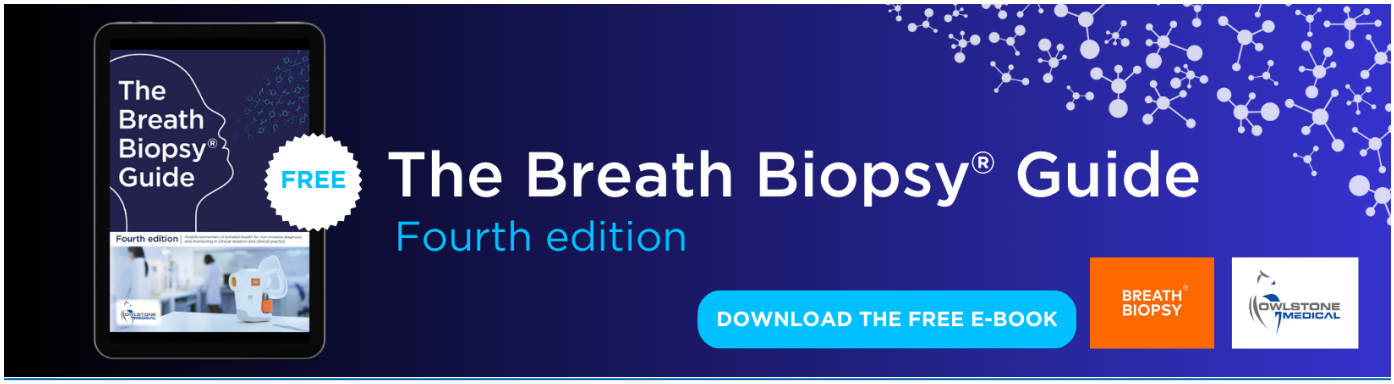
PAPER

Explainable machine learning predictions of perceptual sensitivity for retinal prostheses



Journal of Neural Engineering

PAPER

Explainable machine learning predictions of perceptual sensitivity for retinal prostheses

Galen Pogoncheff[1,*], Zuying Hu[1], Ariel Rokem[2,3] and Michael Beyeler[1,4]

[1] Department of Computer Science, University of California, Santa Barbara, CA, United States of America
[2] Department of Psychology, University of Washington, Seattle, WA, United States of America
[3] eScience Institute, University of Washington, Seattle, WA, United States of America
[4] Department of Psychological & Brain Sciences, University of California, Santa Barbara, CA, United States of America
[*] Author to whom any correspondence should be addressed.

E-mail: galenpogoncheff@ucsb.edu

RECEIVED
27 July 2023

REVISED
24 January 2024

ACCEPTED FOR PUBLICATION
7 March 2024

PUBLISHED
19 March 2024

Keywords: retinal prostheses, perceptual thresholds, electrode deactivation, explainable AI, Argus II

Supplementary material for this article is available online

Abstract

Objective. Retinal prostheses evoke visual precepts by electrically stimulating functioning cells in the retina. Despite high variance in perceptual thresholds across subjects, among electrodes within a subject, and over time, retinal prosthesis users must undergo 'system fitting', a process performed to calibrate stimulation parameters according to the subject's perceptual thresholds. Although previous work has identified electrode-retina distance and impedance as key factors affecting thresholds, an accurate predictive model is still lacking. *Approach.* To address these challenges, we (1) fitted machine learning models to a large longitudinal dataset with the goal of predicting individual electrode thresholds and deactivation as a function of stimulus, electrode, and clinical parameters ('predictors') and (2) leveraged explainable artificial intelligence (XAI) to reveal which of these predictors were most important. *Main results.* Our models accounted for up to 76% of the perceptual threshold response variance and enabled predictions of whether an electrode was deactivated in a given trial with F1 and area under the ROC curve scores of up to 0.732 and 0.911, respectively. Our models identified novel predictors of perceptual sensitivity, including subject age, time since blindness onset, and electrode-fovea distance. *Significance.* Our results demonstrate that routinely collected clinical measures and a single session of system fitting might be sufficient to inform an XAI-based threshold prediction strategy, which has the potential to transform clinical practice in predicting visual outcomes.

1. Introduction

Retinal prostheses evoke visual precepts by electrically stimulating functioning cells in the retina. To use their device, retinal prosthesis users must undergo *system fitting*, a process performed to calibrate stimulation parameters according to the subject's perceptual thresholds (Hu and Beyeler 2021). Specifically, a critical process in system fitting for the Argus II Retinal Prosthesis System (Vivani Medical, Emeryville, CA; formerly Second Sight Medical Products, Inc.) involves measuring perceptual thresholds using psychophysics. These perceptual thresholds are subsequently used to scale the amplitude of the electrical stimuli patterns delivered to represent recorded video frames. Current practices rely on a well-established adaptive up-down staircase procedure, which predicts perceptual thresholds with reasonable accuracy based on approximately 100 trials of a visual detection task (de Balthasar *et al* 2008). During this procedure, electrodes that fail to elicit phosphenes or exhibit mechanical issues are deactivated, preventing their use in future stimulation.

Perceptual thresholds tend to be unstable across subjects, among electrodes within an implant, and over time (de Balthasar *et al* 2008, Ahuja *et al* 2013, Shivdasani *et al* 2014, Yue *et al* 2015, Hu and Beyeler 2021). This instability remains unexplained by common implant or retinal tissue measurements (Yue *et al* 2015). System fitting therefore requires perceptual thresholds to be estimated for each individual electrode and this procedure must be done on a routine

basis to assure proper functioning of the device. This makes system fitting a time-consuming process for Argus II (60 electrodes), but a new approach will be required to handle future devices with hundreds or thousands of electrodes (Palanker *et al* 2020, Chenais *et al* 2021).

Although impedances, electrode-fovea distances, and electrode-retina distances have been suggested to affect perceptual thresholds (de Balthasar *et al* 2008, Ahuja *et al* 2013, Shivdasani *et al* 2014), the lack of accurate, automated threshold estimation frameworks to date may suggest that these two factors alone are far from comprehensive. Further complicating this matter, many of the clinical parameters presumed useful in threshold estimation are difficult or expensive to collect, are prone to measurement error, and may vary drastically between subjects. It is therefore paramount to know which parameters are worth collecting. An explainable predictive model (Adadi and Berrada 2018, Mehta *et al* 2021) fitted to a longitudinal dataset may help elucidate such parameters.

To address these challenges, we set out to develop explainable machine learning (ML) models that could:

- predict perceptual thresholds on individual electrodes as a function of stimulus, electrode, and clinical parameters ('predictors'),
- infer deactivation of individual electrodes as a function of these parameters, and
- identify significant predictors of perceptual thresholds and electrode deactivation.

While previous studies have modeled relationships between clinical measures and perceptual thresholds (de Balthasar *et al* 2008, Ahuja *et al* 2013, Shivdasani *et al* 2014), they have assumed a linear relationship between clinical predictors and threshold measurements. In this study, we additionally demonstrated the efficacy of using lower-bias ML algorithms to model the complex relationships between these variables. We used SHapley Additive exPlanations (SHAP) to quantitatively compare the contribution of each clinical measure across multiple model types, helping to highlight both linear and non-linear relationships between these clinical measures and measures of perceptual sensitivity. Part of this work was previously presented in Hu and Beyeler (2021).

2. Related work

A handful of previous studies have investigated factors affecting perceptual thresholds in retinal prostheses (de Balthasar *et al* 2008, Ahuja *et al* 2013, Shivdasani *et al* 2014), focusing on a range of stimulus (e.g. pulse polarity, pulse rate), electrode (e.g. area), and clinical (e.g. retinal thickness, position of the implant) parameters.

de Balthasar *et al* (2008) correlated perceptual thresholds with electrode impedance, electrode size, electrode-retina distance, and retinal thickness in six recipients of the Argus I epiretinal prosthesis. The study identified impedance and electrode-retina distance as critical factors for determining perceptual thresholds, but did not attempt to develop a predictive model.

Ahuja *et al* (2013) correlated perceptual thresholds with mean electrode-retina distance (averaged across all electrodes of a subject), the mean distance of electrodes from the fovea ('electrode-fovea distance'), and the dark-adapted full-field light threshold in 22 Argus II recipients. In addition to electrode-retina distance, the study identified the residual light threshold as a critical factor, but did not attempt to predict thresholds from these factors on individual electrodes.

Shivdasani *et al* (2014) correlated perceptual thresholds with a number of stimulus (return configuration, pulse polarity, pulse width, inter-phase gap, pulse rate), electrode (area and number of ganged electrodes), and clinical (retinal thickness, electrode-retina distance) parameters in three recipients of a suprachoroidal retinal prosthesis (Bionic Vision Australia). In addition to electrode-retina distance, the study identified the electrode configuration as important (lowest thresholds were achieved with a monopolar return, anodic-first stimulus polarity, short pulse widths with long inter-phase gaps, and high stimulation rates).

In summary, all three studies identified the distance of electrodes from the retinal surface ('electrode-retina distance') as a critical factor, with electrode size and retinal thickness having only a negligible effect on thresholds. However, these studies were either focused on a small number of subjects (de Balthasar *et al* 2008, Shivdasani *et al* 2014) or were limited to predicting only the mean threshold across electrodes from a small number of factors (Ahuja *et al* 2013). A cross-validated predictive model is still lacking.

It is worth noting that some of these parameters are more easily collected than others. For example, retinal thickness can only be inferred from optical coherence tomography (OCT) images, which is 1) difficult to collect as most retinal implant recipients present with nystagmus, and 2) error-prone due to electrodes casting shadows on the b-scan (Ahuja *et al* 2013). It is therefore paramount to know which of these parameters are worth collecting for the purpose of threshold prediction.

3. Methods

3.1. Dataset

The models and analyses presented in this work are based on data collected from 13 Argus II patients over a period of 11 years (2007–2018) (Hu and

Table 1. Summary of the Argus II dataset. *Clean data: Electrode Deactivation* refers to a processed version of the dataset which excluded trials with missing values or invalid impedance readings. *Clean data: Threshold Prediction* refers to a processed version of the dataset which excluded trials associated with deactivated electrode, trials with missing values or invalid impedance reading, and trials with extreme, outlier thresholds. Data cleaning for electrode deactivation and threshold prediction are further detailed in section 3.5. SF: system fitting, LT: life time. © [2021] IEEE. Adapted, with permission, from Hu and Beyeler (2021).

	Raw data			Clean data: Electrode Deactivation					Clean data: Threshold Prediction		
							Deactivated electrodes				
Subjects	Data points	Sessions	Measured electrodes	Data points	Sessions	Measured electrodes	SF	LT	Data points	Sessions	Measured electrodes
12-001	892	44	51	865	43	51	8	39	685	42	45
12-004	369	33	49	369	33	49	15	45	139	27	31
12-005	968	32	56	885	30	56	1	6	863	30	56
14-001	308	16	48	292	14	48	5	43	158	13	28
17-002	418	34	52	398	32	52	2	51	210	27	39
51-001	323	22	54	309	21	54	2	48	157	17	28
51-003	299	15	56	287	14	56	39	54	89	13	20
51-009	391	12	54	391	12	54	1	7	381	12	54
52-001	665	24	60	665	24	60	0	0	661	24	60
52-003	490	19	55	490	19	55	12	53	293	17	43
61-001	84	9	28	—	—	—	—	—	—	—	—
61-004	426	21	59	426	21	59	0	56	220	19	52
71-002	592	72	51	586	71	51	0	41	407	64	43
Total	**6225**	**353**	**673**	**5963**	**334**	**645**	**85**	**443**	**4263**	**305**	**499**

Beyeler 2021), a subset of which was acquired from the Argus II Feasibility Protocol (Clinical Trial ID: NCT00407602). This longitudinal dataset consisted of 6225 perceptual threshold and electrode impedance measurements acquired from 7 implant centers throughout the United States, the United Kingdom, Switzerland, and France (table 1; for demographics see appendix table A.1 and Ahuja *et al* 2013, Dorn *et al* 2013, da Cruz *et al* 2013). This study was deemed exempt from institutional review board (IRB) approval by the IRB at the University of California, Santa Barbara.

Data cleaning and preprocessing was performed in accordance with the two tasks central to our analyses (section 3.3) and is further detailed in section 3.5. Following data cleaning, reliable data from Subject 61-001 was extremely limited. Only 83 samples across 28 electrodes were available for electrode deactivation prediction (15 of these electrodes were deactivated throughout the entirety of the study) and 36 samples from 13 electrodes remained for threshold prediction. Given that clean data from this subject was limited and unlikely to reflect activity across the entire electrode grid, we omitted this subject's data from further analyses. For each of the remaining 12 subjects, each of the 60 electrodes was measured on 12–72 occurrences throughout the course of the study, unless the electrode had been deactivated. Electrodes that failed to elicit precepts below the charge density limit or that had impedance readings indicating a faulty circuit were deemed nonfunctional and deactivated by the manufacturer (Hu and Beyeler 2021). Electrodes rarely needed to be deactivated at system fitting (labeled 'SF' in table 1, *Clean data: Electrode Deactivation* column), however, temporary electrode deactivation was common over the lifetime of the device (labeled 'LT').

3.2. Feature engineering

Following the work of Hu and Beyeler (2021), raw threshold and impedance measurements acquired from the Argus II Feasibility Protocol were supplemented with clinical data acquired from literature and engineered features to establish the datasets used in our ML modeling and analyses. A population-level correlation matrix for this feature set is presented in figure 1. Each feature is additionally described in table 2 and detailed below. These measurements are acquired with varying degree of difficulty and assume various degrees of prior knowledge. For this reason, we followed Hu and Beyeler (2021) to split this feature set into three subsets.

3.2.1. Routinely collected data

Information regarding patient history, demographics, and implantation (i.e. 'SubjectAge', 'SubjectTimeBlind', 'SubjectAgeAtDiagnosis', and 'SubjectAgeAtSurgery') was reported in previous studies (Ahuja *et al* 2013, Dorn *et al* 2013, da Cruz *et al* 2013) or directly from Second Sight. In a few instances, feature values were unavailable from these sources and had to be interpolated from known data (Hu and Beyeler 2021). 'SubjectAge' was computed with the precision of one day and recorded at each clinical visit to retain short-term relationships between age and perceptual sensitivity.

Implant placement and optic disc location were estimated using retinal fundus images obtained either

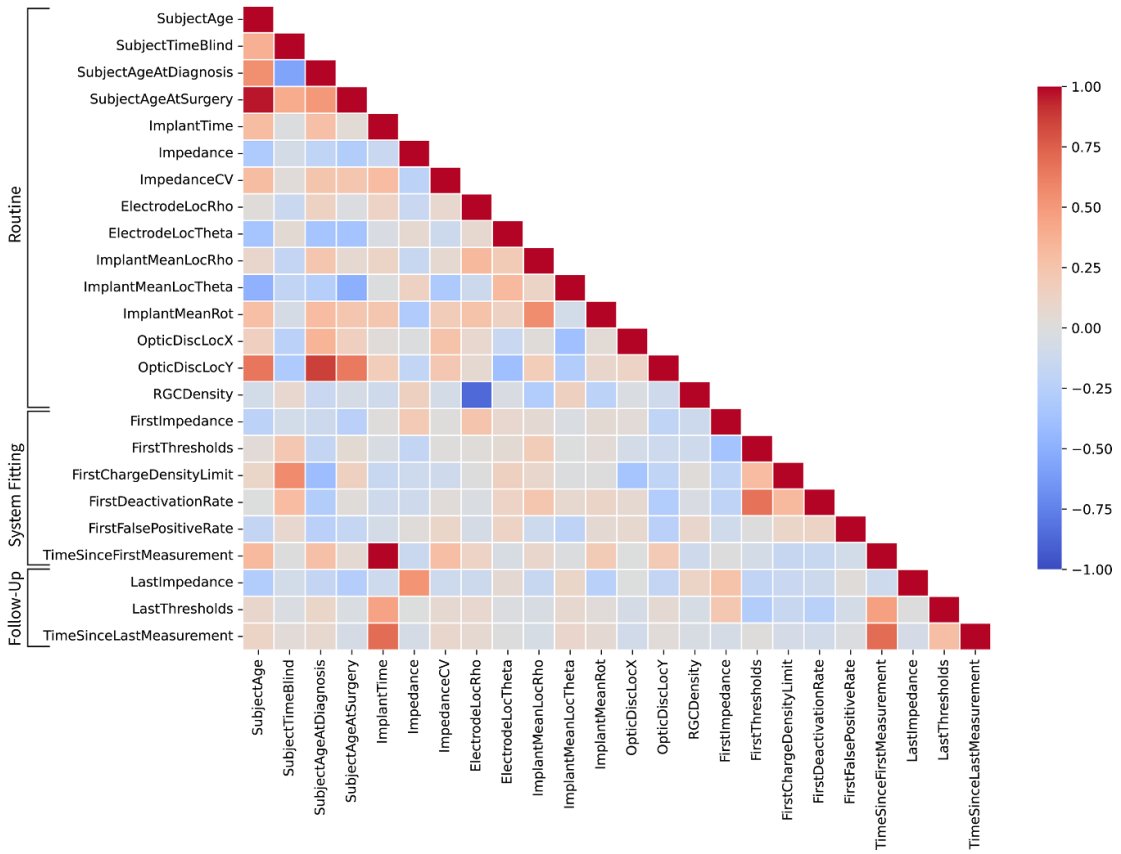


Figure 1. Feature correlation heat map. Feature data from all subjects were aggregated prior to computing each correlation. Each heat map cell depicts the Pearson correlation coefficient between a pair of features. © [2021] IEEE. Adapted, with permission, from Hu and Beyeler (2021).

12 months, 24 months, or 36 months after surgery. Since we did not have access to fundus images from each session, we assumed that the implant did not move over time (Ghani *et al* 2022). We used a procedure described in Beyeler *et al* (2019) and the pulse2percept software (Beyeler *et al* 2017a) to perform image registration and extract the location of the implant and each electrode with respect to the fovea (figure 2). With these locations, we estimated retinal ganglion cell (RGC) density using a previously established method (Curcio and Allen 1990). We note, however, that this method was based on empirical data from healthy retinas, and therefore presumed that this predictor would only be a proxy estimation for true RGC density.

Additional predictors that were available but omitted from this work include a categorical variable specifying the clinic at which the Argus II implant operation was performed for each subject, a binary variable specifying whether the device was implanted in the subject's left or right eye, and the sex of the subject. Although visual outcomes may also depend on surgical precision (as complications during implantation could exacerbate fibrosis), predictors relevant to the surgical center were not considered in our study in an effort to focus our analyses on factors relevant to the subject and implant.

3.2.2. System fitting

Device parameters established during system fitting constituted the second data subset analyzed in this work. These parameters included operational bounds set for the device (i.e. charge densities) and calibration parameters for each of the 60 electrodes of the device, including perceptual thresholds and impedance recordings. It is during this process that the manufacturer would deactivate nonfunctional electrodes (as judged by impedance values). Perceptual false positive rates (the frequency that the subject reported seeing a phosphene while no stimulation was being delivered) were additionally recorded at this time (Hu and Beyeler 2021). As perceptual thresholds can fluctuate substantially over time, we also computed the amount of time elapsed since system fitting for each subsequent visit. Features derived from measurements obtained during system fitting (table 2, middle section) were only used as features, never as labels that the algorithm was supposed to predict.

3.2.3. Follow-up examinations

Given the variability of perceptual thresholds over time, measurements from recent visits are often valuable predictors of future thresholds. This was exemplified in Hu and Beyeler (2021), which demonstrated that threshold and impedance measurements from a

Table 2. Features (measured and engineered) for perceptual sensitivity prediction. Features are binned according to data subset (i.e. Routine, System Fitting, or Follow-Up). © [2021] IEEE. Adapted, with permission, from Hu and Beyeler (2021).

	Feature	Description	Precision
Routine	SubjectAge	Subject age	days
	SubjectTimeBlind	Estimated time since onset of blindness	days
	SubjectAgeAtDiagnosis	Subject age at first retinitis pigmentosa (RP) diagnosis	years
	SubjectAgeAtSurgery	Subject age at time of implant surgery	years
	ImplantTime	Time since implant surgery	days
	Impedance	Manufacturer-provided impedance reading at each electrode	kΩ
	ImpedanceCV	Coefficient of variation for impedance values	float
	ElectrodeLocRho	Location of electrode with respect to fovea (distance component of polar coordinate)	μm
	ElectrodeLocTheta	Location of electrode with respect to fovea (angular component of polar coordinate)	rad
	ImplantMeanLocRho	Mean location of electrode with respect to fovea (distance component of polar coordinate)	μm
	ImplantMeanLocTheta	Mean location of electrode with respect to fovea (angular component of polar coordinate)	rad
	ImplantMeanRot	Implant angle	rad
	OpticDiscLocX	Optic disc location (horizontal component)	μm
	OpticDiscLocY	Optic disc location (vertical component)	μm
	RGCDensity	Retinal ganglion cell (RGC) density in a healthy retina (Curcio and Allen 1990)	RGC deg^{-2}
System Fitting	FirstImpedance	Impedance reading at system fitting for each electrode	kΩ
	FirstThreshold	Perceptual stimulus threshold at system fitting for each electrode	μA
	FirstChargeDensityLimit	Charge density limit at time of system fitting	mCcm^{-1}
	FirstDeactivationRate	Fraction of electrodes deactivivated at system fitting	float
	FirstFalsePositiveRate	False positive phosphene perception rate during system fitting	float
	TimeSinceFirstMeasurement	Elapsed time since system fitting	days
Follow-Up	LastImpedance	Impedance reading from previous session for each electrode	kΩ
	LastThreshold	Perceptual stimulus threshold from previous session for each electrode	μA
	TimeSinceLastMeasurement	Time elapsed previous electrode threshold measurement	days

patient's most recent visit, along with the time elapsed since their last visit, enabled accurate prediction of electrode deactivation. We suspected that these three features would be significant predictors of perceptual thresholds as well and therefore included them in our 'Full' feature subset.

3.3. Prediction tasks

We studied predictors of perceptual sensitivity in the context of two prediction tasks:

- *Threshold prediction:* a regression task in which the goal of the ML model was to predict the measured perceptual threshold on a given electrode at a specific point in time.
- *Electrode deactivation:* a binary classification task in which the goal of the ML model was to correctly predict whether or not an electrode was deactivated during at a specific point in time.

Given the available data, these two tasks enabled us to directly study the impact of each feature on quantitative measures of perceptual sensitivity by means of explainable artificial intelligence (XAI). While we expected that a subset of predictors would be shared across models, we expected the regression models to better reveal predictors that contribute to finer-granularity fluctuations in perceptual sensitivity.

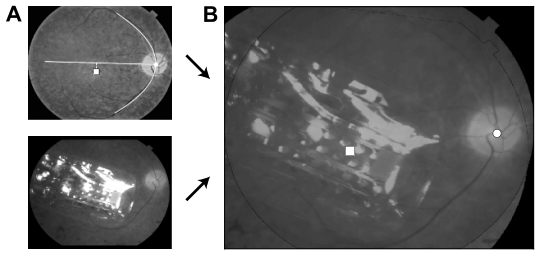


Figure 2. The location and orientation of each subject's implant was estimated by combining their postsurgical fundus photograph (*A*, bottom) with a baseline presurgical image in which the fovea had been identified (*A*, top) to produce a registered image (*B*; □: foveal pit, ○: optic disc). The horizontal raphe (*A*, white line) was approximated by fitting a parabola to the main vascular arcade and finding the tangent to the parabola inflexion point (reproduced from Beyeler *et al* (2019). CC BY 4.0).

3.4. Explainable ML models

As in many neural engineering applications where ML models are used in decision-making processes, it is critical that the predictions made by the model are explainable. Specific to perceptual outcome prediction, we aimed to develop models that can inform clinicians of the most relevant parameters to collect and how such parameters may be used to automate stimulus threshold parameterization. In pursuit of these goals, we considered both linear and nonlinear ML models.

We leveraged logistic regression (LR) with and without L_1 and L_2 loss constraints for electrode deactivation and ordinary least squares (OLS) and elastic net (EN) linear models for threshold regression (in the analyses that follow, the regularized LR model is referred to as LR_{reg} and the non-regularized model as $\text{LR}_{\text{nonreg}}$). Each of these models approximates a target variable (i.e. a perceptual threshold measurement or a binary indicator reflecting electrode deactivation status) as a linear combination of observed clinical measures. The LR classification model applies a final non-linear transformation to this linear function to bound the output range to $[0, 1]$, reflecting the model's confidence that the electrode should be deactivated. We chose these models both for their simplicity and explainability. Furthermore, L_1 regularization (used in a subset of these models) enables greater robustness to correlated features. Meanwhile, in addition to the regularization from any L_1 or L_2 loss penalties, we expected that these low-variance models would elucidate the perceptual outcome predictors that were most generalizable across subjects in our logitudinal dataset.

In addition, we used gradient boosting models (specifically, XGBoost models) and shallow multilayer perceptron (MLP) models in our non-linear modeling analysis. Unlike linear models, whose output is based on a linear, weighted combination of input feature values, XGBoost models are composed of ensembles of decision trees (Chen and Guestrin 2016) defined on multiple subsets of dataset features, allowing them to capture nonlinear patterns in the underlying data. Non-linear modeling is accomplished in MLP models via a composition of linear transformations that are interleaved with non-linear activation functions.

In modeling our data, we aimed to discover the most salient features relevant to perceptual sensitivity, whether these relationships were linear or not. Furthermore, these models were used to establish benchmark results for electrode deactivation and perceptual threshold prediction.

The contributions of features to each model's predictions were evaluated using SHAP (Lundberg and Lee 2017), a post-hoc analysis technique commonly used to compute the relative contributions of a sample's parameter values to its prediction. As SHAP analysis is model-agnostic, it is applicable to linear and non-linear models in both electrode deactivation and threshold prediction tasks. Although these models can be explained through their fitted parameters, SHAP enables a more direct comparison of the predictive behavior between models with varied parameters and assumptions about the underlying data. Additionally, SHAP offers insight into a model's decision at the granularity of a single test sample, regardless of model architecture.

In the analyses that follow, LR, OLS, and EN models were implemented using scikit-learn's LogisticRegression, LinearRegression, and ElasticNet APIs (v1.1.2). XGBoost models were implemented using the XGBClassifier and XGBRegressor APIs of the XGBoost package (v1.6.2). MLP models were built, trained, and evaluated using TensorFlow (v2.11.0). Bayesian hyperparameter optimization was performed using the scikit-optimize package (v0.9.0). Supplementary material and code to run the models and generate the figures can be found at https://github.com/bionicvisionlab/2023-ArgusThresholds.

3.5. Model evaluation and comparison

A significant challenge in developing ML models for biological data is the inherent inter-subject variability of such data. It is not uncommon for a data distribution from one subject to be divergent from the data distribution of another (appendix figure A.1). To estimate the performance of our proposed electrode deactivation and threshold prediction models on data from unobserved subjects, we therefore employed a leave-one-subject-out (LOSO) analysis as follows: for each of the twelve subjects, we instantiated a new model, trained the model on data from the remaining eleven subjects, and generated predictions exclusively for the data of the held-out test-subject. Model training, validation, and testing strategies are further described below and summarized in appendix table A.2.

For regularized linear and gradient boosting models, we performed Bayesian hyperparameter

optimization (Snoek *et al* 2012) in a nested LOSO cross-validation loop to estimate the posterior probability distribution of an objective score (electrode deactivation: F1 score, threshold prediction: R^2 score). This nested LOSO cross-validation loop was executed in a manner similar to the evaluation loop—in each cross-validation fold, data from one of the eleven non-test subjects was withheld for validation testing. The hyperparameters that yielded the highest cross-validation objective score across all validation folds were then selected and the model was re-fit to the training data from all eleven subjects prior to evaluation on the held out test subject. Hyperparameters optimized for the LR_{reg} classification model included 'C' and 'l1_ratio', controlling the total regularization strength and the portion of this regularization contributed by L_1 loss. Similarly, hyperparameters ('alpha' and 'l1_ratio') were optimized for the EN regression model. For both the XGBClassifier (XGB-C) and XGBRegressor (XGB-R) models, we tuned the number of estimators in the ensemble ('n_estimators'), the max depth of each decision tree ('max_depth'), partition criteria 'min_child_weight' and 'gamma', and L_1 and L_2 regularization terms ('reg_alpha' and 'reg_lambda', respectively).

MLP model architectures were established experimentally using validation data. In consideration of computational constraints, the validation data used to establish the MLP model architecture consisted of 20% of each non-test subject's data (i.e. 20% and 80% of the eleven non-test subject data was dedicated to validation and training datasets, respectively, in each LOSO test fold). In both prediction tasks, the MLP featured two hidden layers, each featuring 128 neurons, ReLU non-linearities, post-activation dropout (dropout probability of 0.4) to help mitigate overfitting (Srivastava *et al* 2014), and an output layer composed of a single neuron. Further increasing network depth or width tended to result in overfitting, as evaluated on the validation dataset. For models tasked with predicting electrode deactivation, a sigmoid activation function was applied to the activity of the output neuron to bound its range to $[0, 1]$.

We evaluated model performance when trained with the three subsets of predictors described above (table 2): the 'Routine' dataset, a set composed of the 15 predictors derived from routinely collected clinical data (i.e. the Routine features of table 2); the 'Routine+Fitting' dataset, a subset which contained the predictors of the 'Routine' subset as well as the 6 predictors derived from measurements obtained during system fitting (System Fitting features of table 2); and the 'Full' dataset, which contained all 21 predictors from the 'Routine+Fitting' dataset in addition to the 3 follow-up trial predictors (predictors labeled as Follow-Up features in table 2). For 'Routine+Fitting' and 'Full' datasets, the first measurement from each electrode was removed from the dataset to prevent leaking ground-truth data into the feature vectors.

3.5.1. Threshold prediction

During threshold regression, models inferred a real-valued perceptual threshold for each electrode. These thresholds were recorded following typical clinical procedures and were established based on a subjects' ability to perceive a phosphene at a given stimulation current. In this task, recordings associated with electrodes that were deactivated in the session were removed, as no sensible threshold current could be assigned to the deactivated electrode in this case. Trials with missing data or impedance readings of 0 kΩ were removed. Furthermore, as perceptual threshold estimation is a high-variance procedure, it was not uncommon to observe within-session variability of the threshold for a given electrode or for subjects to report the presence of a phosphene in the absence of stimulation ('catch trials'). To account for such sources of noise, we chose to discard outlier samples with threshold values so far from the subject's perceptual threshold data distribution that they were most likely associated with erroneous measurements. Outliers were discarded using an automated, statistical method based on Chebyshev's data distribution tail bounds (Amidan *et al* 2005). A total of 63 samples were removed from the 'Routine' dataset (1.4% of total dataset) and 60 samples were removed from both the 'Routine+Fitting' and 'Full' datasets (1.6%) in this process. Outlier removal was exclusively performed before model fitting and evaluation (i.e. following the outlier removal process, all model predictions were used in our evaluations and analyses). Following outlier rejection, feature values were normalized according to feature value distributions observed in each training split of LOSO cross validation. Model performance was quantitatively analyzed with the adjusted coefficient of determination (R^2_{adj}) and a variant of the fraction of explainable variance explained (FEVE) (Willeke *et al* 2022).

FEVE $\in (-\infty, 1]$ offers a quantitative measure of explainable variance, like R^2, while also accounting for variability in measurements of the dependent variable of interest (i.e. perceptual threshold) influenced by uncontrolled factors during measurement (e.g. perceptual lapses and false positive perceptions). FEVE was computed as follows:

$$\text{FEVE} = 1 - \frac{\frac{1}{N}\sum_{ij}\left(r_{ij} - \hat{r}_i\right)^2 - \sigma_i^2}{\text{Var}\left[\mathbf{r}\right] - \sigma_i^2}, \qquad (1)$$

where r_{ij} was the jth ground-truth perceptual threshold measurement for electrode i in a single recording session, $\hat{r}_i$ was the predicted perceptual threshold for this electrode during the session, N was the total number of perceptual threshold observations, $\text{Var}[\mathbf{r}]$ was the variance of all perceptual threshold measurements $\mathbf{r}$, and $\sigma_i^2 = \mathbb{E}_i\left[\text{Var}_j[r_{ij}]\right]$ was the expected variance in perceptual threshold measurements for stimuli presented at each electrode of each subject. Given that it was uncommon for

repeated threshold measurements to be made for the same electrode on any given day, Var[**r**] and σ_i^2 were estimated over the entire dataset (that is, they were not computed on a per-subject basis).

We also observed diverging perceptual threshold distributions between subjects (appendix figure A.1). To account for these differences in the 'Routine+Fitting' and 'Full' data subsets, we scaled each threshold measurement according to the first threshold measured for the electrodes of each subject instead of directly predicting perceptual thresholds. This transformation implies that the models fitted to 'Routine+Fitting' and 'Full' feature sets learned to predict *changes* in perceptual thresholds, relative to system fitting measurements. We found that this enabled better model generalization in LOSO threshold prediction.

3.5.2. Electrode deactivation

For the task of electrode deactivation, each electrode was assigned class 1 if it was deactivated in the given recording session and 0 otherwise. The exact time of when an electrode started meeting the criteria for deactivation is unknown, but the date at which it was measured and the decision to deactivate the electrode was made is known for each electrode. Trials with missing values or invalid impedance readings (i.e. 0 kΩ) were removed from the dataset prior to model fitting and analysis.

In each LOSO iteration, the synthetic minority oversampling technique (Chawla *et al* 2002) was applied to the training dataset to balance the number of samples from each class. All feature values were normalized according to the distribution of the training data to have zero mean and unit standard deviation.

4. Results

4.1. Factors affecting perceptual sensitivity

Figure 3 shows the absolute Pearson correlation coefficient for the 24 predictors, organized by feature class ('Routine', 'Routine+Fitting', or 'Full') and ordered by correlation magnitude. Aside from historical threshold measurements (i.e. 'LastThresholds' and 'FirstThresholds'), measures of electrode deactivation rate and charge density limits established at system fitting were among the predictors that had highest correlation with perceptual sensitivity. In terms of demographic factors, time since blindness onset ('SubjectTimeBlind') and time since implantation ('ImplantTime') had the highest correlations.

Figure 4 shows perceptual thresholds plotted against each of the 24 predictors. Immediately observable is the high degree of variability among threshold measurements for any given predictor and the linear correlation between many of these predictors and perceptual sensitivity. When fitting linear regression coefficients between these predictors and their accompanying threshold measurements, the line of best fit for all but two predictors (labeled 'ImplantMeanLocTheta' and 'FirstFalsePositiveRate' in figure 4, respectively) had a significantly non-zero slope ($p < 0.05$). However, such trends were not always consistent across all 12 subjects.

Whereas previous work has demonstrated that thresholds are negatively correlated with impedance readings (labeled 'Impedance' in figure 4; de Balthasar *et al* 2008), our data also highlights correlations with demographic factors. Most notably, thresholds tended to increase with subject age ('SubjectAge'), subject age at implantation ('SubjectAgeAtSurgery'), and time since blindness onset ('SubjectTimeBlind'). Not surprisingly, thresholds also tended to increase with time since implantation ('ImplantTime'), which is consistent with other studies (Yue *et al* 2015). In terms of neuroanatomical parameters, thresholds were positively correlated with electrode-fovea distance ('ElectrodeLocRho') and negatively correlated with proxy estimates of ganglion cell density ('RGCDensity'), which is a nonlinear function of electrode-fovea distance.

Thresholds over time were also strongly correlated with different measures typically obtained during system fitting, such as impedance and threshold readings ('FirstImpedance' and 'FirstThreshold', respectively). Correlations were similar for predictors obtained during follow-up exams.

4.2. Threshold prediction

Table 3 shows aggregated LOSO threshold prediction results observed when modeling 'Routine', 'Routine+Fitting', and 'Full' datasets with OLS, EN, XGB-R, and MLP models. Note that in some cases FEVE values were even more negative than R^2_{adj}, because of the subtraction of σ_i^2 in equation (1). Perceptual threshold estimates compared to ground-truth for all 12 subjects can be found in appendix figures B.1 and B.2.

All models failed to yield accurate perceptual threshold predictions when relying solely on routinely collected data, as indicated by negative R^2_{adj} and FEVE values. Although many of these routine predictors correlated with perceptual thresholds, the regression results suggest that they alone do not carry sufficient information to predict perceptual thresholds over time.

Upon the introduction of measurements recorded during each subject's system fitting session, the predictive power of the XGB-R increased notably ($R^2_{\text{adj}} = 0.327$), more so than the linear model ($R^2_{\text{adj}} = 0.135$). Finally, when considering all collected and engineered predictors in our models (i.e. the 'Full' feature subset), both EN and XGB-R models achieved accurate threshold predictions, explaining more than 70% of the variance in the data.

The improvements in explained variance observed when including 'System Fitting' and

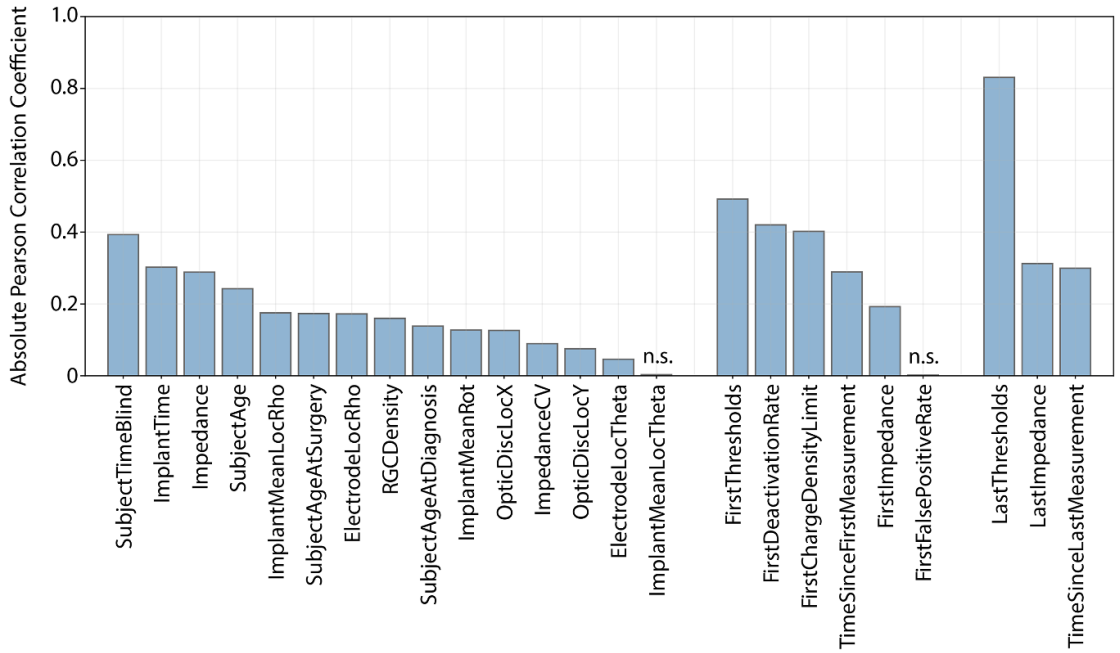


Figure 3. Magnitude of population-level Pearson correlation coefficient between each predictor and perceptual thresholds, grouped by feature set (left: 'Routine', middle: 'Routine+Fitting', right: 'Full') and sorted by correlation magnitude. All predictors had $p < .05$ except for 'ImplantMeanLocTheta' and 'FirstFalsePositiveRate' (labeled n.s.).

'Follow-Up' features can be predominantly attributed to these additional features themselves. A XGB-R model fitted exclusively to the six 'System Fitting' features was observed to explain perceptual threshold variance with an R^2_{adj} value of 0.382. Analogously, EN and XGB-R models fitted to a dataset consisting only of the three 'Follow-Up' features were capable of predicting perceptual thresholds with R^2_{adj} values of 0.761 and 0.697, respectively. Despite these comparable predictive performances, we suggest that including routinely collected clinical features in the 'Routine+Fitting' and 'Full' datasets remains valuable for post-hoc analyses, such as those that follow. Specifically, these analyses enable reasoning about the contribution of routine clinical features to model predictions alongside the 'System Fitting' and 'Follow-Up' features (figure 5) and inform us of dependencies between all features in the context of an accurate predictive model (figure 6).

Post-hoc SHAP analysis (figure 5, top two rows) was performed for the top performing linear and non-linear models (EN and XGB-R) and revealed that among all 'Routine+Fitting' features, the XGB-R predictions were most influenced by initial threshold measurements from system fitting, the time elapsed since implant surgery, and the time elapsed since system fitting ('FirstThresholds', 'ImplantTime', and 'TimeSinceFirstMeasurement', respectively). In these plots, each data point is associated with a threshold prediction from the held-out cross-validation fold (test set). SHAP values indicate each feature's contribution to the model's prediction, with high SHAP values pushing the model towards predicting high thresholds, and low SHAP values pushing the model towards predicting low thresholds. As previously observed in de Balthasar *et al* (2008), measures of impedance were also instrumental in this model's decision process.

The two models, given their inherently different assumptions about the relationships between the predictors and perceptual thresholds, yield different insights into the most impacting predictors of perceptual threshold. The most important predictor for both model types of the 'Full' dataset was the threshold measurement from the previous visit ('LastThresholds'), which is not surprising considering its strong correlation with current thresholds. Greater values for 'LastThresholds' tended to influence the model towards predicting a larger perceptual threshold sensitivity. Additional predictors shared between EN and XGB-R models included 'FirstThresholds', 'FirstImpedance', 'TimeSinceLastMeasurement', 'Impedance', and 'ImpedanceCV'. Of the top five predictors influencing the predictions of the EN model, the remaining four were 'TimeSinceLastMeasurement', 'FirstImpedance', 'FirstThresholds', and 'OpticDiscLocY'. More time elapsed since a subject's previous visit and lower impedance readings often biased the model towards predicting an increased perceptual threshold. Interestingly, a high initial threshold (measured during system fitting) led to decreased threshold predictions over time. Exclusively meaningful to the XGB-R model, advanced age (accounted for in the predictor 'SubjectAge') often led to higher threshold predictions in the XGB-R model.

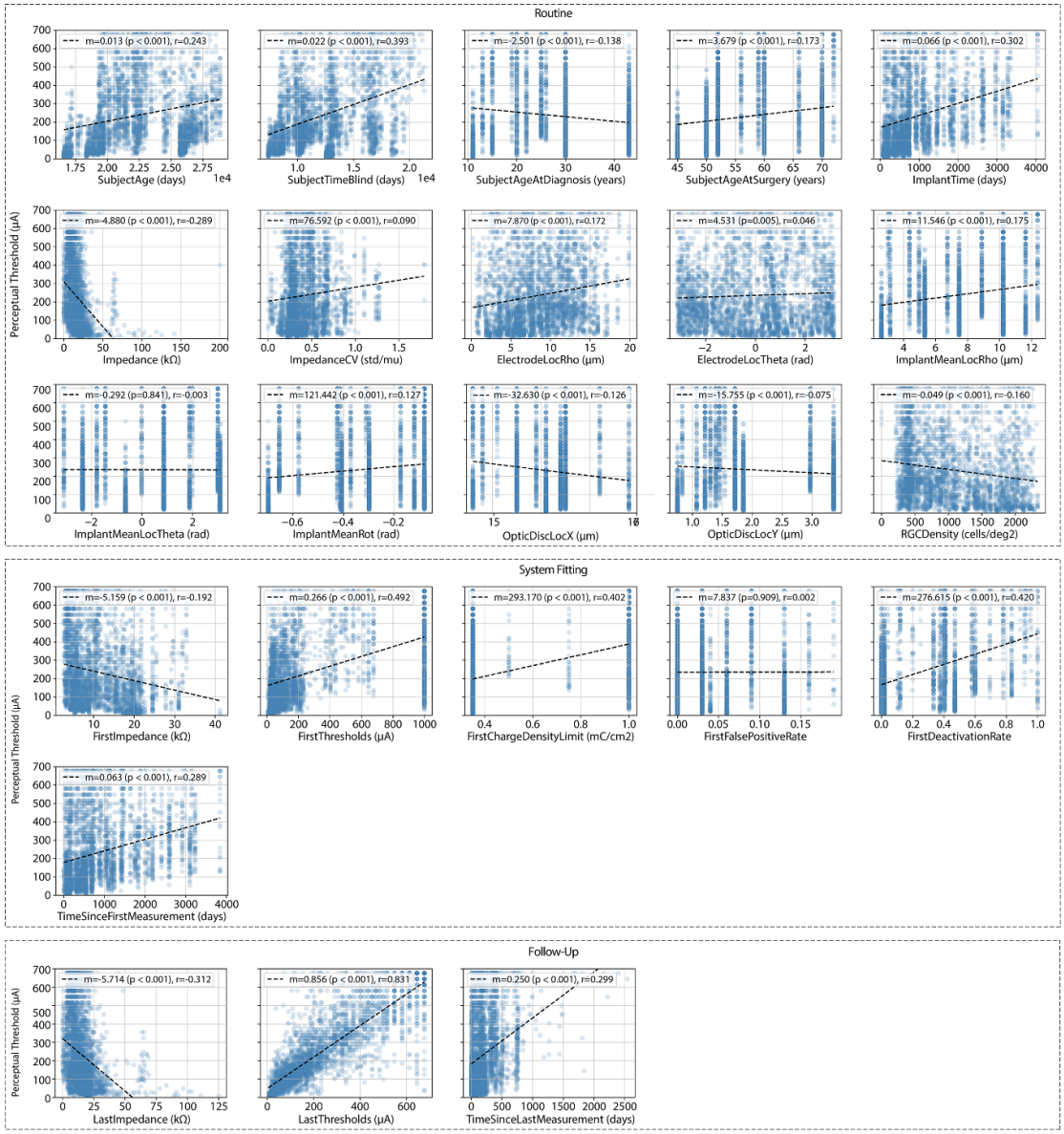


Figure 4. Predictor measurements plotted against perceptual stimulation thresholds and the associated linear-least squares regression line. The corresponding slopes (m), p values (associated with a null hypothesis that the slope of the linear regression line is zero), and Pearson Correlation Coefficients (r) are provided in the legend of each plot. Subplots are grouped by feature set ('Routine', 'Routine+Fitting', or 'Full') and ordered by their listing in table 2. While certain predictors may remain constant over the lifespan of the device for a given participant (e.g. SubjectAgeAtDiagnosis) as perceptual thresholds vary, population-level trends may emerge that are important for the generalization of predictive models to data from patients outside of the training dataset. All computed values are rounded to the nearest thousandth.

Further SHAP analyses reveal interactions between predictors and the dependence of model predictions on such interactions. Interestingly, we found that the importance of impedance for the purpose of threshold prediction was age-dependent (figure 6): whereas high impedances tended to be associated with lower perceptual threshold predictions for the youngest subjects, the opposite was true for the oldest subjects in the dataset. For a wide range of subject ages (between −1.5 and +2 in normalized age), electrode impedance was not predictive of thresholds.

Despite the improved generalization achieved by prediction of a scaled threshold value (as opposed to the exact perceptual threshold current; see section 3.5) in the case of modeling with 'Routine+Fitting' and 'Full' features, impractical predictions were observed in rare occasions. Notably, an unbounded regression model output permitted the prediction of a negative perceptual threshold, but this was observed on a maximum of 10 occassions (evaluated over all EN and XGB-R models).

4.3. Predicting electrode deactivation

Table 4 shows aggregated LOSO classification results observed when modeling 'Routine', 'Routine+Fitting', and 'Full' datasets with each evaluated model. The results presented in this subsection

Table 3. Leave-one-subject-out perceptual threshold regression results. Each metric is evaluated over an aggregated test set (metrics reported as mean ± standard deviation over three randomly initialized and optimized models, with the exception of the OLS regression model, which had a non-stochastic fitting procedure). Per-subject metric means and standard deviations for the median-performing model (as evaluated by R^2_{adj}) of three trials are reported in parentheses. All computed values are rounded to the nearest thousandth. R^2_{adj}: Adjusted coefficient of determination. FEVE: fraction of explainable variance explained.

	Method	R^2_{adj}	FEVE
Routine	OLS	−7.600 (−14.814 ± 16.905)	−7.951 (−15.463 ± 19.491)
	EN	−1.422 ± .001 (−2.574 ± 3.431)	−1.485 ± .001 (−2.504 ± 3.510)
	XGB-R	**−.201** ± .036 (−1.383 ± 1.807)	**−.207** ± .038 (−1.321 ± 1.881)
	MLP	−3.031 ± .396 (−4.232 ± 7.330)	−3.168 ± .415 (−4.203 ± 7.456)
Rout.+Fit.	OLS	−15.888 (−192.366 ± 360.313)	−15.901 (−235.667 ± 478.153)
	EN	.135 ± .000 (−.651 ± 1.478)	.141 ± .000 (−.769 ± 2.206)
	XGB-R	**.327** ± .068 (−.100 ± .369)	**.333** ± .068 (−.025 ± .369)
	MLP	.198 ± .032 (−.181 ± .408)	.204 ± .032 (−.298 ± 1.436)
Full	OLS	.621 (−.144 ± 1.213)	.628 (−.199 ± 1.588)
	EN	**.763** ± .000 (.373 ± .445)	**.770** ± .000 (.418 ± .482)
	XGB-R	.716 ± .029 (.408 ± .318)	.723 ± .029 (.476 ± .284)
	MLP	.661 ± .014 (.215 ± .501)	.667 ± .014 (.270 ± .560)

Note: Metrics of the top-performing model for each dataset are bolded.

include updates and expansions to those presented in Hu and Beyeler (2021).

When predicting electrode deactivation using only 'Routine' measures, subject age, time since blindness onset, subject age at diagnosis, and measures of neuroanatomical and device landmarks ('OpticDiscY', 'ElectrodeLocRho', and 'ImplantMeanRot') were important predictors common to all evaluated models. No model was able to predict electrode deactivation with high fidelity, however, when solely relying on these features. Introducing one or more previous threshold measurements enabled these models to predict electrode deactivation much more reliably. Predictive performance for the LR_{reg} and XGB-C models increased area under the ROC curve (AUC) values above 0.8 when system fitting features were included and peak AUC values of 0.911 and 0.909 were observed for the LR_{reg} and XGB-C models when predicting electrode deactivation using recent, historical measurements in follow-up examinations. The ten most impactful system fitting and follow-up features are shown in figure 5(bottom two rows). The demographic predictor that generalized to each of these data subsets and model types was subject age at diagnosis. This measure stands out as particularly noteworthy given its impact on model prediction and ease of acquisition.

Initial threshold measurements and system fitting parameters (e.g. initial thresholds, proportion of deactivated electrodes) yielded significant improvements in predictive performance. Electrodes with greater initial threshold and higher electrode-fovea distance ('ElectrodeLocRho') were more likely to be deactivated in the future. Similarly, higher proportions of deactivated electrodes during system fitting pushed the model towards predicting deactivation.

In line with earlier work (Hu and Beyeler 2021), model predictions were most influenced by recent threshold measurements in 'Full' experiments. Here, the models clearly learned the strong correlation between recent threshold measurements and perceptual sensitivity. While measurements collected during system fitting may be helpful to provide baseline estimates, the large fluctuations of perceptual thresholds over time (Yue *et al* 2015) limit the long-term usability of these initial measurements. Nonetheless, routinely collected measures still remained important to the predictions of follow-up models.

5. Discussion

The present study is a retrospective investigation of a large clinical dataset and demonstrates the untapped value in clinical recordings taken from neuroprostheses. In this work, we demonstrate the prediction of perceptual thresholds and electrode deactivation using XAI models and the insights into measurable factors influencing perceptual sensitivity that can be gleaned from these models. Automating threshold prediction using imaging and clinical data may be an important and cost-effective strategy for retinal implant calibration.

On a longitudinal dataset composed of data from 12 subjects with Argus II retinal prostheses, electrode deactivations were predicted with AUC values from 0.520 when exclusively using routine clinical measures up to 0.831 when incorporating system fitting data and 0.911 when leveraging information from previous examinations. Additionally, perceptual thresholds were predicted using routine, system fitting, and follow-up measurements, with associated R^2_{adj} values of up to 0.763. An implication of these results is that current devices require consistent monitoring to enable long-term device efficacy and optimal predictive performance. As this

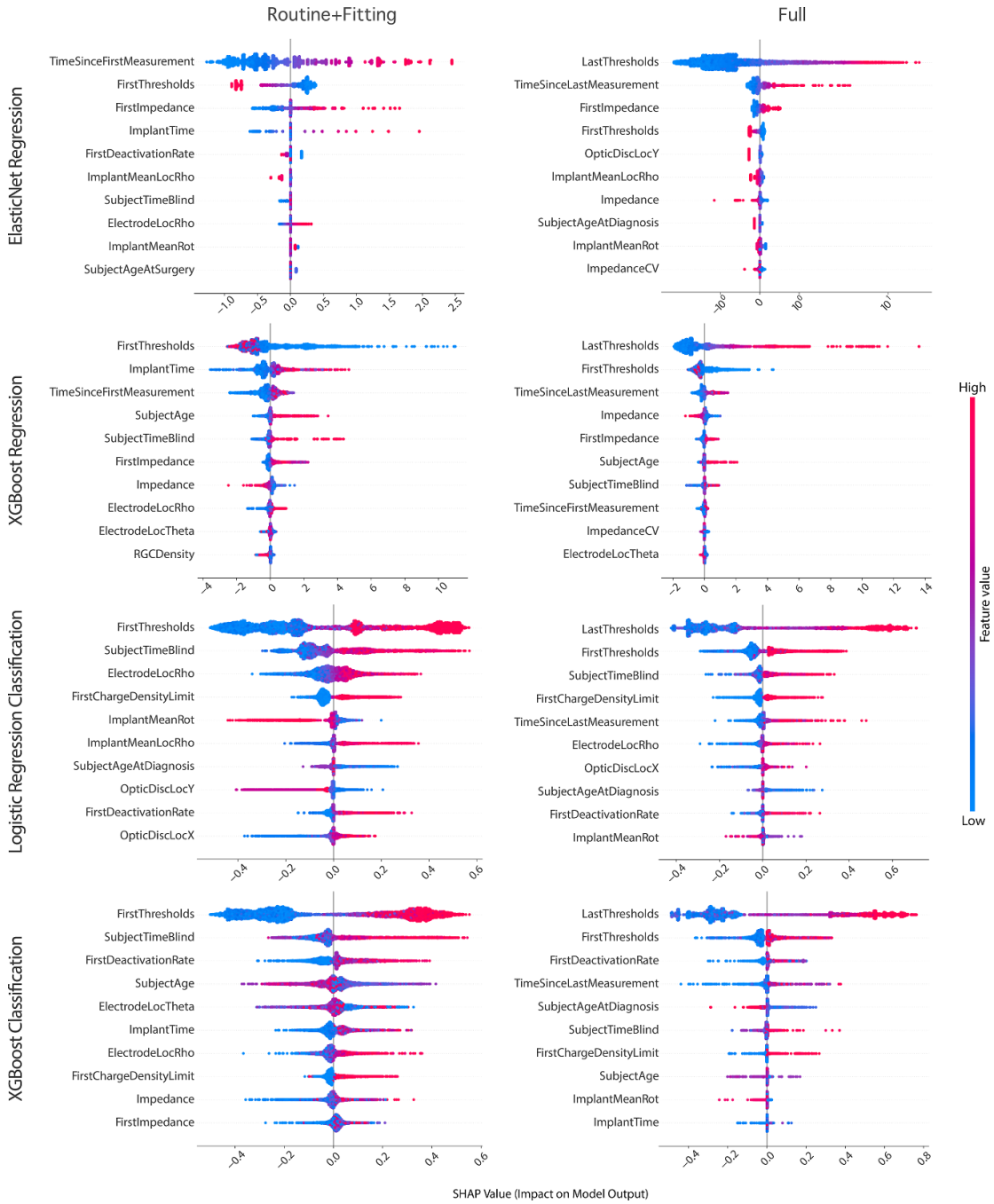


Figure 5. SHapley Additive exPlanations (SHAP) force plots for the ten predictors with greatest contribution to threshold prediction (top two rows) and electrode deactivation inference (bottom two rows). All depicted data points are associated with held-out test set predictions of the median performing model of three randomly initialized models. SHAP values indicate each feature's contribution to the model's output. For threshold prediction, predictors with high SHAP values influenced the model towards predicting greater threshold values while those smaller SHAP values encouraged the model to predict relatively lower threshold values. In the case of electrode deactivation, features with positive SHAP values influenced the model towards predicting deactivation while features with negative values contributed against predicting deactivation. Results are shown for 'Routine+Fitting' and 'Full' feature splits (see table 2). © [2021] IEEE. Adapted, with permission, from Hu and Beyeler (2021).

would become impractical at scale, however, perceptual thresholds could be reasonably predicted with XAI-based models that solely rely on routine clinical measures combined with data from system fitting. Alternatively, the prediction approaches discussed in this paper may be leveraged to reduce the frequency that electrode thresholds must be manually established. Recent measurements could presumably be used as features to adjust threshold parameters over short timescales between clinical visits. As perceptual thresholds can vary over short periods of time, this type of automated calibration process may extend the usability of the device during these periods of fluctuation. Although the results presented in this study were based on measurements exclusively from the Argus II retinal implant, the predictors that we

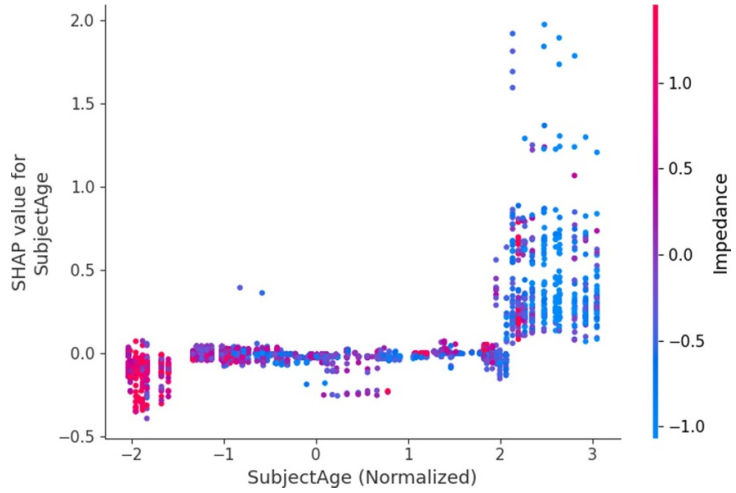


Figure 6. SHAP dependencies between subject age ('SubjectAge') and impedance observed in the full XGB-R model ('Full'). Positive SHAP values indicate that the model was pushed towards predicting higher thresholds. SHAP magnitude indicates the strength of the effect. For instance, the data points on the far left of the plot (i.e. the youngest subjects in the dataset) indicate that high impedances (magenta) were correlated with lower perceptual threshold predictions (negative SHAP values). For subjects above normalized age +2, low electrode impedance was strongly correlated with higher perceptual threshold predictions (large positive SHAP values).

Table 4. Results for leave-one-subject-out (LOSO) electrode deactivation classification. Each metric is evaluated over an aggregated held-out test set. Means and standard deviations across subjects are reported in parentheses. Subjects with no deactivated electrodes were excluded from the mean and standard deviation aggregation reported in parentheses. All computed values are rounded to the nearest thousandth.

	Method	Precision	Recall	F1	AUC
Routine	LR_{nonreg}	.232 ± .003 (.211 ± .248)	.390 ± .009 (.517 ± .447)	.291 ± .005 (.246 ± .283)	.449 ± .002 (.583 ± .153)
	LR_{reg}	.240 ± .004 (.218 ± .257)	.409 ± .011 (.531 ± .444)	.303 ± .006 (.257 ± .291)	.455 ± .001 (.577 ± .164)
	XGB-C	.255 ± .011 (.382 ± .331)	**.438** ± .015 (.547 ± .361)	.323 ± .013 (.346 ± .281)	.519 ± .017 (.646 ± .147)
	MLP	**.266** ± .018 (.317 ± .291)	.411 ± .018 (.489 ± .343)	**.323** ± .019 (.315 ± .279)	**.520** ± .003 (.590 ± .132)
Rout.+Fit.	LR_{nonreg}	.511 ± .010 (.408 ± .277)	.702 ± .023 (.579 ± .414)	.591 ± .002 (.451 ± .311)	.791 ± .003 (.736 ± .124)
	LR_{reg}	.532 ± .007 (.420 ± .290)	.754 ± .007 (.614 ± .410)	**.624** ± .007 (.479 ± .313)	.821 ± .001 (.737 ± .126)
	XGB-C	.515 ± .014 (.466 ± .250)	**.780** ± .045 (.671 ± .341)	.620 ± .023 (.540 ± .276)	**.831** ± .017 (.675 ± .147)
	MLP	**.575** ± .037 (.483 ± .261)	.570 ± .030 (.454 ± .330)	.572 ± .029 (.439 ± .286)	.782 ± .016 (.703 ± .102)
Full	LR_{nonreg}	.616 ± .004 (.545 ± .278)	.778 ± .005 (.638 ± .398)	.687 ± .004 (.539 ± .327)	.870 ± .004 (.810 ± .115)
	LR_{reg}	.645 ± .002 (.541 ± .270)	.843 ± .002 (.692 ± .351)	.731 ± .001 (.599 ± .294)	**.911** ± .001 (.814 ± .134)
	XGB-C	.636 ± .010 (.521 ± .262)	**.862** ± .006 (.715 ± .348)	**.732** ± .005 (.598 ± .292)	.909 ± .002 (.781 ± .163)
	MLP	**.669** ± .005 (.521 ± .289)	.639 ± .048 (.554 ± .362)	.653 ± .024 (.516 ± .329)	.834 ± .016 (.753 ± .106)

Note: Metrics of the top-performing model for each dataset are bolded.

analyzed are likely highly relevant to a wider range of epiretinal prostheses. Further, while a subset of the derived features and conclusions proposed in this work (e.g. the relationship between thresholds and electrode-retina distance) are not applicable to non-epiretinal implants, the methodologies used could easily be extended to other devices.

Post-hoc SHAP analysis revealed the contribution of each clinical measure across multiple model architectures and data subsets. In this analysis, electrode impedance was observed as an important predictor of perceptual sensitivity. Negative correlations between impedance and perceptual threshold measurements have previously been observed in de Balthasar *et al* (2008), wherein the authors suggest that this correlation may stem from the relationship between impedance and electrode-retina distance. In addition, our models discovered correlations with demographic factors, demonstrating that thresholds tend to increase with subject age, time since blindness onset, and time since implantation, which we hypothesize is driven by the progressive dystrophy that occurs in RP. These finding highlight the ability of data-driven approaches to reveal patterns in large datasets that provide support for hypotheses about the modeled data.

Fundus photographs were unavailable for each test session. We therefore had to assume that the location of the array stayed stable over time, which is supported by a recent study highlighting the long-term stability of Argus II (Ghani *et al* 2022).

Our models also suggest electrode-fovea distance (i.e. retinal eccentricity) to be an important threshold predictor. As RP progresses from the periphery inwards, we hypothesize that this and other neuroanatomical markers could stand in as a proxy

for disease progression. Such measurements are easily taken, making them readily applicable to the prediction of visual outcomes using patient-specific computational models (Beyeler *et al* 2019, Finn *et al* 2020, Granley and Beyeler 2021). Incorporating this type of clinical knowledge measured over the duration of the study, not just from system fitting, could add more precise specifications of features that are important for perceptual threshold prediction. However, as our study is limited to Argus II data, future work should focus on extending these results to other retinal implants.

An additional opportunity for future work includes studying the impact of similar clinical measures on phosphene appearance (Hou *et al* 2023) and visual acuity (Spencer *et al* 2023), and how these change as RP progresses (Beyeler *et al* 2017b). This highlights an important direction, as phosphene perception is only one step towards usable prosthetic vision.

Explainable, data-driven approaches that enable accurate, automated inference and yield insights into non-trivial relationships between measured features, such as those studied in this work, may offer great benefits to researchers and practitioners of the neuroprosthetics community. Insights from these models could be leveraged by experts to further improve diagnosis and intervention strategies (Brunton and Beyeler 2019), transforming clinical practice in predicting visual outcomes.

Data availability statement

The data cannot be made publicly available upon publication because they contain sensitive personal information. The data that support the findings of this study are available upon reasonable request from the authors.

Acknowledgments

This work was supported by the National Eye Institute of the National Institutes of Health under Award Number R00-EY029329. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

We would like to thank Arup Roy and Jessy Dorn (who were with Second Sight Medical Products, Inc.) for providing the raw data as well as Profs. Ione Fine and Geoff Boynton at the University of Washington for insightful discussions during an earlier iteration of this work.

Author contributions

M B and A R compiled the dataset and produced preliminary results. G P, Z H, and M B performed data processing, modeling, and evaluation. G P and M B wrote the manuscript. All authors approved the final version of the manuscript.

Conflict of interest

The authors were collaborators with Second Sight Medical Products, Inc. (now Vivani Medical, Inc.), the company that developed, manufactured, and marketed the Argus II Retinal Prosthesis System referenced within this article. Second Sight had no role in study design, data analysis, decision to publish, or preparation of the manuscript.

ORCID iDs

Galen Pogoncheff https://orcid.org/0000-0001-6248-0992
Ariel Rokem https://orcid.org/0000-0003-0679-1985
Michael Beyeler https://orcid.org/0000-0001-5233-844X

References

Adadi A and Berrada M 2018 Peeking inside the black-box: a survey on explainable artificial intelligence (XAI) *IEEE Access* **6** 52138–60

Ahuja A K, Yeoh J, Dorn J D, Caspi A, Wuyyuru V, McMahon M J, Humayun M S, Greenberg R J and Dacruz L 2013 Factors affecting perceptual threshold in Argus II retinal prosthesis subjects *Transl. Vis. Sci. Technol.* **2** 1

Amidan B, Ferryman T and Cooley S 2005 Data outlier detection using the Chebyshev theorem *2005 IEEE Aerospace Conf.* pp 3814–9

Beyeler M, Boynton G M, Fine I and Rokem A 2017a pulse2percept: a Python-based simulation framework for bionic vision *Proc. 16th Science in Python Conf.* ed K Huff, D Lippa, D Niederhut and M Pacer pp 81–88

Beyeler M, Nanduri D, Weiland J D, Rokem A, Boynton G M and Fine I 2019 A model of ganglion axon pathways accounts for percepts elicited by retinal implants *Sci. Rep.* **9** 1–16

Beyeler M, Rokem A, Boynton G M and Fine I 2017b Learning to see again: biological constraints on cortical plasticity and the implications for sight restoration technologies *J. Neural Eng.* **14** 051003

Brunton B W and Beyeler M 2019 Data-driven models in human neuroscience and neuroengineering *Curr. Opin. Neurobiol.* **58** 21–29

Chawla N V, Bowyer K W, Hall L O and Kegelmeyer W P 2002 SMOTE: synthetic minority over-sampling technique *J. Artif. Int. Res.* **16** 321–57

Chen T and Guestrin C 2016 XGBoost: a scalable tree boosting system *Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, KDD '16* (Association for Computing Machinery) pp 785–94

Chenais N A L, Airaghi Leccardi M J I and Ghezzi D 2021 Photovoltaic retinal prosthesis restores high-resolution responses to single-pixel stimulation in blind retinas *Commun. Mater.* **2** 1–16

Curcio C A and Allen K A 1990 Topography of ganglion cells in human retina *J. Comp. Neurol.* **1** 5–25

da Cruz L *et al* (for the Argus II Study Group) 2013 The Argus II epiretinal prosthesis system allows letter and word reading and long-term function in patients with profound vision loss *Br. J. Ophthalmol.* **97** 632–6

de Balthasar C *et al* 2008 Factors affecting perceptual thresholds in epiretinal prostheses *Investigative Ophthalmol. Vis. Sci.* **49** 2303–14

Dorn J D, Ahuja A K, Caspi A, da Cruz L, Dagnelie G, Sahel J A, Greenberg R J, McMahon M J and Grp A I S 2013 The detection of motion by blind subjects with the epiretinal 60-electrode (Argus II) retinal prosthesis *JAMA Ophthalmol.* **131** 183–9

Finn K E, Zander H J, Graham R D, Lempka S F and Weiland J D 2020 A patient-specific computational framework for the Argus II implant *IEEE Open J. Eng. Med. Biol.* **1** 190–6

Ghani N, Bansal J, Naidu A, and Chaudhary K M 2022 Long term positional stability of the Argus II Retinal Prosthesis epiretinal implant.

Granley J and Beyeler M 2021 A computational model of phosphene appearance for epiretinal prostheses *2021 43rd Annual Int. Conf. IEEE Engineering in Medicine Biology Society (EMBC)* pp 4477–81

Hou Y, Nanduri D, Granley J, Weiland J D and Beyeler M 2023 Axonal stimulation affects the linear summation of single-point perception in three Argus II users *medRXiv Preprint* https://doi.org/10.1101/2023.07.21.23292908 (posted online 26 December 2023, accessed 1 January 2024)

Hu Z and Beyeler M 2021 Explainable AI for retinal prostheses: predicting electrode deactivation from routine clinical measures *2021 10th Int. IEEE/EMBS Conf. on Neural Engineering (NER)* pp 792–6

Lundberg S M and Lee S 2017 A unified approach to interpreting model predictions *Advances in Neural Information Processing Systems* vol 30, ed I Guyon, U V Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan and R Garnett (Curran Associates, Inc.)

Mehta P *et al* (UK Biobank Eye and Vision Consortium) 2021 Automated detection of glaucoma with interpretable machine learning using clinical data and multimodal retinal images *Am. J. Ophthalmol.* **231** 154–69

Palanker D, Le Mer Y, Mohand-Said S, Muqit M and Sahel J A 2020 Photovoltaic restoration of central vision in atrophic age-related macular degeneration *Ophthalmology* **127** 1097–104

Shivdasani M N, Sinclair N C, Dimitrov P N, Varsamidis M, Ayton L N, Luu C D, Perera T, McDermott H J and Blamey P J 2014 Factors affecting perceptual thresholds in a suprachoroidal retinal prosthesis *Investigative Ophthalmol. Vis. Sci.* **55** 6467–81

Snoek J, Larochelle H and Adams R P 2012 Practical Bayesian optimization of machine learning algorithms *Advances in Neural Information Processing Systems* vol 25

Spencer M, Kameneva T, Grayden D B, Burkitt A N and Meffin H 2023 Quantifying visual acuity for pre-clinical testing of visual prostheses *J. Neural Eng.* **20** 016030

Srivastava N, Hinton G, Krizhevsky A, Sutskever I and Salakhutdinov R 2014 Dropout: a simple way to prevent neural networks from overfitting *J. Mach. Learn. Res.* **15** 1929–58

Willeke K F *et al* 2022 The Sensorium competition on predicting large-scale mouse primary visual cortex activity (arXiv:2206.08666)

Yue L, Falabella P, Christopher P, Wuyyuru V, Dorn J, Schor P, Greenberg R J, Weiland J D and Humayun M S 2015 Ten-year follow-up of a blind patient chronically implanted with epiretinal prosthesis Argus I *Ophthalmology* **122** 2545–52.e1