DNA Methylation Data to Predict Suicidal and Non-Suicidal Deaths in *Homo Sapiens*

A Machine Learning Approach

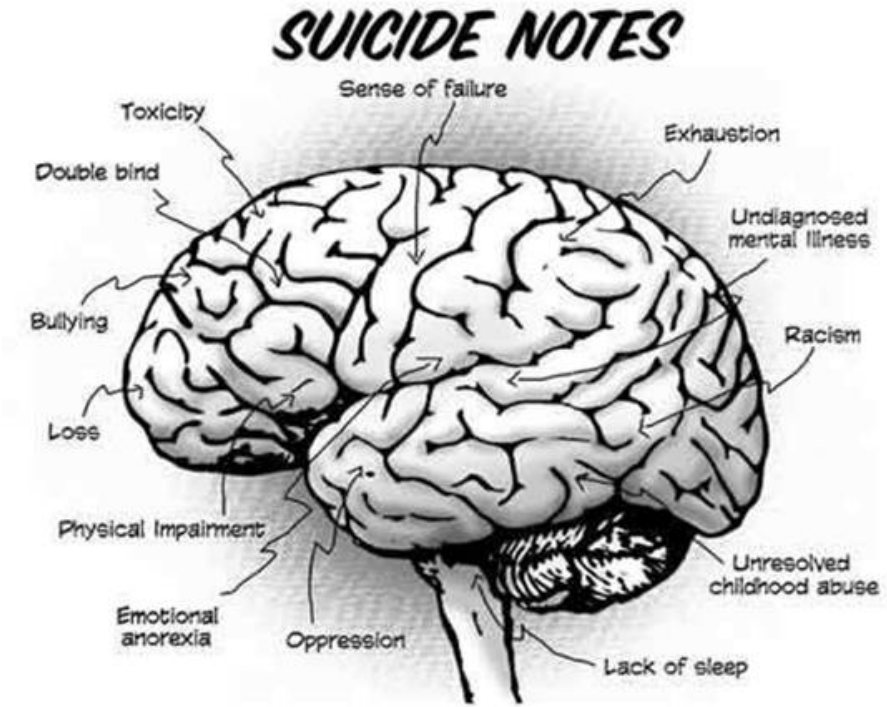
Rifat Zahan

PhD Student
Computational Epidemiology and Public Health Informatics Lab (CEPHIL)
Department of Computer Science
University of Saskatchewan
Saskatoon, SK, Canada

April 11, 2025

Background

- Suicide is one of the leading causes of death, and the rate of suicide-related death is increasing worldwide.
- According to WHO, about 800,000 people commits suicide and the number of actual attempts are still unknown (WHO, 2016).
- About 90% of the people, who commits suicide have *mental disorder*.
- **Major Depressive Disorder (MDD)** is one of the risk factors for committing suicide.
- Genetic data may help to prevent suicides by identifying people, who are at high risk.
- Genetic data may also help to avoid miss-classification of suicide-related deaths during *postmortem*.



PC: Function and information content of DNA methylation.

Retreived from: Epigenetic Blog, Amanda Mayer.

DNA Methylation

- DNA methylation is referred to as addition of methyl group to the cytosine residues of DNA molecule (Haghighi et al., 2014).
- DNA methylation data has successfully been analyzed to predict disease at early stage and to monitor the progression of the disease throughout the treatment (Tost, 2010).
- DNA methylation has successfully been used for monitoring suicide progression and prediction. (Guintivano et al., 2014; Kaminsky et al., 2015).

Objective

The objective of the study is to **predict** *suicidal* and *non-suicidal* deaths from DNA methylation data using modern machine learning algorithm.

Data Description

Data Source

(i) **Primary Source:** Douglas-Bell Canada Brain Bank (DBCBB).

(ii) **Secondary Source:** National Center for Biotechnology Information (NCBI). GEO Accession: GSE88890.

Data Type

(i) **Human Sample:** MDD Suicide cases (n = 20); Non-Psychiatric, sudden death controls (n = 20).

(ii) **Tissue:** Two cortical brain regions; Broadmann Area 11 (BA11) and Broadmann Area 25 (BA25).

(iii) **Extracted Molecule:** Genomic DNA

Genomic Data Extraction

According to Murphy et al. (2017)

- Genomic data was **isolated** from brain cortex.
- DNA was **treated** with sodium bisulfite using DNA methylation kit.
- Samples were **processed** and **assessed** using Illumina Infinium HumanMethylation450K BeadChip (Illumina).
- Raw signal of each probe was **extracted** using Illumina Genome Studio Software.
- CpG contents of **normalized** β values were used for further analysis, where

$$\beta = \frac{\text{Methylated Probe Intensities}}{\text{Methylated Probe Intensities} + \text{Unmethylated Probe Intensities}}$$

Data Preparation for Analysis

Feature Selection

Randomly selected 15K CG contents from 327,616 CG contents.

Sample Split for Cross Validation

Training set: 75% of the data (30 individuals for BA11 and 27 individuals for BA25)

Test set: 25% of the data (10 individuals for BA11 and 8 individuals for BA25)

Methodology

Dimensionality Reduction

Compression of a high-dimensional data into a low dimensional data by keeping the original essence of the data.

(i) Principal Component Analysis (PCA)

(ii) t-distributed Stochastic Neighbor Embedding (t-SNE)

Classifier

In machine learning, classification refers to as categorizing a new set of data based on feature(s).

(i) Support Vector Machine (SVM)

Principal Component Analysis (PCA)

- PCA is a **linear algorithm** that *maximizes* the *variance* of the projected high dimensional data into low dimension space.
  - e.g., projecting 3-D movie to 2-D screen.
- The **first** PC captures the *most variation* of the data.
- PCA captures **global structure** of the data.

t-distributed Stochastic Neighbor Embedding (t-SNE)

- t-SNE is a **non-linear algorithm** that captures the **global**, as well as **local** structure of high-dimensional data.
- t-SNE projects the neighborhood structure of the data in a way that points that are *similar* will **remain closer** in the space.
- t-SNE is NOT a *clustering algorithm*, since the input features are no longer identifiable.
- The *output* from low dimensional features can be used for training the classifier.
- Time and space complexity.
  - Cannot work beyond 10K features.

Dimensionality Reduction for Broadmann Area 11: Two Dimensions

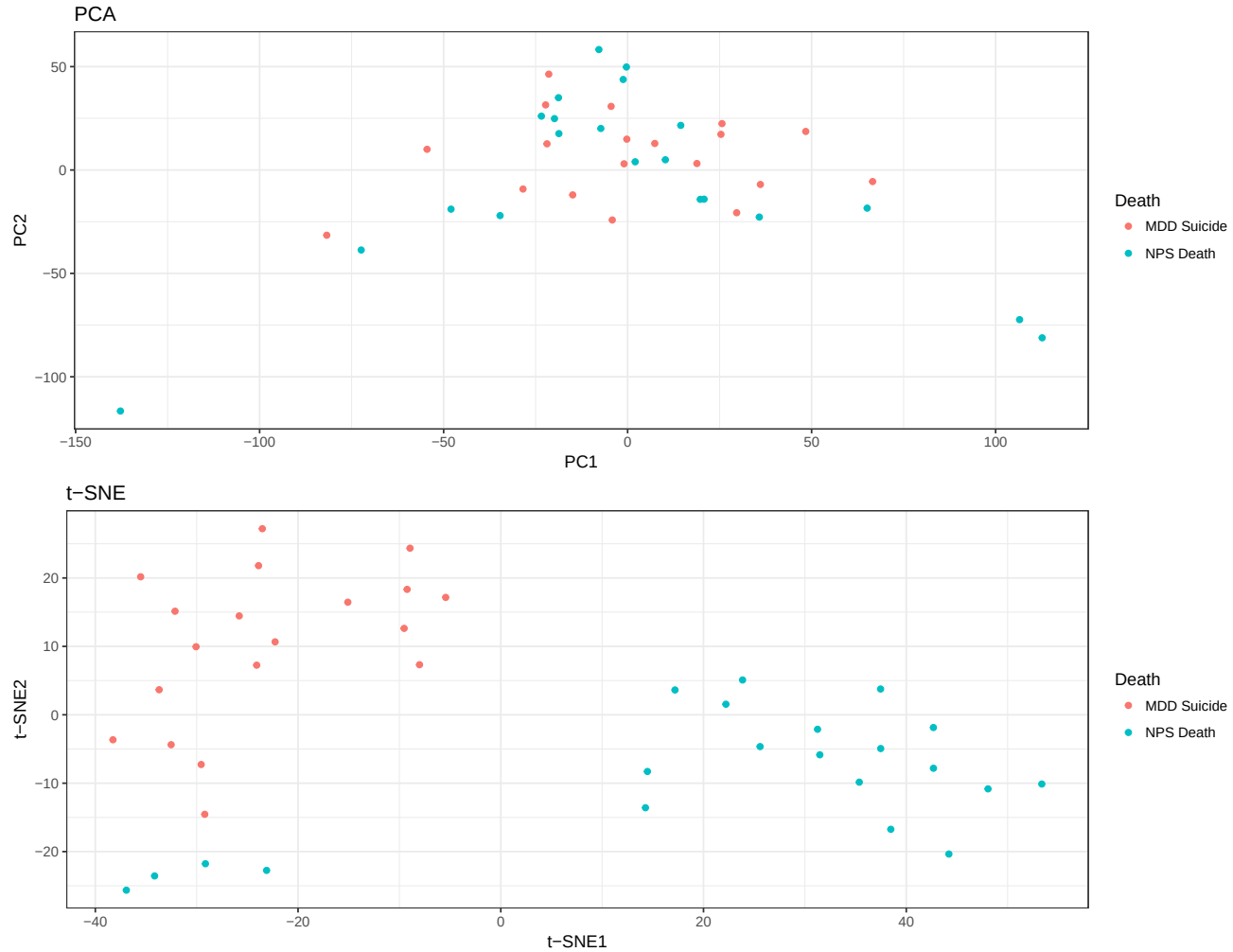


Figure: Two Dimensional PCA and t-SNE for BA11

Dimensionality Reduction for Broadmann Area 11: Three Dimensions

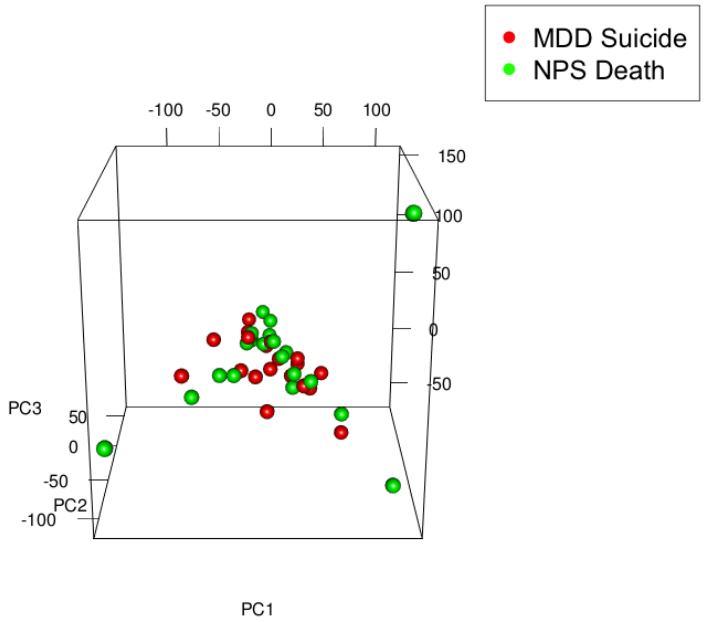


Figure: Three Dimensional PCA for BA11

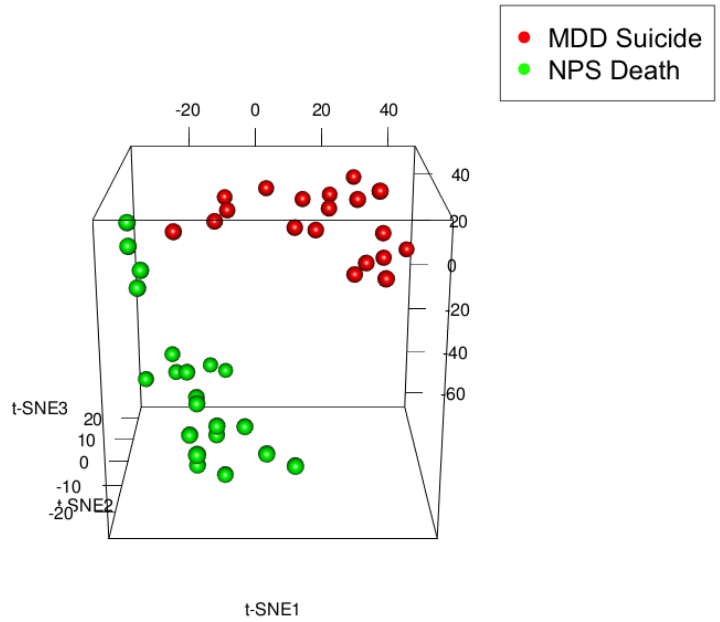


Figure: Three Dimensional t-SNE for BA11

Dimensionality Reduction for Broadmann Area 25: Two Dimensions

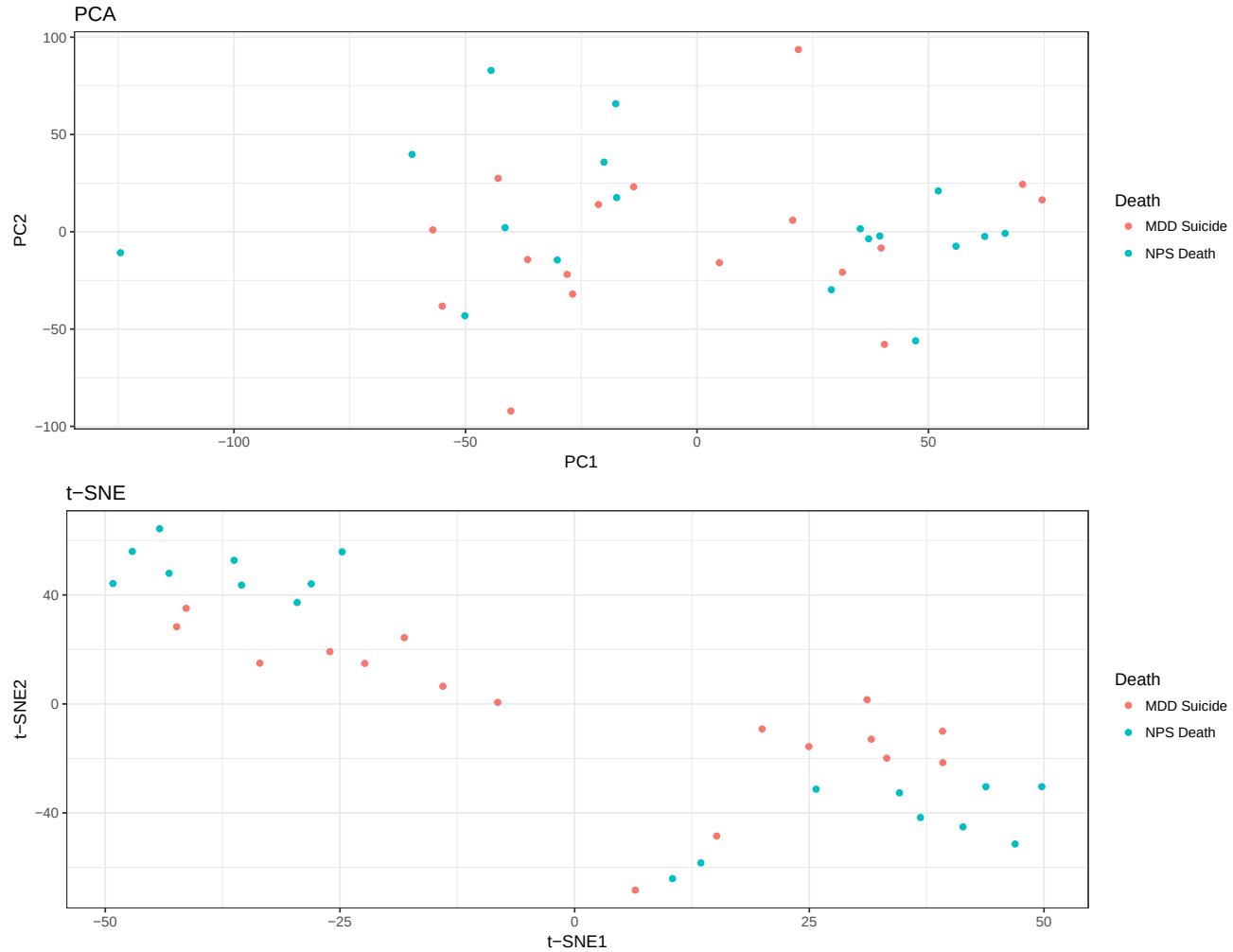


Figure: Two Dimensional PCA and t-SNE BA25

Dimensionality Reduction for Broadmann Area 25: Three Dimensions

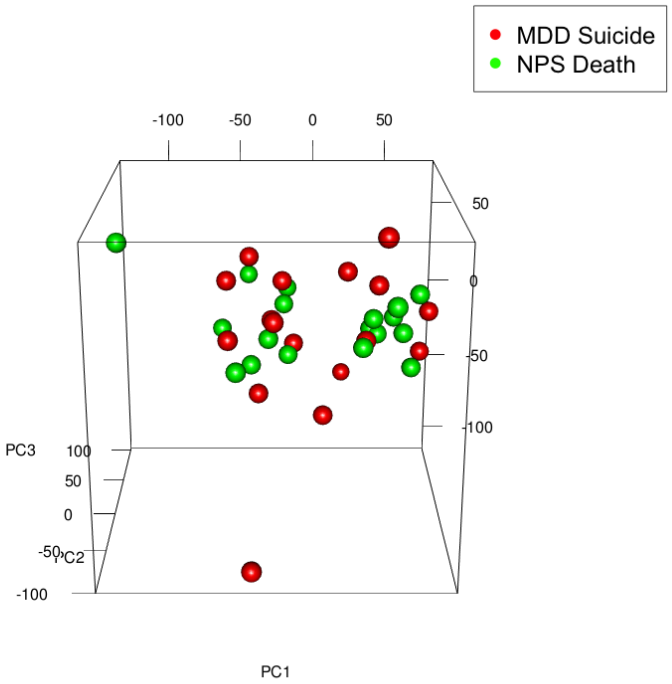


Figure: Three Dimensional PCA for BA25

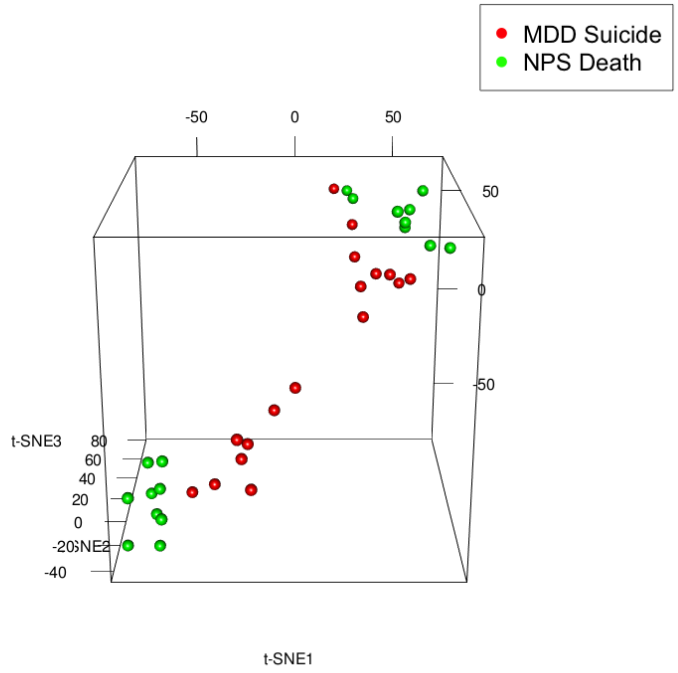


Figure: Three Dimensional t-SNE for BA25

Support Vector Machine (SVM)

- SVM is a classifier, that finds a **hyper-plane** which separates the classes very well.
- SVM has three main tuning parameters:
  - (i) kernel: non-linear plane.
  - (ii) gamma: kernel coefficient.
  - (iii) Cost: penalty parameter.
- SVM is robust to outliers.

Results: SVM for BA11

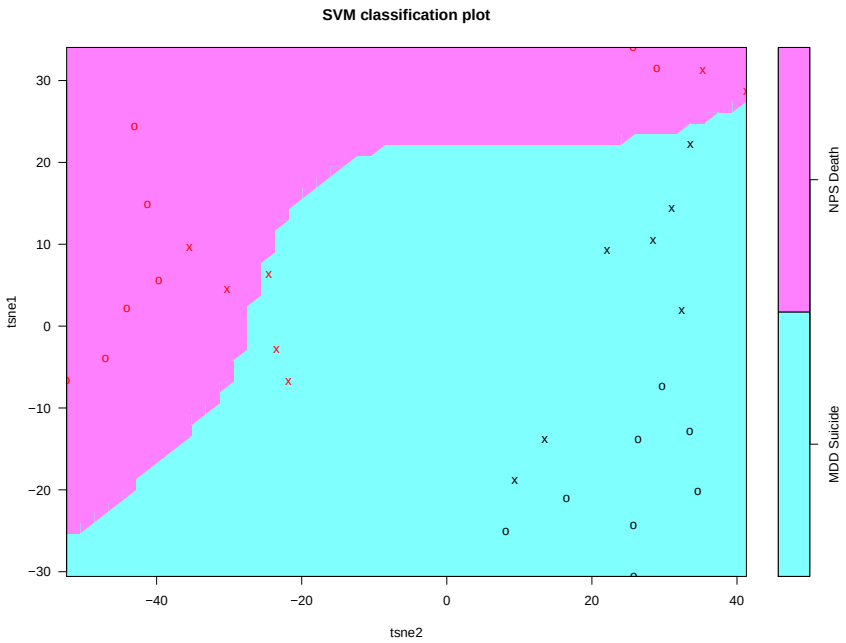


Figure: Fitted SVM model for 2-dimensional t-SNE

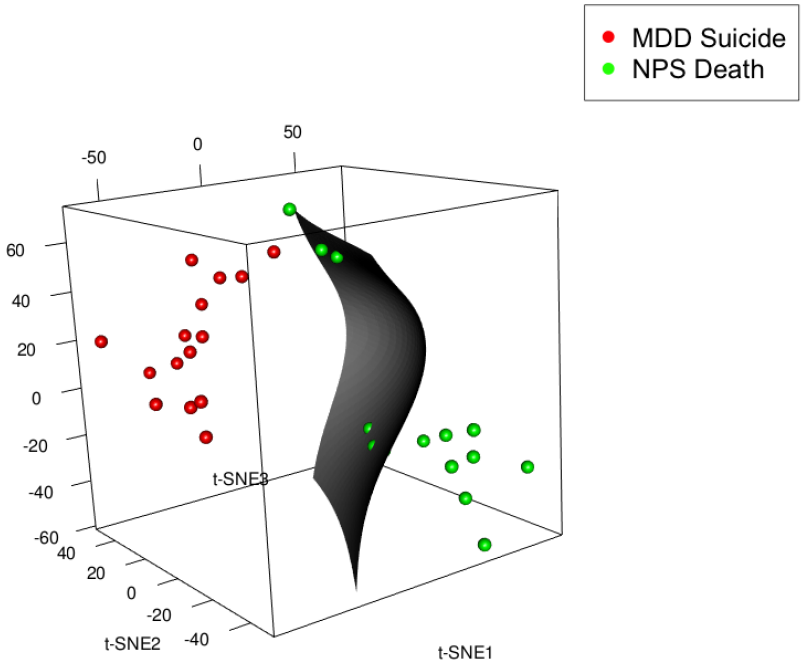


Figure: Fitted SVM model for 3-dimensional t-SNE

Results: SVM for BA25

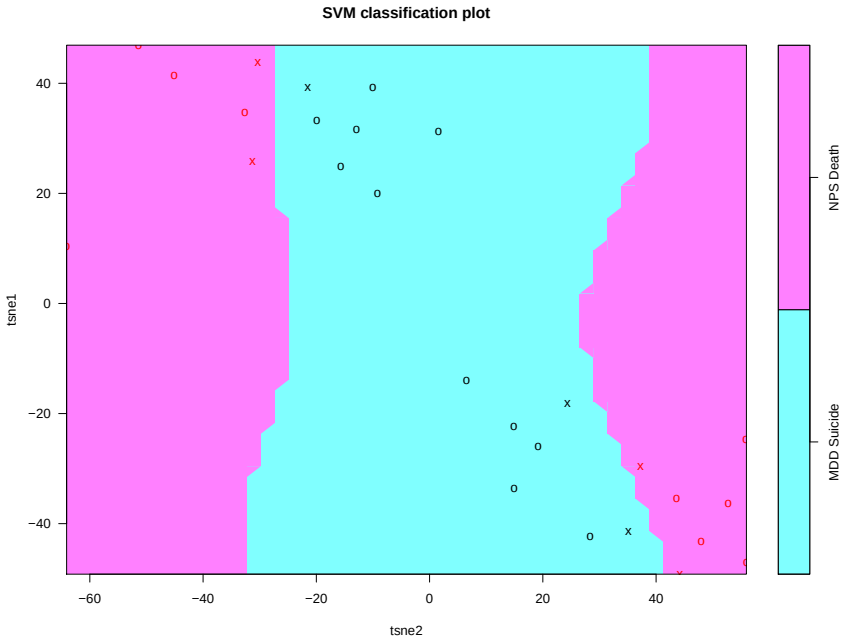


Figure: Fitted SVM model for 2-dimensional t-SNE

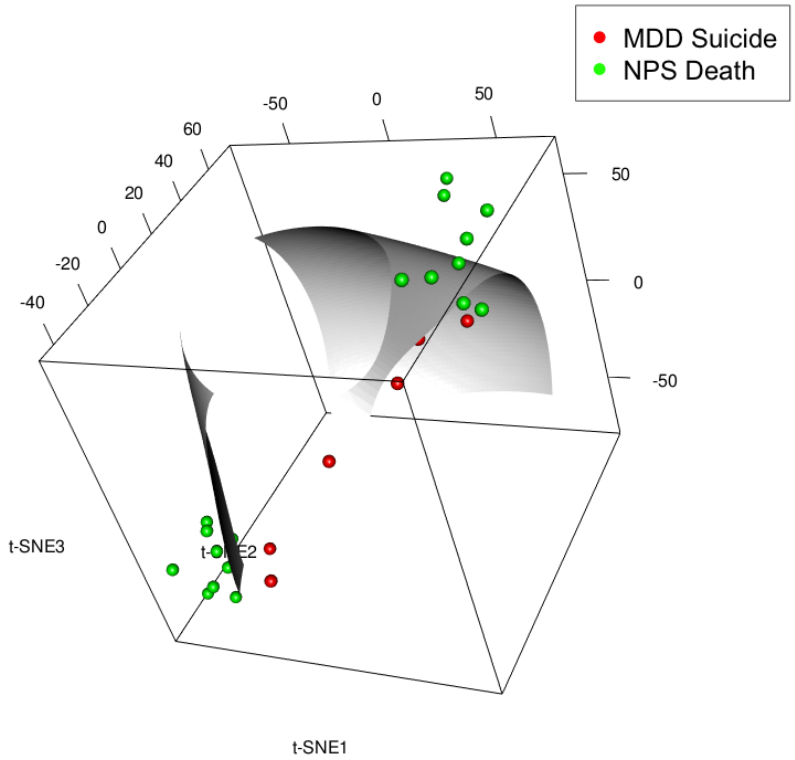


Figure: Fitted SVM model for 3-dimensional t-SNE

Cross Validation

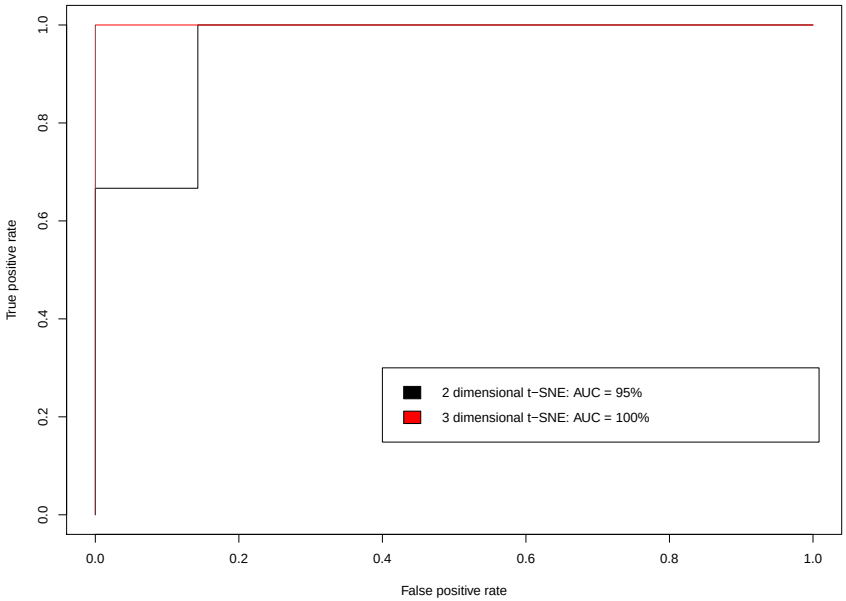


Figure: ROC Curves for Model of BA11

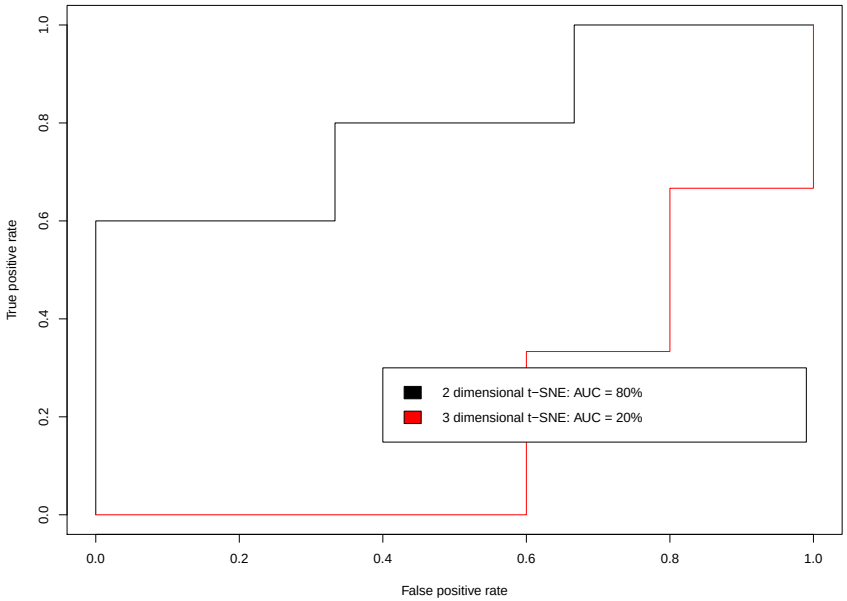


Figure: ROC Curves for Model of BA25

Limitation and Future Study

Limitation of the Study

(i) Small sample size (individuals) for training and testing the data.

(ii) Study suffers from Spectrum Bias.

Future Study

(i) More DNA methylated data should be gathered from both living and non-living individuals (e.g., from blood, saliva, hair, etc.)

(ii) Simulation study can be conducted to generate data, train the model and test it in real data.

Conclusion

- Compared to PCA, t-SNE is found useful in the reduction of dimensionality in high-dimensional DNA methylated data.
- Two-dimensional t-SNE better classifies the suicidal and non-suicidal deaths compared to three-dimensional t-SNE.
- Overcoming the limitation discussed may result in better predictive-ability of the classes in this suicide study.

Reference

Guintivano, J., Brown, T., Newcomer, A., Jones, M., Cox, O., Maher, B. S., Eaton, W. W., Payne, J. L., Wilcox, H. C., and Kaminsky, Z. A. (2014). Identification and replication of a combined epigenetic and genetic biomarker predicting suicide and suicidal behaviors. *American journal of psychiatry*, 171(12):1287–1296.

Haghighi, F., Xin, Y., Chanrion, B., O'Donnell, A. H., Ge, Y., Dwork, A. J., Arango, V., and Mann, J. J. (2014). Increased dna methylation in the suicide brain. *Dialogues in clinical neuroscience*, 16(3):430.

Kaminsky, Z., Wilcox, H., Eaton, W., Van Eck, K., Kilaru, V., Jovanovic, T., Klengel, T., Bradley, B., Binder, E., Ressler, K., et al. (2015). Epigenetic and genetic variation at ska2 predict suicidal behavior and post-traumatic stress disorder. *Translational psychiatry*, 5(8):e627.

Murphy, T., Crawford, B., Dempster, E., Hannon, E., Burrage, J., Turecki, G., Kaminsky, Z., and Mill, J. (2017). Methylomic profiling of cortex samples from completed suicide cases implicates a role for psors1c3 in major depression and suicide. *Translational psychiatry*, 7(1):e989.

Tost, J. (2010). Dna methylation: an introduction to the biology and the disease-associated changes of a promising biomarker. *Molecular biotechnology*, 44(1):71–81.

WHO (2016). *Suicide Data*. `http://www.who.int/mental_health/prevention/suicide/suicideprevent/en/` [Accessed: October 26, 2017].

Thank You!