A Comparison of Frame-by-Frame and Aggregation Approaches for Gesture Classification Using Machine Learning

Michał Wierzbicki[1,†], Jakub Osuchowski[1,*,†]

[1] *Opole University of Technology, Prószkowska 76 St., Opole, 45-758, Poland*

Abstract

In medicine, gesture and body pose analysis, especially in the context of telemedicine and rehabilitation, has gained importance after the COVID-19 pandemic. Gesture recognition and body pose estimation demand high computational power to process large and complex data. To address this problems Authors examined machine learning methods and aggregation techniques for sequential data.

This paper compares two gesture analysis methods: frame-by-frame and gesture sequence analysis. The iMiGUE dataset, which contains skeleton data obtained using the OpenPose tool, is used. In this paper, the gesture classification results obtained using the RandomForestClassifier model with default and optimized parameters are evaluated in detail.

Sequential gesture analysis methods outperformed the classical frame-by-frame analysis in terms of precision and computational efficiency.

Keywords

sequential analysis, machine learning, frame-by-frame analysis, gesture recognition, skeleton data

1. Introduction

Recent years have yielded the most advanced solutions in the domain of artificial intelligence (AI) to date, just to mention transformers architecture [1] and plethora of large language models (LLMs) based on that idea namely ChatGPT (GPT stands for Generative Pretrained Transformer) or Gemini [2],[3],[4]. While results obtained by those models are marvelous, they have incurred significant costs [5],[6], impossible to bear by many institutions. Costs are mainly related to the number of parameters used in the training process - in some cases they reach billions - as well as length of training itself [5]. LLMs have found applications in a variety of domains, for example in healthcare [7],[8].

Yet solely analyzing language itself does not exhaust all possibilities of advanced AI solutions for healthcare. One area that proved to be beneficiary of AI development is gesture and body position recognition. [9] provides distinctions of three groups of gestures that are of our interests: head, hand and body.

Gestures convey more information that can be inferred from speech alone [10]. Significant effort has been put towards development of more robust and precise techniques for gesture recognition whether it is hand specifically [11],[12] or body and head [13],[14],[15]. Those techniques proved to be crucial for rehabilitation purposes for post-stroke patients [16] or people with cerebral palsy [17] allowing medical practitioners to remotely assess a patient's condition and state. In a more general sense gesture and body recognition yield new

ITTAP'2024: 4th International Workshop on Information Technologies: Theoretical and Applied Problems, November 20–22, 2024, Ternopil, Ukraine, Opole, Poland

* Corresponding author.

† These authors contributed equally.

michal.wierzbicki@student.po.edu.pl (M. Wierzbicki); j.osuchowski@po.edu.pl (J. Osuchowski)

0009-0006-5167-0866 (M. Wierzbicki); 0000-0002-9404-966X (J. Osuchowski)

opportunities for rehabilitation processes [18],[19] providing more approachable ways of monitoring progress and overall condition of the patient. As demonstrated by the authors in [20], virtual reality (VR) offers a broad spectrum of applications in neurological rehabilitation. This is largely attributed to its ability to easily replicate natural environments, design specific movement patterns, and create engaging exercises in which patients can actively participate. During these exercises, it is crucial for the patient to be monitored through gesture recognition, allowing for effective tracking of their progress and identification of any obstacles.

COVID-19 pandemic presented a plethora of challenges and obstacles for the healthcare workforce [21]. Unexpected circumstances forced the healthcare sector to adjust to an unfamiliar environment, rendering then-present methods of rehabilitation impossible to execute in a new context. Post-COVID-19 rehabilitation has been deemed "an effective therapeutic strategy to improve the functional capacity and quality of life of patients" [22] yielding improvements in quality of both physical and psychological aspects of life [23]. Pandemic circumstances put emphasis on the development of the niche of telemedicine and remote healthcare [24] with rehabilitation being one of the most crucial aspects. Remote rehabilitation has been implemented during pandemic and to this day it is relied upon by the medical practitioners as a mean that "was safe, feasible, and acceptable for those who accessed it" [25],[26]. In the context of remote rehabilitation gesture recognition and body pose estimation can be an important element of monitoring a patient's wellbeing and recovery [27], [28]. At the same time, in the context of pandemic, the proposed methods may not be accessible for some patients due to limitations of possessed hardware unable to perform required work for proper assessment based on gestures of body pose. Therefore, methods to decrease computational load are necessary to ensure availability of the service and its quality.

In this paper we examine simple methods based on aggregation of the gestures, that prove to be beneficial both in terms of required computational power and storage as well as a performance. Given a dataset [29] consisting of skeleton data we aggregate data that describes each gesture presented in a dataset using five methods: minimum, maximum, integral, average and regression, and we compare obtained results with results obtained on original data.

2. Related works

Significant effort has been put towards examining techniques of processing the data for gesture recognition or body pose estimation. Authors in [30] state that the most focus in research is put on "RGB data, depth data, or skeleton data". In this paper we will solely focus on skeleton data. Ionescu et al. [31] proposed a strategy for segmentation of image for body pose estimation relying on regression to obtain joints coordinates. Wang et al. [32] distinct two regression approaches in case of single person body pose estimation in 2D: "direct regression-based approach, which involves regressing key points directly from features" and heatmap approach that infers joints positions from the heatmap. For the 3D case two mentioned approaches found applications, as well as a third approach that combines 2D and 3D approaches into one complex framework. This work focus on 3D skeleton data.

While there is much effort put into discovery of the new techniques and methods of processing the data, there is significantly less effort put towards examining ways of easing computational load for recognizing gestures or body pose estimation. In the same vein, the

aspect of compressing, like in our case, skeleton data or methods of aggregating such data leave still much to be desired for. While new techniques are providing impressive results (e.g. variations on Spatio-Temporal Graph Convolutional Network [33]), they offer very little in terms of improving the general methodology for the data processing process.

3. Methodology

3.1. Dataset

In this study, the iMiGUE dataset [29] was utilized. This dataset was specifically created to analyze micro-gestures in the context of emotional AI. It contains videos of press conferences following tennis Grand Slam matches, where players respond to questions from journalists. iMiGUE was designed to investigate hidden emotions by analyzing micro-gestures, which are small, often unconscious movements that reflect internal emotional states. The videos were collected from various open video platforms, such as YouTube, and included 359 videos, including 258 winning and 101 losing matches, for a total of 2092 minutes of footage. All videos have a resolution of 1280x720 pixels and were recorded at a rate of 25 frames per second. The data is labeled at two levels: micro-gesture categories at the video clip level and emotion categories at the entire video level. A total of 18,499 micro-gesture samples were labeled, assigning them 32 different categories. It's worth noting that iMiGUE is a dataset that protects individuals' privacy by removing biometric data such as face and voice. It contains data from 72 athletes from 28 countries, allowing for analysis of micro-gestures in the context of diverse cultures and genders. Additionally, the dataset was notably imbalanced, which led to significantly low performance in both detection and classification tasks. Nonetheless, the application of class balancing techniques, as discussed in [34],[35], could potentially enhance the performance in these areas.

In the research, the RGB material contained in the iMiGUE dataset was not used. Instead, the focus was solely on the skeleton data. The skeleton data in the iMiGUE dataset is constructed to facilitate the recognition and understanding of micro-movements. This data is obtained using the OpenPose tool [14], which extracts pose data for each frame of a video sequence. The pose data includes key points corresponding to different body parts, creating a skeletal representation of a person's posture and movements over time. The dataset uses a sequence of key body points (or pose data) for each micro-movement instance, where each frame in the sequence contains the coordinates of key joints. These key points capture the spatial configuration of the body, allowing for the analysis of subtle movements that characterize micro-movements. Skeletal data is advantageous because it is insusceptible to dynamic background changes, making it more suitable for gesture recognition tasks in different environments. By focusing on skeleton data, the iMiGUE dataset provides a detailed and private way to analyze micro-movements, which are crucial for understanding hidden or suppressed emotions.

In this study, Authors used the training and validation data split proposed in the MiGA (Micro-gesture Analysis for Hidden Emotion Understanding) challenge [36]. The data ware restricted to a few selected micro-gestures that the Authors believed had sufficient support in validation part of the dataset to perform correct inference. We selected the following micro-gestures: ear touching (1720 samples, denoted as gesture 8), torso touching (3329 samples,

denoted as gesture 20), finger crossing (184 samples, denoted as gesture 24), lip pressing (2746 samples, denoted as gesture 29), shoulder shaking (4261 samples, denoted as gesture 31), and unspecified gestures (9670 samples, denoted as gesture 99). These classes were selected because they had enough samples, which is crucial for performing correct analysis and inference.

3.2. Sequence data processing

The data in this dataset was originally divided into sequences of gestures (e.g., touching ears sequence, touching torso sequence, crossing fingers sequence, etc.). In the study, the Authors investigated whether frame-by-frame inference (denoted as Base) would yield worse results than using simple methods that allow for the analysis of entire sequences. The simple methods proposed by the Authors included calculating the mean value for the entire sequence (denoted as Avg), determining the minimum (denoted as Min) and maximum (denoted as Max) values for the entire sequence, and performing linear regression on each sequence (denoted as Reg) and taking the slope value (a).

The formula for linear regression in this context can be represented as:

$$y = ax + b \tag{1}$$

Where, a is the slope and b is the intercept.

Additionally, Authors used the integral (trapezoidal rule) for calculating the sequence value (denoted as Int). In this case the vector of sequence values $y = ¿]$ is uniformly distributed over the interval [0,1] and the integral value is calculated using the formula:

$$Integral \approx \int_0^1 f(x)\,dx \approx \frac{1}{n}\sum_0^{n-1} \frac{y_i + y_{i+1}}{2} \tag{2}$$

Where:

- y_i and y_{i+1} are successive values in the vector y,
- n is the number of intervals.

When sequence-based inference was used, the support for the data changed. The new support values were touching ears (34 samples), touching torso (55 samples), crossing fingers (10 samples), pressing lips (82 samples), shaking shoulders (193 samples), and undefined gestures (258 samples). The application of these methods for aggregating gestures belonging to the same sequence allowed for the analysis of entire gesture sequences, rather than analyzing each frame individually.

3.3. First Phase - Used model

In the first part of the study, the model *RandomForestClassifier* [37] with parameters $n_{estimators} = 100$ and $random_{state} = 42$ was utilized. This model is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes

(classification) or mean prediction (regression) of the individual trees. The parameter $n_{estimators}$ specify the number of trees in the forest, while $random_{state}$ ensures reproducibility of the results. This approach was considered the initial step in the study, allowing for a preliminary verification of the research hypothesis.

3.4. Second Phase - Grid Search

In the second part of the study, a Grid Search method was applied to find the best parameters. The parameter grid included the following:

- $n_{estimators}$: number of trees in the forest, with values [100, 200, 300, 400, 500],
- $max_{features}$: number of features to consider when looking for the best split, with options [None, 'sqrt', 'log2'],
- max_{depth}: maximum depth of the tree, with values [None, 10, 20, 30, 40, 50],
- $min_{¿}$: minimum number of samples required to split an internal node, with values [2, 5, 10],
- $min_{¿}$: minimum number of samples required to be at a leaf node, with values [1, 2, 4],
- $bootstrap$: whether bootstrap samples are used when building trees, with options [True, False],
- $criterion$: function to measure the quality of a split, with options ['gini', 'entropy', 'log_loss'],
- oob_{score}: whether to use out-of-bag samples to estimate the generalization accuracy, with options [True, False],
- $class_{weight}$: weights associated with classes, with options [None, 'balanced', 'balanced_subsample'].

The parameters were not searched in a brute-force manner (every combination of parameters), but instead, to save time (as the calculation for Base model was time consuming), it was decided to check each parameter sequentially. First, the best value for the first parameter was selected, then using this value, the best value for the second parameter was determined, and so on. This strategy, which can be used in model optimization [38], allowed each of the proposed models to select parameters that fit best its structure.

3.5. Used metrics

In the study, the following metrics were used to evaluate the algorithms: $Precision, Recall, F1\,Score \wedge Accuracy$. $Precision$measures the exactness of the positive predictions made by the model. It represents the proportion of correctly predicted positive outcomes (true positives) to all outcomes that the model predicted as positive (true positives + false positives) [39]. The formula for precision is as follows:

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \tag{3}$$

where:

- True Positives are the number of correctly predicted positive cases,
- False Positives are the number of incorrectly predicted positive cases.

Recall measures the ability of a model to correctly identify all instances of an object in a dataset. It represents the ratio of the number of correctly detected instances of an object (true positives) to the total number of actual instances of the object in the dataset (true positives + false negatives) [39]. The formula for recall is as follows:

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \tag{4}$$

where:
- True Positives are the number of correctly predicted positive cases,
- False Negatives are the number of actual positive cases that were incorrectly predicted as negative.

The $F1\,Score$ is the harmonic average of *Precision* and *Recall*. It strikes a balance between *Precision* and *Recall*, which is especially important when we want to account for both false positives and false negatives in our model evaluation. The $F1\,Score$ ranges from 0 to 1, with higher values indicating better performance. An ideal $F1\,Score$ of 1 means that the model has achieved both perfect *Precision* and perfect *Recall*, suggesting that it is able to correctly detect all instances of an object without generating false positives [39],[40]. The formula for $F1\,Score$ is as follows:

$$F1\,Score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{5}$$

where:
- *Precision* is the ratio of true positive predictions to the total number of positive predictions made (both true and false positives).
- *Recall* is the ratio of true positive predictions to the total number of actual positive cases (both true positives and false negatives).

The *Accuracy* metric is one of the simplest and most commonly used metrics for evaluating a classification model. It measures the percentage of correctly predicted results over the total number of cases in the data set. *Accuracy* is particularly useful when the data is balanced, i.e. when the number of examples of each class is similar [40]. The formula for *Accuracy* is as follows:

$$Accuracy = \frac{True\ Positives + True\ Negatives}{True\ Positives + True\ Negatives + FalsePositives + False\ Negatives} \tag{6}$$

where:
- True Positives (TP) are the number of correctly predicted positive cases.
- True Negatives (TN) are the number of correctly predicted negative cases.
- False Positives (FP) are the number of incorrectly predicted positive cases.

- False Negatives (FN) are the number of incorrectly predicted negative cases.

4. Results

Authors have conducted examinations that aimed to test how the results obtained by proposed aggregation compare to the Baseline results. By the Baseline results we understand results obtained by performing fitting of the model on the Base dataset, whereas the aggregated results refer to results obtained by fitting models on each of the aggregated datasets. For each of the Phases and for each of the dataset one fitting was performed accordingly.

In the tables below results of each Phase are presented by each chosen metric. For each metric the best score was marked in red, while the best score was marked in green.

4.1. First Phase

In the first Phase every aggregated dataset as well as a Base dataset were fitted on *RandomForestClassifier*. **Table 1** presents obtained results:

- for $F1\,Score$the best result was obtained by regression method (0.55) while the worst result was produced by Baseline model (0.43)
- for *Precision* score surprisingly three methods managed to obtain the same result (0.57), namely Regression, Integral and Maximum methods, while the Minimum method generated the worst result (0.47)
- for *Recall* score the best result was produced by Regression and Integral methods (0.57) and the worst one was produced by Baseline method (0.44)
- for *Accuracy* metric the best score was obtained by Regression and Integral methods (0.57) and the worst one was obtained by Baseline method (0.44)

Table 1
Results for Phase I

	F1 Score	Precision	Recall	Accuracy
Base	0.43	0.49	0.44	0.44
Avg	0.49	0.5	0.52	0.52
Reg	0.55	0.57	0.57	0.57
Int	0.54	0.57	0.57	0.57
Min	0.45	0.47	0.47	0.47
Max	0.54	0.57	0.56	0.56

In the first Phase of the examination the most consistent and best-scoring method turned out to be Regression, closely followed by Integral method, which scored slightly worse in $F1\,Score$ terms and kept equal on all other metrics. Maximum method scored almost as well as two mentioned methods, placing third. The Minimum and Average methods scored noticeably worse than former methods, yet better than the

Baseline results. Baseline results turned out to be the worst scores in three out of four metrics, making it the worst performing approach.

4.2. Second Phase

In the second Phase every aggregated dataset as well as a Base dataset were fitted on *RandomForestClassifier* with Grid Search. **Table 2** presents obtained results:

- for the $F1\,Score$ the best result was obtained by Maximum method (0.56), second best score was Regression (0.55), at the same time Baseline model was the worst scoring one (0.45)
- for *Precision* two methods were able to produce the best results, namely Integral and Maximum (0.60), Regression was able to produce second best result (0.58), the lowest scoring was Baseline (0.47)
- for *Recall* metric again the best score was obtained by Maximum method (0.59) with second best scoring method Integral (0.58), while the lowest score was obtained by Baseline (0.46)
- in terms of *Accuracy* the best scoring method was Maximum (0.59), second best was Integral (0.58) and the lowest scoring was Baseline (0.46)

Table 2
Results for Phase II

	F1 Score	Precision	Recall	Accuracy
Base	0.45	0.47	0.46	0.46
Avg	0.51	0.53	0.55	0.55
Reg	0.55	0.58	0.55	0.57
Int	0.55	0.60	0.58	0.58
Min	0.48	0.50	0.5	0.5
Max	0.56	0.60	0.59	0.59

In Phase II of the examination the best scoring and most consistent method across all metrics was Maximum, which obtained the best results in all four metrics. It was followed by Regression and Integral methods, which produced overall good results, slightly worse than the former method. The overall worst performance in this Phase was performed by Baseline method, which placed last in all considered metrics. Minimum and Average methods both performed better than the Baseline model, but also noticeably worse than the first two mentioned.

4.3. Comparison

In the final part of the study, the $F1\,Scores$ from the I and II Phases were compared separately for each gesture, taking into account each presented method of sequential data processing. Figure 1 illustrates the $F1\,Scores$ for the ear-touching gesture.

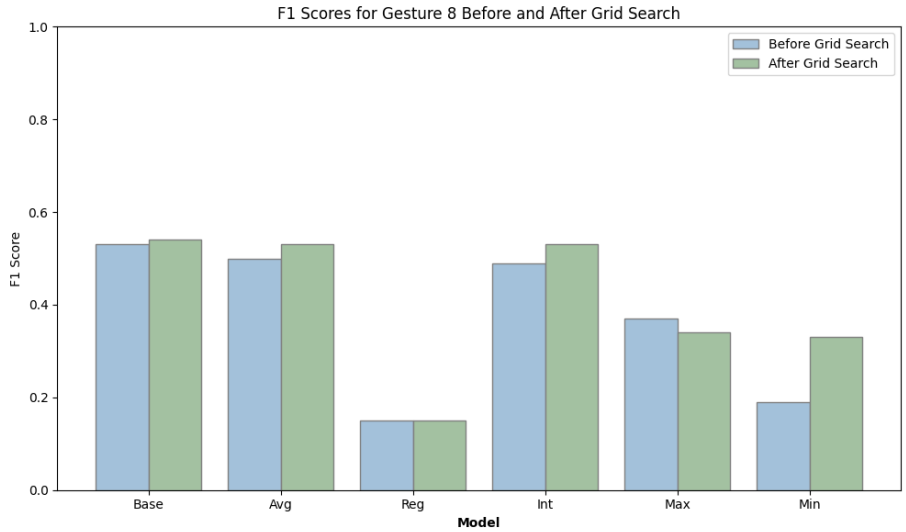

Figure 1: $F1\,Score$ value for gesture 8 (ear touching) before and after Grid Search

After applying Grid Search, better results were obtained for the methods: Base, Avg, Int, and Min. Worse results were observed for the Max method, while the Reg method yielded the same results. The results for gesture 20 (torso touching) are presented in Figure 2.

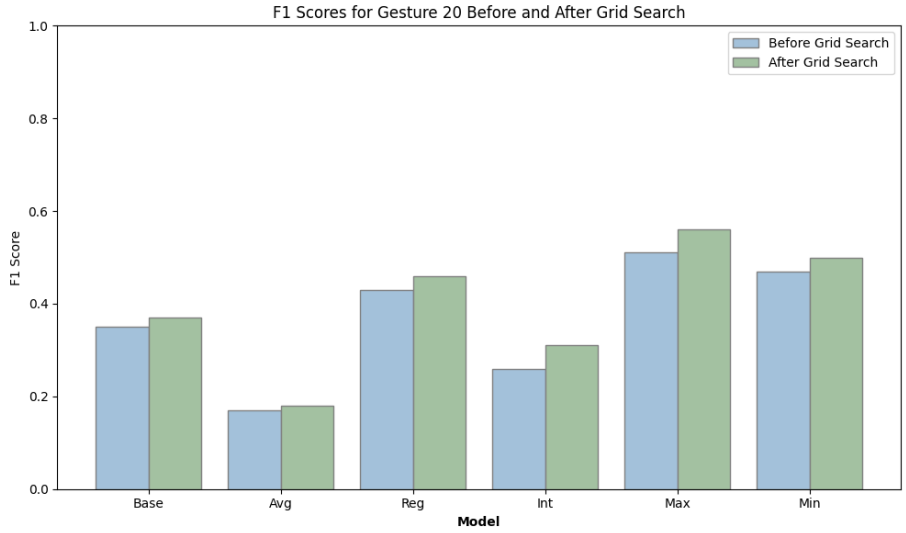

Figure 2: $F1\,Score$ value for gesture 20 (torso touching) before and after Grid Search

For gesture 20 - torso touching, all results improved after applying Grid Search. The best result for this gesture was obtained with the Max method, while the worst was with the Avg method. Notably, before applying Grid Search, the best result was also with Max, and the worst with Avg. Figure 3 presents the results for gesture 24 - finger crossing.

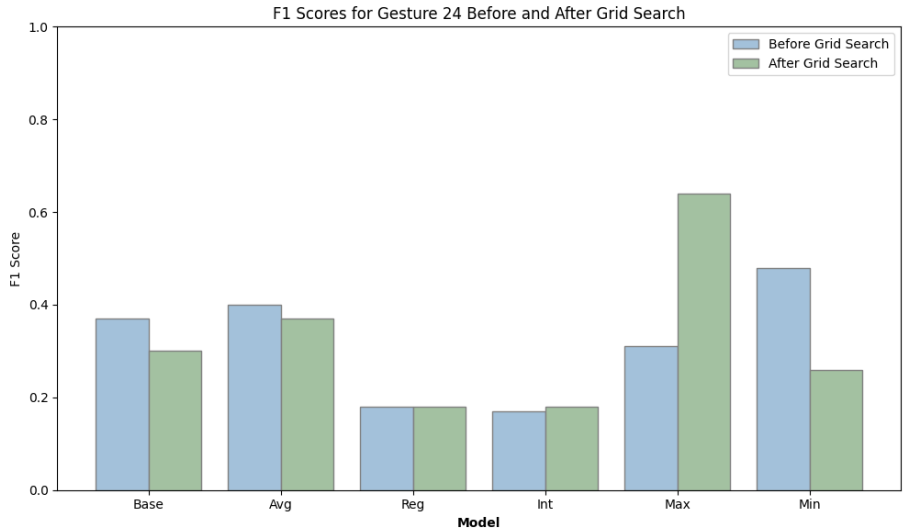

Figure 3: $F\,1\,Score$ value for gesture 24 (finger crossing) before and after Grid Search

A significant improvement after applying Grid Search can be observed for the Max method, while a substantial decline is seen for the Min method. The best result was achieved with the Max method, and the worst with the Reg method. Notably, before applying Grid Search, the best result was with Min, and the worst with Int. Figure 4 presents the results for gesture 29, which represents lip pressing.

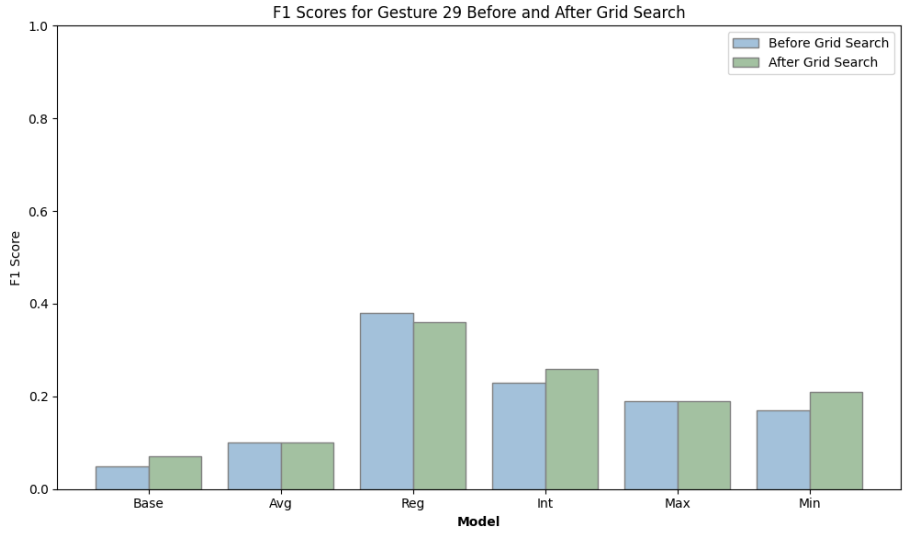

Figure 4: $F\,1\,Score$ value for gesture 29 (lip pressing) before and after Grid Search

Also, a significant improvement after applying Grid Search can be observed for the Base, Int, and Min methods, while a decline is seen for the Reg method, and comparable results for the Avg and Max methods. The best result both before and after applying Grid Search was achieved with the Reg method, while the worst result in both cases was with the Base method. Figure 5 shows the results for gesture 31 - shoulder shaking.

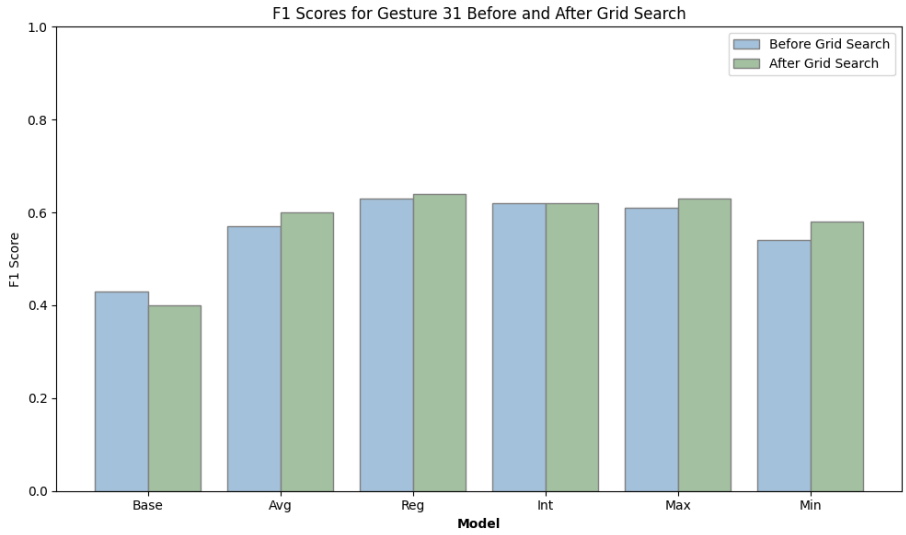

Figure 5: $F1\,Score$ value for gesture 31 (shoulder shaking) before and after Grid Search

All models with the applied sequential analysis method performed better than the Base model for gesture 31, both before and after Grid Search. The best performance, both before and after Grid Search, was achieved by the Reg method, while the worst was by the Base method. After Grid Search, the models that showed improvement were Avg, Reg, Max, and Min. Int model remained the same, while Base performed worse. Figure 6 presents the results for gesture 99 (unspecified gestures).

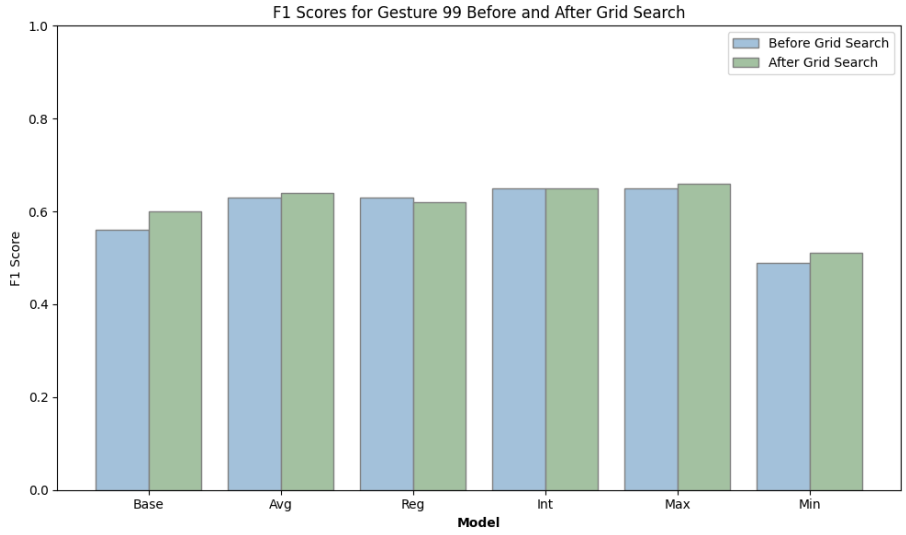

Figure 6: $F1\,Score$ value for gesture 99 (unspecified) gestures before and after Grid Search

For gesture 99 - unspecified gestures, Grid Search led to overall improvements. Most models showed slight score increases, with the model using Maximum value performing best. The Base model also improved significantly. The Regression model experienced a slight decline,

and the model with Integration remained stable. In this case the worst results were achieved by the Min model.

Overall, the application of Grid Search generally improved model performance across various gestures. Most models showed enhancements in their scores, with the models using Maximum value and Integration methods consistently performing well. The Averaging model also demonstrated notable improvement. However, the Regression model occasionally experienced slight declines, and while the Minimum value model showed some improvement, it often remained the lowest performer.

5. Conclusion

Conducted examination proved that using simple aggregating methods for sequential data can be beneficial. The best scoring methods were consistently better than Baseline results, while using a more robust model with Grid Search improved them further. Furthermore, using said methods results in more benefits. One of them is the space it takes to store the data. The original dataset (restricted to just 6 gestures) takes around 400 MB of memory. The aggregated datasets take from around 6 MB up to 12 MB. In the worst case scenario it proves over 33 times reduction in size. Another aspect is the time it takes to calculate a model. For the Baseline method with Grid Search it took over 71 hours to compute. At the same time, for aggregation methods it took on average 1 hour and 28 minutes to compute a model, which resulted in over 48 times improvement. Computing all 5 models took over 9 times less time than the one Baseline model.

In both Phases aggregation methods Regression, Integral and Maximum proved to be worth considerations for further research. All three of them were outperforming the remaining two aggregations methods - Minimum and Average. Nonetheless, all aggregation methods were performing better than the Baseline results.

6. Future works

This paper serves as an introduction for further analysis that examines more complex approaches to aggregating data. While only simple methods are presented, this publication serves as a basis for future works in this direction. An example of such a complex method is the TCIP method, which is described in more detail in [41]. We aim to examine some methods that would allow us to significantly reduce the size of the dataset, while at the same time maintaining, or even in some cases improving, the level of obtained results.

Although obtained results are promising further work is necessary to establish scale of application of those solutions. Presented methods can be further tested on the Baseline dataset to establish comparison solely on the full data instead of aggregated. Authors would also like to acknowledge that the examination was performed on one dataset, further examination on other datasets may be beneficial for estimating usefulness of aggregation methods. Moreover, conducted examination focus on readily available data, we do not process data on the fly. While this might be an interesting approach, it is outside of the scope of this paper.

References

[1] A. Vaswani et al., “Attention Is All You Need,” arXiv.org, Jun. 12, 2017. https://arxiv.org/abs/1706.03762

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” arXiv.org, Oct. 11, 2018. https://arxiv.org/abs/1810.04805

[3] T. B. Brown et al., “Language Models Are Few-Shot Learners,” arxiv.org, vol. 4, May 2020, Available: https://arxiv.org/abs/2005.14165

[4] S. S. Gill and R. Kaur, “ChatGPT: Vision and Challenges,” Internet of Things and Cyber-Physical Systems, vol. 3, pp. 262–271, 2023, doi: https://doi.org/10.1016/j.iotcps.2023.05.004.

[5] O. Sharir, B. Peleg, and Y. Shoham, “The Cost of Training NLP Models: A Concise Overview,” arXiv:2004.08900 [cs], Apr. 2020, Available: https://arxiv.org/abs/2004.08900

[6] S. Samsi et al., “From Words to Watts: Benchmarking the Energy Costs of Large Language Model Inference.” Available: https://arxiv.org/pdf/2310.03003

[7] M. Sallam, “ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns,” Healthcare, vol. 11, no. 6, p. 887, Mar. 2023, doi: https://doi.org/10.3390/healthcare11060887.

[8] P. P. Ray, “Timely need for navigating the potential and downsides of LLMs in healthcare and biomedicine,” Briefings in bioinformatics, vol. 25, no. 3, Mar. 2024, doi: https://doi.org/10.1093/bib/bbae214.

[9] T. Acharya, “Gesture Recognition: A Survey,” IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews), Jan. 2000, Accessed: Apr. 26, 2024. [Online]. Available: https://www.academia.edu/32594485/Gesture_Recognition_A_Survey

[10] S. Goldin-Meadow, “The role of gesture in communication and thinking,” Trends in Cognitive Sciences, vol. 3, no. 11, pp. 419–429, Nov. 1999, doi: https://doi.org/10.1016/s1364-6613(99)01397-2.

[11] H.-J. Kim, J. S. Lee, and J.-H. Park, “Dynamic hand gesture recognition using a CNN model with 3D receptive fields,” Workshop on Neural Networks for Signal Processing, 2008, Accessed: Jun. 24, 2024. [Online]. Available: https://www.semanticscholar.org/paper/Dynamic-hand-gesture-recognition-using-a-CNN-model-Kim-Lee/1ae971c25646d003e15dd9eb706650e58c21d900

[12] G. Devineau, F. Moutarde, W. Xi, and J. Yang, “Deep Learning for Hand Gesture Recognition on Skeletal Data,” IEEE Xplore, May 01, 2018. https://ieeexplore.ieee.org/document/8373818 (accessed Nov. 26, 2022).

[13] E. Samkari, M. Arif, M. Alghamdi, and M. A. Al Ghamdi, “Human Pose Estimation Using Deep Learning: A Systematic Literature Review,” Machine Learning and Knowledge Extraction, vol. 5, no. 4, pp. 1612–1659, Dec. 2023, doi: https://doi.org/10.3390/make5040081.

[14] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, “OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields,” arXiv:1812.08008 [cs], May 2019, Available: https://arxiv.org/abs/1812.08008

[15] L. Pishchulin et al., “DeepCut: Joint Subset Partition and Labeling for Multi Person Pose Estimation,” arXiv:1511.06645 [cs], Apr. 2016, Available: https://arxiv.org/abs/1511.06645

[16] O. N. Zestas, D. N. Soumis, K. D. Kyriakou, K. Seklou, and N. D. Tselikas, "A computer-vision based hand rehabilitation assessment suite," AEU - International Journal of Electronics and Communications, vol. 169, p. 154762, Sep. 2023, doi: https://doi.org/10.1016/j.aeue.2023.154762.
[17] Y.-J. Chang, W.-Y. Han, and Y.-C. Tsai, "A Kinect-based upper limb rehabilitation system to assist people with cerebral palsy," Research in Developmental Disabilities, vol. 34, no. 11, pp. 3654–3659, Nov. 2013, doi: https://doi.org/10.1016/j.ridd.2013.08.021.
[18] V. Tsakanikas et al., "Automated Assessment of Balance Rehabilitation Exercises With a Data-Driven Scoring Model: Algorithm Development and Validation Study," JMIR Rehabilitation and Assistive Technologies, vol. 9, no. 3, p. e37229, Aug. 2022, doi: https://doi.org/10.2196/37229.
[19] Y. Peng, "Smart Home based on Kinect Gesture Recognition Technology," International Journal of Performability Engineering, 2019, doi: https://doi.org/10.23940/ijpe.19.01.p26.261269.
[20] D. Mikolajewski et al., "The Most Current Solutions using Virtual-Reality-Based Methods in Cardiac Surgery -- A Survey," Computer Science, vol. 25, no. 1, Mar. 2024, doi: https://doi.org/10.7494/csci.2024.25.1.5633.
[21] R. Filip, R. G. Puscaselu, L. Anchidin-Norocel, M. Dimian, and W. K. Savage, "Global Challenges to Public Health Care Systems during the COVID-19 Pandemic: a Review of Pandemic Measures and Problems," Journal of Personalized Medicine, vol. 12, no. 8, p. 1295, Aug. 2022, doi: https://doi.org/10.3390/jpm12081295.
[22] T. Sakai, C. Hoshino, M. Hirao, M. Nakano, Y Takashina, and A. Okawa, "Rehabilitation of Patients with Post-COVID-19 Syndrome: A Narrative Review," Progress in rehabilitation medicine, vol. 8, no. 0, p. n/a-n/a, Jan. 2023, doi: https://doi.org/10.2490/prm.20230017.
[23] B. Kesikburun et al., "The effect of comprehensive rehabilitation on post-COVID-19 syndrome," Egyptian Rheumatology and Rehabilitation, vol. 50, no. 1, Dec. 2023, doi: https://doi.org/10.1186/s43166-023-00227-4.
[24] D. Joyce, Aoife De Brún, Sophie Mulcahy Symmons, R. Fox, and É. McAuliffe, "Remote patient monitoring for COVID-19 patients: comparisons and framework for reporting," BMC Health Services Research, vol. 23, no. 1, Aug. 2023, doi: https://doi.org/10.1186/s12913-023-09526-0.
[25] H. Hawley-Hague et al., "Exploring the delivery of remote physiotherapy during the COVID-19 pandemic: UK wide service evaluation," Physiotherapy Theory and Practice, pp. 1–15, Aug. 2023, doi: https://doi.org/10.1080/09593985.2023.2247069.
[26] T. Sakai, C. Hoshino, R. Yamaguchi, M. Hirao, R. Nakahara, and A. Okawa, "Remote rehabilitation for patients with COVID-19," Journal of Rehabilitation Medicine, p. 0, 2020, doi: https://doi.org/10.2340/16501977-2731.
[27] K. Guo, M. Orban, J. Lu, M. S. Al-Quraishi, H. Yang, and M. Elsamanty, "Empowering Hand Rehabilitation with AI-Powered Gesture Recognition: A Study of an sEMG-Based System," Bioengineering, vol. 10, no. 5, p. 557, May 2023, doi: https://doi.org/10.3390/bioengineering10050557.
[28] J. Xu, L. Leng, and B.-G. Kim, "Gesture Recognition and Hand Tracking for Anti-Counterfeit Palmvein Recognition," Applied Sciences, vol. 13, no. 21, p. 11795, Jan. 2023, doi: https://doi.org/10.3390/app132111795.

[29] X. Liu, H. Shi, H. Chen, Z. Yu, X. Li, and G. Zhaoz, "iMiGUE: An Identity-free Video Dataset for Micro-Gesture Understanding and Emotion Analysis," arXiv.org, Jul. 01, 2021. https://arxiv.org/abs/2107.00285 (accessed Jun. 25, 2024).
[30] H.-B. Zhang et al., "A Comprehensive Survey of Vision-Based Human Action Recognition Methods," Sensors, vol. 19, no. 5, p. 1005, Feb. 2019, doi: https://doi.org/10.3390/s19051005.
[31] C. Ionescu, F. Li, and C. Sminchisescu, "Latent Structured Models for Human Pose Estimation," 2011. Accessed: Jun. 26, 2024. [Online]. Available: https://vision.imar.ro/human3.6m/ils_iccv11.pdf
[32] C. Wang and J. Yan, "A comprehensive survey of RGB-Based and skeleton-based human action recognition," IEEE Access, vol. 11, pp. 53880–53898, 2023, doi: https://doi.org/10.1109/ACCESS.2023.3282311.
[33] W. Zhong, W. Xiong, Y. Zhang, M. Zhang, and P. Fu, "A Spatio-Temporal Graph Convolutional Network for Gesture Recognition from High-Density Electromyography." Accessed: Jul. 01, 2024. [Online]. Available: https://arxiv.org/pdf/2312.00553
[34] C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," Journal of Big Data, vol. 6, no. 1, Jul. 2019, doi: https://doi.org/10.1186/s40537-019-0197-0.
[35] M. Tomaszewski and J. Osuchowski, "Effectiveness of Data Resampling in Mitigating Class Imbalance for Object Detection." Accessed: Jul. 02, 2024. [Online]. Available: https://ceur-ws.org/Vol-3628/paper14.pdf
[36] A. Mostafa, A. Shah, H. Chen, and Marko Savic, "The 2nd MiGA-IJCAI Challenge Track 1. Kaggle.," MiGA-IJCAI Challenge, Apr. 27, 2024. https://kaggle.com/competitions/2nd-miga-ijcai-challenge-track1 (accessed Jun. 29, 2024).
[37] S. Dimitriadis, D. Liparas, and ADNI, "How random is the random forest? Random forest algorithm on the service of structural imaging biomarkers for Alzheimer's disease: from Alzheimer's disease neuroimaging initiative (ADNI) database," Neural Regeneration Research, vol. 13, no. 6, p. 962, 2018, doi: https://doi.org/10.4103/1673-5374.233433.
[38] F. Hutter, H. Hoos, and K. Leyton-Brown, "Sequential Model-Based Optimization for General Algorithm Configuration." Available: https://ml.informatik.uni-freiburg.de/wp-content/uploads/papers/11-LION5-SMAC.pdf
[39] R. Yacouby and D. Axman, "Probabilistic Extension of Precision, Recall, and F1 Score for More Thorough Evaluation of Classification Models," ACLWeb, Nov. 01, 2020. https://aclanthology.org/2020.eval4nlp-1.9/ (accessed May 13, 2022).
[40] Ž. Đ. Vujovic, "Classification Model Evaluation Metrics," International Journal of Advanced Computer Science and Applications, vol. 12, no. 6, 2021, doi: https://doi.org/10.14569/ijacsa.2021.0120670.
[41] M. Tomaszewski, R. Gasz, S. S. Kasana, J. Osuchowski, S. Singh, and S. Zator, "TCIP: Transformed Colour Intensity Profiles analysis for fault detection in power line insulators," Multimedia tools and applications, Mar. 2024, doi: https://doi.org/10.1007/s11042-024-18901-w.