

Enhanced control of a brain–computer interface by tetraplegic participants via neural-network-mediated feature extraction

Corresponding author: Benyamin Haghi

Editorial note

This document includes relevant written communications between the manuscript's corresponding author and the editor and reviewers of the manuscript during peer review. It includes decision letters relaying any editorial points and peer-review reports, and the authors' replies to these (under 'Rebuttal' headings). The editorial decisions are signed by the manuscript's handling editor, yet the editorial team and ultimately the journal's Chief Editor share responsibility for all decisions.

Any relevant documents attached to the decision letters are referred to as **Appendix #**, and can be found appended to this document. Any information deemed confidential has been redacted or removed. Earlier versions of the manuscript are not published, yet the originally submitted version may be available as a preprint. Because of editorial edits and changes during peer review, the published title of the paper and the title mentioned in below correspondence may differ.

Correspondence

Wed 08 Mar 2023

Decision on Article nBME-23-0067

Dear Mr Haghi,

Thank you again for submitting to *Nature Biomedical Engineering* your manuscript, "Convolutional neural network feature extraction dramatically improves brain-machine interface control for tetraplegic participants".

The manuscript has been seen by 4 experts, whose reports you will find at the end of this message, and please excuse the unusual delay in reaching a decision (1 Reviewer needed additional time to assess the work and provide a report and we had difficulties in securing reviewers).

You will see that although the reviewers have good words for aspects of the work, they raise serious concerns about the overall significance of the study. Additionally, they raise some technical concerns specifically regarding the lack of comparison of FENet to widely used state-of-the-art band pass filters and feature extraction methods. Moreover, the reviewers expressed concerns about the lack of in-depth analysis of the neural features particularly direct evidence of their correspondence to neural activity. They also questioned the degree of validation of the generalizability of the approach and the explainability of the methodology.

Having considered the reviewers' advice, we have reached the conclusion that the work does not, at present, reach the significance that we look for in manuscripts that we consider for further external peer review.

However, should further work allow you to compare FENet to an appropriate comparison baseline as suggested by the Reviewers, evaluate the mechanisms behind the approach, and provide clarification of the methodology used, including sufficient code and data for evaluation, we would be willing to assess an appeal for reconsideration.



Open Access This file is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. In the cases where the authors are anonymous, such as is the case for the reports of anonymous peer reviewers, author attribution should be to 'Anonymous Referee' followed by a clear attribution to the source work. The images or other third-party material in this file are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0>.

We hope that you will find the referee reports helpful when revising the work
Best wishes,

Michelle

Dr Michelle Korda
Senior Editor, [Nature Biomedical Engineering](#)

* The manuscript may be appropriate for another journal in the Nature Portfolio. If you wish to explore the journals and transfer your manuscript, please use our [manuscript transfer portal](#). You will not have to re-supply manuscript metadata and files. This link can only be used once and remains active until used. For more information, please see our [manuscript transfer FAQ](#) page.

Note that any decision to opt in to *In Review* at the original journal is not sent to the receiving journal on transfer. You can opt in to [In Review](#) at receiving journals that support this service by choosing to modify your manuscript on transfer. *In Review* is available for primary research manuscript types only.

Reviewer #3 (Report for the authors (Required)):

Summary:

The authors approach the challenge of feature extraction from extracellular recordings in intracortical brain-machine interfaces. The authors propose Feature Extraction Network (FENet), which is a 1D CNN that maps raw electrical signals to unstructured neural features, which could then be fed into downstream decoders. The design criteria for FENet includes the ability to use it in current brain-machine interface pipelines; generalize across time, tasks, subjects, and electrodes; run in real-time with low power consumption; and not require additional complexity or increases in volume of training data.

To accomplish these goals, the authors trained FENet to extract neural features from raw electrical signals. The parameters of FENet were fixed across electrodes and sessions, and the mapping from neural features to behavior was allowed to vary across electrodes and sessions. They demonstrated FENet's capabilities offline with data from three human participants. FENet demonstrated substantial improvements in closed-loop BCI performance, relative to threshold crossings (shown for one participant). They show FENet generalizes to new datasets, tasks, and brain regions, but when trying to generalize FENet across participants, performance was relatively low. Importantly, their participant reported a strong preference for the FENet-based decoders. The authors additionally report a brief but impactful testimonial that a participant, who had formerly requested to be explanted, later reversed their decision due to the improvements of FENet.

Degree of Advance:

As described in Major Comments, in many instances the authors did not provide sufficient detail to understand the robustness of their conclusions. As such, it would be important to see a revised manuscript with these concerns addressed before being able to fully assess the degree of advance. However, if sufficient detail could be provided to back up the claims of the manuscript, this work may represent a substantial advance with potentially broad impact in the BCI space. To my knowledge, FENet is the first demonstration that a neural network operating on high-bandwidth intracortical electrophysiology recordings (30 kHz) can extract neural features that are more informative than traditional approaches (threshold crossings, wavelets). It is exciting that such features may also be robust to the degradation in signal quality that can lead to a decrease in performance throughout the lifetime of a chronically implanted array (improvement demonstrated offline in one subject).

Implications of Findings:

FENet may serve as a proof of principle that neural networks for feature detection may improve the performance and longevity of intracortical brain-machine interfaces.

MAJOR Comments:

1 There are many instances where the method's details are insufficient to completely evaluate the

manuscript. We are open to reviewing a revised manuscript, but the lack of many details in the manuscript's current form makes it impossible to evaluate many of the claims. Some examples where additional details may be helpful:

- 1 What decoding strategy was used for threshold crossings (TCs) and wavelet transforms (WTs)? Only the decoding strategy for FENet was indicated.
- 2 What type of smoothing was applied (i.e. for the comparisons in Extended Figure 6)?
- 3 How was partial least-squares regression applied? Since this was a critical step in deciding the final FENet architecture, more details regarding how it was applied would be helpful.
- 4 There are many details of the finger task that are not discussed in the methods despite being a main figure. For example, what do muscle model and somatotopy model mean? What task(s) and subject(s) was FENet trained on to test generalization to the finger task? What do RDA and RDM mean?
- 5 How were bitrates calculated?

2 For work that is heavily computational, to robustly assess, it is important to have a copy of the code, documentation, example trained models, and example data on which to test the models. This is also critical in assessing the reproducibility of the work. We recommend that the authors provide the code and data in a containerized form so that there are no issues with libraries/software versioning/dependencies, etc. In the paper the authors mention that it is difficult to make the broadband data available because the data are "too large and confidential to be publicly shared." However, it is not required to provide all data, some sample datasets to train and test the models and reproduce key conclusions should be sufficient. I was unclear what the "confidential" concern referred to. If there is a concern about participant privacy, there should be ways of anonymizing small stretches of data sufficient for replicating the work. Alternatively, potentially the authors could use previously-recorded monkey data from the same lab or open source, publicly-available broadband data (e.g., https://zenodo.org/record/3854034#.Y_0RbuzMJeg). There should be many possibilities for addressing this concern.

1 Related, in the methods, the authors should specify all software and library version info required to reproduce this work.

3 The authors claim that a major design decision for FENet is that it remains computationally simple and relevant to devices, but there is no data or analysis presented to confirm they achieved this goal. To back up this claim, the authors should include computation time results, as well as estimates of computational complexity (in the form of Big O notation), including the full feature extraction pipeline of FENet against the full feature extraction pipeline of TCs and WT. Readers should have concrete information on the suitability of FENet for low-power implantable devices.

4 It is difficult to interpret how the features FENet extracts relate to the underlying neural activity. I recommend that the authors provide snippets of actual features output by FENet and show comparisons to TCs and WT for the same data snippets, to provide readers with some intuition of what FENet finds are helpful features.

5 What are the critical architecture / hyperparameter choices that affect model performance?

1 Values for hyperparameters are not provided (e.g. dropout rate, learning rate).

2 As with any neural network study, it is important to conduct ablation experiments with FENet to clarify critical components of the model architecture. This would allow other researchers to build off the work.

6 Open-loop and closed-loop TC decoders did not apply the same preprocessing steps as current state-of-the-art methods presented by others. Many groups have applied "non-causal filtering" (Masse 2014 <https://doi.org/10.1016/j.jneumeth.2014.08.004>) and exponential smoothing (Willett 2019 <https://doi.org/10.1016/j.jneumeth.2014.08.004>) of the binned spiking rates during closed-loop control, but based on the methods, it appears these preprocessing steps were not included. For the offline analyses, it would be important to compare specifically against the non-causal threshold crossing approach, at a minimum.

7 The authors repeatedly claim "state-of-the-art" decoding performance (e.g., L66, L149). I recommend this language be adjusted, because it is hard to claim state-of-the-art without a benchmark for comparison. The performance demonstrated is certainly not on par with published performance, both for open-loop decoding (see Pei*, Ye* 2021 <https://doi.org/10.48550/arXiv.2109.04463>, results here

<https://eval.ai/web/challenges/challenge-page/1256/leaderboard>), or closed-loop performance (grid task bitrates, as in the BrainGate work: <https://elifesciences.org/articles/18554>).

8 The authors mention in their methods that closed-loop cursor position was augmented “potentially with some computer assistance” (lines 301-302). The methods by which computers assisted closed-loop cursor control need to be explicitly defined for readers to properly interpret the results presented in the manuscript.

MINOR Comments:

1 24 - Shouldn't the same parameters always be used during training and validation? Otherwise, one isn't validating the trained parameters. This may be rephrased to make the intended point more obvious to the reader.

2 78-83 - “FENet dramatically improved a human clinical BMI participant's ability to use brain signals to control a computer cursor, increasing the bitrate nearly threefold in the closed-loop control. This performance increase was clinically significant: Prior to FENet, the clinical participant requested surgical reimplantation to improve the quality of neural recordings that had degraded substantially since the initial implantation. With FENet, the participant was satisfied with the quality of his neural control. Thus, FENet can extend the functional lifetime of the implanted electrodes, mitigating the need for revision surgeries and thus improving commercial viability.” - while interesting and undeniably motivational, this seems more like a discussion topic rather than an introduction topic

3 90-91 - “FENet can improve the efficacy and safety of implantable electrode systems” - there is no safety data included in the paper to justify this claim.

4 170 - “Interestingly, although improving the correlation” - correlation to what? This statement is unclear.

5 246-250 - “This is why the function of a neuron must be estimated by correlating the number of spikes per unit time against behavior and not the exact features of the waveform shape.” - This implies that studies in the past have correlated waveform shapes with behavior, but I don't know of any studies to have done this. If so, I suggest including a citation. Otherwise, if referring to sorting units first then correlating “number of spikes” for each unit with behavior, then consider rewording the sentence.

6 260-261 - “To the degree that trial-to-trial differences are driven by behavioral factors, no algorithm measuring the local activity of neurons can reduce this variability.” - Don't latent variable models of population activity accomplish this, such as LFADS?

7 305-311 - I am unfamiliar with this method of digital referencing. A citation would be helpful, else validation could be provided in the supplement.

8 309-310 - “we applied an 8-order elliptical high pass filter with the cut-off frequency of 80 Hz” - Was a lower cutoff set at 80Hz to avoid residual 60Hz noise? It would be interesting to read the authors' comments on the decision to avoid beta frequency bands of LFP, which have been useful features for motor control for decades. This would be particularly insightful because of the claim that FENet is given the freedom to extract the features it finds most helpful from the entire available broadband recording.

9 320 - “Suppose that ... represents a one-dimensional sample of the broadband neural data consists of S samples” - consider rewording. This sounds like the algorithm simultaneously needs one sample (i.e. “one-dimensional sample”) or S samples. Should this be one sample per electrode, assuming S electrodes? Or does this mean S samples for one electrode, in which case it isn't a “one-dimensional sample”? Given the following two sentences, I think this means that x is a real number of size [SxN]. The explanation should be redefined from there.

1 331 - I see S has been clarified here. Consider copying this to 320 or consolidating into one description of the input data size.

10 326 - Equation 1 - is a parenthesis misplaced here? Shouldn't the loss be between g using the FENet features and the true labels y? As it is printed, the decoder uses the true labels to predict the true labels...

11 331-332 - “in our case, 30 ms of the recorded broadband neural data, which includes 900 samples” - is this just for participant JJ, and was 50ms (1500 samples) for participant EGS?

12 333 - “the input data, s_{i-1} , is padded with zeros” - what does i refer to here? To clarify notation, consider saying “the input to the ith feature engineering module, s_{i-1} , is padded...”

13 347 - “constraining the feature extraction algorithm to use the same parameters for all the electrodes used in the training set.” - For N electrodes, I understand that there is one FENet per electrode, and each of those FENets uses the same parameters to extract features. The methods provided describe how FENet is trained on one electrode, but it is not detailed how the N sets of FENet parameters are combined into one set of parameters to be used for all electrodes.

14 352-361 - Throughout this explanation, a 30ms moving window is used here, but it seems that this only applies to participant JJ. What about participant EGS?

15 375-378 - The outcomes of these queries were not included in any data presented in the manuscript, please include the results.

16 383 - "C is the number of input neural electrodes" - wasn't this N before? Best to be consistent with variable names.

17 407 - "Using the domain knowledge" - what does this mean?

18 481 - "for each of the standard feature types (TCs and FENet)" - what about WTs?

19 Fig 2d - I recommend extending this in time until the target has been reached to compare against other manuscripts with similar plots

20 Fig 3e,f - it would be helpful to know which session from Fig 3a this is. Also, it would be helpful to plot a common time axis here (in seconds) rather than samples.

21 Extended Fig 7d - The colors in this plot are very difficult to visually differentiate, so conclusions cannot be made about the effect of different PLS dimensionalities.

22 Extended Fig 7d,e - It is unclear what is actually being plotted here. How was the value for each channel determined? (What methods were used?) Are the channels sorted in some way? What does the x=0 point represent?

Stylistic Issues/Copyediting:

1 117 - "Extended Figures 7" should be "Extended Figure 7"

2 191-192 - "These results suggest that FENet can generalize across different time periods, brain areas, and patients, and electrodes" - remove the "and" in "and patients"

3 303-304 - "screen in each trial. Then, the participant immediately attempted to press the corresponding finger of the right hand (Figure 5)." - Unexpected larger font size

4 305 - I think "which have not" should be "which has not"

5 320 - "one-dimensional sample of the broadband neural data consists of S samples" - There is a conjunction missing here, perhaps "and"?

6 323 - missing space between psi and "corresponds"

7 332 - missing space between x and "is"

8 375-376 - "and blinded queries to research participants to evaluate responsiveness" - consider putting a colon here

9 402-403 - "Linear decoder predicts the velocity of the cursor movement as smoother prediction than the cursor positions that can provide a more stable prediction for a BMI system." - Confusing sentence, possible grammatical errors

10 479-480 - "human research participants our during double-blind testing" - order of "our" and "during" should be swapped

11 Fig 1b - Electrode 1 is listed twice

12 Fig 4 caption - "These results suggest that FENet can generalize across different time periods, brain areas, and patients, and electrodes" - remove the "and" in "and patients"

Missing Citations:

1 Willsey et al. recently used a convolutional neural network to improve decoding performance in monkeys, and should be discussed in the Introduction or Discussion sections. <https://doi.org/10.1038/s41467-022-34452-w>

Reviewer #4 (Report for the authors (Required)):

This paper presents a neural network based method for creating better neural features for an intracortical brain machine interface using an older Utah array with low amplitude signals. This would be used with existing downstream algorithms. While they report a dramatic performance increase, this is impossible to evaluate because they are leaving out a very common technique that is in use across at least 3 other groups for the same purpose, with known large effects. Specifically, doubling of performance with units present was reported in 2007 by Stark and Abeles using a simple band pass filter from 300-6000 Hz, and groups since then have discovered that this has a disproportionately positive effect on older Utah arrays. Instead, this group compared their approach to threshold crossing, which was among the control features in Stark and Abeles and also with wavelets, which have not been compared with the more commonly used band pass filter to my knowledge. It seems likely that a convolutional neural network can discover how to implement a band pass filter, and this must be addressed directly. This may explain most of the effect as a known simple technique.

For the effect that remains, it is also important to explore the mechanism. This group discusses previous datasets with spikes present, and these could be used to determine what exactly the network is finding. For

example, is it identifying spectral features that correspond to neural units? The time window of 30 ms is quite short so this seems likely. But the brief claim about finding multiple sources per electrode could be and should be evaluated directly with spiking datasets and or simulation.

Please state within the video captions whether the clock is running in real time. A few movements of the participant look slightly odd, making this reviewer concerned that there could be an unstated clock adjustment. Also, please do not include a video of calibration, which could be mistaken by a less careful reviewer for a BMI experiment.

Reviewer #5 (Report for the authors (Required)):

Brief summary of results:

This paper tries to improve feature extraction from neural activity by using a feature extraction net (FENet). The network is small, and shared across electrodes. It results in better BMI performance on a cursor control task, and generalizes across participants and brain areas.

Degree of advance:

The performance improvements are large, however as it stands this is not a huge advance. Many papers have shown improvement over linear decoding of threshold crossings, which is what the paper is comparing FENet to. (see comments for more)

Implications:

A small neural network can be used to extract good features for decoding arm movements from Utah arrays. This means that a light, flexible system could enable better decoding across participants

Major criticisms:

1. The paper has a weak baseline comparison to just decoding TCs using a linear model and lacks comparisons to other methods
2. Basically all the papers cited to justify that TCs are the standard for closed-loop control underlying best-in-world closed loop performance use either a kalman filter (6, 32) or RNN based model (22, 26, 27, 31) to decode kinematics from neural activity. If these are cited as the benchmark, then they should be what's being used when comparing the FENet extraction algorithm to TCs in my opinion. E.g, even if we use a SOTA RNN, or MRNN, are the results with the FENet better than just using TCs? This would make the paper a significant advance.
3. Furthermore, references 32 and 4 both use HF-LFP features and TCs for decoding, which I suggest adding as a comparison to the FENet extraction algorithm.
4. Another potential comparison would be to PSID (Sani et. al. Nature Neuro 2020). Note this feature extraction paper also used a Kalman filter to compare its method to other extraction methods and TCs.
5. Is the FENet specifically what's important to good performance? Would a neural net that is not shared across elecs achieve similar or better performance? If the whole point is just to extract features from neural activity, why limit it in this way? Comparison would be good.
6. Figure 4: Why should I care about this if I am a BCI user? How much time will it save me to do the fenet transfer, vs just using a new model? Train over 10 sessions to get good performance isn't that many (extended data, fig 7)

Minor criticisms:

1. You apply partial least squares regression to the output of the FENet features of each electrode. I am curious what performance is like if you don't do this? Could be nice for Extended Fig 1e. Alternatively, does apply PLS like this to the TCs to reduce dimensionality improve performance? Even though there are not multiple features per electrode, I am concerned this is a critically useful step that makes the FENet/TC comparisons unfair.
2. More interpretation of why FENET succeeded would be helpful. Can we see the output of FENet compared to TCs? Do the authors have any guesses as to what its doing?

Unclear details: The details were clear.

Missing citations: PSID (Sani et. al. Nature Neuro 2020) should be added.

Optional suggestions:

Comparison to algorithms like LFADS or MYOW (from Eva Dyer's group) would make the FENet even more compelling, but these are non-causal algorithms so it is not necessary to do this.

Stylistic Issues:

Line 117: This line doesn't really make sense. Explain what the exploration was and what it meant

Line 150 explain more why this is the case.

Fig 1b. Fixed mapping is a bit confusing. Its a learnable network right? Different terminology is needed.

Fig 1e utah arrays are way too big? Should be more to scale.

Fig1f can see participants face, make sure that's okay.

Fig 3: Debaucheries wavelets instead of debauchies wavelets in one fig caption

Figure 5: Line 819: Robust brain-computer interface (BMI) - should be BCI

Line 401: doesn't make sense grammatically

Fig 4: Consider different visualization, the grey colors used look sloppy and differing grid sizes make this a bit hard to read.

Reviewer #6 (Report for the authors (Required)):

The work presents a novel subject and task independent general-purpose feature extraction method for decoding of (motor) intent for brain-machine interface control applications. The performance of the new method, based on a 1-D convolutional neural network (CNN) architecture, is evaluated by several offline simulations and online experiments using data recorded from tetraplegic participants. One participant participated in a double-blind online experiment study, conducted to compare the new method with state of the art threshold decoder. Authors conclude that the new method improves by at least 50% all the tested performance metrics.

The work exploits the strengths of convolutional neural networks in identifying features in the re-referenced and 80 Hz high-pass filtered time series neural data. The approach to reduce the hyper-parameters of the method by using the same for all channels and the 2-step joint optimisation of feature extractor and decoder (behaviour) seems to contribute towards a generalised solution that works between subjects and tasks. The authors have taken a smart approach, well done.

The methodology used, and the justifications and argumentations of the authors are sound and comprehensible. There is a lot going on in the paper, which makes it a challenging read. However, the methods are described in enough detail and code will be made available after acceptance.

The proposed method, if the results are reproducible, represents a significant methodological and translational contribution to the field, and the implications can be manifold. In particular, if the method is able to provide neurophysiological plausible explanations and allows us to gain new insights into brain functioning.

There are, however, some issues that authors need to address in order to improve the quality of the work and the results.

1. First, the authors use qualitative terms rather than quantitative values. For example, the title claims the

new method “dramatically” improved control. The reviewer understands this makes a great title, however, also leaves a lot of room for interpretation. Authors are encouraged to be specific and clear, and not use such superlatives if possible.

2. Secondly, and most importantly, authors use the method as a black box and do not address explainability. The article would have more impact and readers would be more likely to trust the results if the authors could show why there was an improvement in control performance.

Please find more comments and suggestions below. The numbers at the beginning refer to the line numbers in the manuscript:

3. #67-72: Why these considerations and the given order? The computational complexity, consideration 3, was not addressed or discussed in enough detail in the work. Related to consideration 4), authors are encouraged to commend and discuss the generalizability of the method and their pro and cons in more detail.

4. #85: Authors state “improvement of ~50%”. Can the reviewer suggest that authors provide more context or a baseline performance, so that reader understand what this means.

5. #87: Authors analysed data from 3 patients. How confident are they that FENet generalises over regions, patients and task? This should be discussed in more detail. See comment above.

6. #121: “Dramatically” Please use quantities whenever possible and not qualitative statements. Or precisely define what “dramatically” means.

7. #187/Fig4: (a) Data from year 2019 yields 0.15 higher R^2 compared to data from other years. Why is this the case? And why is this not the case in Fig4(c)?

8. #288-304: What was required of the participants to gain control? Were instructions given on what specifically to do? What to intend? Adding this information would give a clearer picture of the task.

9. #332: typo, replace “xis” with “x is”

10. #363/425: Please clarify what the “ground truth values” are.

11. #498: It would be helpful if at least one dataset could be released so that colleagues can test the code on the data.

Wed 22 Nov 2023

Decision on Article NBME-23-0067A-Z

Dear Mr Haghi,

Thank you for your patience in waiting for the feedback on your revised manuscript, "Convolutional neural network feature extraction improves brain-machine interface control for tetraplegic participants", which has been seen by the original reviewers. However, despite our chasing efforts, Reviewer #5 has not provided further feedback, and it is unlikely that they will. At the end of this message, I have included the report from Reviewer #3, in addition to the two reports that I had already forwarded to you.

You will see that Reviewers #3 and #6 appreciate your efforts to address their earlier criticisms, and that they offer a few additional technical suggestions and requests for clarification. However, Reviewer #4 remains unconvinced that the performance of the model is substantially better than those in your refs. 20 and 30. I hope that you can provide a thoughtful reply to the reviewer's points.

As before, when you are ready to resubmit your manuscript, please [upload](#) the revised files, a point-by-point rebuttal to the comments from all reviewers, the [reporting summary](#), and a cover letter that explains the main improvements included in the revision and responds to any points highlighted in this decision.

As a reminder, please follow the following recommendations:

- * Clearly highlight any amendments to the text and figures to help the reviewers and editors find and understand the changes (yet keep in mind that excessive marking can hinder readability).
- * If you and your co-authors disagree with a criticism, provide the arguments to the reviewer (optionally, indicate the relevant points in the cover letter).
- * If a criticism or suggestion is not addressed, please indicate so in the rebuttal to the reviewer comments and explain the reason(s).
- * Consider including responses to any criticisms raised by more than one reviewer at the beginning of the rebuttal, in a section addressed to all reviewers.
- * The rebuttal should include the reviewer comments in point-by-point format (please note that we provide all reviewers with the reports as they appear at the end of this message).
- * Provide the rebuttal to the reviewer comments and the cover letter as separate files.

We look forward to receive a further revised version of the work. Please do not hesitate to contact me should you have any questions.

Best wishes,

Pep

Pep Pàmies
Chief Editor, [Nature Biomedical Engineering](#)

Reviewer #3 (Report for the authors (Required)):

The authors should be commended for their thorough response to a lengthy review. Most of my previous comments have been addressed. I have a few remaining comments that I believe the authors could quickly address.

Major comments:

1. Bitrate: from the definition of bitrate now given, it appears that the metric was not set up to penalize incorrect target selections. This differs from the standard "Achieved bitrate" metric definition ([1], and see also [2] for a discussion of the importance of assessing error rates in information-based performance metrics). This formulation breaks the idea of the metric being used as a measure of channel capacity / information transfer rate for a discrete set of symbols, as measuring information transferred requires knowing when symbols were incorrectly transferred also.

If incorrect selections were possible, the bitrate metrics should be corrected to account for that.

If incorrect selections were not possible, the bitrate numbers from this manuscript will not be directly comparable to other measurements in the literature. Therefore this discrepancy should be clearly stated in the main text.

2. Previous comment 12 on non-causal filtering - this comment seems to have been ignored. (as noted, non-causal filtering was a more important comparison point than the exponential smoothing that the authors seem to have focused on). Please provide a comparison of FENet features and the threshold crossing features extracted via non-causal filtering.

Minor comments:

1. 663-681: ReFIT-KF methods are absent. Evaluation was done on open-loop kinematics, so it is unclear how the authors estimated intention to recalibrate the decoder for use with open-loop data. Did the intention come from subsequent closed-loop trials? Or was this just a standard Kalman filter? If the former, please state so in the methods. If the latter, please update the text throughout to accurately reflect the decoder.

2. 199-200: The authors claim FENet improves SNR, but the evidence (from Extended Figure 8) is qualitative. I suggest computing SNRs to quantitatively show FENet improves SNR, or remove the claim about SNR. Perhaps the conclusion comes from FENet features yielding higher values in the tuning curves, but higher feature values should not be confused with higher SNR (i.e. a constant gain increases feature values but does not improve SNR).

3. 306-309: "Furthermore, it is worth noting that future BMI systems may not universally support the capture of raw, high-sample rate broadband data (e.g., Neuralink) to enable training of new per-channel models. In such cases, our approach can be seamlessly utilized without the reliance on this type of data." - Assuming "this type of data" refers to broadband data, it is not immediately clear to readers how FENet can be "seamlessly" adapted to systems that do not record broadband data, since only broadband data was discussed in the manuscript. Please discuss how FENet could be adapted (i.e. to systems that only provide spike times, like Neuralink, or data collected with lower sampling rates).

4. In the authors' response to comment 2, they stated:

`Response: After extracting the features from the 30ms windows, we averaged the resulting windows, with the number being averaged defined by the bin length. This will result in a smoothing effect, as explained in figure 5's caption:

"Extended Figure 5. The effect of changing the window size after averaging the features over larger window sizes.`

I could not understand what this statement means "The effect of changing the window size after averaging the features over larger window sizes". Please revise the text to clarify

[1] <https://ieeexplore.ieee.org/document/6891265>

[2] <https://iopscience.iop.org/article/10.1088/1741-2560/10/2/026014>

Reviewer #4 (Report for the authors (Required)):

The authors did substantial offline analysis to compare to different features already in the literature and common use, which did indeed make up a lot of the distance between threshold crossing and FENET. In

doing so, I think it is clear that the headline results, and the only online data in the paper, was a comparison of a new method with an outdated method rather than the state of the art. No new data was added to show that there is still a big online performance increase compared to the state of the art, and it's inappropriate to include data comparing to an older approach. It does seem like all reviewers agree on this point. Similarly, the absolute performance is not high enough to conclude that there is something fundamentally new here, since it still remains below the best performance seen in the literature. It's true that this can be from a variety of factors, but that makes an online comparison within subject all the more critical. Overall, while there are interesting offline observations that are of interest to the neural engineering community, without a clear and substantial performance improvement online, this is better suited for publication in a more specialty journal.

Seeing what the FENet actually produces compared to the other features, it appears to be doing basically an emphasis of spikes, like wavelets and MUA. The fact that the authors can do it with 30 ms of data instead of 50ms or more is of interest, but again more suited to a specialty journal. I agree with the other reviewer that the approach here is very close to Willsey et al which has a similar convolutional layer that may be achieving the same end in fewer operations, and this should be checked. The MUA presented here in Extended Figure 8 does seem oddly close to TCs in SNR, which is not consistent with the observation of others. Zhang et al represents a post-hoc offline analysis, so I recommend replicating the details of MUA implementation from any study that utilized these features online, which may involve shorter timebins and a higher highpass.

Reviewer #6 (Report for the authors (Required)):

The reviewer thanks the authors for taking into account the reviewer's comments and adding information that improves the clarity of the paper and the significance of the findings. Most of the comments have been addressed satisfactorily. Please find the reviewers reply to the authors rebuttal on the major and selected minor comments and a few more questions below:

Major Comment 1:

The tone of the revised manuscript was changed to leave little room for interpretation of the results, thus improving clarity.

Major Comment 2:

The added paragraphs and new figures reporting on feature extraction and the explainability of the extracted features are very helpful. The filtering properties shown in the expanded Figure 9 seem to indicate that the CNN does what one would expect when analysing neural signals.

The WT Filter shown in Fig. 9 seems to suggest that authors used a one-level decomposition, and not a multi-level decomposition. Is this the case? If so, is this standard in the field?

On page 5, paragraphs 2 and 3, the authors use the "best electrode" or "top electrode" from selected datasets to show the effectiveness of the proposed FENet feature extraction capabilities. It is not clear to the reviewer how the top electrodes were selected. What criterion was used (the one described in the extended Figure 6 caption?), and was the best electrode for each feature extraction method used independently for comparison, or the best electrode with respect to FENet? Also, how were the sessions/datasets chosen? Were the data selected arbitrarily, or are they a good general characterization of the data? Are they part of the evaluation data set?

Minor Comment 1:

Thanks for adding a section with equations to estimate the computational complexity of the method expressed in the number of operations needed.

Minor Comment 2:

Thank you for clarifying the baseline and for including other methods in the comparison.

Minor Comment 3 & 9:

Thank you for adding a statement on clarifying the baseline and for including other methods in the comparison. Making data and code available is very welcome, increases confidence in the authors work, and allows readers interested in the method to get a better understanding of the methods and compare them with others.

Thu 19 Sep 2024

Decision on Article NBME-23-0067B

Dear Mr Haghi,

Big apologies for the really lengthy round of peer review for your revised manuscript, "Convolutional neural network feature extraction improves brain-machine interface control for tetraplegic participants". One of the reviewers, who we had chased for a few months, finally provided their feedback a couple of days ago; another of the reviewers (Reviewer #4) did not provide a report, yet we are overall satisfied and in agreement with the essence of the two reports below. I am therefore pleased to write that we shall be happy to publish the manuscript in *Nature Biomedical Engineering*, provided that the points specified in the attached instructions file are addressed.

When you are ready to submit the final version of your manuscript, please [upload](#) the files specified in the instructions file.

Also, please consider the remaining minor points from the reviewers.

We encourage authors to take up [transparent peer review](#). If you are eligible and opt in to transparent peer review, we will publish, as a single supplementary file, all the reviewer comments for all the versions of the manuscript, your rebuttal letters, and the editorial decision letters. **If you opt in to transparent peer review, in the attached file please tick the box 'I wish to participate in transparent peer review'; if you prefer not to, please tick 'I do NOT wish to participate in transparent peer review'.** In the interest of confidentiality, we allow redactions to the rebuttal letters and to the reviewer comments. If you are concerned about the release of confidential data, please indicate what specific information you would like to have removed; we cannot incorporate redactions for any other reasons. [More information on transparent peer review is available.](#)

Please do not hesitate to contact me should you have any questions. Once more, we apologize for the exceptionally lengthy turnaround time, and for failing to reply to your chasing messages (my workload has been over the roof for many months, as we were recruiting and training new editors while facing historical records of new submissions; unfortunately I routinely receive more e-mail messages than I can possibly reply to).

Best wishes,

Pep

Pep Pàmies
Chief Editor, [Nature Biomedical Engineering](#)

Reviewer #3 (Report for the authors (Required)):

Overall the authors have largely addressed my concerns. My one piece of feedback is that the authors seem to have misunderstood some of the recent literature with this comment:

> Our primary focus within this study was to test techniques that can be applied in a causal real-time system. Thus, while non-causality could potentially enhance a decoder's performance operating on threshold crossing features by employing a filter over the data twice - once forward and once in reverse, thereby providing future information to the feature extraction technique, we think such a comparison is outside the scope of the current manuscript.

The non-causal filtering approach described in Masse et al. is used in high-performing causal real-time systems, it simply requires a small amount of buffering of the signal (e.g. a few milliseconds). To name a few early examples, it is used online in Jarosiewicz et al., Science Trans Med 2015, Gilja et al., Nature Med

2015, and Pandarinath et al., eLife 2017. It has been used in many subsequent demonstrations.

Therefore, when the authors state:

> While non-causality might offer potential performance enhancements by employing a forward-reverse data filtration process, it necessitates defining a system delay to enable its applicability. Such delay contradicts the requirements of a real-time system, particularly when the feature extraction window size is comparable with the delay required by the non-causal filtering techniques.

This assertion is clearly inaccurate given the literature, and should be removed. Given the prominence of this filtering approach in the clinical BCI literature, it would be preferred that the authors' do the requested comparison rather than claiming it is out-of-scope.

Reviewer #6 (Report for the authors (Required)):

The authors have revised the work and have adequately addressed all the reviewer's comments and suggestions. The performance of the novel CNN-based feature extraction network is clearly presented, and the advantages of the proposed approach are demonstrated. There are no technical issues, and the methods are sound. The data basis is limited, but this is a common challenge in this research area and comparable with other works with invasive BMIs.

I would just like to make one small comment. In lines 186-189, it would be helpful if the authors could add a statement confirming that the comparisons are fair and not just done with some standard parameters, but optimised for each method.

Nature Biomedical Engineering is a Transformative Journal. Authors may publish their research with us through the traditional subscription access route, or make their paper immediately open access through payment of an article-processing charge. More [information about publication options](#) is available.

You may need to take specific actions to [comply](#) with funder and institutional open-access mandates. If the work described in the accepted manuscript is supported by a funder that requires immediate open access (as outlined, for example, by [Plan S](#)) and your manuscript was originally submitted on or after January 1st 2021, then you will need to select the gold OA route. Authors selecting subscription publication will need to accept our standard licensing terms (including our [self-archiving policies](#)), and these will supersede any other terms that the author or any third party may assert apply to any version of the manuscript.

Rebuttal 1

Response to Reviewers

Reviewer 3

Major Comments

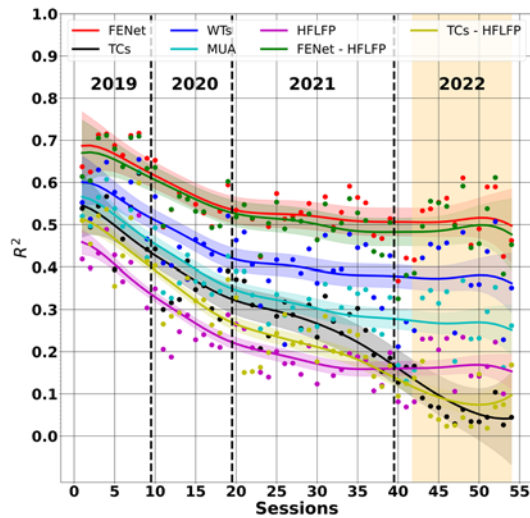
Comment 1: What decoding strategy was used for threshold crossings (TCs) and wavelet transforms (WTs)? Only the decoding strategy for FENet was indicated.

Response: In the original manuscript, the kinematics were decoded using a linear decoder across all feature extraction techniques. Decoding neural signals into control kinematics typically involves two stages: one, feature extraction – the process of transforming raw electrical recordings into neural features; and two, feature decoding – the process of decoding neural features into the behavior of interest (e.g. cursor kinematics). This manuscript seeks to innovate a new solution to the feature extraction process. Therefore, we held the feature decoding stage constant across all feature extraction techniques to minimize confounds to interpretation. In the revised manuscript, following the suggestions of reviewers, we have included the performance of additional feature decoders, namely Long-Short Term Memory Recurrent Neural Network (LSTM), Support Vector Regressor (SVR), Recalibrated Feedback Intention-Trained Kalman filter (ReFIT-KF), and Preferential Subspace Identification (PSID). This was done to ensure that improvements in our novel feature extraction method generalize across feature decoding methods. To provide a comprehensive evaluation of different feature extraction techniques, we present the performance of these decoders in Figure 4.

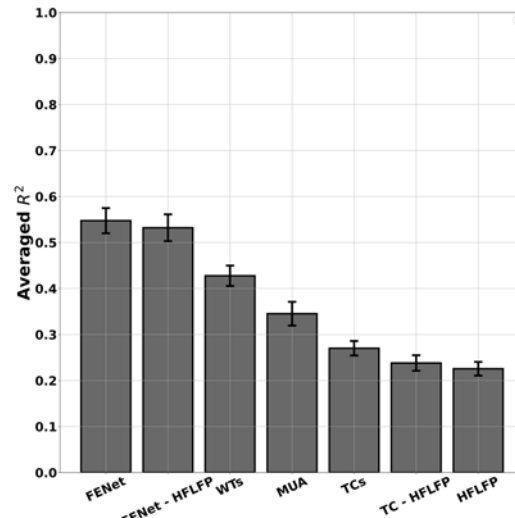
We have added this explanation to the “FENet provides improved open-loop decoding performance” section:

To ensure that improvements in our novel feature extraction method generalize across feature decoding methods, we have also included the performance of additional feature decoders, namely Support Vector Regression^{36–38}, Long-Short Term Recurrent Neural Network (LSTM)³⁹, Recalibrated Feedback Intention-Trained Kalman filter (ReFIT-KF)^{5,24}, and Preferential Subspace Identification (PSID)⁴⁰. Figure 4 provides a comprehensive evaluation of the performance of these decoders operating on different feature extraction techniques. As we see in this figure, FENet improves decoding R^2 compared to the other feature extraction techniques for all the tested decoders.

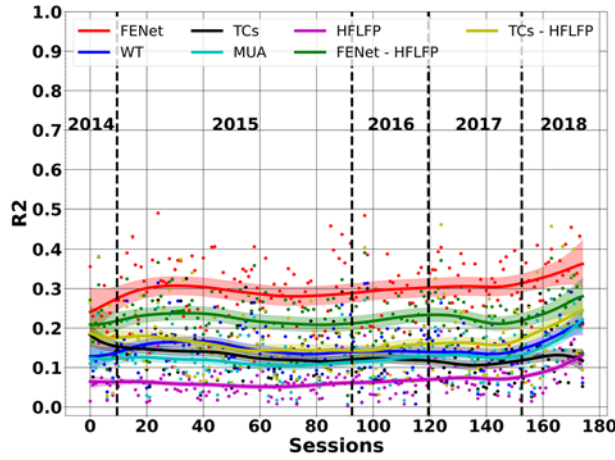
Figure 3 a-d have been replaced by the below figures, including different feature extraction techniques:



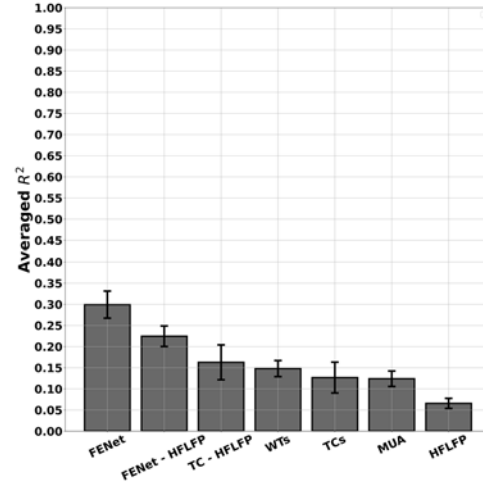
(a)



(b)



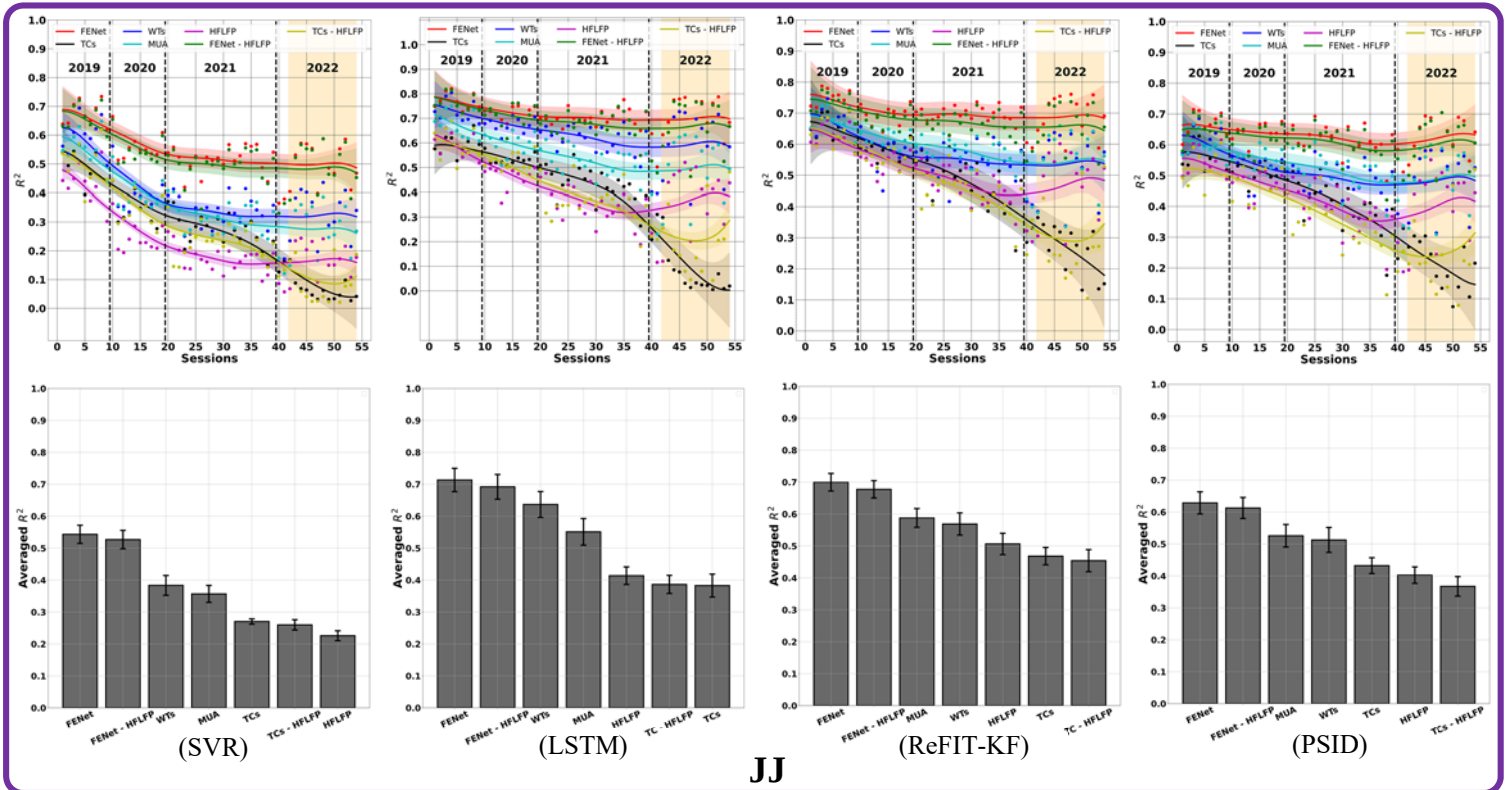
(c)



(d)

Figure 3. Open-loop multi-electrode performance on FENet, [threshold crossing events \(TCs\)](#), [Debaucheries wavelets \(WTs\)](#), [Multi-Unit Activities \(MUA\)](#), [High Frequency LFP \(HFLFP\)](#), and the combination of FENet and TCs with HFLFP (FENET – HFLFP and TCs – HFLFP) on (a-b) the research participant JJ over 54 recorded sessions spanning from 2019 to 2022 (shaded region shows the closed-loop sessions) (c-d) on the research participant EGS over 175 recorded sessions spanning from 2014 to 2018. The dashed lines separate the sessions of different years. The band in each time series shows the range of its 95% confidence interval of a LOESS^{48,49} fit.

Also, a new main figure (Figure 4) has been added to present the performance of other decoders operating on different feature extraction techniques:



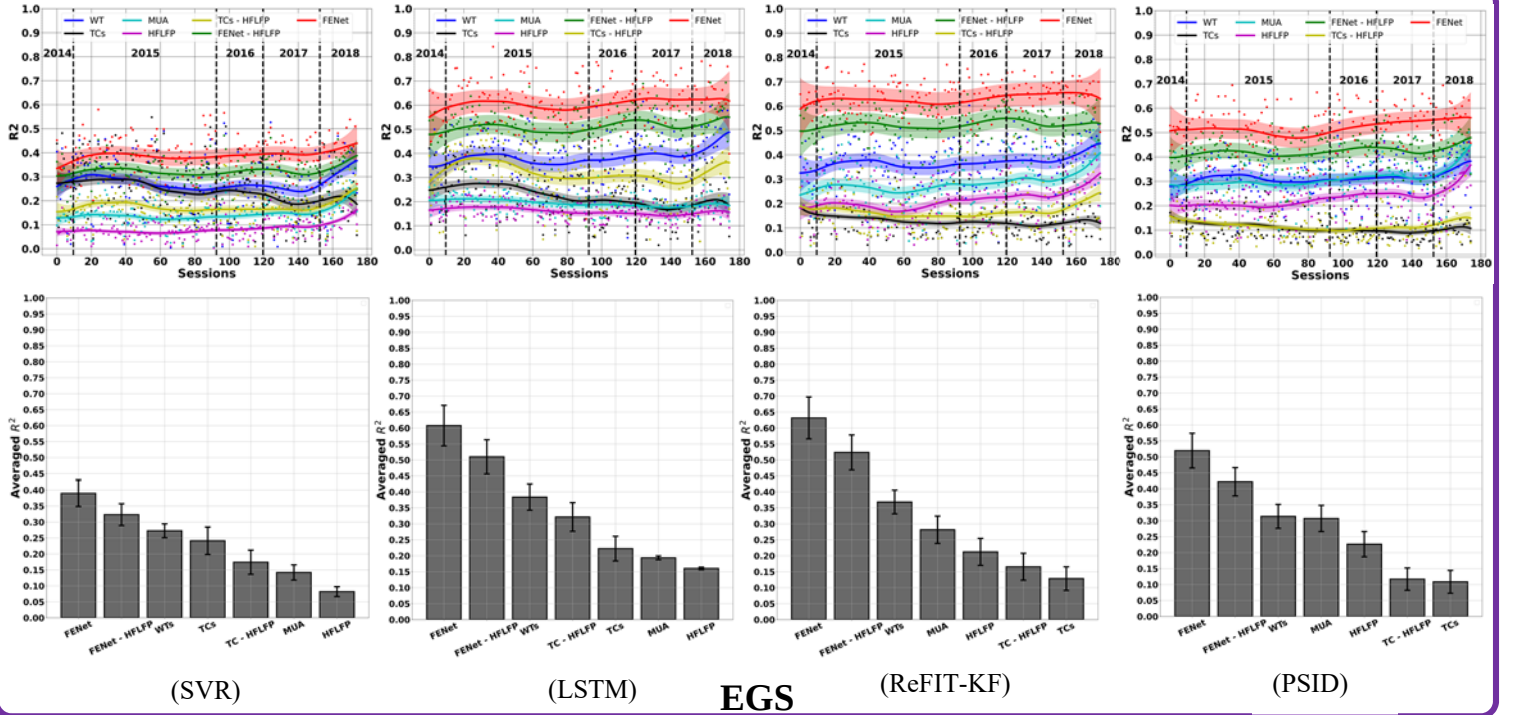


Figure 4. FENet features improve off-line decoding performance across all tested decoding algorithms. The upper and lower panels display the decoder performance for participants JJ and EGS, respectively, in a center-out task. Performance is shown for 5 total decoding algorithms, each applied to features generated by 7 feature extraction techniques. The analyzed feature extraction techniques include FENet (red), Wavelet Transform with db20 mother wavelet (WTs) (blue), Threshold Crossings (TCs) (black), Multi-Unit Activity (MUA) (cyan), High-Frequency Local Field Potentials (HFLFP) (purple), the combination of FENet and HFLFP (green), and the combination of TCs and HFLFP (yellow). The tested decoding algorithms are Support Vector Regression (SVR) (first column), Long-Short Term Recurrent Neural Network (LSTM) (second column), Recalibrated Feedback Intention-Trained Kalman filter (ReFIT-KF) (third column), and Preferential Subspace Identification (PSID) (fourth column). The performance of every decoding algorithm was improved when trained on FENet features versus features generated using the other feature extraction techniques.

Furthermore, we have included a detailed description of each decoder in the methods section for the sake of clarity and comprehensiveness:

Decoders. To evaluate the performance of different feature extraction techniques, we passed them to different types of decoders, including the Linear Decoder^{23,67}, Support Vector Regression³⁶⁻³⁸, Long-Short Term Recurrent Neural Network (LSTM)³⁹, Recalibrated Feedback Intention-Trained Kalman filter (ReFIT-KF)^{5,24}, and Preferential Subspace Identification (PSID)⁴⁰.

The Linear Decoder uses a standard linear regression model where we can predict kinematics ($\hat{y}$) from the extracted neural features (u) by using:

$$\hat{y} = b + \sum_{i=1}^N W_i u_i \quad \text{Equation 4}$$

We find the weights W_i and the bias term b through a least squares error optimization to minimize mean squared error between the model's predictions and ground-truth kinematics during training. The parameters are then used to predict new kinematics given extracted neural features during test.

Support vector regression (SVR) is the continuous form of support vector machines where the generalized error is minimized, given by the function:

$$\hat{y} = \sum_{i=1}^N (\alpha_i^* - \alpha_i) k(u_i, u) + b \quad \text{Equation 5}$$

where α_i^* and α_i are Lagrange multipliers and k is a kernel function, where we use the radial basis function kernel in this paper. The Lagrange multipliers are found by minimizing a regularized risk function:

$$R_{reg}[f] = \frac{1}{2} ||w||^2 + C \sum_{i=1}^l L_\varepsilon(y) \quad \text{Equation 6}$$

where $||w||^2$ represents the model complexity, C is a constant that determines the trade-off between the ε insensitive loss function $L_\varepsilon(y)$. For SVR, we employed a RBF kernel with C set to 1.

It is well-known that Simple RNN⁷³ units cannot remember long term dependencies in sequential data because of the vanishing gradients problem. Another version of RNNs that is widely used in the literature are RNNs with Long-Short Term Memory (LSTM) units. By denoting $\circ$ as Hadamard product, the LSTM is defined as:

$$\begin{cases} f_k = \sigma(W_{fu}u_k + W_{fr}r_{k-1} + b_f) \\ i_k = \sigma(W_{iu}u_k + W_{ir}r_{k-1} + b_i) \\ o_k = \sigma(W_{ou}u_k + W_{or}r_{k-1} + b_o) \\ c_u = \tanh(W_{cu}u_k + W_{cr}r_{k-1} + b_c) \\ c_k = f_k \circ c_{k-1} + i_k \circ c_{k-1} \\ r_k = o_k \circ \tanh(c_k) \\ \hat{y}_k = W_{yr}r_k + b_y \end{cases} \quad \text{Equation 7}$$

r_k is the hidden state as in Simple RNN, c_u is the output from the cell update activation function, c_k is the LSTM cell's internal state, f_k , i_k , and o_k are the output matrices from the respective forget, input, and output activation functions, which act as the LSTM's gates, W and b represent the weights and biases, and σ is the sigmoid function. Following our parameter sweeps, the settings for the number of layers, the number of recurrent nodes, and the history of LSTM were determined as 1, 50, and 10, respectively.

The Kalman Filter combines the idea that kinematics are function of neural firings as well as the idea that neural activity is a function of movements, or the kinematics. This can be represented by two equations:

$$\begin{cases} \hat{y}_{k+1} = A_k \hat{y}_k + w_k \\ u_k = H_k \hat{y}_k + q_k \end{cases} \quad \text{Equation 8}$$

These represent how the system evolves over time as well as how neural activity is generated by the system's behavior. The matrices A , H , Q , and W can be found through a training process (where $q \sim \mathcal{N}(0, Q)$ and $w \sim \mathcal{N}(0, W)$). Using properties of the conditional probabilities of kinematics and neural data, we get a closed form solution for maximizing the joint probability $p(Y_M, U_M)$. Using the physical properties of the problem, we get matrix A to be of the form:

$$A = \begin{bmatrix} 1 & 0 & dt & 0 \\ 0 & 1 & 0 & dt \\ 0 & 0 & a & b \\ 0 & 0 & c & d \end{bmatrix} \quad \text{Equation 9}$$

where A_v is defined as:

$$A_v = \begin{bmatrix} a & b \\ c & d \end{bmatrix} = V_2 V_1^T (V_1 V_1^T)^{-1} \quad \text{Equation 10}$$

V_1 consists of the velocity kinematics points except for the last time step, V_2 consists of the velocity kinematics points except for the first time step, and dt is the time step size used (in our case, 30ms for JJ and NS, and 50ms for EGS).

Furthermore, W is a zero matrix with the matrix $W_u = \frac{1}{N-1} (V_2 - AV_1)(V_2 - AV_1)^T$ in the bottom right corner. H and Q are given by:

$$\begin{cases} H = U^T Y (Y Y^T)^{-1} \\ Q = \frac{1}{N} (U - H Y)(U - H Y)^{-1} \end{cases} \quad \text{Equation 11}$$

Then, we can use the update equations:

$$\begin{cases} \hat{y}_k^- = A\hat{y}_{k-1} \\ P_k^- = AP_{k-1}A^T + W \\ \hat{y}_k = \hat{y}_k^- + K_k(u_k - H\hat{y}_k^-) \\ P_k = (1 - K_kH)P_k^- \end{cases} \quad \text{Equation 12}$$

Here, P is the covariance matrix of the kinematics. K_k , the Kalman filter gain is given by:

$$K_k = P_k^- H^T (HP_k^- H^T + Q)^{-1} \quad \text{Equation 13}$$

PSID models the brain's state as a high-dimensional latent variable influencing neural activity and behavior. PSID, an algorithm built upon the Kalman Filter equations, utilizes a dynamic linear state space model to describe the association between the latent state and the recorded neural activity (u_k) and behavior (y_k). The model consists of a latent state $x_k \in \mathbb{R}^{n_x}$, which includes behaviorally relevant ($x_k^{(1)} \in \mathbb{R}^{n_1}$) and irrelevant ($x_k^{(2)} \in \mathbb{R}^{n_2}$) components as below:

$$\begin{cases} x_{k+1} = Ax_k + w_k \\ u_k = C_y x_k + v_k \\ y_k = C_z x_k + \epsilon_k \end{cases}, \quad x_k = \begin{bmatrix} x_k^{(1)} \\ x_k^{(2)} \end{bmatrix} \quad \text{Equation 14}$$

PSID employs a two-stage identification approach. In the first stage, it directly learns the behaviorally relevant component ($x_k^{(1)}$) from training data without simultaneously learning the irrelevant component ($x_k^{(2)}$), which is optional in the second stage. This prioritization enables PSID to learn behaviorally relevant neural dynamics using low-dimensional states (only $x_k^{(1)}$). Similar to Kalman Filter, the PSID model formulation includes noise terms (ϵ_k , w_k , and v_k) representing behavior dynamics not present in the recorded neural activity. The parameters of the model (A , C_y , C_z , and noise statistics) are learned by PSID using training samples of neural activity and behavior⁴⁰. After the parameter sweep, we adjusted the latent space dimension to 10.

Comment 2: *What type of smoothing was applied (i.e. for the comparisons in Extended Figure 6)?*

Response: After extracting the features from the 30ms windows, we averaged the resulting windows, with the number being averaged defined by the bin length. This will result in a smoothing effect, as explained in figure 5's caption:

“Extended Figure 5. The effect of changing the window size [after averaging the features over larger window sizes](#). FENet performance is robust against the change of the recording window size length in the center-out trajectory task. (a) The averaged R^2 [of a linear decoder operating on](#) FENet and TCs when we increase the feature extraction window size, which has a smoothing effect on the extracted features.”

We have revised lines 172 to 178 as follow:

“We evaluated the open-loop results with FENet using neural data and behavior binned at a fine temporal resolution (30ms bins) and without smoothing the extracted features. This was motivated by our primary concern that FENet be maximally useful for closed-loop control where smoothing decreases the responsiveness of the closed-loop system by using potentially outdated neural information. However, recognizing that FENet could also be used for slow-timescale applications, we tested how FENet performed against TCs when smoothing the extracted features [from a larger window size](#). Extended Figure 6a, b show the robustness of FENet against the change in the window size used to update the feature extraction process in our trajectory tasks.”

Comment 3: *How was partial least-squares regression applied? Since this was a critical step in deciding the final FENet architecture, more details regarding how it was applied would be helpful.*

Response: To add more details about the partial least square regression, we have added the following text to a new section in methods named “Preprocessing of the generated features for open- and closed-loop analysis”:

Each feature extraction technique results in a different number of features per channel. For example, TCs result in one feature per electrode while both FENet and WTs result in 8 features per electrode. The increase in number of features per electrode necessitates dimensionality reduction to prevent decoder overfitting given the limited amount of training data for individual sessions. To address this concern, we utilized Partial Least Square Regression⁶⁶ (PLSR). Specifically, PLSR was applied to the features extracted from each channel, independently, to condense the 8 extracted features obtained from each electrode into a smaller set of features, specifically 2 features in this case. Importantly, the mapping from 8 features per electrode to 2 features

per electrode was learned from the training data and applied to the test data within our cross-validation architecture and was performed in the same way for FENet and WTs based features.

Comment 4: *There are many details of the finger task that are not discussed in the methods despite being a main figure. For example, what do muscle model and somatotopy model mean? What task(s) and subject(s) was FENet trained on to test generalization to the finger task? What do RDA and RDM mean?*

Response: To explain the muscle model and somatotopy model, we have added this to the “Behavioral tasks” section of the methods, where we explain the finger grid task:

To model the multi-finger tasks, we considered the muscle model and somatotopy model⁴⁷. The muscle activation model posits that the representational structure should align with the coactivation patterns observed in muscle activity during individual finger movements. Conversely, the somatotopy model suggests that the representational structure should correspond to the spatial arrangement of the body, wherein neighboring fingers exhibit similar representations. Although somatotopy typically pertains to physical spaces resembling the body, in this context, we employ the term broadly to encompass encoding spaces that resemble the body⁴⁶.

To explain RSA and RDA, we have added this to the “Open-loop evaluation measures” section of methods to explain the measures for the finger grid task:

To assess the performance on the finger-grid task⁴⁶, we employed the framework of representational similarity analysis (RSA)^{75,76} and representational dynamics analysis (RDA)⁷⁷. RSA quantifies the neural representational structure by measuring the pairwise distances between the neural activity patterns associated with each finger. These distances are used to construct the representational dissimilarity matrix (RDM), which provides a concise summary of the representational structure. Notably, these distances are independent of the original feature types, such as electrode or voxel measurements, enabling us to compare finger organizations across subjects and different recording modalities⁷⁸. Additionally, we utilized representational dynamics analysis (RDA) to explore the temporal evolution of the representational structure. This involved modeling the representational structure of finger movements at each timepoint as a non-negative linear combination of potentially predictive models.

Comment 5: *How were bitrates calculated?*

Response: We have added this to the “Closed-loop evaluation measures” section of the methods:

In addition, for the grid-task, we have included the bitrate in our findings. The calculation of the bitrate is outlined below⁷ (Supplemental Video 2):

$$B = \frac{\log_2(N) \times S_c}{t} \quad \text{Equation 17}$$

where N is the number of total targets on the screen, S_c is the number of completed trials, and t is the time elapsed in seconds.

Comment 6: *For work that is heavily computational, to robustly assess, it is important to have a copy of the code, documentation, example trained models, and example data on which to test the models. This is also critical in assessing the reproducibility of the work. We recommend that the authors provide the code and data in a containerized form so that there are no issues with libraries/software versioning/dependencies, etc. In the paper the authors mention that it is difficult to make the broadband data available because the data are “too large and confidential to be publicly shared.” However, it is not required to provide all data, some sample datasets to train and test the models and reproduce key conclusions should be sufficient. I was unclear what the “confidential” concern referred to. If there is a concern about participant privacy, there should be ways of anonymizing small stretches of data sufficient for replicating the work. Alternatively, potentially the authors could use previously-recorded monkey data from the same lab or open source, publicly-available broadband data (e.g., https://zenodo.org/record/3854034#.Y_0RbuzMJeg). There should be many possibilities for addressing this concern.*

Response: We have uploaded 11 sessions (JJ) and 4 sessions (EGS) of broadband data for offline analysis, the extracted features for these sessions, and the code have been uploaded at

https://drive.google.com/drive/folders/1wGZ4BSZTsoKH8p-rgSFOhUJA4A3_g7lh?usp=drive_link.

After acceptance, the code will become publicly available at github.com/BenyaminHaghi/FENet

Comment 7: *Related, in the methods, the authors should specify all software and library version info required to reproduce this work.*

Response: We have discussed a part of software requirements in the “Algorithm implementation requirements” section of the discussion (see below). Moreover, we have shared a readme file that includes more details about the software and library version info as well as the python kernel that can be installed and used.

Algorithmic implementation requirements. We used PyTorch, a deep-learning API for Python⁷⁹, as the programmatical framework to train and operate neural networks. We configured Pytorch to use CUDA, a parallel computing platform and programming model developed by NVIDIA, which can accelerate many of the computations involved in training neural networks with commercially available graphics processing units (GPU)⁸⁰. For offline training and evaluation of the FENet, we used a single Tesla V100 GPU⁸⁰ **and for the closed-loop runs, we used a single NVIDIA GeForce RTX 3080 GPU⁸⁰.**

Comment 8: The authors claim that a major design decision for FENet is that it remains computationally simple and relevant to devices, but there is no data or analysis presented to confirm they achieved this goal. To back up this claim, the authors should include computation time results, as well as estimates of computational complexity (in the form of Big O notation), including the full feature extraction pipeline of FENet against the full feature extraction pipeline of TCs and WTs. Readers should have concrete information on the suitability of FENet for low-power implantable devices.

Response: To answer this question, we have added the following to the discussion section:

It is important to note that FENet was designed to maintain a small computational footprint in comparison to ultradeep RNN feature extraction techniques and other convolutional network designs. This was achieved by extracting features from single electrodes using the same trained parameters for all electrodes. We deliberately constrained the architecture to an algorithm with complexity that allows for computation within 5 milliseconds in closed-loop BMIs. The presented version of FENet, based on the wavelet with db20 mother wavelet architecture described in the paper, consists of only 560 learnable parameters. This significantly reduces its size compared to more complex deep-network alternatives. Additionally, we conducted experiments by swapping hyperparameters of FENet, demonstrating that we can achieve comparable benefits and performance even with a smaller architecture. Exact hardware details will be explored further in our upcoming works, which will focus specifically on hardware implementation.

We have added this section to the methods:

Computational cost of FENet. In order to determine the computational complexity of various FENet architectures, we quantify the total count of multiplicative and additive operations performed for the feature extraction within the network. Assume that S_i , k_i , and s_i are input size, kernel size, and stride of the i^{th} feature engineering module of FENet, respectively. The size of the input for the i^{th} feature engineering module of FENet can be calculated as below⁴⁵:

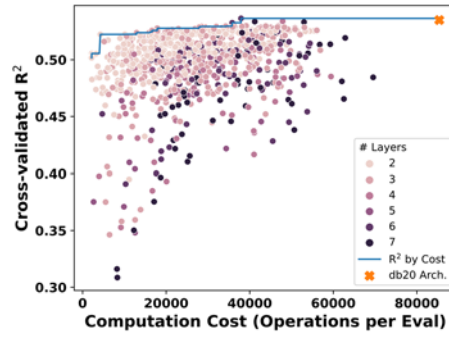
$$S_i = \left\lfloor \frac{S_{i-1} + \max(k_i - s_i, 0) + (k_i - 1) - k_i}{s_i} \right\rfloor \quad \text{Equation 18}$$

where $\max(k_i - s_i, 0)$ and $(k_i - 1)$ represent the left and right paddings, respectively. Then, we can calculate the cost for all the FENet layers as:

$$Cost = \sum_{i=0}^{n-1} 2k_i S_i \quad \text{Equation 19}$$

Given that n represents the quantity of feature engineering modules within the FENet, it is necessary to consider the dual cost incurred by both the upper and lower branches of these modules. As such, the computational cost is effectively doubled to encompass the collective operations of these components.

We have calculated the computational complexity as above and added this figure to Extended Figure 7:



Extended Figure 7. The cross-validated R^2 of linear decoder operating on features extracted by using different FENet architectures vs the computational cost of these different architectures.

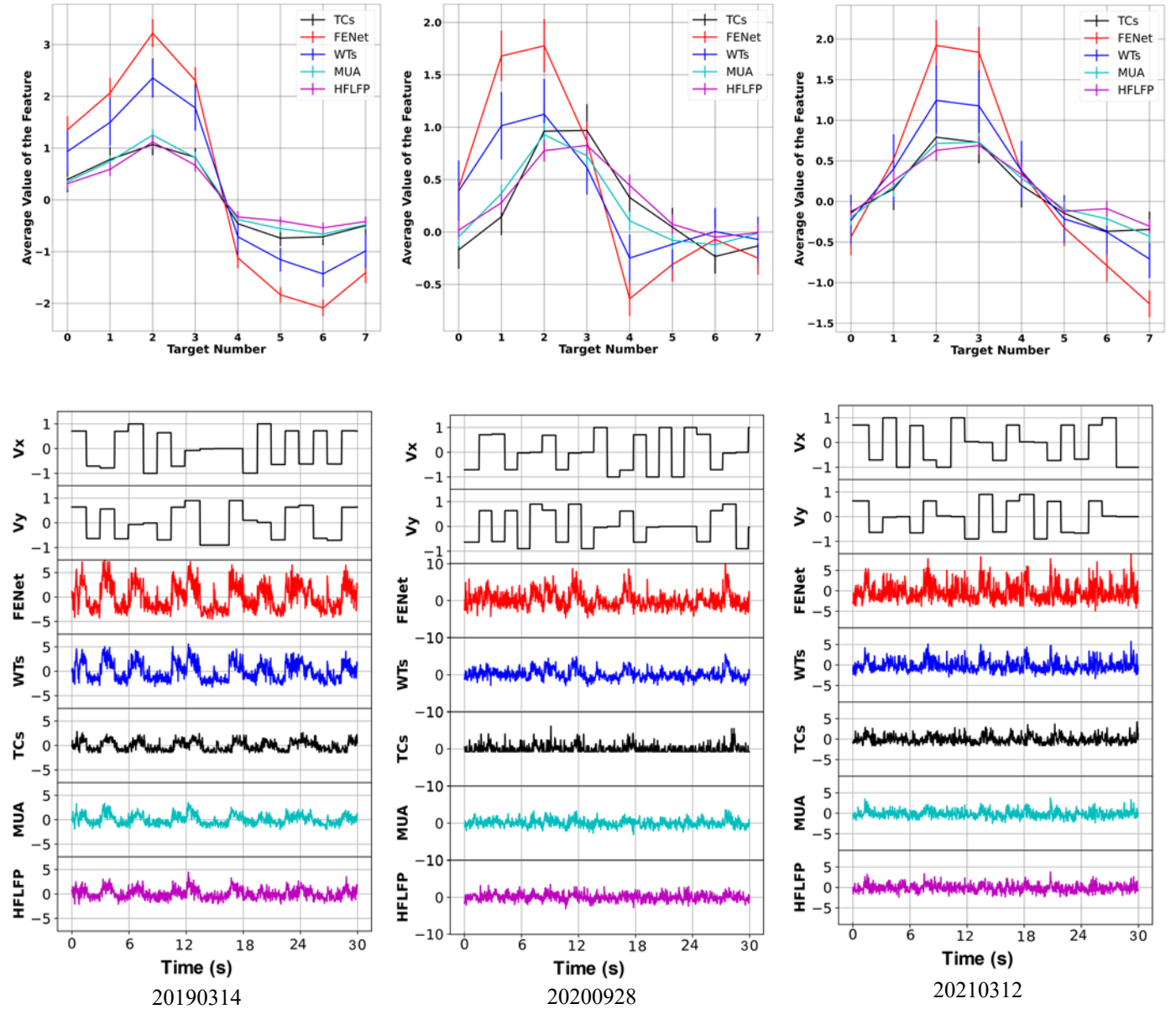
Comment 9: *It is difficult to interpret how the features FENet extracts relate to the underlying neural activity. I recommend that the authors provide snippets of actual features output by FENet and show comparisons to TCs and WTs for the same data snippets, to provide readers with some intuition of what FENet finds are helpful features.*

Response: To increase the interpretability of FENet, we have done multiple extra analyses. We believe that these analyses contribute significantly to our understanding of FENet's learning process relative to other feature extraction techniques. It is important to note that comprehending the precise nature of neural network learning is still an ongoing research endeavor.

- First, we have added this text to the section “FENet provides improved open-loop decoding performance”:

To assess and understand the effectiveness of the extracted features obtained through diverse feature extraction techniques, we conducted a rigorous analysis using offline data of the best electrode from three sample sessions labeled as 20190314, 20200928, and 20210312 (Extended Figure 8). The data of each session consisted of eight distinct trials, each corresponding to a unique target location in a center-out task. To differentiate between these targets, we designated the location where $x > 0$ and $y = 0$ as Target0. Our analysis focused primarily on the feature values derived from the top electrode recorded during these sessions. The results of our analysis demonstrated the remarkable preservation of the fundamental tuning characteristics of the neurons across the various employed feature extraction techniques. Notably, FENet exhibited significant improvements in terms of the Signal-to-Noise Ratio (SNR) and the ability to distinguish individual trials. These findings indicate that FENet provides a more robust and distinctive representation of the neural activity, thereby enhancing the performance for decoding neural signals.

Added “Extended Figure 8” represents the corresponding results of this analysis:



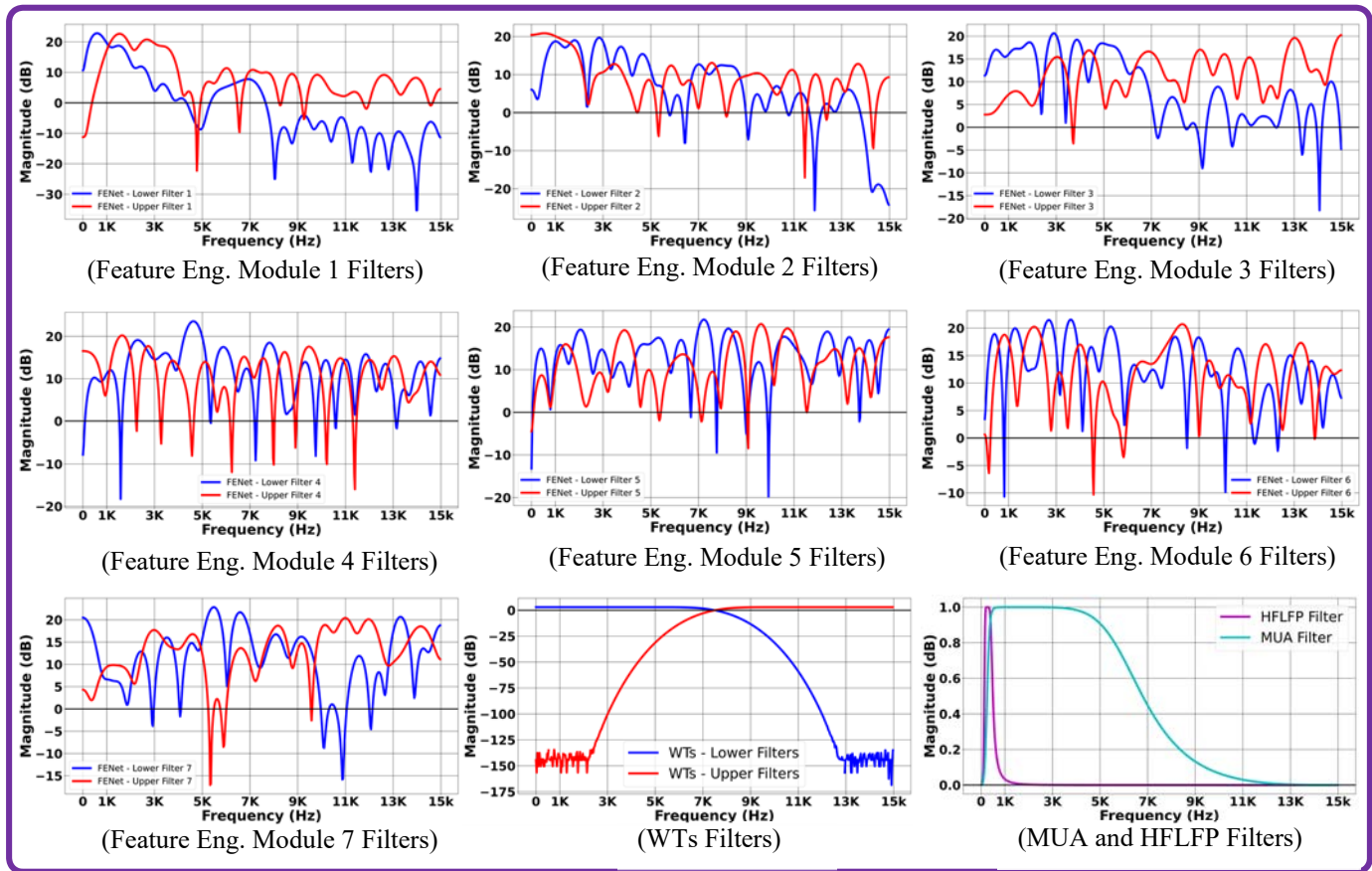
Extended Figure 8. To assess the extracted features obtained from various feature extraction techniques, the offline data of the top electrode of three sample sessions 20190314, 20200928, and 20210312 were partitioned into eight center-out task trials, each trial corresponding to a different target. We named the target with $x>0$ and $y=0$ as Target0. The upper figures depict the average values of features obtained from multiple repetitions of a trial in their respective sessions, while the lower figures display the actual values of the extracted features during the initial 30 seconds of the recorded data. The results demonstrate the preservation of the fundamental tuning characteristics of the neurons, with an enhanced SNR and better trial separability for FENet compared to the other feature extraction techniques.

- Second, we have included the shape of the trained FENet filters in relation to the other feature extraction techniques and added this to the section “FENet provides improved open-loop decoding performance”:

Interpreting machine learning algorithms, especially deep learning, in medical applications is a significant challenge⁴¹. In order to enhance our understanding of the differential characteristics of FENet in comparison to WTs, MUA, and HFLFP feature extraction techniques, we turn to the existing literature on interpretability in deep learning^{42–45}. A key aspect worth exploring is the utilization of filter shapes that exhibit discernibly distinct frequencies (Extended Figure 9). Specifically, we

examined the gain, or the amplification capability, of a sample set of FENet trained convolutional filters across its feature engineering modules. In contrast to the other filters, FENet displayed a unique characteristic of dynamically amplifying specific frequency bands during its training process. FENet's training mechanism takes into account the encoded information within each frequency band, allowing it to selectively enhance relevant features within different frequency ranges. This ability to dynamically amplify distinct frequency bands sets FENet apart from conventional filters such as WTs, MUA, and HFLFP. This deviation from conventional approaches indicates that FENet operates in a manner fundamentally distinct from these conventional feature extraction techniques such as WTs, MUA, and HFLFP. While the exact nature of this divergence requires further investigation, it is evident that FENet encompasses novel elements in its functioning by adaptively adjusting its filters based on the specific frequency information, which exhibits a more nuanced and refined approach of feature extraction and leads to improved performance in analyzing neural data.

We have also added Extended Figure 9 upper panel to present the result of this analysis:



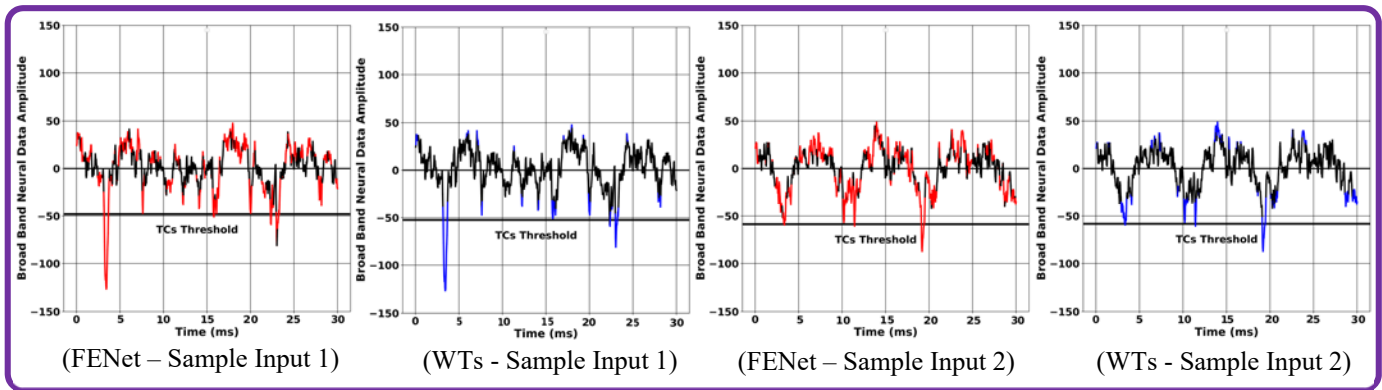
Extended Figure 9. (Upper panel) the gain of sample trained convolutional filters of FENet for all the feature engineering modules compared to the gain of WTs, MUA, and HFLFP filters. In contrast to these filters, FENet exhibit the ability to dynamically amplify distinct frequency bands during its training process, considering the encoded information within each specific frequency band.

- Thirdly, we depicted the sections of the input data that hold greater significance and exert a more substantial impact on the predictions made by the linear decoder. We added this to the “FENet provides improved open-loop decoding performance” section:

To gain more insight into the specific regions of input data that receive more attention from FENet during its prediction process, we present two illustrative examples of single electrode input samples obtained from FENet and WTs (Extended Figure 9 Lower Panel). These samples were collected during a specific session identified as 20190625. To highlight the segments of

higher importance in the predictions made by the linear decoder, we utilize color-coded visual representations. To accurately depict the most relevant sections of the input signals, we calculated the average Shapley value^{42,43} (see Methods) across all samples. Subsequently, we selectively colored the samples whose Shapley values surpassed this calculated average threshold. Additionally, a horizontal line is included in the figures to denote the threshold utilized for extracting features associated with Threshold Crossings (TCs) from each input sample. The presented figures demonstrate that FENet, following its training, leverages not only spike information (TCs) and Wavelet Transforms (WTs), but also exhibits superior capabilities in identifying local patterns within the input data. Furthermore, FENet demonstrates an exceptional proficiency in accurately tracking rapid and abrupt changes present in the input signals when compared to WTs. These empirical findings collectively suggest that FENet possesses the remarkable ability to capture more intricate and localized information, which can result in enhanced feature extraction capabilities when compared to the conventional WTs and TCs approaches.

We have also added Extended Figure 9 Lower Panel:

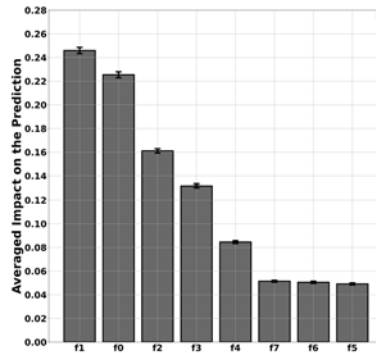


Extended Figure 9. (Lower Panel) We present two single electrode input samples from FENet and WTs obtained during a sample session of participant JJ, labeled as 20190625. In these inputs, the colored sections highlight the segments that hold greater importance in the predictions made by the linear decoder operating on each of FENet and WTs as the feature extraction technique. To represent the more crucial sections of the input signals, we calculated the average Shapley value across all samples of these inputs and colored the samples with Shapley values exceeding this average. Additionally, the horizontal lines indicate the threshold used to extract Threshold Crossings (TCs) features for each input sample. These figures provide evidence that the trained FENet not only utilizes spike information (TCs) and Wavelet Transforms (WTs) to extract features, but also identifies local patterns more effectively. Moreover, FENet demonstrates superior ability to track rapid and abrupt changes in the input signals compared to WTs. These findings indicate that FENet captures more nuanced and localized information, resulting in enhanced feature extraction capabilities when compared to WTs and TCs.

- Fourth, we evaluated the importance of the feature extracted from each FENet feature engineering module. We have added this to the section “FENet provides improved open-loop decoding performance”:

To assess the significance of each extracted feature by FENet for every electrode, we employed the Shapley value⁴² as a measure of importance. The Shapley value allows us to determine the contribution of each input feature in the decoding process when utilizing a linear decoder. The computation of the Shapley value involves comparing the decoder's output with and without the inclusion of a specific feature. The discrepancy between these two cases reflects the contribution of the feature to the decoding process. This calculation is repeated for all possible combinations of features per electrode, and the Shapley value for a given feature is determined by averaging these contributions across all possible combinations, taking into account the number of combinations that include the feature. In this manner, we can evaluate the incremental contribution of each feature to the decoder's output while considering the interactions between features. Features with higher Shapley values are deemed more important since they make a greater contribution to the output variable compared to other features. Extended Figure 7 presents the relative Shapley values of the eight FENet-extracted features. These values represent the average contribution of each feature to the decoding process, calculated and averaged across all electrodes and sessions, using offline data recorded

from human subject JJ during the center-out task. Extended Figure 7 illustrates that the features extracted at the initial stage play a more crucial role in predicting outcomes through the linear decoder.

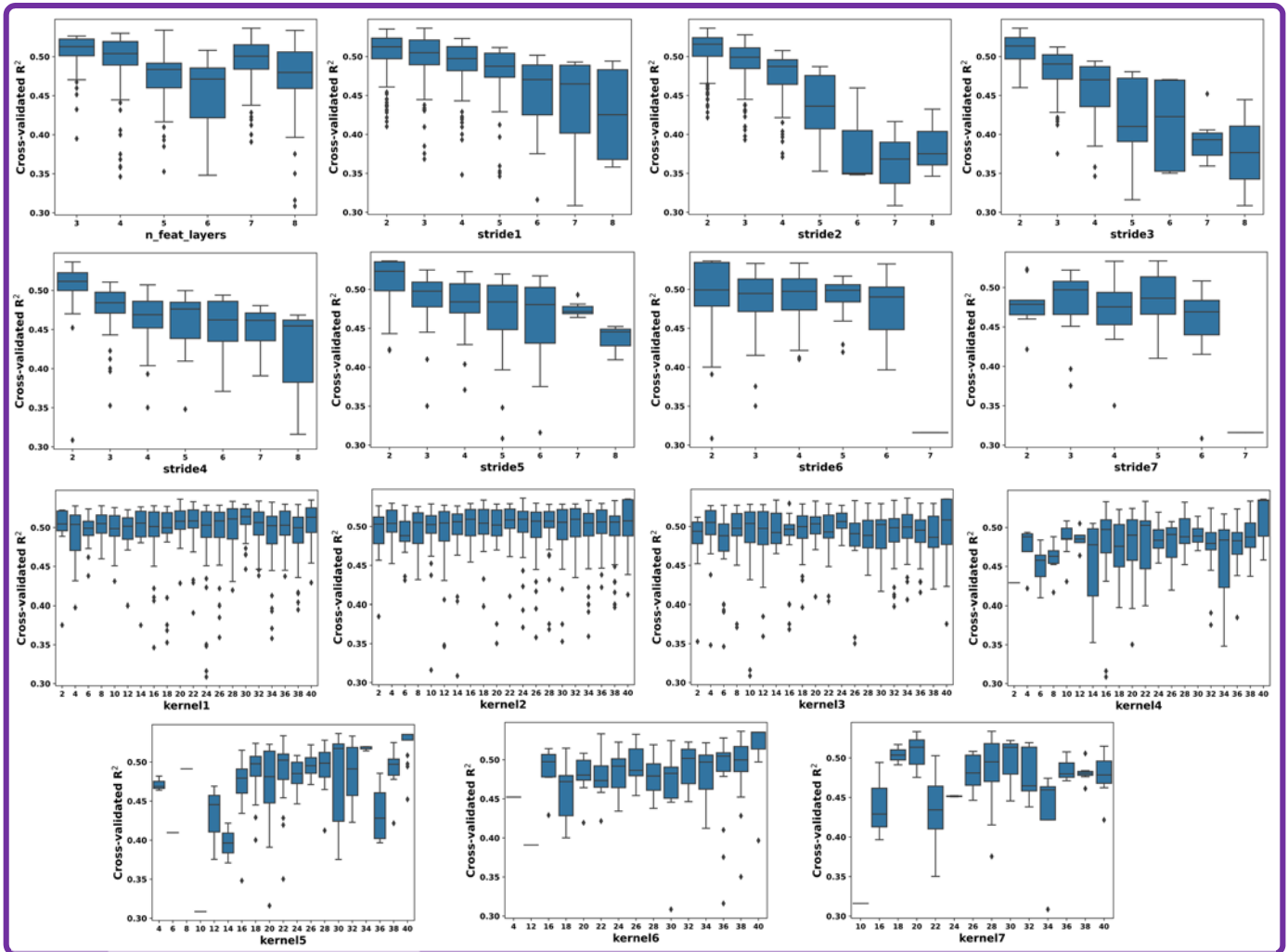


Extended Figure 7. The importance of each extracted FENet feature per electrode. We averaged the proportional Shapley values of all the electrodes over all the sessions for participant JJ.

Comment 10: *What are the critical architecture / hyperparameter choices that affect model performance? As with any neural network study, it is important to conduct ablation experiments with FENet to clarify critical components of the model architecture. This would allow other researchers to build off the work.*

Response: We have added this to the section “Training and Inference for FENet” of the methods (Extended Figure 7):

We conducted parameter sweeps using Bayesian optimization⁷² on the FENet model to assess the importance and impact of each hyperparameter in the model's architecture. The results indicate a correlation between the R^2 values and the parameter values (Extended Figure 8). Our sweeps based on using Bayesian optimization shows that the strides of the initial layers emerge as the most influential parameters, with smaller strides yielding higher performance. One possible interpretation of this phenomenon is that smaller strides enable convolutional kernels to encompass a broader range of patterns within the input. In contrast, larger strides restrict the coverage between successive kernel movements, leading to a reduced capacity for filters to learn diverse patterns⁴⁵. Additionally, we observed that the kernel size becomes more crucial in later layers compared to the initial layers. This suggests that the inputs to later layers summarize information from multiple samples in the preceding layers. Consequently, the network becomes more sensitive to kernel size when combining richer features with different kernel sizes, as these layers combine samples providing less abstract information than deeper layers.



Extended Figure 7. (Upper panel) We conducted parameter sweeps using Bayesian optimization on the FENet model to assess the importance and impact of each hyperparameter in the model's architecture. The results indicate a correlation between the R^2 values of the linear decoder and the parameter values. Notably, the strides of the initial layers emerge as the most influential parameters, with smaller strides yielding higher performance. This is because smaller strides allow the convolutional kernels to cover a greater variety of patterns in the input. Conversely, larger strides limit the coverage between consecutive kernel movements, resulting in the filters learning fewer patterns. Additionally, we observed that the kernel size becomes more crucial in later layers compared to the initial layers. This suggests that the inputs to later layers summarize information from multiple samples in the preceding layers. Consequently, the network becomes more sensitive to kernel size when combining richer features with different kernel sizes, as these layers combine samples providing less abstract information than deeper layers.

Comment 11: Values for hyperparameters are not provided (e.g. dropout rate, learning rate).

Response: The value of learning rate has been explained in the methods section under “Training and Inference for FENet as: “The learning rate α starts at $\alpha=0.1$, which is divided by 2 every ten epochs using a linear scheduler”. For the dropout, we have added this to the same section: “The value for the drop-out has been set to 0.2 for all the layers.”

Comment 12: Open-loop and closed-loop TC decoders did not apply the same preprocessing steps as current state-of-the-art methods presented by others. Many groups have applied “non-causal filtering” (Masse 2014 <https://doi.org/10.1016/j.jneumeth.2014.08.004>) and exponential smoothing (Willett 2019 <https://doi.org/10.1016/j.jneumeth.2014.08.004>) of the binned spiking rates during closed-loop control, but based on the

methods, it appears these preprocessing steps were not included. For the offline analyses, it would be important to compare specifically against the non-causal threshold crossing approach, at a minimum.

Response: During closed-loop control, we have added exponential smoothing similar to Willet 2019 to reduce high frequency variations that occur at a time scale too fast for the patient to control. However, since it is too easy to “juice” and increase R^2 by adding extra smoothing offline, we did not apply any smoothing to the off-line analyses.

We also tested the offline results with smoothing (Extended Figure 6a, b) to show that we maintain the benefits, offline, even with smoothing. We have explained the reason of why we have not added any kind of smoothing for open-loop analysis in lines 172 to 178:

“We evaluated the open-loop results with FENet using neural data and behavior binned at fine temporal resolution (30ms bins) and without smoothing the extracted features. This was motivated by our primary motivation that FENet be maximally useful for closed-loop control where smoothing decreases the responsiveness of the closed-loop system by using potentially outdated neural information. However, recognizing that FENet could also be used for slow-timescale applications, we tested how FENet performed against TCs when smoothing the extracted features. Extended Figure 6a, b show the robustness of FENet against the change in the window size used to update the feature extraction process in our trajectory tasks.

We have added this text to the methods section:

Preprocessing of the generated features for open- and closed-loop analysis. During our offline analysis, we intentionally refrained from applying smoothing to the features under investigation. Smoothing techniques have the potential to enhance the R^2 of decoder output in ways that do not generalize to online control. This is because smoothing works by averaging across time, effectively introducing control delays that render the cursor uncontrollable or undesirably sluggish. Therefore, we opted to evaluate performance in the absence of smoothing. In contrast, during closed-loop control, we employed exponential smoothing⁶⁴ as a preprocessing step for the extracted features. This was done to mitigate abrupt changes and jitters for improved stability⁶⁵. With the patient in the loop, we could ensure that the level of smoothing was not burdensome to the patient. Finally, we also evaluated off-line performance with smoothing (either explicitly, or implicit in the architecture of the decoding algorithm) to ensure that improvements with FENet were not restricted to high-frequency components of the signal that could easily be removed by subsequent signal processing.

Comment 13: *The authors repeatedly claim “state-of-the-art” decoding performance (e.g., L66, L149). I recommend this language be adjusted, because it is hard to claim state-of-the-art without a benchmark for comparison. The performance demonstrated is certainly not on par with published performance, both for open-loop decoding (see Pei*, Ye* 2021 <https://doi.org/10.48550/arXiv.2109.04463>, results here <https://eval.ai/web/challenges/challenge-page/1256/leaderboard>), or closed-loop performance (grid task bitrates, as in the BrainGate work: <https://elifesciences.org/articles/18554>).*

Response: To provide a more nuanced perspective, we have used the term “improved performance” instead. We have also added this text to the discussion section:

It is important to clarify that our reference to improved performance specifically pertains to the feature extraction component, where the patient serves as their own control. Furthermore, it is crucial to acknowledge that the performance of a BMI system can be influenced by various factors, both within and outside of our control. These factors may include the nature of the subject (human or non-human primate), implant site, recording yield, task, severity of spinal cord injury, and the specific decoder employed. In our specific experimental setup, considering our defined problem and requirements, we observed enhanced performance when utilizing features extracted by FENet for decoding in the tasks under investigation.

Comment 14: *The authors mention in their methods that closed-loop cursor position was augmented “potentially with some computer assistance” (lines 301-302). The methods by which computers assisted closed-loop cursor control need to be explicitly defined for readers to properly interpret the results presented in the manuscript.*

Response: We have changed these lines in the “Behavioral Tasks” section of the methods for clarification:

Each of these tasks was conducted in either open-loop, in which the cursor movements were fully generated by the computer and the participant did not directly control the cursor's position, but instead imagines control over a visually observed, computer-controlled cursor, or closed-loop, in which the cursor movements were under the participant's full control with 0% assistance from the computer.

MINOR Comments

Minor Comment 1: 24 - *Shouldn't the same parameters always be used during training and validation? Otherwise, one isn't validating the trained parameters. This may be rephrased to make the intended point more obvious to the reader.*

Response: We agree that this sentence is confusing. We have changed the sentence as below:

FENet jointly optimizes feature extraction and decoding under the constraint that all the electrodes must use the same FENet parameters during BMI decoding.

Minor Comment 2: 78-83 - *"FENet dramatically improved a human clinical BMI participant's ability to use brain signals to control a computer cursor, increasing the bitrate nearly threefold in the closed-loop control. This performance increase was clinically significant: Prior to FENet, the clinical participant requested surgical reimplantation to improve the quality of neural recordings that had degraded substantially since the initial implantation. With FENet, the participant was satisfied with the quality of his neural control. Thus, FENet can extend the functional lifetime of the implanted electrodes, mitigating the need for revision surgeries and thus improving commercial viability." - while interesting and undeniably motivational, this seems more like a discussion topic rather than an introduction topic.*

Response: We agree and have moved this to the discussion section.

Minor Comment 3: 90-91 - *"FENet can improve the efficacy and safety of implantable electrode systems" - there is no safety data included in the paper to justify this claim.*

Response: To provide some context, we understood improved safety as emerging from the fact that the patient regained control of the cursor without the need for a reimplantation surgery. However, we have removed safety to avoid confusion.

Minor Comment 4: 170 - *"Interestingly, although improving the correlation" - correlation to what? This statement is unclear.*

Response: We have changed the sentence as follow:

Interestingly, although FENet improves the R^2 between neural features and kinematics compared to WTs and TCs, Extended Figure 4d-j shows that FENet reports similar tuning preferences to TCs and WTs at the same electrodes.

Minor Comment 5: 246-250 - *"This is why the function of a neuron must be estimated by correlating the number of spikes per unit time against behavior and not the exact features of the waveform shape." - This implies that studies in the past have correlated waveform shapes with behavior, but I don't know of any studies to have done this. If so, I suggest including a citation. Otherwise, if referring to sorting units first then correlating "number of spikes" for each unit with behavior, then consider rewording the sentence.*

Response: This was poor phrasing on our part – the point we were trying to make is that constraining the feature extraction algorithm to be the same across electrodes is reasonable. Waveform shape is something that might be idiosyncratic from one electrode to the next, but waveform shape does not carry task-relevant information (as you point out). We have removed this sentence to avoid confusion and updated the corresponding paragraph with more explanations to convey our assumption better as below:

Furthermore, the waveform of an action potential is predominantly a function of the relative location of the electrode tip and the neuron, not the neuron's role in behavior⁵³. We assume that the precise spatiotemporal features of broadband data report the relative activity of the local neural ensemble and not details of the behavior per session. Therefore, in a similar vein to conventional feature extraction methods that apply uniform operations to individual electrodes, FENet employs a consistent feature extraction process across all electrodes. Furthermore, it is worth noting that future BMI systems may not universally support the capture of raw, high-sample rate broadband data (e.g., Neuralink) to enable training of new per-channel models. In such cases, our approach can be seamlessly utilized without the reliance on this type of data. FENet was designed to be a feature extraction method that could easily be dropped into current decoding pipelines with existing decoding approaches and thus must generalize across electrodes, patients, brain areas, and time, despite potential recording non-stationarities. It is possible that neural networks that are electrode specific could improve performance, but this would occur at the cost of generalizability and ease of use. Further, given recording non-stationarities and limited access to training data, the practical ability to train electrode-specific realizations of FENet is non-trivial. Consequently, this limitation constrains the potential benefits of hyper-specific solutions tailored to individual electrodes. Therefore, FENet is designed to be agnostic to the specific number and

configuration of electrodes within different Brain-Machine Interface (BMI) systems, making it readily adoptable by users, particularly those who prefer to avoid setting up their own training protocols.

Minor Comment 6: 260-261 - *“To the degree that trial-to-trial differences are driven by behavioral factors, no algorithm measuring the local activity of neurons can reduce this variability.” - Don’t latent variable models of population activity accomplish this, such as LFADS?*

Response: We make this statement exclusively from the perspective of a feature extraction technique using recorded activity from a single electrode. The trial-to-trial variability is an accurate report of how behavioral factors have changed the activity measured at the single electrode level. You are correct that the ability to record populations of neurons can be used to remove/reduce this variability through an estimate of the task-relevant latent variable, but this is a separate point. LFADS should not be considered as a feature extraction technique, as it relies on threshold crossings or other pre-extracted features as its input. Furthermore, LFADS does not perform causal and real-time processing of these features, as it utilizes data recorded for an entire trial. Hence, comparing LFADS, which is not inherently a feature extraction technique but utilizes extracted features as input, with FENet, a dedicated feature extraction technique, leads to an unfair comparison.

To clarify the point, we revised the sentence as follows:

To the degree that trial-to-trial differences are driven by behavioral factors, no algorithm measuring the activity from a single electrode can reduce this variability, although populations of such electrode recordings can reduce variability through estimates of latent variables²².

Minor Comment 7: 309-310 - *“we applied an 8-order elliptical high pass filter with the cut-off frequency of 80 Hz” - Was a lower cutoff set at 80Hz to avoid residual 60Hz noise? It would be interesting to read the authors’ comments on the decision to avoid beta frequency bands of LFP, which have been useful features for motor control for decades. This would be particularly insightful because of the claim that FENet is given the freedom to extract the features it finds most helpful from the entire available broadband recording.*

Response: We have added this to the “Preprocessing the broadband neural data” section of the methods and “Figure 4a, b” caption:

Given that we estimate FENet features in a window duration of 30 ms, the theoretical lower frequency limit is approximately 33.33 Hz. However, to accurately estimate frequency content, we need at least two cycles of the lowest frequency within our window. Setting a high-pass cutoff at 80 Hz ensures more than two cycles within the 30 ms window, providing a more reliable frequency estimation.

In lines 167,168, we explained that: “Extended Figure 3a, b shows that FENet does not rely on the low-frequency (< 250 Hz) local-field potentials to achieve its enhanced decode performance.”

In the Extended Figure 3 caption, we have explained that: “We have used an 80 Hz filter since window size of 30ms, which has been used for filtering, is small enough to assume that the lower frequency activities are excluded from the broadband neural activity in the 30ms window.”

A drawback to increase the windows size to include the lower frequency features is that it effectively smooths and introduces delays into the decoding process. In this work, we were specifically focused on the fine temporal resolutions of 30ms and 50ms.

Minor Comment 9: 320 - *“Suppose that ... represents a one-dimensional sample of the broadband neural data consists of S samples” - consider rewording. This sounds like the algorithm simultaneously needs one sample (i.e. “one-dimensional sample”) or S samples. Should this be one sample per electrode, assuming S electrodes? Or does this mean S samples for one electrode, in which case it isn’t a “one-dimensional sample”? Given the following two sentences, I think this means that x is a real number of size [SxN]. The explanation should be redefined from there.*

Response: We have reworded this sentence as below:

Suppose that $x \in \mathbb{R}^S$ represents a one-dimensional time series consisting of S samples of the broadband neural data recorded from one electrode.

Minor Comment 10: 331 - *I see S has been clarified here. Consider copying this to 320 or consolidating into one description of the input data size.*

Response: It has been clarified in the previous comment.

Minor Comment 11: 326 - *Equation 1 - is a parenthesis misplaced here? Shouldn't the loss be between g using the FENet features and the true labels y? As it is printed, the decoder uses the true labels to predict the true labels...*

Response: It has been corrected by moving the parentheses as below:

$$\psi^*, \theta^* = \operatorname{argmin}_{\psi, \theta} E_{(x, y) \in D} \mathcal{L}(g_{\theta}(F_{\psi}(x)), y)$$

Minor Comment 12: 331-332 - *“in our case, 30ms of the recorded broadband neural data, which includes 900 samples” - is this just for participant JJ, and was 50ms (1500 samples) for participant EGS?*

Response: That is correct. 30ms is for both JJ and NS. For EGS, the time interval is 50ms, which is 1500 samples with the sampling rate of 30 KHz. We have changed this sentence to:

(e.g., 30ms of the recorded broadband neural data, which includes 900 samples for JJ and NS, and 50ms of the recorded broadband neural data, which includes 1500 samples for EGS).

Minor Comment 13: 333 - *“the input data, s_{i-1}, is padded with zeros” - what does i refer to here? To clarify notation, consider saying “the input to the ith feature engineering module, s_{i-1}, is padded...”*

Response: The sentence has been changed accordingly.

Minor Comment 14: 347 - *“constraining the feature extraction algorithm to use the same parameters for all the electrodes used in the training set.” - For N electrodes, I understand that there is one FENet per electrode, and each of those FENets uses the same parameters to extract features. The methods provided describe how FENet is trained on one electrode, but it is not detailed how the N sets of FENet parameters are combined into one set of parameters to be used for all electrodes.*

Response: We have added this to the “FENet pipeline” section of the methods:

According to our assumption that the generative process that produces the broadband neural activity is statistically similar across electrodes, we design FENet such that it learns a single set of parameters ψ for all the electrodes. Thus, the same instantiation of FENet, as defined by the parameter set ψ , is applied independently to broadband data recorded from all electrodes.

Minor Comment 15: 352-361 - *Throughout this explanation, a 30ms moving window is used here, but it seems that this only applies to participant JJ. What about participant EGS?*

Response: Correct. For EGS, we have used window size of 50ms. We have added this at the beginning of the section and did change the section as follow:

Generation of other features. We have extracted the features from the 30ms bins for JJ and NS, and from each 50ms bin for EGS. To extract wavelet features (WTs)^{19,20}, we used a db20 mother wavelet with 7 scales on moving windows (no overlap) of the time series recorded from each electrode. A db20 mother wavelet was selected as it contains filters with length 40 and can model WTs high pass and low pass filters more accurately compared to other Daubechies wavelet families²⁰. Moreover, in our preliminary experiments, we found that it outperformed other wavelet variants for the datasets tested in the current study¹⁹ (e.g., db4, Haar). The mean of absolute-valued coefficients for each scale was calculated to generate $M = 8$ time series per electrode, including seven detailed coefficients and one approximation coefficient generated by the WT high-pass filters and the final stage WT low-pass filter, respectively. To generate threshold crossing features (TCs), we thresholded the neural data at -3.5 times the root-mean-square (RMS) of the noise of the broadband signal, independently computed for each electrode, after band-pass filtering the broadband signal between 250Hz and 5KHz. We did not sort the action potential waveforms²¹. TCs events were counted using the same intervals as WTs and FENet. To derive the MUA features, the raw broadband neural data underwent a bandpass filtering process with a range of 300 to 6000 Hz. Following this, customized root mean square (RMS) values were calculated to generate the MUA signal for each bin^{20,35}. To generate the High Frequency Local Field Potentials (HFLFP) features, the raw broadband neural data from each electrode underwent a second-order band-pass filtering process using a Butterworth filter with low and high cutoff frequencies set at 150 Hz and 450 Hz. The power of the filter's

output was then calculated and used as the HFLFP feature for each electrode^{4,33}. For FENet – HFLFP and TCs-HFLFP, we simply concatenate the corresponding features together to generate a larger feature matrix that include both types of extracted features.

Minor Comment 16: 375-378 - The outcomes of these queries were not included in any data presented in the manuscript, please include the results.

Response: Correct. In each session, we consistently sought feedback from the participant after each task. However, we have decided to remove this sentence from the manuscript since we do not include the participant feedback in the manuscript.

Minor Comment 17: 383 - “C is the number of input neural electrodes” - wasn’t this N before? Best to be consistent with variable names.

Response: C has been changed to N in Extended figure 1b a and the corresponding paragraph mentioned in the comment.

Minor Comment 18: 407 - “Using the domain knowledge” - what does this mean?

Response: This was unnecessarily confusing. We have changed this sentence as below:

For the symmetric replication of the feature engineering modules of the FENet, we designed FENet to have a hierarchical and symmetric architecture similar to the db20 wavelet transform.

Minor Comment 19: 481 - “for each of the standard feature types (TCs and FENet)” - what about WTs?

Response: Following the reviewer comment, we conducted sessions to compare WTs with FENet in closed-loop sessions. Our focus was to compare WTs, as the second best feature extraction technique, and TCs, as the gold standard used in other labs, with FENet features in closed-loop analysis. We have updated Figure 2 as below and updated the “FENet improves closed-loop control” section accordingly:

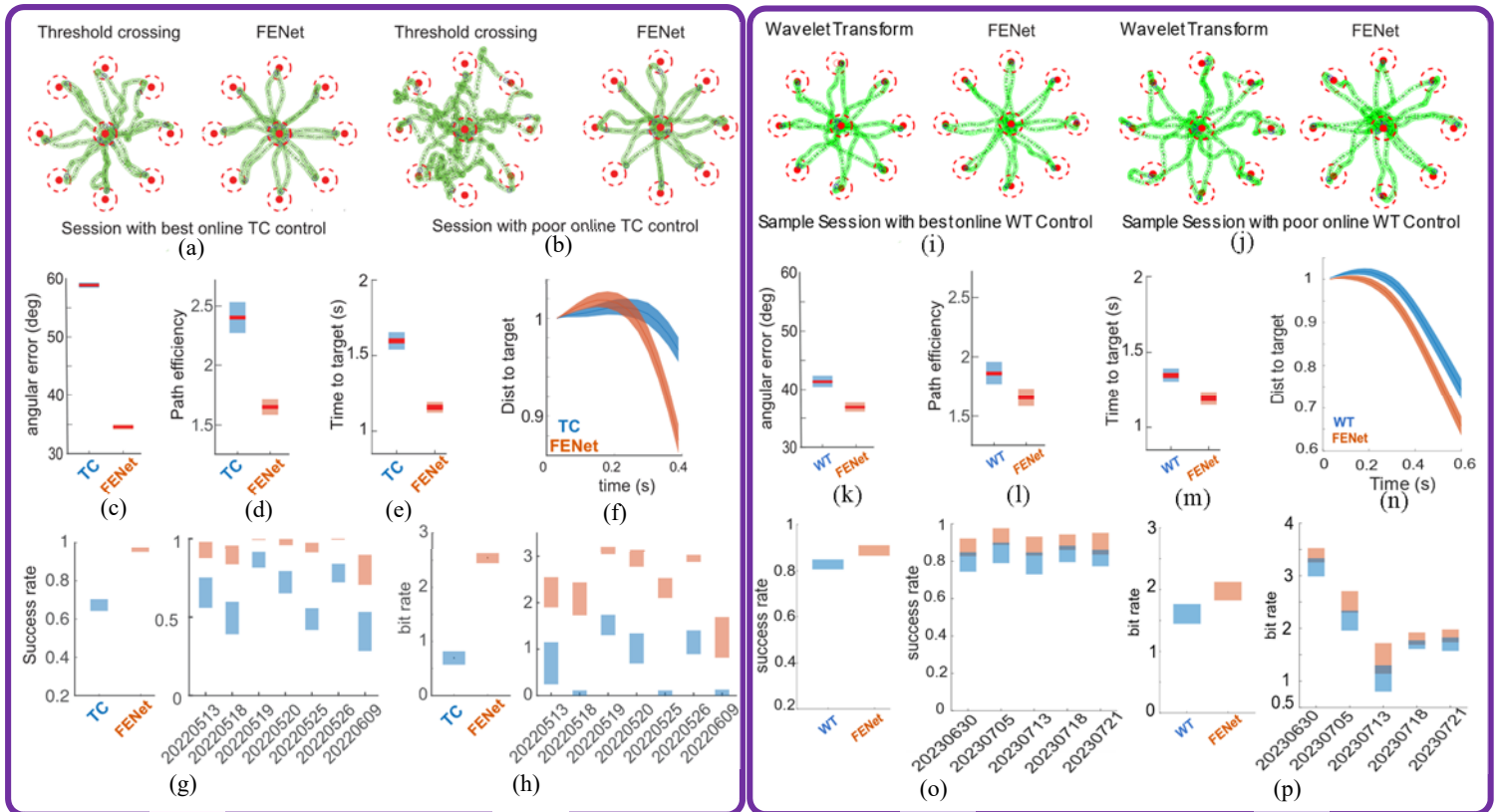


Figure 2: Closed-loop performance evaluation for JJ. Online trajectories comprising one movement out and back to each of eight targets in a center-out paradigm. This figure illustrates the effectiveness of a linear decoder when performs on FENet features, compared to the threshold crossings (TC) features (left panel) and the Wavelet transform (WT) features (right panel).

At the top, left and right figures show trajectories using the other feature extraction technique and FENet-based features respectively. Trajectories were sampled from the same experimental run, as part of an interleaved-block design. There is high daily variability in control quality using TCs. Additionally, due to the significant instability encountered during the operation of WTs, we had to manually introduce bias in both the X and Y directions to prevent the cursor from going beyond the screen after prolonged use. (a), (i) results from the best recent day using TCs and WTs. (b), (j) results for another experimental session with poor TCs and WTs performances. TCs control quality has degraded substantially while FENet has largely preserved performance. Furthermore, the control achieved through the use of WTs has exhibited inconsistent stability between sessions following five years of implantation. (c), (k) the averaged angular error, (d), (l) path efficiency, and (e), (m) time to target over the closed-loop sessions as the closed-loop control metrics. Instantaneous angular error captures the angle between the vector pointing towards the target and the instantaneous velocity of the cursor. Path efficiency is measured as the total distance traveled end route to the target normalized by the straight-line distance from the starting location to the target. Distance to target (mean $\pm$ 95%CI) was used to quantify cursor responsiveness to the participant's intent. Here, latency from target onset to goal-directed movements is shorter for FENet-based features as compared to TCs and WTs. (f), (n) the averaged distance to target for the center-out task. To account for variations in trial lengths, the figure depicts the average distance to the target across multiple trials has been generated using the duration of the smallest trial as a reference. Consequently, the figure does not extend in time until the target has been reached. (g), (o) the success-rate, and (h), (p) the bit rate (see methods) within an 8-by-8 grid task (for FENet versus TCs, $t = -11.850$, $p = 0.0000$, and for FENet vs. WTs, $t = -4.252$, $p = 0.0000$). Success was measured as the ability to move the cursor to and hold a target (0.5 second hold time) within 4 seconds.

We have compared the performance of other decoders operating on multiple feature extraction techniques in our offline analysis. We have also mentioned this at the beginning of the section "FENet provides improved open-loop decoding performance":

"Direct comparison in closed-loop testing is ideal but opportunities for such testing are relatively limited. To increase the scope of comparison across time and feature extraction techniques, we evaluated the ability of FENet to reconstruct the movement kinematics using previously collected neural data recorded from implanted electrode arrays."

We have also added this to the discussion:

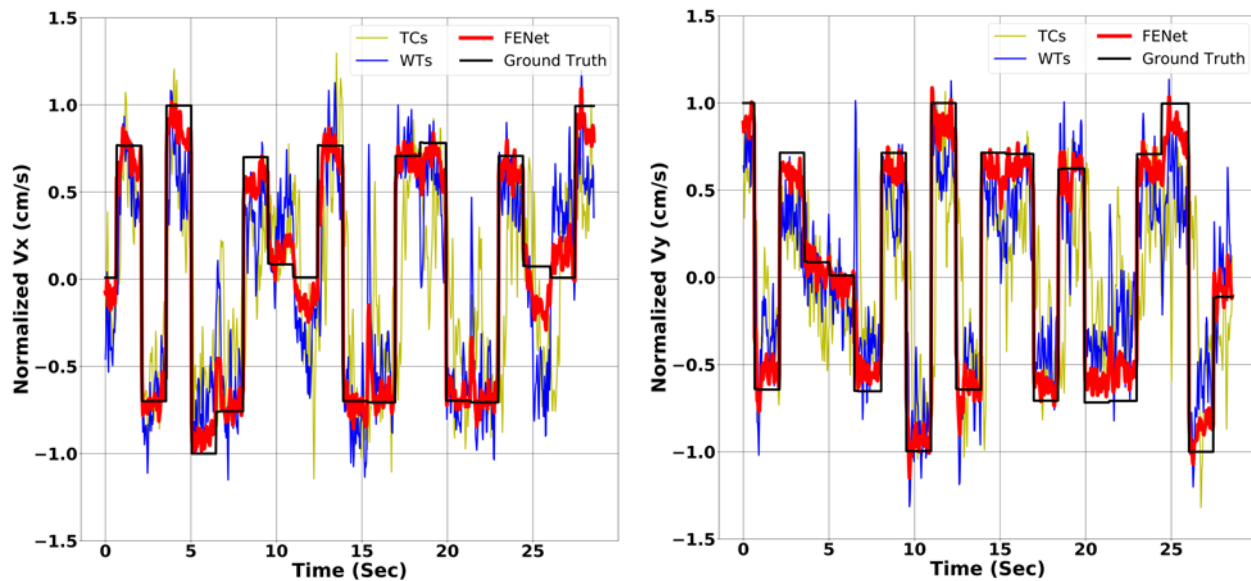
In this study, we conducted an offline analysis to compare FENet's performance with multiple other feature extraction techniques using different linear and non-linear decoders. To expand the scope of comparison across different time periods and feature extraction techniques, we assessed FENet's capability to reconstruct movement kinematics using previously recorded neural data from implanted electrode arrays. A retrospective analysis over years of recordings showed that FENet significantly improved the cross-validated R^2 and the SNR of extracted neural features compared to the other feature extraction techniques across multiple patients and through the lifetime of the arrays. Further, FENet generalized well across cortical brain regions, patients, and tasks, demonstrating its ability to serve as a drop-in replacement for other feature extraction techniques. Moreover, the population-level analysis demonstrated that FENet preserves the representational structure and temporal dynamics of sorted neural populations and, thus, provides an accurate measure of brain activity. Consequently, according to the results of our offline analysis, we evaluated and compared the performance of threshold crossings (TCs), as the current standard for closed-loop control^{4,6,22,26,27,31,33}, and Wavelet transforms (WTs), which have also demonstrated performance improvements in our offline analysis and in the recent studies on BMIs^{20,29}, against FENet features in our closed-loop analysis. Neural decoders employing FENet-based features outperformed TC-based and WT-based features across all metrics. Additionally, FENet enhanced the cursor's responsiveness to the participant's intent, reducing the time it took for the cursor to move towards the target after its onset. These improvements in both aspects led to significant enhancements in overall task performance during the closed-loop sessions. While we reported both the open- and closed-loop performances for participant JJ, our evaluation of the presented feature extraction techniques was limited to the recorded open-loop neural data from participants EGS and NS in specific tasks, as their participation in the clinical trial concluded, and their electrodes were explanted. Nevertheless, the principles derived from analyzing the recorded open-loop neural data from EGS and NS will hold relevance for future subjects with electrodes in the same cortical areas. Taken together, FENet can improve the efficacy of implantable electrode systems while delivering improved performance and ease of use.

Minor Comment 20: *Fig 2d - I recommend extending this in time until the target has been reached to compare against other manuscripts with similar plots*

Response: Our objective with this figure is to illustrate the response time, from the time of target onset to the time the cursor starts moving in the direction of the target. Thus we used a smaller window to better emphasize this point.

Minor Comment 21: Fig 3e,f - it would be helpful to know which session from Fig 3a this is. Also, it would be helpful to plot a common time axis here (in seconds) rather than samples.

Response: We have plotted a common time axis here. Also, Fig 3e, f are showing the fourth session, 20190507, which is the fourth session in Figure 3a. We have added this information in the figure caption as well.



Minor Comment 22: Extended Fig 7d - The colors in this plot are very difficult to visually differentiate, so conclusions cannot be made about the effect of different PLS dimensionalities.

Response: The observed trend in the figure is characterized by a progressive darkening of colors within each family of curves. This visual representation effectively highlights the impact of increasing PLS (Partial Least Squares) dimensions. Specifically, the plot clearly demonstrates that higher PLS dimensions correspond to a greater contribution from the first PLS component in predicting the output.

Minor Comment 23: Extended Fig 7d,e - It is unclear what is actually being plotted here. How was the value for each channel determined? (What methods were used?) Are the channels sorted in some way? What does the $x=0$ point represent?

Response: We have changed the caption to:

Results are shown separately for each PLS-based latent dimension after sorting the electrodes by maximum per-session R^2 , and then averaging across the sessions. Electrodes were sorted based on the R^2 value between the ground-truth and the linearly regressed movement kinematics using each single electrode.

Stylistic Issues/Copyediting Comments

Stylistic Issues Comment 1: 117 - “Extended Figures 7” should be “Extended Figure 7”

Response: We have corrected this.

Stylistic Issues Comment 2: 191-192 - “These results suggest that FENet can generalize across different time periods, brain areas, and patients, and electrodes” - remove the “and” in “and patients”

Response: We have corrected this.

Stylistic Issues Comment 3: 303-304 - “screen in each trial. Then, the participant immediately attempted to press the corresponding finger of the right hand (Figure 5).” - Unexpected larger font size

Response: We have corrected this.

Stylistic Issues Comment 4: 305 - I think “which have not” should be “which has not”

Response: We have corrected this.

Stylistic Issues Comment 5: 320 - “one-dimensional sample of the broadband neural data consists of S samples” - There is a conjunction missing here, perhaps “and”?

Response: We have corrected this by adding “which” as a conjunction.

Stylistic Issues Comment 6: 323 - missing space between ψ and “corresponds”

Response: We have corrected this.

Stylistic Issues Comment 7: 332 - missing space between x and “is”

Response: We have corrected this.

Stylistic Issues Comment 8: 375-376 - “and blinded queries to research participants to evaluate responsiveness” - consider putting a colon here

Response: We have corrected this.

Stylistic Issues Comment 9: 402-403 - “Linear decoder predicts the velocity of the cursor movement as smoother prediction than the cursor positions that can provide a more stable prediction for a BMI system.” - Confusing sentence, possible grammatical errors

Response: We have changed the sentence to make it clearer:

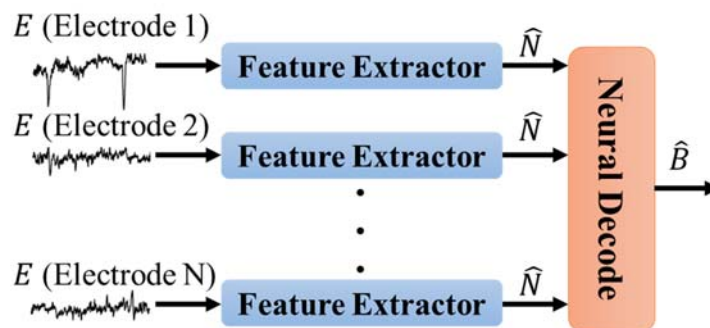
Since predicting the velocity of the cursor movements in a BMI system is more stable and smoother than predicting the cursor position⁶⁴, we first predict the cursor velocity by using the decoder.

Stylistic Issues Comment 10: 479-480 - “human research participants our during double-blind testing” - order of “our” and “during” should be swapped

Response: We have corrected this.

Stylistic Issues Comment 11: Fig 1b - Electrode 1 is listed twice

Response: It has been corrected as below:



Stylistic Issues Comment 12: Fig 4 caption - “These results suggest that FENet can generalize across different time periods, brain areas, and patients, and electrodes” - remove the “and” in “and patients”

Response: We have corrected this.

Missing Citations Comment

Missing Citations Comment 1: *Willsey et al. recently used a convolutional neural network to improve decoding performance in monkeys, and should be discussed in the Introduction or Discussion sections.* <https://doi.org/10.1038/s41467-022-34452-w>

Response: The primary objective of our research is to develop a feature extraction technique. While the mentioned paper primarily emphasizes decoder design, which falls outside the scope of our manuscript, it is still relevant to our introduction section where we discuss existing techniques involving threshold crossings (TCs) and wavelet transforms (WTs). It should be noted that the paper in question utilizes spike band power as the feature and employs convolutional and fully connected layers for decoding, rather than extracting features directly from raw broadband recordings. Therefore, we have included a citation to this paper in introduction to acknowledge its contribution and contextualize our work (reference 30):

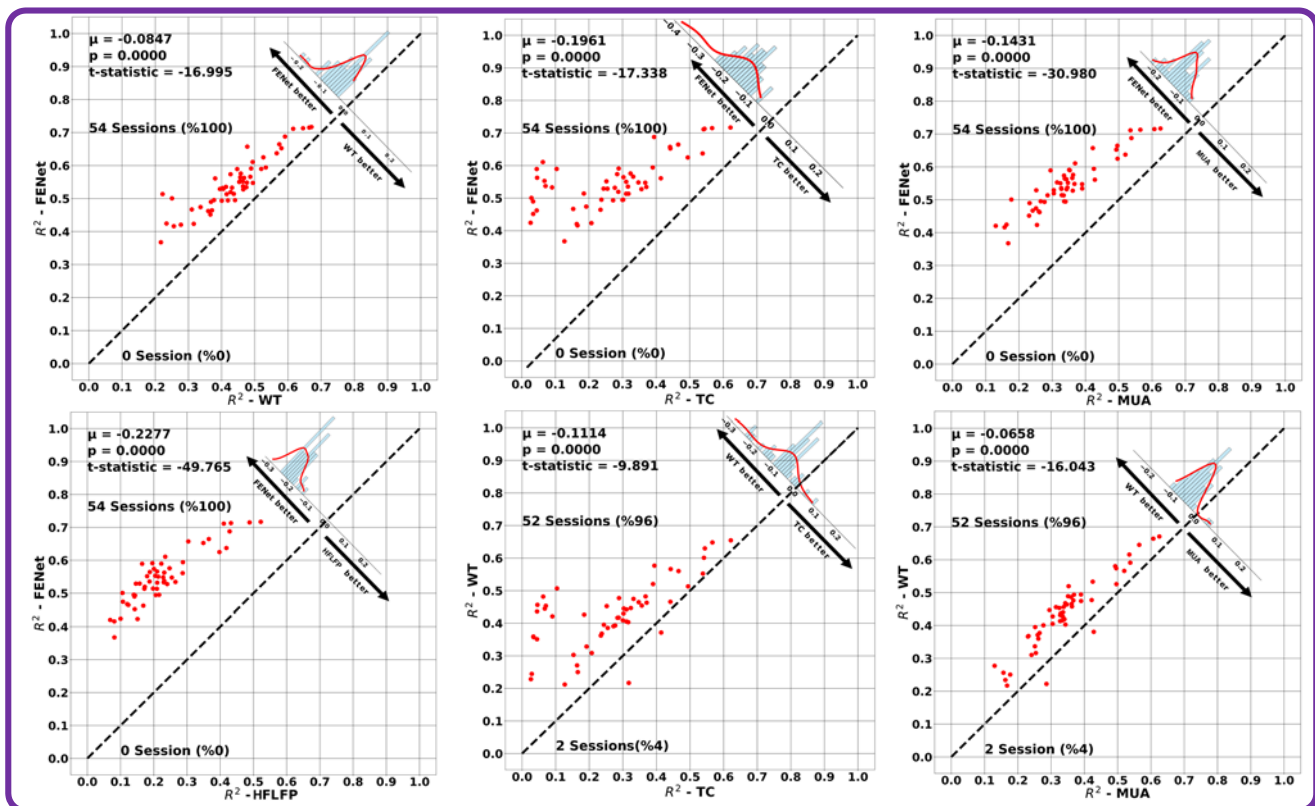
Current studies usually compare the decoders across a limited set of feature extraction techniques, such as neural threshold crossings (TCs) or wavelets (WTs)^{4,17,19–30}.

Reviewer 4

Comment 1: This paper presents a neural network based method for creating better neural features for an intracortical brain machine interface using an older Utah array with low amplitude signals. This would be used with existing downstream algorithms. While they report a dramatic performance increase, this is impossible to evaluate because they are leaving out a very common technique that is in use across at least 3 other groups for the same purpose, with known large effects. Specifically, doubling of performance with units present was reported in 2007 by Stark and Abeles using a simple band pass filter from 300-6000 Hz, and groups since then have discovered that this has a disproportionately positive effect on older Utah arrays. Instead, this group compared their approach to threshold crossing, which was among the control features in Stark and Abeles and also with wavelets, which have not been compared with the more commonly used band pass filter to my knowledge. It seems likely that a convolutional neural network can discover how to implement a band pass filter, and this must be addressed directly. This may explain most of the effect as a known simple technique.

For the effect that remains, it is also important to explore the mechanism. This group discusses previous datasets with spikes present, and these could be used to determine what exactly the network is finding. For example, is it identifying spectral features that correspond to neural units? The time window of 30 ms is quite short so this seems likely. But the brief claim about finding multiple sources per electrode could be and should be evaluated directly with spiking datasets and or simulation.

Response: In Zhang et al.'s study (2018) [reference 20], Bandpass filtering (MUA) was found to be less effective than Wavelets, leading to its exclusion from our initial analysis. However, in response to Comment 1 raised by Reviewer 3, we have updated our manuscript to incorporate the bandpass results, allowing for a direct comparison in the summary figure (Figure 3a,b). Additionally, Extended Figure 9 showcases the filter shapes learned by FENet and compares them with Wavelet filters and the bandpass filter (Multi Unit Activities or MUA). Notably, FENet's filters exhibit distinct shapes and amplify different frequency ranges compared to bandpass filtering, wavelet transform, and threshold crossings. Furthermore, in Extended Figure 8, we explore the average feature values across different trials, highlighting the stronger response elicited by FENet compared to other feature extraction techniques. The performance of different decoders operating on various feature extraction techniques is extensively analyzed in Figures 3, 4, and 5, consistently demonstrating the superior performance of FENet. Figure 5, displayed below, specifically showcases the performance of the linear decoder using different feature extraction techniques, further emphasizing the superiority of FENet. For a more detailed understanding of the interpretation of features extracted by FENet and the performance of different decoders across the employed feature extraction techniques, we encourage readers to refer to Comments 1 and 9 provided by Reviewer 3.



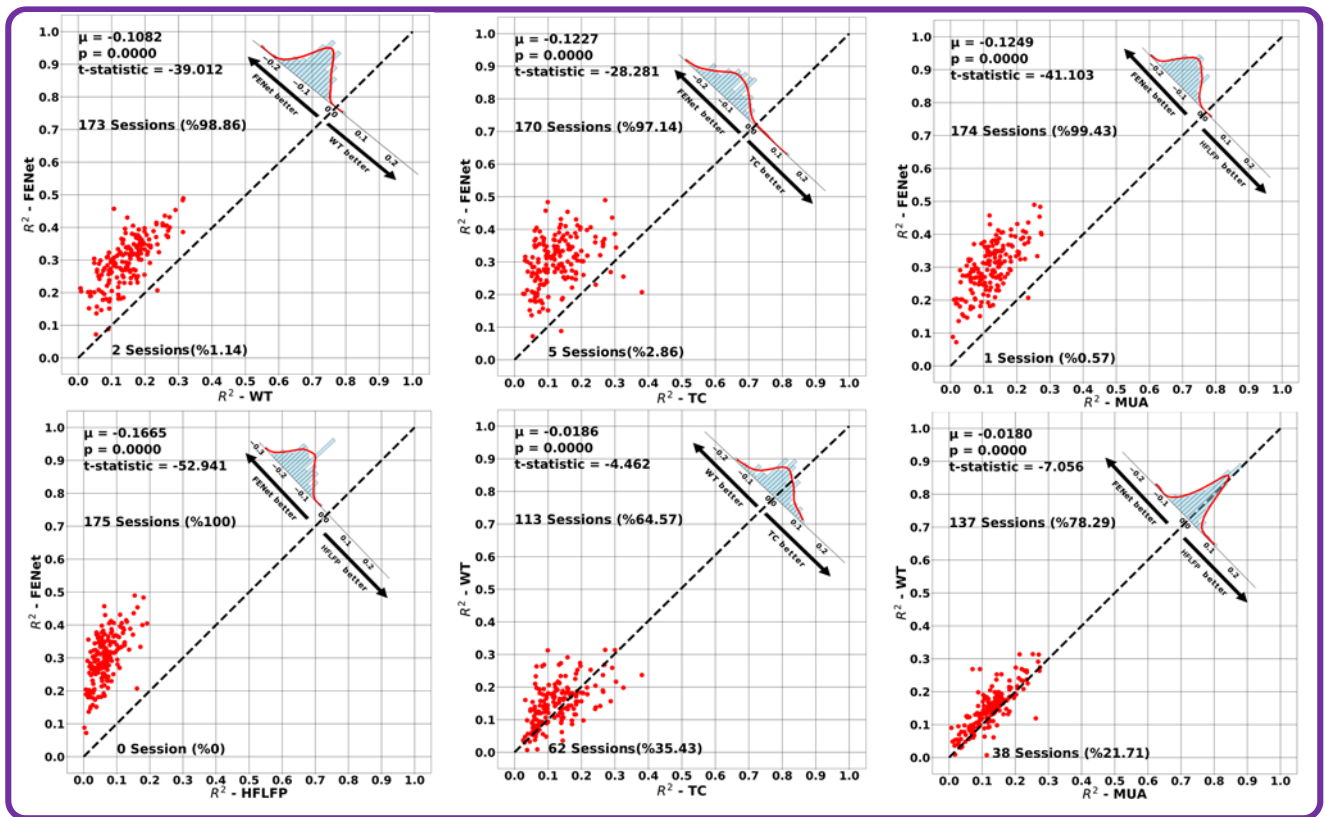


Figure 5. Comparison of the cross-validated R^2 of linear decoder operating on one feature extraction technique vs. the other feature extraction technique for (top) participant JJ, and (down) participant EGS. Red dots show the sessions. The dashed line shows $y = x$. The percentage of dots on each side of $y = x$ shows the number of sessions in favor of the corresponding feature extraction technique. The t-test statistics have been calculated to show the confidence level of the reported statistics. According to these figures, linear decoder operating on the FENet-based features provides superior performance in term of R^2 compared to other feature extraction techniques for both human participants.

Comment 2: Please state within the video captions whether the clock is running in real time. A few movements of the participant look slightly odd, making this reviewer concerned that there could be an unstated clock adjustment. Also, please do not include a video of calibration, which could be mistaken by a less careful reviewer for a BMI experiment.

Response: The clocks are running in real time in all recordings. We have mentioned this in the video captions. We have excluded the calibration video to avoid confusion.

Reviewer 5

Major Comments

Comment 1: *The paper has a weak baseline comparison to just decoding TCs using a linear model and lacks comparisons to other methods*

Response: In order to enhance the comparative analysis and address the suggestions made by the other reviewers, we expanded our investigation beyond threshold crossings (TCs) and wavelet transforms (WTs) by including Multi Unit Activity (MUA), High-Frequency Local Field Potentials (HF-LFP), and combinations of FENet with TCs and HF-LFP (FENet - HF-LFP and TCs - HF-LFP) as additional feature extraction techniques commonly used in the field. Furthermore, to explore alternative decoder models beyond linear decoders, we applied Long-Short Term Memory RNN (LSTM), Support Vector Regressor (SVR), ReFIT-Kalman Filter, and PSID to all the extracted features. The manuscript has been updated to incorporate the findings from these analyses. The accompanying figures, based on the results of these latest investigations, demonstrate the superior performance of all decoders when operating on FENet features compared to other feature extraction techniques. For a more detailed response, we direct the reviewer to the answers provided in response to "Comment 1" of the major comments from "Reviewer 3" and "Comment 1" from "Reviewer 4."

Comment 2: *Basically all the papers cited to justify that TCs are the standard for closed-loop control underlying best-in-world closed loop performance use either a kalman filter (6, 32) or RNN based model (22, 26, 27, 31) to decode kinematics from neural activity. If these are cited as the benchmark, then they should be what's being used when comparing the FENet extraction algorithm to TCs in my opinion. E.g, even if we use a SOTA RNN, or MRNN, are the results with the FENet better than just using TCs? This would make the paper a significant advance.*

Response: We agree that comparing the performance of FENet to these established benchmarks would indeed contribute to the significance and advancement of the paper. We have incorporated this suggestion into our evaluation and analysis to provide a comprehensive comparison between FENet and the other feature extraction techniques, including TCs. We refer the reviewer to the answer to "Comment 1" of major comments of "Reviewer 3", where we applied a variety of decoders on different feature extraction techniques.

Comment 3: *Furthermore, references 32 and 4 both use HF-LFP features and TCs for decoding, which I suggest adding as a comparison to the FENet extraction algorithm.*

Response: We have added HF-LFP as well as the concatenation of FENet and TCs with HF-LFP, named as FENet – HFLFP and TCs – HFLFP, respectively. We have reported the performance of different decoders operating on these newly added features. We refer the reviewer to the answer to "Comment 1" of major comments of "Reviewer 3" and "Comment 1" of "Reviewer 4".

Comment 4: *Another potential comparison would be to PSID (Sani et. al. Nature Neuro 2020). Note this feature extraction paper also used a Kalman filter to compare its method to other extraction methods and TCs.*

Response: We have added PSID as a neural decoding algorithm applied to the different feature processing techniques. We refer the reviewer to "Comment 1" of "Reviewer 3" and "Comment 1" of "Reviewer 4" for the results of PSID.

Comment 5: *Is the FENet specifically what's important to good performance? Would a neural net that is not shared across elec achieve similar or better performance? If the whole point is just to extract features from neural activity, why limit it in this way? Comparison would be good.*

Response: FENet was designed to be a feature extraction method that could easily be dropped into current decoding pipelines with investigators existing decoding approaches and thus must generalize across electrodes, patients, brain areas, and time, despite potential recording non-stationarities. It is possible that neural networks that are electrode specific could improve performance, but this would occur at the cost of generalizability and ease of use. Further, given recording non-stationarities and limited access to training data, the practical ability to train electrode-specific realizations of FENet is non-trivial. To explain this, we have added this to the discussion section:

Therefore, in a similar vein to conventional feature extraction methods that apply uniform operations to individual electrodes, FENet employs a consistent feature extraction process across all electrodes. Furthermore, it is worth noting that future BMI systems may not universally support the capture of raw, high-sample rate broadband data (e.g., Neuralink) to enable

training of new per-channel models. In such cases, our approach can be seamlessly utilized without the reliance on this type of data. FENet was designed to be a feature extraction method that could easily be dropped into current decoding pipelines with existing decoding approaches and thus must generalize across electrodes, patients, brain areas, and time, despite potential recording non-stationarities. It is possible that neural networks that are electrode specific could improve performance, but this would occur at the cost of generalizability and ease of use. Further, given recording non-stationarities and limited access to training data, the practical ability to train electrode-specific realizations of FENet is non-trivial. Consequently, this limitation constrains the potential benefits of hyper-specific solutions tailored to individual electrodes. Therefore, FENet is designed to be agnostic to the specific number and configuration of electrodes within different Brain-Machine Interface (BMI) systems, making it readily adoptable by users, particularly those who prefer to avoid setting up their own training protocols.

Comment 6: *Figure 4: Why should I care about this if I am a BCI user? How much time will it save me to do the fenet transfer, vs just using a new model? Train over 10 sessions to get good performance isnt that many (extended data, fig 7)*

Response: The primary point of our paper is to demonstrate that FENet provides significant performance improvements when compared against current state-of-the-art feature extraction algorithms. This is the primary benefit to the user. The primary point of this figure is to demonstrate that the substantial improvement provided by FENet can be realized immediately across tasks, individuals, and recording quality of the array. In other words, the benefits are likely to generalize beyond the specific set of patients or tasks studied in this paper. Thus, the user will care because they are likely to realize the performance improvements of FENet. The fact that they can realize these benefits immediately, without the need for days of training, can be considered an additional benefit, that, in our opinion, is non-trivial.

Minor Comments

Minor Comment 1: *You apply partial least squares regression to the output of the FENet features of each electrode. I am curious what performance is like if you dont do this? Could be nice for Extended Fig 1e. Alternatively, does apply PLS like this to the TCs to reduce dimensionality improve performance? Even though there are not multiple features per electrode, I am concerned this is a critically useful step that makes the FENet/TC comparisons unfair.*

Response: We only apply PLS on a per-channel basis – to reduce the increase in dimensionality that comes with FENet or Wavelets so as not to lead to overfitting when training the decoding model that maps features to kinematics. Given that TCs only have one feature dimension per electrode, we cannot further reduce the dimensionality. We have attempted to improve the clarity of our explanation in the methods section “Preprocessing of the generated features for open- and closed-loop analysis”:

Given the flexibility of FENet’s and WTs’ design to accommodate varying numbers of feature extraction levels, the resulting impact on the number of features extracted from each electrode necessitates the reduction of dimensionality. This reduction is essential to prevent overfitting of the decoder during individual sessions. To address this concern while maintaining the single channel architecture of the feature extraction technique, we utilized Partial Least Square regression⁶⁶ (PLSR). Specifically, PLSR was independently applied to the features extracted from each channel. The objective was to condense the 8 extracted features obtained from each electrode into a smaller set of features, specifically 2 features in this case.

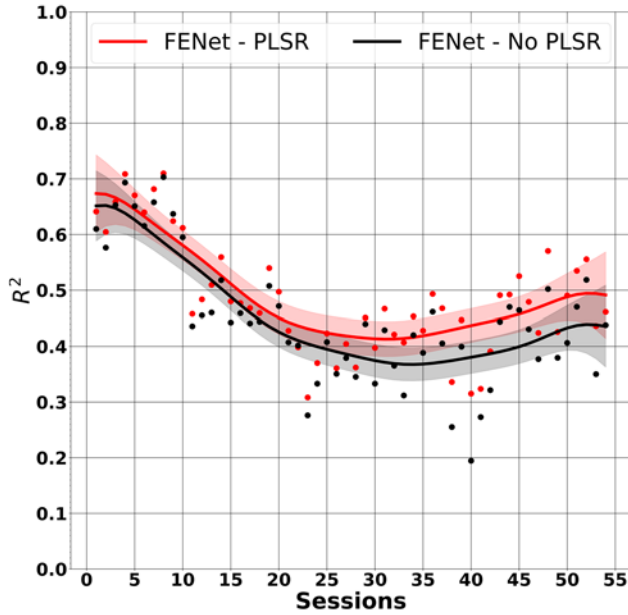
In addition, we have compared the performance of FENet with and without PLSR on all the 54 sessions of participant JJ’s open-loop data.

We have added this to the “Training and Inference for FENet” section of the methods:

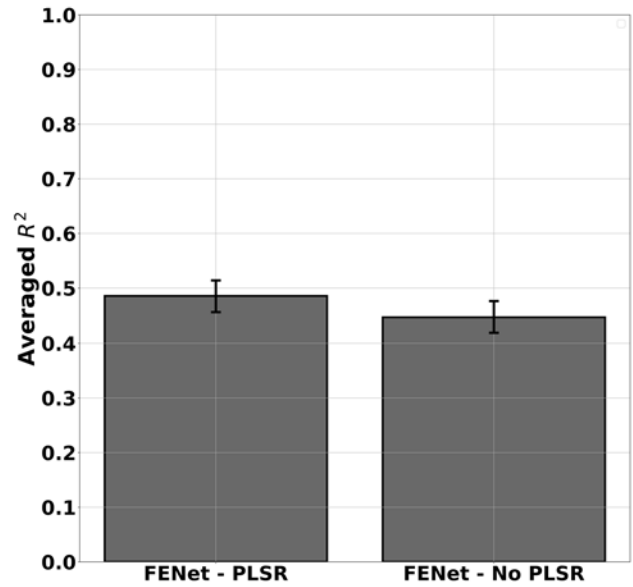
In order to evaluate the impact of Partial Least Squares Regression (PLSR) on the performance of the linear decoder operating on FENet, we conducted a rigorous analysis utilizing data from all 54 sessions of participant JJ (Extended Figure 3c, d). Our evaluation involved a comparison of FENet’s performance with and without the application of PLSR, specifically applied to the top 40 electrodes within these sessions. The selection of these top 40 electrodes was motivated by the goal of mitigating potential overfitting issues that may arise in the linear decoder, particularly in scenarios where PLSR is not employed. The results depicted in this figure provide compelling evidence demonstrating that FENet exhibits the ability to effectively capture informative features from the vast neural data, irrespective of the presence or absence of PLSR. Additionally, the application of PLSR plays a vital role in reducing the dimensionality of the extracted features. This dimensionality reduction step is crucial as it helps prevent overfitting of the decoders, particularly when working with limited neural data obtained from human

participants within each session. These findings highlight the robustness and efficacy of FENet as a feature extraction technique in neural decoding tasks. Furthermore, they underscore the importance of employing dimensionality reduction methods such as PLSR, which can enhance the performance and generalizability of the linear decoder by mitigating the risk of overfitting when working with limited neural data.

We have also added the corresponding results to the Extended Figure 3:



(c)



(d)

Extended Figure 3. (c, d) we conducted a comprehensive evaluation of the effect of PLSR on the performance of linear decoder operating on FENet's using all 54 sessions of participant JJ. We compared FENet's performance with and without Partial Least Squares Regression (PLSR) applied to the top 40 electrodes in these sessions. We selected the top 40 electrodes to mitigate overfitting in the linear decoder, particularly in cases where PLSR is not applied. The results presented in this figure demonstrate that FENet, regardless of PLSR, effectively captures informative features from the broad neural data. The application of PLSR serves to reduce feature dimensionality and prevent overfitting of the decoders when working with limited neural data from human participants per session.

Minor Comment 2: *More interpretation of why FENET succeeded would be helpful. Can we see the output of FENet compared to TCs? Do the authors have any guesses as to what its doing?*

Response: We refer the reviewer to the answer to “Comment 9” of major comments of “Reviewer 3”.

Missing citations

Citation Comment 1: *PSID (Sani et. al. Nature Neuro 2020) should be added.*

Response: We have added this citation and our results with PSID to the manuscript. We refer the reviewer to “Comment 1” of “Reviewer 3”.

Optional suggestions

Optional Comment 1: *Comparison to algorithms like LFADS or MYOW (from Eva Dyer's group) would make the FENet even more compelling, but these are non-causal algorithms so it is not necessary to do this.*

Response: In this paper we find a method to improve the information content of neural features. We are not trying to improve the mapping from features to kinematics. PSID, LFADS, and MYOW are all techniques that promise to improve the feature to kinematic mapping. However, this is outside the scope of the current paper. It is our assumption that more informative features will assist any algorithm that transforms features to kinematics. This assumption has been evaluated when different decoders operate on different studied feature extraction techniques ("Comment 1" of "Reviewer 3"). However, since among all the mentioned technique, PSID is causal and agrees with the presented decoding techniques, we have tested the performance of PSID operating on the presented feature extraction techniques. FENet combined with PSID is still providing a higher performance on all the sessions compared to other feature extraction techniques. We refer the reviewer to the answer to "Comment 1" of major comments of "Reviewer 3", and to "Comment 1" and "Comment 2" of Reviewer 5.

Stylistic Issues

Stylistic Comment 1: *Line 117: This line doesnt really make sense. Explain what the exploration was and what it meant.*

Response: We have clarified it by adding to Extended Figures 6, 7:

The amount of training data, hyperparameters, the importance of FENet extracted features per electrode, and the computational cost of FENet per evaluation for different FENet architectures were explored in Extended Figure 6, 7.

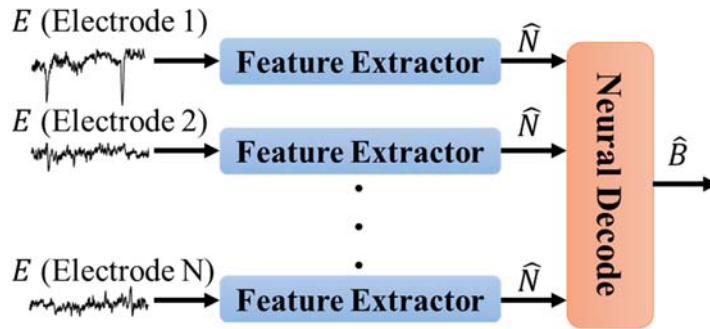
Stylistic Comment 2: *Line 150 explain more why this is the case.*

Response: Please refer to "Minor Comment 19" of "Reviewer 3". To explain more, we have added this to the discussion:

In this study, we conducted an offline analysis to compare FENet's performance with multiple other feature extraction techniques using different linear and non-linear decoders. To expand the scope of comparison across different time periods and feature extraction techniques, we assessed FENet's capability to reconstruct movement kinematics using previously recorded neural data from implanted electrode arrays. A retrospective analysis over years of recordings showed that FENet significantly improved the cross-validated R^2 and the SNR of extracted neural features compared to the other feature extraction techniques across multiple patients and through the lifetime of the arrays. Further, FENet generalized well across cortical brain regions, patients, and tasks, demonstrating its ability to serve as a drop-in replacement for other feature extraction techniques. Moreover, the population-level analysis demonstrated that FENet preserves the representational structure and temporal dynamics of sorted neural populations and, thus, provides an accurate measure of brain activity. Consequently, according to the results of our offline analysis, we evaluated and compared the performance of threshold crossings (TCs), as the current standard for closed-loop control^{4,6,22,26,27,31,33}, and Wavelet transforms (WTs), which have also demonstrated performance improvements in our offline analysis and in the recent studies on BMIs^{20,29}, against FENet features in our closed-loop analysis. Neural decoders employing FENet-based features outperformed TC-based and WT-based features across all metrics. Additionally, FENet enhanced the cursor's responsiveness to the participant's intent, reducing the time it took for the cursor to move towards the target after its onset. These improvements in both aspects led to significant enhancements in overall task performance during the closed-loop sessions. While we reported both the open- and closed-loop performances for participant JJ, our evaluation of the presented feature extraction techniques was limited to the recorded open-loop neural data from participants EGS and NS in specific tasks, as their participation in the clinical trial concluded, and their electrodes were explanted. Nevertheless, the principles derived from analyzing the recorded open-loop neural data from EGS and NS will hold relevance for future subjects with electrodes in the same cortical areas. Taken together, FENet can improve the efficacy of implantable electrode systems while delivering improved performance and ease of use.

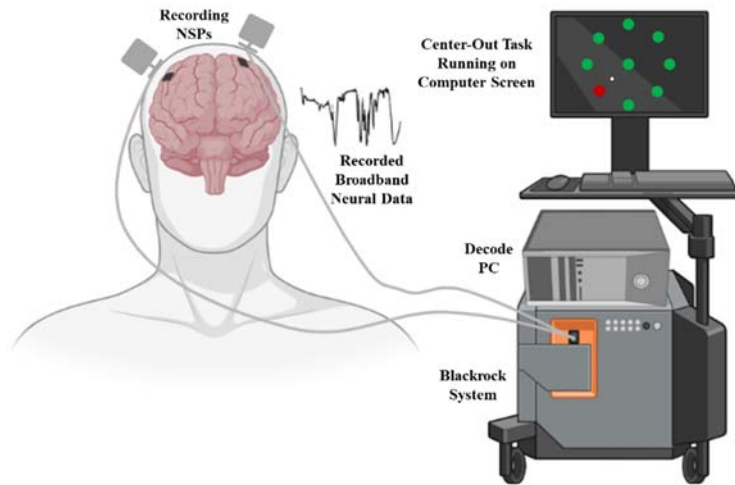
Stylistic Comment 3: *Fig 1b. Fixed mapping is a bit confusing. Its a learnable network right? Different terminology is needed.*

Response: What we meant by the term "Fixed Mapping" was that FENet is trained over all the sessions, while the decoder is trained from scratch on each single session. However, to avoid confusion, we have removed this term and we believe that the terms "feature extraction" and "Decoder" are enough to convey what we meant in this figure.



Stylistic Comment 4: Fig 1e Utah arrays are way too big? Should be more to scale.

Response: We have made the Utah Arrays smaller and updated the corresponding figure.



Stylistic Comment 5: Fig1f can see participants face, make sure thats okay.

Response: We have submitted an approval with this manuscript mentioning that participant agreed to show his face in the images and the supplemental videos.

Stylistic Comment 6: Fig 3: Debaucheries wavelets instead of debauchies wavelets in one fig caption

Response: We have corrected this.

Stylistic Comment 7: Figure 5: Line 819: Robust brain-computer interface (BMI) - should be BCI

Response: We meant Brain-Machine Interfaces (BMI). We have corrected this.

Stylistic Comment 8: Fig 4: Consider different visualization, the grey colors used look sloppy and differing grid sizes make this a bit hard to read.

Response: In order to enhance the readability of the printed manuscript, efforts were made to minimize the inclusion of colorful figures. In this particular figure, darker colors represent lower performances, while brighter colors indicate higher performance levels. To optimize legibility, a strategy was employed wherein contrasting colors were used, such as black for brighter colors and white for darker colors. This approach was implemented with the aim of improving the overall readability and interpretability of the figure.

Reviewer 6

Major Comments

Comment 1: *First, the authors use qualitative terms rather than quantitative values. For example, the title claims the new method “dramatically” improved control. The reviewer understandx this makes a great title, however, also leaves a lot of room for interpretation. Authors are encouraged to be specific and clear, and not use such superlatives if possible.*

Response: We have removed the word “dramatically” from the title. Also, we have used more qualitative terms. Please refer to “Comment 13” of “Reviewer 3”.

Comment 2: *Secondly, and most importantly, authors use the method as a black box and do not address explainability. The article would have more impact and readers would be more likely to trust the results if the authors could show why there was an improvement in control performance.*

Respond: Please refer to the responses to “Comment 9” of “Reviewer 3” and “Comment 1” of “Reviewer 4”.

Minor Comments

Minor Comment 1: #67-72: *Why these considerations and the given order? The computational complexity, consideration 3, was not addressed or discussed in enough detail in the work. Related to consideration 4), authors are encouraged to commend and discuss the generalizability of the method and their pro and cons in more detail.*

Response: Please refer to the response to “Comment 8” of “Reviewer 3” for more discussion on the computational cost of FENet.

Minor Comment 2: #85: *Authors state “improvement of ~50%”. Can the reviewer suggest that authors provide more context or a baseline performance, so that reader understand what this means.*

Response: The baseline has been considered to be the performance of threshold crossings (TCs) as the most used techniques in the actual clinical settings. Moreover, we have added and compared the offline performance of more decoders (SVR, LSTM, REFIT Kalman Filter, and PSID) operating on FENet featurers versus other feature extraction techniques (WTs, MUA, HF-LFP, TCs – HF-LFP, and FENet- HF-LFP). Please refer to the response to “Comment 1” of “Reviewer 3”. We have changed this sentence as below:

FENet demonstrated a minimum improvement of ~50% in cross-validated coefficient of determination (R^2) [compared to TCs, as the current standard feature extraction technique in labs work with human subjects^{4,6,22,26,27,31}](#), across multiple patients and through the lifetime of the arrays.

Minor Comment 3: #87: *Authors analysed data from 3 patients. How confident are they that FENet generalises over regions, patients and task? This should be discussed in more detail. See comment above.*

Response: In discussion, we mention that three is still a small number. By providing code publicly, this can be tested out. Please refer to “Minor Comment 19” of “Reviewer 3” and “Stylistic Comment 2” of “Reviewer 5”. We have also added this to the discussion to discuss challenges of working with SCI human participants:

[In the current study we found that FENet generalized well between three patients, three brain regions, closed- and open-loop settings, and up to five years of recordings. This provides preliminary confidence that FENet provides a generalizable improvement to current feature extraction methods. However, it remains possible that FENet will not improve performance across all subjects, tasks, and array technologies. We therefore provide our code in a public repository in the hope that additional clinical sites will test and ultimately improve FENet.](#)

Minor Comment 4: #121: *“Dramatically” Please use quantities whenever possible and not qualitative statements. Or precisely define what “dramatically” means.*

Response: Considering this suggestion, we have softened the language accordingly. Please refer to the response to “Comment 13” of the response to “Reviewer 3”.

Minor Comment 5: #187/Fig4: (a) Data from year 2019 yields 0.15 higher R^2 compared to data from other years. Why is this the case? And why is this not the case in Fig4(c)?

Response: It is because of the well-known signal degradation over time for implantable electrode arrays (Barrese et al. 2013). This has been discussed at the end of the “FENet improves closed-loop control” section as below:

“This poor performance was the consequence of significant degradation in the quality of the neural signals over the lifetime of the recording arrays³⁴ (see Extended Figure 2). As shown in the next section, we find that FENet can improve the performance across the lifetime of the array (even when TCs produce excellent performance) and across the participants.”

Also, the change in the quality of neural activities has been shown in Extended Figure 2. Despite JJ, the neural data was recorded only from PPC brain regions for EGS, which provide different behavior than M1 (Aflalo et al. 2015).

Minor Comment 6: #288-304: What was required of the participants to gain control? Were instructions given on what specifically to do? What to intend? Adding this information would give a clearer picture of the task.

Response: We have explained the details of the instructions given to the participants in the “Behavioral tasks” section of the methods. We have also added more details as below:

Behavioral tasks. Data was collected while participants performed various two-dimensional control tasks, including center-out^{5,19,28}, grid, and finger-grid⁴¹ tasks using [pseudo-random interleaving of targets](#) to ensure balanced statistical sampling of movement directions⁸. In the center-out task, a cursor moves in two dimensions on a computer screen from a central target outward to one of the eight targets located around a circle, and back to the center (Figure 1e, 1f top). We define a trial to be one trajectory, either from the central location outward to the peripheral targets, or from the peripheral targets back to the center target. In the grid task, the target appears in a random location in an 8-by-8 squared grid on the computer screen and the cursor moves starting from the old target to the newly appeared target (Figure 1f). Cursor movement kinematics are updated every 30ms for JJ and every 50ms for EGS. For the purposes of this study, we extracted trajectories from 200 ms after target presentation to 100 ms before the cursor overlapped the target. This segment of time captures a window where the subject’s intent should be well defined, after reacting to the presented target and before possibly slowing down as the cursor approaches the target⁸. Neural features were regressed against cursor velocity, which, for simplicity, was modeled as constant amplitude. [Each of these tasks was conducted in either open-loop, in which the cursor movements were fully generated by the computer and the participant did not directly control the cursor’s position, but instead imagines control over a visually observed, computer-controlled cursor, or closed-loop, in which the cursor movements were under the participant’s full control with 0% assistance from the computer.](#)

For the finger-grid task, a text cue (e.g., ‘T’ for thumb) was displayed to the participant on a computer screen in each trial. Then, the participant immediately attempted to press the corresponding finger of the right hand^{46,62} (Figure 7). [To model the multi-finger tasks, we considered the muscle model and somatotopy model⁴⁷. The muscle activation model posits that the representational structure should align with the coactivation patterns observed in muscle activity during individual finger movements. Conversely, the somatotopy model suggests that the representational structure should correspond to the spatial arrangement of the body, wherein neighboring fingers exhibit similar representations. Although somatotopy typically pertains to physical spaces resembling the body, in this context, we employ the term broadly to encompass encoding spaces that resemble the body⁴⁶.](#)

Minor Comment 7: #332: typo, replace “xis” with “x is”

Response: We have corrected this typo.

Minor Comment 8: #363/425: Please clarify what the “ground truth values” are.

Response: We have changed the “ground-truth values” to “ground-truth kinematics”.

Minor Comment 9: #498: It would be helpful if at least one dataset could be released so that colleagues can test the code on the data.

Response: We have uploaded 10 sessions of broadband data for offline analysis, the extracted features for these sessions, and the code itself in the submission system.

Rebuttal 2

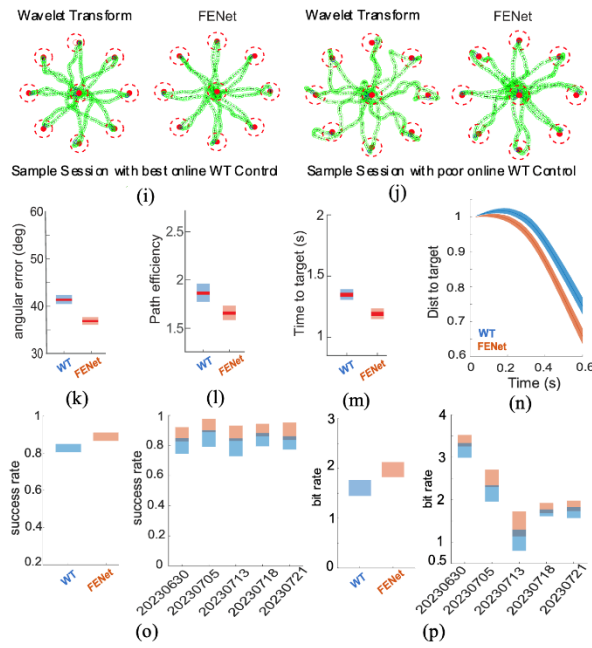
Response to Reviewers

Reviewer 4

Comment 1: *The authors did substantial offline analysis to compare to different features already in the literature and common use, which did indeed make up a lot of the distance between threshold crossing and FENET. In doing so, I think it is clear that the headline results, and the only online data in the paper, was a comparison of a new method with an outdated method rather than the state of the art. No new data was added to show that there is still a big online performance increase compared to the state of the art, and it's inappropriate to include data comparing to an older approach. It does seem like all reviewers agree on this point. Similarly, the absolute performance is not high enough to conclude that there is something fundamentally new here, since it still remains below the best performance seen in the literature. It's true that this can be from a variety of factors, but that makes an online comparison within subject all the more critical. Overall, while there are interesting offline observations that are of interest to the neural engineering community, without a clear and substantial performance improvement online, this is better suited for publication in a more specialty journal.*

Seeing what the FENet actually produces compared to the other features, it appears to be doing basically an emphasis of spikes, like wavelets and MUA. The fact that the authors can do it with 30 ms of data instead of 50ms or more is of interest, but again more suited to a specialty journal. I agree with the other reviewer that the approach here is very close to Willsey et al which has a similar convolutional layer that may be achieving the same end in fewer operations, and this should be checked. The MUA presented here in Extended Figure 8 does seem oddly close to TCs in SNR, which is not consistent with the observation of others. Zhang et al represents a post-hoc offline analysis, so I recommend replicating the details of MUA implementation from any study that utilized these features online, which may involve shorter timebins and a higher highpass.

Response: We appreciate the time and effort reviewer 4 has dedicated evaluating our manuscript and providing feedback. As an initial point, we agree with the reviewer that current BMI studies, particularly clinical studies, often rely on small sample sizes making within-subjects comparisons critical for evaluating the performance of new decoding approaches. For example, many studies report the performance of ~2-subjects and in these cases, performance differences between subjects vary up to 42% [Young et al. 2019], 90% [Pandarinath et al. 2017], and 42% [Fan et al. 2014]. We have added clarifying notes in the introduction and discussion to emphasize this point (see below). We also agree that measuring within-subject performance differences is critical. In the initial submission, we compared online performance of FENet against threshold-crossings and demonstrated a significant improvement (e.g., average bit rate improvement of ~257%, $t = -11.85$, $p = 0.0000$). The studies that have shown the best performance in terms of bitrate have used threshold crossings and thus we thought this was a reasonable comparison (see references 4, 6, 22, 26, 27, 31, and 34 in our manuscript). However, based on reviewers' feedback, we updated the manuscript to include additional online performance testing that compares FENet against wavelet-based features extraction. We chose wavelets since they performed significantly better than all other current feature extraction approaches in our offline analysis (see figure 3, figure 4, and figure 5). This new data shows that FENet significantly outperforms wavelet-based features in online testing (see figure 2, copied below, e.g., average bit rate improvement of ~21%, $t = -4.252$, $p = 0.0000$). Additionally, the advancement in performance seen in publications within the Nature family journals, introducing fresh decoding techniques, such as Willsey et al. 2022, Sani et al. 2021, and Schwemmer et al. 2018 typically falls between 2% and 15%. This range aligns comparably with our own research. Thus, we disagree with the reviewer's point that "No new data was added to show that there is still a big online performance increase compared to the state of the art".



The reviewer raises an additional point that the performance differences we observe using FENet could also be realized using alternate decoding approaches. This is likely the case, but we think may miss the mark of the current manuscript. Every BMI system currently used can be understood as involving two stages. The first stage, feature extraction, maps raw broadband data into neural features (e.g., threshold crossing, WT, MUA, etc) and the second stage, feature decoding, which maps features into the decoded variable of interest (e.g., cursor velocity, which finger to move, etc). We agree with the reviewer that advanced feature decoding algorithms can theoretically make up the difference in feature extraction methods. However, it is our position that current feature extraction methods are sub-optimal at extracting task-relevant neural information from raw broadband data. There is more juice to be squeezed! Thus, better feature decoding methods will always be improved by better feature extraction methods because they simply have more information to work with [Chandrashekar et al. 2014, Khalid et al. 2014, Cai et al. 2018]. To help make this point, we added figure 4 showing that all feature decoding methods are significantly improved when using FENet features as compared to other features. We think the lack of emphasis on feature extraction has led to sub-optimal decoding. We hope our manuscript helps shift the spotlight on this underdeveloped aspect of neural decoding.

Based on the above considerations we have made the following revisions to introduction

Due to the inherent variability in absolute performance of BMI systems across subjects, labs, tasks, and implant sites and age, we employ within-subject comparisons to assess the efficacy of FENet. This approach aligns with current recommendations for evaluating BMI performance and underscores the importance of considering subject-specific factors when interpreting results^{5,32–35}.

And this to the discussion section:

Furthermore, it is crucial to acknowledge that the performance of a BMI system can be influenced by various factors, both within and outside of our control. These factors may include the nature of the subject (human or non-human primate), implant site and age, recording yield, task, and the specific decoder employed^{5,32–35}. Recognizing this inherent heterogeneity in BMI performance across subjects, tasks, and labs, we adopt a within-subject experimental setup to evaluate FENet. We observed enhanced performance when utilizing features extracted by FENet with improvements across all datasets included in this study (three human participants, 192 total electrodes, and many hours of neural data representing multiple years of implantation). We found that FENet generalized well between three patients, three brain regions, closed- and open-loop settings, and up to five years of recordings. This provides preliminary confidence that FENet provides a generalizable improvement to current feature extraction methods. However, it remains possible that FENet will not improve performance across all subjects, tasks, and array technologies. We therefore provide our code in a public repository in the hope that additional clinical sites will test and ultimately improve FENet.

Finally, the reviewer points to 2 papers, Willsey et al. and Zhang et al., that we should evaluate against our current paper. Willsey et al. uses band-power as a feature extraction method and a convolutional neural network as a feature decoding method.

The primary point of our paper is that significant information is lost when using band-power as a feature extraction method and thus the papers are fundamentally different in approach. As we show in figure 4 of the revised manuscript, using more sophisticated feature decoding approaches can improve performance beyond a simple linear decoder (consistent with Willsey et al. and not the emphasis of our paper), however, every feature decoding approach benefits from FENet, including neural network-based approaches (which is the novelty of our paper and what distinguishes it from the Willsey et al.'s paper). About comparison with Zhang et al., the participants in our work are focused on center-out and grid tasks as regression tasks and we are reporting correlation of determination (R^2) as the measure in our analysis, which differs from classification accuracies reported by Zhang et al. for hand-wrist movements task as a classification task. Ultimately, MUA features are extracted from neural data bins (30 ms for JJ and NS, and 50 ms for EGS) following the approach outlined by Stark and Abeles 2007. Notably, we refrained from implementing post-hoc offline steps observed in Zhang et al., such as the 1-second moving average and 15-second baseline subtraction. Our detailed approach for extracting MUA features is elucidated in the “Generation of other features” section within the Methods:

We have extracted the features from the 30 ms bins of [recorded broadband neural data](#) for JJ and NS, and from 50 ms bins for EGS [without any post-hoc offline steps](#).

- [1] Young et al. Closed-loop cortical control of virtual reach and posture using Cartesian and joint velocity commands. *J Neural Eng*, 2019.
- [2] Pandarinath et al. High-performance communication by people with paralysis using an intracortical brain-computer interface. *Elife*, 2017.
- [3] Fan et al. Intention Estimation in Brain Machine Interfaces. *J Neural Eng*, 2014.
- [4] Sani OG. Et al. Modeling behaviorally relevant neural dynamics enabled by preferential subspace identification. *Nature Neuroscience*. 2021.
- [5] Schwemmer et al. Meeting brain-computer interface user performance expectations using a deep neural network decoding framework. *Nature Medicine*, 2018.
- [6] Willsey et al. Real-time brain-machine interface in non-human primates achieves high-velocity prosthetic finger movements using a shallow feedforward neural network decoder. *Nature Communication*, 2022.
- [7] Chandrashekar et al. A Survey on Feature Selection Methods. 2014.
- [8] Khalid et al. A Survey of Feature Selection and Feature Extraction Techniques in Machine Learning. 2014.
- [9] Cai et al. Feature Selection in Machine Learning: A New Perspective. 2018.

Reviewer 3

Major Comments

Comment 1: *from the definition of bitrate now given, it appears that the metric was not set up to penalize incorrect target selections. This differs from the standard “Achieved bitrate” metric definition ([1], and see also [2] for a discussion of the importance of assessing error rates in information-based performance metrics). This formulation breaks the idea of the metric being used as a measure of channel capacity / information transfer rate for a discrete set of symbols, as measuring information transferred requires knowing when symbols were incorrectly transferred also. If incorrect selections were possible, the bitrate metrics should be corrected to account for that. If incorrect selections were not possible, the bitrate numbers from this manuscript will not be directly comparable to other measurements in the literature. Therefore this discrepancy should be clearly stated in the main text.*

Response: Thank you for your comment. We agree with the reviewer and report the achieved bitrate in Figure 2 as

$$B = \frac{\log_2(N) \times \max(S_c - S_i, 0)}{t}$$

where N signifies the total number of targets on the screen, S_c denotes the correct selections, S_i represents the incorrect selections, and t stands for the elapsed time in seconds. We have updated the methods section of the manuscript to reflect that we are using achieved bitrate.

Comment 2: *Previous comment 12 on non-causal filtering - this comment seems to have been ignored. (as noted, non-causal filtering was a more important comparison point than the exponential smoothing that the authors seem to have focused on). Please provide a comparison of FENet features and the threshold crossing features extracted via non-causal filtering.*

Response: Our primary focus within this study was to test techniques that can be applied in a causal real-time system. Thus, while non-causality could potentially enhance a decoder’s performance operating on threshold crossing features by employing a filter over the data twice - once forward and once in reverse, thereby providing future information to the feature extraction technique, we think such a comparison is outside the scope of the current manuscript. Further, Masse et al. did not offer comparisons of closed-loop performance, making it difficult to directly compare effects on online performance where e.g., latency added by non-causal filtering are likely to impact online control. We have added the below paragraph to “Generation of other features” in Methods section to answer this question:

In this study, our primary objective has centered on maintaining causality in our filtering methodology to prevent system latency. While non-causality might offer potential performance enhancements by employing a forward-reverse data filtration process, it necessitates defining a system delay to enable its applicability. Such delay contradicts the requirements of a real-time system, particularly when the feature extraction window size is comparable with the delay required by the non-causal filtering techniques.

Minor Comments

Minor Comment 1: 663-681: *ReFIT-KF methods are absent. Evaluation was done on open-loop kinematics, so it is unclear how the authors estimated intention to recalibrate the decoder for use with open-loop data. Did the intention come from subsequent closed-loop trials? Or was this just a standard Kalman filter? If the former, please state so in the methods. If the latter, please update the text throughout to accurately reflect the decoder.*

Response: Thank you for your feedback. In our offline analysis, we utilize the Kalman filter without incorporating any feedback during the decoding process. Consequently, ReFIT-KF simplifies to the standard Kalman filter. As a result, we have substituted ReFIT-KF with KF in our manuscript and have updated the caption for corresponding figures accordingly.

Minor Comment 2: 199-200: *The authors claim FENet improves SNR, but the evidence (from Extended Figure 8) is qualitative. I suggest computing SNRs to quantitatively show FENet improves SNR, or remove the claim about SNR. Perhaps the conclusion comes from FENet features yielding higher values in the tuning curves, but higher feature values should not be confused with higher SNR.*

Response: Acknowledging the potential confusion related to the term SNR, we have reworked these lines to demonstrate that FENet generates higher values within tuning curves and achieves improved trial separability compared to other feature extraction techniques. We have updated the first sentence of the last paragraph of introduction as below:

Additionally, a retrospective analysis over years of recordings showed that FENet generates a higher magnitude peak-to-trough within tuning curves and achieves improved trial separability compared to other feature extraction techniques over the entire lifetime of the array.

We have also updated the corresponding sentence in the fourth paragraph of discussion as below:

Notably, FENet exhibited significant improvement in the amplitude of the preferred versus anti-preferred directions in the tuning curves thus improving the ability to distinguish individual trials.

Finally, we have updated the last sentence of the caption of Extended Figure 8 as below:

The results demonstrate the preservation of the fundamental tuning characteristics of the neurons, FENet generates higher values within tuning curves and achieves improved trial separability compared to other feature extraction techniques.

Minor Comment 3: 306-309: *“Furthermore, it is worth noting that future BMI systems may not universally support the capture of raw, high-sample rate broadband data (e.g., Neuralink) to enable training of new per-channel models. In such cases, our approach can be seamlessly utilized without the reliance on this type of data.” - Assuming “this type of data” refers to broadband data, it is not immediately clear to readers how FENet can be “seamlessly” adapted to systems that do not record broadband data, since only broadband data was discussed in the manuscript. Please discuss how FENet could be adapted (i.e. to systems that only provide spike times, like Neuralink, or data collected with lower sampling rates).*

Response: We understand the confusion. The point we attempted to make is that implantable systems simply will not have the wireless bandwidth to transmit raw broadband data off-device to enable training new models. The generalizability of FENet across patients and implants is powerful as it can be deployed on device and will work without the need for additional training using newly collected patient specific broadband data. For clarity, we have elaborated upon this sentence in the discussion section as follows:

Furthermore, it is crucial to consider the bandwidth limitations of implantable BMIs. Even with systems that currently record high sample-rate broadband data, transmitting such data off-device for continuous model retraining remains impractical due to bandwidth and power constraints. This scenario underscores the critical advantage of FENet, which can be directly deployed without requiring further training on newly collected broadband data. This capability paves the way for robust BMI performance across diverse patients and implants.

Minor Comment 4: *Response: After extracting the features from the 30ms windows, we averaged the resulting windows, with the number being averaged defined by the bin length. This will result in a smoothing effect, as explained in figure 5’s caption: “Extended Figure 5. The effect of changing the window size after averaging the features over larger window sizes.” I could not understand what this statement means “The effect of changing the window size after averaging the features over larger window sizes”. Please revise the text to clarify.*

Response: The aim of this analysis was to evaluate how altering the feature extraction window size influences the performance of decoders operating on the extracted features. In this investigation, FENet continued to be trained on data segmented into bins with a window size of 30 ms for JJ. However, we augmented the window size to observe whether FENet demonstrates superior performance compared to TCs. Additionally, the utilization of a larger history of broadband data for feature extraction with increased window sizes introduces a smoothing effect in the decoding process. Consequently, we anticipate improvements in decoder performance for both feature extraction techniques due to this inherent smoothing effect. For clarity, we have updated the figure caption as below:

The impact of modifying window sizes on decoder performance subsequent to averaging the extracted features across larger temporal windows.

We have also added these to the section “FENet provides improved open-loop decoding performance” in Results section:

This analysis aimed to assess the impact of varying feature extraction window sizes on the performance of decoders using the extracted features. Throughout this study, FENet remained trained on data partitioned into 30 ms bins. However, we expanded the window size to ascertain if FENet exhibited superior performance compared to TCs during inference. Furthermore, the utilization of a more extensive history of broadband data for feature extraction with larger window sizes introduces a smoothing effect in the decoding process. Consequently, we observe that both feature extraction techniques demonstrate improved decoder performance owing to this inherent smoothing effect.

Reviewer 6

Comment 1: *The WT Filter shown in Fig. 9 seems to suggest that authors used a one-level decomposition, and not a multi-level decomposition. Is this the case? If so, is this standard in the field?*

Response: Thank you for your input. To provide further clarification, we have previously outlined the specifics of our multi-level Wavelet Transform (WT) feature extraction in the “Generation of other features” section within the Methods segment:

To extract wavelet features (WTs)^{19,20}, we used a db20 mother wavelet with 7 scales on moving windows (no overlap) of the time series recorded from each electrode. A db20 mother wavelet was selected as it contains filters with length 40 and can model WT's high pass and low pass filters more accurately compared to other Daubechies wavelet families²⁰. Moreover, in our preliminary experiments, we found that it outperformed other wavelet variants for the datasets tested in the current study¹⁹ (e.g., db4, Haar). The mean of absolute-valued coefficients for each scale was calculated to generate $M = 8$ time series per electrode, including seven detailed coefficients and one approximation coefficient generated by the WT high-pass filters and the final stage WT low-pass filter, respectively.

Comment 2: *On page 5, paragraphs 2 and 3, the authors use the "best electrode" or "top electrode" from selected datasets to show the effectiveness of the proposed FENet feature extraction capabilities. It is not clear to the reviewer how the top electrodes were selected. What criterion was used (the one described in the extended Figure 6 caption?), and was the best electrode for each feature extraction method used independently for comparison, or the best electrode with respect to FENet? Also, how were the sessions/datasets chosen? Were the data selected arbitrarily, or are they a good general characterization of the data? Are they part of the evaluation data set?*

Response: To provide further clarity regarding the session and top electrode selection process, we have included this in the caption of Extended Figure 8 and within the “FENet provides improved open-loop decoding performance” section:

To identify the top electrode within a session, we organized electrodes based on their individual electrode R^2 values, indicating the linear predictability of kinematics for each electrode using each distinct feature extraction technique. Subsequently, we randomly chose three sample sessions spanning 2019, 2020, and 2021 from those where the index of the top electrodes remained consistent across all feature extraction techniques.

Rebuttal 3

Response to Reviewers

Reviewer 3

Comment 1: Overall, the authors have largely addressed my concerns. My one piece of feedback is that the authors seem to have misunderstood some of the recent literature with this comment:

> Our primary focus within this study was to test techniques that can be applied in a causal real-time system. Thus, while non-causality could potentially enhance a decoder's performance operating on threshold crossing features by employing a filter over the data twice - once forward and once in reverse, thereby providing future information to the feature extraction technique, we think such a comparison is outside the scope of the current manuscript.

The non-causal filtering approach described in Masse et al. is used in high-performing causal real-time systems, it simply requires a small amount of buffering of the signal (e.g. a few milliseconds). To name a few early examples, it is used online in Jarosiewicz et al., *Science Trans Med* 2015, Gilja et al., *Nature Med* 2015, and Pandarinath et al., *eLife* 2017. It has been used in many subsequent demonstrations.

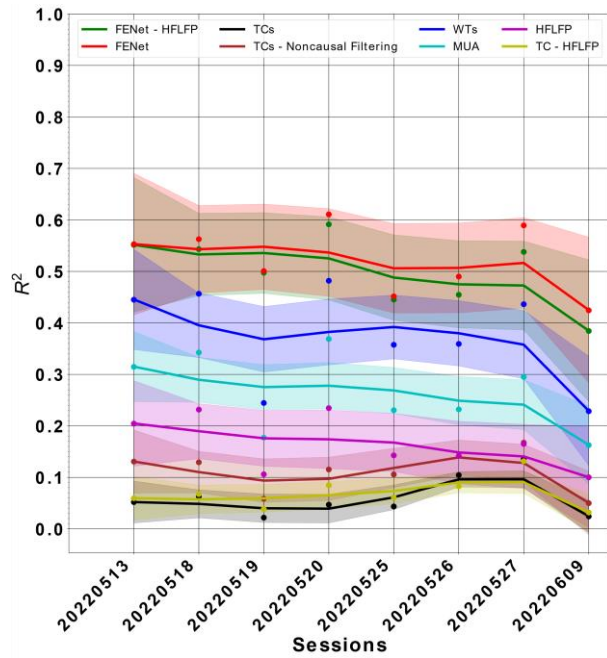
Therefore, when the authors state:

> While non-causality might offer potential performance enhancements by employing a forward-reverse data filtration process, it necessitates defining a system delay to enable its applicability. Such delay contradicts the requirements of a real-time system, particularly when the feature extraction window size is comparable with the delay required by the non-causal filtering techniques.

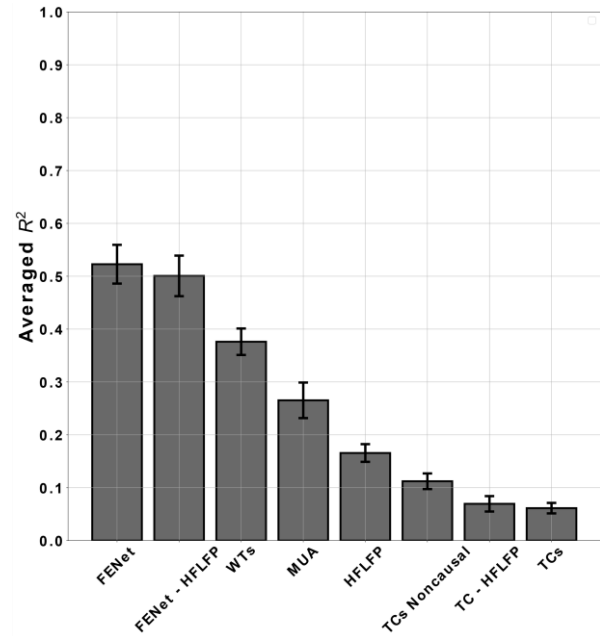
This assertion is clearly inaccurate given the literature and should be removed. Given the prominence of this filtering approach in the clinical BCI literature, it would be preferred that the authors' do the requested comparison rather than claiming it is out-of-scope.

Response: Thank you for your feedback. Following this comment, we have evaluated the performance of the decoder operating on the Threshold-Crossing features (TCs) by applying a non-causal filter described in Masse et al. Figure below shows the performance of the decoder on eight sessions. While the average R^2 of the decoder increases from 0.0614 to 0.1123 when applying a non-causal filter before extracting TCs, still there is a large gap between the performance of non-causal filtered TCs and FENet, which has the average R^2 of 0.5228 on these eight sessions. We have added the below paragraph in the discussion:

Our primary focus within this study was to test techniques that can be applied in a causal real-time system. To explore the effect of applying a non-causal filter when extracting TCs⁴⁴, we evaluated the decoder's performance when using FENet as the feature extraction technique versus TCs preceded by a non-causal filter that was applied over the data twice - once forward and once in reverse. Supplemental Fig. 1 shows the performance of the decoder on eight sessions. This figure highlights the relative strengths of different feature extraction approaches in enhancing decoder performance, with FENet emerging as the most effective technique. The non-causal filtering applied to TCs improves performance but is still notably less effective than FENet in terms of decoding accuracy.



(a)



(b)

Supplemental Fig. 1| Decoder performance comparison across eight sessions using different feature extraction techniques. (a) The session-wise R^2 values of the linear decoder operating feature extraction techniques applied over the course of eight sessions. Each line represents a different feature extraction method: FENet (red), FENet + HFLFP (green), TCs - Noncausal Filtering (black), WT (blue), MUA (cyan), HFLFP (purple), and TC + HFLFP (pink). The shaded regions around each line represent the standard deviation across the sessions. FENet consistently achieves the highest performance, while Threshold-Crossing (TC) techniques, both with and without non-causal filtering, lag behind. Notably, applying non-causal filtering to the TC features improves the decoder's performance slightly, but the performance still remains far behind the FENet method. (b) Averaged R^2 performance. The bar plot demonstrates the average R^2 values across all sessions for each feature extraction technique, with error bars indicating the standard error of the mean. FENet provides the highest overall averaged performance, with an average R^2 of 0.5228 on these eight sessions, notably outperforming the TC-based methods. The TCs - Noncausal Filtering method improves the decoder's performance (average R^2 of 0.1123) compared to standard TCs (0.0614), however, there remains a substantial gap between this method and the FENet decoder. This figure highlights the relative strengths of different feature extraction approaches in enhancing decoder performance, with FENet emerging as the most effective technique. The non-causal filtering applied to TCs improves performance but is still notably less effective than FENet in terms of decoding accuracy.

Reviewer 6

Comment 1: *The authors have revised the work and have adequately addressed all the reviewer's comments and suggestions. The performance of the novel CNN-based feature extraction network is clearly presented, and the advantages of the proposed approach are demonstrated. There are no technical issues, and the methods are sound. The data basis is limited, but this is a common challenge in this research area and comparable with other works with invasive BMIs.*

I would just like to make one small comment. In lines 186-189, it would be helpful if the authors could add a statement confirming that the comparisons are fair and not just done with some standard parameters but optimized for each method.

Response: Thank you for your comment. The optimized values of the parameters have been mentioned under “Decoders” section of the method. We have updated this section as below:

To ensure that improvements in our novel feature extraction method generalize across feature decoding methods, we have also included the performance of additional feature decoders, namely Support Vector Regression^{39–41}, Long-Short Term Recurrent Neural Network (LSTM)⁴², Kalman filter (KF)^{5,24}, and Preferential Subspace Identification (PSID)⁴³. Figure 4 provides a comprehensive evaluation of the performance of these decoders operating on different feature extraction techniques. [The comparisons were conducted using optimized parameters for each decoder, ensuring fair evaluations rather than relying on standard configurations \(see methods\).](#) As we see in this figure, FENet improves decoding R^2 compared to the other feature extraction techniques for all the tested decoders.