# Term Project Proposal
# **Music Generation using Recurrent Neural Network**

Fawwaz Dzaky Zakiyal (201899213)
Tanjung Dion (201883621)

Deep Learning (Spring 2019)

## 1 Introduction

Prostate cancer affects a significant proportion of the male population, but in most cases, the disease is harmless.

## 2 Methodology

### 2.1 Concepts

- **DNA methylation**: the process of adding methyl groups to DNA, in this process modification of covalent nucleotides in the human genome, namely cytosine and also guanine. It is one of epigenetic modification which takes an important role in the development of cancer.

- **Gleason Score**: a measure (from 2-10) of the aggression of prostate cancer cells based on clinical pathology of prostate tissue [1].

- **High-Grade Prostate Cancer (HGPCa)**: A Gleason score from 8-10 is indicative of High-Grade prostate cancer, while scores below 8 are considered not as severe, HGPCa generally results in poor patient outcomes mortality, complications, and long-term disease-free survival) [1].

- **Cancer gene catalogue**: a catalogue those genes which contain mutations that have been causally implicated in cancer.

### 2.2 Dataset

The dataset comes from the National Cancer Institute GDC Data Portal and can be found online under Project TCGA-PRAD (The Cancer Genome Atlas Prostate Adenocarcinoma) [3]. This dataset consists of genomic information and clinical pathology reports belonging to 549 patients who have been diagnosed with prostate cancer. It contains both clinical (recurrence, survival & treatment resistance) and molecular profiles (Exon (mRNA) expression, DNA methylation, Copy Number Variations (CNV) & Single Nucleotide Polymorphism (SNP)). The file that we used are:

- **PRAD.meth.by_mean.data**: the mean signal values among each gene corresponding to the patient barcode. The size is 20111 genes x 549 patients (52 MB). This data become our learning features.

- **All_CDEs**: clinical data elements corresponding to the patient barcode. It contains many related information, but we only consider the gleason score to be the class label.

### 2.3 System Design

The process consists of preprocessing of TCGA-PRAD dataset, consist of feature selection, missing data imputation & adding the label. After that, split the dataset into training & testing data, next we train & evaluate the classifier models. The system design can be seen in Figure 1. The system specification where this project running on:

- **Programming language**: Python

- **Support library**: Scikit-Learn & Pandas

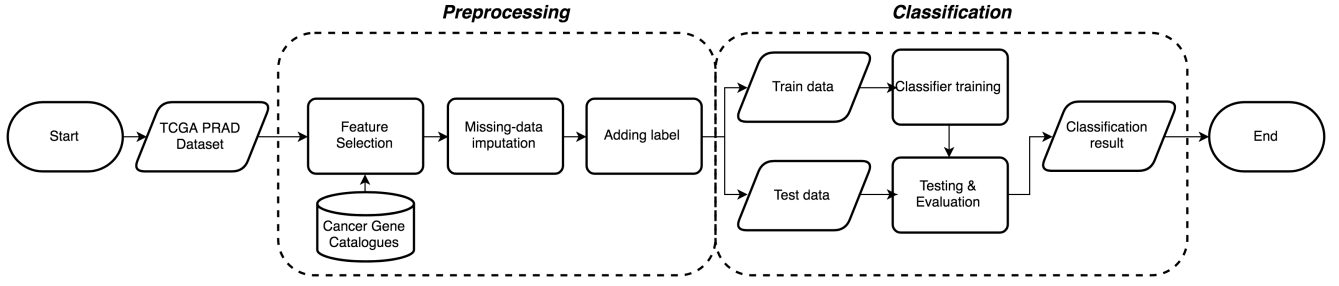- **PC**: Windows 10 64-bit, Intel i5-7500, RAM 4 GB

Figure 1: System Design

## 2.4 Preprocessing

- **Feature selection**: To reduce the feature space of the dataset, only those genes for which mutations have been causally implicated in cancer are considered, these are obtained through cancer gene catalogue like *COSMIC* (Catalogue of Somatic Mutations in Cancer) that consist of 3946 genes information [4], and *CIVIC* (Clinical Interpretation of Variants in Cancer) that consist of 366 genes information [5]. Originally, our dataset has 20111 genes, after feature selection it has only 821 genes.

- **Missing data imputation**: It is the process of replacing missing data of dataset features with substituted values. This project using mean imputation.

- **Adding the label**: We categorized into 2 classes, between score of 2 to 7 (not severe) and 8 to 10 (HGPCa / severe) from All_CDEs data. We added class labels to the dataset with correspond to the patient id.

## 3 Experiments

- **Evaluation on Logistic Regression (LR)**: The experiment cases are LR using L1 or L2 penalty. We fixed the other parameters, such as the regularization parameter $C$ (1.0). The result is LR+L1 accuracy (76.36%) outperform LR+L2. In general, the L1 is expected to be very effective with sparse data, thus it can give a better result. The result shown at Figure 3.

- **Evaluation on K-Nearest Neighbor (KNN)** : The experiment cases are KNN with number of neighbors 2, 3, 4, or 5. We fixed the other parameters, such as the distance metric (Minkowski). The result is KNN with 3 neighbors accuracy (68.48%) outperform the other cases. It seems k=3 is the optimum number of neighbors in this experiment. The result shown at Figure 4.

- **Evaluation on Support Vector Machine (SVM)**: The experiment cases are SVM using Linear kernel, RBF kernel, or Poly kernel (degree 2, 3, or 4). We fixed the other parameters, such as the regularization parameter $C$ (1.0), and the gamma kernel coefficient for RBF & Poly (1/number of features). The result is SVM using linear kernel accuracy (76.97%) outperform the other cases that get the same accuracy (61.21%). It seems that linear kernel get high accuracy indicates the data is linearly separable and the non-linear kernel (RBF & Poly) low accuracy indicates an over-fitting issue. The result shown at Figure 5.

- **Evaluation on Multilayer Perceptron (MLP)**: The experiment cases are MLP using Sigmoid/Logistic, Tanh or ReLU activation function with number of hidden layers 1 (100 neurons) or 2 (100 neurons & 100 neurons). We fixed the other parameters, such as the optimizer algorithm (Adam), the batch size (200), and the learning rate (0.001). The result is MLP using Tanh activation function with 2 hidden layer accuracy (76.97%) outperform the other cases. It seems that increasing number of hidden layers lead to increase the accuracy, and generally the Tanh activation function is expected to be very effective with two classes classification task. The result shown at Figure 6.

## 4 Project Statements

**Contributions from Fawwaz Dzaky Zakiyal**:

- Dataset split

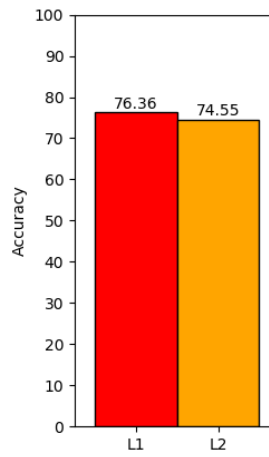**Contributions from Tanjung Dion**:

- Dataset split

Figure 2: Evaluation on Logistic Regression
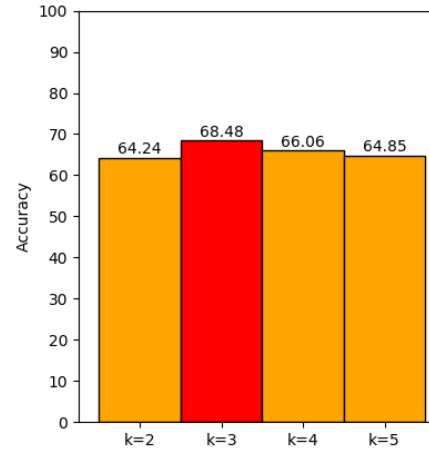


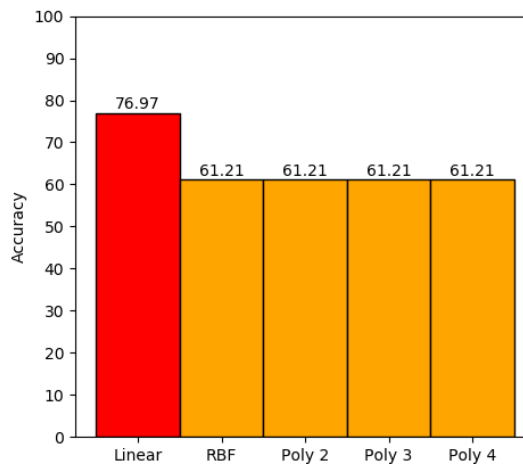Figure 3: Evaluation on K-Nearest Neighbor



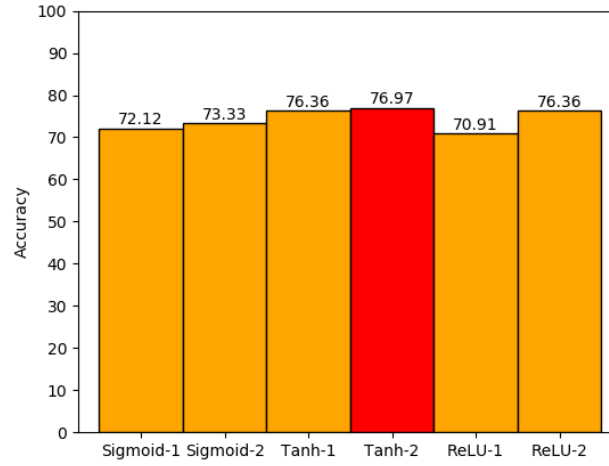Figure 4: Evaluation on Support Vector Machine



Figure 5: Evaluation on Multilayer Perceptron

# References

[1] Ilir Agalliu, Robert Gern, Suzanne Leanza, and Robert D. Burk. 2009. *Associations of High-Grade Prostate Cancer with BRCA1 and BRCA2 Founder Mutations.* Clinical Cancer Research. Vol. 15, No. 3.

[2] Hao, Xiaoke, et al. *DNA methylation markers for diagnosis and prognosis of common cancers.* Proceedings of the National Academy of Sciences 114.28 (2017): 7414-7419.

[3] Genome.ifmo.ru. (2018). TCGA PRAD Dataset. [online] Available at: *https://genome.ifmo.ru/files/software/phantasus/tcga/PRAD/* [Accessed 29 Nov. 2018].

[4] Cancer.sanger.ac.uk. (2018). Cancer Gene Census. [online] Available at: *https://cancer.sanger.ac.uk/census* [Accessed 29 Nov. 2018].

[5] Civicdb.org. (2018). CIViC - Clinical Interpretations of Variants in Cancer. [online] Available at: *https://civicdb.org/releases* [Accessed 29 Nov. 2018].