mathematics

MDPI

Article

Transformer Models and Convolutional Networks with Different Activation Functions for Swallow Classification Using Depth Video Data

Derek Ka-Hei Lai [1,†], Ethan Shiu-Wang Cheng [2,†], Bryan Pak-Hei So [1], Ye-Jiao Mao [1], Sophia Ming-Yan Cheung [3], Daphne Sze Ki Cheung [4,5], Duo Wai-Chi Wong [1,*] and James Chung-Wai Cheung [1,5,*]

[1] Department of Biomedical Engineering, Faculty of Engineering, The Hong Kong Polytechnic University, Hong Kong 999077, China
[2] Department of Electronic and Information Engineering, Faculty of Engineering, The Hong Kong Polytechnic University, Hong Kong 999077, China
[3] Department of Mathematics, School of Science, The Hong Kong University of Science and Technology, Hong Kong 999077, China
[4] School of Nursing, The Hong Kong Polytechnic University, Hong Kong 999077, China
[5] Research Institute of Smart Ageing, The Hong Kong Polytechnic University, Hong Kong 999077, China
* Correspondence: duo.wong@polyu.edu.hk (D.W.-C.W.); james.chungwai.cheung@polyu.edu.hk (J.C.-W.C.); Tel.: +852-2766-7669 (D.W.-C.W.); +852-2766-7673 (J.C.-W.C.)
† These authors contributed equally to this work.

Abstract: Dysphagia is a common geriatric syndrome that might induce serious complications and death. Standard diagnostics using the Videofluoroscopic Swallowing Study (VFSS) or Fiberoptic Evaluation of Swallowing (FEES) are expensive and expose patients to risks, while bedside screening is subjective and might lack reliability. An affordable and accessible instrumented screening is necessary. This study aimed to evaluate the classification performance of Transformer models and convolutional networks in identifying swallowing and non-swallowing tasks through depth video data. Different activation functions (ReLU, LeakyReLU, GELU, ELU, SiLU, and GLU) were then evaluated on the best-performing model. Sixty-five healthy participants (n = 65) were invited to perform swallowing (eating a cracker and drinking water) and non-swallowing tasks (a deep breath and pronouncing vowels: "/eɪ/", "/iː/", "/aɪ/", "/oʊ/", "/uː/"). Swallowing and non-swallowing were classified by Transformer models (TimeSFormer, Video Vision Transformer (ViViT)), and convolutional neural networks (SlowFast, X3D, and R(2+1)D), respectively. In general, convolutional neural networks outperformed the Transformer models. X3D was the best model with good-to-excellent performance (F1-score: 0.920; adjusted F1-score: 0.885) in classifying swallowing and non-swallowing conditions. Moreover, X3D with its default activation function (ReLU) produced the best results, although LeakyReLU performed better in deep breathing and pronouncing "/aɪ/" tasks. Future studies shall consider collecting more data for pretraining and developing a hyperparameter tuning strategy for activation functions and the high dimensionality video data for Transformer models.

Keywords: dysphagia; aspiration pneumonia; computer-aided screening; gerontechnology; deep learning

MSC: 68T01; 68U10

Citation: Lai, D.K.-H.; Cheng, E.S.-W.; So, B.P.-H.; Mao, Y.-J.; Cheung, S.M.-Y.; Cheung, D.S.K.; Wong, D.W.-C.; Cheung, J.C.-W. Transformer Models and Convolutional Networks with Different Activation Functions for Swallow Classification Using Depth Video Data. *Mathematics* **2023**, *11*, 3081. https://doi.org/10.3390/math11143081

Academic Editors: Xujuan Zhou, Lemai Nguyen and Guohun Zhu

Received: 9 June 2023
Revised: 6 July 2023
Accepted: 10 July 2023
Published: 12 July 2023

1. Introduction

Eating is an essential part of everyone's life. However, older adults may have difficulty eating or swallowing because of sarcopenia, declining cognitive functions, tissue elasticity, and neuromuscular control of the neck [1,2], or other health conditions such as strokes, age-related neurological conditions, and gastroesophageal reflux [3,4]. Dysphagia, classified as a sign or symptom, is defined as the difficulty in swallowing [5], in which foods/liquids may obstruct the passage towards the stomach [6]. Individuals with dysphagia may have

problems with drinking, eating, controlling saliva, and taking medications. A quarter of the adult population manifested swallowing problems, while the prevalence of dysphagia in stroke, institutionalized dementia, and Parkinson's patients was 41%, 45%, and 60%, respectively [2,7,8].

Complications of dysphagia are major causes of mortality and morbidity in the elderly and include aspiration pneumonia, malnutrition, and dehydration [9]. Dysphagic individuals reported a mortality rate that was 1.7 times higher and spent approximately USD 6000 more in hospitalization expenses compared to the non-dysphagic group [1]. Moreover, the fear and anxiety of choking also severely impacted their quality of life and psychological wellbeing [10]. Over one-third of dysphagic older adults avoid eating because of their conditions [11]. In fact, up to 68% of dysphagic elderly people lived in nursing homes, and about one-third of them lived independently [12], which inherited a significant burden and risk to the healthcare system and society.

Screening or assessment is crucial to prompting immediate management and rehabilitative interventions to reduce complication risks. Clinically, fiber-optic endoscopic evaluation of swallowing (FEES) and the video-fluoroscopy swallowing study (VFSS) are standard methods for dysphagia screening [13]. The procedure of FEES involves passing the endoscopic instrument through the nose to observe the pharynx and larynx when the individual is swallowing saliva with and without food consistencies [13]. Similarly, VFSS evaluates the swallowing function with different food consistencies, but through fluoroscopy over the oral cavity, pharynx, and cervical esophagus [13]. There are some drawbacks to these two methods. FEES induces pain and discomfort, while topical anesthesia may be applied sometimes. The VFSS exposes patients to radiation hazards and contrast agents [13]. Moreover, FEES and the VFSS are expensive and require professionals to operate.

It is demanding to develop alternative bedside methods that are valid and reliable [14]. Non-instrumental bedside assessments relied heavily on experts or therapists to conduct anamnesis, morphodynamical, and gustative function evaluations [15], whereas other related tests, such as the 3-ounce water swallowing test [16] and cough reflex test [17], lacked sensitivity and predictive strength [18] despite being routinely carried out in nursing homes or care homes. The use of acoustic and accelerometric sensors has been one of the common approaches to analyze swallowing [19,20]. The accelerometer is positioned on the surface of the skin above the larynx, where muscle movements take place when an individual swallows [21]. On the other hand, through a microphone near the throat, the chewing and swallowing sounds could be collected and analyzed to determine the food consistencies and viscosities and thus the swallowing conditions. Hidden Markov or other deep learning models were used for signal processing and analysis [22–24]. However, the approach was subject to background noise and may require additional pre-processing and segmentation of the acoustic data [25]. For the piezoelectric sensors, they were in the form of necklaces or patches that were versatile and light. The sensors detected physical strain and movement, which were subsequently processed with deep learning models to recognize chewing and swallowing motions [26–28]. It might also be challenging to implement contact-based sensors for older adults, especially those with dementia [29].

Recently, noncontact optical-based approaches using infrared depth cameras have emerged and been adopted for different mobile health applications [30–33]. Specific to dysphagia, An et al. [34] developed a liquid viscosity estimation model using the built-in camera of the smartphone with a convolutional neural network (CNN). Some other researchers also attempted to estimate the swallowing time using a depth camera [35]. Another study focused on measuring laryngeal movement via depth images, modeled by a decision tree [36]. We believed that the infrared depth camera could analyze the swallowing movement of the throat and compromise privacy. With the advancement of deep learning models in computer vision, we anticipated that they could help identify swallowing and thus abnormalities of swallowing. While CNN was a class of models commonly used for image classification, recent studies demonstrated that another cutting-edge class of models,

Transformers, which lay upon the core of natural language processing (NLP) [37], could effectively model the spatiotemporal relationship of image data (i.e., video data) and thus improve the classification performance.

On the other hand, regardless of the kind of model, an activation function is often required after the linear transformation of each layer. It is essential to provide nonlinearity in order to facilitate the learning of complicated input–output interactions [38]. In more technical terms, activation functions turn the weighed sum of inputs into an output value and transmit it to the nodes of the next layer. During model training, the choice of activation function is often determined by compromising convergence, complexity, smooth gradient flow, and data preservation during model training [38]. A Rectified Linear Unit (ReLU) is one of the common activation functions utilized by renowned networks, including AlexNet, GoogleNet, ResNet, and MobileNets. Other more recent activation functions, such as Swish, Exponential Linear Unit (ELU), Gaussian Error Linear Unit (GELU), and Gated Linear Units (GLU) have garnered attention for being superior to ReLU in certain tasks, despite the fact that the majority of model developments still adhere to the well-established ReLU [39]. Contemporary CNN networks often incorporate residual blocks with a Rectified Linear Unit (ReLU) as the default activation function. Nevertheless, the developers of ResNet and their successors did not justify or evaluate the choice of activation functions.

To this end, the objective of this study was to evaluate the performance of deep learning models (CNNs and Transformers) in classifying swallowing events from infrared depth camera video data. For the model with the best performance, we would then analyze the activation function that may further enhance the performance. The goal was to select the appropriate model and activation functions for this application at the outset and to facilitate a full-scale study for deployment in the future. This work represented the initial step to pave the road towards affordable and accessible instrumented dysphagia screening.

2. Materials and Methods

2.1. Participant Recruitment

We recruited 65 healthy adults (28 males and 37 females) from the university campus. Inclusion criteria were adults with no prior swallowing deprivation or disorder and no operation history for the head or neck within three months. Exclusion criteria were adults with difficulties in communication due to consciousness disturbance and patients with a tracheotomy hole. The participants had a mean age of 43.2 years (SD: 17.7, range: 18 to 77), an average height of 164.6 cm (SD: 8.19 cm, range: 148 cm to 183 cm), and a weight of 62.9 kg (SD: 13.5 kg, range: 40 kg to 100 kg). The experiment was approved by the Institutional Review Board of the university (reference No.: HSEARS20210416005). Prior to the start of the experiment, all participants were provided with oral and written descriptions of the experimental procedures, and they signed an informed consent form indicating their understanding and agreement to participate.

2.2. System Setup

An infrared Red-Blue-Green (RGB) stereo-based depth camera (Realsense D435i, Intel Corp., Santa Clara, CA, USA) was positioned to capture the entire swallowing process using the RealSense viewer program. Some preliminary tests were previously carried out and it was determined that the camera should be oriented at a 45° angle from the horizontal plane and placed 30 cm away from the neck of the participant (Figure 1a) to acquire the lower face and neck regions. The depth image data were captured at a resolution of 640 × 480 pixels, a frame rate of 30 frames per second, and a pixel depth of 2 bytes per pixel (or 1 mm per depth unit). The data were transmitted and processed on a personal computer.

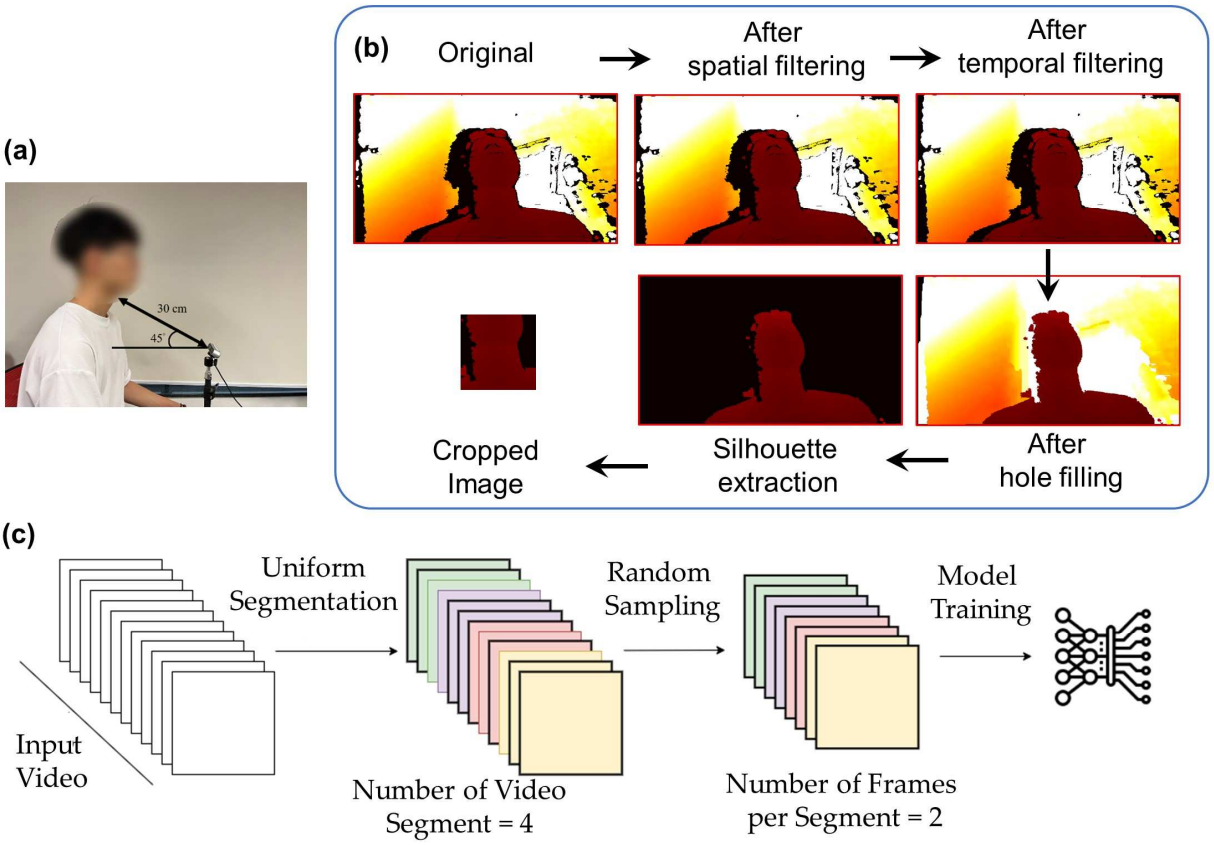

Figure 1. (**a**) System setup of the depth camera, (**b**) overall data processing framework, and (**c**) the temporal segment network [40].

2.3. Experimental Procedure

During the experiment, we recorded the lower face and neck (lip, mandible, and throat) motions for both non-swallowing and swallowing tasks. For the non-swallowing tasks, participants were asked to pronounce vowels, "/eɪ/", "/iː/", "/aɪ/", "/oʊ/", "/uː/" (i.e., /a/, /e/, /i/, /o/, /u/), in addition to performing a deep breath. After completing the non-swallowing motion tests, participants were asked to perform swallowing tasks. The first swallowing task was to eat (swallow) a cracker, approximately 45 mm × 45 mm in size. The second task was to drink (swallow) a cup of 10 mL of water. Participants were asked to consume as much as possible while taking bites/boluses at their comfortable size/volume.

The recording time depended on the actual duration of the tasks for each participant and trial. The swallowing time was approximately 1.0 to 1.5 s. Similarly, all tasks were repeated four times. Therefore, there was a total of 520 and 1560 sample data for all participants on the swallowing and non-swallowing tasks, respectively. The full dataset, with both swallowing and non-swallowing tasks, constituted 2080 sample data. The actual swallowing or non-swallowing tasks performed by the participants (i.e., ground truth) were manually labeled on each clip.

2.4. Data Processing

The overall data processing framework was shown in Figure 1b, which consisted of frame-by-frame filtering and video sampling. After data collection, we processed the data to improve the image (frame) quality and reduce noise. We followed the processing pipeline using RealSense SDK, as recommended by the official documents. For each frame, we first transformed the depth domain of the images to the disparity domain. Next, we applied spatial and temporal filters to denoise. The spatial filter was a one-dimensional edge-preserving spatial filter using a high-order domain transformation [41]. It aimed to

smooth the depth noise while preserving the edges. The temporal filter was similar to the spatial filter but suppressed artifacts across consecutive frames of the depth video sequence. The strength of smoothing was controlled by the parameters α and δ, for calculating the one-dimensional exponential moving average (EMA). It is defined by the recursive Equation (1):

$$S_t = \begin{cases} Y_1, & t = 1 \\ \alpha Y_t + (1-\alpha) S_{t-1}, & t > 1 \; and \; \Delta = |S_t - S_{t-1}| < \delta_{thresh} \\ Y_t, & t > 1 \; and \; \Delta = |S_t - S_{t-1}| > \delta_{thresh} \end{cases} \quad (1)$$

where coefficient α refers to the degree of weighting decrease, Y_t represents the latest recorded value for disparity or depth, and S_{t-1} is the value of the EMA at a previous time period, denoted as t.

When α is set to 1, no filtering is applied, while an α of zero results in an infinite history for the filtering. Additionally, the delta threshold (δ_{thresh}) was introduced. If the difference in depth values between neighboring pixels exceeds δ_{thresh}, α would be temporarily reset to one, which disables the filtering. In other words, if an edge is detected, the smoothing function is temporarily turned off. However, this may result in artifacts, depending on the direction of the edge traversed (i.e., right-to-left or left-to-right). To mitigate this, two bi-directional passes would be employed in both the vertical and horizontal directions of the images.

In temporal filtering, the same EMA smoothing was employed in the time domain. Similar to the spatial filter, α was used to represent the extent of the temporal history that should be averaged. The advantage of this approach is that it allows fractional frames to be effectively averaged. By setting $\alpha = 1$, there would be no filtering, while α = 0 would increase the averaging effect and result in a smoother output, allowing fine-grained smoothing beyond simple discrete frame averaging. Moreover, it is also important to incorporate the delta threshold, δ_{thresh}, to reduce the temporal smoothing effects near edges and ensure that missing depth information is not included in the averaging. We applied RealSense SDK default values for α and δ_{thresh}, where $\alpha = 0.5$ and $\delta_{thresh} = 20$ for the spatial filter, and $\alpha = 0.4$ and $\delta_{thresh} = 20$ for the temporal filter. Since image reconstruction of the stereo depth camera is based on a triangulation technique, the noise would appear at a level correlated with the squared rate of the camera–subject distance. In this context, α and δ would need to be adjusted based on the camera–subject distance, such that over-smoothing of near-range data and under-smoothing of far-range data could be avoided. We adopted a simpler approach by transforming the data into disparity domains before applying the filter.

After the filtering process, the images (frames) were back-transformed to the depth domain. We applied the hole-filling filter (boundary fill from Realsense SDK) to gaps or missing regions in depth maps that might result from occlusions and reflections. Subsequently, we removed the image background by zeroing the data with depth values larger than 60 cm and segmenting the silhouette of the subject. The region of interest (ROI) was located by first identifying the centroid of the silhouette $(\overline{x}, \overline{y})$ based on the image moment, the weighted averages of the image pixels' values, which are defined in Equations (2) and (3).

$$Image\ moment = M_{ij} = \sum_x \sum_y x^i y^j I(x, y) \quad (2)$$

$$Centroid = (\overline{x}, \overline{y}) = \left(\frac{M_{10}}{M_{00}}, \frac{M_{01}}{M_{00}} \right) \quad (3)$$

where i and j constitute the order of the moment, and $I(x,y)$ represents the pixel value of row x and column y. The first-order moments M_{10} and M_{01} normalized by the zero-order moment M_{01} would yield the centroid of the silhouette $(\overline{x}, \overline{y})$ and crop out a 224 × 224 pixel region from $(\overline{x} - 112, \overline{y} - 168)$ to $(\overline{x} + 112, \overline{y} + 56)$.

This setting was assigned based on our pilot analysis to ensure that the throat, mandibular (jaw), and lip regions were covered. To avoid excessive memory and computational requirements associated with utilizing the entire sequence of video frames in training, frames were sampled from the video using the temporal segment network [40]. As shown in Figure 1c, depth video frames were sampled by dividing the entire footage into several snippets, followed by a random selection of frames from each snippet. In our case, we decided to divide the depth videos into four snippets and randomly sample two frames from each snippet, as determined by our pilot analysis. The approach could ensure that every part of the video was representative of the loaded frames, and the method would be flexible enough to accommodate arbitrary and varying video lengths [40]. The pseudocode of the process is illustrated in Algorithm A1.

2.5. Activation Functions

While ReLU was the default activation function for X3D, Slowfast, and R(2+1)D, GELU was utilized by ViViT and TimeSFormer. In this study, we tested five activation functions, including ReLU [42], LeakyReLU [43], GELU [44], ELU [45], a Sigmoid-weighted Linear Unit (SiLU) [46], and a Gated Linear Unit (GLU) [47], on the model with the best performance. For the best-performing activation function, we would further conduct hyperparameter tuning of the activation function. The formulations with an input to a neuron (x) for all activation functions are illustrated in Equations (4)–(9) and compared in Figure 2.

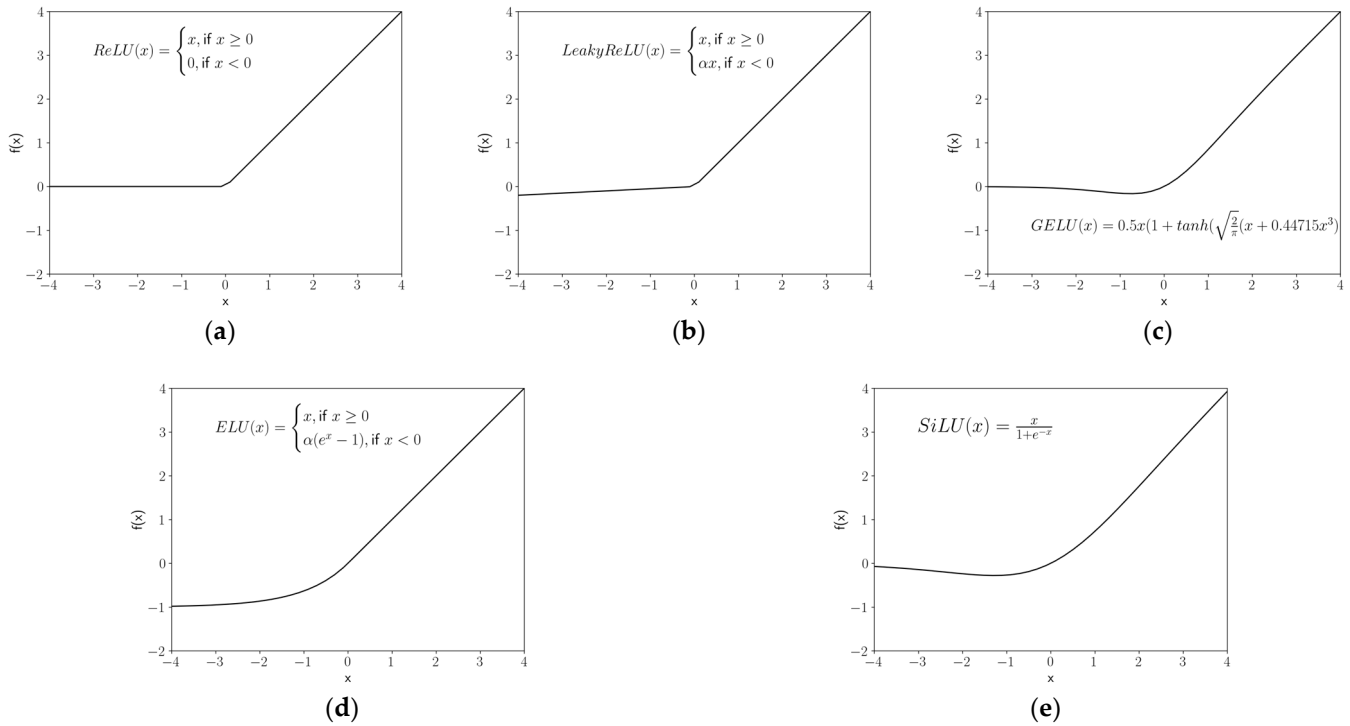

Figure 2. Activation function responses of (**a**) ReLU; (**b**) LeakyReLU; (**c**) GELU; (**d**) ELU; and (**e**) SiLU.

ReLU is a piecewise linear function that outputs the input directly if it is positive, and zero if it is negative, which is the default for most of networks due to its simplicity and high performance.

$$ReLU(x) = x^+ = \max(0, x) = \begin{cases} x & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases} \tag{4}$$

$$LeakyReLU(x) = \begin{cases} x & \text{if } x \geq 0 \\ \alpha \cdot x & \text{if } x < 0 \end{cases} \tag{5}$$

LeakyReLU (a particular kind of Parametric ReLU) is based on ReLU but returns a small negative value or slope (default α = 0.01) if the input is negative to account for the situation in which a large number of neuron inputs are negatives. Therefore, some information is "leaked" to prevent information loss (dead neurons) [48]. The ELU adopted a similar strategy but introduced exponential nonlinearity on negative inputs to mitigate the vanishing gradient problem (α default is one), whilst the SiLU utilized a Sigmoid function (σ). Vanishing gradient problems appear when lower layers of a network have gradients that are close to zero because higher layers are virtually saturated at -1 or 1 due to the *tanh* function [49].

$$ELU(x) = \begin{cases} x & \text{if } x > 0 \\ \alpha(exp(x) - 1) & \text{if } x \leq 0 \end{cases} \qquad \alpha > 0 \tag{6}$$

$$SiLU(x) = x \cdot \sigma(x) = \frac{x}{1 + e^{-x}} \tag{7}$$

GELU multiplies the input neuron by a random value from 0 to 1, calculated by the cumulative distribution function of the Gaussian distribution $\Phi(x)$. When the value of the input neuron is small, there is a large likelihood that the function's output would be zero (i.e., $Pr(X \leq x)$). GeLU is based on the assumption that the input neuron follows a normal distribution, especially after batch normalization.

$$\begin{gathered} GELU(x) = xPr(X \leq x) = x \cdot \Phi(x) \cong 0.5x(1 + tanh\left[\sqrt{\tfrac{2}{\pi}}(x + 0.044715x^3)\right] \\ X \sim N(0,1) \end{gathered} \tag{8}$$

$$\begin{gathered} GLU(X) = (X \cdot W + b) \otimes \sigma(X \cdot V + c) \\ X \in \mathbb{R}^{k \times m},\ W \in \mathbb{R}^{k \times m \times n}, V \in \mathbb{R}^{k \times m \times n}, b \in \mathbb{R}^n, c \in \mathbb{R}^n \end{gathered} \tag{9}$$

where k is the patch size, and m and n are the number of input and output feature maps, respectively.

The GLU is constructed by the linear project of the neuron input $(x \cdot W + b)$, multiplied by the Sigmoid gates $\sigma(x \cdot V + c)$. The element-wise multiplication of the gates on the input projection matrices could control the information passed on the hierarchy.

2.6. Model Training

Five cutting-edge deep learning models were trained for swallowing/non-swallowing classification, including two models of the Transformer class (TimeSFormer [50], Video Vision Transformer (ViViT) [51],) and three models of the CNN class (SlowFast [52], X3D [53] and R(2+1)D [54]). Model architectures are illustrated in Figure 3. Explanations of the models were provided in the discussion section. The models were trained using a computational unit with Intel Core i7 12700 and Nvidia RTX 4090. The total parameters and training time referenced to our computer are provided in Table 1.

Table 1. Total parameters and training time of models.

	X3D	**SlowFast**	**R(2+1)D**	**TimeSFormer**	**ViViT**
Total parameters	2.99 M	6.19 M	31.51 M	121.26 M	3.05 M
Training speed	35 s/epoch	51 s/epoch	70 s/epoch	94 s/epoch	37 s/epoch

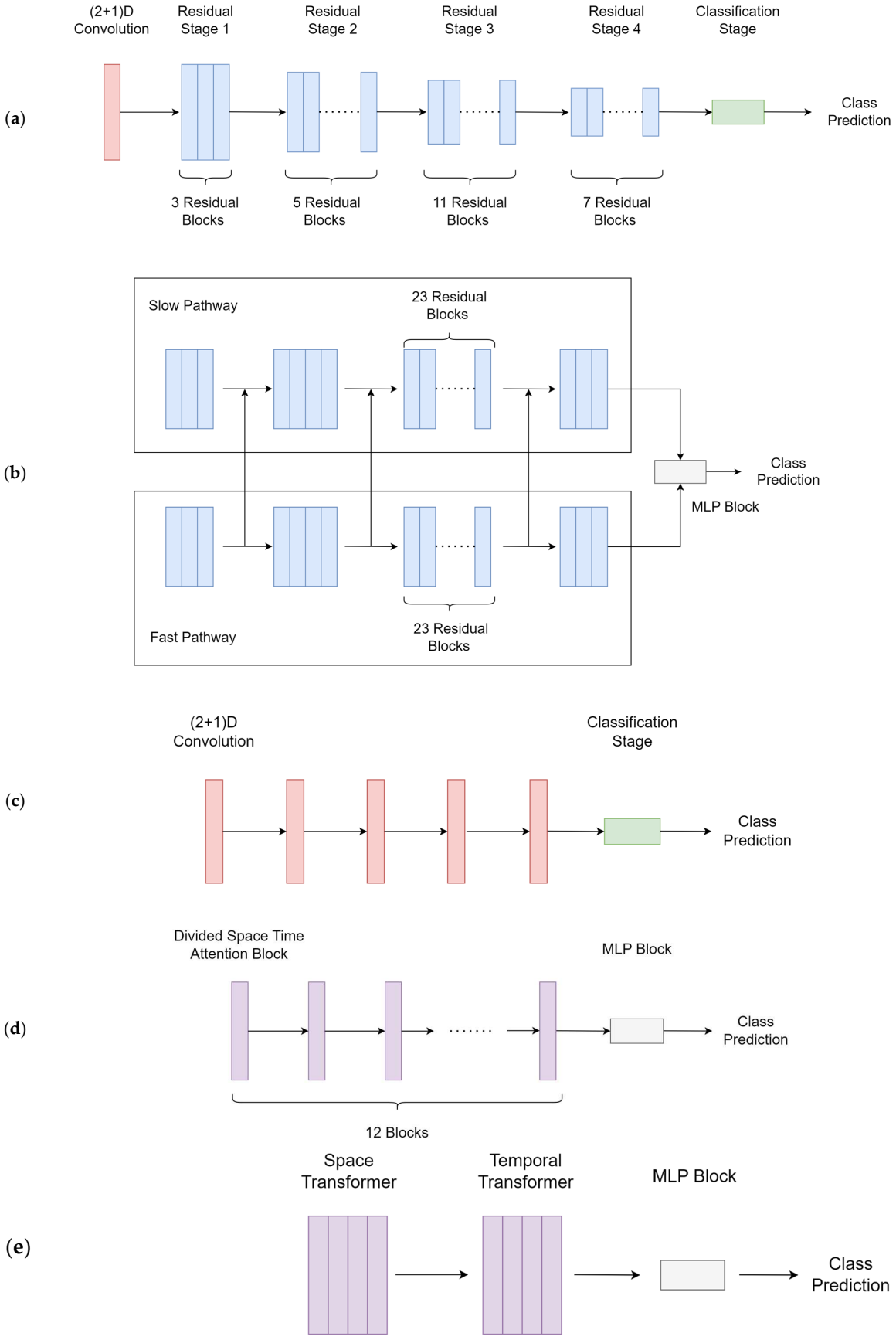

Figure 3. Model architectures for the five models: (**a**) X3D; (**b**) SlowFast; (**c**) R(2+1)D; (**d**) TimeSFormer; (**e**) ViViT [50–54].

We split the data into training, validation, and testing datasets at a ratio of 7:2:1. The models were trained using the training datasets. The performance of the model on the validation set during training was monitored to prevent overfitting. We performed 200 training epochs, with early stopping if the best performance did not improve in the next 20 iterations. The Adam optimizer was used for all models at a 0.0001 learning rate using cross entropy as the loss function. The pseudocode for the Adam optimizer is included in Algorithm A2.

The training batch size was set to four. We performed 100 training epochs, with early stopping if the best validation loss value did not improve in the next 20 iterations. For hyperparameters, TimeSFormer's attention mechanism was divided into space-time attention, where temporal attention and spatial attention were separately applied one after the other [50]. The patch size of ViViT was set to eight. The ResNet101 backbone was employed in the SlowFast model. All other unspecified hyperparameters were set to default, corresponding to each of the models. All processes were implemented using the PyTorch library [55].

2.7. Outcome Measures and Data Analysis (Model Evaluation)

Model evaluation was conducted by making predictions by inputting testing datasets onto the models. The primary analysis involved the overall performance in classifying the swallowing and non-swallowing tasks (i.e., coarse classification). Thereupon, two fine-grained classifications (subgroup analyses) on four classes and eight classes were performed. The former involved vowel pronunciation, deep breathing, eating, and drinking, while the latter involved the eight swallowing and non-swallowing tasks. On the best model, the same analysis would be conducted to compare the performance of various activation functions.

The F1-score was used as the primary outcome, which was believed to be less prone to an imbalanced class bias [56]. It is the harmonic mean of precision and recall, which is calculated by reciprocating the arithmetic mean of the reciprocals of precision and recall, as shown in Equations (10)–(12). Precision was defined as the proportion of positive predictions that were correct, while recall was the proportion of true positives that were correctly identified [57]. These outcome measures were derived from the confusion matrix (i.e., contingency table) that visualized the relationship between the predicted and actual (ground truth) class labels for the testing dataset. The cells of the table consisted of counts of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). The confusion matrix of the best-performing model was presented, in addition to the precision and recall for the other models and subgroup analyses. The counts were also used to analyze the source of the misclassification. Moreover, the Area under the receiver-operating characteristics curve (AUC) was used to evaluate the discrimination power of a binary classifier model. As a rule-of-thumb, we considered an F1-score over 0.70 as acceptable, 0.85 as good, and 0.9 as excellent.

The F1-score was calculated in Equations (10)–(12).

$$F1 = \frac{2}{\frac{1}{Pc} + \frac{1}{Rc}} = \frac{2 \times Pc \times Rc}{Pc + Rc} \tag{10}$$

$$Pc = \frac{TP}{TP + FP} \tag{11}$$

$$Rc = \frac{TP}{TP + FN} \tag{12}$$

where Pc is precision and Rc is recall. TP, FP, and FN are true positive, false positive, and false negative, respectively.

For model evaluation, an adjustment of the F1-score, precision, and recall were supplemented by bootstrapping (n = 26) on the major class to accommodate the imbalance in class size because of multiclass subgroup analyses (Algorithm A3). Confidence intervals of precision and recall for bootstrapping were estimated by their standard errors assuming a binomial distribution.

3. Results

3.1. Model Performance

For the coarse classification of swallowing and non-swallowing, X3D was the best-performing model with an average F1-score (adjusted F1-score) of 0.920 (0.885) (Figure 4a). The F1-scores (adjusted F1-scores) for detecting swallowing and non-swallowing were 0.878 (0.880) and 0.962 (0.889), respectively. CNNs were apparently better than Transformers. The other two CNNs, SlowFast and R(2+1)D, achieved average F1-scores (adjusted F1-scores) of 0.902 (0.884) and 0.866 (0.863), respectively, whereas the F1-scores (adjusted F1-scores) of TimeSFormer and ViViT were 0.648 (0.707) and 0.683 (0.766), as shown in Figure 4a). The adjusted F1-scores are shown in Table 2, calculated by the bootstrapped precision and recall.

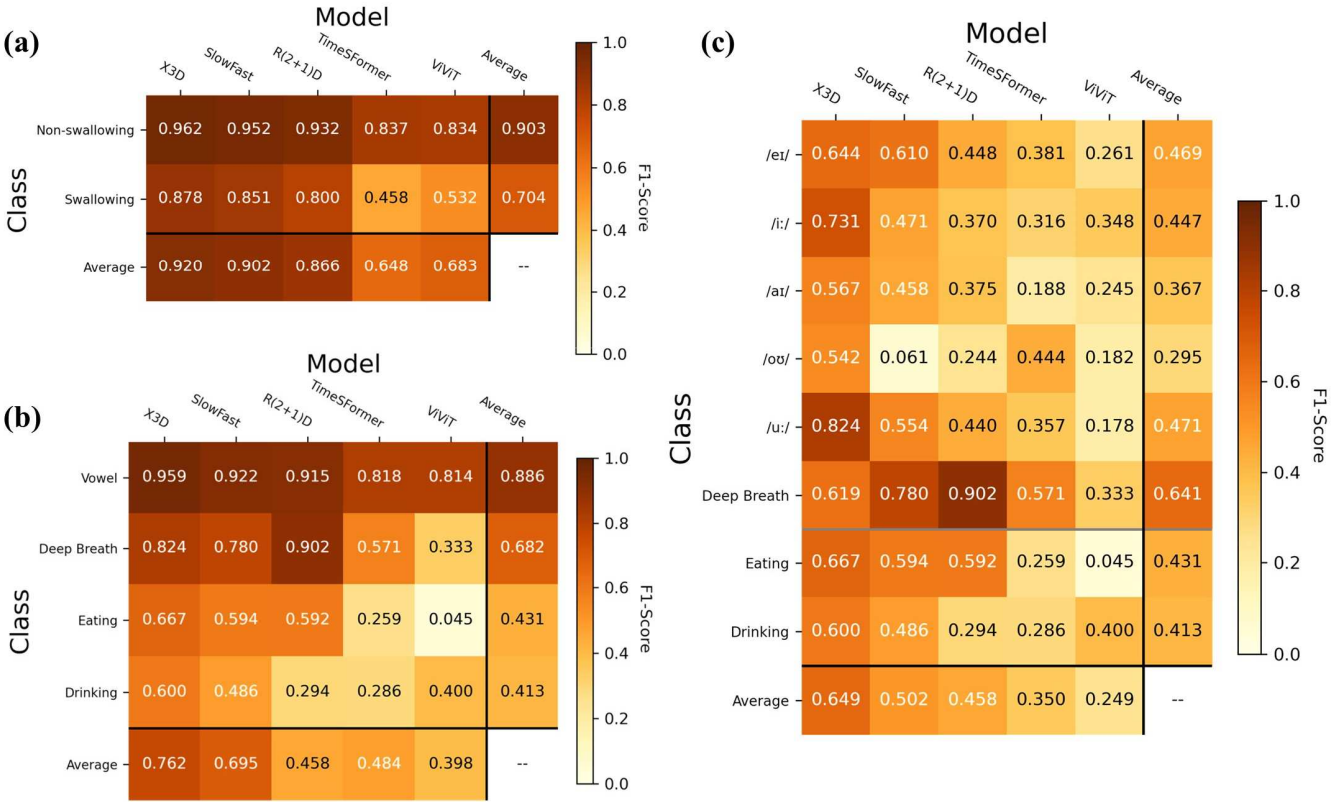

Figure 4. F1 scores for all models in (**a**) coarse classification; (**b**) 4-class fine-grained classification; (**c**) 8-class fine-grained classification; and confusion matrices of X3D, SlowFast, R(2+1)D, TimeSFormer, and ViViT.

Table 2. Adjusted F1-score of the five models, calculated by bootstrapping.

	X3D	SlowFast	R(2+1)D	TimeSFormer	ViViT
			Coarse Classification		
Swallowing	0.880	0.875	0.844	0.667	0.739
Non-swallowing	0.889	0.893	0.881	0.746	0.793
Coarse Average:	0.885	0.884	0.863	0.707	0.766

Table 2. *Cont.*

	X3D	SlowFast	R(2+1)D	TimeSFormer	ViViT
			Fine-grained Classification		
Eating	0.711	0.769	0.783	0.650	0.294
Drinking	0.732	0.556	0.364	0.476	0.412
Deep Breathing	0.894	0.939	0.939	0.894	0.488
Vowel Pronunciation	0.926	0.923	0.889	0.755	0.720
Pronouncing "/eɪ/"	0.809	0.809	0.727	0.564	0.462
Pronouncing "/iː/"	0.756	0.683	0.389	0.756	0.188
Pronouncing "/aɪ/"	0.800	0.564	0.353	0.333	0.242
Pronouncing "/oʊ/"	0.783	0.207	0.444	0.526	0.343
Pronouncing "/uː/"	0.894	0.863	0.585	0.714	0.194
4-class Average:	0.816	0.797	0.744	0.694	0.479
8-class Average:	0.797	0.674	0.573	0.614	0.328

As shown in Figure 5, fine-grained classification imposed additional challenges to the model prediction accuracy. X3D remained the best-performing model and produced average F1-scores (adjusted F1-scores) of 0.762 (0.816) and 0.649 (0.797), respectively, for the four-class and eight-class analyses (Figure 4b,c). Although the Transformers performed worse, their average F1-scores managed to get over the probability of random guess in four classes (0.250) and eight classes (0.125).

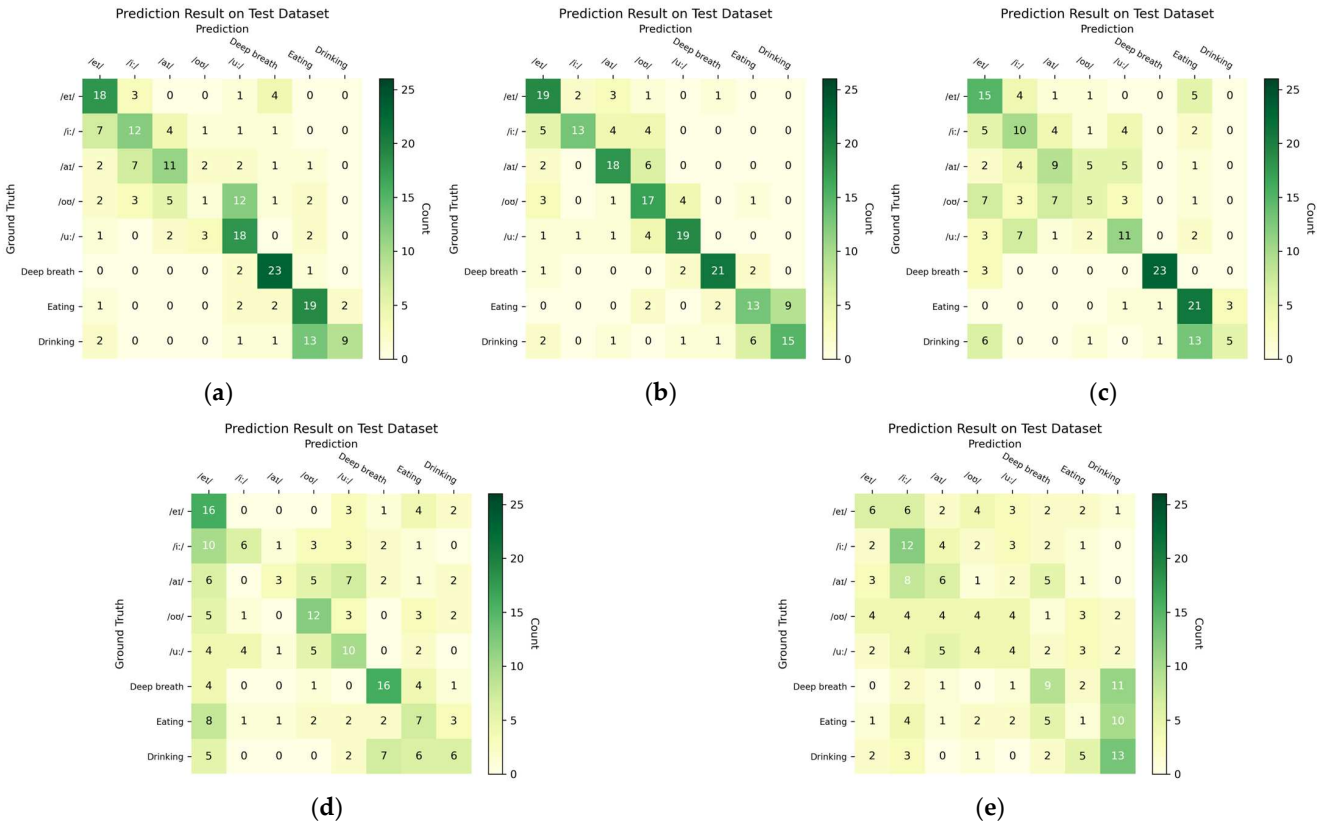

Figure 5. 8-class fine-grained classification; and confusion matrices of (**a**) X3D; (**b**) SlowFast; (**c**) R(2+1)D; (**d**) TimeSFormer; and (**e**) ViViT.

3.2. Task Prediction Performance

The prediction performance for non-swallowing was better than swallowing in coarse classification. The average F1-score across models for non-swallowing was 0.903, with a

range from 0.834 to 0.962 (adjusted F1-score: 0.840, range from 0.746 to 0.893), compared to that of swallowing, which was 0.704, with a range from 0.458 to 0.878 (adjusted F1-score: 0.801, range from 0.667 to 0.880). Among the eight swallowing and non-swallowing tasks, deep breath was the most distinctive, and the best model for this event, R(2+1)D attained an F1-score of 0.902. However, the highest adjusted F1-score for deep breath was 0.939, achieved by both SlowFast and R(2+1)D models. It was simpler to recognize vowel pronunciation from other tasks, but it was more difficult to pinpoint each individual vowel pronunciation. X3D attained an F1-score of 0.959 in classifying vowel pronunciation, but that of recognizing each vowel ranged from 0.542 to 0.824. The SlowFast model tended to misclassify "/oʊ/" as "/u:/" with a low F1-score of 0.061 for this event. On the other hand, it was also difficult to classify eating and drinking. For X3D, the F1-score (adjusted F1-score) to identify eating was 0.667 (0.711), while that of drinking was 0.600 (0.732). Figure 6 details the precision and recall for each model and task.

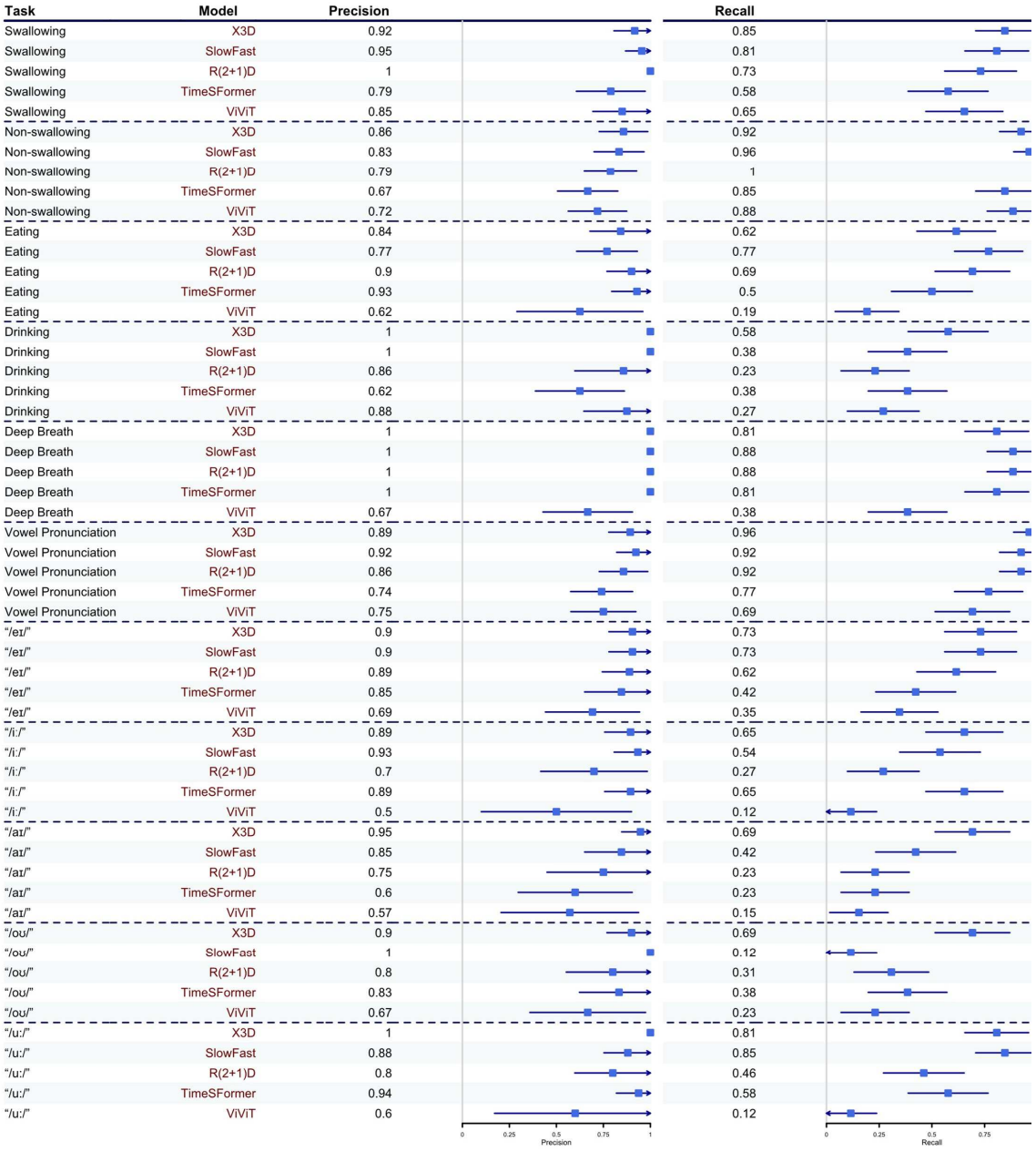

Figure 6. Estimated precision and recall of coarse and fine-grained classification by bootstrapping for all models and tasks.

3.3. Evaluation of Activation Functions on the Best Model

Evaluation of activation functions was performed on the X3D model. The default activation function, ReLU, produced the best overall performance (F1-score: 0.920), followed by LeakyReLU. Nevertheless, LeakyReLU, GELU, and ELU had higher F1-scores in recognizing non-swallowing events, which were 0.927, 0.911, and 0.899, respectively.

For fine-grained classification (Table 3), ReLU and LeakyReLU fared similarly, although LeakyReLU performed slightly better. Their F1-scores for four-class classification were 0.711 and 0.718, respectively, while those for eight-class classification were 0.649 and 0.656. The ReLU appeared to outperform the other functions in articulating the pronunciation of "/iː/" (F1-score: 0.731) and "/uː/" (F1-score: 0.824). However, the ReLU had the lowest performance (F1-score: 0.619) among the functions for identifying breathing, which was largely improved by using LeakyReLU (F1-score: 0.816).

Table 3. F1-score of X3D on different activation functions.

	ReLU (Default)	**LeakyReLU**	**GELU**	**ELU**	**GLU**	**SiLU**
			Coarse Classification			
Swallowing	0.962	0.807	0.681	0.673	0.906	0.925
Non-swallowing	0.878	0.927	0.911	0.899	0.763	0.750
Coarse Average:	0.920	0.867	0.796	0.786	0.835	0.838
			Fine-grained Classification			
Eating	0.667	0.551	0.418	0.407	0.488	0.593
Drinking	0.600	0.578	0.458	0.462	0.250	0.429
Deep Breathing	0.619	0.816	0.760	0.627	0.800	0.857
Vowel Pronunciation	0.959	0.925	0.895	0.839	0.895	0.924
Pronouncing "/eɪ/"	0.644	0.644	0.543	0.462	0.654	0.538
Pronouncing "/iː/"	0.731	0.596	0.519	0.545	0.596	0.378
Pronouncing "/aɪ/"	0.567	0.723	0.578	0.585	0.510	0.500
Pronouncing "/oʊ/"	0.542	0.591	0.474	0.500	0.511	0.341
Pronouncing "/uː/"	0.824	0.750	0.506	0.516	0.588	0.429
4-class Average:	0.711	0.718	0.633	0.584	0.608	0.701
8-class Average:	0.649	0.656	0.532	0.513	0.550	0.508

Subsequently, we further evaluated LeakyReLU with different α values (Table 4). α of Tan12° and Tan6° produced better performances. Their average F1-scores for eight-class classification were 0.697 and 0.691, respectively. In addition, the α of Tan6° showed higher performance in recognizing eating and drinking events (F1-scores of 0.692 and 0.652, respectively), whereas the α of Tan12° was superior in identifying non-swallowing events.

Table 4. Evaluation of hyperparameter, α, on the performance of LeakyReLU.

		Tan18°	**Tan15°**	**Tan12°**	**Tan9°**	**Tan6°**	**Tan3°**	**Tan0°**
F1	Pronouncing "/eɪ/"	0.585	0.585	0.667	0.548	0.543	0.677	0.644
	Pronouncing "/iː/"	0.577	0.510	0.750	0.578	0.605	0.681	0.731
	Pronouncing "/aɪ/"	0.711	0.627	0.846	0.621	0.698	0.650	0.567
	Pronouncing "/oʊ/"	0.667	0.625	0.708	0.486	0.625	0.679	0.542
	Pronouncing "/uː/"	0.742	0.632	0.691	0.698	0.808	0.727	0.824
	Deep breathing	0.939	0.826	0.830	0.894	0.902	0.816	0.619
	Eating	0.691	0.607	0.571	0.679	0.692	0.571	0.667
	Drinking	0.600	0.476	0.510	0.549	0.652	0.553	0.600
	8-class Average	0.689	0.611	0.697	0.632	0.691	0.669	0.649

Table 4. *Cont.*

		Tan18°	Tan15°	Tan12°	Tan9°	Tan6°	Tan3°	Tan0°
Precision	Pronouncing "/eɪ/"	0.487	0.487	0.588	0.472	0.400	0.583	0.576
	Pronouncing "/iː/"	0.577	0.520	0.818	0.684	0.765	0.762	0.731
	Pronouncing "/aɪ/"	0.842	0.640	0.846	0.562	0.882	0.929	0.500
	Pronouncing "/oʊ/"	0.727	0.682	0.773	0.818	0.682	0.667	0.591
	Pronouncing "/uː/"	0.639	0.581	0.655	0.595	0.808	0.690	0.840
	Deep breathing	1.000	0.950	0.815	1.000	0.920	0.870	0.813
	Eating	0.655	0.567	0.609	0.667	0.692	0.486	0.643
	Drinking	0.857	0.625	0.520	0.560	0.750	0.619	0.625
	8-class Average	0.723	0.632	0.703	0.670	0.737	0.701	0.665
Recall	Pronouncing "/eɪ/"	0.731	0.731	0.769	0.654	0.846	0.808	0.731
	Pronouncing "/iː/"	0.577	0.500	0.692	0.500	0.500	0.615	0.731
	Pronouncing "/aɪ/"	0.615	0.615	0.846	0.692	0.577	0.500	0.654
	Pronouncing "/oʊ/"	0.615	0.577	0.654	0.346	0.577	0.692	0.500
	Pronouncing "/uː/"	0.885	0.692	0.731	0.846	0.808	0.769	0.808
	Deep breathing	0.885	0.731	0.846	0.808	0.885	0.769	0.500
	Eating	0.731	0.654	0.538	0.692	0.692	0.692	0.692
	Drinking	0.462	0.385	0.500	0.538	0.577	0.500	0.577
	8-class Average	0.688	0.611	0.697	0.635	0.683	0.668	0.649

4. Discussion

The novelty of this research lies in the application of depth cameras, in addition to state-of-the-art deep learning techniques including CNNs and Transformer models, to analyze and classify swallowing and non-swallowing tasks, which paves the road towards accessible instrumental dysphagia screening. We believed that this may be one of the first works of its kind. Moreover, the swallow monitoring system could be expanded to evaluate patients with eating behavioral and malnutritional problems and to facilitate biofeedback training [58,59].

Five cutting-edge deep learning models were used and compared, including TimeSFormer, ViViT, SlowFast, X3D, and R(2+1)D. These models were specialized in leveraging both spatial and temporal information from video sequences to perform tasks such as action recognition, object detection, and video segmentation, while addressing various challenges unique to video analysis, such as the temporal variability and the need for efficient and scalable architectures. [60]. The two Transformer models differed from one another in the design of the attention scheme. TimeSformer embedded frame-level patches and learned spatiotemporal features by dividing temporal and spatial attention schemes within each block [50]. On the other hand, ViViT proposed multiple-head self-attention architectures that accounted for the factorization of spatial and temporal dimensions of the input [51].

For the CNNs, SlowFast consisted of a slow and a fast pathway processing the same input with different temporal resolutions. The slow pathway was a standard 3D CNN, while the fast pathway integrated a 2D CNN with a temporal down-sampling unit. The two pathways were joined with a Time-strided convolution (T-conv) [52]. X3D was built using a ResNet structure and the Fast pathway of the SlowFast model, along with degenerated (single frame) temporal input [53]. Moreover, the characteristics of R(2+1)D were the utilization of a 2D convolutional filter with a 1D temporal convolutional filter, governed by the hyperparameter related to the intermediate subspace between the spatial and temporal convolutions [54].

X3D was the best model in our study with good-to-excellent performance (F1-score: 0.920; adjusted F1-score: 0.885) in classifying swallowing and non-swallowing conditions despite the fact that the performance was just acceptable. The model focused on one data dimension at a time in building up the model blocks to accommodate the level of complexity, which might be appropriate and efficient for our occasion. For the other two CNNs, R(2+1)D manifested spatiotemporal representation through temporal convolutions, while the SlowFast model captured high-level semantics and spatiotemporal information through the slow and fast pathways. These approaches could be vulnerable to the predefined layer size and number of layers, which might require strategies for extensive hyperparameter optimization to arrest critical spatiotemporal features. The hyperparameter tuning process could be very time-consuming and demanding of computational power because of the higher dimensionality of video data, compared to those working on numeric and image data. On the other hand, our initial hypothesis was that the Transformers could outperform the CNNs because of their long-range capturing capacity and attention mechanism. Nevertheless, Transformers exhibited poor performance in our study because of the small dataset size. In fact, Transformers placed a very high demand on the size of the dataset [61]. We did not pre-train the Transformers because a large-depth video dataset was not available in the public domain.

The classification of the depth camera relied on manifested morphological motions of the lip (mouth), mandibular (jaw), and neck (throat) regions. Swallowing and non-swallowing could be easier to classify because of the discernible depth of the throat, with and without bolus. Although eating behaviors can be represented by "periodic" mandibular (jaw) activities (i.e., chewing) [62], our study found it difficult to discriminate between eating and drinking, probably due to their comparable lip and throat motions. Capturing hand movements might help distinguish the type of foods/liquids. On the other hand, while pronunciation could be recognized by lip movements, some vowels had subtle lip apertures and might vary depending on individuals' speaking habits or speaking countries [63]. This could be the reason for the low accuracy in the fine-grained classification of vowel pronunciation. Nevertheless, the success in recognizing talking (pronunciation), breathing, and eating/drinking might facilitate monitoring systems for sleep apnea and somniloquy.

Real-time and continuous extraction and identification of high-level spatial and temporal features were the challenges in this study. The experimental protocol itself might confound the data features. In particular, swallowing tasks generally had a shorter duration than non-swallowing tasks. We endeavored to apply the temporal segment network [40] to equalize the amount of information in the temporal domain to ensure that the model was analyzing the spatiotemporal features of the data instead of the length of the recording. Nevertheless, the approach might not account for the dynamic time wrapping issue [64]. For example, the variations on the start/stop instants of the recording and features might fail to "synchronize and align" the temporal features corresponding to each task. These would lead to bias during random sampling within the temporal segment network.

The optimal training/validation/testing ratio for machine learning was mostly empirical and lacked precise recommendations [65,66]. While Joseph [65] and Dubbs [66] suggested that the number of parameters and the size of the dataset could be used to estimate the splitting ratio for linear models and Ridge and Lasso regression, a general law for the splitting ratio, determined analytically or asymptotically for all models, has not yet been established [66]. A rule of thumb was to divide the data in an 80/20 ratio based on the Pareto principle, while some advised allocating 70% of data for model training and distributing the remaining data evenly for model validation and testing. Reducing the size of the training dataset, especially for small datasets, would increase the variance of the parameter estimates of the model, while the trade-off between the validation and testing datasets was decided by the need to prevent over-fitting [67]. Guyon [68] proposed that the training size determines the model inference, while the validation set (or cross-validation) would serve to indicate which family of feature patterns (recognizer) works best.

In this study, we adopted a 70/20/10 approach because our dataset was small, and a larger training set ratio was preferred. In fact, an optimal splitting ratio may also depend on the type of models, data dimensionality, and validation methods, such as cross-validation and bootstrapping [69,70], posing difficulties for deep learning models with complex model architecture and high data dimensionality, and warranting further investigations on the theories behind hunches.

Activation functions contribute to the advance in deep learning [71] and have a substantial effect on the behavior and performance of deep learning models [72–74]. However, numerous studies overlooked the activation function and associated hyperparameters (e.g., slope coefficient, α) [48,75] and relied on model defaults. In fact, selecting the activation function is exceedingly difficult and typically requires extensive trial-and-error attempts. It depends on the dataset and the problem at hand [75,76]. From a different point of view, it is dependent on the input-output relationship of each node and each layer, but it is hard to trace since the neural network may be a direct rich space of ill-posed functions [77]. The challenge is exacerbated by the notion of the "edge of chaos", which states the model should neither run in an overly ordered nor overly random state [78]. Several studies attempted to offer solutions to this issue. Dushkoff and Ptucha [79] employed more than one activation function depending on the classification error. Jagtap, et al. [76] integrated a basic activation function, using a gated or hierarchical structure to adapt to the inputs, while Li et al. [80] utilized a differential evolution algorithm to determine the activation function based on the input data. Through a "smart search" method, Marchisio et al. [74] realized an automatic selection of the best possible activation functions for each layer. Nevertheless, the optimization of activation functions and associated hyperparameters requires considerable computing power and time.

Imbalanced classes were one of the challenges in different fields using machine learning/deep learning, including medical imaging [81,82], digital health [83–85], and machine learning-driven instruments [19,86]. In fact, imbalanced class scenarios often skew towards the negative cases, since disease cases (positives) are generally rarer than non-disease cases (negatives). Models tend to predict "negative" in a highly imbalanced class problem in order to maximize their probability of making a correct "guess" [87]. In such cases, the loss function of the models could be penalized using a class-weighted inverse proportion of the class size [88]. Nevertheless, we avoid the imbalanced class problem on the training dataset by collecting the same amount of data for each task. For the multiclass issue in the subgroup analysis [89], we mitigate the imbalanced class problem in testing with a bootstrapping approach.

There were some limitations in this research. Firstly, the relatively small size of the testing set may restrict the robustness of the model. In our study, a single incorrect prediction of the testing data would deflate the model accuracy by about 0.5%. A k-fold cross-validation could improve the model robustness upon deployment [90]. Secondly, our protocol design did not purport to cover every swallowing task. While we took reference from the comprehensive assessment protocol for swallowing (CAPS) [91], identifying the fewest swallowing tasks necessary to accurately depict swallowing functions would be helpful to develop the instrument for dysphagia screening and lessen the time and inconvenience during the assessment, which warrant further investigations. The inclination of the camera was determined based on our pilot experiment that better captured the frontal view of the neck area. We believed that our model would be insensitive to the variations of the camera orientation since the model could accommodate the variations by the affine transformation nature of the convolutional layer. Moreover, with respect to subject recruitment, gender could be a significant confounder and critical feature in the study because of the larger Adam's apple in males that might need to be input into the model. Secondly, the duration of the data samples (i.e., sample/sequence length) was about 1.0 to 1.5 s. To learn the temporal features effectively and produce accurate predictions, some models, especially Transformers, need sufficient temporal duration for each data sample [37,92]. Although the sample length requirement could be task-specific, longer data sequences provide more con-

text for the model to learn complex relationships between inputs and outputs [37,92]. Data augmentation techniques might be used to prolong the data sequences [93]. For example, repeating the short video frames to lengthen the video clip. Moreover, data augmentation could help resolve the demand for large datasets in Transformers. Data augmentation for depth frames could be achieved by adding rotations about the three-dimensional axes to simulate different orientations or viewpoints of the depth camera [94]. Alternatively, the Synthetic Minority Over-sampling Technique (SMOTE) could be one way to create synthetic samples by interpolating neighboring instances of that class, which could also be used to resolve the imbalanced class problem [95]. Lastly, we have not constructed explainability maps to understand the attention of the network on salient features and locations since there are no available libraries that could be applied directly to four-dimensional data in our cases, which warrants further investigations.

5. Conclusions

In this study, we developed a stereo-depth camera system to recognize swallowing and non-swallowing through deep learning models. The innovation paves the way towards accessible instrumental dysphagia assessment by expanding our data collection on dysphagic and non-dysphagic populations. Our study determined that X3D was the best model with good-to-excellent performance (F1-score: 0.920; adjusted F1-score: 0.885) in classifying swallowing and non-swallowing conditions using its default activation function. However, the model was only marginally acceptable if individual tasks (fine-grained classification) needed to be recognized (F1-score: 0.649, adjusted F1-score: 0.797). Changing the activation function to LeakyReLU might enhance the classification performance on deep breathing and pronouncing "/aɪ/" tasks. A large dataset, hyperparameter tuning on the activation function, and extensive hyperparameter optimization across high dimensionality are necessary to further improve the system performance.

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

Funding: This study was supported by the Health and Medical Research Fund from the Health Bureau, Hong Kong (reference number: 19200461); and internal fund from the Research Institute for Smart Ageing, The Hong Kong Polytechnic University.

Data Availability Statement: The data and model presented in this study are openly available in GitHub, with the link https://github.com/BME-AI-Lab/Depth_Video_Data_Swallow_Classification_Dataset (accessed on 9 July 2023).

Acknowledgments: We would like to express our sincere gratitude to Chinese medicine practitioner (bonesetter), Ming-Sang Wong, from Wong Ming Sang's Clinic for his assistance in the recruitment of participants and support during the experiment.

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

Appendix A

The pseudocodes used in the study are included in Algorithms A1–A3.

Algorithm A1: The pseudocode of the video frame sampling

Input: *V, the input video with* N_V *number of frames*
s, the number of segments
n, the number of frames per segments
Output: *Array of sampled video frames*
$d \leftarrow floor((N_v - n + 1) \ / \ s)$ *(Compute distance between segments that are approximately evenlyspread across the video frames)*
*Initialized array A [0 . . . s*n]*
for *each segment in video segment* $V[k \times d]$ *to* $V[i + 1) \times d]$ *where k from 0. s* ***do***
$i \leftarrow k \times d + rand(d)$ *(Get the start index of the segment)*
for *j in 0 . . . n* ***do***
$A[n \times s + j] \leftarrow V[i + j]$ *(Append sampled frames to the array)*
end for
end for
return *A*

Algorithm A2: The pseudocode of Adam optimization method

Input: α, *learning rate*
$\beta_1, \beta_2 \in [0, 1)$, *exponential decay rates for the moment estimates*
$f(\theta)$, *stochastic objective function with parameters* θ
θ_0, *initial parameter vector*
ϵ, *small value to prevent division by zero*
Output: θ_t, *resulting parameter vector after t timesteps*
$m_0 \leftarrow 0$
$v_0 \leftarrow 0$
$t \leftarrow 0$
while θ_t *not converged do*
$t \leftarrow t + 1$
$g_t \leftarrow \nabla_\theta f_t(\theta_{t-1})$ *(Compute gradient for the parameters)*
$m_t \leftarrow \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t$ *(Update biased first moment estimate)*
$v_t \leftarrow \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot g_t^2$ *(Update biased second raw moment estimate)*
$\hat{m}_t \leftarrow m_t / \left(1 - \beta_1^t\right)$ *(Compute bias-corrected first moment estimate)*
$\hat{v}_t \leftarrow v_t / \left(1 - \beta_2^t\right)$ *(Compute bias-corrected second raw moment estimate)*
$\theta_t \leftarrow \theta_{t-1} - \alpha \cdot \hat{m}_t / \left(\sqrt{\hat{v}_t} + \epsilon\right)$ *(Update parameters)*
end while
return θ_t

Algorithm A3: The pseudocode of Bootstrapping

Input: $D(V_a)$, *dataset of depth videos with label a*
$D(V_{\sim a})$, *dataset of depth videos without label a*
n, size of each bootstrap sample
Output: *B, array containing the bootstrap samples*
Initialize empty array B
For t from 0 to n − 1 do
Randomly select an index from $D(V_a)$
Add the depth video with index i in $D(V_a)$ *to B*
Randomly select an index from $D(V_{\sim a})$
Add the depth video with index i in $D(V_{\sim a})$ *to B*
end for
return *B*

References

1. Patel, D.; Krishnaswami, S.; Steger, E.; Conover, E.; Vaezi, M.; Ciucci, M.; Francis, D. Economic and survival burden of dysphagia among inpatients in the United States. *Dis. Esophagus* **2018**, *31*, 131. [CrossRef]
2. Rogus-Pulia, N.; Malandraki, G.A.; Johnson, S.; Robbins, J. Understanding dysphagia in dementia: The present and the future. *Curr. Phys. Med. Rehabil. Rep.* **2015**, *3*, 86–97. [CrossRef]
3. Smukalla, S.M.; Dimitrova, I.; Feintuch, J.M.; Khan, A. Dysphagia in the elderly. *Curr. Treat. Options Gastroenterol.* **2017**, *15*, 382–396. [CrossRef]
4. Warnecke, T.; Labeit, B.; Schroeder, J.; Reckels, A.; Ahring, S.; Lapa, S.; Claus, I.; Muhle, P.; Suntrup-Krueger, S.; Dziewas, R. Neurogenic dysphagia: Systematic review and proposal of a classification system. *Neurology* **2021**, *96*, e876–e889. [CrossRef] [PubMed]
5. World Health Organization. *The ICD-10 Classification of Mental and Behavioural Disorders: Clinical Descriptions and Diagnostic Guidelines*; World Health Organization: Geneva, Switzerland, 1992.
6. Malagelada, J.-R.; Bazzoli, F.; Boeckxstaens, G.; De Looze, D.; Fried, M.; Kahrilas, P.; Lindberg, G.; Malfertheiner, P.; Salis, G.; Sharma, P. World gastroenterology organisation global guidelines: Dysphagia—Global guidelines and cascades update September 2014. *J. Clin. Gastroenterol.* **2015**, *49*, 370–378. [CrossRef]
7. Crary, M.A.; Carnaby, G.D.; Sia, I.; Khanna, A.; Waters, M.F. Spontaneous swallowing frequency has potential to identify dysphagia in acute stroke. *Stroke* **2013**, *44*, 3452–3457. [CrossRef]
8. Auyeung, M.; Tsoi, T.; Mok, V.; Cheung, C.; Lee, C.; Li, R.; Yeung, E. Ten year survival and outcomes in a prospective cohort of new onset Chinese Parkinson's disease patients. *J. Neurol. Neurosurg. Psychiatry* **2012**, *83*, 607–611. [CrossRef]
9. Takizawa, C.; Gemmell, E.; Kenworthy, J.; Speyer, R. A systematic review of the prevalence of oropharyngeal dysphagia in stroke, Parkinson's disease, Alzheimer's disease, head injury, and pneumonia. *Dysphagia* **2016**, *31*, 434–441. [CrossRef] [PubMed]
10. Baijens, L.W.; Clavé, P.; Cras, P.; Ekberg, O.; Forster, A.; Kolb, G.F.; Leners, J.-C.; Masiero, S.; Mateos-Nozal, J.; Ortega, O. European Society for Swallowing Disorders–European Union Geriatric Medicine Society white paper: Oropharyngeal dysphagia as a geriatric syndrome. *Clin. Interv. Aging* **2016**, *11*, 1403. [CrossRef] [PubMed]
11. Ekberg, O.; Hamdy, S.; Woisard, V.; Wuttge-Hannig, A.; Ortega, P. Social and psychological burden of dysphagia: Its impact on diagnosis and treatment. *Dysphagia* **2002**, *17*, 139–146. [CrossRef] [PubMed]
12. Bhattacharyya, N. The prevalence of dysphagia among adults in the United States. *Otolaryngol. Head Neck Surg.* **2014**, *151*, 765–769. [CrossRef] [PubMed]
13. Warnecke, T.; Dziewas, R.; Langmore, S. FEES and Other Instrumental Methods for Swallowing Evaluation. In *Neurogenic Dysphagia*; Warnecke, T., Dziewas, R., Langmore, S., Eds.; Springer: Boston, MA, USA, 2021; pp. 55–107.
14. Kertscher, B.; Speyer, R.; Palmieri, M.; Plant, C. Bedside screening to detect oropharyngeal dysphagia in patients with neurological disorders: An updated systematic review. *Dysphagia* **2014**, *29*, 204–212. [CrossRef] [PubMed]
15. Maccarini, A.R.; Filippini, A.; Padovani, D.; Limarzi, M.; Loffredo, M.; Casolino, D. Clinical non-instrumental evaluation of dysphagia. *Acta Otorhinolaryngol. Ital.* **2007**, *27*, 299–305.
16. Suiter, D.M.; Leder, S.B. Clinical utility of the 3-ounce water swallow test. *Dysphagia* **2008**, *23*, 244–250. [CrossRef] [PubMed]
17. Lee, J.Y.; Kim, D.-K.; Seo, K.M.; Kang, S.H. Usefulness of the simplified cough test in evaluating cough reflex sensitivity as a screening test for silent aspiration. *Ann. Rehabil. Med.* **2014**, *38*, 476. [CrossRef] [PubMed]
18. O'Horo, J.C.; Rogus-Pulia, N.; Garcia-Arguello, L.; Robbins, J.; Safdar, N. Bedside diagnosis of dysphagia: A systematic review. *J. Hosp. Med.* **2015**, *10*, 256–265. [CrossRef]
19. So, B.P.-H.; Chan, T.T.-C.; Liu, L.; Yip, C.C.-K.; Lim, H.-J.; Lam, W.-K.; Wong, D.W.-C.; Cheung, D.S.K.; Cheung, J.C.-W. Swallow Detection with Acoustics and Accelerometric-Based Wearable Technology: A Scoping Review. *Int. J. Environ. Res. Public Health* **2023**, *20*, 170. [CrossRef]
20. Lai, D.K.-H.; Cheng, E.S.-W.; Lim, H.-J.; So, B.P.-H.; Lam, W.-K.; Cheung, D.S.K.; Wong, D.W.-C.; Cheung, J.C.-W. Computer-aided screening of aspiration risks in dysphagia with wearable technology: A Systematic Review and meta-analysis on test accuracy. *Front. Bioeng. Biotechnol.* **2023**, *11*, 1205009. [CrossRef]
21. Zahnd, E.; Movahedi, F.; Coyle, J.L.; Sejdić, E.; Menon, P.G. Correlating Tri-Accelerometer Swallowing Vibrations and Hyoid Bone Movement in Patients with Dysphagia. In Proceedings of the ASME 2016 International Mechanical Engineering Congress and Exposition, Phoenix, AZ, USA, 11–17 November 2016.
22. Päßler, S.; Wolff, M.; Fischer, W.-J. Food intake monitoring: An acoustical approach to automated food intake activity detection and classification of consumed food. *Physiol. Meas.* **2012**, *33*, 1073. [CrossRef]
23. Kuramoto, N.; Ichimura, K.; Jayatilake, D.; Shimokakimoto, T.; Hidaka, K.; Suzuki, K. Deep Learning-Based Swallowing Monitor for Realtime Detection of Swallow Duration. In Proceedings of the 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Montreal, QC, Canada, 20–24 July 2020; pp. 4365–4368.
24. Dudik, J.M.; Coyle, J.L.; El-Jaroudi, A.; Mao, Z.-H.; Sun, M.; Sejdić, E. Deep learning for classification of normal swallows in adults. *Neurocomputing* **2018**, *285*, 1–9. [CrossRef]
25. Taniwaki, M.; Kohyama, K. Fast fourier transform analysis of sounds made while swallowing various foods. *J. Acoust. Soc. Am.* **2012**, *132*, 2478–2482. [CrossRef]
26. Farooq, M.; Fontana, J.M.; Sazonov, E. A novel approach for food intake detection using electroglottography. *Physiol. Meas.* **2014**, *35*, 739. [CrossRef]

27. Tajitsu, Y.; Suehiro, A.; Tsunemine, K.; Katsuya, K.; Kawaguchi, Y.; Kuriwaki, Y.; Sugino, Y.; Nishida, H.; Kitamura, M.; Omori, K. Application of piezoelectric braided cord to dysphagia-detecting system. *Jpn. J. Appl. Phys.* **2018**, *57*, 11UG02. [CrossRef]
28. Nguyen, D.T.; Cohen, E.; Pourhomayoun, M.; Alshurafa, N. SwallowNet: Recurrent neural network detects and characterizes eating patterns. In Proceedings of the 2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops), Kona, HI, USA, 13–17 March 2017; pp. 401–406.
29. Cheung, J.C.-W.; So, B.P.-H.; Ho, K.H.M.; Wong, D.W.-C.; Lam, A.H.-F.; Cheung, D.S.K. Wrist accelerometry for monitoring dementia agitation behaviour in clinical settings: A scoping review. *Front. Psychiatry* **2022**, *13*, 913213. [CrossRef] [PubMed]
30. Tam, A.Y.-C.; Zha, L.-W.; So, B.P.-H.; Lai, D.K.-H.; Mao, Y.-J.; Lim, H.-J.; Wong, D.W.-C.; Cheung, J.C.-W. Depth-Camera-based Under-blanket Sleep Posture Classification using Anatomical Landmark-guided Deep Learning Model. *Int. J. Environ. Res. Public Health* **2022**, *19*, 13491. [CrossRef] [PubMed]
31. Tam, A.Y.-C.; So, B.P.-H.; Chan, T.T.-C.; Cheung, A.K.-Y.; Wong, D.W.-C.; Cheung, J.C.-W. A Blanket Accommodative Sleep Posture Classification System Using an Infrared Depth Camera: A Deep Learning Approach with Synthetic Augmentation of Blanket Conditions. *Sensors* **2021**, *21*, 5553. [CrossRef] [PubMed]
32. Bian, Z.-P.; Hou, J.; Chau, L.-P.; Magnenat-Thalmann, N. Fall detection based on body part tracking using a depth camera. *IEEE J. Biomed. Health Inform.* **2014**, *19*, 430–439. [CrossRef]
33. Procházka, A.; Charvátová, H.; Vyšata, O.; Kopal, J.; Chambers, J. Breathing analysis using thermal and depth imaging camera video records. *Sensors* **2017**, *17*, 1408. [CrossRef]
34. An, K.; Zhang, Q.; Kwong, E. ViscoCam: Smartphone-based Drink Viscosity Control Assistant for Dysphagia Patients. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2021**, *5*, 3. [CrossRef]
35. Yoshida, J.; Kozawa, K.; Moritani, S.; Sakamoto, S.-I.; Sakai, O.; Miyagi, S. Detection of Swallowing Times Using a Commercial RGB-D Camera. In Proceedings of the 2019 IEEE 8th Global Conference on Consumer Electronics (GCCE), Osaka, Japan, 15–18 October 2019; pp. 1154–1155.
36. Sugimoto, C.; Masuyama, Y. Elevation Measurement of Laryngeal Prominence from Depth Images for Evaluating Swallowing Function. In Proceedings the of 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Honolulu, HI, USA, 18–21 July 2018; pp. 1562–1565.
37. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is All You Need. *arXiv* **2017**, arXiv:1706.03762. [CrossRef]
38. Dubey, S.R.; Singh, S.K.; Chaudhuri, B.B. Activation Functions in Deep Learning: A Comprehensive Survey and Benchmark. *arXiv* **2021**, arXiv:2109.14545. [CrossRef]
39. Nwankpa, C.; Ijomah, W.; Gachagan, A.; Marshall, S. Activation Functions: Comparison of trends in Practice and Research for Deep Learning. *arXiv* **2018**, arXiv:1811.03378. [CrossRef]
40. Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; Van Gool, L. Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. *arXiv* **2016**, arXiv:1608.00859. [CrossRef]
41. Gastal, E.S.L.; Oliveira, M.M. Domain transform for edge-aware image and video processing. *ACM Trans. Graph.* **2011**, *30*, 69. [CrossRef]
42. Fukushima, K. Cognitron: A self-organizing multilayered neural network. *Biol. Cybern.* **1975**, *20*, 121–136. [CrossRef]
43. He, K.; Zhang, X.; Ren, S.; Sun, J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. *arXiv* **2015**, arXiv:1502.01852. [CrossRef]
44. Hendrycks, D.; Gimpel, K. Gaussian Error Linear Units (GELUs). *arXiv* **2016**, arXiv:1606.08415. [CrossRef]
45. Clevert, D.-A.; Unterthiner, T.; Hochreiter, S. Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs). *arXiv* **2015**, arXiv:1511.07289. [CrossRef]
46. Elfwing, S.; Uchibe, E.; Doya, K. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Netw.* **2018**, *107*, 3–11. [CrossRef] [PubMed]
47. Dauphin, Y.N.; Fan, A.; Auli, M.; Grangier, D. Language Modeling with Gated Convolutional Networks. *arXiv* **2016**, arXiv:1612.08083. [CrossRef]
48. Varshney, M.; Singh, P. Optimizing nonlinear activation function for convolutional neural networks. *Signal Image Video Process.* **2021**, *15*, 1323–1330. [CrossRef]
49. Bengio, Y.; Simard, P.; Frasconi, P. Learning long-term dependencies with gradient descent is difficult. *IEEE Trans. Neural Netw.* **1994**, *5*, 157–166. [CrossRef] [PubMed]
50. Bertasius, G.; Wang, H.; Torresani, L. Is Space-Time Attention All You Need for Video Understanding? *arXiv* **2021**, arXiv:2102.05095. [CrossRef]
51. Arnab, A.; Dehghani, M.; Heigold, G.; Sun, C.; Lučić, M.; Schmid, C. ViViT: A Video Vision Transformer. *arXiv* **2021**, arXiv:2103.15691. [CrossRef]
52. Feichtenhofer, C.; Fan, H.; Malik, J.; He, K. SlowFast Networks for Video Recognition. *arXiv* **2018**, arXiv:1812.03982. [CrossRef]
53. Feichtenhofer, C. X3D: Expanding Architectures for Efficient Video Recognition. *arXiv* **2020**, arXiv:2004.04730. [CrossRef]
54. Tran, D.; Wang, H.; Torresani, L.; Ray, J.; LeCun, Y.; Paluri, M. A Closer Look at Spatiotemporal Convolutions for Action Recognition. *arXiv* **2017**, arXiv:1711.11248. [CrossRef]
55. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *arXiv* **2019**, arXiv:1912.01703. [CrossRef]

56. Saito, T.; Rehmsmeier, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE* **2015**, *10*, e0118432. [CrossRef]
57. Fortuna-Cervantes, J.M.; Ramírez-Torres, M.T.; Mejía-Carlos, M.; Murguía, J.S.; Martinez-Carranza, J.; Soubervielle-Montalvo, C.; Guerra-García, C.A. Texture and Materials Image Classification Based on Wavelet Pooling Layer in CNN. *Appl. Sci.* **2022**, *12*, 3592. [CrossRef]
58. So, B.P.-H.; Lai, D.K.-H.; Cheung, D.S.-K.; Lam, W.-K.; Cheung, J.C.-W.; Wong, D.W.-C. Virtual Reality-Based Immersive Rehabilitation for Cognitive-and Behavioral-Impairment-Related Eating Disorders: A VREHAB Framework Scoping Review. *Int. J. Environ. Res. Public Health* **2022**, *19*, 5821. [CrossRef]
59. Imperatori, C.; Mancini, M.; Della Marca, G.; Valenti, E.M.; Farina, B. Feedback-based treatments for eating disorders and related symptoms: A systematic review of the literature. *Nutrients* **2018**, *10*, 1806. [CrossRef] [PubMed]
60. Selva, J.; Johansen, A.S.; Escalera, S.; Nasrollahi, K.; Moeslund, T.B.; Clapés, A. Video Transformers: A Survey. *arXiv* **2022**, arXiv:2201.05991. [CrossRef] [PubMed]
61. Park, N.; Kim, S. How do vision transformers work? *arXiv* **2022**, arXiv:2202.06709.
62. Bedri, A.; Li, R.; Haynes, M.; Kosaraju, R.P.; Grover, I.; Prioleau, T.; Beh, M.Y.; Goel, M.; Starner, T.; Abowd, G. EarBit: Using wearable sensors to detect eating episodes in unconstrained environments. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2017**, *1*, 1–20. [CrossRef]
63. Noiray, A.; Cathiard, M.A.; Ménard, L.; Abry, C. Test of the movement expansion model: Anticipatory vowel lip protrusion and constriction in French and English speakers. *J. Acoust. Soc. Am.* **2011**, *129*, 340–349. [CrossRef]
64. Yadav, M.; Alam, M.A. Dynamic time warping (dtw) algorithm in speech: A review. *Int. J. Res. Electron. Comput. Eng.* **2018**, *6*, 524–528.
65. Joseph, V.R. Optimal ratio for data splitting. *Stat. Anal. Data Min. ASA Data Sci. J.* **2022**, *15*, 531–538. [CrossRef]
66. Dubbs, A. Test Set Sizing Via Random Matrix Theory. *arXiv* **2021**, arXiv:2112.05977. [CrossRef]
67. Amari, S.; Murata, N.; Muller, K.R.; Finke, M.; Yang, H.H. Asymptotic statistical theory of overtraining and cross-validation. *IEEE Trans. Neural Netw.* **1997**, *8*, 985–996. [CrossRef]
68. Guyon, I.M. A Scaling Law for the Validation-Set Training-Set Size Ratio. Available online: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.33.1337&rep=rep1&type=pdf (accessed on 9 July 2023).
69. Afendras, G.; Markatou, M. Optimality of Training/Test Size and Resampling Effectiveness of Cross-Validation Estimators of the Generalization Error. *arXiv* **2015**, arXiv:1511.02980. [CrossRef]
70. Xu, Y.; Goodacre, R. On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning. *J. Anal. Test.* **2018**, *2*, 249–262. [CrossRef] [PubMed]
71. Erickson, B.J.; Korfiatis, P.; Kline, T.L.; Akkus, Z.; Philbrick, K.; Weston, A.D. Deep learning in radiology: Does one size fit all? *J. Am. Coll. Radiol.* **2018**, *15*, 521–526. [CrossRef] [PubMed]
72. Marcu, D.C.; Grava, C. The impact of activation functions on training and performance of a deep neural network. In Proceedings of the 2021 16th International Conference on Engineering of Modern Electric Systems (EMES), Oradea, Romania, 10–11 June 2021. [CrossRef]
73. Farzad, A.; Mashayekhi, H.; Hassanpour, H. A comparative performance analysis of different activation functions in LSTM networks for classification. *Neural Comput. Appl.* **2019**, *31*, 2507–2521. [CrossRef]
74. Marchisio, A.; Abdullah Hanif, M.; Rehman, S.; Martina, M.; Shafique, M. A Methodology for Automatic Selection of Activation Functions to Design Hybrid Deep Neural Networks. *arXiv* **2018**, arXiv:1811.03980. [CrossRef]
75. Basirat, M.; Roth, P.M. The Quest for the Golden Activation Function. *arXiv* **2018**, arXiv:1808.00783. [CrossRef]
76. Jagtap, A.D.; Kawaguchi, K.; Karniadakis, G.E. Adaptive activation functions accelerate convergence in deep and physics-informed neural networks. *J. Comput. Phys.* **2020**, *404*, 109136. [CrossRef]
77. Parhi, R.; Nowak, R.D. The role of neural network activation functions. *IEEE Signal Process. Lett.* **2020**, *27*, 1779–1783. [CrossRef]
78. Hayou, S.; Doucet, A.; Rousseau, J. On the Selection of Initialization and Activation Function for Deep Neural Networks. *arXiv* **2018**, arXiv:1805.08266. [CrossRef]
79. Dushkoff, M.; Ptucha, R. Adaptive activation functions for deep networks. In Proceedings of the IS&T International Symposium on Electronic Imaging, San Francisco, CA, USA, 14–18 February 2016; pp. COIMG-149.1–COIMG-149.5.
80. Li, B.; Li, Y.; Rong, X. The extreme learning machine learning algorithm with tunable activation function. *Neural Comput. Appl.* **2013**, *22*, 531–539. [CrossRef]
81. Mao, Y.-J.; Zha, L.-W.; Tam, A.Y.-C.; Lim, H.-J.; Cheung, A.K.-Y.; Zhang, Y.-Q.; Ni, M.; Cheung, J.C.-W.; Wong, D.W.-C. Endocrine Tumor Classification via Machine-Learning-Based Elastography: A Systematic Scoping Review. *Cancers* **2023**, *15*, 837. [CrossRef]
82. Mao, Y.-J.; Lim, H.-J.; Ni, M.; Yan, W.-H.; Wong, D.W.-C.; Cheung, J.C.-W. Breast tumour classification using ultrasound elastography with machine learning: A systematic scoping review. *Cancers* **2022**, *14*, 367. [CrossRef]
83. Solares, J.R.A.; Raimondi, F.E.D.; Zhu, Y.; Rahimian, F.; Canoy, D.; Tran, J.; Gomes, A.C.P.; Payberah, A.H.; Zottoli, M.; Nazarzadeh, M. Deep learning for electronic health records: A comparative review of multiple deep neural architectures. *J. Biomed. Inform.* **2020**, *101*, 103337. [CrossRef] [PubMed]
84. Ebrahimi, A.; Wiil, U.K.; Schmidt, T.; Naemi, A.; Nielsen, A.S.; Shaikh, G.M.; Mansourvar, M. Predicting the risk of alcohol use disorder using machine learning: A systematic literature review. *IEEE Access* **2021**, *9*, 151697–151712. [CrossRef]

85. Artetxe, A.; Beristain, A.; Grana, M. Predictive models for hospital readmission risk: A systematic review of methods. *Comput. Methods Programs Biomed.* **2018**, *164*, 49–64. [CrossRef] [PubMed]
86. Paganelli, A.I.; Mondéjar, A.G.; da Silva, A.C.; Silva-Calpa, G.; Teixeira, M.F.; Carvalho, F.; Raposo, A.; Endler, M. Real-time data analysis in health monitoring systems: A comprehensive systematic literature review. *J. Biomed. Inform.* **2022**, *127*, 104009. [CrossRef] [PubMed]
87. Ling, C.X.; Sheng, V.S. Cost-sensitive learning and the class imbalance problem. *Encycl. Mach. Learn.* **2008**, *2011*, 231–235.
88. Sinha, S.; Ohashi, H.; Nakamura, K. Class-wise difficulty-balanced loss for solving class-imbalance. *arXiv* **2020**, arXiv:2010.01824.
89. Abd Elrahman, S.M.; Abraham, A. A review of class imbalance problem. *J. Netw. Innov. Comput.* **2013**, *1*, 332–340.
90. Lei, J. Cross-validation with confidence. *J. Am. Stat. Assoc.* **2020**, *115*, 1978–1997. [CrossRef]
91. Lim, H.-J.; Lai, D.K.-H.; So, B.P.-H.; Yip, C.C.-K.; Cheung, D.S.K.; Cheung, J.C.-W.; Wong, D.W.-C. A Comprehensive Assessment Protocol for Swallowing (CAPS): Paving the Way towards Computer-Aided Dysphagia Screening. *Int. J. Environ. Res. Public Health* **2023**, *20*, 2998. [CrossRef]
92. Tay, Y.; Dehghani, M.; Abnar, S.; Shen, Y.; Bahri, D.; Pham, P.; Rao, J.; Yang, L.; Ruder, S.; Metzler, D. Long range arena: A benchmark for efficient transformers. *arXiv* **2020**, arXiv:2011.04006.
93. Feng, S.Y.; Gangal, V.; Wei, J.; Chandar, S.; Vosoughi, S.; Mitamura, T.; Hovy, E. A survey of data augmentation approaches for NLP. *arXiv* **2021**, arXiv:2105.03075.
94. Dawar, N.; Ostadabbas, S.; Kehtarnavaz, N. Data Augmentation in Deep Learning-Based Fusion of Depth and Inertial Sensing for Action Recognition. *IEEE Sens. Lett.* **2019**, *3*, 7101004. [CrossRef]
95. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-sampling Technique. *arXiv* **2011**, arXiv:1106.1813. [CrossRef]