Automated novel neuronal type discoveries by machine learning

Michael Jones, Hideki Sasaki, Chi-Chou Huang, James S.J. Lee

DRVision Technologies LLC, 15921 NE 8th St. Suite 200, Bellevue, WA 98008, USA

Introduction

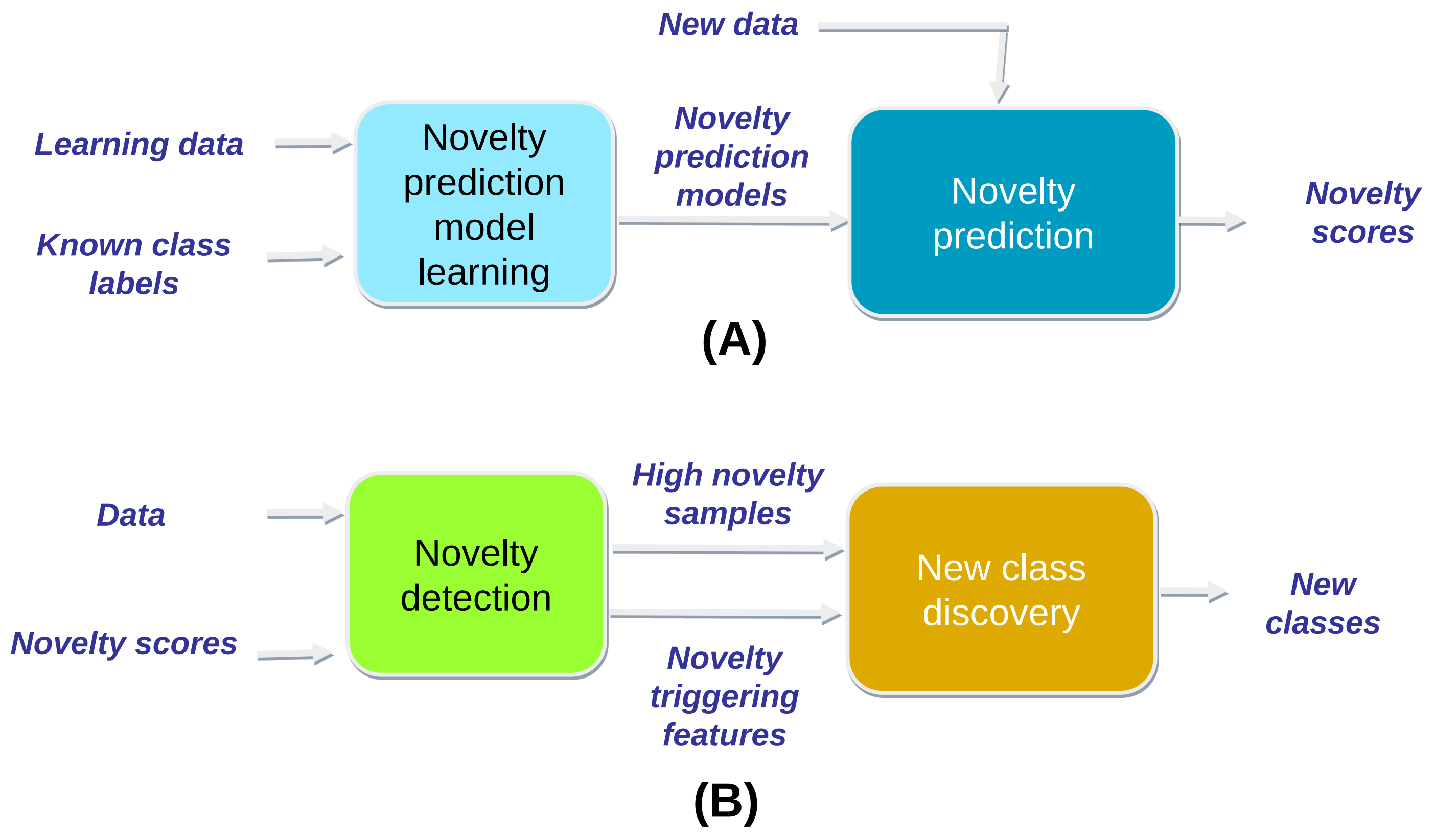
The structures of neuronal dendrites and axons play fundamental roles in synaptic integration and network connectivity. The neuron morphological characterization enables comparative anatomical investigations, morphometric analysis of cells, or brain modeling.

Advancement in neurobiology, microscopy, and imaging software are rapidly transforming the three-dimensional (3D) reconstruction of neuronal morphology into a mainstream technique. However, classification and quantitative characterization of morphologies from 3D microscopy neuronal reconstruction is challenging since it is still unclear how to delineate a neuronal cell types and the best features to define them. There is a critical need for analytical tools to enable the discovery of novel neuronal patterns and cell types automatically.

We developed a machine learning framework for automatic object classification. The framework supports neuron classification and classification of individual dendrite segments, dendrite branches, spines, and other 3D objects such as cells and nuclei. However, it requires the teaching of known neuron types. To facilitate novel neuronal type discoveries, we are extending the framework for novelty detection to identify new or unknown types that the framework has not been taught.

The objective of this study is to assess the effectiveness of the automated novelty detection tool and its application to novel neuronal type discoveries.

Novelty Detection Application Scenario


New data

Learning data

Known class labels

Novelty prediction model learning

Novelty prediction models

Novelty prediction

Novelty scores

(A)

Data

Novelty scores

Novelty detection

High novelty samples

Novelty triggering features

New class discovery

New classes

(B)

Fig 1. New class discovery scenario: A possible novelty detection application scenario including novelty scoring and new class discovery. (A) Novelty scoring : A novelty prediction model learning step inputs learning data, each data point is labeled with a known class. It generates novelty prediction models using the learning data and the known class labels. The models are applied to new data in a novelty prediction step. This step assign a novelty score to each data. (B) New class discovery: A set of data and its associated novelty scores are assessed by a novelty detection step to detect high novelty samples and their novelty triggering features. The information is processed by human analysis in a new class discovery step. This step could find new classes that have not been previously uncovered. The new classes can be added to the known class set and their data points can be included in the new learning data set for updated novelty scoring and next iteration of new class discovery. This iteratively process can be repeated until no new classes can be found from the available data. In this way, new phenotypes , biomarkers or biological events can be efficiently discovered.

Novelty Prediction Models

Novelty prediction identifies data containing new or unknown classes that a machine learning classifier has not been trained on. For example, if there are three classes and the classifier is taught the first two classes, a new samples from the first two classes should be predicted as not novel yet new samples from the third, unknown, class should be predicted as novel. Novelty prediction is more challenging than classification where all classes are known because the decision boundary between unknown and known classes are hard to define.

We implemented novelty prediction models in the AIVIA 6, a software for microscopy image visualization, morphometric analysis and phenotype classification / discovery. Three models are implemented:

- ❑ One Class Random Forest (OCRF) [1,2]
- ❑ One Class Support Vector Machines (OCSVM) [3]
- ❑ Kernel Null Folley-Sammon (KNFS)[4]

Novelty Prediction Model Learning

The novelty prediction model is learned using labeled data. The data are stored as feature vectors where each feature is a measurement. The labels are from known classes. We will explain the learning for each of the three models below.

One Class Random Forest (OCRF)

The OCRF model creates a novelty predictor for each class. For a sample to be considered novel, the novelty predictors of all different classes must predict the sample as novel. The novelty predictor for a class is constructed from trees of random subsets of the features.

From the random subset of features, histograms are generated for each feature using the training data. The inverse of the histogram is then used to create a cumulative distribution function and generate synthetic data representing unknown classes (see Fig. 2). The tree is then taught using the original training data and the synthetic data. To avoid the impacts of noisy features, a subset of discriminate features could be used for novelty prediction.

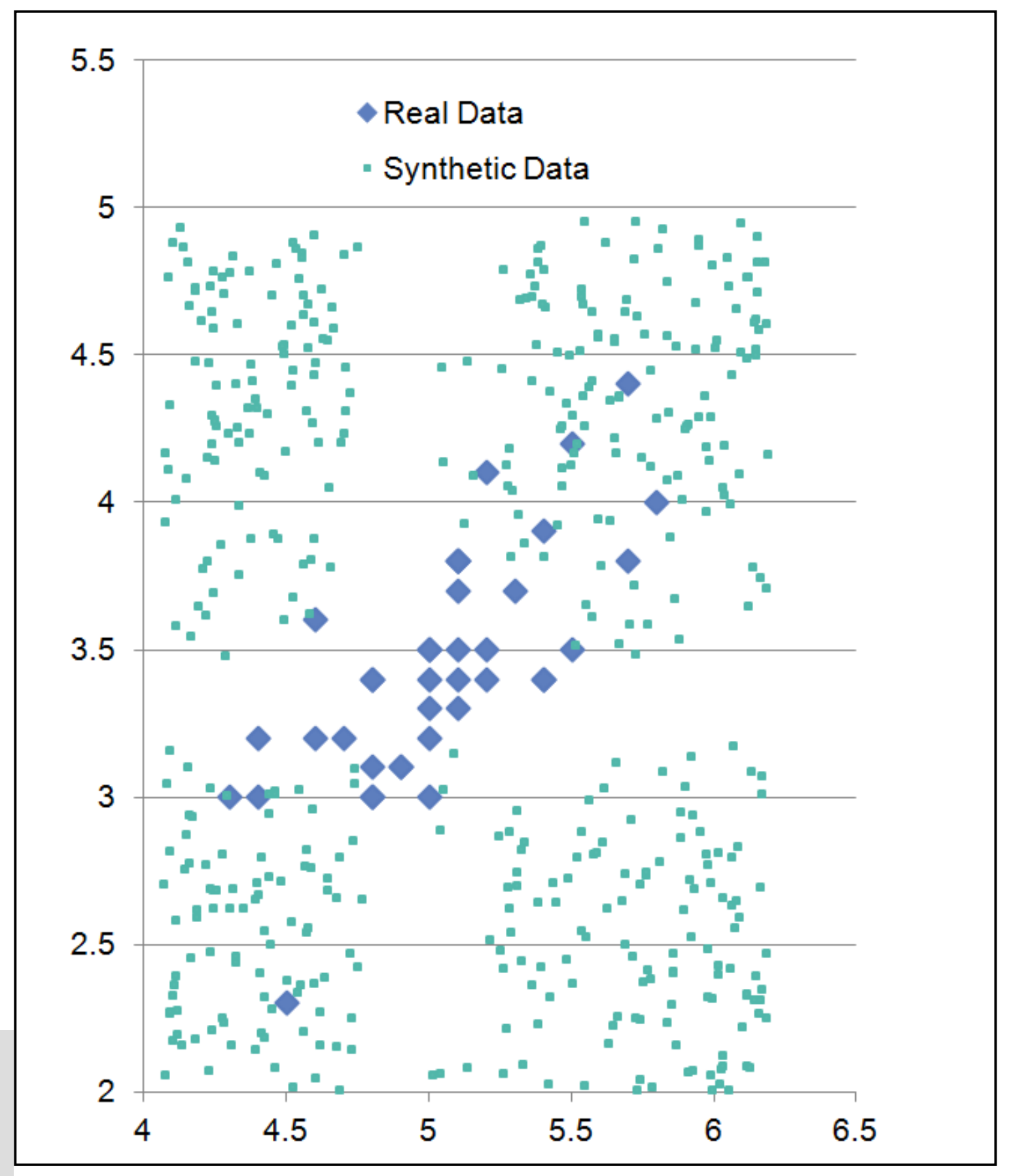


Fig 2. Illustrative two feature example of synthetic data generated from inverse of histograms

One Class Support Vector Machines (OCSVM)

This is a standard model commonly used[3]. The model works by examining each class individually like OCRF. Each SVM returns a value a likelihood value when given a sample, positive values indicate the sample belongs to the learning data.

Kernel Null Folley-Sammon (KNFS)

This model learning applies a null space method to map all training samples of one class to a single point[4]. New samples are mapped to the null space and their distance to the points of each class are used to determine their novelty. If a new sample is far from all of the points it is considered novel. How far a sample must be to be considered novel is determined automatically with the learning data.

Study Materials and Methods

The novelty detection tool is applied to neuronal type discoveries. It is tested on several data sets to validate the tool and to help direct improvement and compare strengths of different models. The test data sets include a human set (101 neurons, 4 types), a mouse set (287 neurons, 6 types) and a rat set (354 neurons, 6 types). 71 neuronal features were extracted by AIVIA for the testing. The tests were performed by training different number of types and evaluating the untaught types as novel types. 80% of the data were used for training and 20% were used for testing. The test metrics are detection accuracy(A): correct detection/ total detection, precision (P): accuracy of samples detected as novel, and recall (R): accuracy of novel samples detected as novel. The precision is the most important metric.

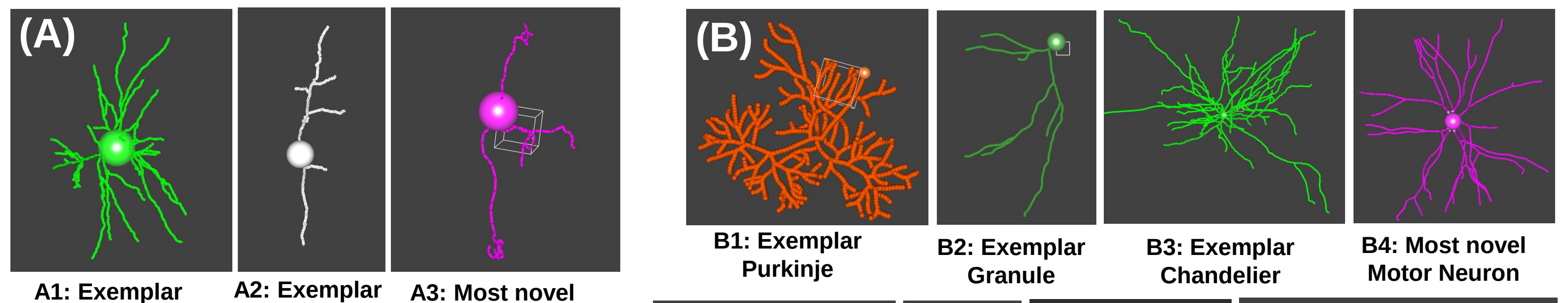
Results

(A)	Human					
	1 Type			2 Types		
Model	A	P	R	A	P	R
OCRF	0.52	0.88	0.34	0.68	0.75	0.25
OCSVM	0.77	0.75	1	0.54	0.46	1
KNFS	0.79	0.84	0.86	0.55	0.45	0.83

(B)	Mouse																	
	1 Type			2 Types			3 Types			4 Types			5 Types			6 Types		
Model	A	P	R	A	P	R	A	P	R	A	P	R	A	P	R	A	P	R
OCRF	0.79	0.97	0.77	0.67	0.88	0.59	0.66	0.77	0.45	0.7	0.58	0.32	0.79	0.33	0.27	0.9	NA	NA
OCSVM	0.91	0.91	0.98	0.83	0.81	0.97	0.75	0.68	0.95	0.69	0.52	0.93	0.64	0.3	0.92	0.59	NA	NA
KNFS	0.86	0.92	0.90	0.58	0.83	0.46	0.72	0.67	0.87	0.61	0.46	0.94	0.48	0.24	0.97	0.33	NA	NA

(C)	Rat																	
	1 Type			2 Types			3 Types			4 Types			5 Types			6 Types		
Model	A	P	R	A	P	R	A	P	R	A	P	R	A	P	R	A	P	R
OCRF	0.7	0.94	0.69	0.6	0.85	0.51	0.59	0.71	0.36	0.67	0.56	0.32	0.75	0.36	0.29	0.91	0.25	0.66
OCSVM	0.85	0.9	0.92	0.73	0.78	0.85	0.65	0.63	0.79	0.6	0.46	0.73	0.57	0.28	0.7	0.58	0.08	1
KNFS	0.57	0.944	0.52	0.32	0	0	0.50	0.56	0.16	0.63	0.58	0.36	0.75	0.35	0.31	0.948	0	0

Table 1. Test metrics results for (A) human set, (B) mouse set and (C) rat set



A1: Exemplar Pyramidal

A2: Exemplar Spindle

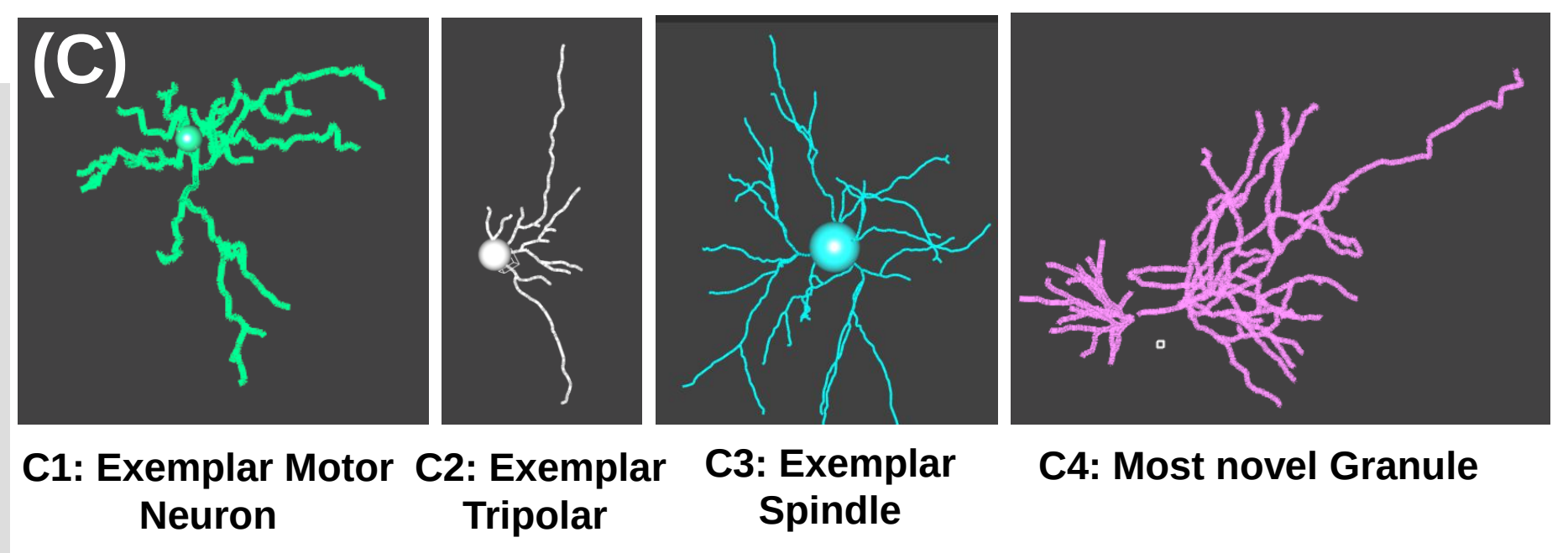
A3: Most novel Granule

B1: Exemplar Purkinje

B2: Exemplar Granule

B3: Exemplar Chandelier

B4: Most novel Motor Neuron



C1: Exemplar Motor Neuron

C2: Exemplar Tripolar

C3: Exemplar Spindle

C4: Most novel Granule

Fig 3. Novel samples returned in test data sets (A) in human set, Pyramidal and Spindle were learned and Granule and Stellate are considered novel types. A1 and A2 are exemplar Pyramidal and Spindle samples. A3 is a Granule sample that has the highest novelty score. (B) in mouse set, Purkinje, Granule and Chandelier were learned and Motor Neuron, Tripolar and Pyramidal are considered novel types. B1, B2 and B3 are exemplar Purkinje, Granule and Chandelier samples. B4 is a Motor Neuron sample that has the highest novelty score. (C) in rat set, Motor Neuron, Tripolar and Spindle were learned and Purkinje, Granule, Pyramidal and Chandelier are considered novel types. C1, C2 and C3 are exemplar Motor Neuron, Tripolar and Spindle samples. C4 is a Granule sample that has the highest novelty score.

Discussions and Conclusion

The preliminary results indicate that OCRF is a good machine learning method for the neuron applications. The results also provide a promising direction of automated novel neuronal type discoveries using machine learning. We will further test and improve the methods using additional data sets including dendritic spine classifications.

References

1. **Chesner D´esir, Simon Bernard, Caroline Petitjean, Heutte Laurent.** One class random forests. *Pattern Recognition, Elsevier*, 2013, 46, pp.3490-3506
2. **Shi, Zhongkui et al.**, An Outlier Generation Approach for One-Class Random Forests: An example in One-Class Classification of Remote Sensing Imagery, *IEEE International Geoscience and Remote Sensing Symposium*, 2016
3. **Mennatallah Amer et al,** Enhancing One-class Support Vector Machines for Unsupervised Anomaly Detection, *ODD'13, August 11th*, 2013
4. **Bodesheim, Paul et al.**, Kernel Null Space Methods for Novelty Detection, *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition***,** 2013

Acknowledgments

This research was supported in part by grant no. 4R44NS097094-02 from the NINDS and grant no. 5R43MH100780-02 from the NIMH