










## Research

# Investigation of an Open AI Model in the Analysis of STN Microelectrode Recordings: Consistency With Clinicians and Potential for DBS Targeting

STN Analizinde Açık Bir Yapay Zeka Modelinin Araştırılması Mikroelettrot Kayıtları: Klinisyenlerle Tutarlılık ve DBS Hedefleme Potansiyeli

 Ozan Haşımoğlu<sup>1</sup>,  Ayça Altinkaya<sup>2</sup>,  Taha Hanoğlu<sup>1</sup>,  Tuba Özge Karaçoban<sup>1</sup>,  Nur Bahar Geylan<sup>1</sup>,  
 Fırat Demir<sup>1</sup>,  Bekir Tuğcu<sup>1</sup>

<sup>1</sup>University of Health Sciences Türkiye, Başakşehir Çam and Sakura City Hospital, Clinic of Neurosurgery, İstanbul, Türkiye

<sup>2</sup>University of Health Sciences Türkiye, Başakşehir Çam and Sakura City Hospital, Clinic of Neurology, İstanbul, Türkiye

### ABSTRACT

**Objective:** Deep brain stimulation (DBS) of the subthalamic nucleus (STN) requires precise electrode placement, often assisted by microelectrode recording (MER). However, the interpretation of MER remains highly subjective, varying among clinicians based on experience. This study evaluates the ability of an artificial intelligence (AI) model (ChatGPT 4.0) to classify STN MER recordings and evaluate its consistency with experienced and less experienced clinicians.

**Methods:** A total of 32 STN MER recordings were independently evaluated by two experienced clinicians, two less experienced clinicians, and the AI model. Classifications were assigned to artifact, thalamus, silent, STN, suspicious STN, substantia nigra, and N/A (no recording), categories. Fleiss' Kappa was used to assess inter-rater consistency, while Cohen's Kappa measured agreement between generative pre-trained transformer (GPT) and each clinician. Additionally, precision and recall were calculated for each category.

**Results:** The overall Fleiss' Kappa among all evaluators was 0.544, with higher agreement among experienced clinicians (0.738) compared to less experienced ones (0.631). GPT showed low agreement with both groups, with Cohen's Kappa values ranging from 0.341 to 0.375. GPT demonstrated the highest accuracy in detecting STN (73.47%), but its performance was significantly lower for other categories. Within-category consistency (14.28%) indicated variability in transition zones, with a misclassification rate of 45.87% compared to the majority opinion of clinicians.

**Conclusion:** While GPT exhibited partial consistency with clinicians in identifying the STN, its reliability in classifying transition zones and adjacent structures was low. For AI to serve as a reliable tool in STN targeting, further refinement of its algorithms and expanded training datasets is necessary. Although GPT is not yet suitable for clinical decision making, its potential for future DBS applications is promising.

**Keywords:** Deep brain stimulation, subthalamic nucleus, microelectrode recordings, artificial intelligence, machine learning, ChatGPT

### ÖZ

**Amaç:** Subtalamik nükleusun (STN) derin beyin stimülasyonu (DBS), genellikle mikroelettrot kayıtları (MER) ile desteklenen hassas elektrot yerleşimi gerektirir. Ancak, MER'nin yorumlanması oldukça öznel ve klinisyenler arasında deneyime bağlı olarak değişkenlik gösterebilir. Bu çalışma, bir yapay zeka (YZ) modelinin (ChatGPT 4.0) STN MER kayıtlarını sınıflandırma yetisini değerlendirerek, deneyimli ve daha az deneyimli klinisyenlerle tutarlılığını analiz etmektedir.

**Gereç ve Yöntem:** Toplam 32 STN MER kaydı, iki deneyimli klinisyen, iki daha az deneyimli klinisyen ve YZ modeli tarafından bağımsız olarak değerlendirildi. Kayıtlar artefakt, talamus, sessiz, STN, şüpheli STN, substantia nigra ve N/A (kayıt yok) kategorilerine ayrıldı. Değerlendiriciler arasındaki tutarlılık Fleiss' Kappa, üretici önceden eğitilmiş dönüştürücü (GPT) ile her bir klinisyen arasındaki uyum ise Cohen's Kappa ile ölçüldü. Ayrıca, her kategori için kesinlik ve duyarlılık hesaplandı.

**Bulgular:** Tüm değerlendiriciler arasında genel Fleiss' Kappa 0,544 olarak bulundu; deneyimli klinisyenler arasında tutarlılık 0,738, daha az deneyimli klinisyenler arasında ise 0,631 idi. GPT'nin her iki grup ile uyumu düşük olup, Cohen's Kappa 0,341 ile 0,375 arasında değişti. GPT,

**Address for Correspondence:** Ozan Haşımoğlu, MD, University of Health Sciences Türkiye, Başakşehir Çam and Sakura City Hospital, Clinic of Neurosurgery, İstanbul, Türkiye  
**E-mail:** ozanhasim@hotmail.com **ORCID ID:** orcid.org/0000-0003-1394-5188

**Cite as:** Haşımoğlu O, Altinkaya A, Hanoğlu T, Karaçoban TÖ, Geylan NB, Demir F, et al. Investigation of an Open AI model in the analysis of STN microelectrode recordings: consistency with clinicians and potential for DBS targeting. Med J Bakirkoy. 2025;21(3):288-295

**Received:** 05.02.2025

**Accepted:** 08.04.2025

**Publication Date:** 03.09.2025

STN tespitinde en yüksek doğruluğu (%73,47) gösterse de, diğer kategorilerde performansı belirgin şekilde daha düşüktü. Kategori içi tutarlılığı %14,28 olarak belirlenen modelin, geçiş bölgelerindeki değişkenliği yüksek olup, klinisyenlerin çoğunluk görüşüne göre yanlış sınıflandırma oranı %45,87 idi.

**Sonuç:** GPT, STN'yi tanımlamada klinisyenlerle kısmi bir tutarlılık gösterse de, geçiş bölgeleri ve komşu yapıları sınıflandırmadaki güvenilirliği düşüktü. YZ'nin STN hedeflemede güvenilir bir araç olabilmesi için algoritmalarının iyileştirilmesi ve eğitim veri setlerinin genişletilmesi gereklidir. GPT henüz klinik karar vermeye uygun olmasa da, gelecekteki DBS uygulamaları için umut vadetmektedir.

**Anahtar Kelimeler:** Derin beyin stimülasyonu, subtalamik nükleus, mikroelektrot kaydı, yapay zeka, makine öğrenmesi, ChatGPT

## INTRODUCTION

The subthalamic nucleus (STN) is one of the most frequently targeted structures in deep brain stimulation (DBS) for Parkinson's disease (1). Given that the accuracy of electrode placement in DBS surgery directly impacts clinical outcomes, microelectrode recordings (MER) obtained intraoperatively serve as a crucial tool for delineating the boundaries of the STN (2-4). The interpretation of MER data relies on clinical expertise to distinguish the STN from surrounding structures (5-7). However, precisely defining the exact boundaries of the STN and ensuring consistency among different evaluators remain challenging (8). The interpretation of MER recordings is inherently subjective, often yielding varying results when assessed by different clinicians. These discrepancies can arise due to differences in evaluator experience and anatomical variations in the STN among individuals (9). In particular, identifying the onset and termination points of the STN, as well as accurately interpreting transition zones involving silent areas or structures such as the thalamus and substantia nigra, presents a significant challenge (10). Therefore, it is crucial to develop methods that enhance consistency among clinicians and make the decision-making process more objective.

In recent years, artificial intelligence (AI) and machine learning-based models have been increasingly utilized as supportive tools in the analysis of neurophysiological recordings. The automated classification of MER recordings can both accelerate the surgical process by saving time and provide a more objective analytical approach by minimizing human-induced variations in interpretation (11,12). However, the reliability of AI models and their consistency with clinicians remain subjects of ongoing debate (2,8,9).

The aim of this study is to compare how four clinicians (two experienced and two less experienced) and the OpenAI ChatGPT 4.0 model evaluate STN MER recordings, and to statistically analyze the consistency of their interpretations. The study examines the agreement among clinicians with different levels of experience, the consistency between clinicians and AI, and the distribution of errors across specific depth levels. In particular, it investigates the extent

to which uncertainties in the boundary regions of the STN hinder the establishment of a shared interpretation, and it assesses the consistency of evaluations across different classification categories. By selecting ChatGPT as the AI model, the study aims to assess the reliability of a widely accessible software that clinicians can use. These analyses are expected to provide valuable insights into improving clinical decision-making in STN targeting, evaluating the potential of AI models, and developing new strategies for achieving a more objective interpretation of STN MER recordings.

## METHODS

### Participant Selection

This study was conducted with the approval of the Scientific Research Ethics Committee of the University of Health Sciences Türkiye, Başakşehir Çam and Sakura City Hospital (approval no: 47, date: 04.12.2024). Between 2021 and 2024, a total of thirty-two MERs from patients diagnosed with Parkinson's disease who underwent STN-DBS performed by the same surgical team were included in this study. The surgical decision was made by a multidisciplinary movement disorder board, including a neurosurgeon, movement disorder neurologist, psychiatrist, neuropsychologist, speech therapist, and physiotherapist. Patients were evaluated based on objective criteria established by the board, and those deemed suitable for DBS surgery.

All patients underwent comprehensive preoperative clinical evaluations, including the Unified Parkinson's Disease Rating Scale, Parkinson's Disease Questionnaire, and an extensive neuropsychological test battery. Patients included in the study underwent these assessments both preoperatively and postoperatively. Only those who demonstrated significant clinical benefit from DBS, had confirmed STN stimulation in all postoperative evaluations, and continued to live with an active stimulator system, were considered for analysis.

Patients with incomplete, inaccurate, or unreliable electrophysiological recordings, those who did not experience the expected benefit from DBS, or those whose stimulation systems were deactivated were excluded from

the study. Written informed consent was obtained from all of the participants.

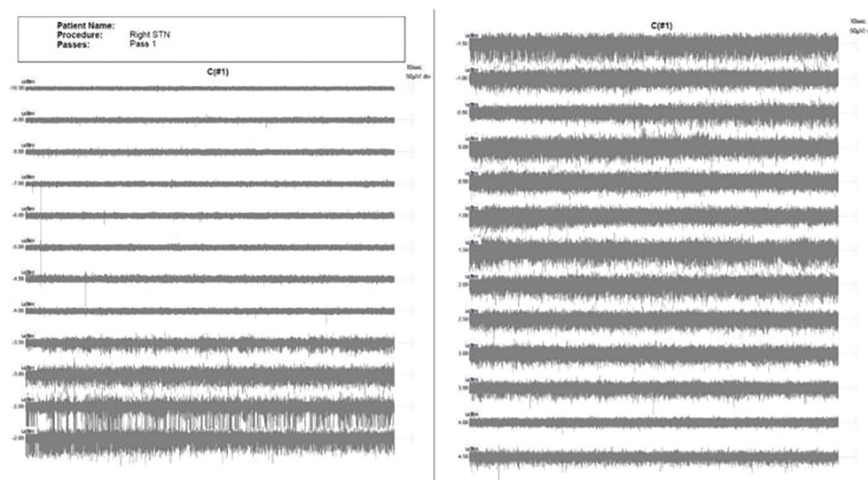
### Surgical Planning and Electrophysiological Recording

All patients underwent high-resolution 1.5 Tesla or 3 Tesla magnetic resonance imaging (MRI) 1 to 3 days prior to surgery. The MRI protocol included T1-weighted images, T2-weighted images, and contrast-enhanced T1 images, as well as diffusion tensor imaging (DTI). The DTI sequences were used to map white matter tracts. Targeting was performed using both direct and indirect methods. The dorsolateral region of the STN was identified as the optimal target for controlling motor symptoms in Parkinson's disease. The initial coordinates were determined based on the anterior commissure-posterior commissure (PC) line, with a starting reference of X:  $\pm 12$ , Y: -2, Z: -4 mm. Adjustments were made using direct MRI visualization to precisely target the dorsolateral STN. All surgical planning was conducted using Stealth and BrainLab Elements™ surgical navigation software. All patients remained awake during surgery, and a stereotactic frame was placed on the morning of the procedure. A 1-mm slice computed tomography (CT) scan was obtained with the frame in place and fused with preoperative MRI to determine the stereotactic coordinates. MER and macrostimulation were performed in all patients. Electrophysiological recordings were obtained intraoperatively using Alpha Omega, AlphaRS, and FHC Guideline 5 systems. The MER technique had been detailed in a previous study (13). MER was initiated 10 mm above the target coordinate, advancing in 1-mm increments until reaching 5 mm below, and then in 0.5 mm increments until the final depth was reached. Recordings were terminated when the STN electrophysiological activity ended, when a substantia nigra pars reticulata recording was

obtained, or, at +4.5 mm. Macrostimulation was performed in the orientation where the longest STN recording was observed. MER recordings were stored in PDF format for further analysis (Figure 1). During macrostimulation, motor responses to stimulation were observed and evaluated. The placement of the DBS electrodes was guided by MER and macrostimulation findings, following a Ben-Gun orientation. The final positioning of the electrodes ensured that the middle contact points were aligned with the intended target, while the tip of the electrode was placed in contact with the substantia nigra. The decision to use directional or non-directional electrodes was made intraoperatively based on electrophysiological findings and the patient's motor response. DBS hardware included Boston Scientific Gevia, Genus, and Medtronic Activa RC/PC models. The stimulator implantation was performed during a single surgical session, with the device placed in the midclavicular region. Postoperatively, 1 mm slice CT scans were fused with preoperative MRI to assess electrode placement. Revisions were performed for patients with a deviation greater than 2 mm from the planned target. Deep brain stimulators were typically activated within 3 to 7 days postoperatively. Prior to activation, patients underwent a 12-hour medication withdrawal. Neurology and neurosurgery specialists performed stimulator activation, and motor effects were evaluated to determine stimulation parameters. After optimizing individualized medication and stimulation settings, patients were discharged.

### Artificial Intelligence Analysis Process

In this study, 32 STN MERs were independently evaluated by two experienced and two less experienced MER clinicians. Each clinician was instructed to classify each depth level into one of the following categories: artifact, thalamus,



**Figure 1.** Sample microelectrode recording (MER) output. This figure shows a sample MER output, displaying depth annotations on the left side and 10-second recordings obtained at each depth level

silent, STN, suspicious STN, substantia nigra, and N/A (no recording). The same recordings were also analyzed by the ChatGPT 4.0 model, and the results were recorded. A specially designed prompt was used to enable ChatGPT to analyze MER recordings. Through this prompt, the model independently assessed wave patterns at each depth level and assigned them to the predetermined categories. The prompt structure used for generative pre-trained transformers (GPTs) interpretation of MER recordings was designed as follows (8,9,14).

**Task Description:** This prompt was used to analyze 10-second MER recordings in PDF format, categorizing wave patterns based on depth levels. Each depth level was independently evaluated.

**Depth Values:** Recordings began 10 mm above the target and proceeded down to -4.5 mm. The scale on the left side of the recordings indicated depth levels, and the analysis was conducted separately for each level.

**Categories:** Evaluations were assigned to one of the following categories based on wave patterns:

- **Silent:** Regions with little to no electrical activity. The wave pattern is nearly silent. This activity is typically observed outside high-activity regions such as the STN, the thalamus, or within the zona incerta.
- **Thalamus:** Contains low-frequency, regular patterns. Typical firing frequency ranges from 10 to 30 Hz, with action potential amplitude between 50 and 100 mV. This region is involved in motor and sensory transmission and is typically observed before the STN.
- **Suspicious STN:** Regions where baseline broadening begins or diminishes are included in this category. Changes in frequency and amplitude occur before or after bursts intensify. This phenomenon is generally observed after the thalamus and in transition zones between the STN and substantia nigra.
- **STN:** High-frequency, high-amplitude regions containing intense burst activities are classified as STN. The firing frequency ranges from 15 to 30 Hz, with action potential amplitude between 60 and 80 mV. It plays a critical role in motor control and is observed at depths where baseline broadening becomes prominent.
- **Substantia Nigra:** Refers to specific regions characterized by low-frequency, regular activity. Dopaminergic neurons in the pars compacta fire at 1-8 Hz, while GABAergic neurons in the pars reticulata fire at 20-40 Hz. Action potential amplitude ranges from 40 to 80 mV.

- **Artifact:** Disturbances or unwanted noise arising during electrical recording are classified as artifacts. These may result from electrode movement or environmental factors.

- **N/A:** Depths where no recordings were obtained or where the data were unprocessable were classified as N/A. Cases where measurements were not performed during electrode transitions were also included in this category.

**Analysis Principles:** Each depth level was analyzed independently. All assessments were performed objectively based on predefined wave patterns.

**Output Format:** All analyses were presented in a table format, systematically recording the assigned categories for each depth level.

Throughout the study, manual corrections were made only in cases of technical errors in GPT's decision-making processes. However, no external intervention was applied to the category decisions made by the model.

### Statistical Analysis

To assess the overall consistency between the classifications made by clinicians and GPT, Fleiss' Kappa coefficient was calculated. Fleiss' Kappa is a multi-rater agreement coefficient used to measure the consistency among multiple evaluators. Additionally, the consistency of each clinician with GPT was calculated separately using Cohen's Kappa coefficient. Cohen's Kappa is a statistical method that quantifies the level of agreement between two raters, adjusting for chance agreement to determine the actual level of consistency. Values below 0 indicate no agreement between evaluators, suggesting a level of disagreement worse than randomness. Values between 0.00 and 0.20 indicate very weak agreement, meaning no meaningful association between evaluators' decisions. Values between 0.21 and 0.40 indicate weak agreement, where the evaluators' decisions are only slightly better than random chance. Values between 0.41 and 0.60 indicate moderate agreement, suggesting that evaluators partially share a common perspective. Values between 0.61 and 0.80 indicate strong agreement, indicating evaluators mostly make the same decisions. Values between 0.81 and 1.00 indicate almost perfect agreement, meaning evaluators reach nearly identical classifications. The consistency among experienced clinicians and less experienced clinicians were analyzed separately. Additionally, the agreement between clinician groups and GPT was compared. Furthermore, GPT's reliability for each classification category was assessed using precision and recall calculations to evaluate the accuracy of its assigned classifications. For depth-based analysis, classifications from GPT and clinicians were

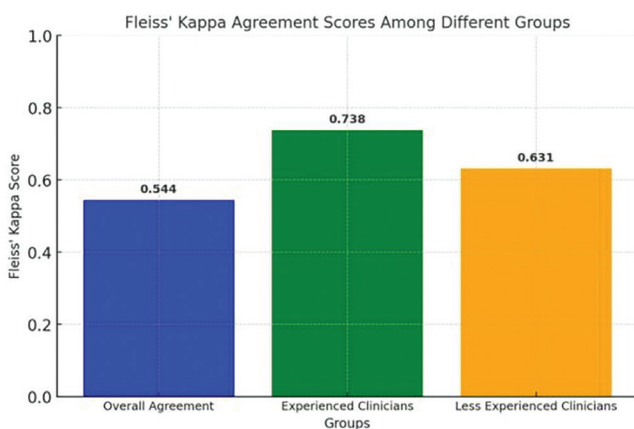


compared across different depth levels, with a particular focus on the transition zones at the STN's entry and exit points. The stability of GPT's classifications was measured by evaluating its consistency with previous and subsequent predictions, providing insights into category transition consistency. This analysis aimed to determine how reliably GPT identified specific categories and to what extent its predictions fluctuated in transition zones. To assess misclassification rates, cases where GPT's classification did not align with the majority opinion were identified. The majority opinion was defined as the most frequently chosen category among the four clinicians, and GPT's consistency with this majority classification was analyzed. Statistical analyses were conducted using SPSS version 22 (IBM Corp., Armonk, NY, USA), and data visualization was performed with Python Matplotlib (Matplotlib Development Team, Python Software Foundation, USA).

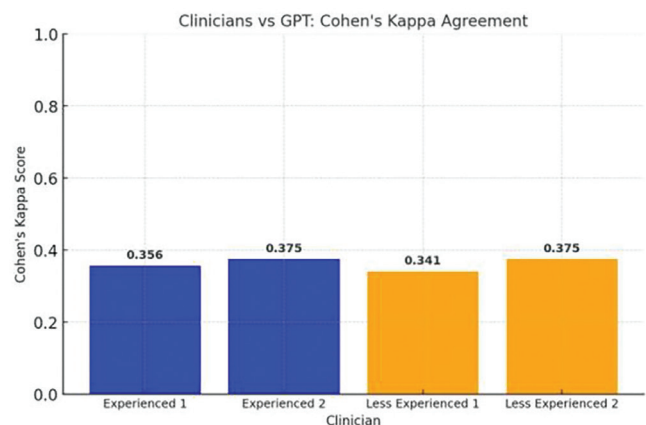
## RESULTS

In this study, 32 STN MERs were independently evaluated by two experienced clinicians, two less experienced clinicians, and an AI model (ChatGPT 4.0), and the results were analyzed statistically. The overall consistency among all evaluators, calculated using Fleiss' Kappa coefficient, was found to be 0.544. The Fleiss' Kappa coefficient between the experienced clinicians was 0.738, while the Fleiss' Kappa coefficient between the less experienced clinicians was 0.631 (Figure 2). The agreement between GPT and clinicians was analyzed using Cohen's Kappa coefficient. Cohen's Kappa coefficient between experienced clinician 1 and GPT was 0.356, while the value between experienced

clinician 2 and GPT was 0.375. Among the less experienced clinicians, Cohen's Kappa coefficient between beginner clinician 1 and GPT was 0.341, and between beginner clinician 2 and GPT was 0.375. These findings indicate that GPT exhibited low consistency with both experienced and less experienced clinicians (Figure 3). In category-based analysis, the accuracy of GPT in identifying the STN category was 73.47%; the silent category, it was 66.41%; the artifact category, it was 42.85%; the thalamus category, it was 7.69%; and the substantia nigra category, it was 3.33%. GPT demonstrated the highest accuracy in identifying the STN category, suggesting significant agreement with clinicians in recognizing the STN region. The 66.41% accuracy for the silent category indicates that the model was relatively successful in identifying low-activity regions. For the artifact category, the accuracy was 42.85%, indicating that the model struggled to distinguish artifacts caused by electrode movement or environmental factors. The accuracy rates for the thalamus and substantia nigra categories were 7.69% and 3.33%, respectively, indicating that GPT was unreliable in classifying these regions. The precision for STN was 0.73, and recall for STN was 0.65. For the silent category, precision was 0.66, and recall was 0.59. For artifact, precision was 0.42, and recall was 0.37. For thalamus, precision was 0.08, and recall was 0.12. For the substantia nigra, precision was 0.03, and recall was 0.05. These findings indicate that GPT was relatively successful in identifying the STN, but it exhibited a significant error margin in transition zones (Figure 4). In depth-based analysis, GPT's category transition consistency was calculated as 51.5%, indicating that the model selected the same category for consecutive depth

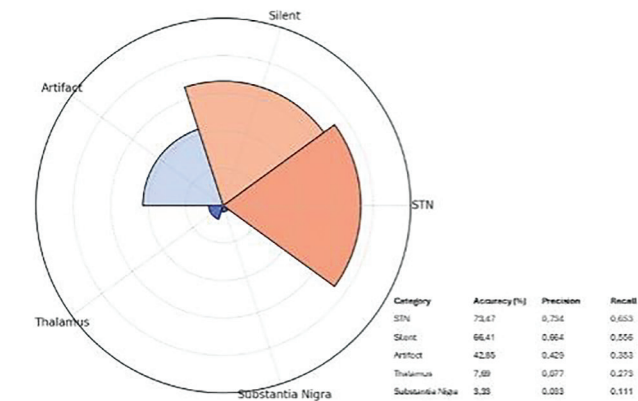


**Figure 2.** Fleiss' Kappa agreement scores among evaluator groups. This figure illustrates the overall Fleiss' Kappa agreement scores among all evaluators. The overall agreement was calculated as 0.544, reflecting moderate consistency among all evaluators. Among the experienced clinicians, the agreement score was 0.738, indicating strong consistency, while the agreement score among the less experienced clinicians was 0.631, showing moderate consistency.



**Figure 3.** Cohen's Kappa agreement between GPT and clinicians. This figure shows the Cohen's Kappa agreement scores between GPT and individual clinicians.

GPT: Generative pre-trained transformer



**Figure 4.** Performance of GPT in identifying categories: accuracy, precision, and recall. This figure presents the accuracy, precision, and recall values for each category evaluated by GPT. The STN category showed the highest performance, with an accuracy of 73.47%, precision of 0.734, and recall of 0.653, indicating relatively successful identification of the STN region

GPT: Generative pre-trained transformer, STN: Subthalamic nucleus

levels 51.5% of the time. The misclassification rate was found to be 45.87%, meaning that GPT produced different results from the majority opinion of the clinicians in 45.87% of cases. To measure the model's tendency to maintain a category assignment over multiple depths, within-category consistency was calculated as 14.28%, indicating that GPT exhibited a high degree of variability in its decision-making process.

## DISCUSSION

DBS procedures are becoming increasingly widespread, and successfully performing these procedures requires a high level of expertise. In some cases, the inability to obtain preoperative imaging of the desired quality or the difficulty in identifying the STN sweet spot solely through radiological methods demonstrate that relying exclusively on imaging techniques does not always lead to optimal targeting (8,9,11,12). MER is a valuable alternative for more reliable targeting. However, MER techniques are still predominantly reliant on clinicians' subjective visual and auditory assessments during surgery, which introduces variability in outcomes. To address this, AI-based technologies have been developed to both assist novice surgical teams in mastering this highly specialized technique and to reduce subjectivity in decision-making. These advancements aim to enable real-time and automated evaluation of MER recordings, enhancing the objectivity and consistency of the process (2,8,15).

These methods have primarily been developed by first implementing preprocessing and artifact removal steps, followed by the application of techniques designed to detect differences in various signal characteristics (9,12). Some approaches focus on spike-related parameters (16,17) others analyze power changes in specific frequency bands (18-20), and yet others rely on wavelet analysis (21,22). Another group of methods is based entirely on deep learning algorithms (23-26). In these studies, the supervised learning method has been predominantly used for AI training. The primary reason for this preference is that supervised learning allows AI to be trained with smaller datasets, making the training process easier and faster. Although supervised learning enables rapid implementation, its performance ceiling is inherently limited by the expertise of the annotators who label the training data. Since AI learns both correct and incorrect classifications, its effectiveness is directly influenced by the accuracy and consistency of the human-provided labels (2,8,15). Despite these limitations, the full integration of AI-driven real-time models into DBS procedures appears to be a realistic near-future possibility. AI could serve as a guidance tool for specialists by providing feedback and fine-tuning suggestions, thereby improving decision-making in MER interpretation (15). When designing this study, our goal was to evaluate an AI model with access to open-source big data and assess its ability to interpret MER recordings independently, without relying on clinician directives or modifications to its source code. The objective was to determine whether the model could function like a human evaluator who retrospectively analyzes MER recordings and to assess its suitability for clinical use. In this regard, our study differs from previous AI-based research in this field. At this point, an unsupervised machine learning approach, which could eliminate the need for human expert input, was considered. While in theory this method could overcome the limitations of supervised learning and lead to more advanced models, the requirement for extremely large datasets in unsupervised learning remains a significant barrier. At present, no clinical setting has access to datasets of the necessary scale, making the application of unsupervised learning in this field an unrealistic goal.

Another significant issue with the existing methods is that they are either commercially available at high costs or, if open source, they require complex technical knowledge and software expertise, making them difficult for clinicians to utilize. At this point, the idea of using ChatGPT, which is easily accessible and does not require extensive technical background knowledge, emerges as a potential tool to assist in the interpretation of MER recordings. In this study, we aimed to explore this possibility by evaluating 32

STN MERs, independently assessed by two experienced clinicians, two less experienced clinicians, and an AI model (ChatGPT 4.0), with statistical analyses conducted on the results. However, the findings indicate that this AI model still exhibits low consistency with experienced and less experienced clinicians.

In this study, GPT demonstrated the highest accuracy in the STN category, indicating partial consistency with clinicians in identifying the STN region. Similar findings have been reported in AI models focused on background neural activity (17,27-29). The STN exhibits twice the background activity compared to its neighboring structures (30), a phenomenon likely associated with the high neuronal density within the STN (14). Since the AI model used in this study primarily focuses on basal activity, it performed better in identifying the STN. However, the study also revealed that the model has a significant margin of error in the STN entry and exit zones as well as in adjacent structures. In the model proposed by Rajpurohit et al. (16), the accuracy for STN entry and exit regions was reported to be between 60% and 80%. Similarly, Chaovalitwongse et al. (28) successfully identified STN and its neighboring structures with approximately 90% accuracy using a combination of seven spike-dependent and six spike-independent approaches. These findings highlight that feature-based machine learning models have demonstrated higher accuracy than the AI model in our study.

In summary, our study demonstrates that while GPT shows partial consistency with clinicians in identifying the STN, it exhibits low accuracy in other categories. The low within-category consistency (14.28%) indicates that the model exhibits significant variability in transition zones and makes unstable decisions. Although GPT has achieved a certain level of accuracy in STN detection, its reliability remains low in differentiating transition zones and low-activity structures. For the AI model to be considered a supportive tool in STN targeting, its training must be expanded, and its algorithms must be refined to enhance precision.

### Study Limitations

This study has several limitations. First, the evaluation of GPT's performance was based on a limited dataset of MER recordings obtained from a single center, which may restrict the generalizability of the results. Second, the ground truth was established through the consensus of clinicians, which, although a common method, may still reflect inter-rater variability and subjective interpretation. Third, the model was not specifically trained or fine-tuned on electrophysiological data related to DBS, which may have affected its performance in distinguishing between critical anatomical regions. Lastly, technical constraints in data

formatting and input length limitations may have impacted the accuracy and consistency of the model's classifications.

## CONCLUSION

AI-based algorithms, which are increasingly assisting us in various fields, also hold significant potential to support DBS procedures. Soon, the impact of open AI models in clinical practice is expected to grow. However, at present, many clinicians remain unable to benefit from these advancements due to both technical and financial barriers. While ChatGPT is widely used in various fields as a practical and cost-effective tool, our findings indicate that it is not yet sufficiently reliable for DBS-related applications. Nonetheless, its potential for future development is promising. At this stage, while the model may serve as a guiding tool, its integration into clinical decision-making processes still appears to be premature. Further studies, in light of ongoing advancements and updates, may reveal whether this situation will change.

## ETHICS

**Ethics Committee Approval:** This study was conducted with the approval of the Scientific Research Ethics Committee of the University of Health Sciences Türkiye, Başakşehir Çam and Sakura City Hospital (approval no: 47, date: 04.12.2024).

**Informed Consent:** Written informed consent was obtained from all of the participants.

## FOOTNOTES

### Authorship Contributions

Surgical and Medical Practices: O.H., T.H., T.Ö.K., B.T., Concept: O.H., T.H., T.Ö.K., B.T., Design: O.H., A.A., N.B.G., F.D., Data Collection or Processing: O.H., A.A., N.B.G., F.D., Analysis or Interpretation: F.D., B.T., Literature Search: A.A., T.H., T.Ö.K., Writing: O.H., T.H., T.Ö.K., B.T.

**Conflict of Interest:** No conflict of interest was declared by the authors.

**Financial Disclosure:** The authors declare that this study received no financial support.

## REFERENCES

1. Hariz M, Blomstedt P. Deep brain stimulation for Parkinson's disease. *J Intern Med*. 2022;292:764-78.
2. Wan KR, Maszczyk T, See AAQ, Dauwels J, King NKK. A review on microelectrode recording selection of features for machine learning in deep brain stimulation surgery for Parkinson's disease. *Clin Neurophysiol*. 2019;130:145-54.
3. Vinke RS, Geerlings M, Selvaraj AK, Georgiev D, Bloem BR, Esselink RA, et al. The role of microelectrode recording in deep brain stimulation surgery for Parkinson's disease: a systematic review and meta-analysis. *J Parkinsons Dis*. 2022;12:2059-69.

4. Al Awadhi A, Tyrand R, Horn A, Kibleur A, Vincentini J, Zacharia A, et al. Electrophysiological confrontation of Lead-DBS-based electrode localizations in patients with Parkinson's disease undergoing deep brain stimulation. *Neuroimage Clin.* 2022;34:102971.
5. Pastor J, Vega-Zelaya L. Can we put aside microelectrode recordings in deep brain stimulation surgery? *Brain Sci.* 2020;10:571.
6. Kinfé TM, Vesper J. The impact of multichannel microelectrode recording (MER) in deep brain stimulation of the basal ganglia. *Acta Neurochir Suppl.* 2013;117:27-33.
7. van den Munckhof P, Bot M, Schuurman PR. Targeting of the subthalamic nucleus in patients with Parkinson's disease undergoing deep brain stimulation surgery. *Neurol Ther.* 2021;10:61-73.
8. Lima J, de Carvalho M, Dias JM. Analysis and classification of microelectrode recordings in deep brain stimulation surgery. *Academia.* 2015:1-13.
9. Coelli S, Levi V, Del Vecchio Del Vecchio J, Mailland E, Rinaldo S, Eleopra R, et al. Characterization of microelectrode recordings for the subthalamic nucleus identification in Parkinson's disease. *Annu Int Conf IEEE Eng Med Biol Soc.* 2020;2020:3485-8.
10. Novak P, Daniluk S, Elias SA, Nazzaro JM. Detection of the subthalamic nucleus in microelectrographic recordings in Parkinson disease using the high-frequency (> 500 Hz) neuronal background. *J Neurosurg.* 2007;106:175-9.
11. Valsky D, Blackwell KT, Tamir I, Eitan R, Bergman H, Israel Z. Real-time machine learning classification of pallidal borders during deep brain stimulation surgery. *J Neural Eng.* 2020;17:016021.
12. Martin T, Jannin P, Baxter JSH. Generalisation capabilities of machine-learning algorithms for the detection of the subthalamic nucleus in micro-electrode recordings. *Int J Comput Assist Radiol Surg.* 2024;19:2445-51.
13. Tugcu B, Hasimoglu O, Altinkaya A, Barut O, Hanoglu T. Comparison of electrophysiological and radiological subthalamic nucleus length and volume. *Turk Neurosurg.* 2023;33:126-33.
14. Benazzouz A, Breit S, Koudsie A, Pollak P, Krack P, Benabid AL. Intraoperative microrecordings of the subthalamic nucleus in Parkinson's disease. *Mov Disord.* 2002;17 Suppl 3:S145-9.
15. Inggas MAM, Coyne T, Taira T, Karsten JA, Patel U, Kataria S, et al. Machine learning for the localization of subthalamic nucleus during deep brain stimulation surgery: a systematic review and meta-analysis. *Neurosurg Rev.* 2024;47:774.
16. Rajpurohit V, Danish SF, Hargreaves EL, Wong S. Optimizing computational feature sets for subthalamic nucleus localization in DBS surgery with feature selection. *Clin Neurophysiol.* 2015;126:975-82.
17. Schiaffino L, Muñoz AR, Martínez JG, Villora JF, Gutiérrez A, Torres IM. STN area detection using K-NN classifiers for MER recordings in Parkinson patients during neurostimulator implant surgery. *J Phys Conf Ser* 2016;705:012050.
18. Vargas Cardona HD, Álvarez MA, Orozco ÁA. Multi-task learning for subthalamic nucleus identification in deep brain stimulation. *Int J Mach Learn Cybern* 2018;9:1181-92.
19. Valsky D, Marmor-Levin O, Deffains M, Eitan R, Blackwell KT, Bergman H, et al. Stop! border ahead: automatic detection of subthalamic exit during deep brain stimulation surgery. *Mov Disord.* 2017;32:70-9.
20. Khosravi M, Atashzar SF, Gilmore G, Jog MS, Patel RV. Intraoperative localization of STN during DBS surgery using a data-driven model. *IEEE J Transl Eng Health Med.* 2020;8:1-9.
21. Karthick P, Wan KR, Qi ASA, Dauwels J, King NKK. Automated detection of subthalamic nucleus in deep brain stimulation surgery for Parkinson's disease using microelectrode recordings and wavelet packet features. *J Neurosci Methods.* 2020;343:108826.
22. Park KH, Sun S, Lim YH, Park HR, Lee JM, Park K, et al. Clinical outcome prediction from analysis of microelectrode recordings using deep learning in subthalamic deep brain stimulation for Parkinson's disease. *PloS One.* 2021;16:e0244133.
23. Peralta M, Bui QA, Ackaouy A, Martin T, Gilmore G, Haegelen C, et al. SepaConvNet for localizing the subthalamic nucleus using one second micro-electrode recordings. *Annu Int Conf IEEE Eng Med Biol Soc.* 2020;2020:888-93.
24. Martin T, Peralta M, Gilmore G, Sauleau P, Haegelen C, Jannin P, et al. Extending convolutional neural networks for localizing the subthalamic nucleus from micro-electrode recordings in Parkinson's disease. *Biomed Signal Process Control* 2021;67:102529.
25. Hosny M, Zhu M, Gao W, Fu Y. Deep convolutional neural network for the automated detection of Subthalamic nucleus using MER signals. *J Neurosci Methods.* 2021;356:109145.
26. Xiao L, Li C, Wang Y, Si W, Lin H, Zhang D, et al. Amplitude-frequency-aware deep fusion network for optimal contact selection on STN-DBS electrodes. *Sci China Inf Sci* 2022;65:140404.
27. Cagnan H, Dolan K, He X, Contarino MF, Schuurman R, van den Munckhof P, Wadman WJ, Bour L, Martens HC. Automatic subthalamic nucleus detection from microelectrode recordings based on noise level and neuronal activity. *J Neural Eng.* 2011;8:046006.
28. Chaovaitwongse WA, Jeong YS, Jeong MK, Danish SF, Wong S. Pattern recognition approaches for identifying subcortical targets during deep brain stimulation surgery. *IEEE Intell Syst* 2011;26:54-63.
29. Ciecierski K, Mandat T, Rola R, Raś ZW, Przybyszewski AW. Computer aided subthalamic nucleus (STN) localization during deep brain stimulation (DBS) surgery in Parkinson's patients. *Ann Acad Med Siles* 2014;68:275-83.
30. Kano T, Katayama Y, Kobayashi K, Kasai M, Oshima H, Fukaya C, et al. Multiple-cell spike density and neural noise level analysis by semimicroelectrode recording for identification of the subthalamic nucleus during surgery for Parkinson's disease. *Neuromodulation.* 2008;11:1-7.