CS4220 Knowledge Discovery Methods in Bioinformatics
Group Project: Predicting Gene Essentiality in Cancer Cell-Lines using a Weighted Naïve Bayes Model

Professors:
Prof Anders Jacobsen Skanderup
Prof Niranjan Nagarajan

Submitted by:
Amanda Ho Shan Rui (A0187715U)
Benjamin Tan Jee Min (A0192270J)
Chan Sheng You (A0135459X)
Chen Tianying, Tiana (A0169834N)

Introduction

A growing number of genetic mutations and features that cause development of cancers are being identified and understood, but translating this understanding into clinical treatments and medication has been slow. Targeting specific genetic alterations require tests to identify biomarkers and effective targeted treatment options because only genes specifically essential to cancer cell-lines present an opportunity for targeted cancer therapy. Thus, the aim of our project is to develop a predictive model that can infer gene essentiality in cancer cell-lines using molecular features. Gene essentiality is defined as the degree to which a gene is essential for the survival and proliferation of an organism, where a loss of function of an essential gene will compromise viability and result in a drastic loss of fitness in cells (Bartha et al., 2018). Identifying essential genes within cancer cell-lines provide potential gene targets for personalized cancer medicine and treatment, enabling selective killing of cancer cells while minimising lethal effects against healthy cells. The need for more efficient identification of essential cancer genes for targeted therapy underpins the importance of predicting gene essentiality within cancer cell-lines.

Specific targeted cancer therapy has been shown to improve prognosis significantly (Baudino, 2015). A common example is the identification of estrogen receptors (ER), progesterone receptors (PR) and human epidermal receptor 2 (HER2) genes in breast cancers. These genes are shown to be essential to specific breast cancers when overexpressed and the availability of treatments towards these genes have significantly improved the life expectancy and the quality of life for patients (Montano-Samaniego et al., 2020). On the other hand, essential genes of a specific breast cancer called triple negative breast cancer (TNBC) have not yet been identified. As a result, current treatments for TNBC are still ineffective with high chances of relapse, culminating in a bleaker outlook for patients (Mehanna et al., 2019).

Several popular ways have been developed to identify essential genes within cancer cell-lines including the use of large-scale functional screens. Such methods include using RNA interference, short hairpin RNA or CRISPR-Cas9 to suppress gene expression for the screening of essential genes (Shao et al., 2013). However, these screens have only identified a small number of essential genes through statistical analysis (Marcotte et al., 2016). To provide a more efficient method of identifying essential genes, we propose implementing a machine learning model to predict gene essentiality in cancer cell-lines using molecular features. Specifically, we demonstrate that our Weighted Gaussian Naive Bayes (WNB) model can accurately predict gene essentiality within cancer cell-lines using gene expression and functional enrichment data.

Methods and Materials

Project Workflow Overview

An overview of our project workflow is shown in Figure 1. Molecular feature selection is first performed on our given data before functional enrichment is carried out using the given gene list. Gene expression data and gene essentiality scores were used to derive an input matrix while functional enrichment results, gene expression data and gene essentiality scores were used to derive a weight matrix. The input matrix and weight matrix is then used to train a WNB model.

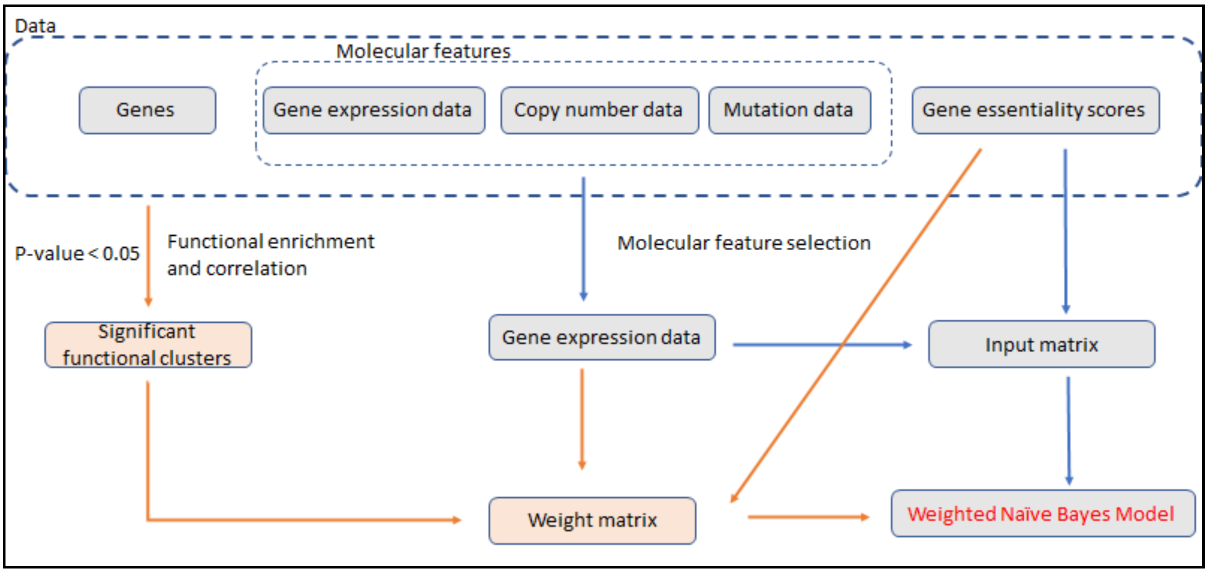

Figure 1 Overview of project workflow. Molecular feature selection is first performed on our given data before an input matrix and weight matrix is derived for our Weighted Gaussian Naïve Bayes Model.

Data

Two independent sets of training and testing data used for this project was provided by the BROAD-DREAM Gene Essentiality Prediction Challenge (Gonen et al., 2017). The training and testing data provided three types of molecular features for usage in prediction, namely gene expression data, copy number variation and mutation data for various cancer cell-lines specific to different cancer types. In addition, the training and testing data provided gene essentiality scores of the corresponding cancer cell-lines as class labels for each gene. Training and testing data were both provided in the form of genes as rows and cancer cell-line names as columns. In total, there were 1369 genes in each of the training and testing data. Molecular features and gene essentiality scores corresponding to 105 and 44 cancer cell-lines were provided in the training and testing data respectively.

Molecular features provided in the training and testing data were obtained from the Broad Institute Cancer Cell Line Encyclopedia (CCLE) (Barretina et al., 2012). CCLE is a compilation of gene expression, copy number variation and sequencing data from nearly 1,000 human cancer cell-lines. Gene expression data on CCLE is obtained through microarray analysis, background correction, normalization and summarization. A higher gene expression value corresponds to greater gene expression. Copy number variation on CCLE was obtained through a high density single nucleotide polymorphism (SNP) array. A positive copy number value indicates amplification while a negative value indicates deletion. Mutation data on CCLE was obtained through sequencing of over 1,600 genes and mass spectrometric genotyping. A value of 1 indicates the occurrence of mutation while 0 indicates the wild type.

Gene essentiality scores provided in the training and testing data were obtained through shRNA knockdown screens (Gonen et al., 2017). Read counts were normalized and $\log_2$ transformed before shRNA with overlapping sequences or low abundances were removed. The fold change of reads were then calculated using DNA reference samples and quantile normalized within each cancer cell-line before mapping to genes. Lastly, the DEMETER algorithm was used to obtain gene essentiality scores from the reads (Tsherniak et al., 2017). A more negative gene essentiality value indicates higher degree of gene essentiality.

Pre-processing of Training and Testing Data

To train and test our WNB model, molecular feature data and gene essentiality scores from two cancer cell-lines of the same cancer type were selected from the training and testing data. Cancer cell-lines of the same cancer type were selected because we expect gene essentiality profiles to be more similar within the same cancer type as opposed to different cancer types. As we expect gene expression data and gene essentiality scores to correlate positively, we performed Pearson correlation on gene expression data and gene essentiality scores within each cancer cell-line from the testing and training data. The lung cancer cell-lines CHAGOK1 and LOUNH91 were chosen for usage as our training and testing data respectively because they displayed the most significant correlation among all cell-lines.

Functional enrichment analysis and subsequent functional cluster selection were carried out. Functional enrichment was first performed on all 1369 genes for testing and training data separately using the Database for Annotation, Visualisation, and Integrated Discovery (DAVID) (Dennis et al., 2003). Significant functional clusters (adj p-val $\leq$ 0.05) were retained and functional cluster columns indicating the occurrence of each gene in each functional cluster were added into the training and testing data. A value of 0 in a functional cluster column indicates a non-occurrence of a gene in the functional cluster, while a value of 1 indicates occurrence of a gene in the functional cluster.

To select only molecular features that were relevant for the prediction of gene essentiality, molecular feature selection was carried out. Pearson correlation was performed between molecular feature data and gene essentiality scores within each of the 3 molecular features. It was found that only gene expression molecular feature showed significant positive correlation with the gene essentiality scores. Thus, only gene expression data was retained for the training and testing of our WNB model.

As the WNB model performs poorly as a regression model but is typically an excellent classifier, we discretized the gene essentiality scores for each gene. The gene essentiality scores were found to follow a normal distribution, thus discretization was done using quartile ranges. Genes with gene essentiality scores within the lower quartile were given the “Essential” class label while the remaining genes were given the “Non-Essential” class label.

Training and Testing of Weighted Gaussian Naive Bayes Model

To train our WNB model, we first derive a weight matrix and an input matrix from the training data. The input matrix contains the gene expression data and the discretized gene essentiality class labels for each gene. To obtain a weight matrix, gene expression was first discretized using quartile ranges because gene expression was found to follow a normal distribution. Genes with gene expression within the upper quartile were labelled ‘High’ while the remaining genes were labelled ‘Low’. The final weight matrix contains the discretized gene expression data, functional cluster columns and discretized gene essentiality class labels. A weight vector containing the weights of each gene was then obtained from the weight matrix using a modified Information Gain formula. The same procedure was followed to derive a weight matrix and an input matrix from the testing data. The model was then trained and tested with the weight vectors and input matrices using the GaussianNB function provided by scikit-learn 0.24.1 in Python.

Benchmarking and Evaluation of Weighted Naive Bayes Model

To assess the performance of our WNB model, it was benchmarked against the Gaussian NB model. Confusion matrices of both models were visualised using plot_confusion_matrix. The performances of both models were evaluated by their accuracy, specificity, sensitivity derived manually from the confusion matrices. The models were also evaluated by their Area Under ROC Curve (AUC) using the roc_auc_score function provided by scikit-learn in Python. In addition, functional enrichment using DAVID was performed on incorrectly and correctly WNB-predicted essential genes separately to investigate the reason behind why certain essential genes were preferentially predicted correctly.

Code Availability

Code used in this project is available on Github at https://github.com/Amandahsr/CS4220_Predicting_Gene_Essentiality.git.

Weighted Gaussian Naive Bayes Model

To build a model that accurately predicts gene essentiality in cancer cell-lines, the Gaussian NB was chosen as a suitable model (Devroye et al., 1996). Gaussian NB is suitable as a classifier for our gene essentiality problem because it does not require a large training dataset and is simple to implement. Additionally, it is shown to outperform other more complex models for classification problems (Cheng et al., 2005; Zelic et al., 1997). The implementation of Gaussian NB is based on the Bayes Theorem:

$$\mathrm{P}(c_i \mid x_i) \; = \; \mathrm{P}(c_i \mid x_i) \; \times \; \mathrm{P}(c_i) \, \mathrm{P}(x_i) — (1)$$

where c_i refers to the gene essentiality of gene i and x_i refers to the gene expression of gene i.

However, Gaussian NB operates on the assumption of class conditional independence, where features are assumed to be independent of one another given their class. In the context of the gene essentiality problem, the class conditional independence assumption is violated because of gene dependency. To account for gene dependency, a modified version of Gaussian NB called the WNB is implemented. In WNB, the probability of each gene is weighted according to its gene expression and gene essentiality in relation to other genes:

$$Probability\ of\ Gene\ i\ being\ Essential = \; w_i \; \times \; \mathrm{P}(c_i \mid x_i)$$

where w_i refers to the weight of gene i and $\mathrm{P}(c_i \mid x_i)$ is defined in equation (1).

We propose a novel weights formula for the calculation of weights in WNB. The weights formula considers three assumptions for gene essentiality. Firstly, we assume that gene essentiality is positively correlated with gene expression. Secondly, we assume that a gene is deemed essential only in relation to other genes. Thus, genes with “High” gene expression are more likely to be essential, and only when it exists within a proportion of other genes with “Low” gene expression. To model these two assumptions, the Gain value of each gene is calculated using the Information Gain formula (Shannon, 1997). The Information Gain formula is particularly useful because it models the gene expression of each gene relative to other genes. The Gain value of each gene can be calculated using the weight matrix:

$$Gain(c_i \, , a_i) = B\left(\frac{P}{P+N}\right) - H(c_i \mid a_i) — (2)$$

where c_i refers to the gene essentiality of gene i, a_i refers to the gene expression of gene i, P and N refers to the total number of essential and non-essential genes respectively. B and H formula is calculated using the Entropy formula.

Secondly, we assume that gene essentiality is related to the number of functions a gene perform. Thus, genes that perform more functions and appear in more functional clusters are more likely to be essential. To model this assumption, the Gain value of each gene within each of its functional cluster is calculated before all Gain values are aggregated for each gene. The final weight vector containing the weights of each gene can then be obtained by:

$$Weight\ of\ Gene\ i = \sum_{j=1}^{N} Gain_j(c_i \, , a_i)$$

where k refers to the total number of j functional clusters gene i is involved in and the Gain formula is as defined in (2).

The WNB model then takes in the input matrix and the final weight vector to predict the gene essentiality of each gene.

Results and Discussion

Benchmarking of Weighted Naive Bayes Model

To benchmark our WNB model, its gene essentiality prediction results is compared to the Gaussian NB model. The two models were evaluated on their accuracy, sensitivity, specificity and AUC. The confusion matrices are shown in Figure 2 and the benchmarking results shown in Table 1.

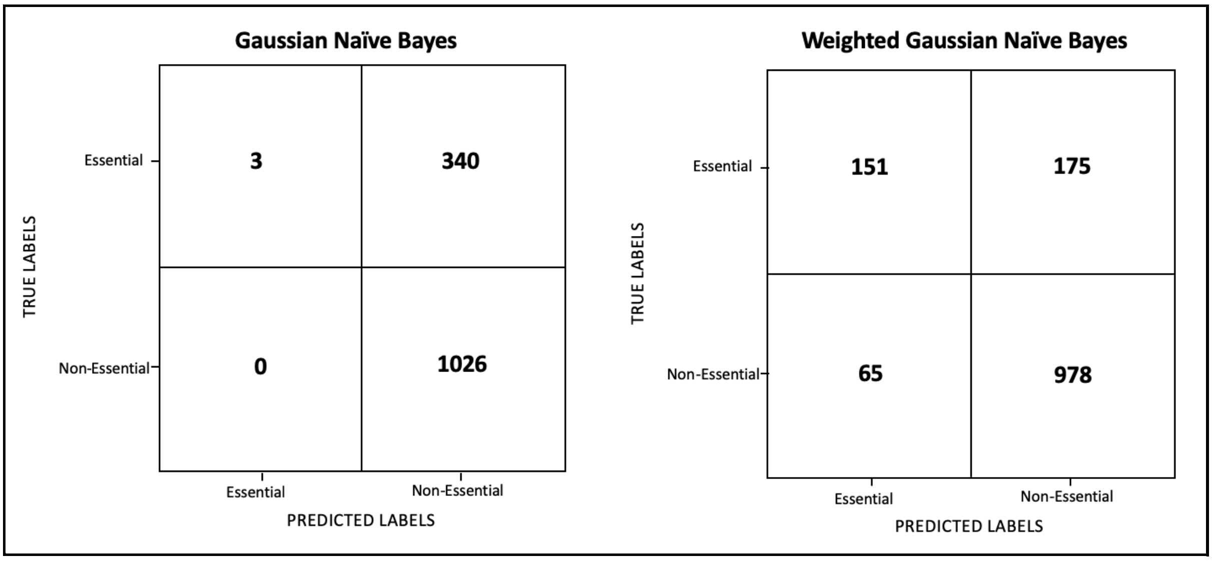

Figure 2 Confusion Matrix of the Gaussian Naive Bayes model and the Weighted Gaussian Naive Bayes model.

Table 1 Accuracy, Sensitivity, Specificity and AUC of the Gaussian Naive Bayes model and the Weighted Gaussian Naive Bayes model.

	Gaussian Naive Bayes	**Weighted Gaussian Naive Bayes**
Accuracy	75.2%	82.5%
Sensitivity	0.87%	45.9%
Specificity	100%	99.3%
AUC	0.5	0.56

Our WNB model is shown to outperform the Gaussian NB model in terms of accuracy, indicating that the WNB model was able to produce an overall increase of correct gene essentiality predictions. The WNB accuracy score translates to 1129 out of 1369 total genes being correctly predicted as their true gene essentiality labels. Our WNB model also displayed a significant increase in sensitivity compared to the Gaussian NB model, indicating that the WNB model was able to correctly predict more essential genes. The WNB sensitivity score translates to 151 out of 326 essential genes being correctly classified as essential. However, our WNB model displayed a small decrease in specificity compared to the Gaussian NB model. This was expected because our WNB model had correctly classified more essential genes than the Gaussian NB model.

To assess the two models on a fair metric, AUC was used to evaluate the performance of the models across all classification thresholds. Although our WNB model had only a marginal increase in AUC compared to the Gaussian NB model, the AUC values suggests that our WNB model performed better at distinguishing essential from non-essential genes. This was an improvement from the Gaussian NB model which displayed no distinction between the two classes of genes. Based on the accuracy, sensitivity, specificity and AUC, it is therefore shown that our WNB model outperformed the Gaussian NB model when predicting gene essentiality in cancer cell-lines.

Evaluation of Weighted Naive Bayes Model

To assess why our WNB model correctly predicted certain essential genes preferentially, we performed functional enrichment analysis on correctly and incorrectly predicted essential genes separately. The functional enrichment results on correctly predicted essential genes produced top significant enrichment clusters for more functional roles such as cell-cell adhesion (adj p-val = 2.0E-2) and apoptosis (adj p-val = 5.1E-3). On the other hand, functional enrichment results on incorrectly predicted essential genes produced top significant enrichment clusters for more molecular roles such as protein kinase catalytic domain (adj p-val = 8.8E-29) and protein phosphorylation (adj p-val = 7.2E-22). Thus, we speculate that our WNB model was preferentially predicting essential genes that performed more functional roles as opposed to those that performed more molecular roles. This preferential prediction was not entirely unexpected because the functional clusters retained for weights calculation in our WNB model were all biological pathways, therefore missing out on essential genes that could have performed other types of gene function.

Limitations

To apply a WNB model, gene essentiality scores were discretized using quartile ranges. However, discretization by quartile ranges may not capture the true proportion of essential and non-essential genes within the chosen cancer cell-lines. Similarly, discretization of gene expression data was also performed using quartile ranges to obtain a weight matrix. This method of discretization may not have captured the true proportion of genes with low and high expression within the chosen cancer cell-line. Our method of discretization was limited by the lack of standardized quantitative threshold for which gene essentiality scores and gene expression data are considered essential and high respectively. Other methods of discretization shall be explored in future work to assess the best method for the most accurate gene essentiality predictions.

Our WNB model was trained and tested on two chosen lung cancer cell-lines. Thus, predictions made on gene essentiality may not be generalized to other cancer types or lung cancer cell-lines. To determine if the predictions produced by our WNB model are reproducible and applicable to other cancer type cell-lines, further training and testing using different combinations of training and testing cancer type cell-lines can be explored.

Our WNB model makes use of significant functional enrichment clusters for the prediction of gene essentiality. However, not all significant functional clusters are clinically significant because essential genes in cancer cells may also be house-keeping genes in healthy cells. Thus, our WNB model may have predicted overlapping essential genes in cancer cells and healthy cells. There exists a need to improve our WNB model to better identify cancer-specific essential genes such that healthy cells are minimally impacted when essential cancer genes are targeted for treatments. To better identify cancer-specific essential genes, future training and testing of our WNB model using cancer and healthy cell-lines should be explored.

Conclusion

Our WNB model has been shown to successfully predict gene essentiality in cancer cell-lines, allowing for more effective and accurate predictions of cancer gene essentiality. However, gene essentiality in cancer cells is a complex problem made difficult by additional factors such as intra-tumour heterogeneity within cancer types. Further improvements focused on identifying cancer-specific essential genes and obtaining additional predictive molecular features will enable us to build a more robust WNB model for the accurate prediction of gene essentiality in cancer cell-lines.

References

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., . . . Garraway, L. A. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature, 483*(7391), 603-607. doi:10.1038/nature11003

Bartha, I., di Iulio, J., Venter, J. C., & Telenti, A. (2018). Human gene essentiality. *Nat Rev Genet, 19*(1), 51-62. doi:10.1038/nrg.2017.75

Baudino, T. A. (2015). Targeted Cancer Therapy: The Next Generation of Cancer Treatment. *Curr Drug Discov Technol, 12*(1), 3-20. doi:10.2174/1570163812666150602144310

Cheng, B. Y., Carbonell, J. G., & Klein-Seetharaman, J. (2005). Protein classification based on text document classification techniques. *Proteins, 58*(4), 955-970. doi:10.1002/prot.20373

Dennis, G., Jr., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C., & Lempicki, R. A. (2003). DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol, 4*(5), P3. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/12734009

Devroye, L., Györfi, L. s., & Lugosi, G. b. (1996). *A probabilistic theory of pattern recognition*. New York: Springer.

Gonen, M., Weir, B. A., Cowley, G. S., Vazquez, F., Guan, Y., Jaiswal, A., . . . Margolin, A. A. (2017). A Community Challenge for Inferring Genetic Predictors of Gene Essentialities through Analysis of a Functional Screen of Cancer Cell Lines. *Cell Syst, 5*(5), 485-497 e483. doi:10.1016/j.cels.2017.09.004

Marcotte, R., Sayad, A., Brown, K. R., Sanchez-Garcia, F., Reimand, J., Haider, M., . . . Neel, B. G. (2016). Functional Genomic Landscape of Human Breast Cancer Drivers, Vulnerabilities, and Resistance. *Cell, 164*(1-2), 293-309. doi:10.1016/j.cell.2015.11.062

Mehanna, J., Haddad, F. G., Eid, R., Lambertini, M., & Kourie, H. R. (2019). Triple-negative breast cancer: current perspective on the evolving therapeutic landscape. *Int J Womens Health, 11*, 431-437. doi:10.2147/IJWH.S178349

Montano-Samaniego, M., Bravo-Estupinan, D. M., Mendez-Guerrero, O., Alarcon-Hernandez, E., & Ibanez-Hernandez, M. (2020). Strategies for Targeting Gene Therapy in Cancer Cells With Tumor-Specific Promoters. *Front Oncol, 10*, 605380. doi:10.3389/fonc.2020.605380

Shannon, C. E. (1997). The mathematical theory of communication. 1963. *MD Comput, 14*(4), 306-317. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/9230594

Shao, D. D., Tsherniak, A., Gopal, S., Weir, B. A., Tamayo, P., Stransky, N., . . . Mesirov, J. P. (2013). ATARiS: computational quantification of gene suppression phenotypes from multisample RNAi screens. *Genome Res, 23*(4), 665-678. doi:10.1101/gr.143586.112

Tsherniak, A., Vazquez, F., Montgomery, P. G., Weir, B. A., Kryukov, G., Cowley, G. S., . . . Hahn, W. C. (2017). Defining a Cancer Dependency Map. *Cell, 170*(3), 564-576 e516. doi:10.1016/j.cell.2017.06.010

Zelic, I., Kononenko, I., Lavrac, N., & Vuga, V. (1997). Induction of decision trees and Bayesian classification applied to diagnosis of sport injuries. *J Med Syst, 21*(6), 429-444. doi:10.1023/a:1022880431298

Contributions

Member	**Contributions**	**Contribution (%)**
Amanda Ho Shan Rui	Involved in formulation and implementation of the Weighted Naïve Bayes Model. Presented “Weighted Naïve Bayes Model” of presentation. Wrote “Weighted Naïve Bayes Model” and “Results and Discussion” of report.	25
Benjamin Tan Jee Min	Involved in formulation and implementation of the Weighted Naïve Bayes Model. Presented “Methodology” of presentation. Wrote “Project Overview”, “Pre-Processing of Data” and “Limitations” of report.	25
Chan Sheng You	Involved in formulation and implementation of the Weighted Naïve Bayes Model. Presented “Introduction” and “Dataset” of presentation. Wrote “Introduction”, “Data” and “Conclusion” of report.	25
Chen Tianying, Tiana	Involved in formulation and implementation of the Weighted Naïve Bayes Model. Presented “Benchmarking” and “Conclusion” of presentation. Wrote “Weighted Naïve Bayes Model” and “Benchmarking and Evaluation of Weighted Naïve Bayes Model” under Methods and Materials section of report.	25