Finding Biomarkers for Type 2 Diabetes

Aratrika Das[1]

Supervisors: Dr. Thomas Abeel[1]**, Eric van der Toorn**[1]**, David Calderón Franco**[1]

[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 25, 2023

Name of the student: Aratrika Das
Final project course: CSE3000 Research Project
Thesis committee: Dr. Thomas Abeel, Eric van der Toorn, David Calderón Franco, Thomas Höllt

Abstract

Type 2 Diabetes is a very prevalent disease in current times and leads to significant adverse effects. Recently, there has been a growing interest in the association of the human gut microbiome with respect to chronic diseases like Type 2 Diabetes with the aim to identify biomarkers. In this study, we researched the effect of different machine learning and feature selection techniques to identify biomarkers for Type 2 Diabetes that can later be used for diagnosis and prediction. The main methods that we explored were Random Forests, Linear Regression, Support Vector Machines and XGBoost along with mRMR and CMIM as feature selection techniques. These methods were applied to data taken from Europe and China. We found that mRMR improved the performance of the Random Forest classifier compared to CMIM. Apart from finding biomarkers specific to one location, we found that *Clostridiales*, *Clostridium*, *Roseburia* and *Lactobacillus* could be of interest in the prediction of Type 2 Diabetes irrespective of location. This study verified biomarkers found in previous literature and evaluated several techniques for the prediction of the disease across different regions.

1 Introduction

Recent research has shown that the gut microbiota has an important role in the development of diseases and is linked to health (Wang et al., 2017). The human gut microbiota consists of a complex range of microorganisms like fungi, viruses, archea and bacteria. These microorganisms are associated with nutrient metabolism, drug metabolism, and antimicrobial protection (Jandhyala et al., 2015). Studies have shown that the gut microbiota composition is affected by factors like diet and host genetics (Benson et al., 2010). The dysbiosis of the gut microbiome has negative effects on human health (Wang et al., 2017). This imbalance in the gut microbial composition has been associated with chronic diseases like inflammatory bowel disease (IBD), obesity and neurological disorders (Halfvarson et al., 2017; Wang et al., 2017). We will take a closer look at the association of gut microbiota with Type 2 Diabetes.

Type 2 Diabetes (T2D) is a chronic disease that occurs when the glucose level in the blood is too high (Griffin et al., 2000). This is caused due to the cells in the body becoming resistant to insulin and being unable to effectively utilise it or due to the pancreas being unable to produce sufficient insulin. Major contributing factors that lead to the onset of T2D include diet, lifestyle, genetics and socioeconomic factors (Kolb & Martin, 2017). As of 2021, there were around 537 million adults (10% of adults) that were living with diabetes (Magliano et al., 2021). Around 90% of these cases were T2D. 6.7 million people died due to diabetes that year. Early detection is key in the treatment and may prevent harmful outcomes like cardiovascular diseases, forms of neuropathy and other co-morbidities (Griffin et al., 2000; Safieddine et al., 2021).

Several studies have been conducted on the association of the gut microbiome with T2D to identify biomarkers (Bakir-Gungor et al., 2021; Ge et al., 2022). A study by (N. Larsen et al., 2010) found that there was a significant reduction of phylum *Firmicutes* and *Clostridia* class. Both Karlsson et al. (2013) and Qin et al. (2012) observed an increase in *Clostridium clostridioforme* and a decrease in *Roseburia* in patients with T2D. Several studies also noted an increase in species of *Lactobacillus* in T2D samples (Karlsson et al., 2013; N. Larsen et al., 2010; Lê et al., 2013).

The advancement in sequencing technologies has led to an increase in the availability of microbial data for countless phenotypes. New technologies like shotgun sequencing have increased the accuracy of the classification of microbes (Marcos-Zambrano et al., 2021). Recently many studies have used machine learning to analyse this microbial data in terms of taxonomical and functional diversity to observe the relationship between these communities and human health (Bakir-Gungor et al., 2021). Techniques like Linear Regression, Support Vector Machines and Random Forests to name a few, are widely popular in this field. In particular, Random Forests have been used by both (Qin et al., 2012) and (Karlsson et al., 2013) to predict T2D using genus level and strain level gut microbial data respectively. Studies have also focused on the effect of different feature selection techniques for the identification of biomarkers for T2D. Some techniques include maximum relevance minimum redundancy (mRMR), conditional mutual information maximisation (CMIM), Lasso (Marcos-Zambrano et al., 2021).

In this project we aim to use different machine learning techniques on metagenomic shot- gun sequenced data of samples affected with Type 2 Diabetes and control samples to effectively identify biomarkers that can be used to predict Type 2 Diabetes. In order to effectively fulfil this goal we look into the following aspects:- (i) How does the gut microbiome data in samples with T2D differ from control samples, (ii) Can the results previously found in the literature on the data be reproduced using a Random Forest as described (iii) Can we identify other techniques that improve or produce better results than the Random Forest model (iv) What are the most important features that can be identified as biomarkers and do they correspond to the existing biomarkers found (v) Can we identify biomarkers irrespective of the geographical location of the samples. We hope that the results of this study will help improve the early detection of T2D as well as improve understanding of the association of the gut microbiome with the disease.

2 Methods

2.1 Language and Framework

To answer our research question we have conducted several experiments on the data as described in the following sections. The code that we have used in this study has been written in Python version 3.8. The main libraries that were required are: scikit-learn (Pedregosa et al., 2011), pandas, numpy, scipy and xgboost (Chen & Guestrin, 2016). For visualisation, we used matplotlib, seaborn and alphashape. For feature selection, we used the library scikit-feature (J. Li et al., 2018). Additionally, an overview of our workflow has been depicted in Figure 1.

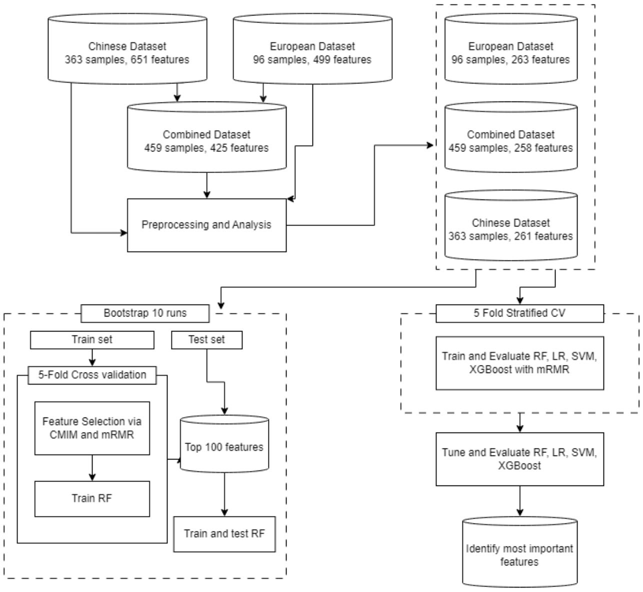

Chinese Dataset
363 samples, 651 features
European Dataset
96 samples, 499 features
European Dataset
96 samples, 263 features
Combined Dataset
459 samples, 425 features
Combined Dataset
459 samples, 258 features
Preprocessing and Analysis
Chinese Dataset
363 samples, 261 features
Bootstrap 10 runs
Train set
Test set
5-Fold Cross validation
Feature Selection via CMIM and mRMR
Train RF
Top 100 features
Train and test RF
5 Fold Stratified CV
Train and Evaluate RF, LR, SVM, XGBoost with mRMR
Tune and Evaluate RF, LR, SVM, XGBoost
Identify most important features

Figure 1: Workflow diagram describing the steps of this study

2.2 Data

The data used in this study was gathered from the curatedMetagenomicsProject (Pasolli et al., 2017). We decided to use 2 data sets, with samples from 2 different locations. Both the data sets that have been used were obtained by metagenomic shotgun sequencing on faecal DNA samples. The data corresponding to the study conducted by Qin et al. (2012) contains 363 samples and 651 features taken from Chinese participants. 170 samples were taken from patients with T2D and the remaining 193 were control samples. Another data set that has been used was published by Karlsson et al. (2013) and contains 145 samples and 499 features from 70-year-old European women. 53 of these samples were affected with T2D, 43 were healthy and 49 were taken from people with Impaired Glucose Tolerance (IGT). IGT is often a precursor to T2D. However, the IGT samples were not used in this project as we did not know whether these patients later developed T2D or were healthy. We also combined the data sets for further analysis by intersecting the features of the European and Chinese sets. This resulted in the creation of a combined data set with 459 samples and 425 features. In this project, we have analysed the relative species abundance of these samples along with their corresponding metadata.

2.3 Pre-processing and Analysis

To get a better understanding of the data, we conducted an initial exploratory analysis. Initially, we analysed the distribution of the class labels with the metadata for each data set. After that, we took a closer look at the features. We calculated the frequency of the value 0 across all samples for all features in the data sets. Features for which the frequency of 0 was more than 95% were removed. This led to the removal of 391 features from the Chinese data set leaving 261 features. From the European data set 237 features were removed resulting in 263 features being left. For the combined data 258 features were left remaining after filtering out 167. We then scaled the data using z-score scaling through the StandardScalar class in Python. We further inspected the mean and variance of the features based on the target class.

In order to visualise the high-dimensional data in 2 dimensions, we used PCA and t-SNE. PCA is a linear dimensionality reduction technique that tries to preserve the global structure of the data, while t-SNE is a non-linear technique that tries to preserve the local structure of the data. We plotted the first 2 dimensions for each method for all the data sets and also graphed the cumulative explained variance for the principal components obtained during PCA.

2.4 Feature Selection

We compared 2 feature selection methods:- minimal redundancy maximum relevance (mRMR) (Peng et al., 2005), Conditional Mutual Information Maximization (CMIM) (Fleuret, 2004). Proposed by Peng et al, mRMR involves the maximum redundancy criterion to select features that have a high correlation with the target class and the minimum redundancy criterion to select features with low correlation between each other (Radovic et al., 2017). CMIM maximises the information of the features selected with the target by iteratively picking features conditional to the features already picked (Fleuret, 2004).

To identify the best method, we performed bootstrapped feature selection 10 times for different train test splits. In each run, we performed 5-fold cross-validation. The 100 features were selected by each feature selection method for the training set in each fold. With this selected set of features, our baseline model, the random forest classifier, was trained. We then tested the set on the validation set in each fold and calculated the cross-validation scores for accuracy and F1 score. When we were selecting the top 100 features, we also obtained the corresponding mutual information score between the selected features and the target variable. The mutual information score is the amount of information shared between 2 variables (Brown et al., 2012). We then calculated the average mutual information for all the features selected in the 5 folds. Using the 100 features that had the highest average mutual information score, we tested the test set. This process was repeated 10 times and the average scores for the test set across all runs were calculated. Finally, to identify the best-performing feature selection algorithm, we compared these scores for both CMIM and mRMR. We used

Wilcoxon signed-rank test to determine the significance of the difference between the scores of both feature selection methods obtained on the test sets.

2.5 Machine Learning Models

In this study, we compared several machine learning techniques in order to evaluate which method can be best used to predict T2D. The baseline method we used was the Random Forest (RF) method. Several existing studies that have used taxanomical data of the gut microbiome found that RFs outperformed other methods in the prediction of phenotypes (Bakir-Gungor et al., 2021; Karlsson et al., 2013). The other methods that we compared were Logistc Regression (LogReg), Support Vector Machines (SVM), and XGBoost (XGB). This comparison allowed us to ensure that the predictions we made and the biomarkers we identified were as accurate as possible.

We used 5-fold cross-validation to obtain a generalised score for the performance of our models on each data set. In each fold, we first performed feature selection using mRMR and selected the top 100 features. After that, we trained and tested each model and recorded the values for the evaluation metrics for that fold. The averages of the scores across all folds were calculated. We also graphed the ROC curve and the Precision-Recall curve. Following this, we performed hyper-parameter tuning on our best-performing model to optimise our results based on the evaluation metrics as described in **Section 2.6**. We then checked the statistical significance of our results to determine if there was a classifier that was performing better than the rest. We then identified the top 20 most important features in each model as potential biomarkers and compared them to find any resemblance between the sets. Furthermore, we compared the features across the 3 data sets to assess their similarities. We then compared our results to that of existing literature and researched more into the biological significance of the potential biomarkers.

2.6 Evaluation Criteria

The following metrics were used to compare and evaluate the models: Accuracy, Area under the ROC curve, F1-Score and Area under the Precision-Recall graph.
Accuracy is the fraction of predictions the model correctly classified. It is defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

AUC is the Area under the Receiver Operating Characteristic (ROC) curve. The ROC curve plots the True Positive Rate against the False Positive Rate at all decision thresholds. The AUC measures the classifiers ability to differentiate between classes. A higher AUC means that the model is better at separating positive and negative classes.
Precision is a metric used to measure what fraction of the positive identification were accurate.

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall, also known as the True Positive Rate, is the fraction of the actual positive class samples were correctly identified.

$$\text{Recall} = \frac{TP}{TP + FN}$$

We use the area under the Precision-Recall curve (AUPRC) as an evaluation metric. It is desirable for the model to have a high AURPRC as it is desirable to have both high Precision and Recall.
The F1 Score is the harmonic mean of the precision and recall and is used for evaluation particularly when the data is unbalanced. Ideally, a higher F1 Score is desirable.

$$\text{F1 Score} = \frac{2 * Precision * Recall}{Precision + Recall}$$

2.7 Statistical Tests

Throughout this study, we have used statistical tests to draw conclusions from our observations and to confirm or reject our hypothesis.

We used the Shapiro-Wilk test to determine if our data followed a Gaussian distribution. The Shapiro-Wilk test is a method of hypothesis testing to see if the data follows a normal distribution. The null hypothesis is that the data set has been generated from a normal distribution. By using this test on all of our data sets, we got p-vlaues lower than 0.05. Hence, we could reject the null hypothesis and our data was likely not normal.

Since our data did not follow a Gaussian distribution, we used Wilcoxon signed-rank test to evaluate if the difference between our feature selection methods were significant. The Wilcoxon signed-rank test is the non-parametric equivalent of the paired t-test. The null hypothesis in this test is that there is no difference between the methods. During the feature selection process, we used this to test if the evaluation metrics generated by the RF model when using mRMR and CMIM, as described in **Section 2.4**, were significantly different and thereby conclude if one was performing better than the other.

In model selection, we used the Friedman test to compare the different classifiers that we used across all 3 data sets (Demšar, 2006). The parametric equivalent of this test is f the repeated-measures ANOVA. This test is based on ranking classifiers based on their performance on each data set independently. This test was conducted separately for the accuracy, F1 score and the AUROC.

3 Results and Discussion

3.1 PCA and t-SNE: Samples cluster based on location

We conducted PCA and t-SNE to visualise the data in 2 dimensions and observe any relationship between the samples. The plots generated by PCA and t-SNE did not show any significant formation of clusters based on the target class for any data set as can be seen in **Figure 2**. However, for the combined data set we did observe that the European and the Chinese samples did cluster separately for both methods, indicating that there is a distinct difference in gut microbial composition based on the geographical location of the patients. This observation corresponds to that of Karlsson et al. (2013), where they also observed the Chinese and European data forming distinct clusters. Studies have shown that a cause for this is due to the difference in dietary habits for different geographical locations (Senghor et al., 2018). It is important to note here that the composition of the data sets differs on other factors like age range and gender. The data from China was taken from adult females and males (Qin et al., 2012). The data from Europe was taken only from 70-year-old females (Karlsson et al., 2013).

3.2 mRMR performs best for feature selection

We compared the effect of using mRMR and CMIM as feature selection methods. We evaluated the performance of the RF model obtained when using these methods to determine which one is best suitable for our data. As described in **Section 2.4**, we calculated the average score on the test set over 10 runs. **Table 1** depicts these observations. Based on these metrics we can see that the mean scores are slightly higher for mRMR as compared to CMIM. However, the intervals of the standard deviations do overlap. Therefore, to be certain that the results were significantly different, we conducted the Wilcoxon signed-rank test. **Table 2** shows the p-values for different metrics obtained using this test. As we can see, all the p-values are lower than 0.05 and hence, are significant. Therefore, we can reject the null hypothesis that the RF model using mRMR performs the same as the RF model combined with CMIM. From this result, we draw the conclusion that the RF model gives better results with mRMR as the feature selection method as compared to CMIM for our data.

Table 1: Average value and standard deviation for the metrics calculated on the RF model for all 10 iterations using different feature selection techniques. These metrics were calculated on the test set of each run before being aggregated

Metrics	**European**		**Chinese**		**Combined**	
	mRMR	**CMIM**	**mRMR**	**CMIM**	**mRMR**	**CMIM**
Accuracy	0.68 ± 0.06	0.54 ± 0.10	0.67 ± 0.04	0.58 ± 0.04	0.64 ± 0.04	0.58±0.06
F1 Score	0.74 ± 0.06	0.59 ± 0.10	0.62 ± 0.05	0.53 ± 0.03	0.62 ± 0.04	0.56± 0.04
AUROC	0.68 ± 0.05	0.54 ± 0.10	0.66 ±0.04	0.58± 0.04	0.64 ± 0.04	0.58± 0.03
AUPR	0.69 ± 0.10	0.61 ± 0.12	0.58 ± 0.05	0.51 ± 0.02	0.58 ± 0.06	0.54 ± 0.04

Table 2: The p-values obtained by performing Wilcoxon signed-rank test to compare the performance of RF with mRMR as a feature selection technique and RF with CMIM have been tabulated. This test has been conducted independently for all 3 data sets.

Metrics	European	Chinese	Combined
Accuracy	0.0098	0.0059	0.0125
F1 Score	0.0059	0.0195	0.0059
AUC	0.0273	0.0098	0.0108
AUPRC	0.0488	0.0098	0.0108

3.3 No significant difference between classifiers

As described in before, we performed 5-fold stratified cross-validation to compare the LogReg, RF, SVM and XGB classifiers. We used mRMR as the feature selection method to select 100 features. **Figure 3** shows the mean ROC curve obtained during the 5-fold cross-validation for all the classifiers as well as the area under the curve for each data set. In all the data sets, the RF and XGB models have a marginally higher AUROC compared that of LogReg and SVM. The AUROC for RF and XGB on the European data are 0.73 and 0.69 respectively. Similarly, the scores for the Chinese data are 0.69 and 0.65, and for the Combined data, 0.64 and 0.66. **Figure 4** graphs the accuracy, F1 scores and AUPRC of the models in the form of a box plot in order to portray the mean and spread of the data. As we can see, there is a lot of overlap between the thresholds of the average scores of the classifiers. Hence, we cannot deduce if any classifiers perform better than the others from these results.

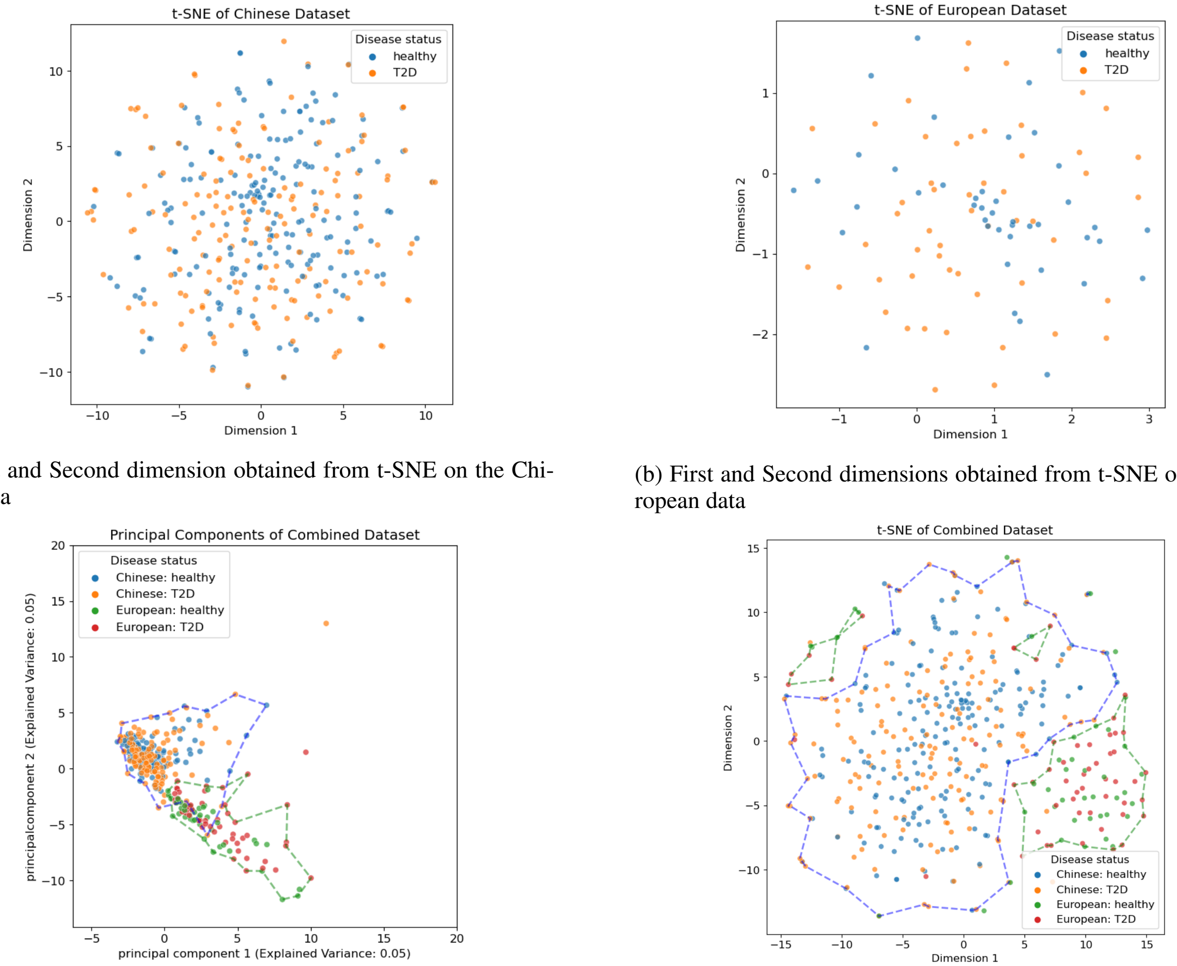
We then tuned the models to optimise the hyper-parameters. **Table 3** compares the performance of the tuned models on our test set. After this, we performed a Friedman test which is a method of hypothesis testing used when multiple classifiers are compared across multiple data sets. The p-value calculated by the Friedman test was not significant. This means that we cannot reject the null hypothesis that there is no significant difference between the classifiers.

Table 3: The table shows the Accuracy, F1 Score, AUROC, and AUPRC for the four classifiers on all the data sets after hyper-parameter tuning. The scores were calculated for a test set after doing cross-validated hyper-parameter tuning on the train set.

		RF	**LogReg**	**SVM**	**XGB**
European	Accuracy	0.76	0.72	0.66	0.72
	F1 Score	0.79	0.73	0.67	0.75
	AUROC	0.76	0.74	0.67	0.73
	AUPRC	0.76	0.75	0.69	0.74
Chinese	Accuracy	0.67	0.72	0.69	0.61
	F1 Score	0.56	0.60	0.54	0.48
	AUROC	0.65	0.68	0.65	0.58
	AUPRC	0.49	0.53	0.50	044
Combined	Accuracy	0.66	0.62	0.66	0.63
	F1 Score	0.63	0.50	0.64	0.62
	AUROC	0.66	0.62	0.66	0.63
	AUPRC	0.60	0.58	0.60	0.58

3.4 Important features were found common to all classifiers

We compared the top 20 features identified by the tuned classifiers. The Venn diagrams depicted in **Figure 5** enlist the 20 most important features identified by the different classifiers for the European and Chinese sets. The intersection of these sets shows the markers that have been identified by multiple models and hence, may be of relevance. The features present in the top 20 features of all the classifiers for each data set have been enlisted in **Table 4**. As we can see there are no common biomarkers identified by all methods across both sets. However, the biomarkers identified per dataset do corroborate with existing literature. For the European data



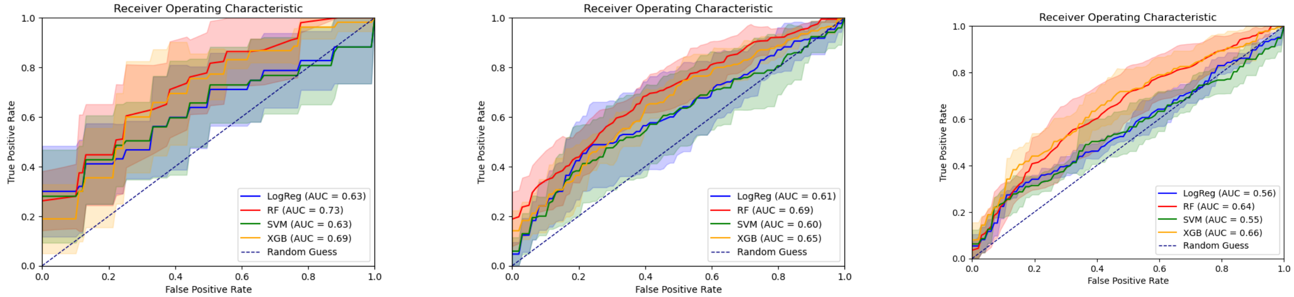
(a) First and Second dimension obtained from t-SNE on the Chinese data

(b) First and Second dimensions obtained from t-SNE on the European data

(c) First and Second principal components obtained from PCA on the Combined data

(d) First and Second dimensions obtained from t-SNE on the combined data

Figure 2: The plots help visualise the high dimensional combined data in 2 dimensions. In plots (a) and (b), the blue samples represent the healthy patients while the orange points correspond to T2D. For plots (c)-(d), the blue and orange points represent the Chinese samples while the red and green points depict the European samples. The blue dashed line outlines the cluster formed by the Chinese data while the green line demarcates the European clusters.

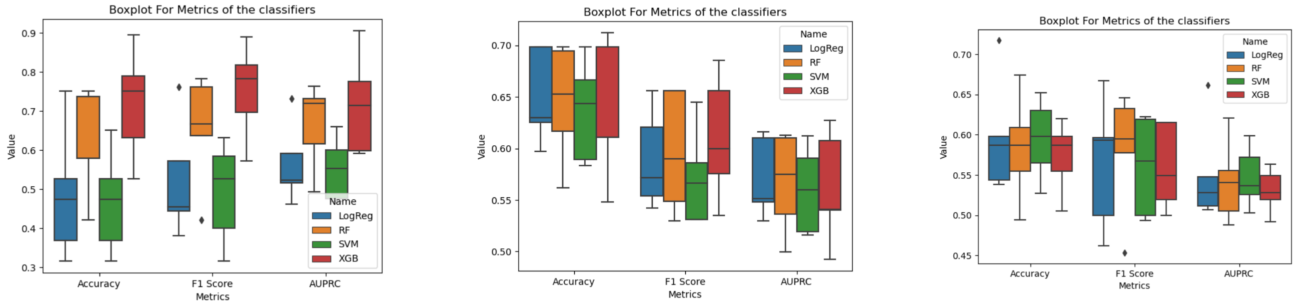


(a) ROC curve of the classifiers on the European data

(b) ROC curve of the classifiers on the Chinese data

(c) ROC curve of the classifiers on the Combined data

Figure 3: The plots show the ROC curve for the four classifiers (LogReg, RF, SVM, XGB) across five folds of cross-validation on each data set. The shaded regions depict the error bands for each curve.

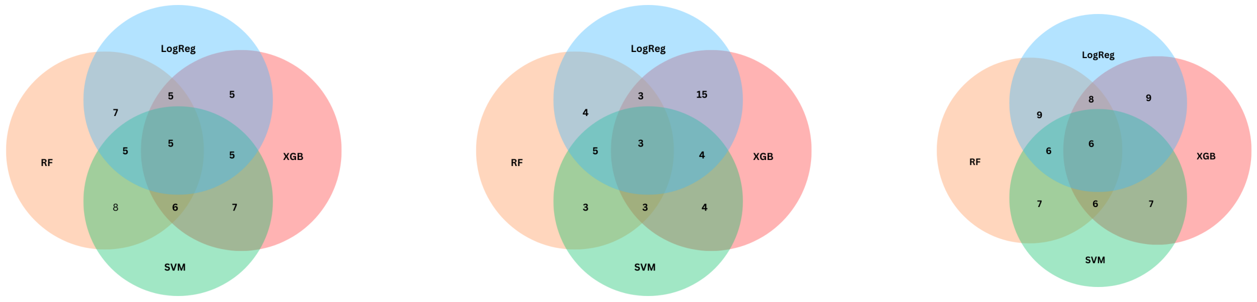


(a) Boxplot for the performance of the classifiers on the European data

(b) Boxplot for the performance of the classifiers on the Chinese data

(c) Boxplot for the performance of the classifiers on the Combined data

Figure 4: Boxplots depicting the distribution of Accuracy, F1 Score, and AUPRC for the four classifiers (LogReg, RF, SVM, XGB) across five folds of cross-validation. Their mean and variability along with outliers are shown.



(a) Venn diagram of the intersected features on the European data

(b) Venn diagram of the intersected features on the Chinese data

(c) Venn diagram of the intersected features on the Combined data

Figure 5: The Venn diagrams show the number of important features common to different combinations of the four classifiers (LogReg, RF, SVM, XGB) after tuning.

Table 4: The features identified in the European and Chinese data set respectively by all tuned classifier

European	Chinese
Alistipes inops *Ruminococcaceae bacterium_D16* *Alistipes shahii* *Faecalibacterium prausnitzii* *Roseburia sp_CAG_182*	*Acidaminococcus sp_CAG_542* *Prevotella bivia* *Lactobacillus mucosae*

Faecalibacterium prausnitzii is a very abundant bacterial species and is known to be a good biomarker for certain conditions (Leylabadlo et al., 2020). It was also identified by Karlsson et al. (2013) for the same data. The genus *Rosburia* was noted by their study as well and this verifies our identification of *Roseburia sp_CAG_182*. Both of these species have been associated with improved insulin sensitivity and are an example of butyrate-producing bacteria (Louis et al., 2010). Although *Alistipes shahii* was not identified as important by Karlsson et al. (2013) for the European data, J. Li et al. (2022) found that the relationship between increased red meat intake and increased HbA1c levels was significantly strengthed by the presence of this species. Since HbA1c is currently used to diagnose patients with T2D, this species may be of relevance (Leong & Wheeler, 2018). The genus *Ruminococcaceae* was also identified in many studies to have a positive correlation with T2D (Esquivel-Hernández et al., 2023; Therdtatha et al., 2021).

For the Chinese data, the genus *Prevotella* was noted to be of importance during PCA, however, Qin et al. (2012) did not find any association between it and T2D. Another study stated that *Prevotella* has a strong association with chronic inflammatory diseases (J. M. Larsen, 2017). Although Qin et al. (2012) did mention any association between *Lactobacillus mucosae* and T2D, Karlsson et al. (2013) did note that many *Lactobacillus* species were associated with the disease.

3.5 *Clostridiales*, *Clostridium*, *Roseburia* and *Lactobacillus* are of interest irrespective of location

Since we were analysing samples taken from different locations, we did expect to see some difference in the biomarkers based on this grouping. We further investigated the generalisation of the markers irrespective of location. The Venn diagram depicted in **Figure 5(b)** shows the number of important features common to different combinations of classifiers for the combined data set. **Table 5** enlists the 6 features identified by all the models for this data. We also compared the

Table 5: The features identified in the Combined data set by all tuned classifiers

Combined
Clostridiales bacterium_1_7_47FAA
Christensenella minuta
Clostridium scindens
Lactobacillus mucosae
Blautia producta
Roseburia sp_CAG_182

most important features identified in all 3 data sets. It is important to note that the initial features of each data set were not identical. In all the data sets the genus *Roseburia* has shown to be of importance. The European and Combined set identifies *Roseburia_sp_CAG_182* and the Chinese set identifies *Roseburia_sp_CAG_303*. While the exact species of these features are unidentified, co-abundance groups have been indicated. Several other genera have been found in the significant features in all data sets. This observation may indicate that genus-level data will provide more generalised biomarkers. Other studies have also identified the genera *Clostridiales*, *Clostridium* and *Lactobacillus* to be of interest (Karlsson et al., 2013; Mora-Ortiz et al., 2019; Qin et al., 2012)

3.6 Responsible Research

To ensure that this research study upholds the integrity and ethos of the scientific community, we have taken the subsequent steps. All external tools, software and libraries in used in this project are open-sourced. The data that we used was published from previous studies and is freely accessible. We have ensured that a detailed description of our methodology and data has been provided, along with all sources from which we obtained any material. We have used all samples of the published data and any filtering and processing of the data has been justified. Through the handling of our data, we try to limit the introduction of any bias. All of these procedures ensure that our work can be easily reproduced and verified.

4 Conclusions

In this study, we aimed to identify biomarkers for T2D that could be later used for early prediction. We utilised techniques like machine learning and feature selection to achieve this goal and we also explored their effects. We took a more generalised approach to the problem and investigated the presence of biomarkers in data taken form 2 different locations.

Through our experiments to answer our third sub-question, we could not identify a classifier that performed significantly better than our baseline model, the Random Forest model. However, we did find that mRMR performed better than CMIM in terms of feature selection. It is also important to note that finding a better classifier is dependent on hyperparameter tuning. Given the short time frame of the project, only a small subset of possible parameters were tested for each model. Hence, to draw more conclusive results, further research can be done in the future.

Our experiments to answer sub-question 4 confirmed several existing biomarkers for each data set independently like *Faecalibacterium prausnitzii* and *Roseburia*. Furthermore, we also confirmed that the genera *Clostridiales*, *Clostridium*, *Roseburia* and *Lactobacillus* were for interest in relation to T2D irrespective of location. Since we conducted species analysis, further investigation of the genus-level abundances would need to be conducted to provide more reliable results.

Overall, our research has shown that there is a significant association between the gut microbiome and T2D. Although machine learning methods may not be accurate enough to predict the presence of T2D based on data obtained from faecal samples, we hope that our results will be of help to clinicians and may assist them in their diagnoses.

References

Bakir-Gungor, B., Bulut, O., Jabeer, A., Nalbantoglu, O. U., & Yousef, M. (2021). Discovering Potential Taxonomic Biomarkers of Type 2 Diabetes From Human Gut Microbiota via Different Feature Selection Methods.

Benson, A. K., Kelly, S. A., Legge, R., Ma, F., Low, S. J., Kim, J., Zhang, M., Oh, P. L., Nehrenberg, D., Hua, K., et al. (2010). Individuality in gut microbiota composition is a complex polygenic trait shaped by multiple environmental and host genetic factors. *Proceedings of the National Academy of Sciences*, *107*(44), 18933–18938.

Brown, G., Pocock, A., Zhao, M.-J., & Luján, M. (2012). Conditional likelihood maximisation: A unifying framework for information theoretic feature selection. *The journal of machine learning research*, *13*, 27–66.

Chen, T., & Guestrin, C. (2016). XGBoost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/2939672.2939785

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine learning research*, *7*, 1–30.

Esquivel-Hernández, D. A., Martınez-López, Y. E., Sánchez-Castañeda, J. P., Neri-Rosario, D., Padrón-Manrique, C., Giron-Villalobos, D., Mendoza-Ortız, C., & Resendis-Antonio, O. (2023). A network perspective on the ecology of gut microbiota and progression of type 2 diabetes: Linkages to keystone taxa in a mexican cohort. *Frontiers in Endocrinology*, *14*, 1128767.

Fleuret, F. (2004). Fast binary feature selection with conditional mutual information. *Journal of Machine learning research*, *5*(9).

Ge, X., Zhang, A., Li, L., Sun, Q., He, J., Wu, Y., Tan, R., Pan, Y., Zhao, J., Xu, Y., Tang, H., & Gao, Y. (2022). Application of machine learning tools: Potential and useful approach for the prediction of type 2 diabetes mellitus based on the gut microbiome profile.

Griffin, S., Little, P., Hales, C., Kinmonth, A., & Wareham, N. (2000). Diabetes risk score: Towards earlier detection of type 2 diabetes in general practice. *Diabetes/metabolism research and reviews*, *16*(3), 164–171.

Halfvarson, J., Brislawn, C. J., Lamendella, R., Vázquez-Baeza, Y., Walters, W. A., Bramer, L. M., D'amato, M., Bonfiglio, F., McDonald, D., Gonzalez, A., et al. (2017). Dynamics of the human gut microbiome in inflammatory bowel disease. *Nature microbiology*, *2*(5), 1–7.

Jandhyala, S. M., Talukdar, R., Subramanyam, C., Vuyyuru, H., Sasikala, M., & Reddy, D. N. (2015). Role of the normal gut microbiota. *World journal of gastroenterology: WJG*, *21*(29), 8787.

Karlsson, F. H., Tremaroli, V., Nookaew, I., Bergström, G., Behre, C. J., Fagerberg, B., Nielsen, J., & Bäckhed, F. (2013). Gut metagenome in European women with normal, impaired and diabetic glucose control - Nature.

Kolb, H., & Martin, S. (2017). Environmental/lifestyle factors in the pathogenesis and prevention of type 2 diabetes. *BMC medicine*, *15*(1), 1–11.

Larsen, J. M. (2017). The immune response to prevotella bacteria in chronic inflammatory disease. *Immunology*, *151*(4), 363–374.

Larsen, N., Vogensen, F. K., Van Den Berg, F. W., Nielsen, D. S., Andreasen, A. S., Pedersen, B. K., Al-Soud, W. A., Sørensen, S. J., Hansen, L. H., & Jakobsen, M. (2010). Gut microbiota in human adults with type 2 diabetes differs from non-diabetic adults. *PloS one*, *5*(2), e9085.

Lê, K.-A., Li, Y., Xu, X., Yang, W., Liu, T., Zhao, X., Tang, Y. G., Cai, D., Go, V. L. W., Pandol, S., et al. (2013). Alterations in fecal lactobacillus and bifidobacterium species in type 2 diabetic patients in southern china population. *Frontiers in physiology*, *3*, 496.

Leong, A., & Wheeler, E. (2018). Genetics of hba1c: A case study in clinical translation. *Current opinion in genetics & development*, *50*, 79–85.

Leylabadlo, H. E., Ghotaslou, R., Feizabadi, M. M., Farajnia, S., Moaddab, S. Y., Ganbarov, K., Khodadadi, E., Tanomand, A., Sheykhsaran, E., Yousefi, B., et al. (2020). The critical role of faecalibacterium prausnitzii in human health: An overview. *Microbial pathogenesis*, *149*, 104344.

Li, J., Li, Y., Ivey, K. L., Wang, D. D., Wilkinson, J. E., Franke, A., Lee, K. H., Chan, A., Huttenhower, C., Hu, F. B., et al. (2022). Interplay between diet and gut microbiome, and circulating concentrations of trimethylamine n-oxide: Findings from a longitudinal cohort of us men. *Gut*, *71*(4), 724–733.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2018). Feature selection: A data perspective. *ACM Computing Surveys (CSUR)*, *50*(6), 94.

Louis, P., Young, P., Holtrop, G., & Flint, H. J. (2010). Diversity of human colonic butyrate-producing bacteria revealed by analysis of the butyryl-coa: Acetate coa-transferase gene. *Environmental microbiology*, *12*(2), 304–314.

Magliano, D. J., Boyko, E. J., Atlas, I. D., et al. (2021). Global picture. In *Idf diabetes atlas [internet]. 10th edition*. International Diabetes Federation.

Marcos-Zambrano, L. J., Karaduzovic-Hadziabdic, K., Loncar Turukalo, T., Przymus, P., Trajkovik, V., Aasmets, O., Berland, M., Gruca, A., Hasic, J., Hron, K., Klammsteiner, T., Kolev, M., Lahti, L., Lopes, M. B., Moreno, V., Naskinova, I., Org, E., Paciência, I., Papoutsoglou, G., … Truu, J. (2021). Applications of machine learning in human microbiome studies: A review on feature selection, biomarker identification, disease prediction and treatment. *Frontiers in Microbiology*, *12*. https://doi.org/10.3389/fmicb.2021.634511

Mora-Ortiz, M., Oregioni, A., & Claus, S. P. (2019). Functional characterisation of gut microbiota and metabolism in type 2 diabetes indicates that clostridiales and enterococcus could play a key role in the disease. *BioRxiv*, 836114.

Pasolli, E., Schiffer, L., Manghi, P., Renson, A., Obenchain, V., Truong, D. T., Beghini, F., Malik, F., Ramos, M., Dowd, J. B., et al. (2017). Accessible, curated metagenomic data through experimenthub. *Nature methods*, *14*(11), 1023–1024.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence*, *27*(8), 1226–1238.

Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., Liang, S., Zhang, W., Guan, Y., Shen, D., Peng, Y., Zhang, D., Jie, Z., Wu, W., Qin, Y., Xue, W., Li, J., Han, L., Lu, D., … Wang, J. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes - Nature.

Radovic, M., Ghalwash, M., Filipovic, N., & Obradovic, Z. (2017). Minimum redundancy maximum relevance feature selection approach for temporal gene expression data. *BMC bioinformatics*, *18*(1), 1–14.

Safieddine, B., Sperlich, S., Epping, J., Lange, K., & Geyer, S. (2021). Development of comorbidities in type 2 diabetes between 2005 and 2017 using german claims data. *Scientific reports*, *11*(1), 11149.

Senghor, B., Sokhna, C., Ruimy, R., & Lagier, J.-C. (2018). Gut microbiota diversity according to dietary habits and geographical provenance. *Human Microbiome Journal*, *7*, 1–9.

Therdtatha, P., Song, Y., Tanaka, M., Mariyatun, M., Almunifah, M., Manurung, N. E. P., Indriarsih, S., Lu, Y., Nagata, K., Fukami, K., et al. (2021). Gut microbiome of indonesian adults associated with obesity and type 2 diabetes: A cross-sectional study in an asian city, yogyakarta. *Microorganisms*, *9*(5), 897.

Wang, B., Yao, M., Lv, L., Ling, Z., & Li, L. (2017). The human microbiota in health and disease. *Engineering*, *3*(1), 71–82.