BAR ILAN UNIVERSITY

Faculty of Engineering

Research proposal towards an M.Sc. degree

ON THE SUBJECT OF

Neuronal cell type classification using deep learning methods

סיווג תאי עצב בעזרת שיטות מבוססות למידה עמוקה

Under the supervision of:

Prof. Orit Shefi
Dr. Ofir Lindenbaum

Author:

Ofek Ophir
207180191

January 2023

Bar-Ilan University
אוניברסיטת בר-אילן

Contents

Abstract

The brain is likely the most complex organ, given the variety of functions it controls, the number of cells it comprises, and their corresponding diversity.
Identifying and studying neurons, the major building blocks of the brain, is a crucial milestone and is essential for understanding brain functionality in health and disease. Previously, the task of identifying and classifying distinct types of neurons relied mostly on their morphological features, requiring massive neuronal tracing. Classification based solely on the neuronal electrophysiological features is still lacking.
Recent developments in machine learning have provided advanced abilities for classifying neurons. However, these methods remain black boxes with no explainability and reasoning. This research proposal aims to provide a robust and explainable deep-learning framework to classify neurons based on their electrophysiological activity. Our analysis is performed on data provided by the Allen Cell Types database. The data contains a survey of biological features derived from single-cell data from both humans and mice. First, neuronal type classification will be performed on the broad binary types of neurons, excitatory or inhibitory. Then, neurons will be classified into sub-types that are based on Cre mouse lines using deep neural networks in an explainable fashion.
We show promising preliminary results in dendrite type classification of excitatory vs. inhibitory neurons and Cre-line classification. These two classifications are performed solely using action potential features as described in the Allen Cell Type database. The model is also inherently interpretable, revealing the correlations between neuronal types and electrophysiological properties.

תקציר

המוח הוא ככל הנראה האיבר המורכב ביותר, בהינתן היקף הפונקציונליות שלו, הכמות ומגוון התאים השונים אשר מהם מורכב. חקר וזיהוי תאי עצב, אבני הבניין המרכזיות של המוח, הוא צעד הכרחי בביולוגיה, וחיוני להבנת תפקוד המוח הבריא והחולה.

בעבר, משימת שיוך תאי העצב לסוגיהם הסתמכה בעיקרה על תכונות מורפולוגיות של התא ולכן היה נדרש מעקב מסיבי אחר התאים. סיווג על בסיס תכונות חשמליות בלבד עדיין לוקה בחסר.

התפתחויות אחרונות בתחום למידת המכונה אפשרו יכולות סיווג תאים מתקדמות בעלות דיוק רב יותר, אולם ללא יכולות הסבר והנמקה.

מטרת המחקר המוצע, היא לספק מסגרת סיווג מבוססת למידת מכונה בעלת יכולות הנמקה באשר להחלטות המודל בעניין סיווג תאי עצב שונים בעכבר ואדם על פי תוויות מוצעות, הן באופן בינארי, בין תאי עצב מעכבים למעוררים, והן במשימת סיווג תאים מרובה תוויות בתתי המחלקות הנלוות לסוג הדנדריט של אותו התא.

סוגים נלווים אלו מבוססים על קווים גנטיים מהונדסים בבעלי חיים ויכולת צביעת תאים ספציפיים ביחס לפרופיל הגנטי שלהם.

אנו מראים תוצאות מבטיחות במשימת הסיווג הבינארית ובמשימת הסיווג מרובת התוויות, שתי משימות הסיווג מתבצעות בהתבסס על נתונים חשמליים של פוטנציאל הפעולה של התא בלבד.

המודל מספק גם יכולות הנמקה להחלטותיו ובכך חושף מתאם בין הסוגים השונים של התא לבין פעילותם החשמלית.

Introduction

The brain is an extremely complex system that contains billions of neurons which propagate signals in order to communicate and share information.
Proper functionality of the nervous system requires mechanisms for information sharing between many neurons in different regions of the brain. Understanding these mechanisms remains an open and challenging problem in biology and requires a detailed and exact description of all brain regions and the neurons composing them.

The task of classifying neurons, the building blocks of the nervous system, has been an ongoing challenge in neuroscience ever since Ramon y Cajal's 'Histology of the Nervous System of Man and Vertebrates' [1] was published, which was to a certain degree, an attempt to classify neurons. Neuroscientists attempting to study the nervous system have hypothesized that the differences in neuron morphology play a role in the neural circuit. For this reason, it is essential to accurately classify the different types of neurons [2].

Defining a solid neuronal cell-type taxonomy is a challenging task, that includes two major obstacles. The first is that classification studies were underpowered and laborious which caused highly biased results. However, in the past decade, technological advances have made it possible to analyze hundreds of neurons accurately and efficiently [3].
The second obstacle is the difficulty of determining how fine and firm the distinctions between neuronal types should be.
If the resolution is too broad (such as the distinction between sensory neurons and motor neurons) then the taxonomy might be too coarse and have little value in experimental purposes. However, if the resolution is too fine, the neuronal taxonomy might have no relevance at all (an extreme case would be to think of each neuron as an independent type).

In our research, we will carry out classification using electrophysiological features such as action potential (AP) threshold, width, height, hyperpolarization voltage, and resting potential. These features are aimed to describe the differences among the observed variability in neuronal activities and can be used to define electrophysiological types of neurons [4], [5]. AP propagation is also impacted by the axonal morphologies and is relevant in the complex axonal ramification patterns of neurons [6], [7].

We seek to provide a deep learning framework for predicting neuronal types in two different classification tasks. The first is classifying neurons to their broad type, excitatory or inhibitory. The second task also includes classifying neurons to their inhibitory subclasses as well as excitatory neurons to their broad class. For both classification tasks, we use data from the Allen Cell Types database [8], a publicly available brain cell database, which contains recordings of electrical stimulus and response in different types of neurons from both human and mouse cells.

Background

At the most fundamental level, cells can be classified into non-neuronal cells and neurons, which can then be further classified into excitatory and inhibitory neurons [9]. The major difference between the two types is that the excitatory neurons release neurotransmitters (most commonly glutamic acid) that fire an action potential in the postsynaptic neuron. In contrast, inhibitory neurons release neurotransmitters (most commonly gamma-aminobutyric acid - GABA) which inhibit the firing of an action potential. Inhibitory interneurons comprise only 10-20% of the total neural population in the cortex but are essential for sensation, movement, and cognition [10].

Excitatory neurons are usually morphologically spiny, with a long apical dendrite, and exhibit less variability in their electrophysiological features. This makes it harder to distinguish between excitatory cell types solely using electrophysiological features. Inhibitory neurons are typically aspiny or sparsely spiny, with a more compact dendritic structure, having a larger variance in electrophysiological properties and tending to spike faster [11], [12]. Neurons can also be classified based on their neurotransmitter, GABAergic neurons which are mostly inhibitory cells, and Glutamatergic neurons which are habitually excitatory and brain-area specific.

Many neuroscientists also consider GABAergic neurons as belonging to one of the four subclasses based on the expression of specific principal markers; these include:

1. Pvalb (parvalbumin) positive.
2. Vip (vasoactive intestinal peptide) positive.
3. Sst (somatostatin) positive.
4. Htr3a (5-hydroxytryptamine receptor) positive but Vip negative.

These subclasses of GABAergic interneurons account for most neurons in specific brain regions. The classes are expressed in a non-overlapping manner, meaning that each neuron belongs to one class in a monovalent fashion, with different cell types accompanied by different physiological properties [13].

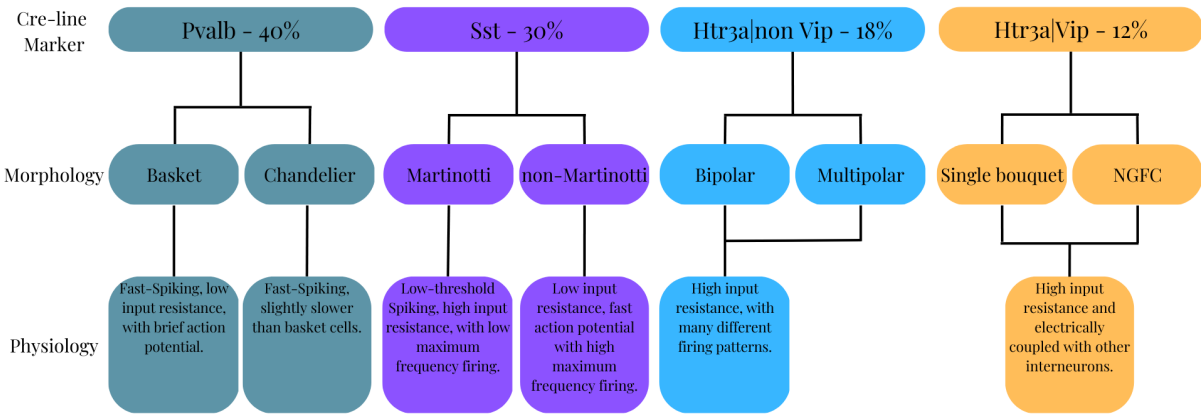


Figure 1. GABAergic neurons in the neocortex express one of four markers: Pvalb (Parvalbumin), Sst (Somatostatin), and the ionotropic serotonin receptor Ht3ar which can be either Vip (Vasoactive intestinal polypeptide) positive or negative.

Furthermore, Glutamatergic neurons can be grouped based on gene markers and laminar locations yet are difficult to distinguish solely using electro-physiological properties [13].

In 2019, The Allen Cell Types database became public [14], [8] and with recent advances in computing capabilities and rapid development of machine and deep learning methods, the domain of neuronal cell-type classification has leaped forward.
From The Allen Cell Types database, 17 electrophysiological neuron types were identified, 4

of which were classified as excitatory subtypes, and 13 were inhibitory. The 13 inhibitory subtypes were further mapped into the four inhibitory interneuron types based on genetic tags: Vip, Ndnf, Sst, and Pvalb. The researchers also identified 38 morphological, and 46 morpho-electric neuron types, all of which were classified using current clamp electrophysiological recordings and the help of dimensionality reduction algorithms such as principal component analysis [15] and t-distributed stochastic neighbor embedding [16].

Using the Allen Cell data Ghaderi et al. [17] developed a semi-supervised method in which neuron classification takes place within three types of neurons.
These types are excitatory pyramidal cells (Pyr), parvalbumin-positive (Pvalb) interneurons, and somatostatin positive (Sst) interneurons from layer 2/3 of the mouse primary visual cortex. The authors achieved accuracies of 91.59 ± 1.69, 97.47 ± 0.67, and 89.06 ± 1.99 for Pvalb, Pyr, and Sst, respectively, which yielded an overall accuracy of 92.67 ± 0.54%.

In 2019, Seo et al [18] used machine learning to predict transgenic markers of neurons using electrophysiology recordings. They evaluated three different methods, namely: Random forest - RF, least absolute shrinkage and selection operator – LASSO, and Artificial neural networks - ANN (Artificial Neural Networks). The prediction performance of the three models was similar and insufficient, with 28.57-46.93% accuracy at predicting the transgenic marker of excitatory neurons (Ctgf, Cux2&Slc17, Nr5a1&Scnn1a, Ntsr1, Rbp4, and Rorb) and 59.03-73.49% accuracy at predicting the transgenic marker of inhibitory neurons (Chrna2, Gad2, Htr3a, Ndnf, Nkx2, Pvalb, Sst, Vip&Chat).

In 2021, Rodríguez et al [19] revealed a circular ordered taxonomy using a transformation of the first two principal components and validated the proposed taxonomy with machine learning models (Linear Discriminant analysis – LDA, RF, Gradient Boosted Decision Tree – GBDT, Support vector machine – SVM, and AvNNet). These models were able to discriminate the different neuron types (4 types of inhibitory neurons – Pvalb, Htr3a, Sst, Vip as well as Glutamatergic excitatory cells) using electrophysiological features with accuracy ranging between 66.1-75.2% for the raw data, and 72.0-80.3% accuracy for a subset of the data that has been cleaned using anomaly detectors.

It is worth noting that these studies only used mouse data, which is most likely due to insufficient human data. To overcome this issue, one can train a machine learning model on data from both human and mouse. But one issue that can arise is that training the model on a certain domain can result in overfitting to the domain features, and lead to a performance gap on other types of data [20] [21]. This is known as the problem of domain shift and can be addressed using tools from domain adaptation. Another issue with using these types of algorithms emerges from the complexity of the neural networks, making it difficult to interpret the model's decisions [22]. Model interpretability is an important aspect in biomedicine, where practitioners need to trust the machine learning model.

This research seeks to address these issues by providing a machine learning framework for predicting neuronal cell types in two steps. The first is classifying excitatory vs inhibitory neurons, and the second is classifying excitatory Glutamatergic cells and the different subclasses (Pvalb, Htr3a, Sst, and Vip) of inhibitory GABAergic cells. Regarding dendrite type classification, we use mus-musculus (house mouse) source data from the Allen Cell data which we have in a larger quantity to learn a distribution over the homo sapiens (human) target data from the Allen Cell data using domain adaptation methods. By doing this we improve results and robustness of the model on the target data. We also explain the importance of the unique features during testing using a concept from cooperative game theory. Then, we use deep neural networks to predict Cre-line labels from mouse data, these account for the different GABAergic neuron subclasses as well as Glutamatergic neurons.

Research Goals

This research has four main objectives:

1. Propose a robust and explainable cell type classification pipeline using AP features.
2. Expand the classification task to neuronal t-type classification and provide the ability to distinguish neurons with different cell marker genes using AP features.
3. Generalize and adapt the classification pipeline to human neuronal cell type classification.
4. Validate our results using a currently known neuron electrophysiological taxonomy.

Methods

Data

The Allen Cell type database [8] https://celltypes.brain-map.org/data contains electrophysiological recordings from 1920 mice and 413 human cells.
The cells from both mouse and human data are categorized by dendrite type: spiny, aspiny, and sparsely spiny, as well as location and layer in the brain.
The cells from mouse samples are further mapped into transgenic targeting such as Pvalb positive, Sst positive, Vip positive, and Htr3a (5-hydroxytryptamine receptor) positive but Vip negative. [13]

The mouse data contains whole-cell current clamp recordings from identified fluorescent Cre-positive neurons or nearby Cre-negative neurons in acute brain slices derived from adult mice. The human data contains whole-cell current clamp recordings from adult human neocortical neurons in brain slices derived from surgical specimens. Each whole-cell current clamp recording is a response to a stimulation consisting of pink noise with a coefficient of variation (CV) equal to 0.2, as it resembles in vivo data. These stimuli consist of 3x3sec noise epochs superimposed on top of square pulses at 0.75, 1, and 1.5 times rheobase.

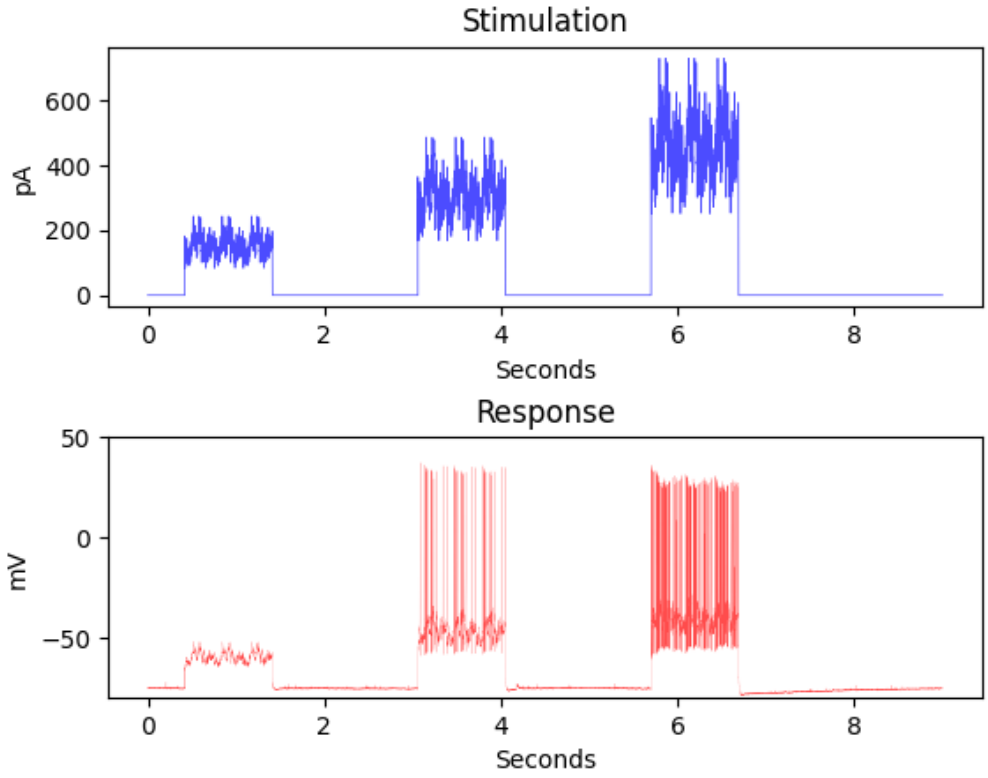


Figure 2. (top) Stimulation of noise pulses with square current injections scaled to three amplitudes, 0.75, 1, and 1.5 times rheobase. (bottom) cell response to the stimulation given.

AP features are extracted from each whole-cell current clamp stimulation response, as described in Table 1 and Figure 3.

mean_threshold_t
mean_threshold_v
mean_threshold_i
mean_peak_t
mean_peak_v
mean_peak_i
mean_trough_t
mean_trough_v
mean_trough_i
mean_upstroke
mean_upstroke_t
mean_upstroke_v
mean_downstroke
mean_downstroke_t
mean_downstroke_v
mean_fast_trough_t
mean_fast_trough_v
mean_fast_trough_i
mean_width
mean_upstroke_downstroke_ratio

Table 1. List of extracted electrophysiological features.

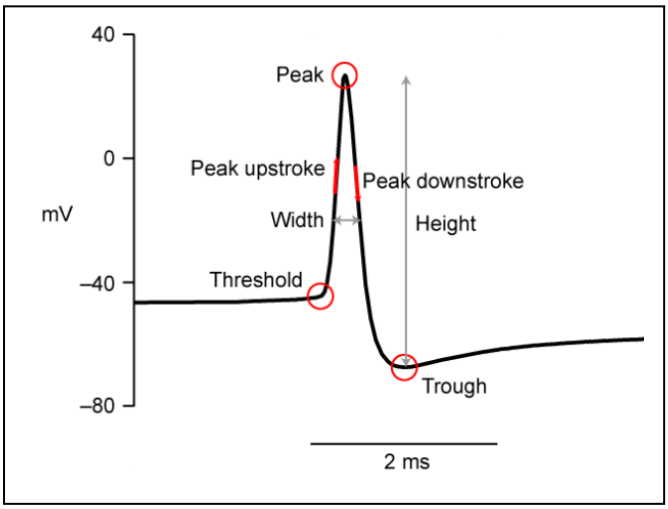


Figure 3. A single action potential and extracted electrophysiological features.

A comprehensive description of these AP features can be found in Table 5.

Artificial Neural Network

Artificial neural networks (ANN) are computing models inspired by biological neural networks. ANNs rely on matrix multiplications followed by nonlinear activation functions to learn complex relations between input and output. ANNs are comprised of artificial neurons which are connected through edges, these edges typically have a weight value that can adjust the strength of the signal at that connection; the weights are 'learned' through an optimizer such as Stochastic Gradient Descent (SGD). [23]
There are many types of architectures regarding artificial neural networks. In this thesis we focus on fully connected neural networks, also referred as multi-layer perceptron (MLP), or just a 'neural network' (NN).

ANN consists of a series of fully connected layers that connect every neuron in a hidden layer i to each neuron in a hidden layer $i + 1$.
The input layer consists of N neurons and is fed data as a tensor of the same size, the output layer can be any number of neurons desired. Usually, in classification tasks, the output layer consists of L neurons that represent each of the L classes (or labels) present in the data.

One major advantage of a NN is that it is agnostic to data structure, meaning that there is no assumption needed to be made about the data input, compared to other neural network architectures which are domain specific, such as Convolutional Neural Networks (CNNs) for vision or Recurrent Neural Networks (RNNs) for sequences of data.

In our research we define two neuronal cell-type classification tasks:

- In the 1st task, cells are taken from both humans and mice and the label is the cell's dendritic type (inhibitory/excitatory).
- In the 2nd task, cells are only taken from mice, and the label is according to the cell's gene marker (Pvalb, Sst, Vip, Htr3a, Glutamatergic).

We use a NN with a sample specific feature selection mechanism for these classification tasks. Furthermore, we evaluate a domain adaptation mechanism to handle measurements from human and mice simultaneously.

Domain Adversarial Neural Network for domain adaptation

Mouse neuronal data is acquired from selected brain areas in adult mice. Cells are identified using transgenic mouse lines harboring fluorescent reporters, with drivers that allow enrichment for cell classes based on marker genes.
On the other hand, human neuronal data is acquired from donated ex vivo brain tissues analyzed from neurosurgical and postmortem sources and is available thanks to the generosity of tissue donors.
Because of this, human neuronal data is difficult to obtain and there is less of it compared to data from mouse (1920 mouse samples vs 413 human samples).

Our aim is to design a model that can classify human neuronal types, but this is difficult due to the scarcity of human samples.
To deal with this issue, we use both mouse data and human data to better classify human samples. This is possible because the two distributions (mouse data and human data) are similar (both come from mammalian brain tissues), but since they are not the same, we use domain adaptation to overcome the domain shift between the two distributions.

Ganin, Yaroslav, et al [24] introduced a technique called 'domain-adversarial neural network' (DANN), that combines both representation learning (i.e., deep feature learning) and unsupervised domain adaptation in an end-to-end training process.
DANN jointly optimizes two adversarial losses:
1. *minimizing* the loss of a *label classifier*.
2. *maximizing* the loss of a *domain classifier*.

Training both losses can be considered as a form of adversarial neural network regularization. On the one hand, the network needs to classify the data into the correct labels. But on the other hand, the predictions made by the network, must be based on features that cannot discriminate between the source domain and target domain.

In our setting, the mouse cells are the source distribution and are more abundant (since it is easier to obtain neurons from the rat brain than the human brain), and the human cells serve as the target distribution.

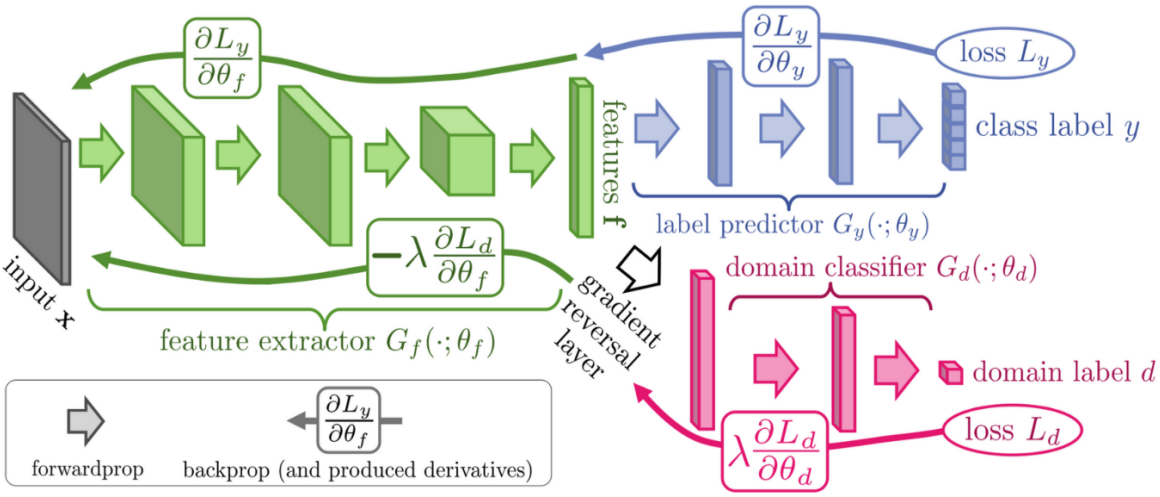


Figure 4. DANN architecture, taken from [24].

Shapley Values

Deep neural networks are often regarded as black boxes, even more so, insights about the predictions made by deep learning models are often opaque.
In certain scientific fields, such as biology, medicine, and others, understanding the reasoning behind a model's predictions is vital [25].

Understanding how neurons differ and interact with each other is a prerequisite to fully comprehend how the brain functions.
Therefore, one must define a neuronal taxonomy that is based on known and measurable features. That being the case for neuron classification, a task which can benefit the ability to explain the predictions made by the classification model. Regarding machine learning, this is considered 'explainable' artificial intelligence.

To address this problem, we use the SHAP library [26] (Shapley Additive exPlanations), the framework assigns each feature an importance value for a particular prediction.
The library is based on a concept from coalitional game theory called 'Shapley values', which states that a prediction cab be explained by assuming that each feature value of the instance is a "player" in a game where the prediction is the payout. Shapley values tells us how to fairly distribute the payout among the features.

Our objective is to 'explain' the classification predictions on both tasks - excitatory/inhibitory classification and T-type classification. We aim that the explainability SHAP provides aid the understanding of the differences between the electrophysiological classes of neurons.

Locally Sparse Neural Network

Collecting whole-cell current clamp recordings is computationally challenging, therefore, the Allen Cell database contains only 1920 mouse cells and 413 human cells.
The size of the data makes it difficult to train an overparametrized NN while avoiding overfitting. To address this obstacle, we implemented the recently proposed intrinsically interpretable network for biomedical data, 'Locally Sparse Interpretable Network' – LSPIN [27]. We use the model to predict GABAergic subclasses in mice data and distinguish them from Glutamatergic neurons.

The model is a locally sparse neural network in which the local sparsity is learned to identify the subset of the most relevant features for each sample.
LSPIN includes two neural networks which are trained in tandem:
The 1st is a gating network, that predicts the sample-specific sparsity patterns.
The 2nd is a prediction network, that classifies the neuron type using the extracted features, that are detailed in *Table 1*.

By forcing the model to select a subset of the most informative features for each sample, we can reduce overfitting in low-sample size data.
Another benefit of this model is that by predicting the most informative features locally, we obtain an interpretation of the model's predictions.

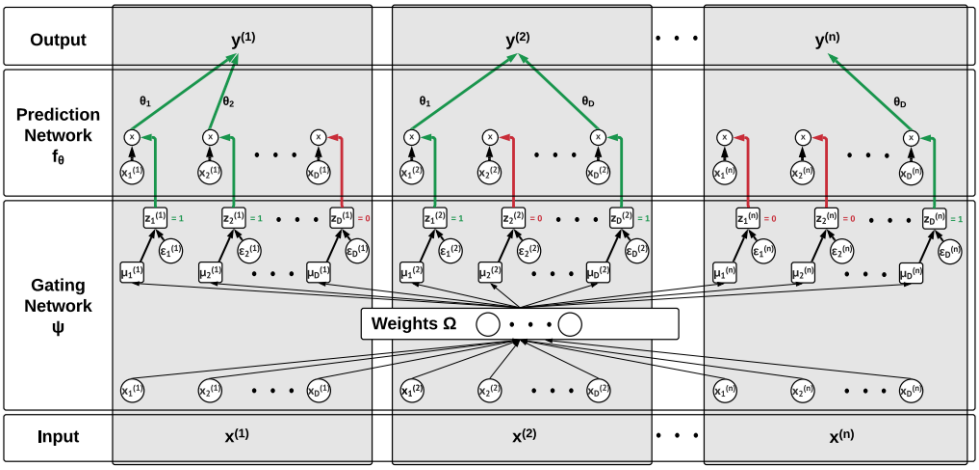


Figure 5. The architecture of Locally Linear SParse Interpretable Networks (LLSPIN) the data $\left\{x^{(i)} = \left[x_1^{(i)}, x_2^{(i)}, .., x_D^{(i)}\right]\right\}_{i=1}^{n}$ *is fed to a gating network and to a prediction network. The gating network learns to predict a set of parameters* $\left\{\mu_d^{(i)}\right\}_{d=1,i=1}^{D,n}$ *which depict the behavior of the local stochastic gates* $z_d^{(i)} \in [0,1]$ *that sparsify the set of features used by the prediction network, Taken from [27].*

Preliminary Results

The data was downloaded from the Allen Cell Type database. Electrophysiological features were extracted into tabular format, then, similar Cre lines were grouped together according to Figure 1.
The data was normalized using 'StandardScaler' and split into 80% train and 20% validation.

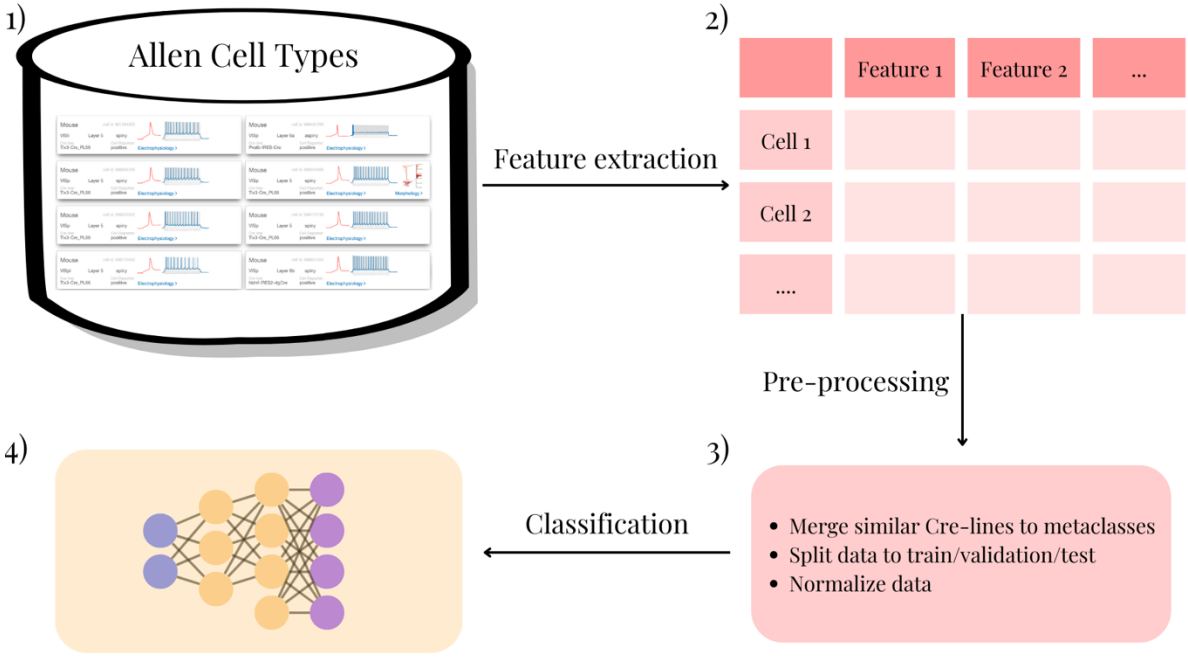


Figure 6. Electrophysiological features pipeline. (1) Data is obtained from the Allen Cell Types Database. (2) Features are extracted in a tabular format. (3) Similar Cre-lines are merged, data is split into train/test and normalized, (4) classification occurs through previously described algorithms.

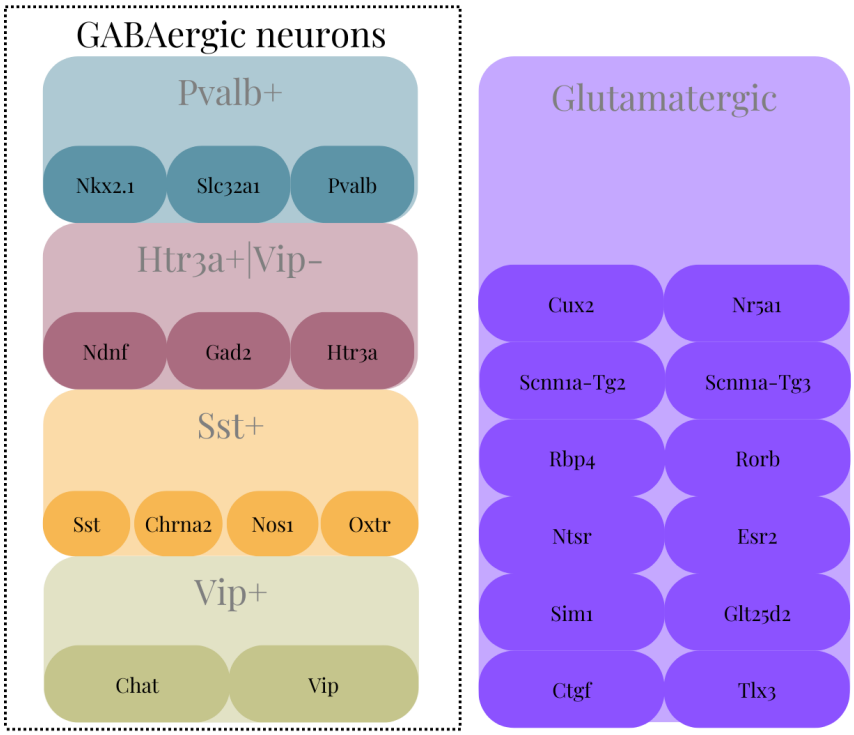


Figure 7. Cre-lines composing the defined metaclasses.

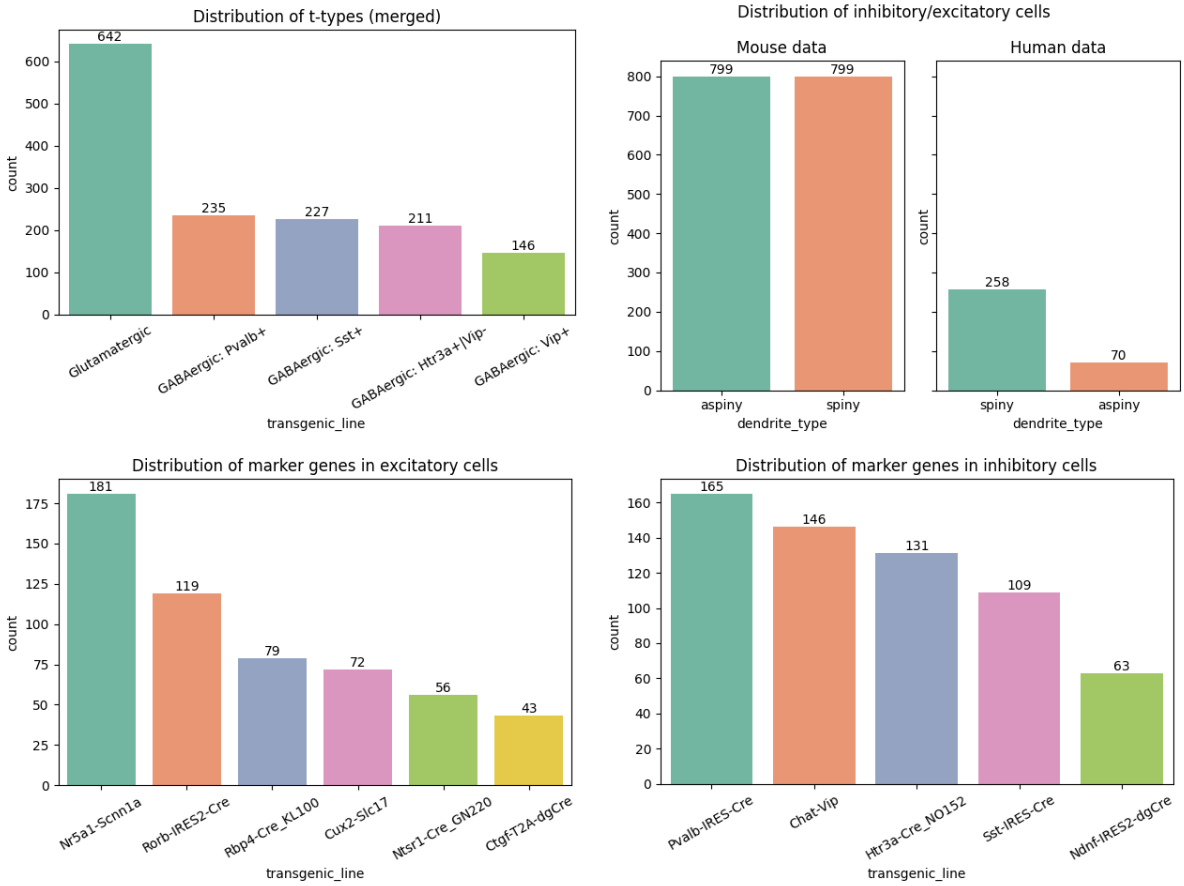


Figure 8. Top left: distribution of neuronal t-types, merged into 5 classes. Top right: distribution of excitatory and inhibitory cells. Bottom left: distribution of marker genes in the excitatory cells. Bottom right: distribution of marker genes in inhibitory cells.

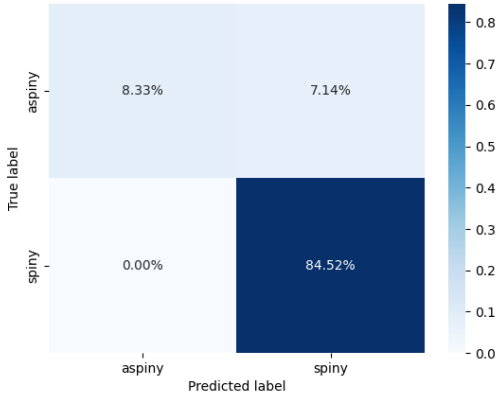
We extracted the electrophysiological features from each sample and fed as inputs to the models described in the 'Methods' chapter.
A fully connected neural network was trained on the mouse cells' data and the human cell's data separately.
The test results for the mouse model over the mouse data are shown in *Table 2*.

Human classification	Mouse classification
Accuracy: 0.928 F1 Score: 0.959 Precision: 0.922 Recall: 1.0 ROC AUC: 0.769	Accuracy: 0.921 F1 Score: 0.923 Precision: 0.941 Recall: 0.905 ROC AUC: 0.921

Table 2. Classification of aspiny vs spiny dendrite type using ANN.

The model achieved ~93% accuracy in classifying dendrite types for human cells, even though the human data is imbalanced meaning that most neurons are spiny and only a minority are aspiny.

Human classification confusion matrix



Mouse classification confusion matrix

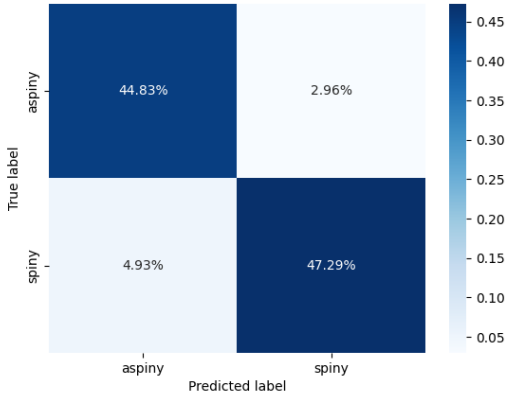


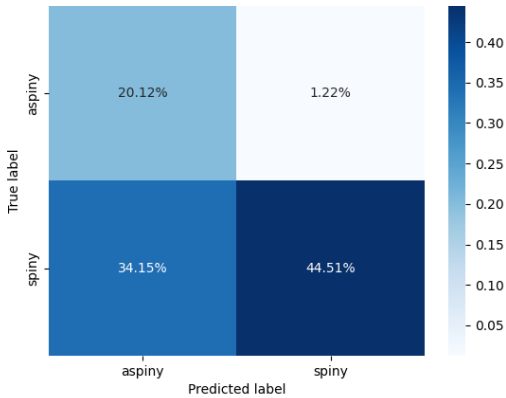
Figure 9. (left) Confusion matrix for human classification. (right) Confusion matrix for mouse classification.

Each network was also tested on the opposite data (trained mouse network on human data and vice-versa) to check whether a domain shift exists.

Human domain (Network trained on mouse data)	Mouse domain (Network trained on human data)
Accuracy: 0.646 F1 Score: 0.715 Precision: 0.973 Recall: 0.565 ROC AUC: 0.754	Accuracy: 0.596 F1 Score: 0.712 Precision: 0.554 Recall: 0.997 ROC AUC: 0.595

Table 3. Domain shift results using Artificial Neural Networks.

Human classification confusion matrix



Mouse classification confusion matrix

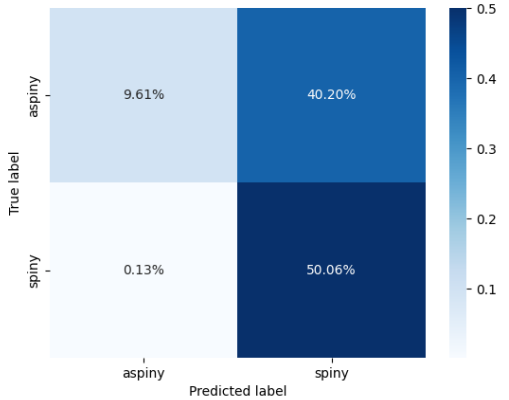


Figure 10. (left) Confusion matrix for human classification with domain shift. (right) Confusion matrix for mouse classification with domain shift.

As can be inferred from the results in *Figure 8* and Table 3, There is a domain shift between the human data and mouse data.
This is both due to the difference in electrophysiological properties as well as the lack of aspiny data for the human domain.

To deal with this problem, and create a robust classification pipeline for each organism, we implemented the DANN model, which deals with domain adaptation through an adversarial process. The model's architecture is shown in *Figure 4*. DANN architecture, taken from .

The data used for training the DANN model consisted of both human cells and mouse cells, additionally the model was tested on mouse/human cells alone, to provide a better understanding of its capabilities in each of the domains.

The model was trained on 1534 cells, out of them, 1271 mouse cells and 262 human cells with a split of 920 training cells, 231 validation cells, and 391 testing cells.
We also tested each one of the domains alone, to understand the performance of each organism with 320 mouse cells and 66 human cells.
Results are shown in Table 4. DANN results for the Human classification task (left) and mouse classification task (right)..

Human domain	Mouse domain
Accuracy: 0.954 F1 Score: 0.972 Precision: 0.964 Recall: 0.981 ROC AUC: 0.899	Accuracy: 0.865 F1 Score: 0.890 Precision: 0.805 Recall: 0.994 ROC AUC: 0.852

Table 4. DANN results for the Human classification task (left) and mouse classification task (right).

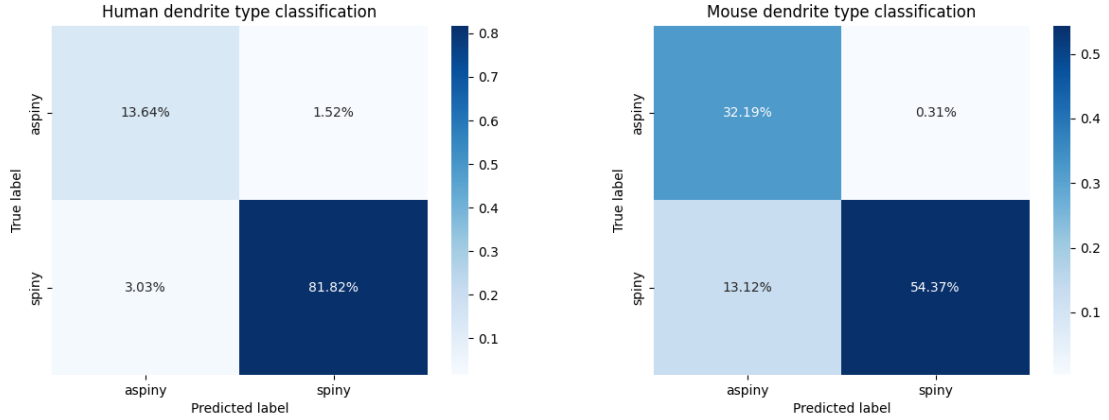


Figure 11. (left) Human classification confusion matrix. (right) Mouse classification confusion matrix.

Using Shapley values, we extracted the most expressive features for the classification task.

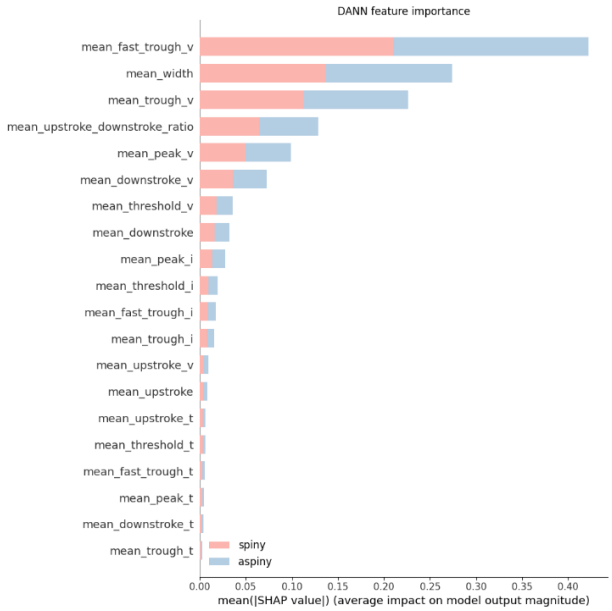


Figure 12. Feature importance using Shapley values.

The hyperpolarization voltage, action potential width and the upstroke/downstroke ratio are the most prominent features in terms of distinguishing inhibitory cells from excitatory cells.

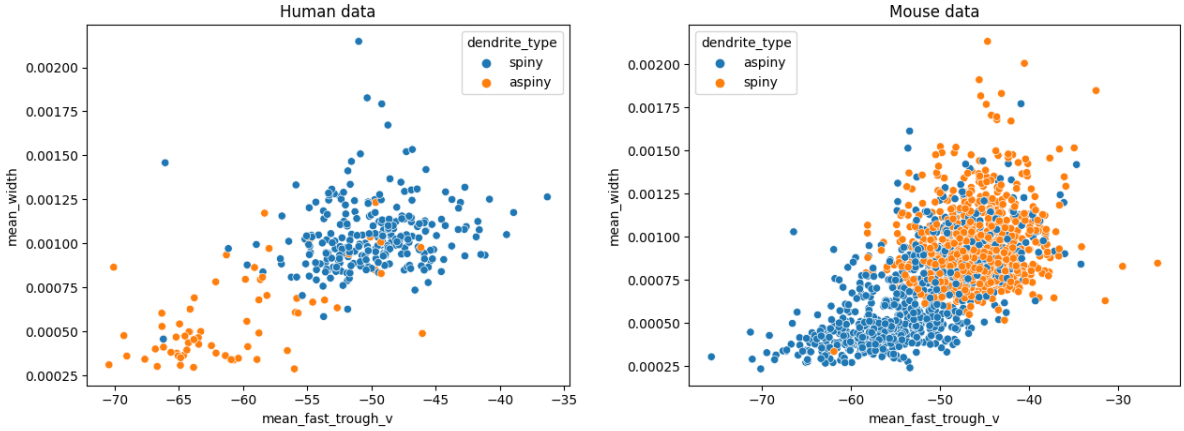


Figure 13. trough voltage and AP width in human cells (left) and mouse cells (right).

Using the locally sparse model we were able to distinguish between different neuron t-types with state-of-the-art results (81% accuracy).

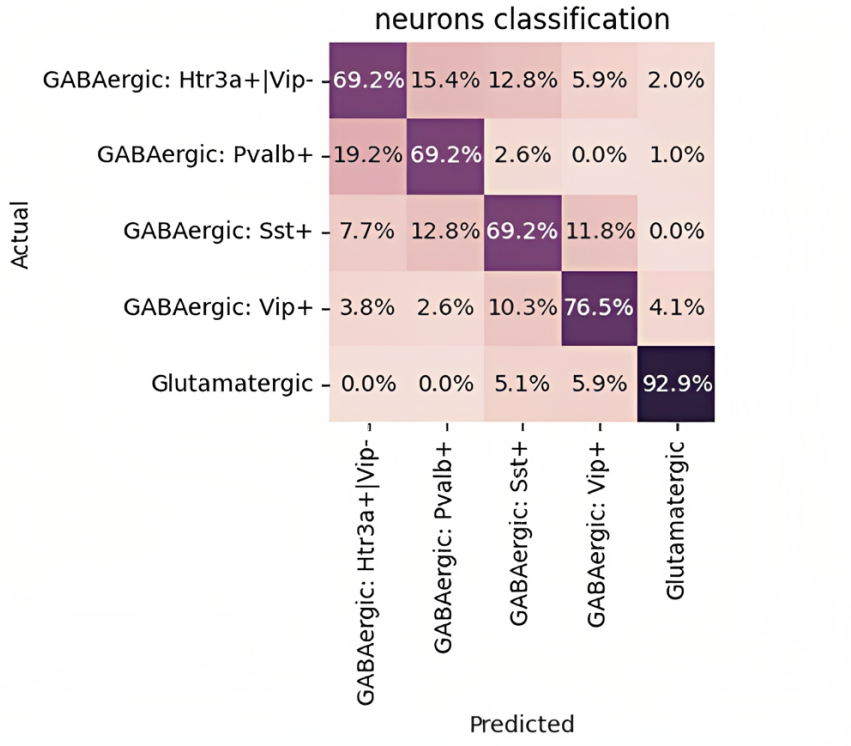


Figure 14. Confusion matrix for neuron t-type classification (Htr3a+, Pvalb+, Sst+, Vip+, Glutamatergic).

References

[1] S. R. Cajal, "Histology of the nervous system of man and vertebrates," *History of Neuroscience,* 1995.

[2] H. Zeng and J. R. Sanes, "Neuronal cell-type classification: challenges, opportunities and the path forward," *Nature Reviews Neuroscience,* pp. 530-546., 2017.

[3] Y. Zhao, S. Inayat, D. A. Dikin, J. H. Singer, R. S. Ruoff and J. B. Troy, "Patch clamp technique: review of the current state of the art and potential contributions from nanoengineering," *Proceedings of the Institution of Mechanical Engineers, Part N: Journal of Nanoengineering and Nanosystems,* pp. 222(1), 1-11., 2008.

[4] M. Beierlein, J. R. Gibson and B. W. Connors, "Two dynamically distinct inhibitory networks in layer 4 of the neocortex.," *Journal of neurophysiology,* pp. 90(5), 2987-3000., 2003.

[5] L. G. Nowak, M. V. Sanchez-Vives and D. A. McCormick, "Lack of orientation and direction selectivity in a subgroup of fast-spiking inhibitory interneurons: cellular and synaptic mechanisms and comparison with other electrophysiological cell types," *Cerebral Cortex,* pp. 18(5), 1058-1078., 2008.

[6] N. Ofer, O. Shefi and G. Yaari, "Axonal Tree Morphology and Signal Propagation Dynamics Improve Interneuron Classification," *Neuroinformatics 18(4),* pp. 581-590, 2020.

[7] N. Ofer and O. Shefi, "Axonal geometry as a tool for modulating firing patterns," *Applied Mathematical Modelling, 40(4),* pp. 3175-3184, 2016.

[8] "Allen Cell Types Database, Technical White Paper: Electrophysiology," (FEB 2018)," 2017.

[9] S. Melzer and H. Monyer, "Diversity and function of corticopetal and corticofugal GABAergic projection neurons," *Nature Reviews Neuroscience,* pp. 21(9), 499-515., 2020.

[10] O. K. Swanson and A. Maffei, "From hiring to firing: activation of inhibitory neurons and their recruitment in behavior," *Frontiers in molecular neuroscience,* pp. 12, 168., 2019.

[11] Y. Kawaguchi, "Neostriatal cell subtypes and their functional roles," *Neuroscience research,* pp. 27(1), 1-8., 1997.

[12] C. Strübing, Ahnert-Hilger, G, J. Shan, B. Wiedenmann, J. Hescheler and A. M. Wobus, "Differentiation of pluripotent embryonic stem cells into the neuronal lineage in vitro gives rise to mature inhibitory and excitatory neurons.," *Mechanisms of development,* pp. 53(2), 275-287., 1995.

[13] R. Tremblay, S. Lee and B. Rudy, "GABAergic interneurons in the neocortex: from cellular properties to circuits," *Neuron,* pp. 91(2), 260-292., 2016.

[14] N. W. Gouwens, S. A. Sorensen, J. Berg, C. Lee, T. Jarsky, J. Ting and S. M. S. e. al., "Classification of electrophysiological and morphological neuron types in the mouse visual cortex," *Nature neuroscience 22.7,* pp. 1182-1195., 2019.

[15] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley interdisciplinary reviews: computational statistics,* pp. 2(4), 433-459., 2010.

[16] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of machine learning research,* p. 9(11), 2008.

[17] P. Ghaderi, H. R. Marateb and M. S. Safari, "Electrophysiological profiling of neocortical neural subtypes: a semi-supervised method applied to in vivo whole-cell patch-clamp data," *Frontiers in neuroscience,* pp. 12, 823, 2018.

[18] I. Seo and H. Lee, "Predicting transgenic markers of a neuron by electrophysiological properties using machine learning," *Brain Research Bulletin,* pp. 150, 102-110., 2019.

[19] A. Rodríguez-Collado and C. Rueda, "Electrophysiological and Transcriptomic Features Reveal a Circular Taxonomy of Cortical Neurons," *Frontiers in Human Neuroscience,* p. 410, 2021.

[20] R. Novak, Y. Bahri, D. A. Abolafia, J. Pennington and J. Sohl-Dickstein, "Sensitivity and generalization in neural networks: an empirical study," *arXiv preprint arXiv,* p. 1802.08760, 2018.

[21] A. Farahani, S. Voghoei, K. Rasheed and H. R. Arabnia, "A brief review of domain adaptation," *Advances in data science and information engineering,* pp. 877-894, 2021.

[22] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf and G. Z. Yang, "XAI—Explainable artificial intelligence," *Science robotics,* pp. 4(37), eaay7120., 2019.

[23] S. Ruder, "An overview of gradient descent optimization algorithms," *arXiv preprint arXiv: 1609.04747.,* 2016.

[24] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette and V. Lempitsky, "Domain-adversarial training of neural networks," *The journal of machine learning research,* pp. 17(1), 2096-2030, 2016.

[25] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (xai): Toward medical xai," *EEE transactions on neural networks and learning systems,* vol. 32, no. 11, pp. 4793-4813, 2020.

[26] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems, 30.,* 2017.

[27] J. Yang, O. Lindenbaum and Y. Kluger, "Locally Sparse Neural Networks for Tabular Biomedical Data," *In International Conference on Machine Learning,* pp. pp. 25123-25153, 2022.

[28] T. Akiba, S. Sano, T. Yanase, T. Ohta and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," *In Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining,* pp. pp. 2623-2631, 2019.

[29] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The bulletin of mathematical biophysics,* vol. 5, no. 4, pp. 115-133, 1943.

Appendix

Electrophysiological Features

Action potential peak	Maximum value of the membrane potential during the action potential (i.e., between the action potential's threshold and the time of the next action potential, or end of the response).
Action potential trough	Minimum value of the membrane potential in the interval between the peak and the time of the next action potential.
Action potential fast trough	Minimum membrane potential value in the interval lasting 5 MS after the peak.
Action potential height	The action potential height was defined as the difference between the action potential peak and the action potential trough.
Action potential full width	The action potential full width was defined as the width at half-height. The points in the voltage trace on either side of the peak that matched half the action potential height were identified, and the width was defined as the time interval between these points.
Action potential peak upstroke	The maximum value of dV/dt between the action potential threshold and the action potential peak.
Action potential peak downstroke	The minimum value of dV/dt between the action potential peak and the action potential trough.
Upstroke/downstroke ratio	The ratio between the absolute values of the action potential peak upstroke and the action potential peak downstroke.

Table 5. Full description of the electrophysiological features extracted from the action potential.

Domain Adversarial training of Neural Networks description

Notation

We can define the notation:

- $X \subseteq \mathbb{R}^d$ input space, $Y \subseteq \mathbb{R}^L$ output space
- P_S source domain, a distribution over $X \times Y$
  - D_S marginal distribution over X

$$S = \{(x_i, y_i)\}_{i=1}^{n} \sim (D_S)^n$$

- P_T target domain, a different distribution over $X \times Y$

D_T marginal distribution over $X \times Y$

$$T = (x_i, y_i)_{i=n+1}^{N} \sim (D_T)^{n'}$$

Goal

- Define a classifier $\eta: X \to Y$
- Define Target risk: $R_{D_T}(\eta) = \Pr(\eta(x) \neq y),\ (x, y) \sim D_T$
- The goal of the algorithm is to build a classifier $\eta: X \to Y$ so that $R_{D_T}(\eta)$ is low.
- If η is learned from the source domain, how will it perform on the target domain?

Domain adaptation is achieved by predictions that are based on features that cannot discriminate between source and target domains.
Final classification decisions are made using features that are both discriminative and invariant to the change of domains.
A good representation for cross-domain transfer is one for which an algorithm cannot identify between the source and target domains.

Loss

We Note the prediction loss and domain loss:

$$L_y^i(\theta_f, \theta_y) = L_y\big(G_y\big(G_f(x_i; \theta_f); \theta_y\big), y_i\big)$$

$$L_d^i(\theta_f, \theta_d) = L_d\big(G_d\big(G_f(x_i; \theta_f); \theta_d\big), d_i\big)$$

Training the DANN consists of optimizing:

$$a\ E(\theta_f, \theta_y, \theta_d) = \frac{1}{n}\sum_{i=1}^{n} L_y^i(\theta_f, \theta_y) - \lambda\left(\frac{1}{n}\sum_{i=1}^{n} L_d^i(\theta_f, \theta_d) + \frac{1}{n'}\sum_{i=n+1}^{N} L_d^i(\theta_f, \theta_d)\right)$$

a $N = n + n' \to$ number of samples
$S = \{(x_i, y_i)\}_{i=1}^{n} \sim (D_s)^n \to$ source sample
$T = \{x_i\}_{i=n+1}^{N} \sim (D^T)^{n'} \to$ target sample
We want to Find the saddle point such that:

$$(\hat{\theta}_f, \hat{\theta}_y) = argmin\ E(\theta_f, \theta_y, \hat{\theta}_d)$$

$$\hat{\theta}_d = argmax\ E(\hat{\theta}_f, \hat{\theta}_y, \theta_d)$$

This can be done using gradient updates:

$$\theta_f \leftarrow \theta_f - \mu\left(\frac{\partial L_y^i}{\partial \theta_f} - \lambda\frac{\partial L_d^i}{\partial \theta_f}\right)$$

$$\theta_y \leftarrow \theta_y - \mu \frac{\partial L_y^i}{\partial \theta_y}$$

$$\theta_d \leftarrow \theta_d - \mu\lambda \frac{\partial L_d^i}{\partial \theta_d}$$

The paper suggests doing so using the 'Gradient Reversal Layer,' which is a novel layer that consists of flipping the gradient during backpropagation and outputting the same input in forward propagation, it can be interpreted as:

- Forward propagation:

$$R(x) = x$$

- Backpropagation:

$$\frac{\partial R}{\partial x} = -\boldsymbol{I}$$

We note the final optimization function as:

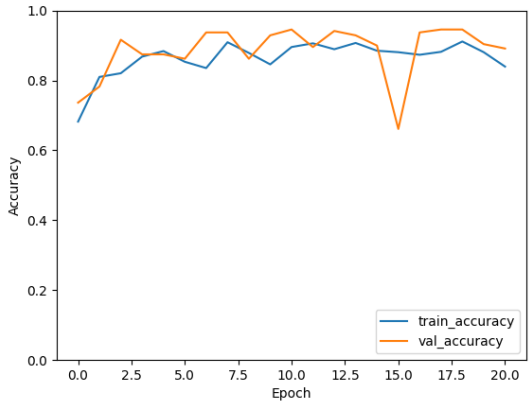
$$\begin{aligned}\tilde{E}(\theta_f, \theta_y, \theta_d) = \frac{1}{n}\sum_{i=1}^{n} L_y\big(G_y\big(G_f(x_i;\theta_f);\theta_y\big), y_i\big) \\ - \lambda\left(\frac{1}{n}\sum_{i=1}^{n} L_d\left(G_d\left(\boldsymbol{R}\left(G_f(x_i;\theta_f)\right);\theta_d\right), d_i\right)\right. \\ \left. + \frac{1}{n'}\sum_{i=n+1}^{N} L_d\left(G_d\left(\boldsymbol{R}\left(G_f(x_i;\theta_f)\right);\theta_d\right), d_i\right)\right)\end{aligned}$$

Hyperparameter optimization

ANN

Hyper-parameters: Learning rate: 0.1, Weight decay: 0.0001, Optimizer: Adam, Architecture: [20 (input), 512, 256, 128, 64, 32, 2 (output)], Dropout rate: 0.2, Activation function: Swish activation for all layers except output which was SoftMax.
These were found optimal through grid search optimization.

ANN training on mouse data



ANN training on human data

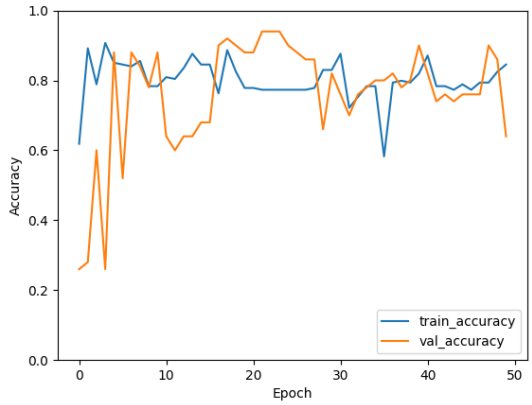


Figure 15. (left) Fully connected neural network training on mouse data. (right) Fully connected neural network training on human data.

DANN

Architecture: 20 (input), 512, 256, 128, 64, 32, 2 (output), Activation function for each hidden layer: Swish, Activation function for each output layer: SoftMax, Optimizer: Adam, Weight decay: 0.0001, Learning rate: 0.1, Batch size: 64, Dropout rate: 0.2, Number of epochs: 1024, λ: 0.7

DANN training process:

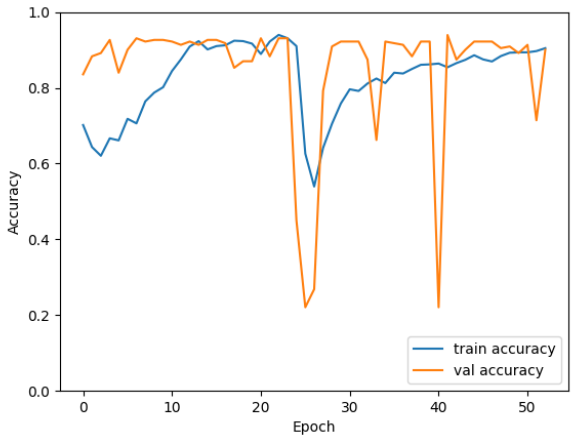


Figure 16. training process of the DANN model.

LSPIN interpretability

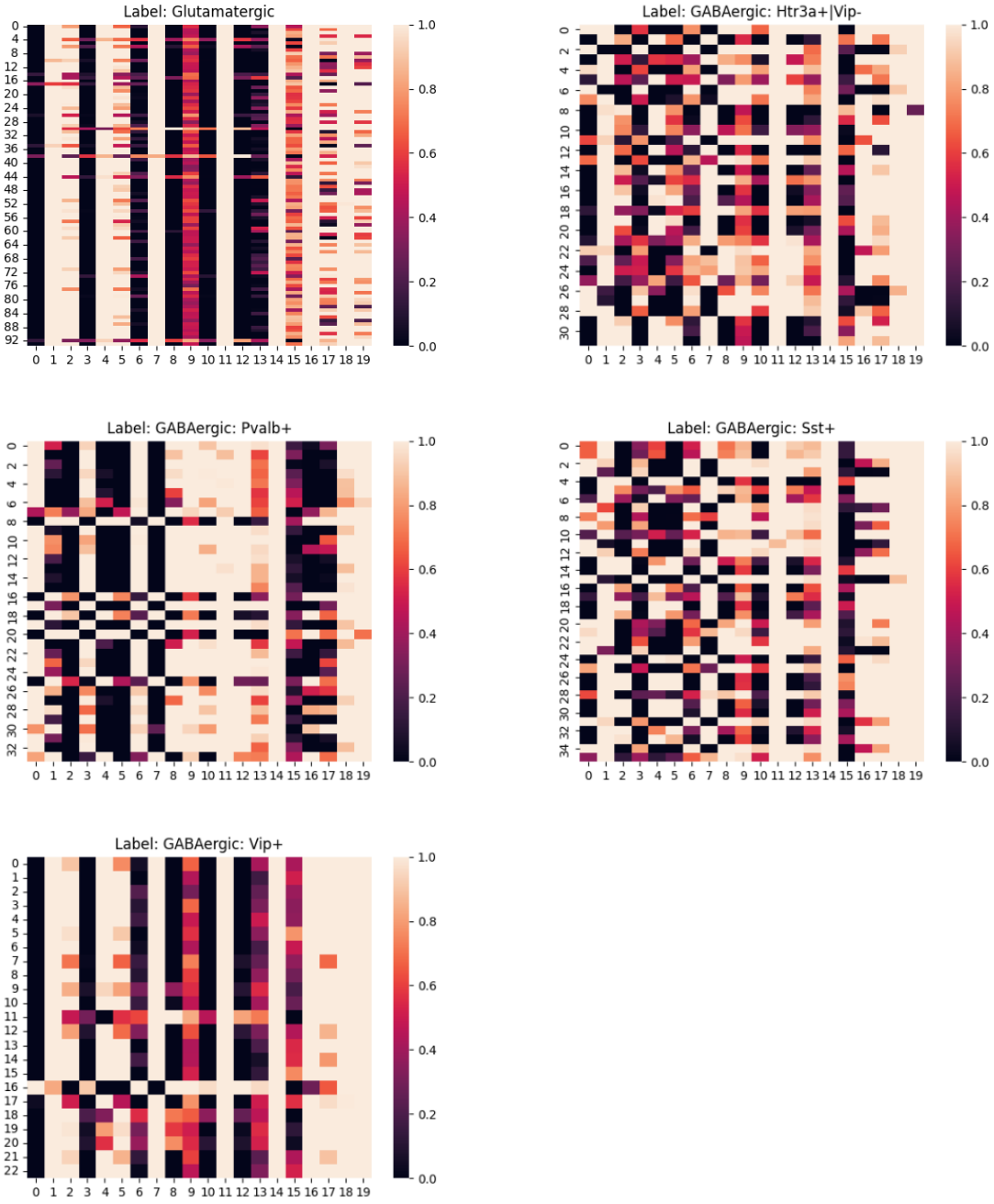


Figure 17. LSPIN feature selection for each class, color describes the weight of each feature that are presented in Table 1.