Automated Classification of Nerve Integrity Monitoring Signals Using a Convolutional Neural Network for Hypoglossal Nerve Stimulator Placement

Vir Patel[1], Dylan Yoon BS[2], Nathan Rasmussen BS[3], Yashnoor Sandhu BS[2], Zachary Whong BS[2], Ryan Nord MD[3]

[1]Mills E. Godwin High School; [2]Virginia Commonwealth University School of Medicine; [3]Virginia Commonwealth University Health, Department of Otolaryngology – Head and Neck Surgery

Introduction

- Hypoglossal nerve stimulation (HGNS) treats obstructive sleep apnea by stimulating nerve branches that control the tongue to keep the airway open. The stimulation cuff must be appropriately placed to stimulate inclusion branches (tongue protrusion) and avoid stimulating exclusion branches (tongue retraction).[1]
- Nerve integrity monitoring (NIM) records the muscle (EMG) response to each branch. These signals can be categorized as inclusion, exclusion, or mixed (both).[1]
- Mixed responses are common and can be difficult to interpret.
- This study aims to determine if an AI image classifier, a convolutional neural network (CNN), can be trained to read and interpret these signals.[2-4]

Methods

- Design: 312 NIM screenshots → split by patient into 5 groups → AI trained on 4 groups and tested on the 5th → repeated 5 times. Every patient was tested by an AI that had never seen them.
- Data: 312 intraoperative screenshots from 33 HGNS patients, categorized by the surgical team as inclusion (n=101), mixed (n=73), or exclusion (n=138). Images were padded to square and resized to 224×224 pixels.
- Model: ResNet18, a CNN pretrained on ImageNet, fine-tuned with class-balanced sampling. Predictions were averaged over 4 modified views of each image (test-time augmentation).
- Comparator: classical machine-learning baseline (HOG features with Random Forest, SVM, and Logistic Regression) on a single patient-level 80/20 split.
- Outcomes: accuracy, precision, recall, and F1.

Figures/Tables

True label	Predicted label: Inclusion	Mixed	Exclusion	Recall	F1
Inclusion (n = 101)	**79**	16	6	0.78	0.77
Mixed (n = 73)	19	**42**	12	0.58	0.60
Exclusion (n = 138)	7	9	**122**	0.88	0.88
Precision	0.75	0.63	0.87	**Accuracy 77.9%**	

Table 1. The AI rarely confused inclusion with exclusion; most errors involved mixed signals. Rows show the true signal type and columns show the AI's call, so the colored diagonal is correct. Recall = share of true signals found. Precision = share of the AI's calls that were correct. F1 = combined score (1.0 = perfect). Outlined: mixed called inclusion (19 of 73). Each patient was tested only by a model trained without that patient (n=312).

Approach	How it was tested	Accuracy	Range
Classical machine learning (HOG features + Random Forest / SVM / Logistic Regression)	One 80/20 patient split	**65%**	–
CNN, class-weighted loss	5 random 80/20 splits (mean)	**71.6%**	67.7 - 79.0%
CNN, balanced sampling + deeper fine-tuning	5 random 80/20 splits (mean)	**70.3%**	66.1 - 75.8%
CNN, 5-fold cross-validation + test-time augmentation	**Out-of-fold, all images**	**77.9%**	**69.8 - 87.3% by fold**
CNN, focal loss + exact-duplicate removal	Out-of-fold, 306 images	**77.1%**	76.7 - 83.6% by fold

Table 2. Accuracy of each approach. Single-split results (rows 1-3) are not directly comparable to the cross-validated results (rows 4-5), which test every image once.

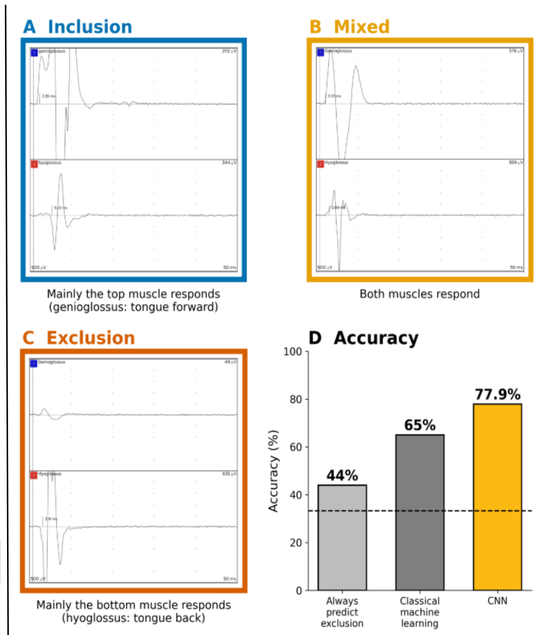


Figure 1. The AI correctly classified about 78 of every 100 nerve monitor signals. (A-C) Example signals. Each panel shows two muscles: genioglossus (top), which pushes the tongue forward, and hyoglossus (bottom), which pulls it back. Inclusion: mainly genioglossus responds. Mixed: both respond. Exclusion: mainly hyoglossus responds. (D) Accuracy of the CNN, a classical machine-learning method, and an always-predict-exclusion rule (n=312 images, 33 patients). Dashed line = chance (33%).

References

1. Sturm JJ, Modik O, Suurna MV. Neurophysiological monitoring of tongue muscle activation during hypoglossal nerve stimulation. The Laryngoscope. 2020 Jul;130(7):1836-1843. DOI: 10.1002/lary.28341. PMID: 31742709.
2. Zha X, Wehbe L, Sclabassi RJ, et al. A Deep Learning Model for Automated Classification of Intraoperative Continuous EMG. IEEE Transactions on Medical Robotics and Bionics. 2021 Feb;3(1):44-52. DOI: 10.1109/tmrb.2020.3048255. PMID: 33997657; PMCID: PMC8117925.
3. Alapati R, Renslo B, Jackson L, Moradi H, Oliver JR, Chowdhury M, Vyas T, Bon Nieves A, Lawrence A, Wagoner SF, Rouse D, Larsen CG, Wang G, Bur AM. Predicting Therapeutic Response to Hypoglossal Nerve Stimulation Using Deep Learning. Laryngoscope. 2024 Dec;134(12):5210-5216. doi: 10.1002/lary.31609. Epub 2024 Jun 27. PMID: 38934474; PMCID: PMC11563902.
4. Masood F, Sharma M, Mand D, Nesathurai S, Simmons HA, Brunner K, Schalk DR, Sledge JB, Abdullah HA. A Novel Application of Deep Learning (Convolutional Neural Network) for Traumatic Spinal Cord Injury Classification Using Automatically Learned Features of EMG Signal. Sensors (Basel). 2022 Nov 3;22(21):8455. doi: 10.3390/s22218455. PMID: 36366153; PMCID: PMC9657335.

Results

- About 78 of every 100 signals were classified correctly, versus 65 for a classical machine-learning method and 33 expected by chance.
- Out-of-fold accuracy was 77.9% (243/312). An always-predict-exclusion rule scored 44% (Fig. 1D).
- F1 was 0.88 for exclusion, 0.77 for inclusion, and 0.60 for mixed (Table 1).
- The two clinically opposite types were rarely confused: 13 of 239 inclusion and exclusion signals (5%). Most of the errors involved mixed signals (56/69, 81.2%).
- 26% of true mixed signals were incorrectly interpreted as inclusion signals, and 16% were incorrectly interpreted as exclusion signals. 16% of inclusion signals were interpreted as mixed signals.
- Accuracy ranged from 70% to 87% across test groups, reflecting the small dataset.
- Focal loss and exact-duplicate removal did not improve performance (77.1%).

Conclusion

- A convolutional neural network AI model can be trained to recognize NIM signal types from a small dataset, supporting the feasibility of automated decision support during HGNS placement.
- The model struggled most with mixed signals (recall 58%) which was the least common signal in the training set (73 of 312 screenshots).
- Limitations include the low number of mixed signals (73) and a relatively small dataset (33 patients and 312 screenshots) compared to high performing neural networks.
- Potential future uses include a second-opinion display during branch testing, and a training aid for surgeons learning NIM interpretation.

Presenter: Vir Patel
Department Otolaryngology – Head and Neck Surgery, VCU Health