applied sciences

MDPI

*Article*

# Data-Driven Stroke Classification Utilizing Electromyographic Muscle Features and Machine Learning Techniques

**Jaehyuk Lee [1], Youngjun Kim [2] and Eunchan Kim [3,*]**

[1] Smart Technology Laboratory, Kongju National University, Cheonan-si 31080, Republic of Korea; jhl9405@kongju.ac.kr
[2] School of Computer Science and Engineering, Kyungnam University, Changwon-si 51767, Republic of Korea; youngjun@kyungnam.ac.kr
[3] Department of Information Systems, Hanyang University, Seoul 04763, Republic of Korea
* Correspondence: eckim@hanyang.ac.kr

**Abstract:** Background: Predicting a stroke in advance or through early detection of subtle prodromal symptoms is crucial for determining the prognosis of the remaining life. Electromyography (EMG) has the advantage of easy and quick collection of biological data in clinical settings; however, its application in data processing and utilization is somewhat limited. Thus, this study aims to verify how simple signal processing and feature extraction utilize EMG in machine learning (ML)-based prediction models. Methods: EMG data were collected from the legs of 120 healthy individuals and 120 stroke patients during gait. Four statistical features were extracted from 16 EMG signals and trained on seven ML-based models. The accuracy of the validation and test datasets was also examined. Results: The model with the best performance was Random Forest. Among the 16 EMG signals, the average and maximum values of the muscle activities involved in knee extension (i.e., vastus medialis and rectus femoris) contributed significantly to the predictions. Conclusion: The results of this study confirmed that the simple processing and feature extraction of EMG signals effectively contributed to the accuracy of ML-based models. Routine use of EMG data collected in clinical environments is expected to provide benefits in terms of stroke prevention and rehabilitation.

**Keywords:** electromyography; machine learning; stroke

check for updates

**Citation:** Lee, J.; Kim, Y.; Kim, E. Data-Driven Stroke Classification Utilizing Electromyographic Muscle Features and Machine Learning Techniques. *Appl. Sci.* **2024**, *14*, 8430. https://doi.org/10.3390/app14188430

Academic Editors: Weijie Wang and Dong Ming

Received: 29 August 2024
Revised: 14 September 2024
Accepted: 17 September 2024
Published: 19 September 2024



## 1. Introduction

Stroke, caused by insufficient cerebral blood supply or hemorrhage, is a leading cause of mortality and functional disability in the elderly population [1]. From a neurophysiological perspective, stroke damage to the brain neurons disrupts motor and sensory integration, leading to paralysis [2], muscle weakness [3], and changes in muscle tone [4]. Although advancements in medical treatment have gradually reduced stroke-related mortality [5], stroke continues to affect high to 85% of survivors with persistent sequelae that significantly impact their quality of life [6]. Among these, gait disorders are among the primary factors limiting mobility and substantially affecting the quality of life of stroke survivors [7]. Various abnormal gait patterns arise not only from hemiplegia following a stroke but also from symptoms such as muscle spasticity, co-contraction, and foot drop, and the extent of these symptoms varies among patients. These gait abnormalities have been demonstrated through changes in the activation of various lower limb muscles, as measured by electromyography (EMG) [8–11].

The most critical factor influencing the recovery of stroke survivors is early detection and treatment, with the best prognosis associated with the rapid identification of prodromal symptoms [12]. In particular, when prodromal symptoms occur, patients often fail to recognize these signs because of mild symptoms or inability to ambulate, leading to delays in seeking medical attention. Statistically, 17.8% of individuals with no prior diagnosis of stroke or transient ischemic attacks (TIAs) over the age of 45 years experience at least one prodromal symptom [13]. In particular, TIAs have been reported to cause motor issues, such as balance and gait disturbances, along with cognitive, sensory, and fatigue-related symptoms [14,15]. Nevertheless, 80% of strokes following TIAs can be prevented by prompt intervention [16].

Among the various tools used to assess the functional status of stroke patients, EMG is a commonly employed technique that measures the electrical activity generated by the muscles. It is widely used in clinical settings to assess motor control abilities quickly and accurately. However, challenges remain in accessing and selecting features from unfiltered time-series data, particularly as the digital healthcare industry advances and the use of medical big data are increasingly developed [17]. Although research on signal processing for EMG data has a long history of addressing issues such as noise and artifacts introduced during the data collection process [18,19], relatively little progress has been observed in practical feature extraction and selection for disease prediction models. Given the large volume of training data required for predictive models, computational efficiency in feature extraction and selection is as crucial as model performance.

Recent advancements in the field of machine learning have seen a growing emphasis on the application of EMG. Studies have demonstrated that machine learning models utilizing features extracted from EMG signals [20]—across various domains such as time, frequency, time-frequency, and statistical—can achieve high classification performance for distinguishing diseases [21–23] and physical movements [24]. Due to the inherent complexity of predicting disease indicators and symptoms or physical movements, enhancing accuracy remains challenging, prompting the exploration of model integration strategies [25,26]. Moreover, in the context of disease prediction and classification, both sensitivity—reflecting the model's ability to accurately detect the presence of a disease—and specificity—reflecting the model's ability to differentiate between diseases—are critical metrics. Thus, ongoing efforts are dedicated to improving these performance indicators. Nevertheless, the choice of data features for training, given its significant impact on model complexity and performance, remains a pivotal area of study within machine learning.

In the development of machine learning models for medical applications such as stroke prediction, an excessive number of data features can lead to several technical challenges. Firstly, incorporating an excessive variety of data can impede the model's ability to discern meaningful signals. Redundant data introduces additional noise, thereby diminishing the model's accuracy. Secondly, as the dimensionality of the data increases, the number of required data points grows exponentially, which heightens the risk of overfitting and reduces the model's generalizability. Furthermore, the increased computational burden renders the model impractical for deployment in local healthcare settings due to the associated high computational costs. These issues present significant barriers to the integration of artificial intelligence technologies into regional medical practice. Consequently, research on simplified data feature extraction and selection methods, based on filtering techniques commonly employed in EMG signal processing, is of critical importance as it holds the potential to alleviate these barriers and enhance clinical applicability.

Numerous traditional predictive studies have reported lifestyle-based stroke risk factors [27–30]. More recently, research has focused on predicting abnormal movements in stroke patients using EMG and machine learning (ML) [31–33]. However, ML-based stroke modeling studies that use the activation levels of multiple muscles as biomarkers remain limited [34]. Predictive modeling using muscle activity as training data could provide valuable insights into data features suitable for the rapid assessment of the neurophysiological state in patients at the early or pre-stroke stages. Therefore, this study aims to examine the impact of the statistical features of muscle activity of the major leg muscles during gait as predictive factors across various models to differentiate between stroke patients and healthy individuals.

## 2. Related Works

Machine learning studies utilizing EMG datasets have predominantly concentrated on movement detection [35–37], the identification of abnormal signal patterns [38–40], and the prediction of diseases [23,41] based on EMG data. Previous studies validating disease prediction models with EMG data have addressed a range of conditions, from musculoskeletal disorders [42], such as osteoarthritis and rheumatoid arthritis, to neurological disorders, including cerebral palsy [43] and Parkinson's disease [44]. These studies have employed predictive models for disease classification using machine learning techniques, including neural networks, linear discriminant analysis, support vector machines, decision trees, and least squares kernel algorithms, with reported accuracy rates ranging from 72% to 98%.

Using EMG datasets for disease classification presents several challenging issues. First, the time burden associated with collecting EMG data from patients is significant. Consequently, many studies focus on unsupervised machine learning techniques for dimensionality reduction or feature extraction due to the limited size of available datasets. Additionally, there are well-known data processing challenges, such as noise and artifact interference, high dimensionality and nonlinearity of the data, correlations between features in a time-series dataset, and variability in the analysis of results due to different filtering methods. Each of these issues represents a distinct research topic, and only after addressing these challenges can a dataset suitable for model training be constructed.

EMG raw data undergoes various filtering methods depending on the measurement environment and equipment. Typically, it passes through a bandpass filter to remove artifacts, followed by rectification to convert the signals to positive values, and then a low-pass filter to create the linear envelope [45]. Feature extraction can be categorized as follows: (1) statistical and physical features in the time domain, frequency domain, and time-frequency domain; (2) linearity features using entropy; and (3) dimensionality reduction features using matrix decomposition and neural networks.

Due to the challenges associated with sensitive signal processing in time-series data, traditional approaches to stroke detection in machine learning have relied on neuroimaging techniques [46], using MRI and CT scans to validate classification model performance. However, these methods reflect spatial rather than temporal characteristics, making early detection modeling more challenging. Furthermore, limitations such as restricted access to data acquisition and the computational costs associated with analysis have emerged as significant constraints. Recently, the potential of EMG data for stroke detection has garnered renewed attention, driven by extensive research on temporal signal processing for early stroke detection and advancements in feature engineering [47]. A recent study reported performance levels ranging from 59% to 80% when statistical features, derived from power spectrum analysis of EMG data from four muscles (bilateral biceps femoris and gastrocnemius muscles), were used to train various machine learning models [48].

## 3. Materials and Methods

### *3.1. Data Acquisition*

In this study, data from a total of 240 participants were used, consisting of 120 healthy controls and 120 stroke patients with an abnormal gait. Stroke patients were selected if they were able to walk, regardless of age, duration since onset, or the affected brain region. Participants with cognitive impairments, neurological damage, musculoskeletal injuries, or pain that prevented them from performing the gait task were excluded from this study. All participants were recruited from multiple medical institutions across South Korea, starting in 2022, following approval from the Institutional Review Board, and written informed consent was obtained from each participant. Surface electromyography (EMG) (Noraxon Telemyo DTS, Scottsdale, AZ, USA) was employed to record signals from both legs during a 10 m straight walk at a comfortable speed. To prevent falls during gait in stroke patients, a physical therapist accompanied the patients while they performed the walking task. Also, EMG electrodes were attached after cleaning the skin with alcohol to reduce skin resistance, and hair was removed if necessary. To ensure accurate and consistent EMG placement, data were collected by a biomechanics researcher with over three years of experience. The following muscles were monitored: the rectus femoris (RFM) and vastus medialis (VMM) acting as knee extensors; the semitendinosus (SEM) and biceps femoris long head (BFM) acting as knee flexors; the adductor magnus (ADD) as a hip adductor; the gluteus medius (GLU) as a hip abductor; the tibialis anterior (TIB) as an ankle dorsiflexor; and the medial gastrocnemius (GCM) as an ankle plantar flexor. EMG electrodes were placed according to SENIAM guidelines [49], and the signals were sampled at 1500 Hz after being bandpass-filtered (20–450 Hz, 4th-order Butterworth filter).

### *3.2. Pre-Processing of Data*

After EMG collection, all data were processed using rectification followed by low-pass filtering (10 Hz, 4th-order Butterworth filter), a standard approach commonly used in clinical research as shown in Figure 1 [50–52]. Ten gait cycles from each participant's 10 m walk were interpolated to 200 time points and averaged. As shown in Figure 2, data splitting was performed after data pre-processing and feature extraction before training the machine learning model. Although sufficient data was secured, in order to understand the classification problem related to the specific condition of stroke, the dataset was divided into training, validation, and test sets with a ratio of 2:6:2. This setup was chosen to avoid artificially inflating the learning effect and to minimize the potential for overfitting in advance. This approach was taken to enhance the model's generalization performance.

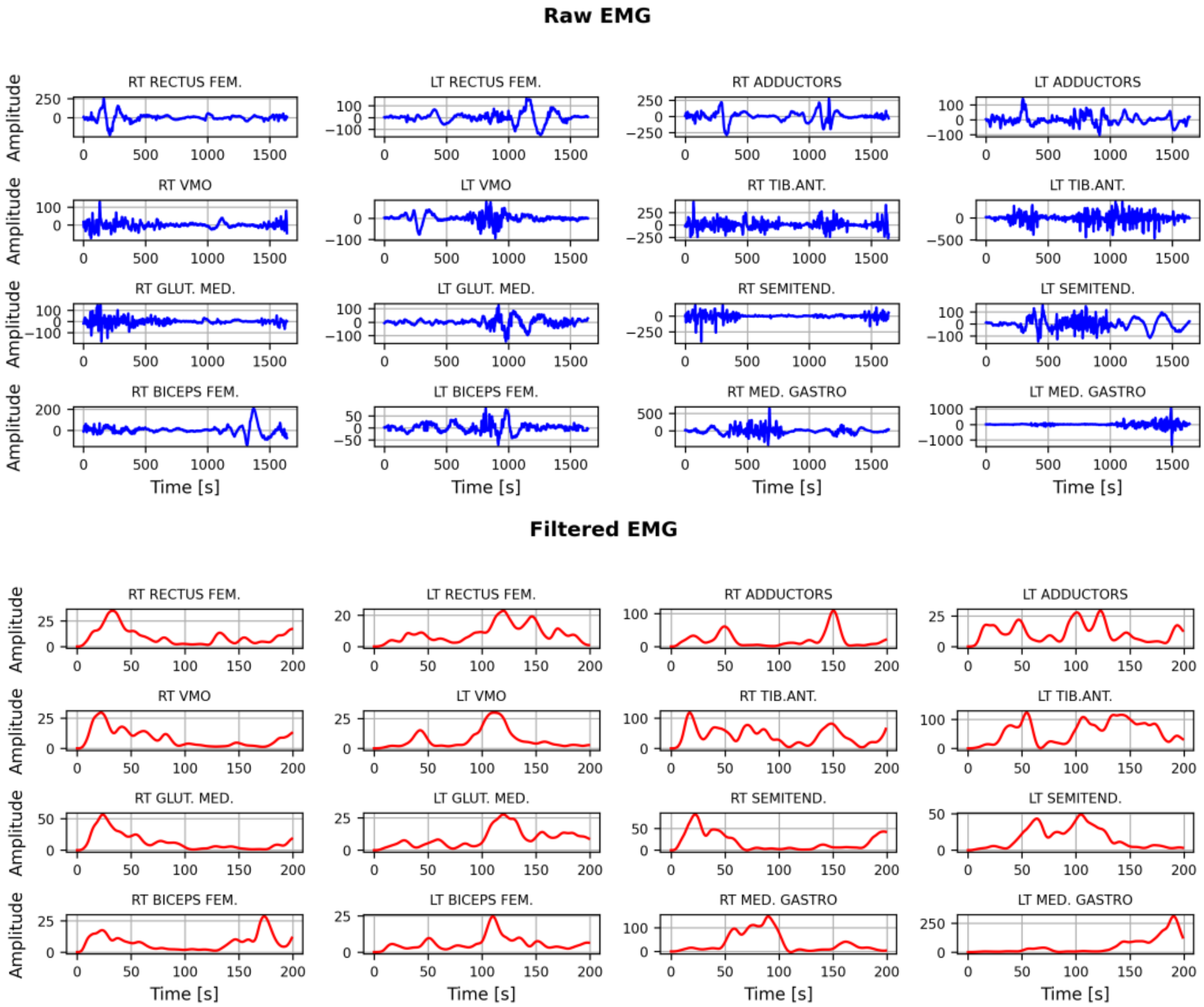


**Figure 1.** A sample of raw EMG data and filtered EMG data used for training 6 machine learning models.

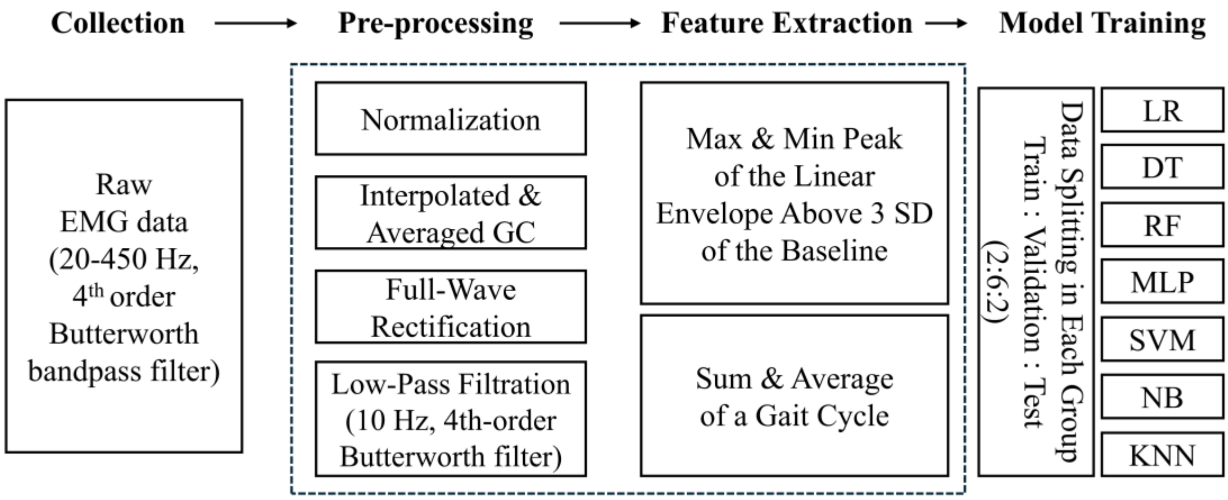


**Figure 2.** Schematic of EMG data processing and feature extraction procedures.

### *3.3. Feature Extraction of EMG*

Each participant's data included basic demographic information such as gender, age, weight, and height. This basic information was useful in the initial analysis to understand the overall characteristics of the participants. The primary focus of this study was on EMG measurements, with 16 EMG signals recorded for each participant. EMG provides valuable insights into the electrical activity of muscles, offering information related to muscle dysfunction. Based on this, for each filtered EMG signal, we extracted computational features, including the minimum peak (min) and maximum peak (max) of muscle activity exceeding 3 SD above the resting activity levels, as well as the total average (avg) and total sum (sum) of activity during the entire gait cycle. Here, the resting activity levels are defined as the average muscle activity during a 100 ms period while the participant is standing still before the gait cycle begins [53]. For example, as illustrated in the bottom panel of Figure 1, using the Lt. VMO as a reference, within a single gait cycle, there are two peak points above the baseline muscle activity, with the larger peak defined as the max and the smaller peak defined as the min. As a result, a total of 64 feature variables were derived for each participant (16 EMG signals $\times$ 4 derived features). These features were chosen because they provide a numeric representation that is both representative and intuitive. Unlike a power spectrum analysis, which may be less straightforward for non-specialists, these features offer an easily interpretable summary of muscle activity. This makes them more accessible in clinical settings where practitioners may not have specialized expertise. Additionally, these features are computationally efficient to extract during pre-processing, which helps streamline the analysis process and enhances the overall efficiency of the model.

### *3.4. Machine Learning Algorithms*

In this study, we classified the stroke group and the healthy group using seven of the most widely used machine learning models based on EMG feature data collected during the 10 m walk. The machine learning models were trained and evaluated using the scikit-learn library in Python, and the results were visualized using matplotlib. In the scikit-learn library, a logistic regression was implemented using the logistic regression module, a decision tree with the decision tree classifier module, random forest with the random forest classifier module, multilayer perceptron with the MLP classifier module, support vector machine with the SVC module, naive Bayes with the Gaussian–NB module, and K-nearest neighbors with the K-neighbors classifier module. The seven machine learning models utilized in this study were selected based on their proven efficacy in learning high-dimensional and non-linear data through each mathematical operation and statistical method. As shown in Table 1, the selection of each model was supported by previous studies that validated disease classification models using EMG data, thereby demonstrating their suitability for the dataset employed in this study.

**Table 1.** Classification performance of each model using EMG reported in previous studies.

| Model | Disease | Performance (%) | | |
|---|---|---|---|---|
| | | ACC | TPR | TNR |
| Logistic Regression [54] | ALS | 95.1 | 96 | 94.3 |
| Decision Tree [55] | ALS | 90.9 | 94 | 88.7 |
| Random Forest [55] | ALS | 96.7 | 99.6 | 94.8 |
| MultiLayer Perceptron [56] | Parkinson | 97 | 100 | 94 |
| Support Vector Machine [54] | ALS | 94.1 | 96 | 94.3 |
| Naive Bayes [56] | Parkinson | 97 | 100 | 94 |
| K-Nearest Neighbors [57] | NMDs | 98.8 | 98.6 | 99.4 |

ALS: amyotrophic lateral sclerosis; NMDs: neuromuscular disorders.

The objective was to identify the model that best learns from the EMG-driven features among these models. The definition and the equation are presented as follows:

- Logistic regression (LR): A widely used model for binary classification problems, logistic regression represents the relationship between feature variables and the outcome using a linear function [58].

$$\text{logit}(P(Y=1)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n \tag{1}$$

  where, logit(P(Y = 1)) represents the log-odds of the probability of belonging to class 1, and $\beta_n$ denotes the weight for each feature variable $X_n$.

- Decision tree (DT): A non-parametric model that performs predictions by partitioning data based on its features. It is known for its intuitive nature and ease of interpretation [59].

$$Gini = 1 - \sum_{k=1}^{K} {P_k}^2 \tag{2}$$

  where, $P_k$ represents the probability of belonging to class $k$.

- Random forest (RF): A model that improves predictive performance by assembling multiple decision trees. It is effective in preventing overfitting [60].

$$\hat{Y} = \text{mode}(\{T_i(X)\}) \tag{3}$$

  where, $T_i(X)$ represents the prediction result of each tree, and the final prediction is determined by majority voting.

- Multilayer perceptron (MLP): An artificial neural network model composed of multiple layers of neurons, known for its ability to effectively learn non-linear relationships [61].

$$\hat{Y} = f(\sum_{i=1}^{n} w_i X_i + b) \tag{4}$$

  where, $f$ represents the activation function, $w_i$ denotes the weights, and $b$ is the bias.

- Support vector machine (SVM): A model that finds a hyperplane to separate data in a high-dimensional space, capable of solving non-linear problems through the use of kernel tricks [62].

$$f(x) = \text{sign}(w \cdot x + b) \tag{5}$$

  where, $w$ represents the weight vector, $b$ is the bias, and the kernel function used is the RBF.

- Naive Bayes (NB): A probabilistic model based on Bayes' theorem, which performs predictions under the assumption that features are independent [63].

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \tag{6}$$

Here, $P(Y|X)$ represents the conditional probability, $P(Y)$ is the prior probability, and $P(X)$ is the evidence.

- K-nearest neighbors (KNN): A model that classifies a new data point by comparing it to the K nearest neighbors. Despite its simplicity, KNN can exhibit robust performance [64].

$$\hat{Y} = mode(\{Y_i | X_i \in N_k(X)\}) \tag{7}$$

Here, $N_k(X)$ represents the K nearest neighbors to $X$.

### 3.5. Model Performance Evaluation Matrices

The performance of the models was evaluated based on the accuracy of classifying each participant as either healthy or a stroke patient. Accuracy, which represents the proportion of correct predictions out of the total predictions, reflects the overall performance of the model. To gain insights into the types of errors made by the model, a confusion matrix was used to illustrate the relationship between the actual class and the predicted class, allowing us to identify the types of errors the model frequently makes.

### 3.6. Hyperparameter Optimization

Initially, the default hyperparameters provided by the scikit-learn library were used for each model to establish a baseline performance. Subsequently, a grid search was employed to find the optimal settings and enhance model performance. Hyperparameter tuning plays a critical role in balancing model complexity and generalization ability. Various configurations were tested, and cross-validation was used to validate the results. The key hyperparameters for each model are as follows:

- LR: penalty = 'l2', C = 1.0, solver = 'lbfgs'
- DT: criterion = 'gini', max_depth = None, min_samples_split = 2
- RF: n_estimators = 100, criterion = 'gini', max_depth = None
- MLP: hidden_layer_sizes = (100,), activation = 'relu', solver = 'adam', max_iter = 500
- SVM: C = 1.0, kernel = 'rbf', gamma = 'scale'
- NB: priors = None, var_smoothing = $1 \times 10^{-9}$
- KNN: n_neighbors = 5, weights = 'uniform', algorithm = 'auto'

## 4. Results

### 4.1. Performance of ML Models

The performance of each model on the validation and test data is summarized in Table 2. The RF model demonstrated the highest performance with 100% accuracy (ACC), true positive rate (TPR), true negative rate (TNR), and positive predictive value (PPV) for both the validation and test data. The MLP, LR, DT, SVM, and KNN models also exhibited excellent predictive performance, suggesting that various models are effective for stroke prediction. In contrast, the NB model exhibited a relatively lower performance, which may have been due to the assumption of independence among the feature variables not being met in the actual data.

### 4.2. Model with the Best Performance: RF

The RF model exhibited the highest performance, with 100% accuracy for both the validation data and the test data. This can be attributed to the ensemble of multiple DTs, which helps prevent overfitting and enhances the generalization performance of the model. The confusion matrices for the RF model validation and test data are summarized in Tables 3 and 4, respectively.

**Table 2.** Performance metrics of each model on validation and test data, where stroke patients are classified as positive cases.

| Model | Validation | | | | Test | | | |
|---|---|---|---|---|---|---|---|---|
| | **ACC** | **TPR** | **TNR** | **PPV** | **ACC** | **TPR** | **TNR** | **PPV** |
| Logistic Regression | 0.96 | 0.96 | 0.96 | 0.96 | 0.92 | 0.97 | 0.90 | 0.97 |
| Decision Tree | 0.94 | 0.93 | 0.96 | 0.96 | 0.96 | 1.00 | 0.90 | 0.78 |
| Random Forest | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| MultiLayer Perceptron | 0.99 | 0.99 | 0.99 | 0.99 | 0.98 | 0.93 | 0.95 | 0.96 |
| Support Vector Machine | 0.94 | 0.97 | 0.66 | 0.72 | 0.94 | 0.93 | 0.76 | 0.83 |
| Naive Bayes | 0.77 | 0.93 | 0.63 | 0.69 | 0.77 | 0.93 | 0.57 | 0.74 |
| K-Nearest Neighbors | 0.85 | 0.97 | 0.75 | 0.77 | 0.96 | 0.96 | 0.95 | 0.96 |

**Table 3.** Confusion matrix for the validation data.

| True \ Prediction | Healthy | Stroke |
|---|---|---|
| Healthy | 76 | 0 |
| Stroke | 0 | 68 |

**Table 4.** Confusion matrix for the test data.

| True \ Prediction | Healthy | Stroke |
|---|---|---|
| Healthy | 21 | 0 |
| Stroke | 0 | 27 |

Feature Importance

Variables with high feature importance in the RF model are shown in Figure 3. The top five variables ranked by importance were as follows: RT VMO avg (0.145), LT VMO avg (0.138), RT RECTUS FEM max (0.112), LT RECTUS FEM avg (0.109), and RT MED. GASTRO max (0.107).

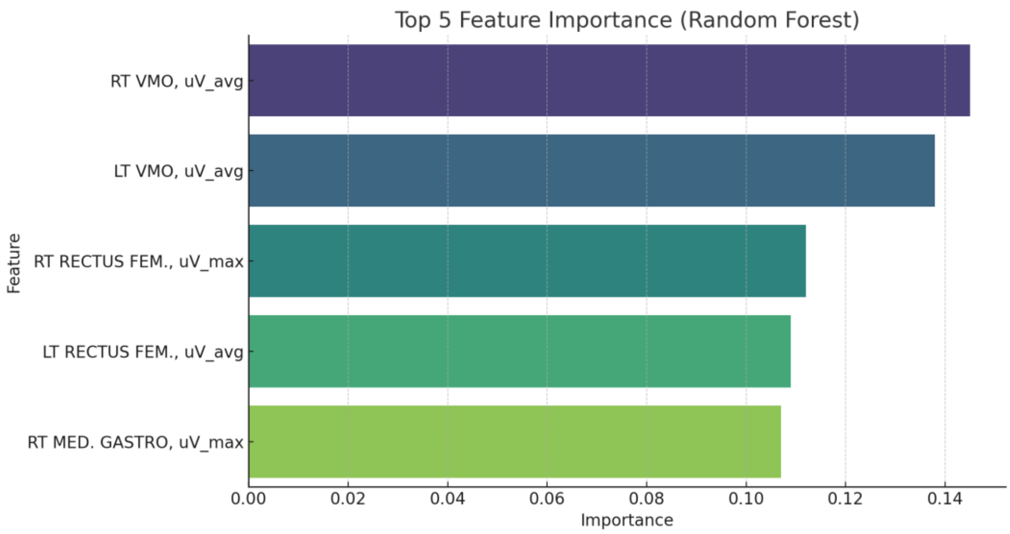


**Figure 3.** Variables with high feature importance in the RF model.

## 5. Discussion

This study examined the performance and feature importance of ML models trained on key statistical features of EMG data collected from both legs of stroke patients and healthy controls while walking. Among the seven models, the RF demonstrated the highest performance, and the confusion matrix showed that the RF model accurately classified all healthy and abnormal cases. This indicates that the model generalizes well to both the training and actual data. The muscle with the highest feature importance was the mean value of the VMM. Additionally, the maximum and mean values of the RFM emerged as important variables, highlighting the significance of EMG measurements and features derived from knee extensors among the various leg muscles for stroke prediction. This suggests that specific muscle EMG signals can be used efficiently to assess the likelihood of stroke in the presence of suspected warning signs or in early stages.

The results of the predictive models using the training data in this study are supported by previous studies that identified quadriceps weakness as a key clinical symptom in patients [3,65–67]. This study emphasizes the importance of the VMM in the quadriceps as a significant predictive factor. Following the VMM, it was found that the RFM and GCM were the significant predictors. This finding is also consistent with clinical observations and studies regarding the impact of stroke on these specific muscle groups.

The RFM, a key muscle in the quadriceps group along with the VMM, plays a critical role in gait, particularly during the weight acceptance phase [40]. This reflects its high importance in our model. Whereas the GCM is essential for ankle plantarflexion and overall gait dynamics during the propulsion phase [40]. Stroke can severely affect this muscle's function, leading to a decrease in gait speed. The feature importance of the GCM in our model corroborates the well-documented impact of stroke on lower limb function and emphasizes the importance of this muscle in maintaining effective mobility and balance.

This finding highlights how machine learning models can effectively capture critical aspects of motor dysfunction associated with stroke, providing valuable insights for clinical assessments and rehabilitation strategies. This correlation not only validates the model's performance but also reinforces the clinical significance of these muscle groups in the context of stroke rehabilitation. Additionally, although beyond the primary scope of this study, feature importance analysis indicated higher values for females, which aligns with global statistics showing a higher stroke risk among women [13]. Furthermore, this study did not differentiate between the paralyzed and non-paralyzed legs of stroke patients, as both legs can experience kinematic damage in stroke patients [66].

In this study, the RF model exhibited the best performance with EMG-driven features, demonstrating its effectiveness in capturing the nonlinearity and feature interactions among the 64 features, as well as its robustness in preventing overfitting. Given that EMG data involves correlations between muscle activations, it is crucial to consider these interdependencies during the learning process. Models like NB, which assume feature independence, showed diminished performance. While models such as SVM and MLP are known for their ability to handle high-dimensional data, their performance is highly sensitive to hyperparameter tuning. As a result, this study alone cannot definitively conclude that RF surpasses SVM and MLP.

Stroke is a major disease that significantly affects the quality of life and leads to severe sequelae in the elderly. Early detection and prompt intervention based on warning signs can be highly beneficial not only in improving outcomes but also in reducing healthcare and socioeconomic costs for the elderly [68]. A recent study reported several barriers to the clinical evaluation and application of EMG [69]. The results of this study demonstrate that predictive models can perform effectively using only the mean or maximum values of muscle activity of the knee extensors without requiring complex analysis of clinical significance or symptom testing. This approach addresses both the critical issue of computational efficiency in time-series data and the challenge of technological accessibility in clinical settings [70]. Rapid initial predictions using ML based on computationally efficient time-series data followed by precise evaluations are expected to alleviate bottlenecks in future medical diagnostic systems [71]. Moreover, because the simultaneous measurement of multiple muscles is not required, managing data quality related to noise and outliers can be managed more easily. Although various studies have proposed the use of ML models for disease prediction, research that validates and examines the performance and impact of stroke prediction models based on the activation of multiple muscles while considering the complex joint degrees of freedom in the lower limbs, as in this study, is rare [72].

This study represents a pilot investigation into the feasibility of utilizing a limited set of muscle activity data for ML-based predictive models. To advance these models for clinical applications, conducting further research that incorporates a range of factors related to stroke patients is crucial, including the etiology of impairment, lesion localization, extent of recovery, and prognosis, by leveraging muscle activity data from specific muscle groups. Moreover, subsequent research should focus on enhancing not only the sensitivity of the models for stroke detection but also their specificity in distinguishing stroke from other diseases. This involves integrating additional risk factors, such as lifestyle habits, to improve diagnostic accuracy and model robustness.

Although 5-fold and 10-fold cross-validation were conducted as supplementary analyses to validate the performance of the random forest (RF), which demonstrated superior results in this study, yielding a performance metric of 1.00 for all indicators (i.e., ACC, TPR, TNR, PPV) as shown in Table 2, it is important to note that an ablation study involving additional datasets and comparisons with other machine learning models was not performed. Consequently, there are limitations in generalizing our findings.

EMG data are continuously accumulated across medical institutions worldwide, and wearable devices for EMG collection are becoming increasingly available in everyday life [73]. This study confirmed that EMG data can be effectively utilized for classifying stroke patients versus healthy individuals using various ML models, highlighting their potential for advancing digital healthcare technologies. Because EMG data do not identify individual participants, sensitive issues related to personal privacy are avoided. Therefore, future research could further enhance the prediction and differentiation performance by leveraging stored EMG data from medical institutions, focusing on advanced feature engineering and model refinement. Furthermore, this study is significant in that it confirmed the effectiveness of a predictive model using only certain EMG data and features, amidst the various medical information and numerous data features available from stroke patients. Specifically, further research is needed to validate the following directions: First, there may be specific muscles that are particularly effective for training purposes when measuring EMG for predictive modeling. Second, among the muscle activity data, certain features may be more effective for training. Third, some predictive models may be more effective for learning from EMG data. Finally, future research in this direction may contribute to addressing issues such as increased noise, overfitting, and computational cost that can arise from using an excessive number of data features in AI model development in the healthcare field.

## 6. Conclusions

This study examined the performance and importance of stroke classification models based on the primary statistical characteristics of EMG data from key muscles during walking. These findings confirm that using the simplest feature extraction techniques (maximum and average) and collecting muscle activity data from a limited number of muscles (knee extensors) can yield highly effective models. This suggests the potential of future wearable devices for predicting stroke risk in both the home and clinical environments through gait analysis. However, this study is limited by the lack of validation of the model's ability to generalize EMG-driven data for distinguishing gait patterns in patients with other neurological disorders. Subsequent research should focus on integrating extensive EMG data collected from specific muscles across medical institutions worldwide with other life-long and medical data to enhance the diagnostic and differentiating capabilities of predictive modeling.

**Author Contributions:** Conceptualization, J.L. and E.K.; methodology, E.K.; software, J.L.; validation, J.L. and Y.K.; formal analysis, E.K.; investigation, J.L. and Y.K.; resources, J.L.; data curation, J.L.; writing—original draft preparation, J.L. and E.K.; writing—review and editing, E.K.; visualization, J.L. and E.K.; supervision, E.K.; project administration, Y.K.; funding acquisition, Y.K. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP)—Innovative Human Resource Development for Local Intellectualization program grant funded by the Korea government (MSIT) (IITP-2024-RS-2024-00436773).

**Institutional Review Board Statement:** This study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board of Korea University (KUIRB2018-0127) for studies involving humans.

**Informed Consent Statement:** Informed consent was obtained from all participants involved in this study.

**Data Availability Statement:** The data presented in this study are available on request from the corresponding author due to privacy restrictions.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Murphy, S.J.; Werring, D.J. Stroke: Causes and clinical features. *Medicine* **2020**, *48*, 561–566. [CrossRef] [PubMed]
2. Hendricks, H.T.; Van Limbeek, J.; Geurts, A.C.; Zwarts, M.J. Motor recovery after stroke: A systematic review of the literature. *Arch. Phys. Med. Rehabil.* **2002**, *83*, 1629–1637. [CrossRef] [PubMed]
3. Harris, M.L.; Polkey, M.I.; Bath, P.M.; Moxham, J. Quadriceps muscle weakness following acute hemiplegic stroke. *Clin. Rehabil.* **2001**, *15*, 274–281. [CrossRef] [PubMed]
4. Sommerfeld, D.K.; Eek, E.U.B.; Svensson, A.K.; Holmqvist, L.W.; Von Arbin, M.H. Spasticity after stroke: Its occurrence and association with motor impairments and activity limitations. *Stroke* **2004**, *35*, 134–139. [CrossRef] [PubMed]
5. Statistics Korea. *Annual Report on the Causes of Death Statistics*; Statistics Korea: Daejeon, Republic of Korea, 2020.
6. Truelsen, T.; Piechowski-Jóźwiak, B.; Bonita, R.; Mathers, C.; Bogousslavsky, J.; Boysen, G. Stroke incidence and prevalence in Europe: A review of available data. *Eur. J. Neurol.* **2006**, *13*, 581–598. [CrossRef]
7. Balaban, B.; Tok, F. Gait disturbances in patients with stroke. *Pmr* **2014**, *6*, 635–642. [CrossRef]
8. Van Criekinge, T.; Vermeulen, J.; Wagemans, K.; Schröder, J.; Embrechts, E.; Truijen, S.; Hallemans, A.; Saeys, W. Lower limb muscle synergies during walking after stroke: A systematic review. *Disabil. Rehabil.* **2020**, *42*, 2836–2845. [CrossRef]
9. Brunner, R.; Romkes, J. Abnormal EMG muscle activity during gait in patients without neurological disorders. *Gait Posture* **2008**, *27*, 399–407. [CrossRef]
10. Banks, C.L.; Huang, H.J.; Little, V.L.; Patten, C. Electromyography exposes heterogeneity in muscle co-contraction following stroke. *Front. Neurol.* **2017**, *8*, 699. [CrossRef]
11. Pilkar, R.; Ramanujam, A.; Nolan, K.J. Alterations in spectral attributes of surface electromyograms after utilization of a foot drop stimulator during post-stroke gait. *Front. Neurol.* **2017**, *8*, 449. [CrossRef]
12. Marler, J.R.; Tilley, B.C.; Lu, M.; Brott, T.G.; Lyden, P.C.; Grotta, J.C.; Broderick, J.P.; Levine, S.R.; Frankel, M.P.; E.C NINDS rt-PA Stroke Study Group; et al. Early stroke treatment associated with better outcome: The NINDS rt-PA stroke study. *Neurology* **2000**, *55*, 1649–1655. [CrossRef] [PubMed]
13. Tsao, C.W.; Aday, A.W.; Almarzooq, Z.I.; Anderson, C.A.; Arora, P.; Avery, C.L.; Baker-Smith, C.M.; Beaton, A.Z.; Boehme, A.K.; Buxton, A.E.; et al. American Heart Association Council on Epidemiology and Prevention Statistics Committee and Stroke Statistics Subcommittee. Heart disease and stroke statistics—2023 update: A report from the American Heart Association. *Circulation* **2023**, *147*, e93–e621. [CrossRef]
14. Batchelor, F.A.; Williams, S.B.; Wijeratne, T.; Said, C.M.; Petty, S. Balance and gait impairment in transient ischemic attack and minor stroke. *J. Stroke Cerebrovasc.* **2015**, *24*, 2291–2297. [CrossRef]
15. Li, N.; Li, J.; Gao, T.; Wang, D.; Du, Y.; Zhao, X. Gait and balance disorder in patients with transient ischemic attack or minor stroke. *Neuropsych. Dis. Treat.* **2021**, *2021*, 305–314. [CrossRef]
16. Coutts, S.B. Diagnosis and management of transient ischemic attack. *Continuum* **2017**, *23*, 82–92. [CrossRef] [PubMed]
17. Kaushik, S.; Choudhury, A.; Sheron, P.K.; Dasgupta, N.; Natarajan, S.; Pickett, L.A.; Dutt, V. AI in healthcare: Time-series forecasting using statistical, neural, and ensemble architectures. *Front. Big Data* **2020**, *3*, 4. [CrossRef] [PubMed]
18. Frigo, C.; Crenna, P. Multichannel SEMG in clinical gait analysis: A review and state-of-the-art. *Clin. Biomech.* **2009**, *24*, 236–245. [CrossRef]
19. Farago, E.; MacIsaac, D.; Suk, M.; Chan, A.D. A review of techniques for surface electromyography signal quality analysis. *IEEE Rev. Biomed. Eng.* **2022**, *16*, 472–486. [CrossRef]

20. Chatterjee, S.; Chatterjee, A.; Saha, S.; Mukherjee, D.; Sen, R.; Banerjee, S.; Mandal, D. EMG Signal Acquisition and Processing for Feature Extraction And Detection of Disease. *J. Eng. Technol. Manag.* **2024**, *19*, 20.
21. Pandey, B.; Mishra, R.B. An integrated intelligent computing model for the interpretation of EMG based neuromuscular diseases. *Expert Syst. Appl.* **2009**, *36*, 9201–9213. [CrossRef]
22. Seyidbayli, C.; Salhi, F.; Akdogan, E. Comparison of machine learning algorithms for EMG signal classification. *Period. Eng. Nat. Sci.* **2020**, *8*, 1165–1176.
23. Tannemaat, M.R.; Kefalas, M.; Geraedts, V.J.; Remijn-Nelissen, L.; Verschuuren, A.J.M.; Koch, M.; Kononova, A.V.; Wang, H.; Bäck, T.H.W. Distinguishing normal, neuropathic and myopathic EMG with an automated machine learning approach. *Clin. Neurophysiol. Pract.* **2023**, *146*, 49–54. [CrossRef] [PubMed]
24. Khairuddin, I.M.; Sidek, S.N.; Majeed, A.P.A.; Razman, M.A.M.; Puzi, A.A.; Yusof, H.M. The classification of movement intention through machine learning models: The identification of significant time-domain EMG features. *PeerJ Comput. Sci.* **2021**, *7*, e379. [CrossRef]
25. Yaman, E.; Subasi, A. Comparison of bagging and boosting ensemble machine learning methods for automated EMG signal classification. *Biomed Res. Int.* **2019**, *2019*, 9152506. [CrossRef]
26. Nasr, A.; Bell, S.; He, J.; Whittaker, R.L.; Jiang, N.; Dickerson, C.R.; McPhee, J. MuscleNET: Mapping electromyography to kinematic and dynamic biomechanical variables by machine learning. *J. Neural Eng.* **2021**, *18*, 0460d3. [CrossRef] [PubMed]
27. Sirsat, M.S.; Fermé, E.; Camara, J. Machine learning for brain stroke: A review. *J. Stroke Cerebrovasc.* **2020**, *29*, 105162. [CrossRef]
28. Shanthi, D.; Sahoo, G.; Saravanan, N. Designing an Artificial Neural Network Model for the Prediction of Thrombo-embolic Stroke. *Int. J. Biom. Bioinform. IJBB* **2009**, *3*, 10–18. Available online: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=73ea8ec7e98573e90c56e5857c04b6ce68ae31fc (accessed on 1 September 2024).
29. Chin, C.L.; Lin, B.J.; Wu, G.R.; Weng, T.C.; Yang, C.S.; Su, R.C.; Pan, Y.J. An automated early ischemic stroke detection system using CNN deep learning algorithm. In Proceedings of the 2017 IEEE 8th International Conference on Awareness Science and Technology (iCAST), Taichung, Taiwan, 8–10 November 2017; pp. 368–372. [CrossRef]
30. Yu, J.; Park, S.; Lee, H.; Pyo, C.S.; Lee, Y.S. An elderly health monitoring system using machine learning and in-depth analysis techniques on the NIH stroke scale. *Mathematics* **2020**, *8*, 1115. [CrossRef]
31. Gandolla, M.; Ferrante, S.; Ferrigno, G.; Baldassini, D.; Molteni, F.; Guanziroli, E.; Cotti Cottini, M.; Seneci, C.; Pedrocchi, A. Artificial neural network EMG classifier for functional hand grasp movements prediction. *J. Int. Med. Res.* **2017**, *45*, 1831–1847. [CrossRef]
32. Jochumsen, M.; Niazi, I.K.; Zia ur Rehman, M.; Amjad, I.; Shafique, M.; Gilani, S.O.; Waris, A. Decoding attempted hand movements in stroke patients using surface electromyography. *Sensors* **2020**, *20*, 6763. [CrossRef]
33. Noor, A.; Waris, A.; Gilani, S.O.; Kashif, A.S.; Jochumsen, M.; Iqbal, J.; Niazi, I.K. Decoding of ankle joint movements in stroke patients using surface electromyography. *Sensors* **2021**, *21*, 1575. [CrossRef] [PubMed]
34. Choi, Y.A.; Park, S.J.; Jun, J.A.; Pyo, C.S.; Cho, K.H.; Lee, H.S.; Yu, J.H. Deep learning-based stroke disease prediction system using real-time bio signals. *Sensors* **2021**, *21*, 4269. [CrossRef]
35. Raurale, S.A.; McAllister, J.; del Rincon, J.M. Real-time embedded EMG signal analysis for wrist-hand pose identification. *IEEE Trans. Signal Process.* **2020**, *68*, 2713–2723. [CrossRef]
36. Unanyan, N.N.; Belov, A.A. Design of upper limb prosthesis using real-time motion detection method based on EMG signal processing. *Biomed. Signal Process. Control.* **2021**, *70*, 103062. [CrossRef]
37. Saikia, A.; Mazumdar, S.; Sahai, N.; Paul, S.; Bhatia, D. Performance analysis of artificial neural network for hand movement detection from EMG signals. *IETE J. Res.* **2022**, *68*, 1074–1083. [CrossRef]
38. Guo, Y.; Gravina, R.; Gu, X.; Fortino, G.; Yang, G.Z. EMG-based abnormal gait detection and recognition. In Proceedings of the 2020 IEEE International Conference on Human-Machine Systems (ICHMS), Rome, Italy, 7–9 September 2020; pp. 1–6.
39. Nishikawa, Y.; Holobar, A.; Watanabe, K.; Takahashi, T.; Ueno, H.; Maeda, N.; Maruyama, H.; Tanaka, S.; Hyngstrom, A.S. Detecting motor unit abnormalities in amyotrophic lateral sclerosis using high-density surface EMG. *Clin. Neurophysiol.* **2022**, *142*, 262–272. [CrossRef]
40. Lee, J. Characteristics of muscle synergy extracted using an autoencoder in patients with stroke during the curved walking in comparison with healthy controls. *Gait Posture* **2024**, *107*, 225–232. [CrossRef]
41. Kalaivani, K.; Kshirsagarr, P.R.; Sirisha Devi, J.; Bandela, S.R.; Colak, I.; Nageswara Rao, J.; Rajaram, A. Prediction of biomedical signals using deep learning techniques. *J. Intell. Fuzzy Syst.* **2023**, *44*, 9769–9782. [CrossRef]
42. Nair, S.S.; French, R.M.; Laroche, D.; Thomas, E. The application of machine learning algorithms to the analysis of electromyographic patterns from arthritic patients. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2009**, *18*, 174–184. [CrossRef]
43. Kamruzzaman, J.; Begg, R.K. Support vector machines and other pattern recognition approaches to the diagnosis of cerebral palsy gait. *IEEE Trans. Biomed. Eng.* **2006**, *53*, 2479–2490. [CrossRef]
44. Kugler, P.; Jaremenko, C.; Schlachetzki, J.; Winkler, J.; Klucken, J.; Eskofier, B. Automatic recognition of Parkinson's disease using surface electromyography during standardized gait tests. In Proceedings of the 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Osaka, Japan, 3–7 July 2013; pp. 5781–5784.
45. Norali, A.N.; Som, M.; Kangar-Arau, J. Surface electromyography signal processing and application: A review. In Proceedings of the International Conference on Man-Machine Systems (ICoMMS), Online, 11–13 October 2009.

46. Ashrafuzzaman, M.; Saha, S.; Nur, K. Prediction of stroke disease using deep CNN based approach. *J. Adv. Inf. Technol.* **2022**, *13*, 604–613. [CrossRef]
47. Ochei, L.C.; Egbono, F. Evaluation of Approaches for Early Stroke Detection and Diagnosis Using EMG Data: Features, Techniques, and Challenges. *Int. J. Intell. Inf. Syst.* **2024**, *12*, 29–42.
48. Hussain, I.; Park, S.J. Prediction of myoelectric biomarkers in post-stroke gait. *Sensors* **2021**, *21*, 5334. [CrossRef] [PubMed]
49. Stegeman, D.; Hermens, H. Standards for Surface Electromyography: The European Project Surface EMG for Non-Invasive Assessment of Muscles (SENIAM). *Enschede Roessingh Res. Dev.* **2007**, *10*, 8–12. Available online: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=b280c4751a2658380a77052b0aab7929e6943a57 (accessed on 1 September 2024).
50. Yeom, H.; Chang, Y.H. Autogenic EMG-controlled functional electrical stimulation for ankle dorsiflexion control. *J. Neurosci. Methods* **2010**, *193*, 118–125. [CrossRef]
51. Suhaimi, R.; Aswad, A.R.; Adnan, N.H.; Asyraf, F.; Wan, K.; Hazry, D.; Shahriman, A.B.; Bakar, J.A.A.; Shahriman, A.B.; Bakar, J.A.A.; et al. Analysis of EMG-based muscles activity for stroke rehabilitation. In Proceedings of the 2014 2nd International Conference on Electronic Design (ICED), Penang, Malaysia, 19–21 August 2014; pp. 167–170. [CrossRef]
52. Sohn, M.H.; Deol, J.; Dewald, J.P. A novel time-based surface EMG measure for quantifying hypertonia in paretic arm muscles during daily activities after hemiparetic stroke. *medRxiv* **2022**. [CrossRef]
53. Tenan, M.S.; Tweedell, A.J.; Haynes, C.A. Analysis of statistical and standard algorithms for detecting muscle onset with surface electromyography. *PLoS ONE* **2017**, *12*, e0177312. [CrossRef]
54. Khan, M.U.; Aziz, S.; Bilal, M.; Aamir, M.B. Classification of EMG signals for assessment of neuromuscular disorder using empirical mode decomposition and logistic regression. In Proceedings of the 2019 International Conference on Applied and Engineering Mathematics (ICAEM), Taxila, Pakistan, 27–29 August 2019; pp. 237–243.
55. Gokgoz, E.; Subasi, A. Comparison of decision tree algorithms for EMG signal classification using DWT. *Biomed. Signal Process. Control.* **2015**, *18*, 138–144. [CrossRef]
56. Jebakumari, V.S.; Shanthi, D.; Sridevi, S.; Meha, P. Performance evaluation of various classification algorithms for the diagnosis of Parkinson's disease. In Proceedings of the 2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS), Srivilliputtur, India, 23–25 March 2017; pp. 1–7.
57. Achmamad, A.; Elfezazi, M.; Chehri, A.; Ahmed, I.; Jbari, A.; Saadane, R. ML-Based Identification of Neuromuscular Disorder Using EMG Signals for Emotional Health Application. *ACM Trans. Internet Technol.* **2023**. [CrossRef]
58. Kleinbaum, D.G.; Dietz, K.; Gail, M.; Klein, M.; Klein, M. *Logistic Regression*; Springer: New York, NY, USA, 2008; p. 536. Available online: http://ndl.ethernet.edu.et/bitstream/123456789/60520/1/131.pdf (accessed on 1 September 2024).
59. Maimon, O.Z.; Rokach, L. *Data Mining with Decision Trees: Theory and Applications*; World Scientific: Singapore, 2014; Volume 81.
60. Qi, Y. Random forest for bioinformatics. In *Ensemble Machine Learning: Methods and Applications*; Springer: New York, NY, USA, 2012; pp. 307–323. [CrossRef]
61. Bengio, Y.; Ducharme, R.; Vincent, P. A Neural Probabilistic Language Model. In Proceedings of the Advances in Neural Information Processing Systems, Denver, CO, USA, 1 January 2000; Volume 13. Available online: https://proceedings.neurips.cc/paper_files/paper/2000/file/728f206c2a01bf572b5940d7d9a8fa4c-Paper.pdf (accessed on 1 September 2024).
62. Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273–297. Available online: https://link-springer-com-443.webvpn.synu.edu.cn/content/pdf/10.1007/BF00994018.pdf (accessed on 1 September 2024). [CrossRef]
63. Domingos, P.; Pazzani, M. On the optimality of the simple Bayesian classifier under zero-one loss. *Mach. Learn.* **1997**, *29*, 103–130. [CrossRef]
64. Altman, N.S. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *Am. Stat.* **1992**, *46*, 175–185. Available online: https://ecommons.cornell.edu/bitstream/1813/31637/1/BU-1065-MA.pdf (accessed on 1 September 2024). [CrossRef]
65. Nozoe, M.; Kanai, M.; Kubo, H.; Kitamura, Y.; Yamamoto, M.; Furuichi, A.; Takashima, S.; Mase, K.; Shimada, S. Changes in quadriceps muscle thickness, disease severity, nutritional status, and C-reactive protein after acute stroke. *J. Stroke Cerebrovasc. Dis.* **2016**, *25*, 2470–2474. [CrossRef]
66. Akazawa, N.; Harada, K.; Okawa, N.; Tamura, K.; Moriyama, H. Muscle mass and intramuscular fat of the quadriceps are related to muscle strength in non-ambulatory chronic stroke survivors: A cross-sectional study. *PLoS ONE* **2018**, *13*, e0201789. [CrossRef]
67. Chow, J.W.; Stokic, D.S. Force control of quadriceps muscle is bilaterally impaired in subacute stroke. *J. Appl. Physiol.* **2011**, *111*, 1290–1295. [CrossRef]
68. Feigin, V.L.; Norrving, B.; George, M.G.; Foltz, J.L.; Roth, G.A.; Mensah, G.A. Prevention of stroke: A strategic global imperative. *Nat. Rev. Neurol.* **2016**, *12*, 501–512. [CrossRef] [PubMed]
69. Merletti, R.; Campanini, I.; Rymer, W.Z.; Disselhorst-Klug, C. Surface electromyography: Barriers limiting widespread use of sEMG in clinical assessment and neurorehabilitation. *Front Neurol.* **2021**, *12*, 642257. [CrossRef] [PubMed]
70. Fu, T.C. A review on time series data mining. *Eng. Appl. Artif. Intell.* **2011**, *24*, 164–181. [CrossRef]
71. Fernandes, L.; Santos, D.; Santos, M.; Rocha, N.P. How to improve emergency information systems to optimize the care of acute stroke. *Procedia Comput.* **2022**, *196*, 606–614. [CrossRef]

72. Dhal, P.; Azad, C. A comprehensive survey on feature selection in the various fields of machine learning. *Appl. Intell.* **2022**, *52*, 4543–4581. [CrossRef]
73. Pregnolato, G.; Rimini, D.; Baldan, F.; Maistrello, L.; Salvalaggio, S.; Celadon, N.; Ariano, P.; Pirri, C.F.; Turolla, A. Clinical Features to Predict the Use of a sEMG Wearable Device (REMO®) for Hand Motor Training of Stroke Patients: A Cross-Sectional Cohort Study. *Int. J. Environ. Res. Public Health* **2023**, *20*, 5082. [CrossRef] [PubMed]