-resolution time series database (InfluxDB [24]). In addition to the high-resolution angle data, pressure readings (measured in PSI) are taken from the pressure management system at 30-50 measurements per second.

Each therapy session for a patient includes a selection of about 30 activities that focus on different types of motor function, including gross motor control, fine motor control, flexor tone reduction, endurance, reaction time, and tracking. A patient can participate in more than one video game during a patient session. The score is recorded and stored once the patient completes the video game. The scores for each game are not necessarily standardized. This means a score of 100 in one game can represent a dramatically different performance than a score of 100 in another. The score is collected each time a player performs an action in the game that would increase or decrease the score, so this field is collected more irregularly and infrequently than angle and pressure data. That being said, we show that the score does have a high contribution to the model performance (see Fig. 9 for a full breakdown of feature importance). Gender and other biometric data such as age, height, and weight are not included in the patient description or the analysis.

Table 1 Session Game Data Dictionary

Variable	Description	Unit	Example
F_{flex}	Maximum centripetal force generated moving in the downward direction (computed from derivatives of angle data)	Newtons	-3.047709105
F_{ext}	Maximum centripetal force generated moving in the upward direction (computed from derivatives of angle data)	Newtons	3.251405759
N_{mov}	The number of completed movements	integer	10
R_{min}	Absolute minimum angle detected by angle sensor during therapy session	degrees	-25
R_{max}	Absolute maximum angle detected by angle sensor during therapy session	degrees	46.41941
t_{game}	Total time spent performing therapy during a session for that game	seconds	15
P_{min}	Minimum Pressure applied by the sensor in a patient session	PSI	-0.04511994
P_{max}	Maximum pressure applied by the sensor in a patient session	PSI	10.30989
P_{mean}	Average pressure applied by the sensor in a patient session	PSI	3.590553432
$Score$	Score achieved by patient per video game	integer	100
h	Peripheral type variable indicating the Hand or Foot	0, 1	Hand
$Class$	Designate stroke severity label for a patient by a clinician (High, Low, No)	0, 1, 2	High
g	Unique identifier for each activity that is available on the Motus Hand/Foot	integer	4
p	Anonymous identifier for each patient using the Motus Hand/Foot in this study	integer	11
s	Unique identifier for each session performed on the Motus Hand/Foot	integer	782302348734

Clinician Labeling

Our supervised learning method requires clinician-labeled data to train a classifier for labeling stroke residual severity appropriately. While these labels are quite broad, the labeling process is hardly a simple algorithm. At the clinician's discretion, quantitative and qualitative factors must apply an appropriate label. By using the potentiometer [25] embedded in the wrist/ankle joint of the Motus Hand/Foot, clinicians can gather data on the patient in Table 2. We define a patient as "Assisted" or "Passive" based on the most arduous assessment performed on the patient. An "Assisted" patient was able to undergo an assessment that used assistance to maximize their active range of motion. A "Passive" patient had little to no active movement, and a passive range of motion assessment was given to determine their passive range of motion. The Low ROM label is mixed between the assisted and passive. Notice that all patients who are classified with a high ROM (low residual stroke severity) were assisted by the clinician. This is important when noticing that patients with ID 2085 and 1781 (blue) have a similar total range of motion ($R_{min} + R_{max}$), but patient ID 1781 requires clinician assistance to reach their maximum ROM. However, there is ambiguity in some labels. For example, take patient ID 2356 (red), where it can be argued that the patient should have a high stroke residual severity (corresponding to low/no ROM) given the low total range of motion with the assistance. This is where the clinician has other outside factors that contribute to the final labeled classification of a patient.

Data Harmonization

To create a more manageable dataset for the labeling task, we generate summary statistics of the high-resolution data for each activity performed during a therapy session. First, we remove outliers from the raw time series data. Then, summarize the angle (relative to a reference midpoint in degrees) and pressure (in PSI) using the following variables: R_{min}, the minimum ROM for a game; R_{max}, the maximum ROM for a game; R_{mean}, the mean ROM for a game; P_{min}, the minimum pressure for a game; F_{flex}, the maximum centripetal force generated while moving downward; F_{ext}, the maximum centripetal force generated while moving upward, P_{max}, the maximum pressure for a game;

ID	R_{max}	R_{min}	Assessment	Classification
2054	30	3	Passive	No
1495	37	-20	Passive	No
2058	50	-24	Passive	No
2273	21	-16	Passive	No
2085	40	-15	Passive	No
2098	44	-9	Passive	No
1864	28	-12	Passive	No
1859	45	-17	Passive	No
1479	45	-20	Passive	No
1865	30	-15	Passive	No
2128	34	-10	Passive	No
1838	33	-15	Passive	No
2183	41	-16	Passive	No
2040	37	-18	Passive	No
2097	43	-18	Passive	Low
2356	-3	-17	Assisted	Low
2356	-3	-17	Assisted	Low
1688	52	-23	Assisted	Low
1876	54	-12	Passive	Low
2029	46	-20	Passive	Low
1458	30	-18	Passive	Low
1113	33	-20	Assisted	High
2262	38	-13	Assisted	High
1637	10	-12	Assisted	High
2282	8	-16	Assisted	High
1781	39	-15	Assisted	High
2360	10	-18	Assisted	High
2035	41	7	Assisted	High
1799	48	1	Assisted	High
2191	40	-20	Assisted	High
1974	38	-6	Assisted	High
2004	41	-20	Assisted	High
2179	49	12	Assisted	High
1470	20	-16	Assisted	High

Table 2 Example patient label table assessed by a clinician using a potentiometer [25]. Note that the final label is at the clinician's discretion and could be based on qualitative factors not accounted for in the test.

P_{mean}, the mean pressure for a game. We finally pair these game-level summary statistics with the number of movements performed in the game (N_{mov}), the maximum score in the game ($Score$), and the total time spent playing that game during a therapy session (t_{game}). This transformation from high-resolution data to game-level summary statistics provides a much more manageable dataset to which we can apply the clinician labels. A low ROM patient (as labeled by the clinician) has little ROM during each game throughout a session. Using this idea, we construct a new dataset from each game a patient plays during a session, with each row having a unique (patient ID, session ID, game ID) tuple. A summary of the data in a given row in the described dataset is presented in Table 1.

After combining the data into this standardized dataset, the data then requires sanitization, analysis, and normalization. To sanitize the data, we need to fill in missing values, correct invalid sensor values, and throw out data that did not represent a meaningful therapeutic exercise. The main variables with missing values are: R_{min}, R_{max}, and t_{game} (game time). For R_{min} and R_{max}, there is often insufficient data per patient per game to fill missing data points with the patient-game average. We used this average when available and the global mean when it was not. We used the global minimum value for missing t_{game} (game time) values.

To isolate games with insufficient activity to draw meaningful conclusions, we restrict the number of movements, N_{mov}, performed during a game. A "movement" is any change of direction recorded in the angle sensor after noise is smoothed out of the time series. We remove any activity with fewer than three movements, as no significant therapeutic exercise can be performed with fewer than three movements (under assistance from the robotic Motus Hand/Foot).

Exploratory Data Analysis

It is well-known that proper data normalization is critical for maximizing model performance across machine learning applications and methods [26]. Knowing the proper normalization technique for each feature requires a cursory dataset analysis. In Fig. 3, we show representative distributions of the features that will be input variables for our comparative model analysis. While some variables are not normally distributed, assuming the data is normally distributed is sufficient considering the results [27].

This study seeks to determine if there exists a key indication variable that is a direct indicator of stroke classification. We analyze the correlation among the features in our dataset to identify potential redundancies. Then, we look at the

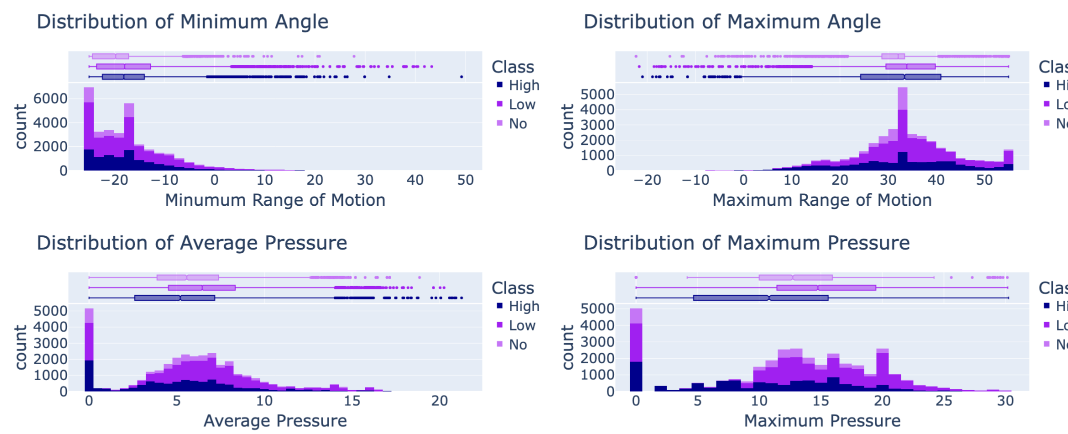

Figure 3 Example of the distribution plots for four of the random variables separated by the class label from the clinician. Subsequently, random variables are distributed normally, which is crucial for using the z-score when inputting into a machine learning algorithm.

principal component decomposition [28] to see if the variation in the data can be meaningfully reduced to a lower dimensional space.

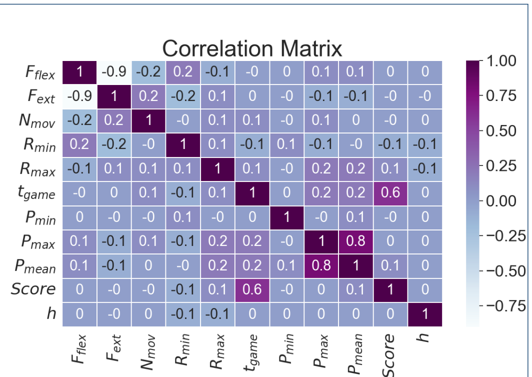

Figure 4 The correlation matrix does not include key variables such as Patient ID, Session ID, Game ID, and Start Time. Importance in dimensionality reduction based on a greater than 0.9 threshold. As seen above, F_{flex} and F_{ext} are highly negatively correlated. However, these variables were both used in the analysis.

The correlation matrix for the feature set, constructed by computing the correlation between each pair of features in the dataset, is shown in Fig. 4. Because a correlation matrix points to potential relationships between features, it can indicate the feasibility of dimensionality reduction when preparing a dataset for building a classifier. If two variables are highly correlated, i.e., $Cor(X, Y) > 0.9$, it suggests that we can drop one of those variables from our analysis [29]. There exists a strong negative correlation between F_{ext} and F_{flex}, but intuitively, we should not drop either one of the variables in the analysis individually. The correlation between the game pressure mean, P_{mean}, and game pressure max, P_{max}, with the value of 0.80, indicates that the Motus Hand or Foot applied more pressure on average in each session; however, because this correlation fails to surpass the threshold of 0.90 we do not drop either variable. Similarly, the correlation (0.60) between game time, t_{game} and game score, $Score$, is intuitive: the longer a patient plays a game, the higher their score. Unfortunately, this correlation also does not meet the threshold for exclusion in the final feature set. Given the inherent redundancy of a correlation matrix, $Cor(X, Y) = Cor(Y, X)$, it suffices only to consider the upper triangular portion of the matrix.

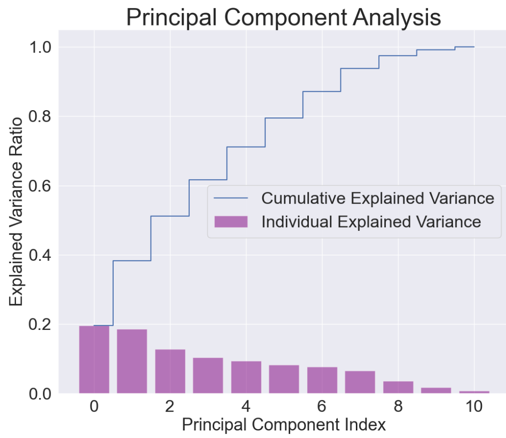

Figure 5 Depiction of the principal components with the explained variance ratio. As shown, 95% of the explained variance is contributed by all principal components. As a result, all variables are used in the machine learning model for the analysis.

Another informative approach for analyzing the potential for dimensionality reduction in a feature set is principal component analysis (PCA). Principal components are new variables constructed as linear combinations of the initial variables. These particular linear combinations ensure that the new variables (i.e., principal components) are uncorrelated and that as few components as possible contain most of the information from the initial variables. Explained variance is a statistical measure of how much variation in a dataset is attributable to each principal component (eigenvectors) generated by the PCA method [30]. Explained variance thus allows us to rank the components in order of importance and to focus on the most important ones when interpreting the results of our analysis.

In Fig. 5, we show the explained variance each principal component contributes to the total variation in the feature set. No component can be described as dominant, as none accounts for more than 20% of the variance in the initial data set. Given this and the results from our correlation analysis, we can conclude that all 11 variables are needed for the analysis.

Model Description

Here, we provide a brief overview of the models compared in the Results.

Logistic regression is a classical statistical technique for binary classification. The technique consists of mapping the probability of an event happening to a logistic curve with the model inputs as dependent variables. Logistic regression is still widely used and is a common first model when performing classification because it is easy to implement and interpret.

Gradient Boosting Decision Tree (GBDT) is a widely-used machine learning algorithm due to its efficiency, accuracy, and interpretability [19]. In essence, the algorithm uses smaller "weaker classifiers" with a number of leaves. By taking a weighted average of these several "weaker classifiers," we then can construct a "stronger classifier" [31]. By training several weaker models, this process is known as AdaBoosting. It results in a stronger model by adding more leaves to the decision tree and taking a weighted combination of these weaker models, where the weights are determined by the performance [32].

The Deep Feed-Forward Neural Network (DNN) is a high-performance deep learning model with varying hidden layers. Several architectures were tested on the training dataset to see if there was an increase in performance by adding hidden layers (from four to eight) or a reduction in nodes in each input layer [33]. The reLU activation function was implemented into the model instead of the sigmoid function. Both were tried. Accuracy results from the computational experiment could surpass 80%, regardless of adding more layers, changing the hidden layer input size, or changing the activation function. The DNN used in the accuracy score of Fig. 6 has three hidden layers with the input size of the hidden layers as $(8, 5, 8)$, respectively.

The Extra Trees Classifier (ETC) is an ensemble learning method for classification. Ensemble learning is a machine learning technique that combines the predictions of multiple individual models to produce a more accurate and robust final prediction. The basic idea is to train multiple models independently, each with a different algorithm or set of hyperparameters, and then combine their predictions at the end [34]. This is similar to the AdaBoosting concept with light gradient boosting, where models can be combined by averaging or weighting their predictions [35]. The model uses entropy as the splitting criterion for the trees, with 100% of the features considered at each split. The maximum number of leaf nodes for each tree is $87, 17$, and the model is comprised of 42 trees [35].

Results

Our computational experiments compare the performance of different machine learning methods and find the best model for identifying stroke residual severity. The original harmonized dataset (described in Table 1) contained all the scores, the minimum and maximum ROM, and minimum and maximum pressure, and we took the maximum score per game per session. Because of the smaller dataset, the training and testing were split on the 80/20 principle, where 20% of the data was the testing data.

Table 3 shows a 10-fold cross-validation of each machine learning classification algorithm with Fig. 6 as a visual display of a single accuracy measure. k-fold cross-validation is used to verify that a high-accuracy model is not necessarily overfitting the training data. The data set is then randomly divided into 10 different subsets, or "folds" [36]. Each of these folds is then used as the training data, while another is used as the testing data for fitting a new model. We then take the mean and standard deviation of the model accuracy across the 10 folds.

Figure 7 presents the confusion matrix of each of the supervised learning methods. Generally, a confusion matrix is used to represent the algorithm's performance visually. Each row of the matrix represents the instances in an actual class, while each column represents the instances in a predicted class, or vice versa. We represent the percentage over the exact numeric number for display purposes. Three performance metrics come from the confusion matrix: precision, recall, and the F1-score. Accuracy measures the proportion of predicted positives that are truly positive. Recall measures the proportion of predicted negatives that are truly negative. The F1-score is the harmonic mean of the precision and recall [23]. In this case, this is macro-averaging (treating all classes equally important). A full breakdown of the performance measures (precision, recall, and F1-score) is shown in Table 4. It is important to notice that while the Extra Tree Classifier has a comparable accuracy (picking the correct label) with the Light Gradient Boosting (LGB) method, LGB performs reliably better than all of the other models when also weighing false positives and false negatives (precision, recall, and $F1$-score). Remarkably, the LGB model best fits the dataset with more than a 50% improvement compared to the classical method, such as the logistic regression.

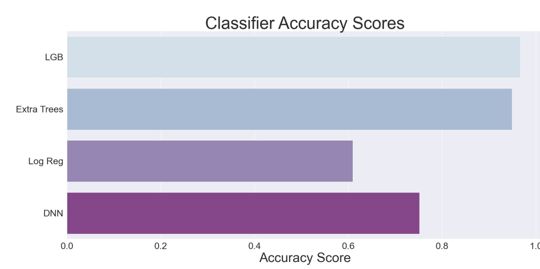

Figure 6 Accuracy score for each machine learning model. The LGB and Extra Trees Classifier have the greatest accuracy score. See Table 3 for the depiction of the cross-validation to prevent over-fitting [36]. Note that accuracy alone is not the best illustration for the performance of a machine learning model. LGB is chosen as the final model in extension for future work.

Discussion

We have demonstrated that machine learning can identify stroke residual severity. We have addressed several challenges arising in healthcare applications of machine learning, such as

Table 3 10-Fold Cross-Validation Mean Scores with Standard Deviation

Classifier	Mean	Std
Extra Trees	96.40%	0.4%
Light Gradient Boosting	94.0%	0.4%
Neural Network	71.70%	0.7%
Logistic Regresssion	61.20%	0.5%

Table 4 Performance Measures

	Precision	Recall	F1-score
Extra Trees			
Low	95.44%	92.83%	94.12%
High	94.46%	91.10%	92.75%
No	94.55%	97.04%	95.78%
LGB			
Low	96.80%	95.61%	96.20%
High	96.49%	94.97%	95.73%
No	96.70%	97.83%	97.26%
DNN			
Low	74.34%	64.94%	69.32%
High	61.93%	28.27%	38.82%
No	71.64%	87.71%	78.86%
Log Reg			
Low	59.61%	37.61%	46.12%
High	48.77%	8.27%	14.15%
No	61.71%	89.42%	73.02%

processing data that contains different physical quantities, errors in sensory data, and ambiguous classifications based on human error.

In previous studies, the algorithms most frequently used among the included models were linear and logistic regressions, confirming a preferable choice toward more conventional and interpretable methods rather than more complex and advanced ones [37, 38]. Unfortunately, with the nature of our data, these methods presented with poor accuracy (less than 80%), and consequently, we considered different approaches. We showed that the LGB method provides substantially high accuracy, albeit on a relatively small dataset. The method provides additional advantages that make it an ideal classifier for online autonomous stroke residual severity classification: it is an easy model to transfer. It requires (relatively) little computational resources.

Figure 7 Confusion Matrices for LGB and Logistic Regression. Considering the false negative column of the No classification, it is seen that the LGB model greatly improved this classification. This is especially important when classifying a patient as having No stroke severity when they are actually a High severity. Mis-classifications can be particularly dangerous, ranging from providing inadequate therapy to a high-range-of-motion patient to injuring a low-range-of-motion user with therapy designed for a high-range-of-motion user.

Building an expanded and more sophisticated dataset remains an area of further study. Real-time processing of sensor data allows a classifier to engage with a user online and recognize and classify subtle changes in their motor function. Subsequently, an up-to-date understanding of a

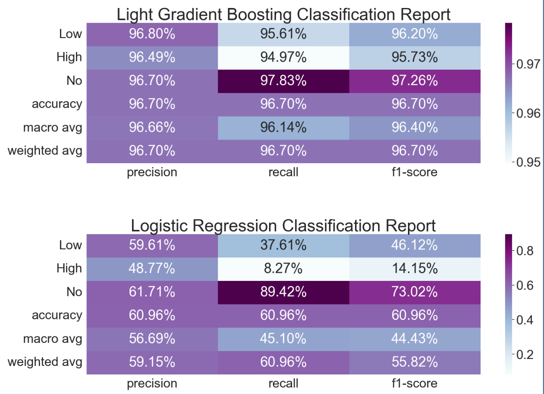

Figure 8 Classification Report for Light Gradient Boosting v.s. Logistic Regression to measure the model performance based on each class's accuracy, macro average, and weighted average. As shown, the LGB model is 73% better in terms of accuracy and F1-score compared to the logistic regression model.

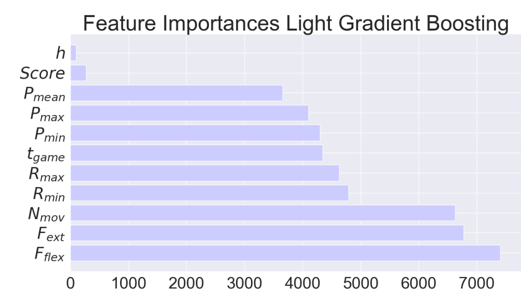

Figure 9 Feature Importances of each model displaying which variables contribute the most to the model. As shown, the F_{ext} and F_{flex} contribute most to the model. Here, we also see that the peripheral type (Hand or Foot) made the least. This could be because only two patients used the foot as a lower extremity. Interestingly, the score variable contributed little compared to the other variables in contrast to one's intuition.

patient's motor function needs allows a clinician (AI or otherwise) to prescribe personalized, targeted interventions that will be the most impactful. Real-time understanding of a patient's needs coupled with an in-home robotic therapy device like the Motus Hand/Foot allows immediate feedback. An AI in the therapeutic games can detect patient needs like fatigue over the course of a therapy session and adapt its strategy accordingly.

Conclusion

Autonomous classification is becoming more important for successful rehabilitation as rehabilitation begins to move out of the clinical setting. Still, it faces challenges with the accessibility and volume of appropriate clinical data for training models and model access to user data for classification.

By leveraging the in-home stroke rehabilitation robotics provided by the Motus Hand and Motus Foot, we have made significant progress in addressing these issues that prevent adequate training of an autonomous classification model. With the data collected from self-guided in-home therapy sessions, we could train a classification model to identify the stroke residual severity in 33 patients. We compared four different models: Extra Trees, Light Gradient Boosting Method, Deep Feed-Forward Neural Network, and Logistic Regression, finding the Light Gradient Boosting Method to outscore the other three with an average F1-score of 94%. The Light Gradient Boosting Method is a particularly powerful model for this case because it combines interpretability and portability.

Because our model relies only on therapy session summary statistics, the proposed method is expected to be successful when applied to a wide range of rehabilitation data sets of similar sizes. Once trained, the model is highly portable and can be integrated into similar rehabilitation settings to provide an autonomous real-time classification of stroke residual severity. Additionally, when paired with something like the Motus Hand and Motus Foot technology, our classifier provides the opportunity to develop personalized training based on the stroke residual severity of the individual and adapt the therapy exercises to each patient's needs. The efficacy of real-time classification and adaptation remains a subject of future study.

Abbreviations
ROM: Range of Motion; PCA: Principal Component Analysis; GBDT: Gradient Boosting Decision Tree; DNN: Deep Feed-Forward Neural Network; ETC: Extra Trees Classifier; LGB: Light Gradient Boosting.

Author's contributions
R.J. and I.B. designed and conceptualized the study; R.J. and S.N.H. collected the data; R.M.G. and R.J. processed and analyzed the data and developed the machine-learning algorithms. All authors interpreted the data and drafted and edited the manuscript.

Acknowledgements
This work was supported by the National Science Foundation (USA) under Grant No. CMMI-1953135.

Ethics approval and consent to participate
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Availability of data and materials
The datasets used and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Author details
[1]Department of Mathematics and Statistics, Georgia State University, Atlanta, Georgia, USA. [2]School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA. [3]Motus Nova, LLC Atlanta, GA. [4]Neuroscience Institute, Georgia State University, Atlanta, Georgia, USA.

References
1. Feigin, V.L., Stark, B.A., Johnson, C.O., Roth, G.A., Bisignano, C., Abady, G.G., Abbasifard, M., Abbasi-Kangevari, M., Abd-Allah, F., Abedi, V., *et al.*: Global, regional, and national burden of stroke and its risk factors, 1990–2019: a systematic analysis for the global burden of disease study 2019. The Lancet Neurology **20**(10), 795–820 (2021)
2. Johnson, C.O., Nguyen, M., Roth, G.A., Nichols, E., Alam, T., Abate, D., Abd-Allah, F., Abdelalim, A., Abraha, H.N., Abu-Rmeileh, N.M., *et al.*: Global, regional, and national burden of stroke, 1990–2016: a systematic analysis for the global burden of disease study 2016. The Lancet Neurology **18**(5), 439–458 (2019)
3. Reyna, M.A., Josef, C., Seyedi, S., Jeter, R., Shashikumar, S.P., Brandon Westover, M., Sharma, A., Nemati, S., Clifford, G.D.: Early prediction of sepsis from clinical data: the physionet/computing in cardiology challenge 2019. In: 2019 Computing in Cardiology (CinC), pp. 1–4 (2019). doi:10.23919/CinC49843.2019.9005736
4. Alabi, R.O., Elmusrati, M., Sawazaki-Calone, I., Kowalski, L.P., Haglund, C., Coletta, R.D., Mäkitie, A.A., Salo, T., Almangush, A., Leivo, I.: Comparison of supervised machine learning classification techniques in prediction of locoregional recurrences in early oral tongue cancer. International Journal of Medical Informatics **136**, 104068 (2020)
5. Cerasa, A., Tartarisco, G., Bruschetta, R., Ciancarelli, I., Morone, G., Calabrò, R.S., Pioggia, G., Tonin, P., Iosa, M.: Predicting outcome in patients with brain injury: differences between machine learning versus conventional statistics. Biomedicines **10**(9), 2267 (2022)
6. Harari, Y., O'Brien, M.K., Lieber, R.L., Jayaraman, A.: Inpatient stroke rehabilitation: prediction of clinical outcomes using a machine-learning approach. Journal of NeuroEngineering and Rehabilitation **17**(1), 1–10 (2020)
7. Campagnini, S., Arienti, C., Patrini, M., Liuzzi, P., Mannini, A., Carrozza, M.C.: Machine learning methods for functional recovery prediction and prognosis in post-stroke rehabilitation: A systematic review. Journal of NeuroEngineering and Rehabilitation **19**(1), 1–22 (2022)
8. Heo, J.N., Yoon, J.G., Park, H., Kim, Y.D., Nam, H.S., Heo, J.H.: Machine learning-based model for prediction of outcomes in acute stroke. Stroke **50**, 1263–1265 (2019)
9. Saber, H., Somai, M., Rajah, G.B., Scalzo, F., Liebeskind, D.S.: Predictive analytics and machine learning in stroke and neurovascular medicine. Neurological Research **41**(8), 681–690 (2019)
10. Lin, W.-Y., Chen, C.-H., Tseng, Y.-J., Tsai, Y.-T., Chang, C.-Y., Wang, H.-Y., Chen, C.-K.: Predicting post-stroke activities of daily living through a machine learning-based approach on initiating rehabilitation. International Journal of Medical Informatics **111**, 159–164 (2018)
11. Gupta, V.P., Garton, A.L.A., Sisti, J.A., Christophe, B.R., Lord, A.S., Lewis, A.K., Frey, H.-P., Claassen, J., Connolly, E.S.: Prognosticating functional outcome after intracerebral hemorrhage: The ichop score. World Neurosurgery **101**, 577–583 (2017)
12. Kim, J.K., Choo, Y.J., Chang, M.C.: Prediction of motor function in stroke patients using machine learning algorithm: Development of practical models. Journal of Stroke and Cerebrovascular Diseases **30**(8), 105856 (2021)
13. Butler, A., Housley, S.N., Chen, Y.-A., Wolf, S.: Increasing access to cost effective home-based robotic telerehabilitation for stroke survivors. 2017 International Symposium on Wearable Robotics and Rehabilitation (WeRob) (2017). doi:10.1109/werob.2017.8383873
14. Butler, A.J., Bay, C., Wu, D., Richards, K.M., Buchanan, S., Yepes, M.: Expanding tele-rehabilitation of stroke through in-home robot-assisted therapy. International Journal of Physical Medicine & Rehabilitation **2**(184), 1–11 (2014)
15. Wolf, S.L., Sahu, K., Bay, R.C., Buchanan, S., Reiss, A., Linder, S., Rosenfeldt, A., Alberts, J.: The haapi (home arm assistance progression initiative) trial. Neurorehabilitation and Neural Repair **29**(10), 958–968 (2015)
16. Cai, S., Li, G., Zhang, X., Huang, S., Zheng, H., Ma, K., Xie, L.: Detecting compensatory movements of stroke survivors using pressure distribution data and machine learning algorithms. Journal of NeuroEngineering and Rehabilitation **16**(1), 1–11 (2019)
17. Thakkar, H.K., Liao, W.-w., Wu, C.-y., Hsieh, Y.-W., Lee, T.-H.: Predicting clinically significant motor function improvement after contemporary task-oriented interventions using machine learning approaches. Journal of NeuroEngineering and Rehabilitation **17**(1), 1–10 (2020)
18. Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., Riedmiller, M.: Playing Atari with deep reinforcement learning. arXiv preprint

arXiv:1312.5602 (2013)
19. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y.: Lightgbm: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems **30** (2017)
20. Geurts, P., Ernst, D., Wehenkel, L.: Extremely randomized trees. Machine Learning **63**, 3–42 (2006)
21. Wang, S., Cao, G., Shang, Z.: Deep neural network classifier for multi-dimensional functional data. arXiv preprint arXiv:2205.08592 (2022)
22. Xie, P., Kim, J.K., Zhou, Y., Ho, Q., Kumar, A., Yu, Y., Xing, E.: Distributed machine learning via sufficient factor broadcasting. arXiv preprint arXiv:1511.08486 (2015)
23. Düntsch, I., Gediga, G.: Confusion matrices and rough set data analysis. Journal of Physics: Conference Series **1229**(1), 012055 (2019)
24. Naqvi, S.N.Z., Yfantidou, S., Zimányi, E.: Time series databases and influxdb. Studienarbeit, Université Libre de Bruxelles **12**, 1–44 (2017)
25. Maceira-Elvira, P., Popa, T., Schmid, A.-C., Hummel, F.C.: Wearable technology in stroke rehabilitation: towards improved diagnosis and treatment of upper-limb motor impairment. Journal of NeuroEngineering and Rehabilitation **16**(1), 1–18 (2019)
26. Sola, J., Sevilla, J.: Importance of input data normalization for the application of neural networks to complex industrial problems. IEEE Transactions on Nuclear Science **44**(3), 1464–1468 (1997)
27. Goldstein, M.L., Morris, S.A., Yen, G.G.: Problems with fitting to the power-law distribution. The European Physical Journal B-Condensed Matter and Complex Systems **41**, 255–258 (2004)
28. Bro, R., Smilde, A.K.: Principal component analysis. Analytical Methods **6**(9), 2812–2831 (2014)
29. Shin, Y., Park, C.: Analysis of correlation based dimension reduction methods. Applied Mathematics and Computer Science **21**, 549–558 (2011)
30. Wold, S., Esbensen, K., Geladi, P.: Principal component analysis. Chemometrics and Intelligent Laboratory Systems **2**(1-3), 37–52 (1987)
31. Schapire, R.E.: The strength of weak learnability. Machine Learning **5**, 197–227 (1990)
32. Chengsheng, T., Huacheng, L., Bing, X.: Adaboost typical algorithm and its application research. In: MATEC Web of Conferences, vol. 139, p. 00222 (2017)
33. Shamshirband, S., Fathi, M., Dehzangi, A., Chronopoulos, A.T., Alinejad-Rokny, H.: A review on deep learning approaches in healthcare systems: Taxonomies, challenges, and open issues. Journal of Biomedical Informatics **113**, 103627 (2021)
34. Ganaie, M.A., Hu, M., Malik, A., Tanveer, M., Suganthan, P.: Ensemble deep learning: A review. Engineering Applications of Artificial Intelligence **115**, 105151 (2022)
35. Arya, M., Sastry G H., Motwani, A., Kumar, S., Zaguia, A.: A novel extra tree ensemble optimized dl framework (Eteodl) for early detection of diabetes. Frontiers in Public Health **9**, 797877 (2022)
36. Shao, J.: Linear model selection by cross-validation. Journal of the American Statistical Association **88**(422), 486–494 (1993)
37. Zariffa, J., Myers, M., Coahran, M., Wang, R.H.: Smallest real differences for robotic measures of upper extremity function after stroke: Implications for tracking recovery. Journal of Rehabilitation and Assistive Technologies Engineering **5**, 205566831878803 (2018)
38. Mostafavi, S.M., Glasgow, J.I., Dukelow, S.P., Scott, S.H., Mousavi, P.: Prediction of stroke-related diagnostic and prognostic measures using robot-based evaluation. In: 2013 IEEE 13th International Conference on Rehabilitation Robotics (ICORR), pp. 1–6 (2013). doi:10.1109/ICORR.2013.6650457