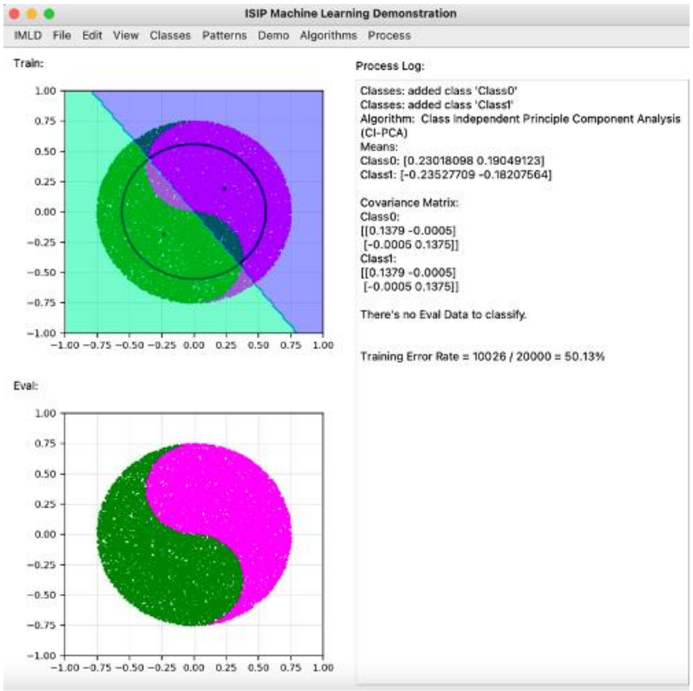
IMLD: A Python-Based Interactive Machine Learning Demonstration

T. Cap[1], A. Kreitzer[2], M. Miranda[2], D. Vadimsky[2] and J. Picone[1]

1. Neural Engineering Data Consortium, Temple University, Philadelphia, Pennsylvania, USA
2. College of Engineering, Temple University, Philadelphia, Pennsylvania, USA
{thao.cap, aaron.kreitzer, matthew.miranda, dakota.vadimsky, picone}@temple.edu

The related fields of machine learning and pattern recognition have enjoyed significant success in recent years due to the impact of deep learning algorithms [1]. Pattern recognition is the automatic recognition of regularities or trends in data. Machine learning, a closely related field that has evolved considerably in the past two decades, refers to the ability of a machine to learn and adapt to data, improving a system's ability to detect patterns and perform inference. These methods are ubiquitous in engineering today impacting diverse fields including signal and image processing, human language technology, bioinformatics, and finance. The ISIP Machine Learning Demo (IMLD) is a tool used to introduce the basics of machine learning using a highly interactive environment in which users can easily visualize the performance of an algorithm. IMLD was first developed as a Java applet in the late-1990's when Java applets were envisioned as the future of interactive computing [2] and there was an emphasis on web-based educational tools [3]. The Institute for Signal and Information Processing, then located at Mississippi State University, developed a suite of interactive demos to teach important concepts in signal processing [4].

However, today, due primarily to security issues, Java applets have fallen in disfavor and are no longer being supported. Instead, it makes more sense to deliver such applications in Python, where the bulk of machine learning research is being done. This allows the application to integrate a wider range of algorithms. Hence, a major focus of this work, being conducted at the Neural Engineering Data Consortium at Temple University, is the conversion of this application from Java to Python. However, as we will discuss in this work, delivering a complex interactive application in Python is not as easy as one might think. Major concerns include the stability of the graphical programming languages available and web accessibility.



```
Classes: added class 'Class0'
Classes: added class 'Class1'
Algorithm:  Class Independent Principle Component Analysis
(CI-PCA)
Means:
Class0: [0.23018098 0.19049123]
Class1: [-0.23527709 -0.18207564]

Covariance Matrix:
Class0:
[[0.1379 -0.0005]
 [-0.0005 0.1375]]
Class1:
[[0.1379 -0.0005]
 [-0.0005 0.1375]]

There's no Eval Data to classify.


Training Error Rate = 10026 / 20000 = 50.13%
```

Figure 1. The IMLD user interface

A typical screenshot of the user interface is shown in Figure 1. IMLD allows users to create unique two-dimensional data sets that can be easily visualized. Users can choose from a wide selection of predefined data sets such as multivariate Gaussian data, draw custom data sets, or create a combination of the two. Data can be saved to a file or uploaded, allowing users to experiment with their own data sets and use IMLD as a reference implementation. A set of standard algorithms are available including fully supervised approaches such as Principal Components Analysis [5], unsupervised approaches such as K-MEANS clustering [6], and popular neural network algorithms such as multilayer perceptrons [7]. The number of classes, classification modules, and other key parameters of the data generation are user defined. A dialog box is included that displays step-by-step computations for the algorithm. Users can step through the algorithms or run them in their entirety. Decision surfaces are rendered, and error rates are computed on the training and evaluation sets.

An overview of IMLD's software architecture is shown in Figure 2. IMLD is dependent on a variety of third-party libraries. Many third-party libraries were considered based on the wide variety of components required to generate the applet. The following libraries were decided based on the specific needs of IMLD. PyQT5 is the framework used to create the front-end design of the application, chosen for its ability to handle large quantities of data. The backend uses NumPy for calculations, Sklearn/Scipy for implementing algorithms, and Matplotlib for interactive graphs. The software is organized into three major components: the graphical user interface (GUI), data handler, and machine learning algorithms.

The first part of the architecture is the GUI, where three key modules handle the window design, events, and parameters. The window module manages all the front-end design for the application. The module includes the menu bar, input/output graphical displays, and output log. The menu bar holds all the functionality for adding/deleting classes, choosing an algorithm, and giving users the option to either import, export, or create test data. The two graphical displays allow users to click and drag their mouse to create either point-like data or Gaussian-like data. The last section for this module is the output log which records all interactions between users and the application, i.e., adding a class or choosing an algorithm. The event module handles execution for the buttons displayed on the GUI. Lastly, the parameter module handles all secondary user information and allows users to configure the data generators.

Another part of the application is the data handler. One of the features that the Python version of IMLD enhances from the Pattern Recognition Applet written in Java is the software that allows users to generate data. The tool provides two underlying mechanisms of the generation of the data: generation by drawing points or generation using a functional form such as a Gaussian distribution. The training and evaluation sets are generated independently in separate windows so that a wide range of machine learning scenarios

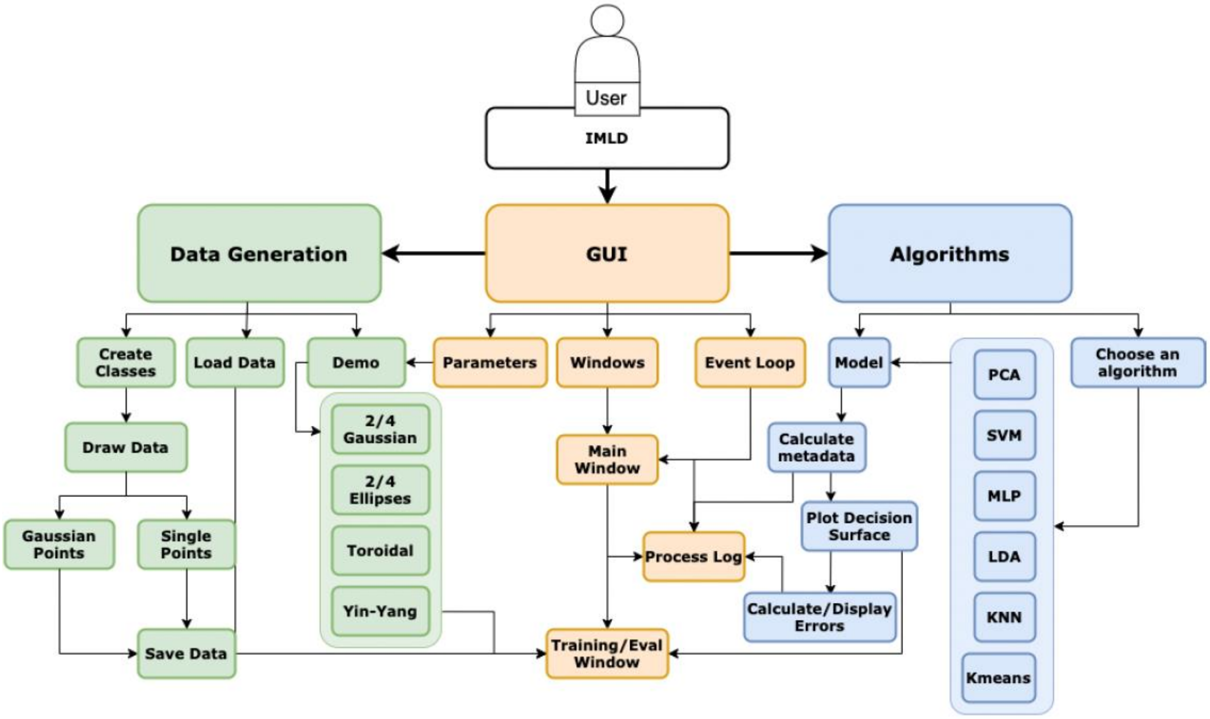


Figure 2. The IMLD software architecture

(e.g., generalization) can be evaluated. A dictionary stores the data where the key is the name of the class added, and the value is an array that holds the class color and an array of x and y coordinates.

Newly created data can be exported into a CSV file where comment fields are used to hold class names and class color, while each line holds the x and y coordinates. This is a simple format we have used for our machine learning class that makes it easy for novice programmers to interface to the data. It was also a preferred format based on a survey we conducted with the user community. The application also allows CSV files to be imported for further study and modification.

IMLD supports the configuration of data sets and analysis parameters through drop-down menus. Class, Color, and Scale are all tools that allow users to uniquely configure their data. While adding a new class, users are prompted to add a name, and then choose from over the 150 colors options. Once set, users can now draw data points. Users can add a new class at any time by navigating to the Classes menu and adding another class. The Scale of the data can also be managed through the Classes tool while in the Classes menu. Users can at any time choose which of the added classes they would like to delete.

Users can also choose from a selection of prestored, or canned, demos that include classic machine learning datasets such as overlapping Gaussian distributions, toroidally shaped distributions that cannot be classified with a linear classifier, and a yin-yang distribution that requires a nonlinear decision surface. Finally, after creating the data, users can choose the algorithm under the algorithm section and select the 'run' or 'run by step' options to start classification.

Currently, IMLD does not allow users to insert their own algorithms. In the future, we plan to implement a method by which users can do this. This feature would significantly enhance the educational capacity of IMLD, as users will not be restricted to the algorithms offered only by the tool.

IMLD is an educational tool with the ability to walk users through a step-by-step process to visualize various machine learning algorithms. It has been used by a machine learning class we have been teaching since the late 1990's (*https://www.isip.piconepress.com/courses/temple/ece_8527/*). It is easily installed on a platform that includes Anaconda v3 and Python v3.7 or later. The source code is available from the course web site at: *https://www.isip.piconepress.com/courses/temple/ece_8527/resources/imld/*. A detailed user manual demonstrating use of the tool and instructional videos are also available. A demonstration will be provided at the symposium.

ACKNOWLEDGMENTS

Research reported in this publication was most recently supported by the National Science Foundation's Industrial Innovation and Partnerships (IIP) Research Experience for Undergraduates award number 1827565. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the official views of any of these organizations. Open source libraries that were used to develop the IMLD are: NumPy v1.15.4, PyQT5 v5.13.1, SkLearn v0.20.0, Scipy v1.1.0, and Matplotlib v3.0.2.

REFERENCES

[1] W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, and K.-R. Müller, "Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications," *Proceedings of the IEEE*, vol. 109, no. 3, pp. 247–278, 2021. *https://ieeexplore.ieee.org/document/9369420*.

[2] D. May and J. Picone, "The ISIP Pattern Recognition Applet," Institute for Signal and Information Processing, College of Engineering, Mississippi State University, 2002. [Online]. Available: *https://www.isip.piconepress.com/projects/speech/software/demonstrations/applets/util/pattern_recognition/current/*. [Accessed: 05-Jul-2021].

[3] J. Picone, R. Duncan, and J. Hamaker, “Internet-Accessible Speech Recognition Technology,” in O’Reilly Open Source Convention, 2001. *http://www.isip.piconepress.com/publications/conference_presentations/2001/oscon/software/*.

[4] J. Picone, “Internet-Accessible Technology Demonstrations,” *Institute for Signal and Information Processing, College of Engineering, Mississippi State University*, 2000. [Online]. Available: *https://www.isip.piconepress.com/projects/speech/software/demonstrations/*. [Accessed: 05-Jul-2021].

[5] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. New York City, New York, USA: John Wiley & Sons, Inc, 2001. *https://www.wiley.com/en-us/Pattern+Classification%2C+2nd+ Edition-p-9780471056690*.

[6] K. Mallapragada, R. Jin, and A. Jain, “Non-parametric Mixture Models for Clustering BT - Structural, Syntactic, and Statistical Pattern Recognition,” E. R. Hancock, R. C. Wilson, T. Windeatt, I. Ulusoy, and F. Escolano, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 334–343. *https://link.springer.com/chapter/10.1007/978-3-642-14980-1_32*.

[7] C. Bishop, *Pattern Recognition and Machine Learning*, 2nd ed. New York, New York, USA: Springer, 2011. *https://www.springer.com/us/book/9780387310732*.

ISIP MACHINE LEARNING DEMO

ABSTRACT

Machine learning (ML) is a field that has evolved considerably in the past two decades. ML refers to the ability of a machine to learn and adapt to data, improving a system's ability to detect patterns and perform inference. These methods are ubiquitous in engineering today impacting diverse fields including signal and image processing, human language technology, bioinformatics, and finance.

The ISIP Machine Learning Demo (IMLD) is a tool used to introduce the basics of ML using a highly interactive environment in which users can easily visualize the performance of an algorithm. IMLD enhances ML education by providing an environment in which students can easily visualize data and algorithm performance.

FEATURES

Generate Data

Users can select from a number of well-known data sets (e.g. overlapping Gaussian distributions), or can create custom data sets using freehand drawing tools.

Import/Export Data

Users can also evaluate data sets of interest by importing them into the tool using a CSV file format. Data can be exported as well, making the tool an ideal way to generate unique and interesting data sets.

Customize Data

Users can augment existing data sets using freehand drawing tools, and can also customize the parameters of several available data generators. Data generators allow users to create historically important data sets.

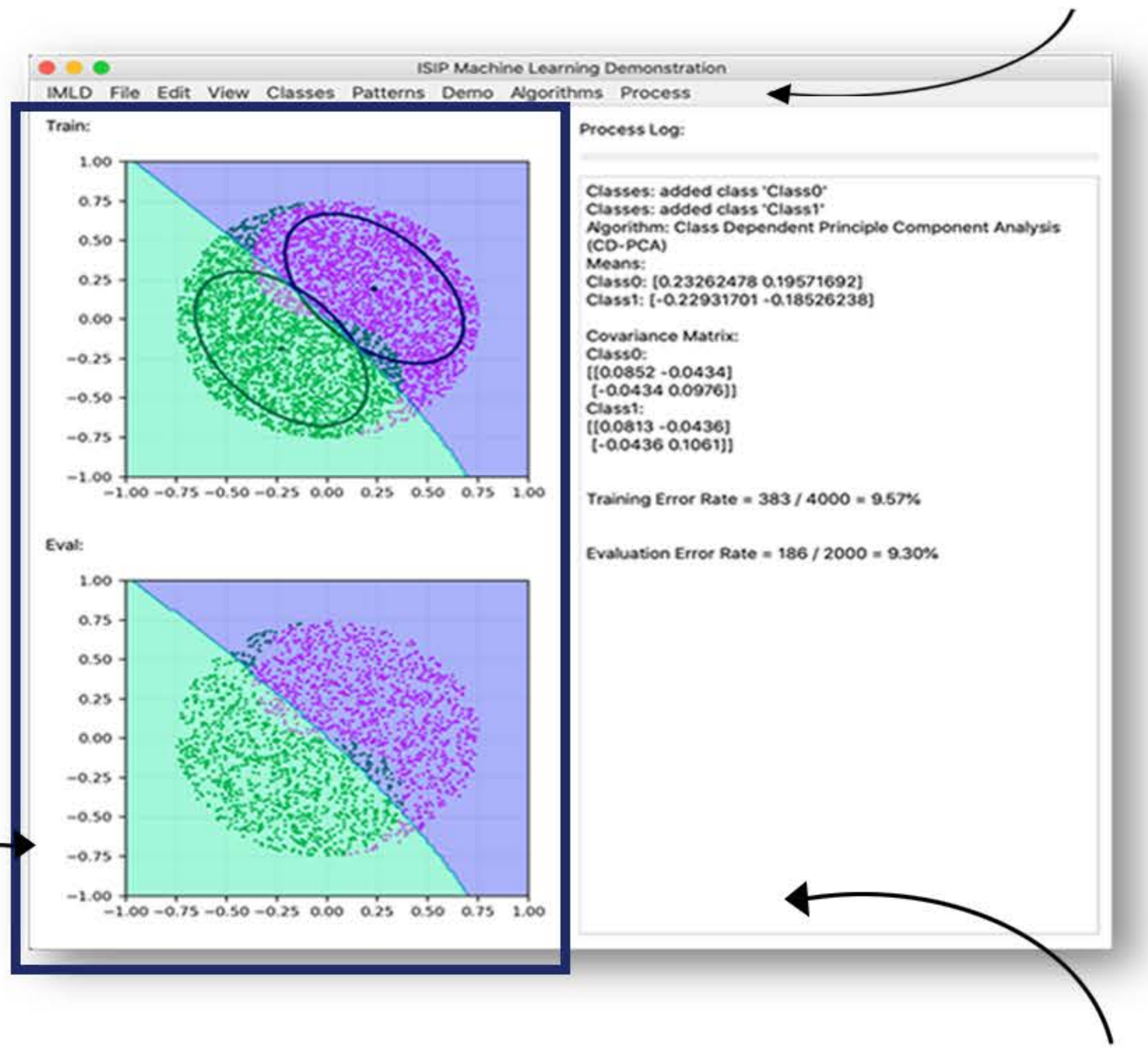
Analyze Data

Users can apply popular ML algorithms to their data. Algorithms can be trained on the data appearing in the "Train" window, and evaluated on unseen data appearing in the "Eval" window. Key parameters of each algorithm can be adjusted and manually optimized.

USER INTERFACE

Menu Ribbon

The Menu Ribbon at the top of the IMLD main window contains the basic functions that allow users to create and manipulate data. For example, under the File option, users can load and store data. Users can add and delete classes, change the parameters of the data generators, select prestored demos, and, of course, select specific algorithms.



```
Classes: added class 'Class0'
Classes: added class 'Class1'
Algorithm: Class Dependent Principle Component Analysis
(CD-PCA)
Means:
Class0: [0.23262478 0.19571692]
Class1: [-0.22931701 -0.18526238]

Covariance Matrix:
Class0:
[[0.0852 -0.0434]
 [-0.0434 0.0976]]
Class1:
[[0.0813 -0.0436]
 [-0.0436 0.1061]]


Training Error Rate = 383 / 4000 = 9.57%


Evaluation Error Rate = 186 / 2000 = 9.30%
```

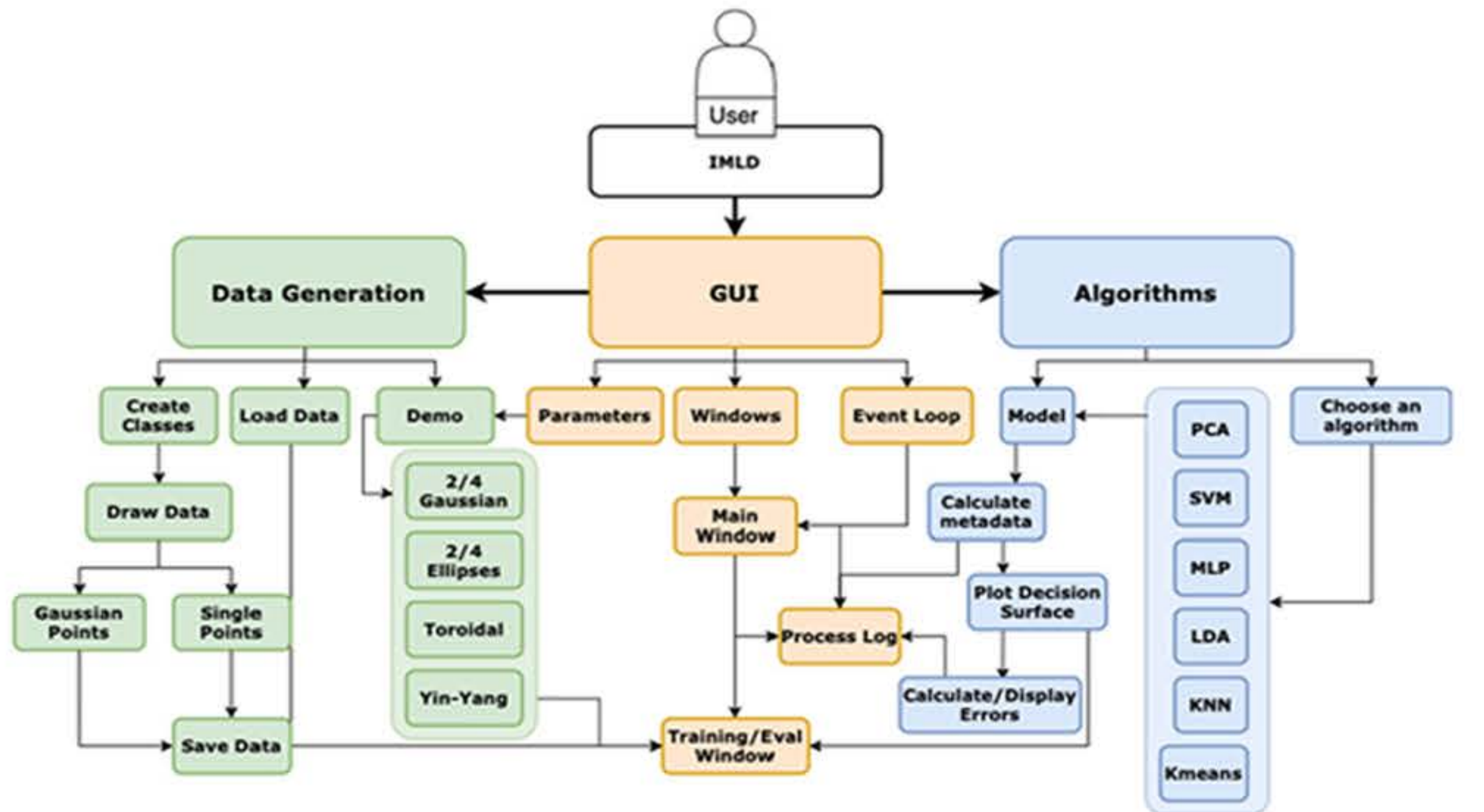
Train and Eval Window

The data input windows are where users can load data files, generate demonstration data sets, and draw their own data using a freehand drawing tools or using a Gaussian-like paint brush. Users can adjust the limits of these windows separately. These window limits will be saved with the data files so that future adjustments are automatically made upon loading.

Process Log

The Process Log keeps the user updated on their own actions within IMLD as well as the actions taken by the program once a classification process is started. The log will display the steps of the selected algorithm, the error rate of the classification process, and warning messages when the user performs actions that are either out of order or disallowed by the program.

SOFTWARE ARCHITECTURE



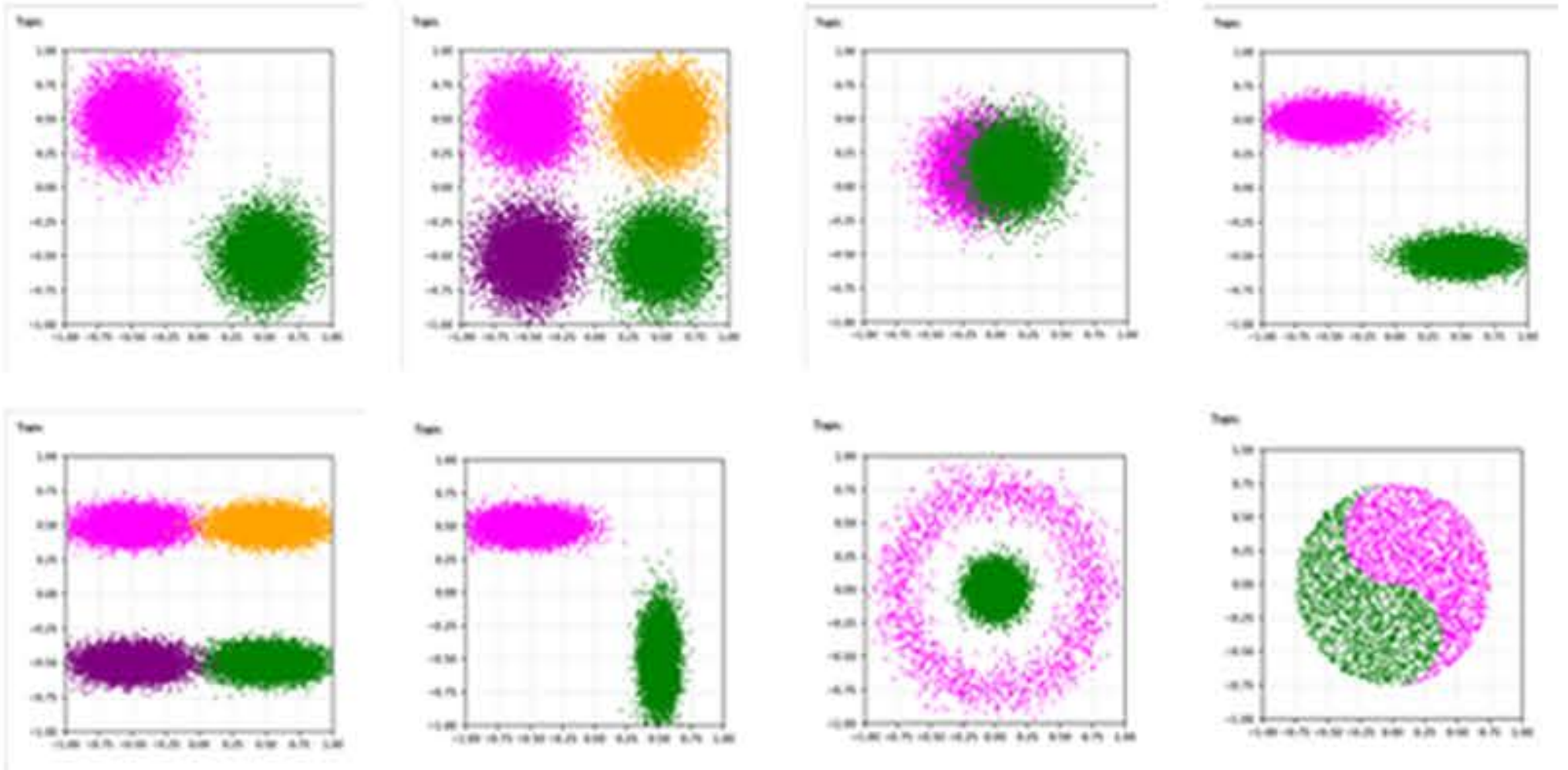
A user's typical interaction with IMLD is shown in the flow chart above. The program is specifically engineered to allow users to add their own demos and algorithms.

The model module takes in four different inputs which consist of the algorithm the user has chosen, the input display, the output display, and the process log. The model API is structured in a way to accept any algorithm and store the results of the prediction, map out the correct decision surface and present both training and evaluation error. The only requirement the model class has for its algorithms is that it contains certain key functions such as prediction, classification, and data extraction.

The default behavior of the data generators can be configured from external parameter files,

These design decisions allow educators to modify the program to suit their specific needs, and allow students to test various algorithms on their own data sets, thereby replicating important industry baselines. This alleviates the need to write code and enables efficient rapid prototyping.

DEMONSTRATION DATA SETS



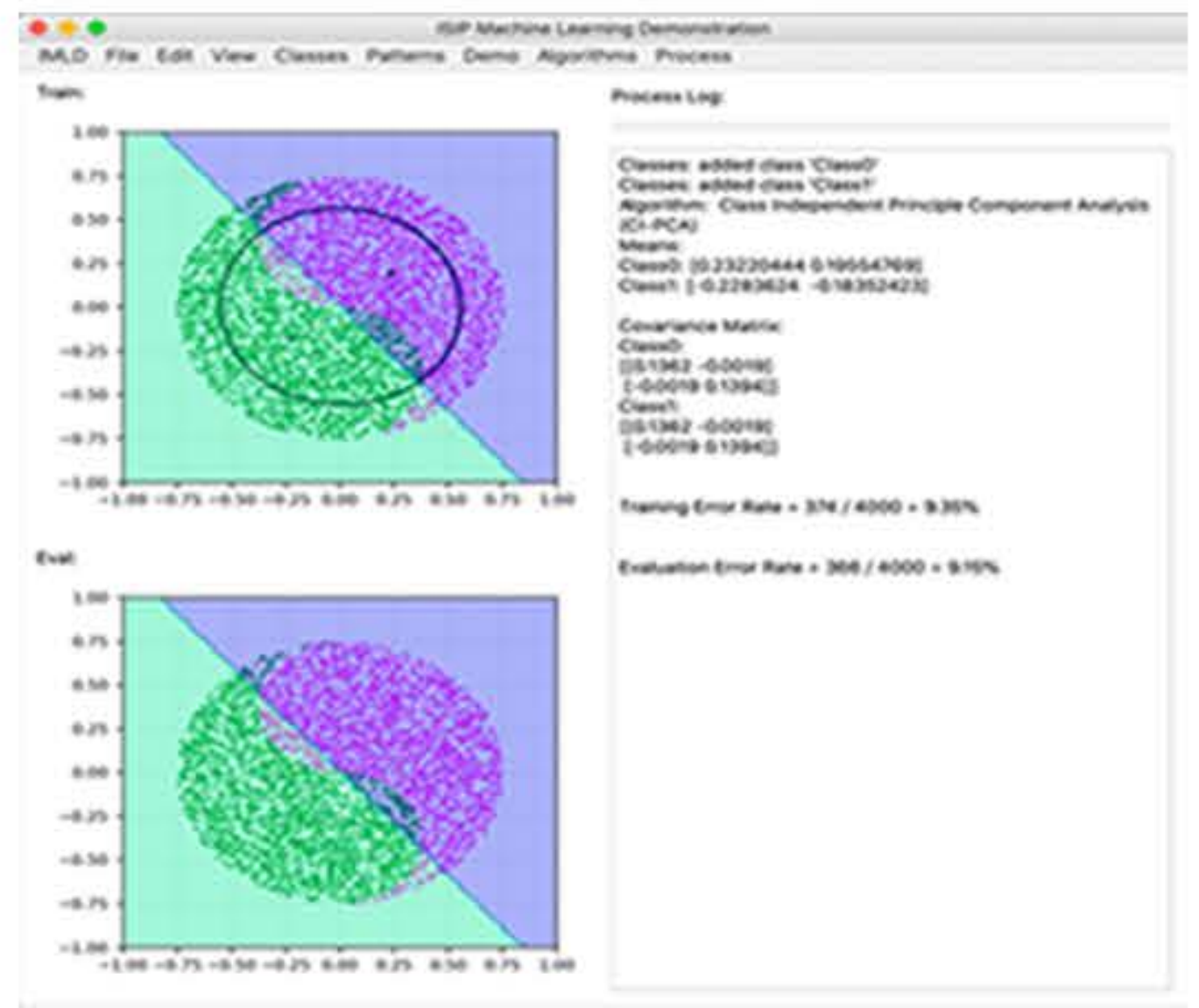
IMLD supports the generation of several preprogrammed data sets that reflect popular statistical models in the field of ML (e.g., a two-class problem involving data shaped as a toroid). These data sets require advanced ML algorithms capable of modeling nonlinear decision surfaces.

These data sets can be adjusted by the user through dialog boxes that allow control of basic statistical parameters (e.g., mean and covariance) and the degree of overlap between the classes. The default parameters can be adjusted by editing an external parameter file.

ALGORITHMS

IMLD consolidates all of its available alogorithms into a class named "imld_model". This class loads data into the training and evaluation windows, and executes a user-selected algorithm on the corresponding data.

IMLD supports most Python-based machine learning algorithms. Users can easily incorporate new algorithms by following a relatively simple API that all algorithms use.


ISIP Machine Learning Demonstration
IMLD File Edit View Classes Patterns Demo Algorithms Process
Train:
Eval:
Process Log:

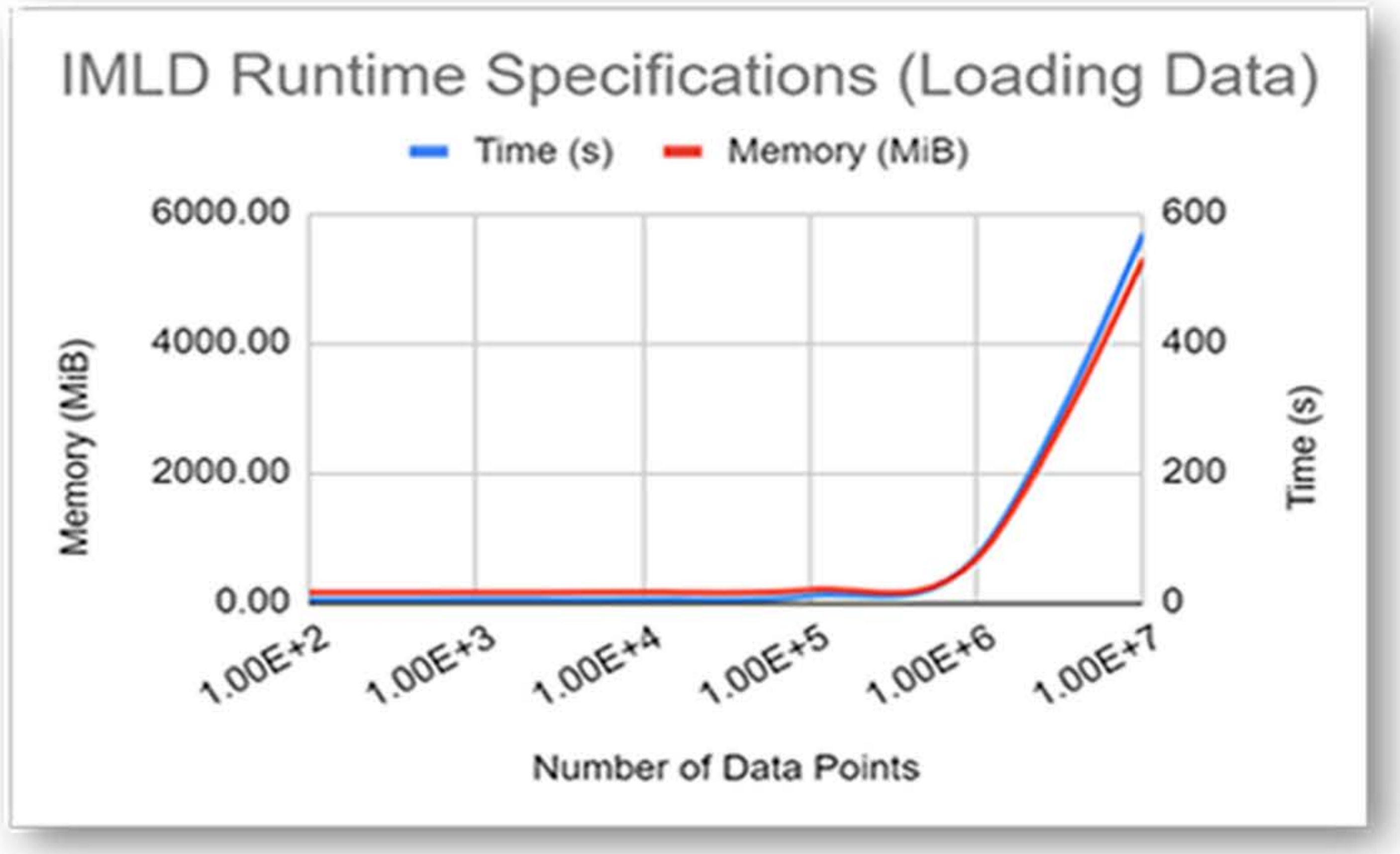
```
Classes: added class 'Class0'
Classes: added class 'Class1'
Algorithm: Class Independent Principle Component Analysis
(CI-PCA)
Means:
Class0: [0.23220444 0.19554769]
Class1: [-0.2283624 -0.18352423]

Covariance Matrix:
Class0:
[[0.1362 -0.0019]
 [-0.0019 0.1394]]
Class1:
[[0.1362 -0.0019]
 [-0.0019 0.1394]]


Training Error Rate = 374 / 4000 = 9.35%


Evaluation Error Rate = 366 / 4000 = 9.15%
```

PERFORMANCE

For small data sets, such as those provided in the prestored demos, IMLD runs instantly.

IMLD has also been designed to handle large data sets and can comfortably handle up to 10M data points using slightly under 6 GBytes of memory and 10 mins of CPU time.

IMLD has been tested on Windows, Mac and Linux platforms and is designed to be on any platform supporting Python and PyQT.



SUMMARY

IMLD is an educational tool with the ability to walk users through a step-by-step process to visualize various machine learning algorithms. It has been used by a machine learning class we have been teaching since the late 1990's. It is easily installed on a platform that includes Anaconda v3 and Python v3.7, The application is available at www.isip.piconepress.com/courses/temple/ece_8527/resources/imld. A detailed user manual demonstrating use of the tool and instructional videos are also available.

ACKNOWLEDGEMENTS

Research reported in this publication was most recently supported by the National Science Foundation’sIndustrial Innovation and Partnerships (IIP) Research Experience for Undergraduates award number 1827565. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the official views of any of these organizations. Open Source libraries that were used to develop the IMLD are: NumPy v1.15.4, PyQT5 v5.13.1, SkLearn v0.20.0 , Scipy v1.1.0, and Matplotlib v3.0.2.

REFERENCES

D. May and J. Picone, “The ISIP Pattern Recognition Applet,” Institute for Signal and Information Processing, College of Engineering, Mississippi State University, 2002. [Online]. Available: https://www.isip.piconepress.com/projects/speech/software/demonstrations/applets/util/pattern_recognition/current/. [Accessed: 05-Jul-2021].

J. Picone, R. Duncan, and J. Hamaker, “Internet-Accessible Speech Recognition Technology,” in O’Reilly Open Source Convention, 2001.

J. Picone, “Internet-Accessible Technology Demonstrations,” Institute for Signal and Information Processing, College of Engineering, Mississippi State University, 2000. [Online]. Available: https://www.isip.piconepress.com/projects/speech/software/demonstrations/. [Accessed: 05-Jul-2021].

T. CAP, A. KREITZER, M. MIRANDA, D. VADIMSKY, AND J. PICONE