

## Article

# Agreement and Reliability Analysis of Machine Learning Scaling and Wireless Monitoring in the Assessment of Acute Proximal Weakness by Experts and Non-Experts: A Feasibility Study

Eunjeong Park <sup>1</sup>, Kijeong Lee <sup>2</sup>, Taehwa Han <sup>3</sup> and Hyo Suk Nam <sup>4,\*</sup>

<sup>1</sup> Integrative Research Center for Cerebrovascular and Cardiovascular Diseases, Yonsei University College of Medicine, Seoul 03722, Korea; eunjeong.ej@gmail.com

<sup>2</sup> Department of Neurology, National Health Insurance Service, Ilsan Hospital, Goyang 10444, Korea; junon8263@gmail.com

<sup>3</sup> Health-IT Center, Yonsei University College of Medicine, Seoul 03722, Korea; taehwa.han@gmail.com

<sup>4</sup> Department of Neurology, Yonsei University College of Medicine, Seoul 03722, Korea

\* Correspondence: hsnam@yuhs.ac; Tel.: +82-2-2228-1617

**Abstract:** Assessing the symptoms of proximal weakness caused by neurological deficits requires the knowledge and experience of neurologists. Recent advances in machine learning and the Internet of Things have resulted in the development of automated systems that emulate physicians' assessments. The application of those systems requires not only accuracy in the classification but also reliability regardless of users' proficiency in the real environment for the clinical point-of-care and the personalized health management. This study provides an agreement and reliability analysis of using a machine learning-based scaling of Medical Research Council (MRC) proximal scores to evaluate proximal weakness by experts and non-experts. The system trains an ensemble learning model using the signals from sensors attached to the limbs of patients in a neurological intensive care unit. For the agreement analysis, we investigated the percent agreement of MRC proximal scores and Bland-Altman plots of kinematic features between the expert- and non-expert scaling. We also analyzed the intra-class correlation coefficients (ICCs) of kinematic features and Krippendorff's alpha of the observers' scaling for the reliability analysis. The mean percent agreement between the expert- and the non-expert scaling was 0.542 for manual scaling and 0.708 for autonomous scaling. The ICCs of kinematic features measured using sensors ranged from 0.742 to 0.850, whereas the Krippendorff's alpha of manual scaling for the three observers was 0.275. The autonomous assessment system can be utilized by the caregivers, paramedics, or other observers during an emergency to evaluate acute stroke patients.

**Keywords:** decision-support system; machine learning; artificial intelligence; sensors; inter-rater reliability; agreement analysis; stroke



**Citation:** Park, E.; Lee, K.; Han, T.; Nam, H.S. Agreement and Reliability Analysis of Machine Learning Scaling and Wireless Monitoring in the Assessment of Acute Proximal Weakness by Experts and Non-Experts: A Feasibility Study. *J. Pers. Med.* **2022**, *12*, 20. <https://doi.org/10.3390/jpm12010020>

Academic Editors: Anton Civit, Manuel Dominguez-Morales and Antonis Billis

Received: 2 December 2021

Accepted: 17 December 2021

Published: 1 January 2022

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

The influence of machine learning in medicine has rapidly increased in decision support for detecting symptoms [1]. Most medical artificial intelligence (AI) studies have shown that the promise of medical AI resides in the data used for informing the care of each patient, and the experience and knowledge of experts used for decision making [2]. Even specialists' decisions can be affected by other factors, including cognitive biases, overconfidence, excessive workloads, and personality traits. Such differences or biases in decision making among specialists and non-specialists can pose a grave problem in care plans, which affects the outcome of treatment, including the mortality of patients [3,4]. To solve this problem, many intelligence solutions have been adopted to reduce medical errors and to meet the efficiency requirements [5,6].

However, the proliferation of medical AI solutions has also raised the debate concerning whether they are reliable in a real environment. As argued by Patrick et al., prematurely released AI solutions can result in increased risks and workloads for clinicians [7]. AI-driven decisions are delicate to maintain the reliability and agreement between healthcare professionals and patients [6]. Specifically, the conversion from patients’ qualitative conditions to quantitative grades in severity scales is confusing and sensitive to determine. As a result, it is challenging to maintain reliability and consistency between graders.

The degree of difficulty in measuring and collecting data determines the availability and the capacity of data for training machine learning models. In addition, ordinal classification suffers from the disparity of data between classes as addressed in [8,9]. To solve this problem, researchers have proposed and achieved prominent solutions to develop high-performance AI systems with the small data set; data augmentation [10,11], transfer learning [12,13], construction of synthetic data [14–16] and ensemble learning [17–19] are the representative approaches that have resulted in outstanding solutions.

In this study, we investigated the agreement and reliability of automated grading system for non-experts who need to determine the severity of symptoms among patients with acute stroke. The autonomous grading system determines the Medical Research Council (MRC) scale, which is widely accepted and frequently used in a clinical environment to assess neurological conditions and muscle strength (Table 1) [20,21].

**Table 1.** Modified MRC scales for assessing proximal weakness. MRC, Medical Research Council.

MRC Scale (6-Point Scale)	Response
9 (V)	Normal power
8 (IV+)	Muscle holds the joint against a combination of gravity and moderate resistance, but muscle holds the joint against moderate to maximal resistance
7 (IV)	Muscle holds the joint against a combination of gravity and moderate resistance
6 (III+)	Muscle holds the joint against a combination of gravity and moderate resistance, but muscle holds the joint only against minimal resistance
5 (III)	Muscle moves the joint fully against gravity and is capable of transient resistance, but collapses abruptly
4 (II+)	Muscle cannot hold the joint against resistance, but moves the joint fully against gravity
3 (II)	Muscle moves the joint against gravity, but not through full mechanical range of motion
2 (I+)	Muscle moves the joint when gravity is eliminated
1 (I)	A flicker of movement is observed or felt in the muscle

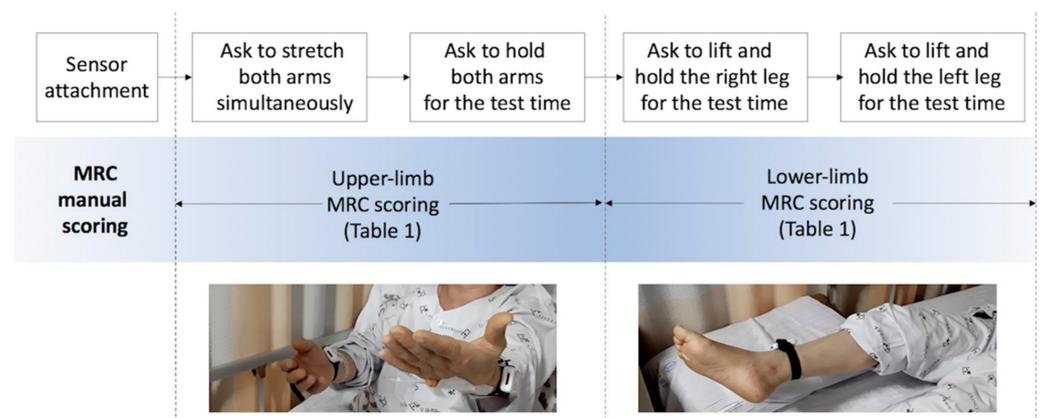
The protocol of this study asked acute stroke patients in the neurological intensive care unit to move and hold their limbs for assessing muscle strength, therefore the degree of difficulty in measuring and collecting data is relatively higher than those of studies for patients in rehabilitation or healthy subjects. Data of rare events or important objects are scarce in many studies, therefore, they need pre-processing to lessen the effect of the data capacity in model construction. This skewed distribution usually suffers from a lower performance than the dichotomous classification [22], and it invokes the problem of imbalanced data between classes [23,24]. Most methods are based on the model adjustment and data balancing approaches that have been popularly utilized in applications [25,26]. According to the review research [24], the hybridization of ensemble methods using sampling and cost-sensitive approaches has proven to be an effective solution to the classification with skewed data distribution. Tanha et al. [27] reviewed methods for multi-class imbalanced data classification as data-level methods, algorithm-level methods, and hybrid methods. For the agreement and reliability analysis, we compared the manual and machine learning

scaling by investigating the percent agreement, Bland-Altman plot with level of agreement (LoA) and Krippendorff's alpha for multiclass ordinal classification of MRC grading.

## 2. Materials and Methods

### 2.1. Participants and Study Protocol

We assessed proximal weakness among patients with acute stroke in a neurological intensive care unit on the day they were conscious enough to cooperate. Patients completed drift test trials that measure unintentional drift and oscillation of limbs caused by neurological deficits as shown in Figure 1. The graders asked patients to move and hold their limbs for the test time and scored MRC scales for each limb while inertial sensors objectively measure the movement.



**Figure 1.** Protocol of proximal-weakness assessment. MRC, Medical Research Council.

To estimate the symptoms of patients in critical condition, we assessed muscle strength according to the criteria of Table 1 shortly after their admission to a neurological intensive care unit.

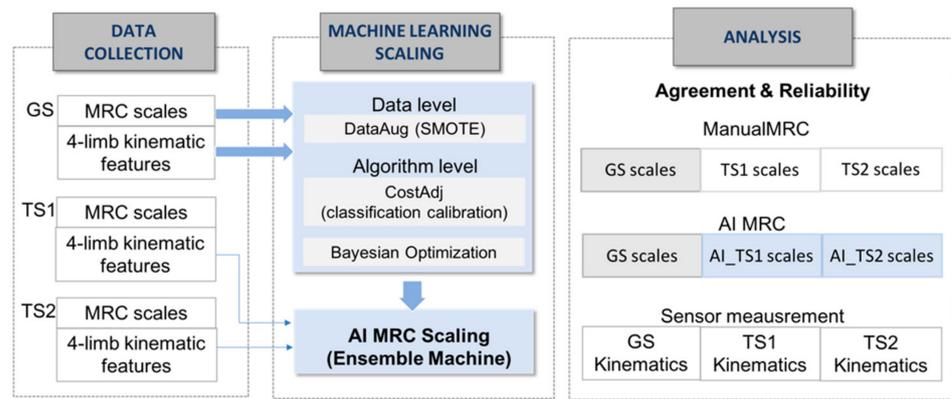
We collected 144 sets of inertial sensing data and their MRC gradings and prepared 600 synthetic data sets for training machine learning models. The participants' ages ranged from 38 to 86 years, with a mean of 65.4 ( $\pm 16.02$ ) years. The test protocol of manual MRC scoring and objective measurement by sensors attached to the arms and legs of patients are depicted in Figure 1.

For the reliability analysis tests, a neurologist who was familiar with patient assessment observed and coded the MRC scales as the gold standard (GS) labels. The observation window was 20 s to monitor the unintended movement of the limbs. Thereafter, two medical students performed additional tests (TS1 and TS2) for each patient as the assessment by non-experts. The decision-making criteria of the MRC scales were provided to the non-experts as Table 1 in advance of the identical test. The detailed protocol of the drift test for stroke patients is described in the works of [20,28].

The collected data were processed to extract kinematic features representing unintended drift and oscillation caused by neurological deficits. The feature set included the mean drift (MeanDrift), maximum drift (MaxDrift), and accumulated oscillation (SumOsc) during the observation window. The demographic features included gender and age. Three graders' tests were performed within 24 h. This study was approved by the Severance Hospital Institutional Review Board, and informed consent was obtained from all participants.

### 2.2. Automated Proximal Weakness Scaling

The data collection in Figure 2 shows the extraction of kinematic features from three graders' test trials for the machine learning MRC scaling. MRC scaling requires an ordinal classification with skewed distribution in multiple classes.



**Figure 2.** Process of data collection, machine learning-based scaling system and analysis. GS, gold standard; TS1, tester1, TS2, tester2, SMOTE, synthetic minority over-sampling technique.

To enhance the performance of MRC scaling with insufficient ordinal-imbalanced data, we adopted a hybrid approach of data-level and algorithm-level methods, as shown in the machine learning scaling part of Figure 2; We used synthetic minority over-sampling technique as the data-level method for constructing 600 data sets with class balancing for training the model, adjusting the mislabeling cost for reflecting distances between predicted and actual classes [27]. The data-level method addresses the disparity in data and adds new minority class instances to a training dataset by finding the k-nearest neighbors of a minority class instances and extrapolating between the original instances and its neighbors to create new instances in each iteration. For the training data set  $T = \{T_1, \dots, T_M\}$  with a skewed distribution in  $M$  classes, the algorithm produced a training instance set  $SB$  with a boosting factor  $SBF_i$  for the balanced  $N$  training instances as follows:

$$SB = \{SB_1, \dots, SB_M\}$$

$$SBF_i = (N/M - n(T_i))/n(T_i), \tag{1}$$

$$\text{and } n(SB_i) = n(T_i) \cdot SBF_i,$$

where  $i = \{1, \dots, M\}$  and  $n(T_i)$  represents the number of instances in the  $i$ th class. We performed the construction of synthetic data set based on the original data  $T$ , and we trained the ensemble machine using the shuffled data of all the original and synthetic data of  $SBT$  ( $SBT_i = SB_i \cup T_i$ ).

Given  $SBT$  with  $N$  instances with ordered classes, the training algorithm should not only maximize the classification accuracy but also minimize the distances between the actual and predicted classes [22,23]. For this problem, machine learning MRC scaling should adjust the cost matrix, which is expressed in terms of average misclassification costs for the multiple classes. Cost adjustment should express relative and unequal distances between classes to give more penalties for the misclassification that is far from the actual classes. The cost matrix  $C_{M \times M}$  is composed of  $C_{ij}$ , which denotes the penalty that misclassifies a class  $j$  instance into class  $i$ . In the linear-weight cost matrix, the cost weights between classes are adjusted as follows:

$$c_{i,j} = |j - i|, \tag{2}$$

for  $i = 1, \dots, M$ , and  $j = 1, \dots, M$ .

If the classification was not sufficiently improved with data-level methods, the misclassification weight of  $C_{M \times M}$  should be corresponding to an imbalance factor as follows [24]:

$$c_{i,j} = \frac{\sum_{i \neq j}^M n(SBT_i)}{n(SBT_i)} |j - i|, \tag{3}$$

for  $i = 1, \dots, M$ , and  $j = 1, \dots, M$ .

In this study, we adopted an ensemble machine learning and cost adjustment with linear weights for the MRC scales.

After sampling and formulating the cost matrix, we applied the Bayesian optimization algorithm selecting models in ensembles, which attempts to minimize a scalar objective function while selecting and tuning machine learning models [25]. The Bayesian update modifies the Gaussian process model at each new evaluation of the objective function  $f(x)$ , and the acquisition function  $a(x)$  based on the Gaussian process model of  $f(x)$  is maximized to determine the next point  $x$  for evaluation. In this study, we updated  $f(x)$  for 50 iterations to select a model among candidate ensemble machines of Bagging, Adaboost, and RUSBoost.

### 2.3. Agreement and Reliability Analysis

Inter-rater agreement and reliability are fundamental to the evaluation of new approaches or methods in various fields [26]. In this study, we calculated the agreement and reliability indices of the kinematic features from measurement and MRC scaling, respectively, as shown in the analysis part of Figure 2. To demonstrate the objective measurement using sensors, we investigated the intraclass correlation coefficients (ICCs) and the Bland-Altman plots with mean-difference and LoA, which are used to compare two measurements of the same variable [27]. Bland-Altman analyses were performed to assess the absolute degree of differences between kinematic features across the entire range of features measured in GS, TS1, and TS2. In the calculation of the ICCs, the average measures and a two-way random model were adopted for each feature [29,30].

To demonstrate the agreement between the expert's MRC scaling (GS) and machine learning MRC scaling of the non-experts (AI\_TS1 and AI\_TS2), we investigated the percent agreement between machine learning and GS scaling (GS:AI\_TS1 and GS:AI\_TS2) [31]. The reliability of machine learning scaling of all the tests was demonstrated using Krippendorff's alpha [32]. Krippendorff's alpha is a chance-adjusted index that estimates the level of agreement between graders that can be expected to have occurred by chance [33]. It is known to be superior to other reliability indices because it is applicable to an unlimited number of observers and categories for binary, nominal, and ordinal assessments, and it is robust to missing data [34,35]. We calculated Krippendorff's alphas to evaluate the reliability of three observers of manual scaling (GS, TS1, and TS2) and machine learning scaling (GS, AI\_TS1, and AI\_TS2).

Machine learning, including data preparation, tests, and evaluation were performed using Matlab2020a (Mathworks, Natick, MA, USA) [36]. The agreement and reliability were analyzed using the AnnotationTask agreement module in the nltk 3.5 library [37,38].

## 3. Results

### 3.1. Sensor Measurement and Features

The collected 3-axis accelerometer signals were converted to the degree of unintentional drift, as shown in Figure 3. Next, for each patient, we collected the drift trajectories measured in GS, TS1, and TS2. The feature extraction process calculates MeanDrift, MaxDrift, and SumOsc from the drift trajectories and the feature values from 144 observations are shown in Figure 4.

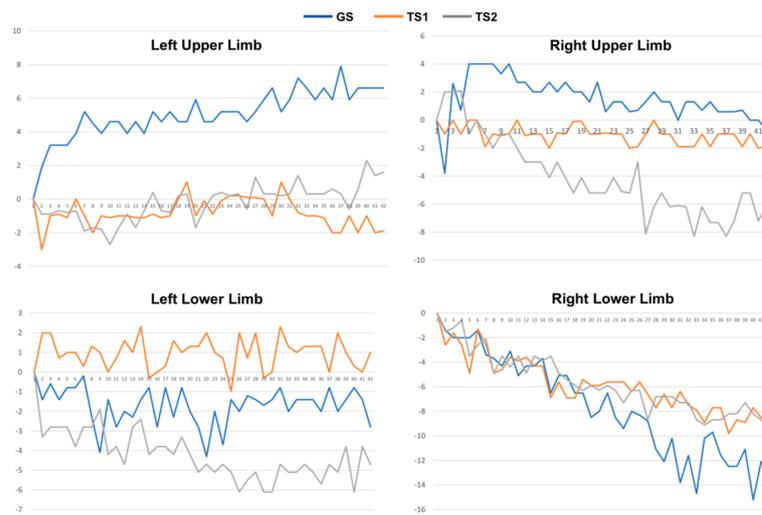


Figure 3. Drift trajectories of 4-limb movement of a patient measured in GS, TS1, and TS2.

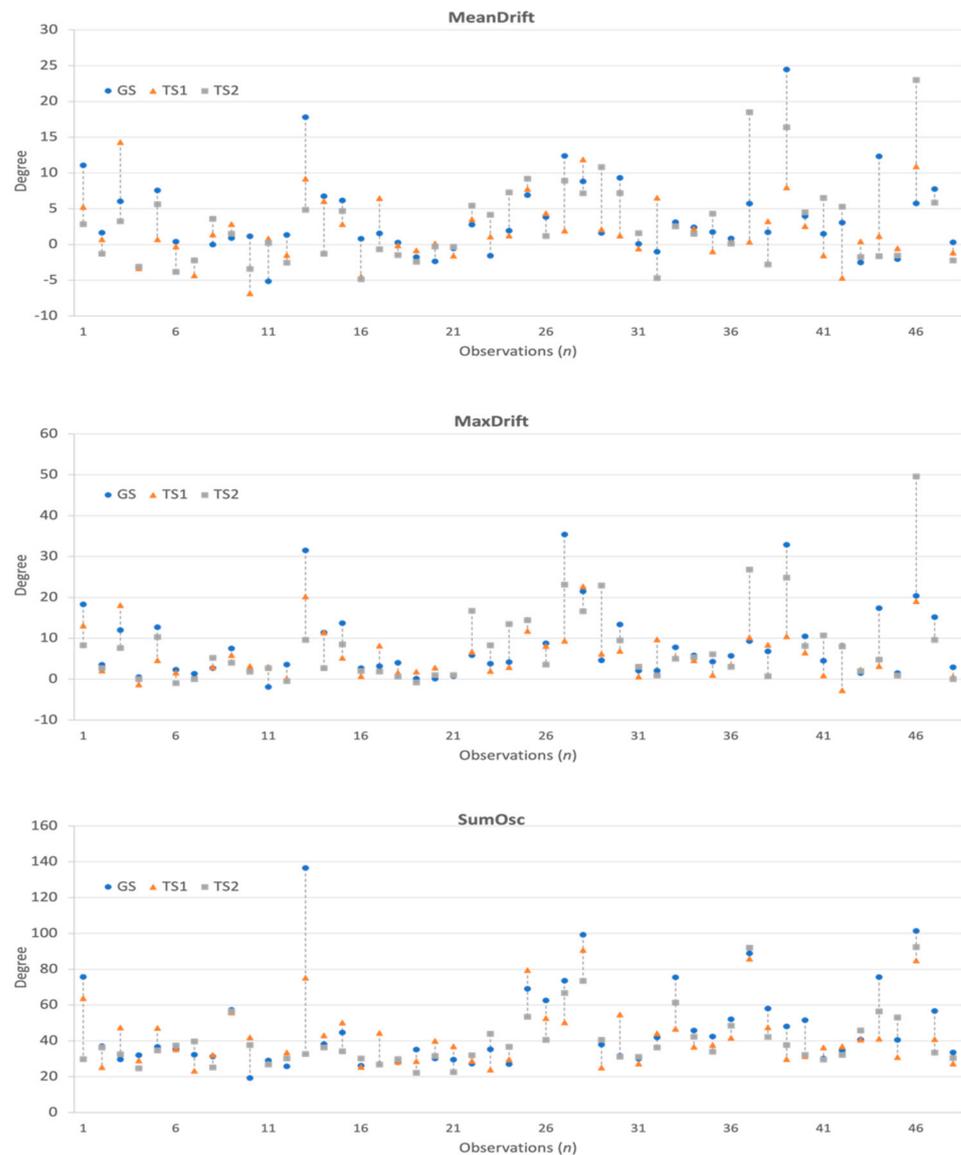
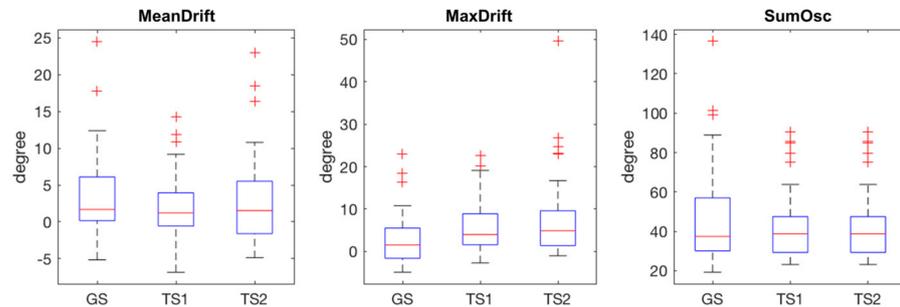


Figure 4. Kinematic features and differences observed in GS, TS1, and TS2.

The statistical plots of Figure 5 show the characteristics of the extracted kinematic features. MeanDrift was  $3.40 \pm 5.57$  (GS),  $2.03 \pm 4.34$  (TS1), and  $2.82 \pm 5.90$  (TS2). MaxDrift was  $8.39 \pm 8.64$  (GS),  $5.78 \pm 5.76$  (TS1), and  $7.61 \pm 9.38$  (TS2). SumOsc was  $46.82 \pm 23.93$  (GS),  $42.42 \pm 17.02$  (TS1), and  $40.09 \pm 15.67$  (TS2).



**Figure 5.** Statistical plots of measured features in GS, TS1, and TS2, red crosses are outliers.

In the agreement analysis, we investigated the agreement in kinematic features between two pairs of GS and TS1, and GS and TS2 using Bland-Altman plots, as shown in Figure 6. The mean values in the GS and TS1 plots of MeanDrift, MaxDrift and SumOsc were  $-1.37$ ,  $-2.617$ , and  $-4.404$ , respectively, which demonstrated a level of retention showing that the measured values of TS1 were slightly larger than those of GS. The LoA of confidence interval (CI) at 95% was  $(-10.96, 8.221)$  for MeanDrift,  $(-14.76, 9.528)$  for MaxDrift, and  $(-33.32, 24.5)$  for SumOsc, respectively. In the Bland-Altman plot of GS and TS2, the mean values were  $-0.582$  (MeanDrift),  $-0.7792$  (MaxDrift), and  $-3.825$  (SumOsc), and the LoA were  $(-17; 15.44)$  (MeanDrift),  $(-13.28; 11.33)$  (MaxDrift), and  $(-17.32; 14,33)$  (SumOsc), respectively.

The reliability of the kinematic features of the three testers was demonstrated using ICCs as shown in Table 2. The ICCs were 0.742 (MeanDrift), 0.798 (MaxDrift), and 0.850 (SumOsc), which are interpreted as being in the good reliability category (ICC, 0.75–0.90) [39].

**Table 2.** Intra-class correlation coefficients (ICCs) of kinematic features (2-way random).

Feature	ICC (2, k)	<i>p</i>	CI (95%)
MeanDrift	0.742	<0.001	(0.59–0.85)
MaxDrift	0.798	<0.001	(0.68–0.88)
SumOsc	0.850	<0.001	(0.76–0.91)

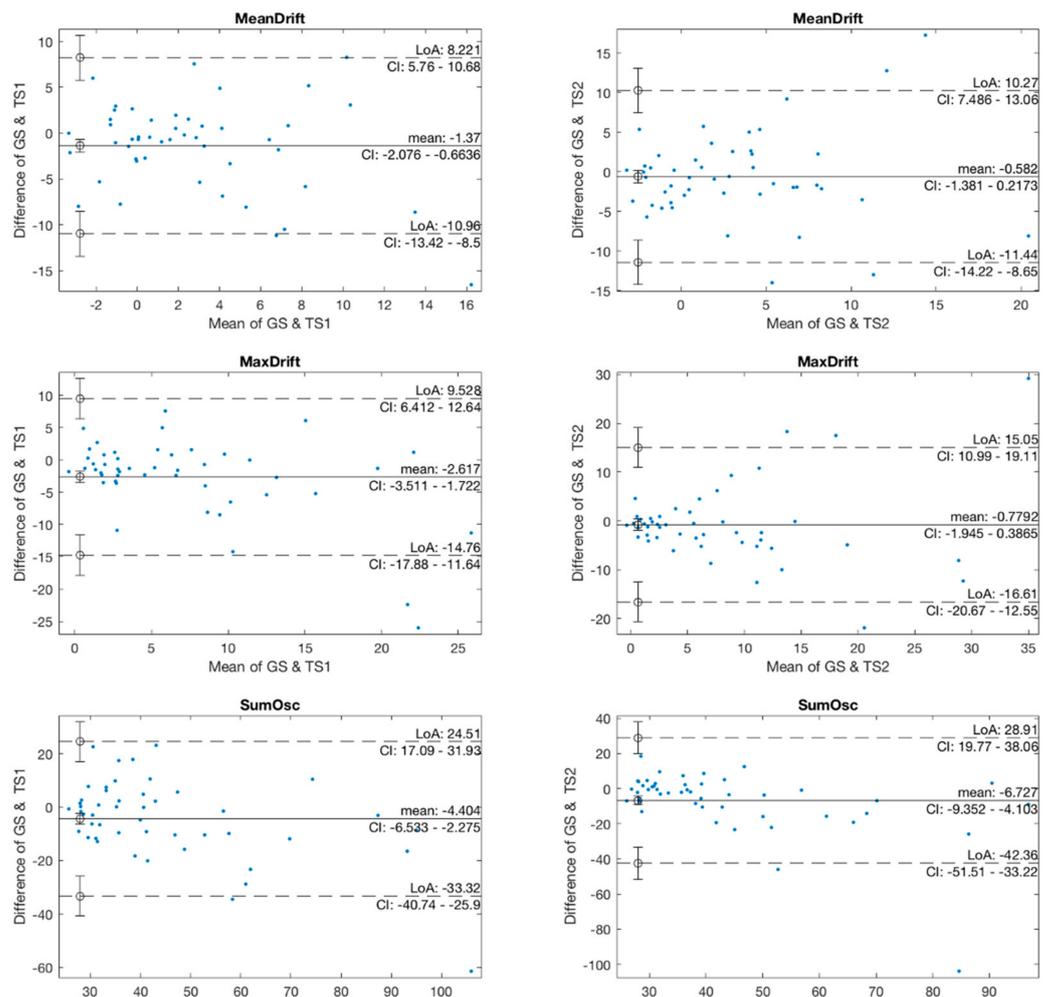


Figure 6. Bland-Altman plots of kinematic features.

### 3.2. Manual and Machine Learning Scaling

The distribution of MRC scales was skewed toward the highest MRC value, with 78 out of 144 observations assessed to MRC 9 (54.2%), as shown in Figure 7. MRC 8, 7, and 5 made up 27.1% (39/144), 15.2% (22/144), and 3.5% (5/144), respectively.

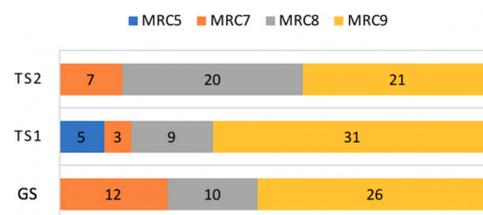
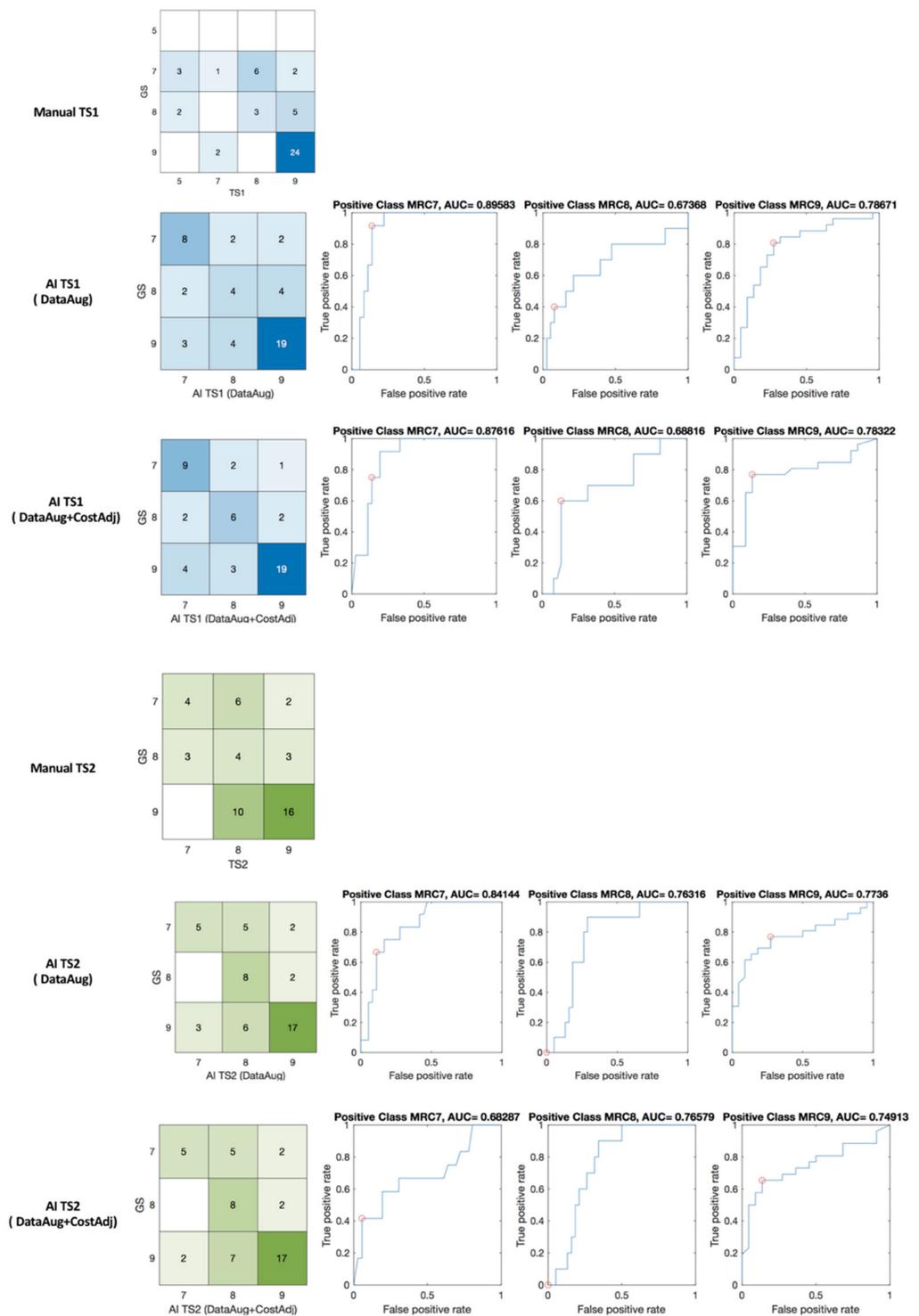


Figure 7. Composition of MRC scales.

The results of MRC scoring by machine learning are shown in Figure 8 with confusion matrices and receiver operating characteristic (ROC) curve and the area under the curve (AUC). The confusion matrix of expert’s scoring and manual scoring of TS1 (GS:TS1) and machine learning grading (GS:AI\_TS1) are shown in blue matrices in Figure 8. The percent agreement of between GS and TS1, and GS and AI\_TS1 was 0.583 and 0.708, respectively. In TS2, the identical patient group was assessed with a percent agreement of 0.5 for GS and TS2, and machine learning scaling enhanced the index to 0.708 for GS and AI\_TS2 as shown in green matrices in Figure 8.



**Figure 8.** Confusion matrices, ROC of manual (GS: TS1 and GS: TS2) and machine learning (GS: AI\_TS1 and GS: AI\_TS2) scaling. red circle—optimal operating point in the ROC curve.

In this study, we more focus on the agreement and reliability of proposed solution than other performance metrics. We analyzed Krippendorff’s alpha of manual and machine learning MRC scaling of the three testers, as shown in Table 3.

**Table 3.** The Krippendorff’s alpha (K-alpha) and Fleiss kappa of manual and machine learning scaling.

Methods	Metrics	GS-TS1	GS-TS2	GS-TS1-TS2
Manual	K-alpha	0.291	0.206	0.275
	Fleiss Kappa	0.300	0.218	0.285
Machine Learning (DataAug)	K-alpha	0.422	0.407	0.381
	Fleiss Kappa	0.416	0.413	0.383
Machine Learning (DataAug + CostAdj)	K-alpha	0.537	0.405	0.445
	Fleiss Kappa	0.534	0.414	0.448

Krippendorff’s alpha values are interpreted as fair in manual scaling with 0.291 for GS-TS1, 0.206 for GS-TS2, 0.275 for GS-TS1-TS2. The enhanced reliability with machine learning scaling was interpreted as moderate when it was hybridized through data augmentation (K-alpha 0.422 and 0.407 for GS-AI\_TS1 and GS\_AI\_TS2, respectively). The reliability of machine learning scaling with data augmentation and cost adjustment increased more to 0.537 for GS-AI\_TS1, 0.405 for GS-AI\_TS2, and 0.455 for GS-AI\_TS1-AI\_TS2. Consequently, the machine learning scaling with data augmentation and cost adjustment increased the reliability index of the three testers’ MRC scaling from 0.275 to 0.445.

#### 4. Discussion

##### 4.1. Agreement and Reliability of AI Model in Clinical Decision Making

Many promising results have accelerated the application of AI in medicine, which assists in making clinical decisions and saving personnel and time during the provision of care. Many studies argue that the first potential roles of medical AI involve triage situations or screening tools [1,40,41]. Specifically, AI prediction with patient monitoring is crucial in real-world environments, such as intensive care units, emergency rooms, and cardiac wards, where timeliness in clinical decision-making can be measured in seconds [6]. In experiments carried out in static environments or in simulations, many medical AI solutions demonstrate impressive performance. Therefore, they still need to be translated to work in realistic clinical settings [42] because the successful implementation of medical AI is premised on its reliable performance in real-world applications. Otherwise, the reliability problem results in the resistance to the application of medical AI, with the concerns being safety and performance [43,44]. To ensure the robust evaluation of performance in a real environment, agreement and reliability are significantly essential factors affecting the adoption of AI-based approaches in medicine. However, many researchers often focused solely on the accuracy of prediction in the use of AI, and they use the terms ‘reliability’ and ‘agreement’ interchangeably despite their technical distinction [26,45]. The selection of a proper evaluation method for reliability is significant in adopting a new approach [7,26,33]. As argued by Nili et al., the evaluation methods for reliability should be determined by considering the type of data, including nominal, ordinal, interval, and ratio data; Missing data of observation, the number of observers, and minimizing the effect of chance in agreement should also be considered [46].

In this pilot study, the Bland-Altman plots and ICCs were used to indicate the agreement and reliability of measured movement through sensors. As shown in the results, the ICCs of the kinematic features between GS, TS1 and TS2 were high (0.742–0.850). Nevertheless, the agreement of manual scaling to the GS was 0.5 for TS1 and 0.508 for TS2. This is caused by the difference in grading qualitative conditions into quantitative scales with cognitive bias in experience. The assessment of facial palsy is also one of the medical decisions that require the recognition of subtle differences among symptoms, as shown in the study on reliability in MRC scaling [21].

For the evaluation of reliability in multiple tests, a Krippendorff’s alpha was used to indicate the reliability between the tests of graders. A negative alpha indicates an inverse agreement less than that expected by chance. Our findings show that the reliability of

manual scaling was in the range of fair reliability (0.21–0.40), and the index of machine learning scaling (0.537) increased to moderate inter-rater reliability (0.41–0.60) using the hybrid approach of data augmentation and cost adjustment [32]. Although the reliability indices in this study are still lower than those in dichotomous classification, the enhancement was demonstrated through machine learning and data preprocessing techniques with respect to the equivocal triage categories; The percent agreement was also enhanced with an average improvement of 30.63%.

This study is to investigate the feasibility of autonomous grading as a consistent and reliable tool to estimate the assessment performed by an expert. To utilize the system as a standard tool, we need to extend the clinical trials to include data from multiple institutes in future works.

#### 4.2. Developing AI for Personalized Medicine with Disparity and Insufficiency of Data

The quality of AI-based decisions is determined by the amount of high-quality training data. However, the availability of data that is representative of the target patient population depends on the environment, urgency, patient conditions, and the availability of facilities for compiling data [6]. As addressed in many studies, AI research in the real environment has the long-tail problem in which data of rare and important events or objects are scarce [8,9,47]. The same problem appears with the availability of medical data, where data extracted from hospital information systems or health data from daily life monitoring are relatively easy to collect and provide sufficient data, however, the systems for monitoring and assessing acute symptoms, especially systems that require patients to actively participate in data collection, suffer from scarce data. The sensor data of limb movement for patients with acute stroke in the intensive care unit, which we measured in this study, were also hard to collect because stroke patients are asked to be attached sensors on the body and to actively stretch and hold limbs to follow protocols after thrombolysis. Consequently, collecting the movement data of emergency patients requires careful interruptions in the streamline of care, and is harder than fetching medical records from hospital information systems or public open data.

To overcome the limitation of small data set, various techniques have been applied and achieved promising results. Data augmentation and transfer learning are representative solutions for deep neural networks [10–13], and synthetic data and ensemble machine learnings also showed prominent performance in the studies with small data sets for rare diseases and acute symptoms [14–19].

In the systems for the treatment of acute symptoms, not only data collection but also the application of autonomous systems results in the inevitable intervention into the current streamline of treatment. Therefore, interventions involving medical AI solutions need to demonstrate their impact on health outcomes by undergoing rigorous and prospective evaluation, as addressed by He and Garcia [39]. The medical AI for acute diseases needs to be deliberately designed for application in real-world environments. Although AI for diagnosis or automatic detection in medical images have achieved promising results [48–51], the application of AI in the treatment of acute diseases should meet the requirements not only to support experts, but to also ensure the timeliness of decisions that face obstacle involving the insertion of new systems in the pipeline of treatment. However, the need for a reliable and consistent assessment during emergencies to save time have been addressed, and the corresponding applications and protocols continuously have been developed [52,53].

## 5. Conclusions

We demonstrated that machine-learning scaling achieved substantial improvement in inter-rater reliability for assessing proximal weakness in clinical scores. The improved agreement in patient assessment between observers can reduce medical errors during decision-making, especially during communication in the streamline of treatment in which experts and non-experts with various roles of care are involved. In our analysis, non-expert assessment with objective measurement using sensors and machine-learning-based scoring

improved agreement and reliability. This can improve the grounded application of reliable AI in the streamlining of care.

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

**Funding:** This research was supported by a grant funded by the Ministry of Science and ICT (NRF-2020R1A2C1013152) of Korea.

**Institutional Review Board Statement:** The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Institutional Review Board of the Severance Hospital Institutional Review Board (420090312).

**Informed Consent Statement:** Informed consent was obtained from all subjects involved in the study.

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

## References

1. He, J.; Baxter, S.L.; Xu, J.; Xu, J.; Zhou, X.; Zhang, K. The practical implementation of artificial intelligence technologies in medicine. *Nat. Med.* **2019**, *25*, 30–36. [CrossRef]
2. Rajkomar, A.; Dean, J.; Kohane, I. Machine learning in medicine. *N. Eng. J. Med.* **2019**, *380*, 1347–1358. [CrossRef]
3. Hess, D.C.; Audebert, H.J. The history and future of telestroke. *Nat. Rev. Neurol.* **2013**, *9*, 340–350. [CrossRef] [PubMed]
4. Sukumaran, M.; Cantrell, D.R.; Ansari, S.A.; Huryley, M.; Shaibani, A.; Potts, M.B. Stroke patient workflow optimization. *Endovasc. Tod.* **2019**, *18*, 46–50.
5. Jiang, F.; Jiang, Y.; Zhi, H.; Dong, Y.; Li, H.; Ma, S.; Wang, Y.; Dong, Q.; Shen, H.; Wang, Y. Artificial intelligence in healthcare: Past, present and future. *Stroke Vasc. Neurol.* **2017**, *2*, 230–243. [CrossRef]
6. Yu, K.H.; Beam, A.L.; Kohane, I.S. Artificial intelligence in healthcare. *Nat. Biomed. Eng.* **2018**, *2*, 719–731. [CrossRef] [PubMed]
7. Patrick, J. How to Check the Reliability of Artificial Intelligence Solutions—Ensuring Client Expectations are Met. *Appl. Clin. Inform.* **2019**, *10*, 269–271. [CrossRef]
8. Shen, T.; Lee, A.; Shen, C.; Lin, C. The long tail and rare disease research: The impact of next-generation sequencing for rare Mendelian disorders. *Genet. Res.* **2015**, *97*, e15. [CrossRef] [PubMed]
9. Winata, G.I.; Wang, G.; Xiong, C.; Hoi, S. Adapt-and-Adjust: Overcoming the Long-Tail Problem of Multilingual Speech Recognition. In Proceedings of the Interspeech 2021, Brno, Czech Republic, 30 August–3 September 2021. [CrossRef]
10. Li, J.; Qiu, L.; Tang, B.; Chen, D.; Zhao, D.; Yan, R. Insufficient Data Can Also Rock! Learning to Converse Using Smaller Data with Augmentation. *Proc. Conf. AAAI Artif. Intell.* **2019**, *33*, 6698–6705. [CrossRef]
11. Ayan, E.; Unver, H.M. Data augmentation importance for classification of skin lesions via deep learning. In Proceedings of the 2018 Electric Electronics, Computer Science, Biomedical Engineering's Meeting (EBBT), Istanbul, Turkey, 18–19 April 2018. [CrossRef]
12. Hagos, T.M.; Kant, S. Transfer learning based detection of diabetic retinopathy from small dataset. *arXiv* **2019**, arXiv:1905.07203. Available online: <https://arxiv.org/abs/1905.07203> (accessed on 10 October 2021).
13. Ravishankar, H.; Sudhakar, P.; Venkataramani, R.; Thiruvengadam, S.; Annangi, P.; Babu, N.; Vaidya, V. Understanding the Mechanisms of Deep Transfer Learning for Medical Images. In *Deep Learning and Data Labeling for Medical Applications. DLMIA 2016, LABELS 2016. Lecture Notes in Computer Science*; Carneiro, G., Mateus, D., Peter, L., Bradley, A., Tavares, J.M.R.S., Belagiannis, V., Papa, J.P., Nascimento, J.C., Loog, M., Lu, Z., et al., Eds.; Springer: Cham, Switzerland, 2016; Volume 10008, pp. 188–196. [CrossRef]
14. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
15. Bunkhumpornpat, C.; Sinapiromsaran, K.; Lursinsap, C. Safe-Level-SMOTE: Safe-Level-Synthetic Minority Over-Sampling TEchnique for Handling the Class Imbalanced Problem. In *Advances in Knowledge Discovery and Data Mining. PAKDD 2009. Lecture Notes in Computer Science*; Theeramunkong, T., Kijssirikul, B., Cercone, N., Ho, T.B., Eds.; Springer: Berlin/Heidelberg, Germany, 2009; Volume 5476, pp. 475–482. [CrossRef]
16. Li, D.; Liu, J.; Liu, J. NNI-SMOTE-XGBoost: A Novel Small Sample Analysis Method for Properties Prediction of Polymer Materials. *Macromol. Theory Simul.* **2021**, *30*, 2100010. [CrossRef]
17. Zhang, Y.; Jin, X. An automatic construction and organization strategy for ensemble learning on data streams. *ACM SIGMOD Rec.* **2006**, *35*, 28–33. [CrossRef]
18. Lei, H.; Li, H.; ElAzab, A.; Song, X.; Huang, Z.; Lei, B. Diagnosis of Parkinson's Disease in Genetic Cohort Patients via Stage-Wise Hierarchical Deep Polynomial Ensemble Learning. In *Predictive Intelligence in Medicine. PRIME 2019. Lecture Notes in Computer Science*; Rekik, I., Adeli, E., Park, S., Eds.; Springer: Cham, Switzerland, 2019; Volume 11843, pp. 142–150. [CrossRef]

19. Sammout, R.; Salah, B.K.; Ghedira, K.; Abdelhedi, R.; Kharrat, N.; Abdelhedi, R.; Kharrat, N. A Proposal of Clinical Decision Support System Using Ensemble Learning for Coronary Artery Disease Diagnosis. In *Wireless Mobile Communication and Healthcare*; Ye, J., O'Grady, M.J., Civitarese, G., Yordanova, K., Eds.; Springer International Publishing: New York, NY, USA, 2021. [CrossRef]
20. Park, E.; Lee, K.; Han, T.; Nam, H.S. Automatic Grading of Stroke Symptoms for Rapid Assessment Using Optimized Machine Learning and 4-Limb Kinematics: Clinical Validation Study. *J. Med. Internet Res.* **2020**, *22*, e20641. [CrossRef]
21. Paternostro-Sluga, T.; Grim-Stieger, M.; Posch, M.; Schuhfried, O.; Vacariu, G.; Mittermaier, C.; Bittner, C.; Fialka-Moser, V. Reliability and validity of the Medical Research Council (MRC) scale and a modified scale for testing muscle strength in patients with radial palsy. *J. Rehabil. Med.* **2008**, *40*, 665–671. [CrossRef] [PubMed]
22. Cardoso, J.S.; Sousa, R. Measuring the performance of ordinal classification. *Int. J. Pat. Rec. Arti. Int.* **2011**, *25*, 1173–1195. [CrossRef]
23. Kotsiantis, S.B.; Pintelas, P.E. A Cost Sensitive Technique for Ordinal Classification Problems. In *Methods and Applications of Artificial Intelligence. SETN 2004. Lecture Notes in Computer Science*; Vouros, G.A., Panayiotopoulos, T., Eds.; Springer: Berlin/Heidelberg, Germany, 2004; Volume 3025, pp. 220–229. [CrossRef]
24. George, N.I.; Lu, T.-P.; Chang, C.-W. Cost-sensitive Performance Metric for Comparing Multiple Ordinal Classifiers. *Artif. Intell. Res.* **2015**, *5*, 135–143. [CrossRef]
25. Lévesque, J.C.; Gagné, C.; Sabourin, R. Bayesian hyperparameter optimization for ensemble learning. In Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence, New York, NY, USA, 25–29 June 2016.
26. Chaturvedi, S.; Shweta, R. Evaluation of Inter-Rater Agreement and Inter-Rater Reliability for Observational Data: An Overview of Concepts and Methods. *J. Ind. Acad. Appl. Psych.* **2015**, *41*, 20–27.
27. Altman, D.G.; Bland, J.M. Measurement in Medicine: The Analysis of Method Comparison Studies. *J. R. Stat. Soc. Ser. D (Stat.)* **1983**, *32*, 307–317. [CrossRef]
28. Darcy, P.; Moughty, A.M. Pronator drift. *N. Engl. J. Med.* **2013**, *369*, e20. [CrossRef]
29. Bartko, J.J. The Intraclass Correlation Coefficient as a Measure of Reliability. *Psychol. Rep.* **1966**, *19*, 3–11. [CrossRef]
30. de Vet, H.C.W.; Terwee, C.B.; Mokkink, L.B.; Knol, D.L. *Measurement in Medicine: A Practical Guide*; Cambridge University Press: Cambridge, UK, 2011.
31. Krippendorff, K. Agreement and Information in the Reliability of Coding. *Commun. Methods Meas.* **2011**, *5*, 93–112. [CrossRef]
32. Krippendorff, K. *Content Analysis: An Introduction to Its Methodology*; Sage Publications: Los Angeles, CA, USA, 2018.
33. Gwet, K.L. *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement among Raters*; Advanced Analytics: Gaithersburg, MD, USA, 2014.
34. Artstein, R.; Poesio, M. Inter-Coder Agreement for Computational Linguistics. *Comput. Linguist.* **2008**, *34*, 555–596. [CrossRef]
35. Allen, M. *The SAGE Encyclopedia of Communication Research Methods*; Sage Publications: New York, NY, USA, 2017.
36. *Matlab, R2020*; Mathworks: Natick, MA, USA, 2020.
37. NLTK. NLTK 3.5 Documentation, Inter-Coder Agreement for Computational Linguistics. Implementations of Inter-Annotator Agreement Coefficients Surveyed by Artstein and Poesio (2007), Inter-Coder Agreement for Computational Linguistics. Available online: <http://www.nltk.org/api/nltk.metrics.html#module-nltk.metrics.agreement> (accessed on 2 June 2021).
38. Vallat, R. Pingouin: Statistics in Python. *J. Open Source Soft.* **2018**, *3*, 1026. [CrossRef]
39. Koo, T.K.; Li, M.Y. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *J. Chiropr. Med.* **2016**, *15*, 155–163. [CrossRef] [PubMed]
40. Levin, S.; Toerper, M.; Hamrock, E.; Hinson, J.S.; Barnes, S.; Gardner, H.; Dugas, A.; Linton, B.; Kirsch, T.; Kelen, G. Machine-Learning-Based Electronic Triage More Accurately Differentiates Patients With Respect to Clinical Outcomes Compared With the Emergency Severity Index. *Ann. Emerg. Med.* **2018**, *71*, 565–574. [CrossRef] [PubMed]
41. Hong, W.S.; Haimovich, A.D.; Taylor, R.A. Predicting hospital admission at emergency department triage using machine learning. *PLoS ONE* **2018**, *13*, e0201016. [CrossRef]
42. Mateen, B.A.; Liley, J.; Denniston, A.K.; Holmes, C.C.; Vollmer, S.J. Improving the quality of machine learning in health applications and clinical research. *Nat. Mach. Intell.* **2020**, *2*, 554–556. [CrossRef]
43. Longoni, C.; Bonezzi, A.; Morewedge, C.K. Resistance to Medical Artificial Intelligence. *J. Consum. Res.* **2019**, *46*, 629–650. [CrossRef]
44. Fraser, H.; Coiera, E.; Wong, D. Safety of patient-facing digital symptom checkers. *Lancet* **2018**, *392*, 2263–2264. [CrossRef]
45. de Vet, H.C.W.; Terwee, C.B.; Knol, D.L.; Bouter, L.M. When to use agreement versus reliability measures. *J. Clin. Epidemiol.* **2006**, *59*, 1033–1039. [CrossRef]
46. Nili, A.; Tate, M.; Barros, A. A critical analysis of inter-coder reliability methods in information systems research. In Proceedings of the 28th Australasian Conference on Information Systems, Tasmania, Australia, 5–6 December 2017.
47. Zang, Y.; Huang, C.; Loy, C.C. FASA: Feature Augmentation and Sampling Adaptation for Long-Tailed Instance Segmentation. *arXiv* **2021**, arXiv:210212867. Available online: <https://arxiv.org/abs/2102.12867> (accessed on 20 June 2021).
48. Armstrong, S. The apps attempting to transfer NHS 111 online. *BMJ* **2018**, *360*, k156. [CrossRef]
49. Bakator, M.; Radosav, D. Deep Learning and Medical Diagnosis: A Review of Literature. *Multimodal Technol. Interact.* **2018**, *2*, 47. [CrossRef]
50. Bates, M. Health Care Chatbots Are Here to Help. *IEEE Pulse* **2019**, *10*, 12–14. [CrossRef]

51. Wong, K.K.; Fortino, G.; Abbott, D. Deep learning-based cardiovascular image diagnosis: A promising challenge. *Futur. Gener. Comput. Syst.* **2019**, *110*, 802–811. [[CrossRef](#)]
52. Fassbender, K.; Balucani, C.; Walter, S.; Levine, S.R.; Haass, A.; Grotta, J. Streamlining of prehospital stroke management: The golden hour. *Lancet Neurol.* **2013**, *12*, 585–596. [[CrossRef](#)]
53. Park, E.; Kim, J.H.; Nam, H.S.; Chang, H.-J.; Park, E. Requirement Analysis and Implementation of Smart Emergency Medical Services. *IEEE Access* **2018**, *6*, 42022–42029. [[CrossRef](#)]