12-15-2021

Building the Path to Early Alzheimer's Prediction Using Machine Learning

Kincaid Rowbotham

Ling Li
University of North Dakota, ling.li.1@UND.edu

Xusheng Wang
University of North Dakota, xusheng.wang@UND.edu

Building the Path to Early Alzheimer's Prediction Using Machine Learning

Kincaid Rowbotham , Ling Li , Xusheng Wang

Biology Department, Arts and Sciences, University of North Dakota, Grand Forks, ND 58202, United States

Introduction

Alzheimer's Disease (**AD**) is the most common form of dementia and one of the most prominent challenges of precision healthcare is early identification of AD. To combat this issue, we plan to implement a method to use machine learning and deep learning to predict Alzheimer's Disease. Through the use of post-mortum frontal cortex proteomic expression data, we have constructed a strong baseline for this type of work into the future.

While Learning methods for AD have already been developed using MRI and RNA in blood this is the first use of one using tissue data. Given the expensiveness of performing MRI and the lack of data for AD RNA in blood, this model is less expensive and has more data to train with, respectively. These optimizations will stand during transitions to other -omics data, types of tissue, and time of affliction.

Methods

- Combinations of Feature Selection Methods and Machine/Deep Learning Methods were used to find the best way to identify AD.
- **Feature Selection Methods**
  - **K-Best** (Control) – Chooses top proteins only
  - **MRMR** – Groups proteins first and chooses the best among a group
- **Learning Methods**
  - Artificial Neural Network (**ANNC**), Gaussian Naive Bayes (**GaussNB**), Gradient Boosting Machine (**GBM**), K-Nearest Neighbors (**KNear**), Random Forest (**Randforest**), Support Vector Machine (**SVM**)

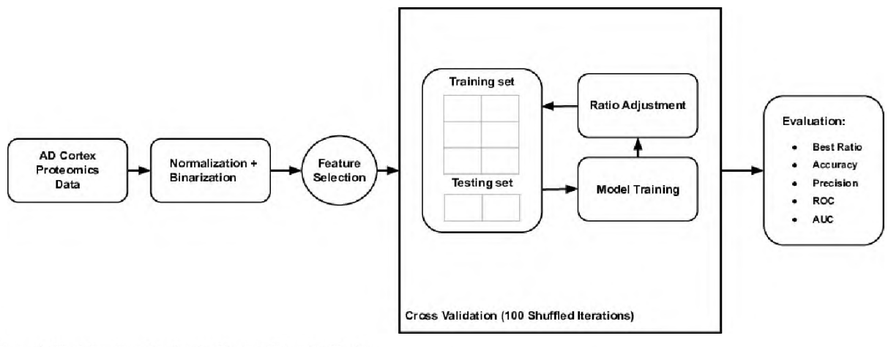


Figure 1. Workflow Visualization for Feature+Learning Testing.

Optimization of Feature+Learning method is performed with 100 shuffles and then tested using two-steps of ROC-Curves (see Figure 2.). All Machine Learning and Feature Selection methods were created using Sci-Kit Learn in Python 3.8.3. The Artificial Binary Neural Network Classifier (**ANNC**) was created using Pytorch in Python 3.8.3.

Results

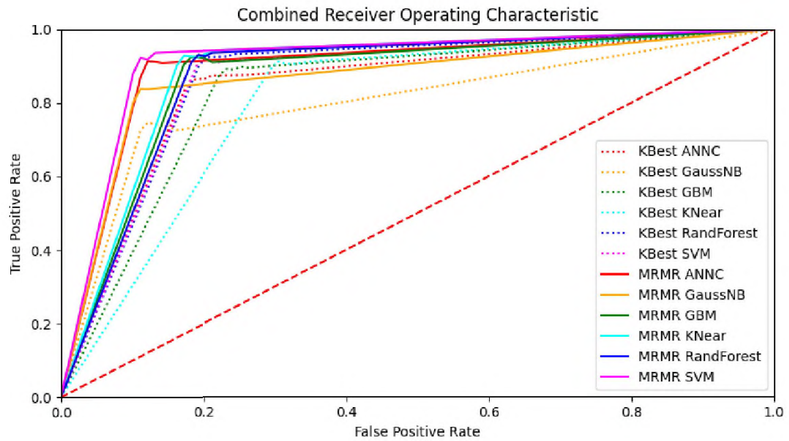
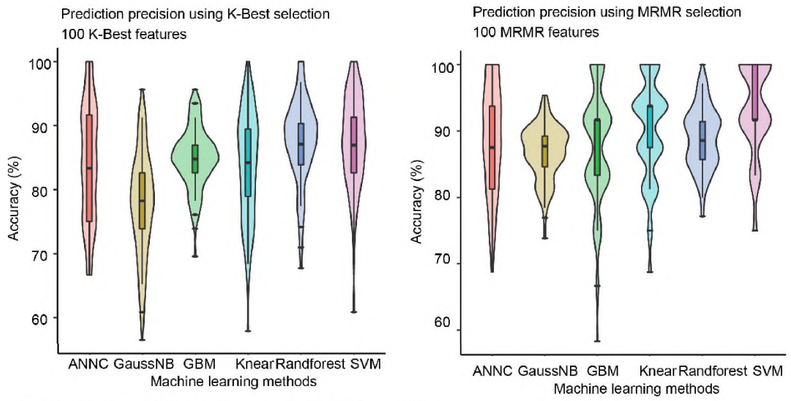


Figure 2. ROC Curve for Selection+Learning Combinations Across Testing-Training ratios

Second-stage ROC Curve, training-testing ratios are used for each combination. Specific ratios used are specified in Column Four in Table 1. The first iteration of ROC Curve goes by ratios of 50% testing to 10% testing in steps of 10% (I.e 50%,40%,30%,20%,10%.) The second stage uses the most performant value of each combination from first-stage and then increments by 2% twice in both directions. For example, if the value is 30% for MRMR+GaussNB then the values for the second-stage ROC Curve Testing ratios are 26%,28%,30%,32%,34%.



Figures 3 Accuracy Profile across Selection+Learning Combinations

Violin and Box plots representing the overall profile of each combination's accuracy outcomes split into K-Best Selection (Left) and MRMR Selection (Right). Each profile is across 100 runs using a shuffled samples for each run.

Discussion

- Despite high accuracy scores some models might have been fit too closely to data and may not perform well cross-cohort.
- Taking into account the risk of over-fitting, precision, and accuracy the most optimal combination from this group is MRMR+RF.
- Inclusion of a sophisticated in-house, multi-omics feature selection method could lead to a better selection of proteins to train the Learning Models on.
- Other parameters such as amount of proteins used in the sets, along with further refinement of the neural network used could lead to different results.

Table 1. Accuracy Testing Results

Optimal Avg Accuracy obtained was obtained purely from performance alone. Evaluations of over-fitting is not evaluated at this stage.

Selection+Learning Combination	Optimal Avg Accuracy at Percent Training	Second Stage AUC	Standard Deviation
K-Best+ANNC	84.25% at 6% testing	84.19%	9.54%
K-Best+GaussNB	77.77% at 12% testing	80.06%	8.18%
K-Best+GBM	84.89% at 24% testing	83.20%	4.74%
K-Best+KNear	83.89% at 10% testing	80.35%	8.42%
K-Best+RF	87.61% at 16% training	86.47%	6.53%
K-Best+SVM	87.35% at 12% training	85.95%	7.65%
MRMR+ANNC	88.89% at 8% testing	89.5%	8.17%
MRMR+GaussNB	86.85% at 34% testing	86.47%	4.01%
MRMR+GBM	87.75% at 6% testing	86.72%	8.82%
MRMR+KNear	90.13% at 8% testing	88.47%	7.49%
MRMR+RF	88.94% at 18% testing	87.35%	4.66%
MRMR+SVM	93.25% at 6% testing	91.32%	6.88%

Acknowledgments

Funding for this project was supplied by ND EPSCoR STEM (UND0025726), the American Society for Pharmacology & Experimental Therapeutics (ASPET) SURF Program, the Chair of the Department of Biomedical Sciences, the Division of Research & Economic Development at the University of North Dakota, an Institutional Development Award (IDeA) from the National Institute of General Medical Sciences of the National Institutes of Health under grant number P20GM103442, and the Dean of the University of North Dakota School of Medicine & Health Sciences.