# Finding biological markers for the prediction of colorectal cancer

## Using machine learning methods to identify functional biomarkers in the human gut microbiome

**Arie Johannes Gijsbert Sloof**[1]
**NetID: Jsloof**

**Supervisors: dr. Thomas Abeel**[1]**, David Calderón Franco**[1]**, Eric van der Toorn**[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 24, 2023

Name of the student: Arie Johannes Gijsbert Sloof
Final project course: CSE3000 Research Project
Thesis committee: dr. Thomas Abeel, David Calderón Franco, Eric van der Toorn, Thomas Höllt

## Abstract

Colorectal cancer (CRC), one of the leading causes of mortality, is challenging to diagnose. By using metagenomic analysis with machine learning methods, this can be done in a non-invasive manner. In this research, a neural network has been trained on relative pathway abundance data, a way to measure the functional potential of a microbiome, in order to find biomarkers for colorectal cancer. The accuracy achieved by the neural network is 57%. The most important features used by the model are compared to established biomarkers in literature. Besides overlapping pathways, this research also found new potential biomarkers for CRC.


## 1 Introduction

Colorectal cancer (CRC), characterized by the uncontrolled growth of cells in the colon or rectum, ranks as the second leading cause of cancer-related mortality in the western world (Vidnes et al., 2013). Diagnosis of CRC with the current methods, like colonoscopy, an invasive procedure, is rather challenging (Vega et al., 2015). Mentioned research also indicates that the available models used for predicting CRC evaluate the symptoms, which presence means the cancer is not in the earliest stages anymore. An non-invasive way of predicting the presence of CRC is by applying machine learning and feature selection on metagenomic data, of which one route is described in this paper.

By looking at human data of microbiome composition, early stages of diseases can be predicted (Xu et al., 2020). The human body contains trillions of microbes, of which a lot of data is available (Dai et al., 2021). Some indicators for the presence of diseases are already found by using different kinds of machine learning (Marcos-Zambrano et al., 2021). According to this research, for CRC in specific a lot of statistic methods and random forests are used to find biological markers.

The problem this research project tackles, is to verify the biological markers that have already been found for CRC by implementing a different type of machine learning which has not been used yet in the process of finding the indicators. This problem is worth to solve, because it is important for the early detection and prevention of CRC. If indicators in the microbiome composition are known, the detection and prevention of CRC is easier and more affordable (Bin Ashraf et al., 2020).

Many studies have been done in this area already, especially in the broader sense for the generation of metagenomic data and search for indicators for different sorts of diseases (Marcos-Zambrano et al., 2021). In this article, a lot of studies that have been done to find biomarkers are listed. The main machine learning method used there is random forests. Different kinds of biomarkers have been found, including 144 species and 75 genera, the taxonomic category between species and families, (Dai et al., 2021) but also functional markers (Loftus et al., 2021). Functional markers refer to specific genetic features that are associated with particular functional traits or activities in the microbiome. Especially functional markers are useful to predict CRC is suggested by multiple studies (Allali et al., 2018) (Wirbel et al., 2019) (Loftus et al., 2021) (N.-N. Liu et al., 2022). The more advanced machine learning techniques like neural networks and deep learning are underrepresented (Marcos-Zambrano et al., 2021). This is why this research focuses on applying a neural network to functional data.

There are some pathway abundances that stand out for people with CRC compared to healthy people according to literature, of which in this paragraph a few are mentioned. Pathway abundances are a way to measure the functional potential of a microbiome, they are a quantitative measurement of the relative levels of pathways (series of interconnected biochemical reactions or processes). First of all virulence factors and peptide degradation are more present, and functions involved in amino-acid biosynthesis are less present in the CRC gut microbiome (Loftus et al., 2021). Wirbel et al. (2019) suggested pathways for the degradation of amino acids, mucins and organic acids have a higher abundance in the CRC gut microbiome. Besides this, Zhang et al. (2020) suggested the pathway involved in cell motility has higher abundance in the CRC gut microbiome than in the control microbiome while the pathway involved in carbohydrate metabolism was less present. Finally, the aromatic amino acid metabolism is associated with CRC (Yachida et al., 2019). Besides this this research also found that sulfide-producing pathways are abundant in CRC gut microbiomes.

This research project aims to investigate whether neural networks with wrapper methods for feature selection can be utilized to analyze a metagenomic dataset in order to verify functional biological markers for the disease CRC. This paper tries to answer (i) how a logistic regression model with a wrapper for feature selection can be implemented, trained, and tested on a metagenomic dataset to classify diagnosed and healthy samples, (ii) how a neural network model with a wrapper for feature selection can be implemented, trained, and tested on a metagenomic dataset to classify diagnosed and healthy samples, (iii) how the neural network model performs on evaluation metrics such as the confusion matrix, accuracy, precision, recall, and F1 score when compared to the logistic regression baseline model and finally, (iv) what the most significant features identified by the feature selection process in the metagenomic dataset are, and if they align with the biomarkers mentioned in existing literature.

This report describes the research leading up to the answers to the questions in multiple sections. Multiple feature selection techniques are compared in terms of performance for the logistic regression and neural network model. The most important selected features are presented and compared to biomarkers for CRC in literature.

## 2 Materials and Methods

### 2.1 Programming language and tools

The software has been developed in the programming language Python (version 3.10), in combination with Scikit Learn (version 1.2.2) (Buitinck et al., 2013), as this package is easy to use, has high performance and is docu-

mented well (Pedregosa et al., 2011). Libraries used for the research are: Pandas (version 1.5.3) (pandas development team, 2020), NumPy (version 1.23) (Harris et al., 2020), Matplotlib (version 3.7.1) (Hunter, 2007), Seaborn (version 0.12.2) (Waskom, 2021) Scipy (version 1.10.1) (Virtanen et al., 2020), Pip (version 23.0.1) ("pip: The Python package installer", n.d.) and mRMR (version 0.2.6) ("mrmr: mRMR (minimum-Redundancy-Maximum-Relevance) for automatic feature selection at scale", n.d.).

## 2.2 Data and metadata

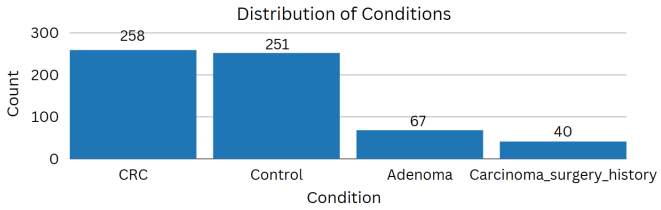
The data used to train the models on is from CuratedMetagenomicData (Pasolli et al., 2017), a package providing curated human microbiome data. The dataset used is a fecal shotgun metagenomic study of CRC with 616 runs ("Available studies", n.d.), called YachidaS_2019 (Yachida et al., 2019). The data is collected from 616 unique people from Japan, with the following study conditions: Adenoma, carcinoma surgery history, CRC and control. The data shows relative abundance for pathways. Pathway abundance data was used as many research suggested this data is useful to predict CRC (Allali et al., 2018) (Wirbel et al., 2019) (Loftus et al., 2021) (N.-N. Liu et al., 2022). The samples with CRC are used as case samples and the ademona and carcinoma surgery history samples are not used, such that 509 samples are left. Only CRC and control data is used because this way the split between case and control is more even, as is visible in **Figure 1a**. Besides this it is unclear if the samples with carcinoma surgery history are healthy now or still have CRC.

The data is further preprocessed by merging columns of species from the same pathway abundance together, adding up the values. Only the pathway description is left in the column names, so the bacterial taxonomy is not there. This is done because now the selected features can be compared directly to pathways associated with CRC in literature and it is reduces the amount of features. Also the unintegrated and unmapped feature columns are removed, as these do not provide useful information and are presumably very different per data set. This ensures the amount of features is reduced from 31291 to 506.

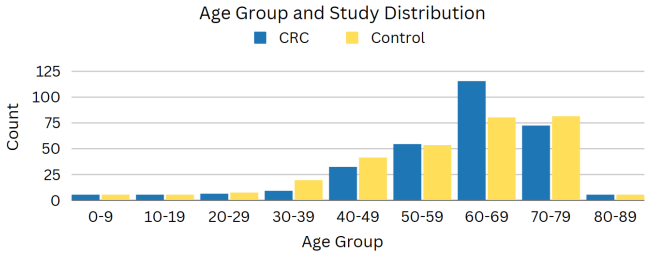
Some insights about study conditions, age and gender of the participants for the data collection are plotted. In **Figure 1a** the counts of samples with the different study conditions is visible, whereas in respectively **Figure 1b** the age is plotted together with the study conditions. The graph shows the data is spread evenly over age, which means the model will not be biased towards a certain age. The genders are also evenly represented.

## 2.3 Feature selection

Feature selection is applied because it can improve the performance of the model, the predictors are faster and the underlying process which generated the data is more understandable (Guyon and Elisseeff, 2003). For this research, amongst others wrapper methods Recursive Feature Elimination (RFE) and Forward Feature Selection (FFS) are applied for feature selection as they are commonly used (Kaushik, 2016). RFE is for both the neural network and the logistic regression model wrapped around a logistic regression model with L1 penalty,




(a) Distribution of the study conditions of the participants involved in the data collection: CRC, control, adenoma and carcinoma surgery history.




(b) Distribution of the age and study condition of the participants involved in the data collection. Per age group, counts of CRC (in blue) and Control (in yellow) samples are plotted.

Figure 1: Distributions of study conditions.

saga solver and a C value of 1. FFS is for both models wrapped around a logistic regression model with L1 penalty, liblinear solver and a C value of 10. Besides wrapper methods, the filter methods variance filtering and minimum Redundancy Maximum Relevance (mRMR) are used in order to compare to in terms of performance.

Wrapper methods are used to select the features for the model training, for two reasons. First of all, wrapper methods can be wrapped around any machine learning model, making it possible to compare the performances of multiple models, opposed to embedded methods, which are only applicable for specific models. Besides this, in wrapper methods, different combinations of features are compared in terms of performance of the machine learning model used in the actual training. This enables wrapper methods to always be able to select the best subset of features, whereas filter methods might fail to do this (Alshamy and Ghurab, 2020).

Comparing feature selection algorithms can only be done after implementing the baseline (logistic regression), which is described in section 2.5, because this way the performance of the model can be tested against the selected features. Comparing the feature selection algorithms is done using 5 fold stratified cross validation based on accuracy. The cross validation is stratified to ensure an equal split between case and control samples in the different partitions. The feature selection techniques are also compared to the performance of the model on all features. Because of the long training times of the neural network, it was impossible to do bootstrapping within the limited time of this research.

## 2.4 Dimension reduction

Dimension reduction techniques Principal Component Analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE) are used to visualise the data in order to have a linear and non-linear visualisation. Silhouette scores, representing the average distance between clusters compared to the average distance between samples in different clusters, are calculated to have an indication of clustering quality.

The dimension reduction technique PCA is also applied to select features. PCA is used to construct 100 linear components from the features, enough for approximately 98% of the variance as can be seen in **Figure 2**.

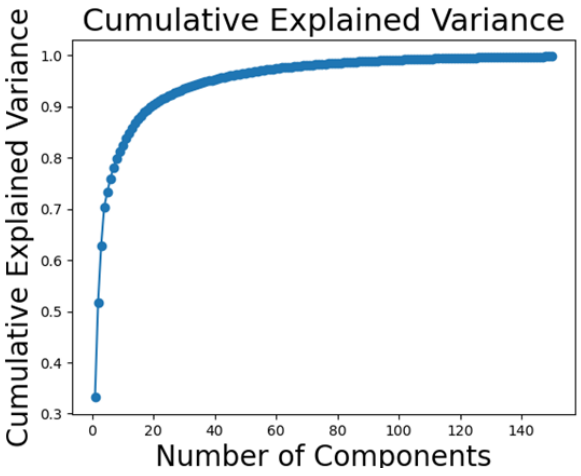



Figure 2: Number of principal components vs. cumulative explained variance.

PCA is only used to compare to as feature selection method in terms of classification performance, not in terms of found biomarkers. Using dimension reduction on the data set means losing the actual features and possibly losing information about important biomarkers. Even though it is possible to get back to information about original features (Amoeba, 2017), the principal components may not correspond to biologically meaningful features.

## 2.5 Machine learning models

To be able to see if the neural network implemented as main focus in this project is a good way to classify the data and find biomarkers, a logistic regression model was implemented as baseline. Accuracy, precision, recall and F1 score metrics are calculated for this algorithm, later described.

Before the algorithm is trained, the features are scaled using min-max scaling because this scores better than standard scaling. The train and test data are scaled separately, in order to prevent for bias.

The best performing hyperparameters for logistic regression are found using grid search, a crossvalidation on different values for the hyperparameters. The performance for C values of 0.1, 1.0 and 10 are determined for the liblinear and saga solvers with L1 regularization and the lbgfs, sag and newton-cg solvers with L2 regularization.

Table 1: Hyper parameter settings used for logistic regression combined with different feature selection techniques.

| Feature Selection Technique | Hyperparameter Settings |
|---|---|
| **No feature selection** | LogisticRegression(random_state=16, max_iter=100, penalty='L2', solver=lbfgs, C=1) |
| **Variance filtering** | LogisticRegression(random_state=16, penalty='l1', solver=liblinear, C=10) |
| **RFE** | LogisticRegression(random_state=16, penalty='l1', solver='saga', C=1) |
| **FFS** | LogisticRegression(random_state=16, penalty='l1', solver=liblinear, C=10) |
| **mRMR** | LogisticRegression(random_state=16, penalty='l1', solver='saga', C=1) |
| **Rfe step 100** | LogisticRegression(random_state=16, penalty='l1', solver='liblinear', C=10) |
| **PCA** | LogisticRegression(random_state=16, penalty='l1', solver='liblinear', C=10) |

Table 2: Hyper parameter settings used for the neural network combined with different feature selection techniques.

| Feature Selection Technique | Hyperparameter Settings |
|---|---|
| **No feature selection** | MLPClassifier(hidden_layer_sizes=(50,), alpha=1, solver='adam', activation='relu', random_state=16) |
| **Variance filtering** | MLPClassifier(hidden_layer_sizes=(50,), alpha=1, solver='adam', activation='relu', random_state=3) |
| **RFE** | (LogisticRegression(random_state=16, penalty='l1', solver='saga', C=1.0), MLPClassifier(hidden_layer_sizes=(35,), alpha=0.01, solver='adam', activation='tanh', random_state=16) |
| **FFS** | LogisticRegression(random_state=16, penalty='l1', solver=liblinear, C=10) MLPClassifier(hidden_layer_sizes=(35,), alpha=0.01, solver='adam', activation='tanh', random_state=16) |
| **mRMR** | MLPClassifier(hidden_layer_sizes=(35,), alpha=0.5, solver='adam', activation='tanh', random_state=16) |
| **Rfe step 100** | MLPClassifier(hidden_layer_sizes=(35,), alpha=0.5, solver='adam', activation='tanh', random_state=16) |
| **PCA** | MLPClassifier(hidden_layer_sizes=(50,), alpha=1, solver='adam', activation='relu', random_state=3) |

The used hyperparameter settings for the logistic regression model in combination with the different feature selection techniques can be found in **Table 1**. RFE step 100 means after the model is trained, 100 features are eliminated, in stead of only 1 with default settings. This is done because it took too much time to select features when eliminating one feature at a time.

The main algorithm that implemented is a neural network, specifically a multi layer perceptron. The network is trained on the same preprocessed data as the logistic regression model is trained on and the metrics used are the same as mentioned before.

There are two main reasons for choosing a neural network as classification model. First of all, limited evidence exists in scientific literature regarding the application of neural networks for identifying biological markers to predict CRC. Besides this, neural networks (in this case a multi layer perceptron) are successfully used for the finding of biological markers for other diseases (B. Liu et al., 2022), so it has potential to be a useful method.

The best performing hyperparameters for the neural network are found using grid search as well. For the activation function relu and tanh have been applied, alpha values used are 0.01, 0.1 and 1.0, adam and stochastic gradient descent solvers are evaluated and performance for different settings for the amount of layers and nodes is determined. The used hyper parameter settings for the neural network model in combination with the different feature selection techniques can be found in **Table 2**.

## 2.6 Statistical analysis

In order to compare cross validation results, T-tests were done. If the p-value would lie below 0.05, the difference would be significant. 0.05 is used, because it is a common value used in statistics (Glen, n.d.). This value is corrected using Bonferonni correction (Hayes, 2010) to prevent comparisons from incorrectly appearing statistically significant.

### 2.7 Ranking features based on importance

For both logistic regression and the neural network a metric is used to determine which of the used features are the most important. For logistic regression the absolute coefficients are used as feature importance (Filho, 2023). The higher the absolute coefficient, the more important the feature is. For neural networks there is no such property, so a different metric is needed: permutation importance (Fisher et al., 2018). Permutation importance calculates importance of a feature by shuffling the values of a feature 100 times, each time calculating the accuracy of the model. The absolute difference between the average accuracy of the 100 permutations and the original accuracy with all unchanged features is the importance of the feature. The bigger the difference between the average permutation accuracy and the original accuracy, the more important the feature is. The importance of the features is determined by running the aforementioned 5 fold stratified cross validation and averaging the importance.

### 2.8 Metrics

Confusion matrix, accuracy, precision, recall and F1 score metrics are calculated for both baseline and the neural network based on unseen test data. These metrics are based on True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN). Confusion matrices show from left top to the right bottom the TP, FN, FP and TN.

- Accuracy:
$$\frac{TP + TN}{TP + TN + FP + FP}$$
- Precision:
$$\frac{TP}{TP + FP}$$
- Recall:
$$\frac{TP}{TP + FN}$$
- F1 score:
$$2 * \frac{Recall * Precision}{Recall + Precision}$$

## 3 Results and Discussion

### 3.1 Case and control data points show a lot of overlap

The plot of the first 2 principal components can be seen in **Figure 3a**. The first principal component accounts for 32.54% of the variance, while the second principal component explains 17.99% of the variance. Around the center of the plot the data points show overlap, which indicates similarity in their characteristics. When moving away from the middle of the plot, the data is more spread out, suggesting more variability in the features. No clear clustering of case or control data points is visible, which means training linear machine learning models can be challenging. The silhouette score is 0.0016, which is relatively low and indicates the clustering quality is low.

The t-SNE plot can be seen in **Figure 3b**. Like in the PCA plot, no clustering is visible and the data is not easily separable non linearly either. The silhouette score of this dimension reduction is 0.0013, which also indicates a low level of clustering quality.

The presence of overlap between case and control data points in both the PCA and t-SNE plots suggests that there is a lack of clear separation between the two groups based on the pathway abundances. This implies that pathway abundances used for analysis may not be strongly indicative of the disease status or that there are other confounding factors influencing the data. A confounding factor can be for example biological variability: individual genetic differences or variations in disease progression can lead to overlapping pathway abundances in case and control groups.

Table 3: Cross validation and testing performance of different feature selection methods for logistic regression.

| Cross Validation (CV) and Testing Scores for Logistic Regression | | | | | | | |
|---|---|---|---|---|---|---|---|
| | No feature selection | Variance filtering | RFE | RFE (step 100) | FFS | MRMR | PCA |
| CV mean accuracy | 0.51 | 0.58 | 0.56 | 0.53 | 0.49 | 0.55 | 0.55 |
| CV SD | 0.057 | 0.046 | 0.063 | 0.056 | 0.073 | 0.042 | 0.046 |
| Test accuracy | 0.60 | 0.64 | 0.56 | 0.53 | 0.56 | 0.52 | 0.57 |
| Test precision | 0.61 | 0.66 | 0.57 | 0.54 | 0.57 | 0.52 | 0.58 |
| Test recall | 0.60 | 0.60 | 0.57 | 0.54 | 0.55 | 0.54 | 0.58 |
| Test F1 score | 0.60 | 0.63 | 0.57 | 0.54 | 0.56 | 0.53 | 0.58 |

Table 4: Cross validation and testing performance of different feature selection methods for the neural network.

| Cross Validation (CV) and Testing Scores for Neural Network | | | | | | | |
|---|---|---|---|---|---|---|---|
| | No feature selection | Variance filtering | RFE | RFE (step 100) | MRMR | PCA | FFS |
| CV mean accuracy | 0.53 | 0.57 | 0.57 | 0.54 | 0.56 | 0.55 | 0.54 |
| CV SD | 0.051 | 0.064 | 0.063 | 0.023 | 0.032 | 0.060 | 0.049 |
| Test accuracy | 0.61 | 0.62 | 0.54 | 0.52 | 0.53 | 0.62 | 0.54 |
| Test precision | 0.63 | 0.62 | 0.55 | 0.53 | 0.54 | 0.62 | 0.55 |
| Test recall | 0.57 | 0.63 | 0.55 | 0.52 | 0.55 | 0.68 | 0.52 |
| Test F1 score | 0.60 | 0.63 | 0.55 | 0.53 | 0.55 | 0.65 | 0.54 |

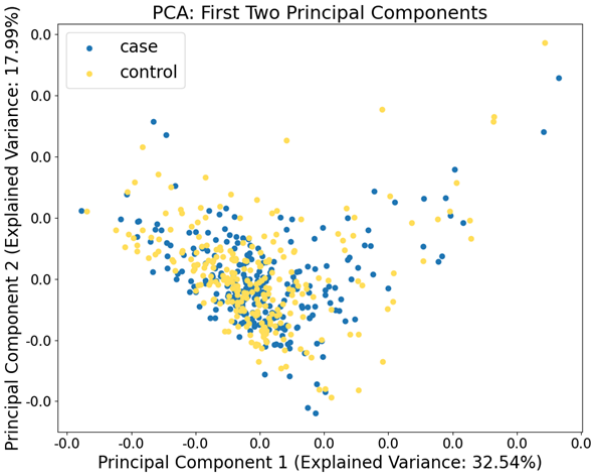
### 3.2 Logistic regression and neural network score similarly low on accuracy with and without feature selection

In **Table 3** the performance of the different feature selection techniques is visible. The columns stand for the methods used for feature selection, described in 2.3. The upper two rows show the stratified cross validation accuracy mean scores for 5 runs and their standard deviation. The 4 lowest rows are scores of the performance of the logistic regression model on the separate test data partition. The performance of the neural network model is visible in **Table 4**.

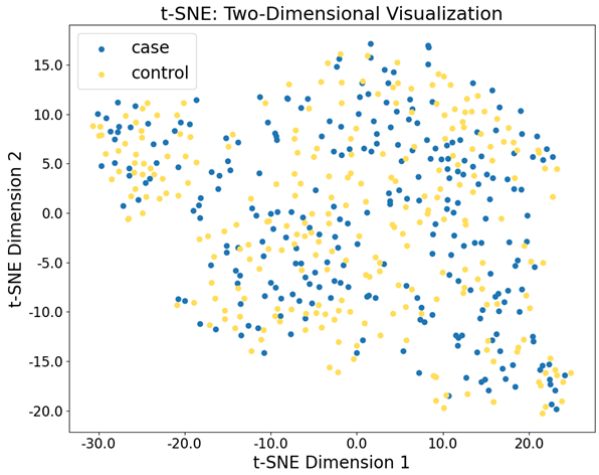
The confusion matrices resulting from runs of the logistic regression model with different feature selection techniques on test data are also visible in **Figure 4**. In **Figure 5** the matrices for the neural network are visible. The matrices show the predictions are not completely random: the number of correctly predicted case and control samples on the are mostly higher than the false positives and negatives, as can be seen from the darker blue colored diagonals.

The cross validation scores are compared for no feature selection against all applied feature selection techniques. This is done for both the logistic regression model, in **Figure 5** and the neural network model in **Figure 6**. All p values lie above 0.0083, so no feature selection technique is significantly better than applying no feature selection at all. 0.0083 is 0.05 divided by 6, the amount of comparisons, for the sake of the Bonferroni correction.

The neural network has been compared to the logistic regression model as well using t-tests. The variance filtering




(a) The pathway abundance data visualized according to the first two principal components, in total explaining 50.53% of the variance. Yellow datapoints represent control samples and blue datapoints represent case samples.




(b) The pathway abundance data visualized using t-SNE. Yellow datapoints represent control samples and blue datapoints represent case samples.

Figure 3: The data visualised with PCA and t-SNE.

Table 5: Logistic regression cross validation scores of no feature selection compared to other feature selection techniques using t-tests.

| Feature selection technique | t-statistic | p-value |
|---|---|---|
| Variance | -2.6 | 0.061 |
| RFE | -1.1 | 0.35 |
| RFE step 100 | -1.1 | 0.32 |
| FFS | 0.64 | 0.56 |
| mRMR | -1.2 | 0.29 |
| PCA | -2.0 | 0.11 |

Table 6: Neural network cross validation scores of no feature selection compared to other feature selection techniques using t-tests.

| Feature selection technique | t-statistic | p-value |
|---|---|---|
| Variance | -1.5 | 0.20 |
| RFE | -1.2 | 0.27 |
| RFE step 100 | -0.38 | 0.72 |
| FFS | -0.93 | 0.41 |
| mRMR | -1.3 | 0.25 |
| PCA | -1.5 | 0.21 |

results got a t-statistic of 0.78 and a p-value of 0.48, and RFE got a t-statistic of -0.54 and a p-value of 0.62. Both p-values lie above 0.025, so no model is significantly performing better than the other. 0.025 is 0.05 divided by 2, the amount of comparisons, for the sake of the Bonferroni correction. If the experiment would be done more often with bootstrapping, the standard deviations might be lower and this results could be significantly different. This is not done, because of the long training times of the neural network.

With scores just above 0.5, the models do not perform much better than random. This could be due to multiple reasons. First of all the chosen models might not be ideal for this classification problem. The data might not be separable linearly, which makes it impossible for the logistic regression model to correctly classify the data. For a neural network enough training data is needed to train on in order to perform well. Besides this CRC is a complex disease (Fearon and Vogelstein, 1990), and the relation between the features and target variable could be complicated, which makes it hard for simple models to make accurate predictions. Finally by merging the features like described in Section 2.2, it is possible the complex relation between features and target variable is oversimplified. By merging the features, important information could have been lost and have made it harder to capture subtle differences between CRC and control samples.

### 3.3 Top 10 most important features from both models selected by variance filtering overlap with 4 pathway abundances from literature

The 10 most important selected features using variance filtering are compared to literature, and 4 pathways appear to overlap with literature. For this comparison variance filtering is used as this gave the biggest overlap with literature. In **Figure 6** the most important pathway abundances are visible. 4 of them overlap with literature, firstly (i) the Su-

perpathway of histidine, purine and pyrimidine biosynthesis (Ugbogu et al., 2022), next to that (ii) the Superpathway of purine nucleotide salvage (Eroglu et al., 2000), thirdly (iii) the Superpathway of UDP-N-acetylglucosamine-derived O-antigen building blocks biosynthesis (Naka et al., 2020) and finally (iv) the superpathway of L-tryptophan biosynthesis (Gonzalez-Mercado et al., 2021). In the venn diagram can be seen 3 other pathways are related to other diseases: diabetes (Chang et al., 2015), prostate cancer (Kim et al., 2019) and gastric cancer (Nie et al., 2021).

3 other pathways are in the top 10 most important features for both models: firstly (i) the superpathway of pyridoxal 5'-phosphate biosynthesis and salvage, secondly (ii) the superpathway of mycolate biosynthesis and finally (iii) peptidoblycan biosynthesis V. Because the models have captured relevant pathways which are known biomarkers for CRC, these three pathways might be relevant for investigation as well. However, with accuracy scores not much above 0.5, the models do not show a lot of predictive power. Therefore, further investigation and validation are necessary to confirm the mentioned pathways are indeed connected to CRC.

### 3.4 Responsible Research

In order to conduct responsible research, the data used in this study has been sourced from CuratedMetagenomicData (Pasolli et al., 2017), a reputable and publicly accessible dataset. The dataset can be downloaded from the repository and the same preprocessing steps as described in section 2 can be applied to obtain the processed dataset used in this research. This availability and transparency contribute to the reproducibility and verifiability of the research findings.

Furthermore, model training and evaluation have been conducted using widely accepted best practices. Multiple models and feature selection algorithms have been tested and compared to identify the most accurate approach. Stratified cross-validation has been used to assess model performance consistently and have a balanced representation of the classes. Also, the data was split into a separate train and test partition, to ensure the performance could be measured and evaluated on unseen data. The used random states have been reported as well, in order to make the results reproducible. These measures help to ensure the reliability and generalizability of the model's performance.

In order to make this research as transparent and reproducible as possible, the code for implementing the machine learning models has been made publicly available on GitLab (Sloof, 2023). Detailed documentation of the steps involved in getting the data and training the models have been provided, allowing other researchers to replicate the experiments and verify the findings.

## 4 Conclusions and Future Work

In this paper, we conducted research on a metagenomic dataset in order to find biomarkers for the disease Colorectal Cancer (CRC). Two machine learning models, logistic regression and a multi layer perceptron were trained on relative pathway abundance data. The models did not perform too well classification wise, with 58% accuracy for the logistic regression model and 57% accuracy for the neural network, which is not significantly different to each other. Using feature importance, the top 10 selected features of both models were compared to literature, showing an overlap of 4 pathways. The models overlapped with each other as well, indicating 3 pathways might have a connection to CRC: firstly (i) the superpathway of pyridoxal 5'-phosphate biosynthesis and salvage, secondly (ii) the superpathway of mycolate biosynthesis and finally (iii) peptidoblycan biosynthesis V. Whether these pathways are connected to CRC should be investigated further, because the predictive power of the models is relatively low and this is only a first indication for being connected to the disease. Further investigation can for example be done on different data sets, to see if the found pathways are more or less dominant for CRC patients compared to control patients.

## References


Allali, I., Boukhatem, N., Bouguenouch, L., Hardi, H., Boudouaya, H. A., Cadenas, M. B., Ouldim, K., Amzazi, S., Azcarate-Peril, M. A., & Ghazal, H. (2018). Gut microbiome of moroccan colorectal cancer patients. *Med. Microbiol. Immunol.*, *207*(3-4), 211–225.

Alshamy, R., & Ghurab, M. (2020). A review of big data in network intrusion detection system: Challenges, approaches, datasets, and tools.

Amoeba. (2017). How to reverse pca and reconstruct original variables from several principal components? https://stats.stackexchange.com/q/229093

Available studies [Accessed: 2023-6-16]. (n.d.). https://waldronlab.io/curatedMetagenomicData/articles/available-studies.html

Bin Ashraf, F., Shafi, M. S. R., & Kabir, M. R. (2020). Host trait prediction from human microbiome data for colorectal cancer [23rd International Conference on Computer and Information Technology (ICCIT), Ahsanullah Univ Sci & Technol, ELECTR NETWORK, DEC 19-21, 2020]. *2020 23RD INTERNATIONAL CONFERENCE ON COMPUTER AND INFORMATION TECHNOLOGY (ICCIT 2020)*. https://doi.org/10.1109/ICCIT51783.2020.9392731

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., VanderPlas, J., Joly, A., Holt, B., & Varoquaux, G. (2013). API design for machine learning software: Experiences from the scikit-learn project. *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 108–122.

Chang, H. .-., Chao, H. .-., Walker, C. S., Choong, S. .-., Phillips, A., & Loomes, K. M. (2015). Renal depletion of myo-inositol is associated with its increased degradation in animal models of metabolic disease. *AMERICAN JOURNAL OF PHYSIOLOGY-RENAL PHYSIOLOGY*, *309*(9), F755–F763. https://doi.org/10.1152/ajprenal.00164.2015

Dai, D., Zhu, J., Sun, C., Li, M., Liu, J., Wu, S., Ning, K., He, L.-j., Zhao, X.-M., & Chen, W.-H. (2021). GMrepo v2: a curated human gut microbiome database with special focus on disease markers and cross-dataset comparison. *Nucleic Acids Research*, *50*(D1), D777–D784. https://doi.org/10.1093/nar/gkab1019

Eroglu, A., Canbolat, O., Demirci, S., Kocaoglu, H., Eryavuz, Y., & Akgul, H. (2000). Activities of adenosine deaminase and 5 '-nucleotidase in cancerous and noncancerous human colorectal tissues. *MEDICAL ONCOLOGY*, *17*(4), 319–324. https://doi.org/10.1007/BF02782198

Fearon, E. R., & Vogelstein, B. (1990). A genetic model for colorectal tumorigenesis. *Cell*, *61*(5), 759–767.

Filho, M. (2023). How to get feature importance in logistic regression [Accessed: 2023-6-24]. https://forecastegy.com/posts/feature-importance-in-logistic-regression/

Fisher, A., Rudin, C., & Dominici, F. (2018). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously.

Glen, S. (n.d.). https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/what-is-an-alpha-level/

Gonzalez-Mercado, V. J., Lim, J., Yu, G., Penedo, F., Pedro, E., Bernabe, R., Tirado-Gomez, M., & Aouizerat, B. (2021). Co-occurrence of symptoms and gut microbiota composition before neoadjuvant chemotherapy and radiation therapy for rectal cancer: A proof of concept. *BIOLOGICAL RESEARCH FOR NURSING*, *23*(3), 513–523. https://doi.org/10.1177/1099800421991656

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *J. Mach. Learn. Res.*, *3*(null), 1157–1182.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hayes, A. (2010). What is the bonferroni test (correction) and how is it used? [Accessed: 2023-6-24]. https://www.investopedia.com/terms/b/bonferroni-test.asp

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Kaushik, S. (2016). Introduction to feature selection methods with an example (or how to select the right variables?) [Accessed: 2023-6-24]. https://www.analyticsvidhya.com/blog/2016/12/introduction-to-feature-selection-methods-with-an-example-or-how-to-select-the-right-variables/

Kim, S., Yang, X., Yin, A., Zha, J., Beharry, Z., Bai, A., Bielawska, A., Bartlett, M. G., Yin, H., & Cai, H. (2019). Dietary palmitate cooperates with src kinase to promote prostate tumor progression. *PROSTATE*, *79*(8), 896–908. https://doi.org/10.1002/pros.23796

Liu, B., Chau, J., Dai, Q., Zhong, C., & Zhang, J. (2022). Exploring gut microbiome in predicting the efficacy of immunotherapy in non-small cell lung cancer. *CANCERS*, *14*(21). https://doi.org/10.3390/cancers14215401

Liu, N.-N., Jiao, N., Tan, J.-C., Wang, Z., Wu, D., Wang, A.-J., Chen, J., Tao, L., Zhou, C., Fang, W., Cheong, I. H., Pan, W., Liao, W., Kozlakidis, Z., Heeschen, C., Moore, G. G., Zhu, L., Chen, X., Zhang, G., … Wang, H. (2022). Multi-kingdom microbiota analyses identify bacterial-fungal interactions and biomarkers of colorectal cancer across cohorts. *NATURE MICROBIOLOGY*, *7*(2), 238+. https://doi.org/10.1038/s41564-021-01030-7

Loftus, M., Hassouneh, S. A.-D., & Yooseph, S. (2021). Bacterial community structure alterations within the colorectal cancer gut microbiome. *BMC Microbiol.*, *21*(1).

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

Mrmr: mRMR (minimum-Redundancy-Maximum-Relevance) for automatic feature selection at scale. (n.d.). https://github.com/smazzanti/mrmr

Naka, Y., Okada, T., Nakagawa, T., Kobayashi, E., Kawasaki, Y., Tanaka, Y., Tawa, H., Hirata, Y., Kawakami, K., Kakimoto, K., Inoue, T., Takeuchi, T., Fukunishi, S., Hirose, Y., Uchiyama, K., Asahi, M., & Higuchi, K. (2020). Enhancement of o-linked n-acetylglucosamine modification promotes metastasis in patients with colorectal cancer and concurrent type 2 diabetes mellitus. *ONCOLOGY LETTERS*, *20*(2), 1171–1178. https://doi.org/10.3892/ol.2020.11665

Nie, S., Wang, A., & Yuan, Y. (2021). Comparison of clinicopathological parameters, prognosis, microecological environment and metabolic function of gastric cancer with or without fusobacterium sp. infection. *JOURNAL OF CANCER*, *12*(4), 1023–1032. https://doi.org/10.7150/jca.50918

pandas development team, T. (2020). *Pandas-dev/pandas: Pandas* (Version latest). Zenodo. https://doi.org/10.5281/zenodo.3509134

Pasolli, E., Schiffer, L., Manghi, P., Renson, A., Obenchain, V., Truong, D. T., Beghini, F., Malik, F., Ramos, M., Dowd, J. B., Huttenhower, C., Morgan, M., Segata, N., & Waldron, L. (2017). Accessible, cu-

rated metagenomic data through ExperimentHub. *Nat. 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.

Pip: The python package installer. (n.d.). https://pip.pypa.io/en/stable/

Sloof, J. (2023). Gitlab [Accessed: 2023-6-24]. https://gitlab.noshit.be/AbeelLab/rp2023/-/tree/main/jsloof

Ugbogu, E. A., Schweizer, L. M., & Schweizer, M. (2022). Contribution of model organisms to investigating the far-reaching consequences of prpp metabolism on human health and well-being. *CELLS*, *11*(12). https://doi.org/10.3390/cells11121909

Vega, P., Valentın, F., & Cubiella, J. (2015). Colorectal cancer diagnosis: Pitfalls and opportunities. *World J. Gastrointest. Oncol.*, *7*(12), 422–433.

Vidnes, T. K., Wahl, A. K., & Andersen, M. H. (2013). Patient experiences following liver transplantation due to liver metastases from colorectal cancer. *EUROPEAN JOURNAL OF ONCOLOGY NURSING*, *17*(3), 269–274. https://doi.org/10.1016/j.ejon.2012.07.004

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., ... SciPy 1.0 Contributors. (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2

Waskom, M. L. (2021). Seaborn: Statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/joss.03021

Wirbel, J., Pyl, P. T., Kartal, E., Zych, K., Kashani, A., Milanese, A., Fleck, J. S., Voigt, A. Y., Palleja, A., Ponnudurai, R., Sunagawa, S., Coelho, L. P., Schrotz-King, P., Vogtmann, E., Habermann, N., Niméus, E., Thomas, A. M., Manghi, P., Gandini, S., ... Zeller, G. (2019). Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nat. Med.*, *25*(4), 679–689.

Xu, F., Fu, Y., Sun, T.-y., Jiang, Z., Miao, Z., Shuai, M., Gou, W., Ling, C.-w., Yang, J., Wang, J., Chen, Y.-m., & Zheng, J.-S. (2020). The interplay between host genetics and the gut microbiome reveals common and distinct microbiome features for complex human diseases. *MICROBIOME*, *8*(1). https://doi.org/10.1186/s40168-020-00923-9

Yachida, S., Mizutani, S., Shiroma, H., Shiba, S., Nakajima, T., Sakamoto, T., Watanabe, H., Masuda, K., Nishimoto, Y., Kubo, M., Hosoda, F., Rokutan, H., Matsumoto, M., Takamaru, H., Yamada, M., Matsuda, T., Iwasaki, M., Yamaji, T., Yachida, T., ... Yamada, T. (2019). Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. *Nat. Med.*, *25*(6), 968–976.

Zhang, S., Kong, C., Yang, Y., Cai, S., Li, X., Cai, G., & Ma, Y. (2020). Human oral microbiome dysbiosis as a novel non-invasive biomarker in detection of colorectal cancer. *Theranostics*, *10*(25), 11595–11606.

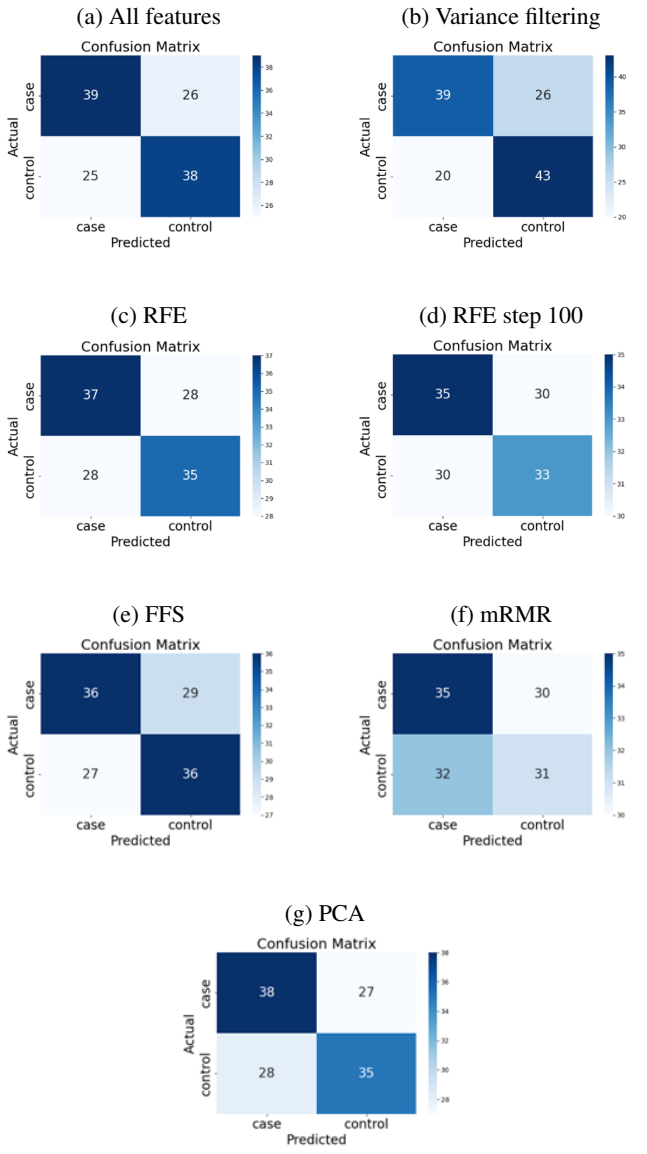



Figure 4: The test confusion matrices of logistic regression combined with different feature selection techniques: no feature selection, variance filtering, Recursive Feature Elimination (RFE), RFE with steps of 100, Forward Feature Selection (FFS), minimum Redundancy Maximum Relevance (mRMR) and Principal Component Analysis (PCA).

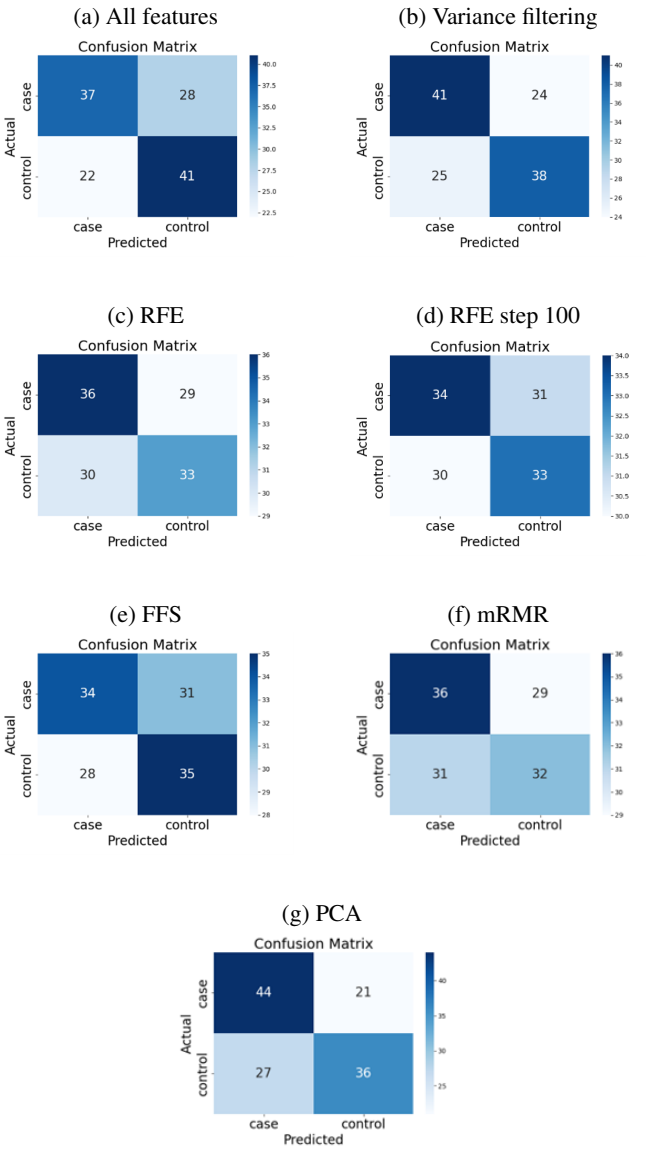



Figure 5: The test confusion matrices of the neural network combined with different feature selection techniques: no feature selection, variance filtering, Recursive Feature Elimination (RFE), RFE with steps of 100, Forward Feature Selection (FFS), minimum Redundancy Maximum Relevance (mRMR) and Principal Component Analysis (PCA).

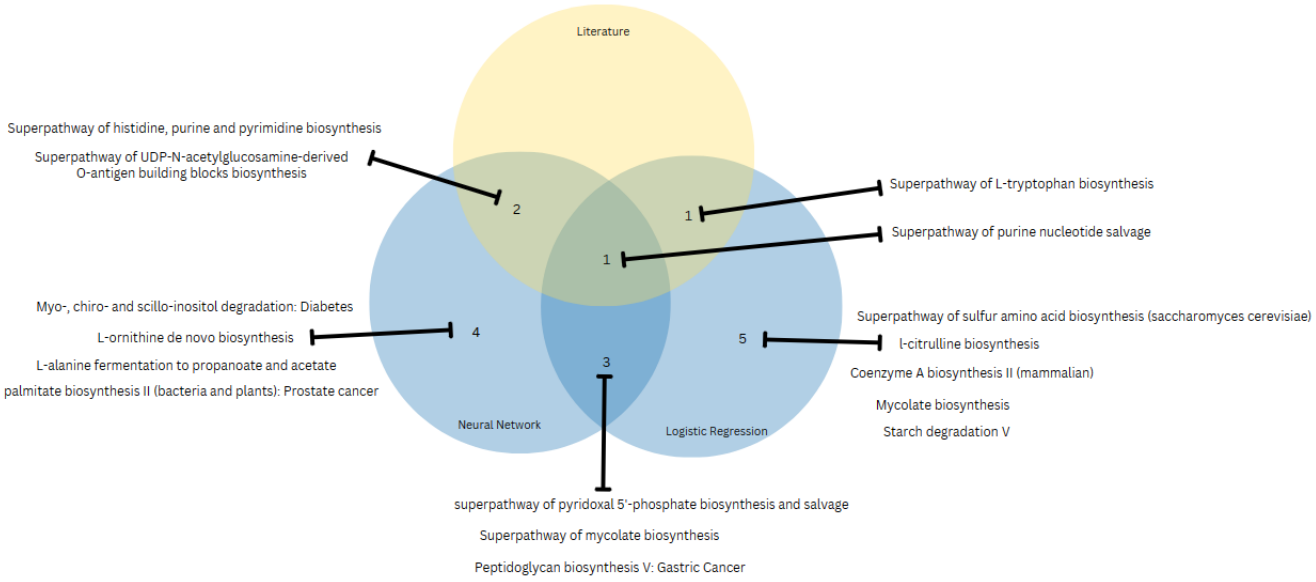



Figure 6: Venn Diagram of most important features selected by variance filtering. The importance is determined by neural network permutation importance.