# On Reproducibility of Graph Neural Network for Facial Palsy and Paresis Assessment: Effects of Pose Variability in Dataset

Zolbayar Shagdar[1,2], Seyed Ali Amirshahi[1], and Kiran Raja[1]

[1] Department of Computer Science, Norwegian University of Science and Technology, 2815 Gjøvik, Norway
{zolbayar.shagdar, s.ali.amirshahi, kiran.raja}@ntnu.no
[2] Department of Research, Innlandet Hospital Trust, 2381 Brumunddal, Norway

**Abstract.** Reproducibility in terms of implementation and performance is a crucial aspect of healthcare applications as they can have high impact consequences on patient welfare and safety. In this paper, we focus on the consistency and reproduction of results for graph neural networks (GNN) based facial palsy and paresis evaluation. Comparative studies between our proposed GNN-based model and state-of-the-art (SOTA) convolutional neural network-based models suggest that the GNN model is sensitive to pose variability within the dataset while the CNN-based models are consistent across the board. With these findings, we propose a sufficiently regularised dataset with pose variability for obtaining consistent and better results. We provide further analysis of the classification behaviour of our model, the results of which suggest potential label ambiguity within the dataset employed. Future improvements regarding the model's performance and consistency are recommended based on the reproducibility analyses.

**Keywords:** Face analysis · graph neural networks · data variability · facial palsy · reproducibility

## 1 Introduction

With the recent advances in artificial intelligence and computer vision, computer-aided diagnostics assistance tools are gaining ground in the healthcare sector [1]. As these tools deal with critical factors like patient well-being and diagnostic accuracy, their aspects of reliability and reproducibility are of utmost importance. Several studies have investigated approaches to automatically detect or evaluate facial palsy [2, 3]. A number of these approaches either design or extract descriptive features from the facial image and feed them into a classifier. Local binary pattern-based features [4], hand-crafted asymmetry features [5, 6], and histogram of oriented gradients [7] have been used with a Support Vector Machine (SVM) [8] classifier for the final facial palsy assessment. k-Nearest Neighbour classifier, random forest, and Linear Discriminant Analysis have been used earlier [9, 10, 11] as well. Researchers also investigated Convolutional Neural Networks (CNNs)

Table 1: Binary classification performance of our proposed model on the Toronto NeuroFace dataset, compared with three state-of-the-art CNN-based methods.

| Metric | Guo et al. [14] | Sajid et al. [17] | Yu et al. [18] | GNN (Our) |
|---|---|---|---|---|
| Accuracy | 0.58 ± 0.08 | 0.57 ± 0.15 | 0.47 ± 0.11 | 0.57 ± 0.09 |
| F2 score | 0.61 ± 0.17 | 0.68 ± 0.11 | 0.60 ± 0.09 | 0.64 ± 0.16 |

for automated facial palsy assessment [12], exploring multiple architectures such as GoogLeNet [13, 14], Darknet [3, 15], VGG-16 [16, 17], and ResNet-34 [4, 18]. Dedicated architectures have been further proposed for the same application [19, 20, 21]. In a similar line, 3-dimensional CNNs [22], as well as Long Short-Term Memory (LSTM) and similar architectures have been employed successfully to leverage temporal data [18, 19, 21].

In a complementary direction, Graph Neural Networks (GNN) have been proposed to deal with certain limitations faced by the current literature [23]. Firstly, the vast majority of the existing studies utilise datasets that are not publicly available, making it impractical to reproduce the results. Furthermore, these recent models do not perform as reported when implemented and tested on publicly available datasets, such as the results presented in Table 1. Secondly, data collection for facial pathology has a plethora of associated complications ranging from the inherent scarcity of such data to patient privacy concerns. This causes the datasets, whether publicly available or not, to generally be small in size. These limitations could potentially be causing the existing computer vision-based methods to have shortcomings such as a tendency to overfit, patient identity memorisation, sensitivity to background information, and so forth. Using facial graphs and GNNs lets us refrain from these issues by processing facial shape and structure only while disregarding unnecessary features including background and colour information.

Shagdar et al. [23] specifically in their work extract the 478 expressive landmarks proposed by Kartynnik et al. [24] that capture the facial details. A Delaunay triangulation [25] is then performed to construct facial graphs with the 478 landmarks as nodes and the triangulation sides as graph edges. The graphs are then fed into a GNN model consisting of 5 graph convolutional layers [26] and two fully connected layers to classify between frames displaying facial pathology against those of healthy individuals.

Intrigued by the performance reported by Shagdar et al. [23], in this work, we take a detailed look at the reproducibility of GNN for facial palsy and paresis assessment. As presented in Table 1, we note that the GNN model performs on par with recently proposed CNN-based models when performing the stroke vs. healthy binary classification task on the Toronto NeuroFace (TNF) dataset [27]. However, we dive further to reproduce the performance of GNNs in this work. As noted in Figure 1, we first note the sensitiveness of the GNNs across different kinds of data in producing consistent and thereby reproducible results. We thus

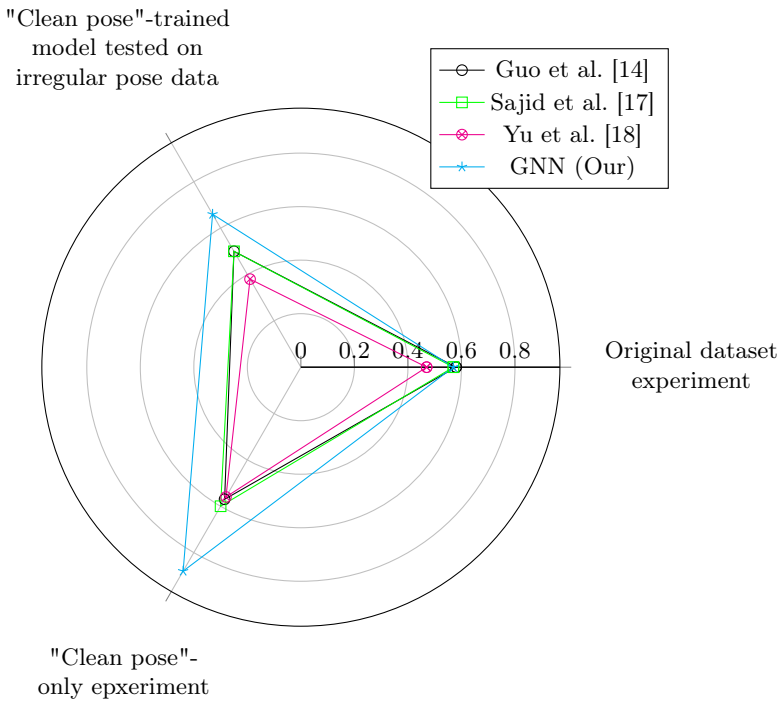



Fig. 1: Web chart visualisation of the GNN model and existing CNN models' classification accuracies in three different experiment settings. The GNN model's performance improves drastically when trained on "clean pose" data.

investigate the factors of reproducibility to provide consistent results by carefully analyzing the performances of GNNs on the TNF dataset [27] dataset.

We first present the settings of our study by discussing the datasets and reporting the settings of various experiments. Then, we present our observations and discuss the findings.

## 2 Dataset and Implementation

### 2.1 Dataset

We employ the Toronto NeuroFace (TNF) dataset [27] for all our experiments. The dataset contains videos of 11 amyotrophic lateral sclerosis (ALS) patients, 14 post-stroke patients, and 11 healthy control subjects performing various facial actions and gestures. There are 261 videos consisting of 3306 frames in total, from which we exclude the ALS patients and work with the remaining 2386 frames in this work. Additionally, two speech-language pathologists rated the videos in

terms of severity of facial symptoms. Scores of one to five are given to the videos according to five categories: symmetry, range-of-motion, speed, variability, and fatigue of facial movements, adding up to a score between 0 to 25.

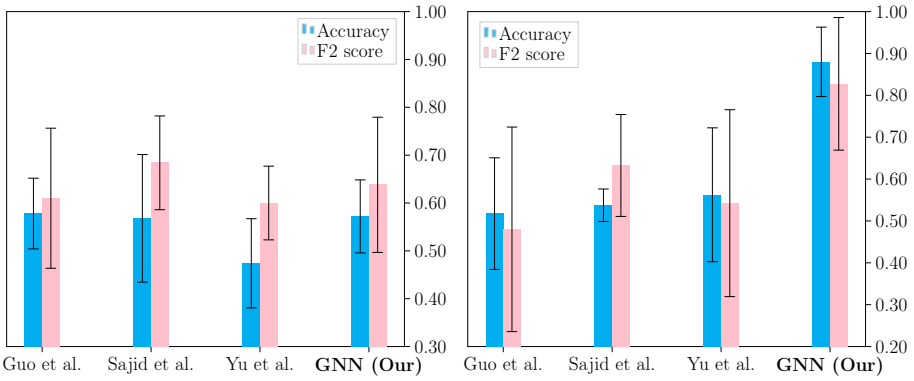
### 2.2 Implementation

The experiment codes and instructions for reproducing our results are released at the following repo: `https://github.com/zo456/gnn-facial-palsy`.

We provide instructions on reproducing the training and evaluation process within the repo. The random seeds used to split the dataset and initialise the model are fixed as customisable variables for reproducibility. Python 3.8 [28] is used throughout the framework and necessary dependencies need to be installed following: `https://github.com/zo456/gnn-facial-palsy/blob/master/requirements.txt`.

## 3 Consistent and Reproducible Results

The universally poor results by CNNs and the proposed GNN, as displayed in Table 1, prompted us to explore potential causes rooted in the dataset rather than the models. Dataset inconsistencies can come in various forms including, but not limited to, data element quality, label accuracy, and label ambiguity.




(a) Barplot: original dataset experiment (b) Barplot: "Clean pose" experiment

Fig. 2: Bar plot visualisations of the experiment results showing the performance increase of our proposed model associated with pose regularity in the dataset. The model accuracies are shown with blue bars, and the F2 scores are shown with pink bars, each with their 95% confidence intervals.

The frames in the TNF dataset have constant dimensions and are fairly regularised in terms of face location within the image frame. However, some subjects

Table 2: Binary classification performance of our proposed model compared with three recent CNN-based methods on three different experiment types with regards to pose irregularity in the dataset.

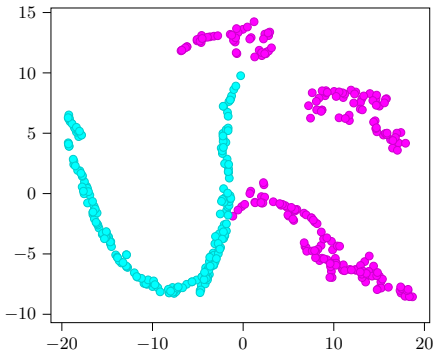
| Training dataset | | Mixed pose | Clean pose | Clean pose |
|---|---|---|---|---|
| Test dataset | | Mixed pose | Bad pose | Clean pose |
| Guo et al. [14] | Accuracy | 0.58 ± 0.08 | 0.50 ± 0.21 | 0.52 ± 0.14 |
| | F2 score | 0.61 ± 0.17 | 0.40 ± 0.32 | 0.48 ± 0.25 |
| Sajid et al. [17] | Accuracy | 0.57 ± 0.15 | 0.50 ± 0.05 | 0.54 ± 0.04 |
| | F2 score | 0.68 ± 0.11 | 0.58 ± 0.07 | 0.63 ± 0.12 |
| Yu et al. [18] | Accuracy | 0.47 ± 0.11 | 0.38 ± 0.07 | 0.56 ± 0.16 |
| | F2 score | 0.60 ± 0.09 | 0.37 ± 0.13 | 0.54 ± 0.23 |
| **GNN (Our)** | Accuracy | 0.57 ± 0.09 | 0.66 ± 0.02 | **0.88 ± 0.08** |
| | F2 score | 0.64 ± 0.16 | 0.59 ± 0.03 | **0.83 ± 0.16** |

had their whole frame sequence recorded with a slouched or tilted position such that their faces were not looking straight at the camera. Out of the 14 subjects with stroke and 11 subjects in the healthy control group, 8 subjects in each category faced the camera straight on while the remaining subjects had various pose irregularities (e.g., tilted head, not facing the camera, etc.). From here on, we refer to the subset consisting of the 16 subjects with straight, regular poses as the "clean pose" dataset.

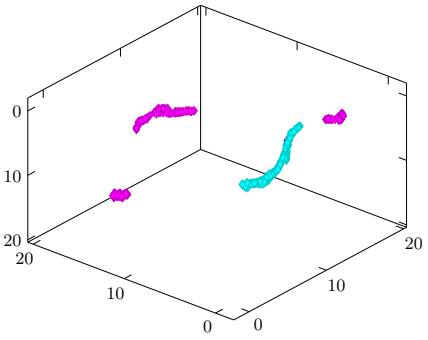
We ran two additional experiments to further explore the effects of pose variability on classification results. Firstly, the models are trained and tested on the "clean pose" dataset to check if the classification performances improve with the exclusion of pose variability in the dataset. Secondly, the models trained on the "clean pose" dataset are tested only on the frames with pose irregularities, that is the rest of the TNF dataset. Together with the original experiment, all three results are presented in Table 2.

Firstly, we can see from Table 2 and Figure 2 that the CNN-based classification models show consistent performance across the board. In other words, the palsy classification results from the CNN-based models are reproducible, albeit poor, regardless of the pose variabilities within the training and test dataset. Going further, the classification performance of our proposed GNN model improves significantly when dealing only with "clean pose" data elements, as illustrated with bar plots in Figure 2. On the other hand, the GNN model performance drops when pose variability is introduced at any stage (training or test) of the experiment. We can conclude here that the classification result reproducibility of our GNN model is closely dependent on the dataset quality, especially the pose irregularities within the dataset.

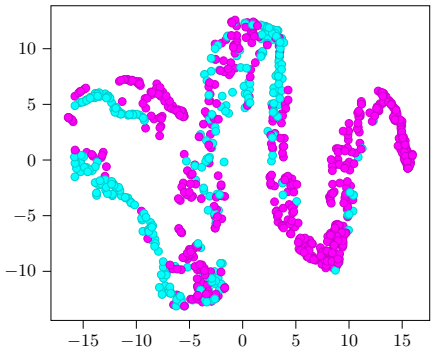
Figure 3 provides the t-SNE [29] and UMAP [30] plots of the features extracted (right before the final classifier layer) by our GNN model. The perfor-



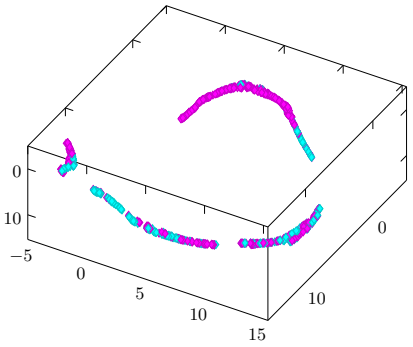

(a) "Clean pose" experiment: t-SNE plot.



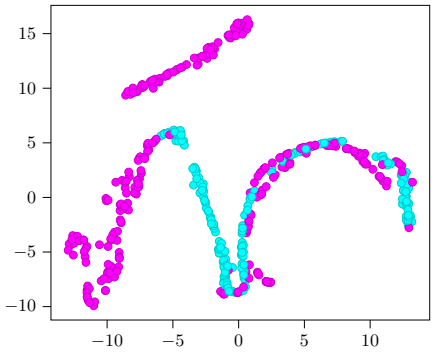

(b) "Clean pose" experiment: UMAP plot.



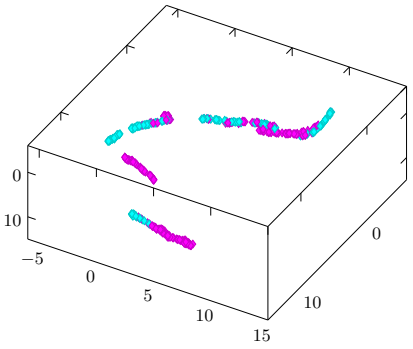

(c) "Clean pose" model tested on irregular pose data: t-SNE plot.




(d) "Clean pose" model tested on irregular pose data: UMAP plot.




(e) Original dataset experiment: t-SNE plot.




(f) Original dataset experiment: UMAP plot.

Fig. 3: t-SNE and UMAP plots derived from the features at the penultimate layer of our GNN model. Healthy control data points are coloured in magenta and palsy data points are in cyan.

mance drop associated with the introduction of pose variability is evident in Figures 3c-3f. On the contrary, the features extracted at inference on the "clean pose" dataset, by our model trained on "clean pose" dataset, are clearly separable for high classification performance, as visualised in Figure 3a and Figure 3b.

### 3.1 GNN classification behaviour

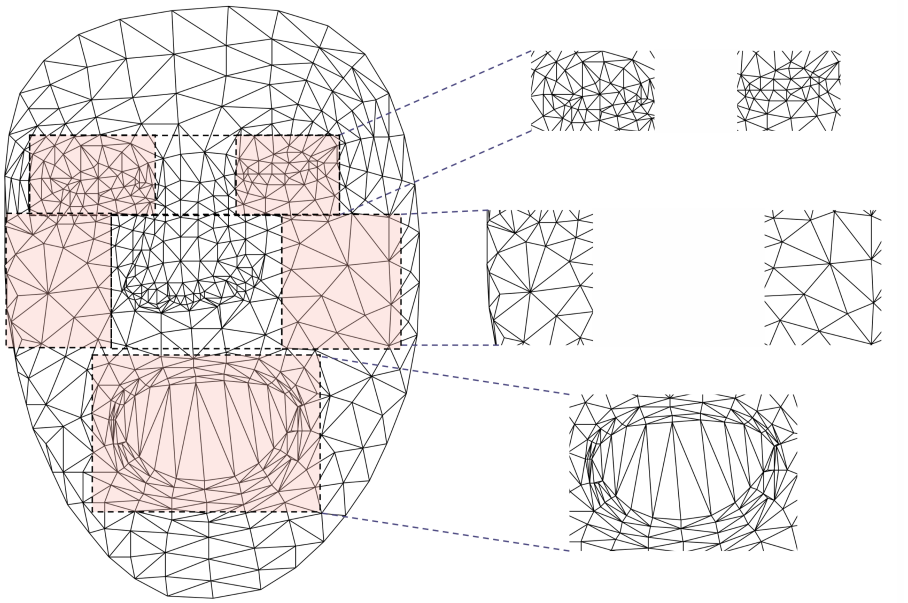
Through the results discussed, it can be noted that the performance of the proposed GNN model is dependent on pose variability in the dataset. Furthermore, in this section, we looked at the classification behaviour of our model when trained and tested solely on "clean pose" data. This could give further insights into the classification result reproducibility by analysing the misclassification characteristics.

Figure 4a illustrates an example face graph that is correctly classified to show facial palsy, in this case, stroke (obtained from subject "S012" of the TNF dataset). As seen from the extracted parts in Figure 4a, there are noticeable differences between the graph regions corresponding to the left and right eyes and cheeks. Furthermore, the subgraph corresponding to the mouth region also has visible asymmetries. On the flip side, the graphs that would correctly be classified as belonging to the healthy control group would naturally be symmetric.

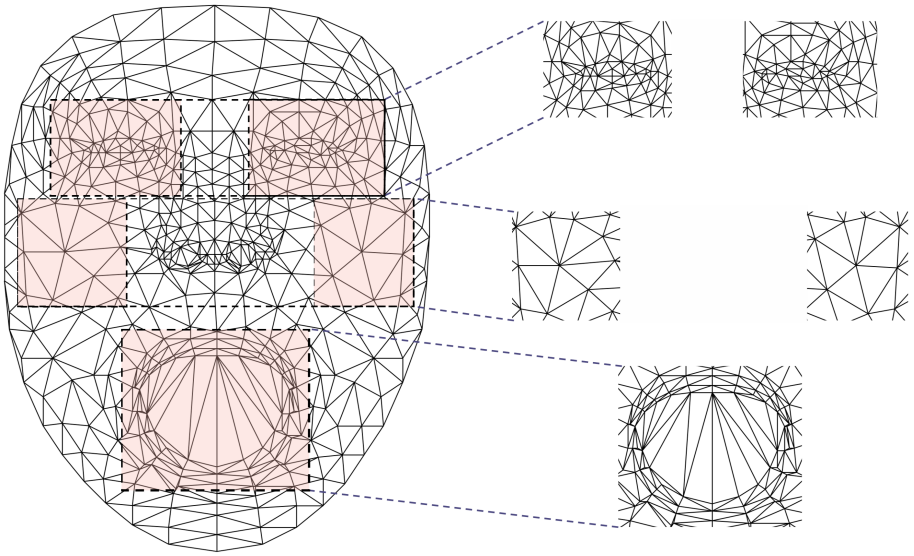
Upon closer examination of the misclassification cases, we found that the reason why the performance of our model comes down to 88% and not higher when operating on the "clean pose" dataset is largely due to a single individual. Namely, all the data collected from a certain individual is consistently misclassified as healthy with our model, leading to the reported overall performance. We provide in Figure 4b an example graph generated from that misclassified individual (subject "OP03" of the TNF dataset) in a frame performing the same action as in Figure 4a being performed. In contrast with the description of the graph in Figure 4a, the graph in Figure 4b is highly symmetric. In detail, the graph has particularly similar eye and cheek regions for the left and right sides, and a symmetric subgraph corresponding to the mouth region, all of which is uncharacteristic for a palsy/stroke-positive case. This is further backed by the original facial dysfunction severity scores given by speech-language pathologists, provided with the TNF dataset. To elaborate, the two specialists gave average dysfunction severity scores of 6.2 and 5.6 for this misclassified individual from a range of 0-25. For comparison, the total averages for the remaining individuals were 9.4 and 8.6, with severe cases having scores as high as 11-12. This finding suggests the possible presence of ambiguous data elements and labels in the TNF dataset. Moreover, it hints at room for improvement regarding the GNN model on the classification of difficult-to-distinguish cases.

## 4 Conclusion

The reproducibility of artificial intelligence-based tools in healthcare is an important factor with significant effects on patient care. With this paper, we report on the reproducibility of results and associated analyses of our proposed



(a) A correctly classified facial graph displaying a palsy-positive face.



(b) A facial graph from a palsy-positive patient wrongly predicted to be healthy.

Fig. 4: Example facial graphs from facial palsy-positive individuals performing the same action. The upper image (obtained from subject "S012" of the TNF dataset) is correctly classified with our model while the lower image is generated from a subject (subject "S012" of the TNF dataset) who is consistently misclassified as healthy.

graph neural network-based framework for facial palsy and paresis evaluation. The results of this study suggest that our method has the potential to outperform significantly the state-of-the-art when dealing with highly regularised "clean" data. Yet, the performance of our model varies drastically with the introduction of pose variability in the dataset. Whereas the existing convolutional neural network-based models show a consistent, albeit poor, performance regardless of the data elements' pose variability. Furthermore, a closer inspection of the classification behaviour of our model suggests the presence of difficult-to-distinguish data elements. Future improvements for the model should focus on dealing with these cases. Beyond that, exploring more advanced graph neural networks utilising complex architectures and state-of-the-art GNN layer types is highly encouraged.

**Acknowledgments.** Portions of the research in this paper uses the Toronto Neuro-Face Database collected by Dr. Yana Yunusova and the Vocal Tract Visualization and Bulbar Function Lab teams at UHN-Toronto Rehabilitation Institute and Sunnybrook Research Institute respectively, financially supported by the Michael J. Fox Foundation, NIH-NIDCD, Natural Sciences and Engineering Research Council, Heart and Stroke Foundation Canadian Partnership for Stroke Recovery and AGEWELL NCE.

Seyed Ali Amirshahi was supported by the project "VQ4MedicS: Video Quality Assessment and Enhancement for Pre-Hospital Medical Services" (grant number 329034, approved on 1 September 2021) from the research council of Norway.

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## References

[1] F. Jiang *et al.*, "Artificial intelligence in healthcare: Past, present and future," *Stroke and vascular neurology*, vol. 2, no. 4, 2017.

[2] Y. Zhuang *et al.*, "Facial weakness analysis and quantification of static images," *IEEE journal of biomedical and health informatics*, vol. 24, no. 8, pp. 2260–2267, 2020.

[3] G.-S. J. Hsu, J.-H. Kang, and W.-F. Huang, "Deep hierarchical network with line segment learning for quantitative analysis of facial palsy," *IEEE Access*, vol. 7, pp. 4833–4842, 2018.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[5] T. Wang, S. Zhang, J. Dong, L. Liu, and H. Yu, "Automatic evaluation of the degree of facial nerve paralysis," *Multimedia Tools and Applications*, vol. 75, no. 19, pp. 11 893–11 908, 2016.

[6] T. H. Ngo, M. Seo, Y.-W. Chen, and N. Matsushiro, "Quantitative assessment of facial paralysis using local binary patterns and gabor filters," in *Proceedings of the 5th Symposium on Information and Communication Technology*, 2014, pp. 155–161.

[7] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, Ieee, vol. 1, 2005, pp. 886–893.

[8] V. Vapnik, *The nature of statistical learning theory*. Springer science & business media, 2013.

[9] W. Kaewmahanin *et al.*, "Automatic facial asymmetry analysis for elderly stroke detection by using cosine similarity," in *2022 19th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, IEEE, 2022, pp. 1–4.

[10] Y. Zhuang *et al.*, "Pathological facial weakness detection using computational image analysis," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, IEEE, 2018, pp. 261–264.

[11] I. Chang, C.-Y. Low, S. Choi, and A. B.-J. Teoh, "Kernel deep regression network for touch-stroke dynamics authentication," *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 1109–1113, 2018.

[12] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep learning for computer vision: A brief review," *Computational intelligence and neuroscience*, vol. 2018, 2018.

[13] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[14] Z. Guo *et al.*, "Deep assessment process: Objective assessment process for unilateral peripheral facial paralysis via deep convolutional neural network," in *2017 IEEE 14th international symposium on biomedical imaging (ISBI 2017)*, IEEE, 2017, pp. 135–138.

[15] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7263–7271.

[16] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[17] M. Sajid, T. Shafique, M. J. A. Baig, I. Riaz, S. Amin, and S. Manzoor, "Automatic grading of palsy using asymmetrical facial features: A study complemented by new solutions," *Symmetry*, vol. 10, no. 7, p. 242, 2018.

[18] M. Yu *et al.*, "Toward rapid stroke diagnosis with multimodal deep learning," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III 23*, Springer, 2020, pp. 616–626.

[19] X. Liu, Y. Xia, H. Yu, J. Dong, M. Jian, and T. D. Pham, "Region based parallel hierarchy convolutional neural network for automatic facial nerve paralysis evaluation," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, no. 10, pp. 2325–2332, 2020.

[20] T. Cai *et al.*, "Deepstroke: An efficient stroke screening framework for emergency rooms with multimodal adversarial deep learning," *Medical Image Analysis*, vol. 80, p. 102 522, 2022.

[21] P. Xu *et al.*, "Automatic evaluation of facial nerve paralysis by dual-path lstm with deep differentiated network," *Neurocomputing*, vol. 388, pp. 70–77, 2020.

[22] G. Storey, R. Jiang, S. Keogh, A. Bouridane, and C.-T. Li, "3dpalsynet: A facial palsy grading and motion recognition framework using fully 3d convolutional neural networks," *IEEE access*, vol. 7, pp. 121 655–121 664, 2019.

[23] Z. Shagdar, S. A. Amirshahi, and K. Raja, "A graph neural network for facial palsy and paresis," in *The 3rd International Workshop on Pattern Recognition in Healthcare Analytics at ICPR-2024*, 2024.

[24] Y. Kartynnik, A. Ablavatski, I. Grishchenko, and M. Grundmann, "Real-time facial surface geometry from monocular video on mobile gpus," *arXiv preprint arXiv:1907.06724*, 2019.

[25] B. Delaunay, "Sur la sphère vide," French, *Bulletin de l'Académie des Sciences de l'URSS. Classe des sciences mathématiques et na*, vol. 1934, no. 6, pp. 793–800, 1934.

[26] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[27] A. Bandini *et al.*, "A new dataset for facial motion analysis in individuals with neurological disorders," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 4, pp. 1111–1119, 2020.

[28] G. Van Rossum and F. L. Drake, *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009, ISBN: 1441412697.

[29] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008.

[30] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv: 1802.03426*, 2018.