Identification of biomarkers for obesity with metabolic syndrome using machine learning models

Identification of biomarkers for obesity with metabolic syndrome using machine learning models

Pin-Yen Chen
BBiomedSc (Hon I)

School of Medical Science
Griffith University

Submitted in fulfilment of the requirements of the degree of
Doctor of Philosophy

August 2020

Summary

Metabolic syndrome (MetS) is a condition that is linked to the increased risk of developing chronic diseases, including type 2 diabetes mellitus (T2DM) and cardiovascular disease (CVD). The association between MetS and chronic disease development lies in the cardiometabolic risk factors that comprise MetS: abdominal obesity, hypertension, hyperglycaemia and dyslipidaemia [1]. The development of MetS is also associated with the dysregulation of many different body systems, such as the immune system [2] and gut microbiome [3]. Due to its multifactorial nature, research in MetS requires the simultaneous analysis of multiple biomarkers across different body systems. As most research thus far have utilised univariate analysis, no biomarker profile has been identified to characterise individuals more at risk of MetS and related diseases. The current study has therefore implemented the use of correlation-based network analysis (CNA) and multiple classification models to identify the biomarkers that collectively link to increased MetS development.

Four variable groups comprising of multiple different measurements were obtained from 117 healthy weight controls and 35 obese with MetS individuals. The four variable groups consisted of: anthropometric measures, haematological measures, gene expression levels and gut microbial counts. The use of CNA allowed a better understanding of the relationships between biomarkers affected by MetS. As expected, the obese with MetS network was denser than the healthy weight control network, demonstrating the complex nature of MetS. The results found molecular interactions supporting the findings of previous literature, particularly correlations that demonstrated the development of anaemia of inflammation in the obese with MetS network. There were also three key hubs identified using gene expression levels, involving transcription factor EB (TFEB), lipocalin 2 (LCN2), and cluster of differentiation- (CD-) 68. The three genes are associated with regulatory T cells and neutrophils, two prominent cells in regulating the

inflammatory state. As obesity and MetS are often described as a state of chronic low-grade inflammation, the findings of CNA correspond with that of previous studies.

Classification models are another type of analytical tool that have demonstrated high predictive ability in many diseases, including T2DM and CVD. The use of classification models for the prediction of diseases allows the risk of disease development to be evaluated. The current study applied classification models for the prediction of MetS using three of the four variable groups measured: haematological measures, gene expression levels and gut microbial counts. Classification models are not only able to assess the relevance of these variable groups to MetS but also identify the specific variables that contributed the most to MetS development. There are a range of classification models that can be used and due to MetS being a relatively new area of research, the most appropriate model for MetS prediction has yet to be determined. As such, the current study predicted MetS using four different types of classification models and compared the predictive abilities of each model. The four models that were used in the current study were: logistic regression (LR), decision tree (DT), support vector machine (SVM) and artificial neural network (ANN). The performance of each classification model was evaluated using 10-fold cross-validation, which splits the dataset into 10 training and testing sets. Each model is then built using the training sets and evaluated using the testing sets to ensure that the model was not fit too closely to the training data.

The model with the highest performance when predicting MetS using haematological measures and gut microbial counts was ANN, while SVM had the highest performance when using gene expression levels. However, ANN was also able to attain a high area under the curve (AUC) value of 0.804 when predicting MetS using gene expression levels. As such, the prediction model that had the highest performance overall was ANN. Each model has their own strengths and limitations dealing with specific types of data and the most appropriate model depends on the research question being asked. Although SVM and ANN are both very powerful algorithms,

capable of handling high-dimensional data, both models have difficulty producing clinically significant results. On the other hand, LR and DT models are both able to identify specific biomarkers that should be further investigated for links to diseases development, deeming them more suitable for clinical applications.

For each of the 10 LR and DT models, constructed using the 10 training sets, the haematological measurement that was found to be most important was triglycerides (TG). Additionally, the best performing LR model, out of the 10 constructed models, found measurements of TG, platelets (PLT), erythrocyte sedimentation rate (ESR), fasting plasma glucose (FPG), haemoglobin (HG) and glycated haemoglobin A1c (HbA1c) to be associated with MetS development. At the same time, high-density lipoprotein-cholesterol (HDL-C) was linked to a reduced risk of MetS development. Using DT, the important measurements in MetS development were TG, PLT, HDL-C, age, HG, C-reactive protein (CRP) and white cell counts. Each variable identified has been found to be linked to either a cardiometabolic risk factor or inflammation and thus the results of the current study are supportive of previous literature on obesity and MetS.

Logistic regression also found the expression of AKT serine/threonine kinase 3 (AKT3), Fc fragment of IgE receptor II (FCER2), cathelicidin antimicrobial peptide (CAMP), interleukin-11 receptor subunit alpha (IL11RA) and granzyme H (GZMH) to increase the odds of developing MetS while C-X-C motif chemokine receptor 6 (CXCR6), C-C motif chemokine ligand- (CCL-)3, suppressor of cytokine signalling 1 (SOCS1) and killer cell lectin like receptor C2 (KLRC2) expression reduces these odds. Consistent with these findings, DTs also predicted individuals with high AKT3, FCER2 and CAMP expression to be obese with MetS while healthy weight controls had higher CXCR6, CCL3 and KLRC2 expression. The findings of the current study were partially supportive of previous literature, with FCER2 and CAMP expression being associated with obesity and inflammation [4, 5] and KLRC2 expression being

inversely associated with obesity and inflammation [6, 7]. On the other hand, AKT3 is associated with glucose and lipid metabolism [8] with evidence of its expression leading to the protection against insulin resistance. As such, the high AKT3 expression found in the obese with MetS cohort was not consistent with current literature. Similarly, the association between the expression of CXCR6 and CCL3 with a healthy weight control classification could not be explained as both genes are typically linked to inflammation [9, 10].

Finally, LR and DT found microbial species belonging to the Firmicutes and Bacteroidetes phyla to both be associated with the increased and reduced risk of developing obesity with MetS. Obesity with MetS is largely characterised by a high Firmicutes-to-Bacteroidetes ratio, particularly when compared to healthy weight controls [11]. While this pattern was not clearly evident in the current study, the cause of discrepancy with previous literature may be due to gut microbial studies in obesity and MetS not being typically reported at the species level.

While LR and DT are both able to identify the variables that are likely to contribute to MetS development in a clinical setting, the performances of either model were not able to compete with that of ANN or SVM. At the same time, despite having the highest performance overall, ANN is unable to produce easily interpretable results with clinical significance. As such, its high predictive ability is not enough to convince researchers to choose ANN for clinical use. To overcome this issue, many researchers combine ANN with a feature selection technique, such as genetic algorithm (GA). Feature selection techniques are able to identify the best combination of biomarkers for the prediction of diseases. In the current study, the variables that were recognised to be significant by the hybrid model supported the findings of LR and DT. The haematological biomarkers that were consistently recognised as important by all three prediction models were measures of TG and HG. Additionally, CCL3 and CXCR6 expression, as well as three gut microbial species belonging to the Firmicutes and Bacteroidetes phyla, were also found to be important for MetS development. Other than the identification of

important variables, the hybrid model was also able to improve the performance of ANN when predicting MetS using gene expression levels and gut microbial counts. Consequently, the current study concluded that the hybrid GA with ANN model was considered to be the most appropriate for MetS prediction.

Another analytical method that was used by the current study was weighted majority voting, which combines the final predicted outcomes of the other classification models to determine whether the performance could be further improved. The weighted majority voting method was able to achieve the highest AUC value for the prediction of MetS using gut microbial counts as well as the second highest AUC when using haematological measures and gene expression levels. However, the dependency of the weighted majority voting method on the performance of individual classification models used was demonstrated in the study. The low sensitivity values attained by DT in the testing set of all three variable groups is likely what prevented the weighted majority voting method from outperforming all the other classification models in the prediction of MetS. In spite of the limitation caused by DT, however, the method was still able to achieve a high performance. As such, the combination of the results from different classification models into a weighted majority voting method to increase the overall predictive ability was found to be a viable choice.

The classification model that was found to be most suitable for the prediction of MetS was the hybrid GA with ANN model. Not only was the model able to achieve high predictive ability due to the ANN portion of the model, it was also able to reveal the optimal combination of variables that contributed the most to an accurate MetS prediction. The variables that were identified were also supportive of the findings of both LR and DT. The measurements used by the current study (haematological measures, gene expression levels and gut microbial counts) were all found to be suitable for the prediction of MetS. Future studies may consider the use of

other biomarkers, including measurements from adipose tissue, for the prediction of MetS using the hybrid GA with ANN model.

References

[1] Expert Panel on Detection, E. and Adults, T.o.H.B.C.i., Executive summary of the third report of the National Cholesterol Education Program (NCEP). JAMA. vol. 285, pp. 2486-2497, 2001.

[2] Ellulu, M.S., et al., Obesity and inflammation: the linking mechanism and the complications. Arch Med Sci. vol. 13, pp. 851-63, 2015.

[3] Warmbrunn, M.V., et al., Gut microbiota: a promising target against cardiometabolic diseases. Expert Rev Endocrinol Metab. vol. 15, pp. 13-27, 2020.

[4] Rastogi, D., Suzuki, M., and Greally, J.M., Differential epigenome-wide DNA methylation patterns in childhood obesity-associated asthma. Sci Rep. vol. 3, 2013.

[5] Li, Y.-X., Li, B.-Z., and Yan, D.-Z., Upregulated expression of human cathelicidin LL-37 in hypercholesterolemia and its relationship with serum lipid levels. Mol Cell Biochem. vol. 449, pp. 73-9, 2018.

[6] Jung, U.J., et al., Differences in metabolic biomarkers in the blood and gene expression profiles of peripheral blood mononuclear cells among normal weight, mildly obese and moderately obese subjects. Br J Nutr. vol. 116, pp. 1022-32, 2016.

[7] Wieser, V., et al., Adipose type I interferon signalling protects against metabolic dysfunction. Gut. vol. 67, pp. 157-65, 2016.

[8] Huang, X., Liu, G., and Su, Z., The PI3K/AKT pathway in obesity and type 2 diabetes. Int J Biol Sci. vol. 14, pp. 1483-96, 2018.

[9] Ma, K.L., et al., Activation of the CXCL16/CXCR6 pathway promotes lipid deposition in fatty livers of apolipoprotein E knockout mice and HepG2 cells. Am J Transl Res. vol. 10, pp. 1802-16, 2018.

[10] Tourniaire, F., et al., Chemokine Expression In Inflamed Adipose Tissue Is Mainly Mediated By NF-κB. PLoS One. vol. 8, 2013.

[11] Koliada, A., et al., Association between body mass index and Firmicutes/Bacteroidetes ratio in an adult Ukrainian population. BMC Microbiol. vol. 17, pp. 1-6, 2017.

Statement of originality

This work has not previously been submitted for a degree or diploma in any university. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made in the thesis itself.

Signed:

Pin-Yen Chen
PhD Candidate
School of Medical Science
Griffith University

Date:

13/08/2020

Acknowledgements

First and foremost, I would like to thank my primary supervisor, Professor Allan Cripps, for all his support during my transition from a biomedical science degree to a data science field. I am particularly grateful for his patience and willingness to learn the terms that are specific to data science as he reviewed countless drafts of my thesis. The support he has provided, including access to essential coursework and workshops, has contributed significantly to the successful completion of my research.

I would also like to give special thanks to Dr Ping Zhang who was always more than willing to answer the many questions I had while navigating my way through the data science field. The support she provided throughout my journey, both professionally and on a personal level, was invaluable. I hope for the opportunity to work alongside Dr Ping Zhang in the future as her guidance has allowed me to grow exponentially as a data scientist.

Furthermore, the excellent assistance and feedback I have received from Dr Nicholas West and other members of the Mucosal Immunology Research Group has helped me to polish my thesis and bridge the gap between biomedical science and data science. My thanks will also be extended to the School of Medical Science, Griffith University, for providing great research facilities and opportunities to further my career in research.

My friends, Mr Kieran Dennis and Mr Brandon Trinh, thank you both for always being willing to read over my work. The support I have received from you both has been instrumental to the successful completion of my thesis.

Last but not least, my endless thanks to my family for their encouragement to help me persevere through this stressful time.

Abbreviations

ABCF1	ATP-binding cassette subfamily F member 1
AKT1	AKT serine/threonine 1
AKT3	AKT serine/threonine kinase 3
ALB	Albumin
ALP	Alkaline phosphatase
ALT	Alanine aminotransferase
ANN	Artificial neural network
APOE	Apolipoprotein E
AST	Aspartate aminotransferase
ATG7	Autophagy related 7
ATM	Adipose tissue macrophage
ATP III	Adult Treatment Panel III
AUC	Area under the curve
AUSDIAB	Australian diabetes, obesity and lifestyle study
BAI	Body adiposity index
BC	Betweenness centrality
BMI	Body mass index
BP	Blood pressure
BUN	Blood urea nitrogen
CAMP	Cathelicidin antimicrobial peptide
CART	Classification and regression tree
CCL-	C-C motif chemokine ligand-
CD-	Cluster of differentiation-
CDH1	Cadherin 1
CEACAM3	CEA cell adhesion molecule 3
CNA	Correlation-based network analysis
CP	Complexity parameter
CREA	Creatinine
CRP	C-reactive protein
CSF3R	Colony stimulating factor 3 receptor
CVD	Cardiovascular disease
CXCL-	C-X-C motif chemokine ligand-
CXCR6	C-X-C motif chemokine ligand 6
DBP	Diastolic blood pressure
DHA	Docosahexaenoic acid
DPF	Diabetes pedigree function
DT	Decision tree
EPA	Eicosapentaenoic acid
ESR	Erythrocyte sedimentation rate
FCAR	Fc fragment of IgA receptor
FCER2	Fc fragment of IgE receptor II
FFA	Free fatty acids

FFSTM	Fat-free soft-tissue mass
FMH	Fat mass-to-height ratio
FMN	Fructosamine
FN	False negative
FP	False positive
FPG	Fasting plasma glucose
FPR1	Formyl peptide receptor 1
G6PD	Glucose-6-phosphate dehydrogenase
GA	Genetic algorithm
GZMH	Granzyme H
GZMM	Granzyme M
Hb1Ac	Glycated haemoglobin A1c
HC	Hip circumference
HCT	Haematocrit
HDL-C	High-density lipoprotein cholesterol
HG	Haemoglobin
HLN	Hidden layer neurons
HMGB1	High Mobility Group Box 1
HMW-adiponectin	High molecular weight-adiponectin
HOMA-IR	Homeostatic model assessment of insulin resistance
IBIL	Indirect bilirubin
IFIT1	Interferon-induced protein with tetratricopeptide repeats 1
IFN-γ	Interferon-gamma
IKK	Iκβ kinase
IL-	Interleukin-
IL11RA	Interleukin-11 receptor subunit alpha
IL1RN	Interleukin-1 receptor antagonist
INSR	Insulin receptor
IQR	Interquartile range
IRF7	Interferon regulatory factor 7
IRS1	Insulin receptor substrate 1
IRS-1	Insulin receptor substrate 1
ITGAE	Integrin subunit alpha E
JNK	c-Jun NH_2-terminal kinase
KLRC2	Killer cell lectin like receptor C2
LCN2	Lipocalin 2
LDL-C	Low-density lipoprotein cholesterol
LPS	Liposaccharide
LR	Logistic regression
LTF	Lactotransferrin
MAPK1	Mitogen-activated protein kinase 1
MCH	Mean corpuscular haemoglobin
MCV	Mean corpuscular volume

MD-2	Myeloid differentiation factor 2
MDA	Malondialdehyde
MetS	Metabolic syndrome
MHCII	Major histocompatibility complex class II
MI	Mutual information
mRNA	Messenger ribonucleic acid
MTOR	Mechanistic target of rapamycin kinase
NF-κβ	Nuclear factor-kappa beta
NRF1	Nuclear respiratory factor 1
NSAIDs	Non-steroidal anti-inflammatory drugs
PCA	Principal component analysis
PIK3CA	Phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha
PIK3R1	Phosphoinositide-3-kinase regulatory subunit 1
PLT	Platelets
PPARγ	Peroxisome proliferator-activated receptor gamma
PYCARD	PYD and CARD domain containing
RBF	Radial basis function
RCC	Red blood cell count
rel_error	Relative error
RMR	Resting metabolic syndrome
ROC	Receiver operating characteristic
RPLP0	Ribosomal protein lateral stalk subunit P0
rprop+	Resilient backpropagation with weight backtracking
RPS6KA1	Ribosomal protein S6 kinase A1
rRNA	Ribosomal ribonucleic acid
S100A12	S100 calcium-binding protein A12
SAT	Subcutaneous adipose tissue
SBP	Systolic blood pressure
SERPING1	Serpin family G member 1
SLC2A4	Solute carrier family 2 member 4
SOCS1	Suppressor of cytokine signalling 1
SVM	Support vector machine
T2DM	Type 2 diabetes mellitus
T-bet	T-box expressed in T cells
TBIL	Total bilirubin
TFEB	Transcription factor EB
TG	Triglycerides
Th-	T-helper-
TLR4	Toll-like receptor 4
TN	True negative
TNFSF13	TNF superfamily member 13
TNF-α	Tumour necrosis factor-alpha
TP	True positive

Treg	T regulatory
TSC1	TSC complex subunit 1
UA	Uric acid
ULK1	UNC-51 like autophagy activating kinase 1
VAT	Visceral adipose tissue
WC	Waist circumference
WCC	White cell count
WHR	Waist-to-hip ratio
xerror	Cross-validation error
xstd	Standard error
α-HBDH	α-hydroxybutyrate dehydrogenase

Symbols and unit of measurements

-	Negative
$	Australian dollars
%	Percentage
|x|	Absolute value
+	Positive
±	Plus or minus
cm	Centimetres
mg/dL	Milligrams per decilitre
mmHg	Millimetre of mercury
mmol/L	Millimoles per litre
ρ	Correlation coefficient
ρ0	Correlation coefficient threshold

Table of Contents

Summary 2
References 7
Statement of originality 8
Acknowledgements 9
Abbreviations 10
Symbols and unit of measurements 13
List of Tables 17
List of Figures 19
List of Appendices 21
CHAPTER 1 Introduction 22
1.1 Research significance 23
1.2 Hypothesis 24
1.3 Project aims 25
1.4 References 25
CHAPTER 2 Literature Review 27
2.1 Obesity and metabolic syndrome 27
2.2 Obesity and inflammation 28
2.2.1 Obesity-related dysregulation of the immune system 29
2.2.2 Innate immunity in obesity 29
2.2.3 Adaptive immunity in obesity 31
2.2.4 Free fatty acids in obesity 35
2.3 Obesity and the gut microbiota 35
2.4 Obesity and metabolic endotoxaemia 37
2.5 Correlation-based network analysis 39
2.6 Multivariate prediction models 41
2.6.1 Logistic regression 42
2.6.2 Decision tree 48
2.6.3 Support vector machine 50
2.6.4 Artificial neural networks 54
2.6.5 Comparing classification algorithms 57
2.7 Feature selection 58
2.8 Future directions 63

2.9 References 65
CHAPTER 3 Discovery of biomarkers in MetS using correlation-based network analysis ... 72
3.1 Abstract 72
3.2 Introduction 73
3.3 Research design 74
3.3.1 Study design and ethics 74
3.3.2 Sample collection 75
3.3.3 Univariate analysis 76
3.3.4 Correlation-based network analysis 76
3.4 Results 79
3.4.1 Descriptive analysis 79
3.4.2 Correlation-based networks 84
3.5 Discussion 89
3.6 Appendices 96
3.7 References 108
CHAPTER 4 Application of classification models for the prediction of MetS 113
4.1 Abstract 113
4.2 Introduction 114
4.3 Research design 115
4.3.1 Study design 115
4.3.2 Cross-validation 116
4.3.3 Logistic regression 119
4.3.4 Decision tree 120
4.3.5 Support vector machine 123
4.3.6 Artificial neural network 124
4.4 Results 128
4.5 Discussion 147
4.6 Appendices 161
4.7 References 182
CHAPTER 5 Combination of artificial neural network with genetic algorithm for the prediction of MetS 184
5.1 Abstract 184
5.2 Introduction 185
5.3 Research design 186

5.3.1 Study design 186
5.3.2 Genetic algorithm 186
5.3.3 Genetic algorithm with artificial neural network 187
5.4 Results 188
5.5 Discussion 195
5.6 Appendices 198
5.7 References 205
CHAPTER 6 Improving MetS prediction through the use of weighted majority voting 206
6.1 Abstract 206
6.2 Introduction 206
6.3 Research design 208
6.3.1 Study design 208
6.3.2 Weighted majority voting 208
6.4 Results 209
6.5 Discussion 212
6.6 References 214
CHAPTER 7 Discussion 215
7.1 References 220

List of Tables

Table 2.1. List of studies which used classification methods for the prediction of MetS and related diseases.....45

Table 2.2. The four different types of kernel functions used in SVM, their formulas and the parameters that need to be specified by the user to optimise the model.....51

Table 3.1. Anthropometric and haematological measures in healthy weight (n = 117) and obese with MetS (n = 35) individuals.....80

Table 3.2. Metabolic effects of genes chosen for comparison between healthy weight and obese with MetS groups.....82

Table 4.1. Example of a complexity parameter table. 121

Table 4.2. Number of healthy weight and obese with MetS participants in each training and testing sets of the three variable groups.....128

Table 4.3. Comparison of the performance by full logistic regression models and models optimised by the forward stepwise technique.....129

Table 4.4. Comparison of the performance of manually pruned trees with that of trees pruned using the grid search approach.....130

Table 4.5. Comparison of the performance by full-sized trees and pruned trees. 130

Table 4.6. Comparison of the 4 kernel functions used to build SVM models using the 3 different variable groups.....134

Table 4.7. Comparison of the averaged best performing neural networks using different hidden layer neuron sizes.....135

Table 4.8. The averaged performance of the 10 best predicted training and testing sets for each prediction model.....137

Table 4.9. Important haematological variables used to build the best performing logistic regression models.....138

Table 4.10. Important haematological variables identified by the best performing decision trees.....139

Table 4.11. Important genes used to build the best performing logistic regression models. The table has been split for editorial purposes.....141

Table 4.12. Important genes identified by the best performing decision trees. The table has been split for editorial purposes.....142

Table 4.13. Important gut microbial species used to build the best performing logistic regression models. The table has been split for editorial purposes.....145

Table 4.14. Important gut microbial species identified by the best performing decision trees. 146

Table 5.1. Comparison of the averaged best performing hybrid genetic algorithm with neural networks using different hidden layer neuron sizes. 189

Table 5.2. The averaged performance of the 10 training and testing sets for each individual classification model and hybrid genetic algorithm with artificial neural network. 193

Table 6.1. The averaged performance of the 10 training and testing sets for each base learner and majority voting. 210

Table 6.2. The performance of the weighted majority voting method after the inclusion of the hybrid model. 211

List of Figures

Figure 2.1. Obesity-associated inflammation through adipocyte enlargement and recruitment of pro-inflammatory biomarkers [9]. 29

Figure 2.2. Example of four molecular pathways resulting from high-fat diet, each contributing to the state of chronic low-grade inflammation associated with obesity, leading to increased risk of chronic diseases. (A) Upregulation of pro-inflammatory biomarkers, (B) downregulation of anti-inflammatory biomarkers, (C) changes in gut microbial composition leading to upregulation of pro-inflammatory biomarkers, and (D) changes to gut permeability exacerbating pro-inflammation. 33

Figure 2.3. Different properties of correlation-based network analysis, including large vertex degree shown by node 7 and high betweenness centrality shown by node 2 and node 6. 40

Figure 2.4. Example of decision tree layout. 48

Figure 2.5. Illustration of the different factors that constitute a support vector machine [77]. 50

Figure 2.6. Demonstration of how kernel functions separate two groups of data at a higher dimensional space [78]. 51

Figure 2.7. Example of the artificial neural network structure, with the input layer, one hidden layer with 14 hidden layer neurons, and the output layer [73]. 55

Figure 2.8. Terminology used in genetic algorithm (GA), with genes representing each input feature, chromosomes representing a subset of features and the population showing the collection of different subsets chosen by GA. 61

Figure 2.9. Demonstration of the (A) crossover and (B) mutation steps of genetic algorithm. 62

Figure 3.1. Example of a multi-analyte network constructed in the current study [6]. 78

Figure 3.2. Multi-level correlation-based network built using measurements from healthy weight individuals. 85

Figure 3.3. Multi-level correlation-based network built using measurements from obese with MetS individuals. 86

Figure 4.1. Example of k-fold cross-validation. 117

Figure 4.2. Example layout of confusion matrix used to calculate the different values that describe the performance of prediction models. 118

Figure 4.3. Example of a ROC curve constructed from plotting sensitivity and specificity values calculated at different thresholds. 119

Figure 4.4. Visual demonstration of how the best model for the 10 training sets were chosen for decision tree, support vector machine and artificial neural network. 122

Figure 4.5. Example of neural network layout using 1 input layer with 4 input layer neurons, 1 hidden layer with 3 hidden layer neurons and 1 output layer of size 1. The activation function applied was logistic function. 126

Figure 4.6. Calculation of the loss function in rprop+ to move towards the global (A) or local (B) optima. Loss function (C) has overstepped the global optimum, resulting in a negative derivative while local function (D) has a positive derivative and the step size must now be adjusted for it to move towards the local optimum.....127

Figure 4.7. Best performing decision tree constructed using haematological measures. 140

Figure 4.8. Best performing decision tree constructed using gene expression level data. 144

Figure 4.9. Best performing decision tree constructed using gut microbial composition data. 147

List of Appendices

Appendix 3.1. List of the gut microbial species, and the phylum to which they belong, that were used for correlation-based network analysis. 96

Appendix 3.2. Previous publication with preliminary data. 97

Appendix 4.1. Correlations between biomarkers from the combined haematological measures variable group. 161

Appendix 4.2. Correlations between biomarkers from the combined gene expression variable group. 162

Appendix 4.3. List of remaining variables from each group after removing highly correlated variables. 165

Appendix 4.4. List of the gut microbial species, and the phylum to which they belong, that were used for the construction of classification models. 167

Appendix 4.5. Complete list of the important gut microbial species identified by the best performing logistic regression models. Table has been split for editorial purposes. 169

Appendix 4.6. Complete list of the important gut microbial species identified by the best performing decision trees. 171

Appendix 4.7 Co-authored published paper. 173

Appendix 5.1. Complete list of the important haematological measures and the generated fitness values identified by genetic algorithm. The table was split for editorial purposes. ... 198

Appendix 5.2. Complete list of the important genes and the generated fitness values identified by genetic algorithm. The table was split for editorial purposes. 199

Appendix 5.3. Complete list of the important gut microbial species and the generated fitness values identified by genetic algorithm. The table was split for editorial purposes 201

CHAPTER 1
Introduction

Metabolic syndrome (MetS) is a collection of cardiometabolic risk factors that increases the risk of an individual developing several chronic diseases, particularly type 2 diabetes mellitus (T2DM) and cardiovascular disease (CVD). The risk factors associated with MetS development include abdominal obesity, hypertension, hyperglycaemia and dyslipidaemia [1]. The development of MetS is accompanied by a multitude of dysregulations in different areas of the body, such as the immune system [2] and gut microbiome [3], which may then contribute to an increased risk of chronic diseases. Studies in both humans and mice have found obese individuals and animals to have higher pro-inflammatory immune markers [2, 12], reduced gut bacterial diversity, altered gut bacterial composition, and increased gut permeability [13] compared to their lean, metabolically-healthy counterparts. While these findings suggest there to be biomarkers that distinguish individuals more at risk of MetS-related diseases, very few biomarkers have been identified for preventive or therapeutic targeting. A key reason is due to the use of univariate analysis rather than simultaneous multivariate analysis. As MetS has a multifactorial nature, the dysregulation of multiple biomarkers across different body systems is what collectively leads to an increased development of MetS. Consequently, there exists an issue of redundancy within biological systems as many biomarkers from different body systems have similar roles. The inactivation of one biomarker may therefore have little to no effect on the overall system and thus simultaneous analysis of different body systems is necessary. While recent technological advances have made simultaneous analysis of multiple biological markers possible, the analysis and interpretation of the data still poses a challenge due to the large amount of data generated by these methods. The current project aims to utilise computational methods, including network-based correlation and classification models, to better understand

the role of biomarkers from different body systems in the development of MetS. Construction of a network analysis allows a global assessment of linked metabolic pathways, providing better insight into MetS-related diseases. Additionally, using classification models will reveal the features that are important in MetS prediction. Identifying important biomarkers in MetS prediction and understanding the relationship between them will help to advance research looking to reduce the incidence of MetS and related diseases.

1.1 Research significance

Abdominal obesity is the risk factor that is of particular concern when it comes to the risk of developing MetS. The Australian diabetes, obesity and lifestyle study (AUSDIAB), conducted in 2012, reported that obese individuals were six times more likely to develop MetS than normal weight individuals [14]. The report also stated that the annual incidence of MetS in overweight and obese individuals is 2.8% and 3.9%, respectively. However, as the number of Australians classified as either overweight or obese has been increasing annually, the associated MetS incidence is also expected to be higher. Between 1995 and 2015, the percentage of overweight or obese Australian adults had risen from 56.3% to 63.4% [15]. More recently, in 2017-8, the figure had increased to 67% [16]. The Australian Burden of Disease Study 2015 labelled high body mass index (BMI) to be the second leading risk factor for disease burden and deaths, second to tobacco use [17]. Furthermore, overweight and obesity were also directly associated with the third, fourth and fifth leading risk factors, which were dietary risks, high blood pressure and high blood plasma glucose. As high blood pressure and blood plasma glucose are both risk factors of MetS, the impact that abdominal obesity has on the development of MetS is further reinforced. Both risk factors also provide the link between MetS and the development of chronic diseases. The indirect medical costs associated with

MetS through CVD and T2DM were reported by the Australian Burden of Disease Study 2015 to be $10.4 billion and $1.56 billion. Again, as the incidence of MetS increases annually, the medical burden is also expected to be greater. The burden of MetS, both directly and indirectly, suggest there is a critical need to better understand biomarkers involved in the development of MetS and related diseases, which may assist in intervention studies looking to reduce its incidence. Studies in MetS are often limited by the inability to undertake detailed, simultaneous analysis to better understand the interactions of biomarkers involved in MetS development. Recent technological advances allow researchers to bypass this limitation by permitting the concurrent analysis of hundreds to thousands of measures across multiple biological systems to better understand the multifactorial and integrated nature of the disease. The research proposed here aims to use integrated analysis of different body systems, including the immune, metabolic and metagenomic systems to better understand MetS to assist with future intervention studies.

1.2 Hypothesis

It is hypothesised that the multifactorial effect of MetS on different systems of the body are connected and thus a profile of biomarkers from different body systems can be identified to characterise individuals at risk of MetS-related diseases.

1.3 Project aims

There are four aims in this study:

1. To investigate any underlying interactions between biomarkers across different body systems that may contribute to disease development through correlation-based network analysis;
2. To assess the ability to accurately predict MetS using the measurements available to build prediction models with different classification methods;
3. To improve the performance of prediction models through a hybrid model; and
4. To identify the best combination of risk factors that predict MetS using a feature selection algorithm.

1.4 References

[1] Expert Panel on Detection, E. and Adults, T.o.H.B.C.i., Executive summary of the third report of the National Cholesterol Education Program (NCEP). JAMA. vol. 285, pp. 2486-2497, 2001.

[2] Ellulu, M.S., et al., Obesity and inflammation: the linking mechanism and the complications. Arch Med Sci. vol. 13, pp. 851-63, 2015.

[3] Warmbrunn, M.V., et al., Gut microbiota: a promising target against cardiometabolic diseases. Expert Rev Endocrinol Metab. vol. 15, pp. 13-27, 2020.

[4] Rastogi, D., Suzuki, M., and Greally, J.M., Differential epigenome-wide DNA methylation patterns in childhood obesity-associated asthma. Sci Rep. vol. 3, 2013.

[5] Li, Y.-X., Li, B.-Z., and Yan, D.-Z., Upregulated expression of human cathelicidin LL-37 in hypercholesterolemia and its relationship with serum lipid levels. Mol Cell Biochem. vol. 449, pp. 73-9, 2018.

[6] Jung, U.J., et al., Differences in metabolic biomarkers in the blood and gene expression profiles of peripheral blood mononuclear cells among normal weight, mildly obese and moderately obese subjects. Br J Nutr. vol. 116, pp. 1022-32, 2016.

[7] Wieser, V., et al., Adipose type I interferon signalling protects against metabolic dysfunction. Gut. vol. 67, pp. 157-65, 2016.

[8] Huang, X., Liu, G., and Su, Z., The PI3K/AKT pathway in obesity and type 2 diabetes. Int J Biol Sci. vol. 14, pp. 1483-96, 2018.

[9] Ma, K.L., et al., Activation of the CXCL16/CXCR6 pathway promotes lipid deposition in fatty livers of apolipoprotein E knockout mice and HepG2 cells. Am J Transl Res. vol. 10, pp. 1802-16, 2018.

[10] Tourniaire, F., et al., Chemokine Expression In Inflamed Adipose Tissue Is Mainly Mediated By NF-κB. PLoS One. vol. 8, 2013.

[11] Koliada, A., et al., Association between body mass index and Firmicutes/Bacteroidetes ratio in an adult Ukrainian population. BMC Microbiol. vol. 17, pp. 1-6, 2017.

[12] Monserrat-Mesquida, M., et al., Metabolic syndrome is associated with oxidative stress and proinflammatory state. Antioxidants vol. 9, pp. 236, 2020.

[13] He, M. and Shi, B., Gut microbiota as a potential target of metabolic syndrome: the role of probiotics and prebiotics. Cell Biosci. vol. 7, pp. 1-14, 2017.

[14] Tanamas, S.K., et al., *AUSDIAB 2012*, in *The Australian diabetes, obesity and lifestyle study*. 2012, Baker IDI Heart and Diabetes Institute: Victoria, Australia. pp. 1-92.

[15] Australian Bureau of Statistics, *4364.0.55.001 - National Health Survey: First Results, 2014-15* 2015: Canberra, Australia.

[16] Australian Bureau of Statistics, *4364.0.55.001 - National Health Survey: First Results, 2017-18*. 2018: Canberra, Australia.

[17] Australian Institute of Health and Welfare, *Australian Burden of Disease Study: impact and causes of illness and death in Australia 2015*, in *Australian Burden of Disease Study*. 2019, AIHW: Canberra.

CHAPTER 2
Literature Review

2.1 Obesity and metabolic syndrome

Many studies have shown that metabolic syndrome (MetS) increases the risk of type 2 diabetes mellitus (T2DM) by five-fold and doubles the risk of cardiovascular disease (CVD) [1]. However, there exists many different definitions for metabolic syndrome (MetS), each with their own criteria. The two most commonly used criteria have been devised by the National Cholesterol Education Program (NCEP) Adult Treatment Panel III (ATP III) [2] and the International Diabetes Foundation (IDF) [3]. According to the ATP III criteria, an individual is identified as having MetS if they have at least three of the five following risk factors:

- abdominal obesity, characterised by a body mass index (BMI) of over 30kg/m^2, a waist circumference of over 94 cm in Caucasian men and over 80 cm in Caucasian women;
- high triglyceride levels of over 150 mg/dL;
- reduced high-density lipoprotein cholesterol (HDL-C) of less than 1.03 mmol/L in men and less than 1.29 mmol/L in women;
- high blood pressure of over 130 mmHg systolic blood pressure (SBP) or over 85 mmHg diastolic blood pressure (DBP); and
- high fasting plasma glucose (FPG) of over 5.6 mmol/L or glycated haemoglobin A1c (Hb1Ac) of over 6.5%.

The IDF criteria for MetS is very similar to that of ATP III, with the exception that central obesity must be present along with two of the four other factors. With the focus on central obesity, IDF accounts for the factors that may change the definition of obesity, such as different ethnicities. The IDF criteria also recognises the differences in the risk of type 2 diabetes mellitus (T2DM) and cardiovascular disease (CVD) development in different ethnic

populations. Nonetheless, there is still debate surrounding whether the emphasis of MetS should be placed on obesity or insulin resistance. As such, the standard criteria for the definition of MetS has not yet been specified.

2.2 Obesity and inflammation

Adipocytes are cells of adipose tissue which store energy in the form of triglycerides, a form of fat present in blood. During the initial stages of obesity, the storage of excess fat occurs through an increase in subcutaneous adipose tissue (SAT) cells, known as a hyperplastic response [4]. The increase in SAT adipocyte numbers occurs through adipogenesis, which is the recruitment and differentiation of new adipocytes. However, as obesity develops, the process of adipogenesis is inhibited by biomarkers such as interleukin- (IL-)17 which impairs the ability of precursor cells to enter adipogenesis. As a result, adipocytes increase in size instead to allow for excess fat storage, which is a hypertrophic response [4]. Once the capacity of SAT to store fat is reached, fat is stored in ectopic fat depots, such as visceral adipose tissue (VAT). An increase in VAT, found around internal organs, causes great concern due to its associated with greater mortality risk, as opposed to SAT which lies just under the skin. Comparisons of VAT and SAT biopsies from 18 lean, 16 obese and 8 obese with T2DM found VAT to contribute to insulin resistance while SAT did not [5]. Similar results were also reported in a study comparing VAT and SAT of 107 Chinese individuals [6]. The difference is likely due to the high level of pro-inflammatory markers, including IL-6 found in VAT compared to SAT or obese individuals [7], resulting in a widespread pro-inflammatory effect on the immune system. The proposed link between inflammation and chronic diseases [8] offers an explanation for the correlation between VAT mass and insulin resistance.

2.2.1 Obesity-related dysregulation of the immune system

As visceral adipocytes enlarge, chemotactic adipokines and chemokines C-C motif chemokine ligand- (CCL-)5, C-X-C motif chemokine ligand- (CXCL-)12 and CCL20 are released (Figure 2.1), recruiting both pro-inflammatory innate and adaptive immune cells into the adipose tissue. Cells are categorised as either pro-inflammatory or anti-inflammatory based on the cytokines they secrete and the receptors they express. Some of the most common pro-inflammatory cytokines are IL-6, tumour necrosis factor-alpha (TNF-α) and interferon-gamma (IFN-γ), and common anti-inflammatory cytokines include IL-4 and IL-10. In obese tissue, a higher pro-inflammatory-to-anti-inflammatory ratio is present which has been associated with increased risk of chronic diseases, including T2DM and CVD.

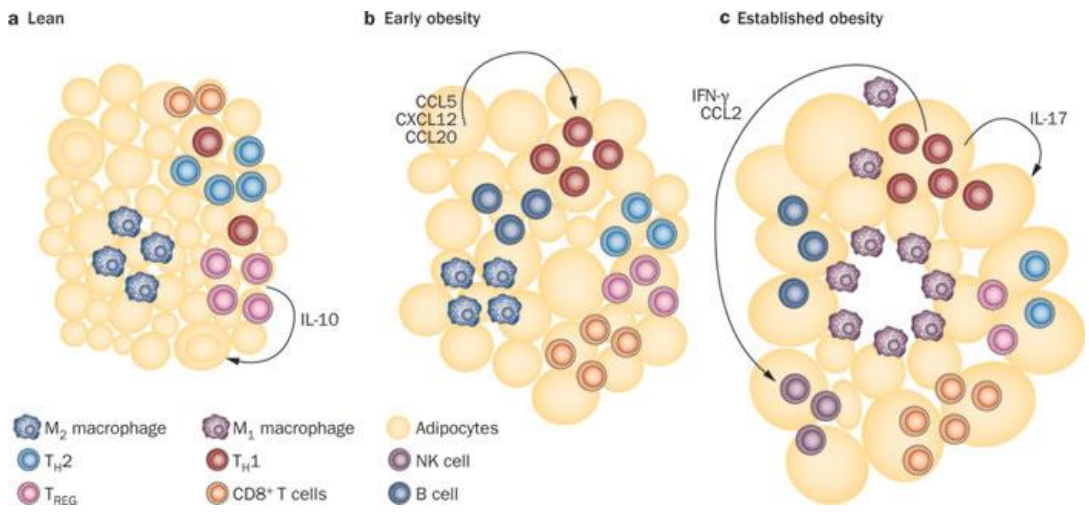


Figure 2.1. Obesity-associated inflammation through adipocyte enlargement and recruitment of pro-inflammatory biomarkers [9]. (Reprinted by permission from Springer Nature Customer Service Centre GmbH: Springer Nature, Nature Reviews Endocrinology, Adaptive immunity in obesity and insulin resistance, Sell et al., © 2012.)

2.2.2 Innate immunity in obesity

Obesity affects the innate immune system which is comprised of cells such as macrophages, dendritic cells and mast cells. Adipose tissue macrophages (ATMs) are innate immune cells which have been studied extensively in obesity research. ATMs can be classified as either M1 or M2 macrophages based on how they are activated and the cytokines they secrete. M1

macrophages are activated by pro-inflammatory mediators such as IFN-γ and secrete pro-inflammatory IL-6 and TNF-α [10]. On the other hand, M2 macrophages are activated by anti-inflammatory IL-4 and IL-13 and secrete anti-inflammatory IL-10. As obesity develops, the number of ATMs that constitute immune cells in adipose tissue increases to 50%, a dramatic increase from the 5% reported in lean tissue [11]. The increase in ATM abundance is accompanied by a shift in M1-to-M2 ratio from low to high. Studies in mice have suggested the change in ratio to be a result of a phenotypic shift from M2 to M1 macrophages. However, human studies did not find ATMs in obese tissue to exhibit functional and cytokine phenotypes that are typical of M1 macrophages but rather a mixed M1/M2 phenotype. Further studies focussed on the different ATM phenotypes in humans are needed to better understand their roles in obesity.

The link between obesity and chronic diseases can be demonstrated by ATMs through the formation of crown-like structures and the activation of effector immune molecules and the adaptive immune system. In lean individuals, ATMs are distributed in interstitial spaces between adipocytes. However, in obese tissue ATMs tend to aggregate around adipocytes that are dead due to hypertrophy, forming crown-like structures [11]. The formation of crown-like structures has been associated with an increase in insulin resistance [11], demonstrating the link between obesity and T2DM. Adipose tissue macrophages also play a role in the activation of the adaptive immune system through inducing naïve cluster of differentiation- (CD-)4^+ T cell differentiation into one of four subsets through antigen presentation. The activation of the adaptive immune system further contributes to the inflammatory state of obesity, which has been linked to increased risk of chronic disease development.

2.2.3 Adaptive immunity in obesity

Obesity also affects the adaptive immune system by inducing naïve $CD4^+$ T cell differentiation into one of four major subsets: T-helper- (Th-)1, Th2, Th17 and T regulatory (Treg). The four subsets are classified based on the cytokines they commonly secrete: Th1 secretes IFN-γ, Th2 secretes IL-4, Th17 secretes IL-17, and Treg secretes IL-10. As IFN-γ and IL-17 are both pro-inflammatory cytokines, Th1 and Th17 cells are referred to as pro-inflammatory immune cells. Conversely, Th2 and Treg are considered anti-inflammatory immune cells as they both secrete anti-inflammatory cytokines. The differentiation of naïve $CD4^+$ T cells occurs through many different mechanisms (Figure 2.2A), including antigen presentation and influence from other cytokines.

Antigen-presentation commonly involves the major histocompatibility complex class II (MHCII) molecule found on cells such as ATMs. The MHCII molecule presents specific antigens to naïve immune cells, triggering differentiation into specific subsets. $CD4^+$ T cell differentiation can also occur through cytokine production from other immune cells. For example, while Th1 differentiation is induced by pro-inflammatory IFN-γ secreted by M1, Th2 differentiation is induced by anti-inflammatory IL-4 secreted by M2. In obesity, the population of anti-inflammatory Th2 and Treg cells decreases [12, 13] while pro-inflammatory Th1 [13, 14] and Th17 [15, 16] profiles increase. The imbalance in pro-inflammatory to anti-inflammatory $CD4^+$ T cell subsets may be due to the high M1-to-M2 ratio in obese subjects, resulting in higher activation of pro-inflammatory immune cells. Other than $CD4^+$ T cells, pro-inflammatory $CD8^+$ T cells have also been largely studied in obesity and has also been reported to increase in frequency concurrently with obesity development [17].

There is a strong association between increases in Th1, Th17 and $CD8^+$ T cells in obesity with the development of chronic disease development. IFN-γ increases the expression of T-bet, a Th1 master regulator T-box transcription factor expressed in T cells, which in turn promotes

differentiation of naïve $CD4^+$ T cells into the Th1 subset [18]. Many knock-out studies have demonstrated the role of these pro-inflammatory immune cells in the metabolic dysregulation that occurs in obesity. In a study by Stolarczyk et al., insulin sensitivity and metabolic restoration was enhanced in T-bet-deficient mice with Th1 deficiency compared to wild-type mice [19]. Another study by Rocha et al. also found IFN-γ-deficient mice to display significant decreases in inflammatory gene expression along with improved glucose intolerance [14]. Coupled with a study showing rapid increases in messenger ribonucleic acid (mRNA) levels of IFN-γ following just one week of high-fat diet, the link between obesity and chronic diseases is demonstrated. High-fat diet also stimulates the development of Th17 cells which secretes IL-17, a blocker of insulin receptor signalling. Chuang et al. neutralised Th17-produced IL-17 in a group of mice, which resulted in decreased glucose levels and improved insulin sensitivity compared to wild-type mice [20]. Obesity dysregulates the immune system, creating a pro-inflammatory profile that is linked to the development of chronic diseases such as T2DM through insulin resistance.

Human studies have also demonstrated the association between obesity and chronic diseases through an increase in pro-inflammatory biomarkers. In one study comparing the percentage of T cell subsets between SAT and VAT of bariatric surgery subjects, Th1, Th17 and $CD8^+$ T cells were all significantly higher in the VAT compared to SAT [21]. This finding not only suggests obesity-related inhibition of adipogenesis, resulting in fat storage in ectopic fat depots, but also the increased inflammatory profile associated with obesity. In the same study, IL-6 expression was 2-fold greater and IL-17 expression was 3-fold higher in the VAT compared to the SAT [21]. Bertola et al. also reported a positive relationship between BMI and IL-17 expression after comparing adipose tissue from lean, overweight and obese humans [16]. A more recent study saw an upregulation of Th1 markers in pregnant women with gestational diabetes compared to pregnant women with normoglycemia [22]. However, further research in

this area is still required due to conflicting findings. While Th1, Th17 and $CD8^+$ T cell abundance were all higher in VAT of obese subjects in a study by McLaughlin et al., there was no association reported between higher pro-inflammatory biomarkers and insulin resistance [21]. In other studies, anti-inflammatory Th2 and Treg abundance increased in morbidly obese subjects while $CD8^+$ T cells were unchanged [23], and lean controls had higher levels of pro-inflammatory $CD8^+$ T cells compared to obese participants [24].

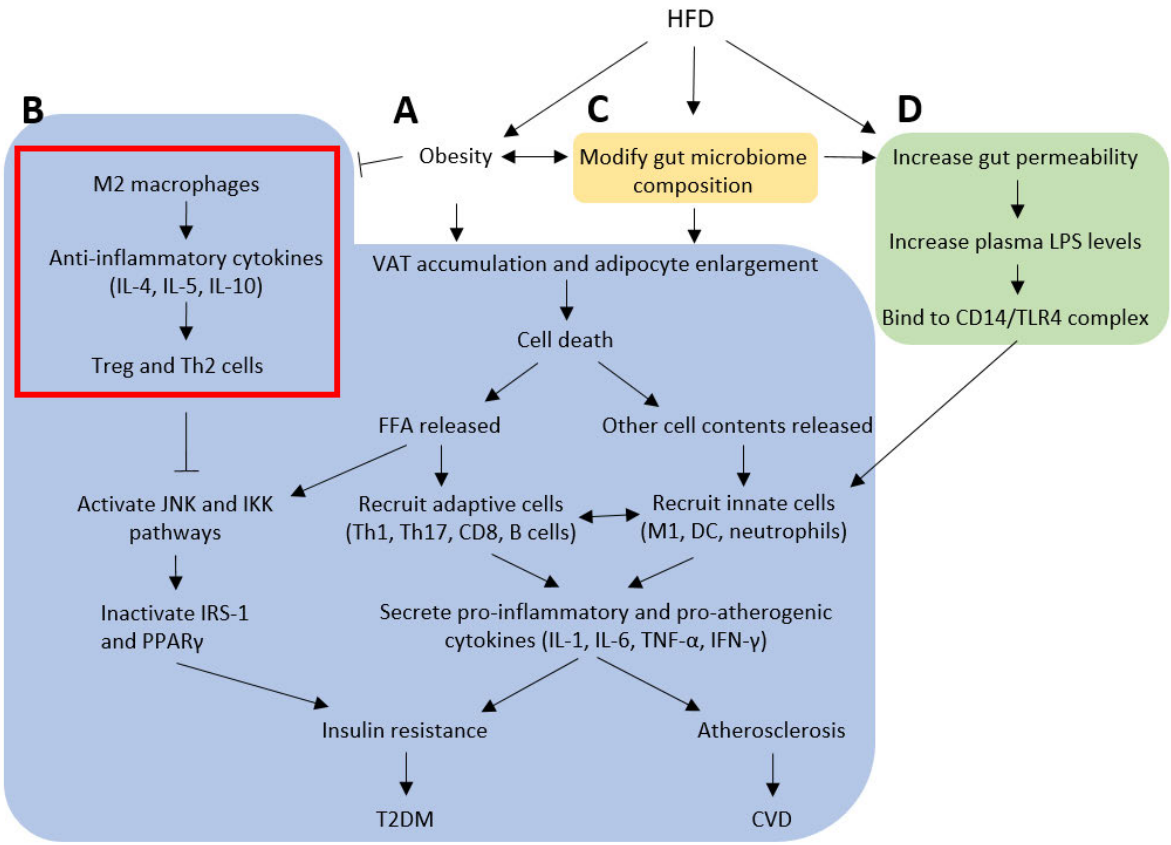


Figure 2.2. Example of four molecular pathways resulting from high-fat diet, each contributing to the state of chronic low-grade inflammation associated with obesity, leading to increased risk of chronic diseases. (A) Upregulation of pro-inflammatory biomarkers, (B) downregulation of anti-inflammatory biomarkers, (C) changes in gut microbial composition leading to upregulation of pro-inflammatory biomarkers, and (D) changes to gut permeability exacerbating pro-inflammation.

Chronic low-grade inflammation associated with obesity is also a result of the downregulation of protective anti-inflammatory biomarkers (Figure 2.2B). Both anti-inflammatory $CD4^+$ Th2 and Treg subsets have shown to improve insulin sensitivity. Two independent studies [25, 26] have both demonstrated increases in Th2 cytokine production in adipose tissue to improve insulin sensitivity. Treg also helps in maintaining insulin sensitivity by producing IL-10 which limits inflammation in adipose tissue, reducing insulin resistance [12]. Th2-produced IL-4 [27] and Treg-produced IL-10 [28] also activate insulin receptor substrate 1 (IRS-1) and peroxisome proliferator-activated receptor gamma (PPARγ) functions, improving insulin sensitivity [29]. Both $CD4^+$ Th2 and Treg have been found in smaller frequencies in obese adipose tissue compared to lean tissue. The pattern of increased pro-inflammatory biomarkers and decreased anti-inflammatory biomarkers in obesity development has been shown in both mice and human studies. After comparing lean and diet-induced obese mice, an increase in Th1-to-Treg ratio was observed, from 1.5-to-1 to 6.5-to-1 [28]. In humans, this ratio increased from 6-to-1 in lean adipose tissue to 12-to-1 in obese tissue [28]. At the same time, Th2 numbers significantly decreased by 50% in the VAT of high-fat fed mice compared to healthy controls [28]. A high Th1-to-Th2 ratio was also reported in pregnant women with gestational diabetes and their newborn macrosomic babies [22]. Additionally, Th1 numbers were 10- to 20-fold higher than both Treg and Th2 numbers in the adipose tissue of bariatric surgery subjects [21]. However, the effect obesity has on Th1-to-Th2 or Th1-to-Treg ratio could not be determined in this study as healthy controls were not included. The high Th1-to-Th2 and Th1-to-Treg ratios are likely to be a result of the high M1-to-M2 ratio mentioned previously as M1 induces Th1 differentiation from naïve $CD4^+$ T cells while M2 induces Th2 and Treg differentiation. A downregulation of protective $CD4^+$ Th2 and Treg, coupled with the upregulation of pro-inflammatory Th1 and Th17 would therefore exacerbate obesity-related insulin resistance, leading to T2DM development.

2.2.4 Free fatty acids in obesity

Other than through the innate and adaptive arms of the immune system, free fatty acids (FFA) can also contribute to adipose tissue inflammation (Figure 2.2A). Visceral adipocyte enlargement renders it susceptible to premature cell death [30]. Upon death, FFA and other cellular contents are released into the extracellular space [31]. Free fatty acids then activates pathways that further promote inflammation, including c-Jun NH_2-terminal kinase (JNK) and Iκβ kinase (IKK) pathway activation [31]. These pathways upregulate inflammation through nuclear factor-kappa beta (NF-κβ) transcription, which in turn inhibits IRS-1 and PPARγ function [32]. IRS-1 and PPARγ are both involved in insulin signalling pathways and thus inactivation of these molecules will lead to insulin resistance and subsequent T2DM [32]. Free fatty acid also promotes pro-inflammation by increasing the expression of genes associated with the MHCII antigen [33]. As MHCII is commonly found on antigen-presenting cells [33], such as macrophages, increased expression would also increase the differentiation of $CD4^+$ T cells into pro-inflammatory Th1 and Th17. Free fatty acid therefore contributes to the inflammation of adipose tissue associated with obesity, leading to an increased risk of chronic disease development.

2.3 Obesity and the gut microbiota

Another mechanism by which obesity-associated inflammation can be activated is through the gut microbiota (Figure 2.2C). The adult body is colonised by trillions of microbes [34], with over 100-fold more genes than the rest of the body [35]. The gut microbiota plays many important roles, such as extraction of energy from common polysaccharides that are indigestible by enzymes encoded from human genome [36]. Regulation of fat storage is another important function of the microbiome, as demonstrated by germ-free mice which were

protected against obesity and MetS [37]. Restoration of normal intestinal flora in the same mice led to significant increases in body fat and insulin resistance [37]. Similar results were reported by Turnbaugh et al. following transplantation of faecal microbiota from 9 lean mice and 13 obese mice into germ-free mice [38]. The mice with obese microbiota were able to extract more energy from food compared to mice with the lean microbiota [38]. In humans, Jumpertz and colleagues overfed 12 lean individuals for 3 consecutive days and observed a shift in gut microbiota that was associated also with increased energy harvest from food [39]. To accommodate the increased energy extracted from diet, adipocytes must increase in size, further adding to the obesity-associated inflammation from abdominal obesity (Figure 2.2A). The impact of the gut microbiome on fat storage was demonstrated by germ-free mice being resistant to weight gain and accumulation of fat [40]. However, what constitutes obese microbiota is still debatable with many studies reporting different gut microbe compositions to be associated with obesity [41].

The four most dominant bacterial phyla in the gut are: Gram-negative Bacteroidetes and Proteobacteria, and Gram-positive Actinobacteria and Firmicutes [37]. Many studies comparing obese and healthy individuals have found differences in gut microbiota composition [42-44]. The main difference is in the Firmicutes-to-Bacteroidetes ratio which has been evident in both murine [45, 46] and human studies [47, 48], demonstrating obesity-related dysregulation of the gut microbiome. In a murine study conducted by Ley et al., obese mice had 50% reduction in Bacteroidetes abundance and a proportional increase in Firmicutes abundance compared to lean mice [45]. Ley et al. also examined differences in gut microbiota composition of obese and lean humans [49]. The study compared the faecal microbiota from 12 obese people, who underwent diet therapy for 52 weeks, and 2 healthy controls [49]. Prior to the intervention, the obese group had fewer Bacteroidetes and higher Firmicutes abundance. Throughout the therapy, the abundance of Bacteroidetes increased while Firmicutes decreased,

which were defining characteristics of the healthy group. A more recent study conducted by Koliada et al. also reported similar findings in 61 Ukrainian adults, with the obese cohort having a higher Firmicutes-to-Bacteroidetes ratio [48]. Most et al. also conducted a similar study with the addition of examining gender-differences, with a sample size of 15 obese men and 14 obese women. The high Firmicutes-to-Bacteroidetes ratio was prevalent in both men and women, with men having a lower ratio compared to women. In addition, the ratio in men was found to be inversely proportional to peripheral insulin sensitivity [47]. However, this correlation was not present in women, suggesting women to be less sensitivity to metabolic aberrations resulting from changes in the gut microbiota. Although each study found higher Firmicutes-to-Bacteroidetes ratios in obese participants, the sample sizes used were all very small. It would therefore be worthwhile to confirm these findings in a study with a larger sample size.

2.4 Obesity and metabolic endotoxaemia

Obesity-associated inflammation can also occur through the elevation of plasma liposaccharide (LPS) [50], often referred to as metabolic endotoxaemia (Figure 2.2D). Liposaccharides, or endotoxins, are fragments from the cell walls of Gram-negative gut bacteria that have a strong affinity for chylomicrons. High-fat feeding increases both the production of chylomicrons by epithelial intestinal cells and intestinal permeability [51]. Liposaccharides attached to chylomicrons are then transported through the gut wall and towards target tissues, including adipose tissue [51]. Toll-like receptor 4 (TLR4) are LPS receptors located on the surface of innate immune cells [51], which are present in large quantities within obese adipose tissue [41]. The binding of LPS to TLR4 is assisted by co-receptor CD14 and myeloid differentiation factor

2 (MD-2), leading to the activation of NF-κβ and subsequent transcription of pro-inflammatory adipocytokines [52].

Translocation of LPS from gut microbes to adipocytes initiates inflammation, suggesting metabolic endotoxaemia to be the link between obesity-related dysregulation of the gut microbiome and the immune system. The association between metabolic endotoxaemia and diet-induced obesity was demonstrated by Cani et al. using LPS injections and CD14 knockout mice. Following injections of LPS into lean mice for 4 weeks, the mice had similar levels of inflammation and insulin resistance as diet-induced obese mice [39]. On the other hand, removal of the LPS receptor CD14 resulted in resistance to diet-induced obesity and related disorders [52]. The involvement of gut microbes was also demonstrated in studies using antibiotics and transplanting faecal microbes from lean to obese mice, resulting in decreases in both plasma LPS and TNF-α levels [53]. Metabolic endotoxaemia and obesity associations is also evident in humans, with plasma endotoxaemia levels of 8 healthy individuals increasing by 71% following a month of consuming high-fat diet [53]. Similarly, following a single high-fat meal, plasma LPS increased by 50% in 12 healthy men [40] and 12% in 10 lean individuals just 3 hours after eating [51]. Each of the human studies [40, 51, 53] have also reported on the elevation in systemic inflammation in conjunction with heightened plasma LPS levels. The direct correlation between metabolic endotoxaemia and chronic diseases was analysed by Pussinen et al. [54]. Liposaccharide levels were higher in the diabetic cohort, consisting of almost 1,000 participants, compared to 6,170 non-diabetics [54]. Due to its link with obesity, metabolic endotoxaemia is an essential factor to consider when analysing the multifaceted interactions involved in obesity and related diseases.

2.5 Correlation-based network analysis

The current literature indicates that obesity results in dysregulation of numerous body systems, including metabolic function, the immune system and the gut. Despite extensive research in this field, no biomarker profiles have been found to stratify individuals at risk of obesity-related diseases into risk groups for targeted preventive or therapeutic intervention. The use of integrated analysis may advance obesity research by overcoming the challenges faced when analysing the complex network associated with obesity-related disease. Recently, the use of correlation-based network analysis (CNA) has increased as researchers begin to recognise the multifactorial nature of disease mechanisms which involves multiple biological systems. Correlation-based network analysis is a data-mining method that reduces multidimensional data while retaining the majority of information needed for interpretation [55]. Correlation-based network analysis allows visualisation of interactions between biomarkers which are altered by disease states, allowing researchers to understand mechanisms that promote the development of disease [56]. Each variable is represented as a node and the link between variables are shown as edges. There are many properties of CNA [55], including:

- vertex degree: the number of edges stemming from a particular node;
- network diameter, the maximum number of possible shortest paths between two nodes;
- network density, the ratio of the total number of edges in the network to the number of all possible edges;
- betweenness centrality (BC), the number of shortest paths between any two nodes that passes through the node in question;
- closeness centrality, the mean number of nodes a particular node must pass through to reach every node in the network; and
- network hub, a cluster of nodes connected by a large number of edges.

Figure 2.3 provides a visual representation of a correlation-based network analysis and the different properties which can be used to describe the relationships between biomarkers within a network. Vertex degree is often represented by different node sizes and is evident in Figure 2.3, with node 7 having the highest vertex degree and thus is the biggest node in the network. Additionally, nodes 2 and 6 are shown to have the highest betweenness centrality scores as they provide the only link between nodes from the two hubs of nodes. The two clusters of nodes, linked together by nodes 2 and 6, form network hubs that may signify an important pathway that should be analysed by researchers when using CNA to better understand multifactorial diseases.

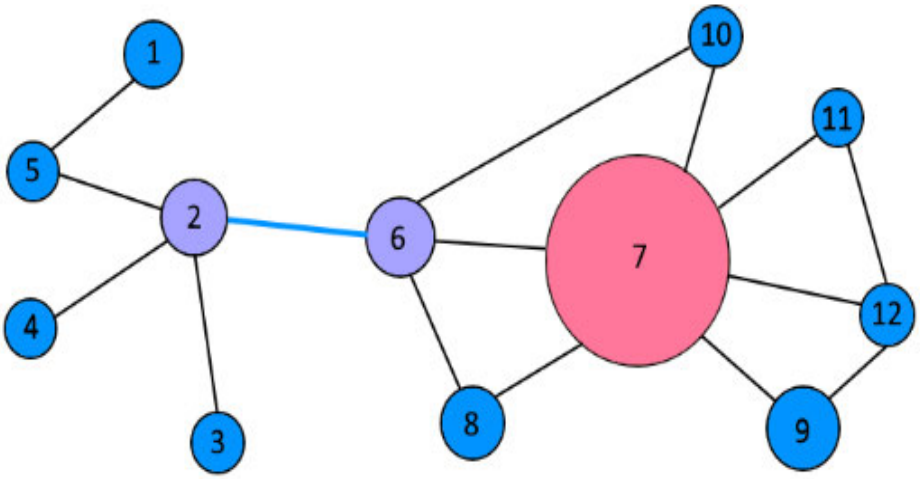


Figure 2.3. Different properties of correlation-based network analysis, including large vertex degree shown by node 7 and high betweenness centrality shown by node 2 and node 6.

Each of these properties is useful in identifying important relationships between biomarkers in diseases that may otherwise be missed in univariate analysis. In a study by Haring et al. [57], CNA was used to assess each MetS risk factor to identify the more prominent ones that drive MetS pathophysiology. The study used data from a five-year follow-up study with 3,187 participants. The network analysis revealed a network hub with healthy individuals that developed MetS after five years. Within the network hub were the risk factors: low HDL-C,

high blood pressure and central obesity. Through the use of CNA, the study was able to identify the three risk factors that individuals should be particularly aware of when looking to prevent the development of MetS and related diseases. Network analysis was also used by Zhang, Xin & Lu [58] to visualise common gene signatures and signalling pathways between MetS, dementia and diabetes. The study found 10 different genes to have high vertex degrees, including apolipoprotein E (APOE) which was common across MetS, dementia and diabetes. Similarly, Su et al. [59] also identified common genes between obesity and six other obesity-related diseases using network analysis. Across the 2,499 assessed genes, 31 were found to be hub genes for the six obesity-related diseases: coronary artery disease, diabetes, hypertension, breast cancer, polycystic ovary syndrome, and kidney cancer. Out of the 31 key genes, 17 were also in the obesity gene set which suggests the effect of obesity on the development of the other six obesity-related diseases. Through the studies by Zhang, Xin & Lu and Su et al., the practicality of network analysis was demonstrated, with its ability to identify common genes which can then serve as targets for intervention purposes. Network analysis is also useful in its ability to highlight relationships between biomarkers, which may not be evident in traditional hypothesis recognition, providing insight into disease pathogenesis to advance prevention research.

2.6 Multivariate prediction models

Beyond the recognition of the association between biomarkers in diseases, which may be targeted for therapeutic intervention, there are also tools that can be used to classify individuals into categories of interest. Through the categorising of individuals into their respective groups, classification models can also identify the features that played a fundamental role in accurate classification. Classification models can be used to predict diseases in individuals by

identifying hard-to-detect patterns in large data. There are a range of multivariate prediction modelling techniques available, including: logistic regression (LR), decision tree (DT), support vector machine (SVM), and artificial neural network (ANN). To evaluate the performance of a prediction model, there are different criteria that can be used, including:

- classification rate: the percentage of data that was correctly classified by the model;
- sensitivity: the ability of the model to correctly identify those with the disease of interest (true positive ratio);
- specificity: the ability of the model to correctly identify those without the disease of interest (true negative ratio); and
- area under the curve (AUC): measure of the classification accuracy of a model through sensitivity and specificity numbers.

2.6.1 Logistic regression

Logistic regression is a type of classification method that displays the relationship between the dependent and predictor variables using a logit function, forming an S-shaped curve. In a binary LR model, the asymptotes of the curve are '0' and '1', representing the two categories of the dependent variable, such as 'healthy' and 'obese', respectively. The function will then calculate the probability of each data point, or participant, being closer to either '0' or '1', based on the values of the predictor variables used. The points are then placed on the curve accordingly. Through this calculation, LR has the ability to conclude which predictor variables allowed the model to classify data points into their respective categories more accurately. Logistic regression has therefore been widely used in many studies, such as those looking to identify the top predictor variables involved in disease development. In a study by Al-Thani et al [60], LR was used to evaluate which obesity measurement best predicts at least two other factors of MetS in 2,496 Qatari individuals. The three measurements of obesity included BMI, waist

circumference (WC) and waist-to-hip ratio (WHR). Waist circumference was found to be the best predictor, followed by WHR and then BMI. However, the sensitivity and specificity values for detecting MetS in men was only 57.8% and 58.4%, respectively. The low predictive performance is likely a result of the lack of haematological measures used to build the LR model, as MetS is a condition comprised mostly of haematological risk factors. Al-Thani et al. stated that many participants in the study had refused to provide blood samples, largely due to cultural reasons. With this limitation, many measurements would have had missing data and thus features that may have predicted MetS better than WC may have been missed. Another study, conducted by Gradidge et al. [61], also used LR to predict MetS but took a different approach. Gradidge et al. raised the concern that WC has been found to be highly correlated with many CVD risk factors and therefore deliberately excluded this feature from the analysis. By doing so, the study has eliminated the risk of finding other features to be associated with MetS solely through their relationship with WC. The study concluded subcutaneous fat, homeostatic model assessment of insulin resistance (HOMA-IR), age, smoking, adiponectin and leg fat to be important features for MetS prediction. The same study also built a model with the inclusion of WC and found that subcutaneous fat was no longer considered to be a significant risk factor. Unfortunately, as the aim of the study was to identify the important features in MetS prediction, the classification accuracy, sensitivity and specificity values were not reported. Comparisons between the models built by Al-Thani et al. could therefore not be made. Both Al-Thani et al. and Gradidge et al. did not mention the use of any cross-validation techniques to optimise the models constructed. On the other hand, a more recent study by Mao et al. [62] evaluated the ability of LR to predict CVD using MetS- and CVD-related risk factors with both internal and external validation. The predictive model was built with 706 Kazakh participants and 18 risk factors (Table 2.1). Using LR, the study identified WC, weight, body adiposity index, bilirubin, HDL-C and APOE to be the important features for CVD prediction.

For internal validation, with a sample size of 384, the model achieved an AUC value of 0.857, 80.49% sensitivity and 81.71% specificity. The study then used data collected from a previous study, with 243 participants, for external validation and the model produced an AUC value of 0.914, 81.82% sensitivity and 88.48% specificity. While Al-Thani et al. constructed a LR model with very poor performance, Mao et al. was able to build one with high predictive ability. The difference in the performance of the models in the two studies is likely due to the health parameters that was used for training the models. Al-Thani et al. did not use any haematological health parameters for the prediction of MetS, a condition comprised mostly of haematological risk factors. Overall, the outcomes of each of the three studies suggest that LR is a powerful tool that is very useful in identifying important features in disease prediction. However, the use of relevant risk factors for the disease being predicted is important for producing a model with high predictive ability.

Table 2.1. List of studies which used classification methods for the prediction of MetS and related diseases.

Authors	Sample	Parameters used	Important parameters identified	Classification method	Performance
Al-Thani et al., 2016 [60]	Qatari cohort: 1,123 Control, 1,373 MetS	BMI, WC, WHR	WC	LR	Sensitivity: 57.8% (men), 65.7% (women); Specificity: 58.4% (men), 66.7% (women)
Gradidge et al., 2016 [61]	African women cohort: 552 Control, 90 MetS	Age, education, employment status, smoking, menopausal status, HIV status, adiponectin, leptin, HOMA-IR, subcutaneous and visceral adipose thickness, total body fat, total FFSTM, HC, WC	Age, smoking, adiponectin, HOMA-IR, subcutaneous fat, total FFSTM	LR	
Mao et al., 2018 [62]	706 Xinjiang Kazakhs	Weight, WC, BAI, SBP, DBP, HDL-C, APOA, FPB, FMN, ALT, AST, A-HBDH, TBIL, IBIL, ALB, UA, CREA, BUN	Weight, WC, BAI, TBIL, IBIL, HDL-C, APOA	LR	AUC: 0.857; Sensitivity: 80.49%; Specificity: 81.71%
Worachartcheewan et al., 2010 [63]	Thai cohort: 2,377 Control, 2,690 MetS	Gender, age, SBP, DBP, BMI, FPG, BUN, CREA, UA, cholesterol, TG, HDL-C, LDL-C, AST, ALT, ALP, HG, HCT, WCC, PLT	Gender, SBP, DBP, FPG, TG, HDL-C	DT	Accuracy: 99.86%; Sensitivity: 99.89%; Specificity: 99.83%
AlJarullah, 2011 [64]	Pima Indian female cohort: 518 Control, 206 MetS	Age, BMI, DBP, plasma glucose, DPF, number of times pregnant	Age, BMI, plasma glucose, DPF, number of times pregnant	DT	Accuracy: 78.18%
Karimi-Alavijeh, Jalili & Sadeghi, 2016 [65]	Isfahan cohort: 1,511 Control; 596 MetS	Gender age, weight, BMI, WC, WHR, HC, physical activity, smoking history, hypertension, antihypertensive medication use, SBP, DBP, FPG, 2-hour glucose, TG, total cholesterol, LDL, HDL-C, MCV, MCH	TG, BP, BMI	DT	Accuracy: 73.9%; Sensitivity: 75.8%; Specificity: 72%
Chen & Chen, 2019 [66]	Taiwanese cohort: 187,002 Control,	WC, BP, FPG, TG, HDL-C	WC, BP, TG, HDL-C	DT	

	14,085 MetS				
Kim et al., 2012 [67]	1013 Korean cohort	WC, BP, FPG, TG, HDL-C		DT	
				SVM (Polynomial)	Accuracy: 75.7%; Sensitivity: 77.4%; Specificity: 74%
Kumari & Chitra, 2013 [68]	460 Pima Indian cohort	Age, BMI, plasma glucose, DBP, DPF, 2-hour serum insulin, number of times pregnant, triceps skinfold thickness		SVM (RBF)	Accuracy: 78%; Sensitivity: 80%; Specificity: 76.5%
Chung et al., 2014 [69]	Korean population	Gender, Age, WC, BMI, family history of diabetes, hypertension, alcohol consumption, smoking, physical activity	Gender, Age, WC, BMI, family history of diabetes, hypertension, alcohol consumption	LR	
				SVM (RBF)	Accuracy: 64.9%; AUC: 0.761; Sensitivity: 78.9%; Specificity: 61.2%
Fernández-Navarro et al., 2019 [70]	Northern Spain cohort: 20 Control, 34 Pre-obese/Obese	Age, gender, %Fat, energy intake, alcohol consumption, smoking, leptin, MDA, CRP, total FFA, arachidonic acid, DHA, EPA, stearic acid, linoleic acid, linolenic acid, oleic acid, palmitoleic acid, gamma-linolenic acid, palmitic acid, Akkermansia, Bacteroides group, Bacteroides-Prevotella-Porphyromonas, Bifidobacterium, Faecalibacterium, Clostridia XIVa group, Blautia coccoides-Eubacterium rectale, Lactobacillus group	Gender, EPA, Palmitic acid, Bifidobacterium, Faecalibacterium, Bacteroides, DHA	DT	Sensitivity: 71%; Specificity: 86%
				SVM	Sensitivity: 57%; Specificity: 83%
Hirose et al., 2011 [71]	410 Japanese male cohort	Age, BMI, SBP, DBP, FPG, TG, HDL-C, LDL-C, AST, ALT, HMW-adiponectin, total adiponectin, glycated ALB, cholesterol, FFA, insulin, HOMA-IR, smoking, MetS	Age, BMI, DBP, HDL-C, LDL-C, HOMA-IR	LR	Sensitivity: 27%; Specificity: 95%
				ANN	Sensitivity: 93%; Specificity: 91%

Meng et al., 2013 [72]	752 Control, 735 diabetic or prediabetic	Age, BMI, gender, family history of diabetes, marital status, education level, work stress, sleep duration, physical activity, preference for salty food, fish consumption, coffee consumption		LR	Accuracy: 76.54%; Sensitivity: 79.40%; Specificity: 73.54%
				DT	Accuracy: 76.97%; Sensitivity: 78.11%; Specificity: 75.78%
				ANN	Accuracy: 72.59%; Sensitivity: 79.40%; Specificity: 65.47%
Alić et al., 2017 [73]	Bosnia and Herzegovina cohort: 150 Control, 150 MetS	WC, BP, glucose, HDL-C, TG		ANN	Accuracy: 96%; Sensitivity: 96%; Specificity: 92.7%
Disse et al., 2017 [74]	565 Obese	Age, gender, height, body weight		ANN	Accuracy: 68.1%; Percentage error: 8.6%

ALB: albumin; ALP: alkaline phosphatase; ALT: alanine aminotransferase; APOA: apolipoprotein A; AST: aspartate aminotransferase; BAI: body adiposity index; BMI: body mass index; BP: blood pressure; BUN: blood urea nitrogen; CREA: creatinine; CRP: C-reactive protein; DBP: diastolic blood pressure; DHA: docosahexaenoic acid; DPF: diabetes pedigree function; EPA: eicosapentaenoic acid; FFA: free fatty acid; FFSTM: fat-free soft-tissue mass; FMN: fructosamine; FPG: fasting plasma glucose; HC: hip circumference; HCT: haematocrit; HDL-C: high-density lipoprotein cholesterol; HG: haemoglobin; HMW-adiponectin: high molecular weight-adiponectin; HOMA-IR: homeostatic model assessment of insulin resistance; IBIL: indirect bilirubin; LDL-C: low-density lipoprotein cholesterol; MCH: mean corpuscular haemoglobin; MCV: mean corpuscular volume; MDA: malondialdehyde; MetS: metabolic syndrome; PLT: platelet; SBP: systolic blood pressure; TBIL: total bilirubin; TG: triglycerides; UA: uric acid; WC: waist circumference; WCC: white blood cell count; WHR: waist-to-hip ratio; α-HBDH: α-hydroxybutyrate dehydrogenase

2.6.2 Decision tree

Decision trees are a type of supervised machine learning algorithm that can be used to classify datapoints into categories of interest. The algorithm is built on a set of if-then rules and has a hierarchical tree-like structure, with the most important variable at the top as the root node. The root node then bifurcates into other decision nodes before reaching the terminal node, or leaf, which represents the classification result (Figure 2.4).

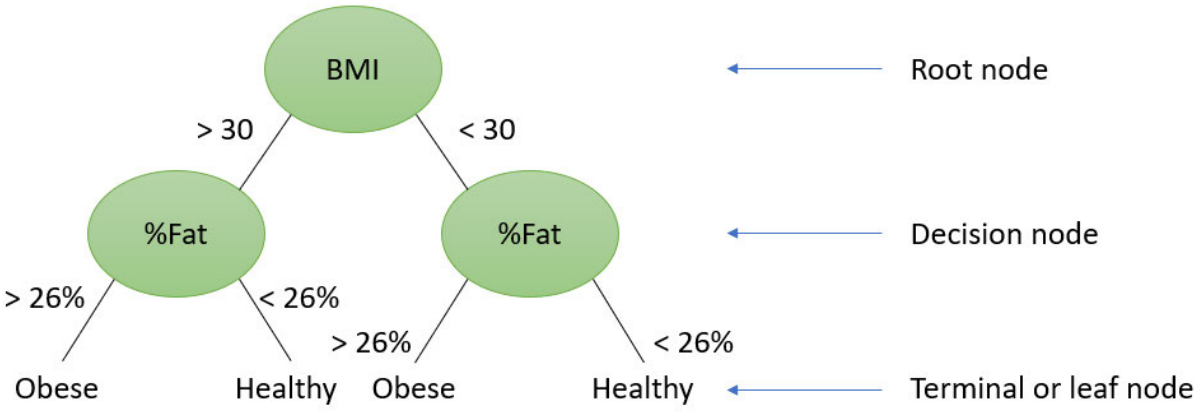


Figure 2.4. Example of decision tree layout.

Full-sized decision trees, which contain all the variables within a database for the prediction of a disease, are often overfitted whereby the model fits too closely to the data and thus is unable to accurately predict unseen data. To prevent issues of overfitting and increase the classification accuracy or the model, as well as decrease computational time and cost [75], DTs are often optimised through pruning. During the process of pruning, sections of the tree with little power to classify instances are removed. A common way of pruning a tree is using the complexity parameter (CP). Prior to building a DT, the user will decide on a CP. If the cost of splitting the tree by another variable exceeds the CP, the tree building will cease, limiting the overall size of the tree. Due to its comprehensible nature, DTs have been used in many studies to identify

the features that are instrumental in predicting diseases such as metabolic syndrome. Chen & Chen [66] used the DT algorithm to evaluate the importance of the five MetS criteria for the prediction of MetS in a Taiwanese population and found triglycerides to be the most important feature, followed by WC, blood pressure and HDL-C. Interestingly, fasting plasma glucose did not appear in the DT, which may be attributable to the imbalance in the number of participants in the two studied categories. While the study used a large sample size of 201,087, only 7% of the participants had MetS, with the remaining being healthy controls. On the other hand, other studies have reported fasting plasma glucose to either be the root node [64] or a decision node [63, 65, 67] in predicting MetS. The classification accuracies of these studies were all relatively high, with the lowest being 75.8% [65]. It's also worth noting that studies which used more variables for input into the DT yielded higher classification accuracies. A study by Worachartcheewan et al. [63] attained an accuracy of 99.86%, sensitivity of 99.89% and a specificity of 99.83% from a DT built by 20 health parameters. At the same time, a study [64] that only included six attributes yielded an accuracy of 78.18%, sensitivity of 80.5% and specificity of 72.33%. A list of the parameters used by each study can be found in Table 2.1. Although the DT used by AlJarullah performed more poorly than in the study Worachartcheewan et al., it still produced relatively high accuracy, sensitivity and specificity values. Additionally, neither studies specifically mentioned the use of pruning to prevent overfitting, which DTs are very prone to, thus the validity of the performance of each model is unknown. Regardless, DTs have shown to be a useful tool in identifying key features that assist researchers to predict diseases with high accuracy and ease of interpretability.

2.6.3 Support vector machine

Support vector machines are another type of supervised machine learning tool. The algorithm plots each data point of a dataset onto n-dimensional space, where n is the number of features in the dataset. The resulting plot then displays two clusters of data points for the two categories of interest, with each point now being referred to as a vector. The vectors nearest the space separating the two classes are known as support vectors. These support vectors are used as reference points for drawing the line, or hyperplane, that best separates the two classes. The optimal hyperplane is one that is furthest away from the support vectors of each class, optimising the margin to lower misclassification rate [76]. While Figure 2.5 demonstrates a linear separation of data, many datasets cannot be separated linearly as data points will fall into the wrong category.

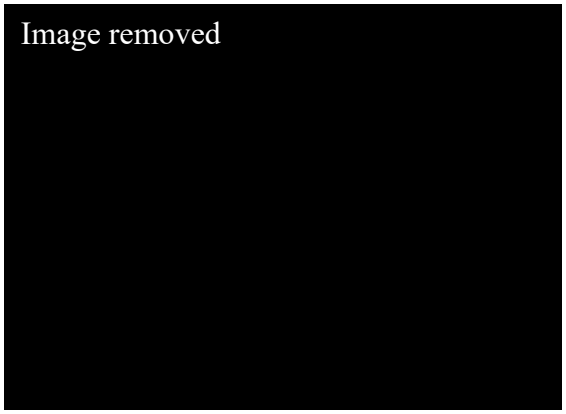


Figure 2.5. Illustration of the different factors that constitute a support vector machine [77].

Depending on the type of data that is being used to build the SVM, the way in which the data needs to be separated by the hyperplane may differ. Kernel functions may be applied to the SVM model to adjust how the data can be separated and deciding the most suitable kernel function is part of the optimisation process. Figure 2.6 shows how the data may be separated differently through the use of kernel functions which adds more dimensions to the data.

Figure 2.6. Demonstration of how kernel functions separate two groups of data at a higher dimensional space [78].

To adjust the separation of data, there are different kernel functions that may be applied. The four common kernel functions used in SVM are linear, polynomial, sigmoid and radial basis function (RBF). Deciding which kernel function to apply to SVM is important as it will dictate the predictive ability of the model. Each kernel function has its own formula which can be optimised using different shape parameters shown in Table 2.2.

Table 2.2. The four different types of kernel functions used in SVM, their formulas and the parameters that need to be specified by the user to optimise the model.

Kernel function	Formula	Optimisation parameter
Linear	$K(X,Y) = X^T Y$	C
Polynomial	$K(X,Y) = (\gamma \cdot X^T Y + r)^d$	C, γ, r, d
Sigmoid	$K(X,Y) = \tanh(\gamma \cdot X^T Y + r)$	C, γ, r
RBF	$K(X,Y) = \exp(-\gamma \cdot \|X - Y\|^2)$	C, γ

C: cost; γ: gamma; r: coefficient; d: degree

Each shape parameter is able to change the hyperplane and margin of the SVM in a particular way, which then changes the decision boundary. In doing so, the model may fit more closely to the data, leading to a high classification accuracy. However, the risk of overfitting also increases as the model will fit too closely to the data on which is was trained. As such, the model will be unable to accurately predict any unseen data. The tuning of shape parameters is therefore a very important optimisation step to produce a model with high classification

accuracy while reducing the risk of overfitting. In addition to being computationally complex, SVM also have a "black box" nature in the sense that the results generated are not directly sensible to researchers. Despite the complexity of SVM, however, it's often used by researchers due to its robustness to noise and overall high generalisation performance [79]. The algorithm was used by Karimi-Alavijeh, Jalili & Sadeghi [65] to predict the 7-year incidence of MetS in 2,107 Iranian participants using 21 different health parameters Al-Thani et al. stated that many participants in the study had refused to provide blood samples, largely due to cultural reasons. With this limitation, many measurements would have had missing data and thus features that may have predicted MetS better than WC may have been missed. Another study, conducted by Gradidge et al. [61], also used LR to predict MetS but took a different approach. Gradidge et al. raised the concern that WC has been found to be highly correlated with many CVD risk factors and therefore deliberately excluded this feature from the analysis. By doing so, the study has eliminated the risk of finding other features to be associated with MetS solely through their relationship with WC. The study concluded subcutaneous fat, homeostatic model assessment of insulin resistance (HOMA-IR), age, smoking, adiponectin and leg fat to be important features for MetS prediction. The same study also built a model with the inclusion of WC and found that subcutaneous fat was no longer considered to be a significant risk factor. Unfortunately, as the aim of the study was to identify the important features in MetS prediction, the classification accuracy, sensitivity and specificity values were not reported. Comparisons between the models built by Al-Thani et al. could therefore not be made. Both Al-Thani et al. and Gradidge et al. did not mention the use of any cross-validation techniques to optimise the models constructed. On the other hand, a more recent study by Mao et al. [62] evaluated the ability of LR to predict CVD using MetS- and CVD-related risk factors with both internal and external validation. The predictive model was built with 706 Kazakh participants and 18 risk factors (Table 2.1). The study decided to use the polynomial kernel and yielded a classification

accuracy of 75.7%, sensitivity of 77.4% and specificity of 74%. Although the study has recognised the impact the selected kernel function may have on the classification accuracy, there was no justification through the comparison of results obtained by the other kernel functions to explain why the polynomial function was used. There is a possibility that the performance of the model may be increased or decreased if a different kernel function was used. While studies using SVM to predict MetS are limited, many studies have used the algorithm to predict the incidence of the associated chronic disease, T2DM. In 2014, Chung et al. [69] screened for pre-diabetes in a Korean population using SVM and 7 predictor variables: age, BMI, hypertension, gender, alcohol intake, WC and family history of diabetes. The algorithm was trained with a sample size of 3,134 and predicted pre-diabetes in 1,551 participants with an accuracy of 64.9%, 78.9% sensitivity and 61.2% specificity. The study utilised the RBF kernel and attained a classification accuracy of 64.9%, with 78.9% sensitivity and 61.2% specificity. Although the choice of the kernel function was also not justified by Chung et al., it was reported that the model constructed performed better than the existing clinical score model used for screening pre-diabetes. It cannot be concluded whether the classification accuracy will be improved through the use of a different kernel function or the inclusion of better health parameters, such as haematological biomarkers. Kumari & Chitra [68] used blood chemical parameters such as plasma glucose and serum insulin the prediction of diabetes and achieved a higher performance compared to Chung et al., with 78% accuracy, 80% sensitivity and 76.5% specificity. The higher SVM performance in the study by Kumari & Chitra may be due to the inclusion of blood chemical parameters, such as plasma glucose and serum insulin, which may explain the variance better than features used by Chung et al. The authors of the study stated that RBF was the kernel function of choice as it handles high dimensional data very well, though it is unclear whether it was compared to the results that may be obtained with other kernel functions. Although Kumari & Chitra used a smaller sample size of 460 Pima Indians,

evidently the performance of the model was not compromised. On the other hand, Fernández-Navarro et al. [70] used a sample size of 66 participants to investigate the relationship between serum FFA and faecal microbiota in obesity. The model achieved a sensitivity of 57% and a specificity of 83% which is likely a result of inadequate optimisation due to the small sample size used. Additionally, the kernel function used by the study was not reported, which may have also impacted the predictive performance of SVM. Collectively, the studies have demonstrated that, when relevant features are used to predict diseases in a large cohort, SVM has a high generalisation performance. Although the choice of RBF in each of the studies mentioned was not clearly justified, it is likely due to its ability to map and approximate almost any nonlinear function through the fine-tuning of its optimisation parameters.

2.6.4 Artificial neural networks

Similar to the support vector machine, artificial neural networks also have a reputation of having a "black box" nature. Artificial neural network derives its name from the way it mimics how the human brain functions. The neurons of our brain form a network that is able to process the information received from our senses and in turn produce an appropriate reaction. In the same way, information is fed into an ANN and the importance of each input variable, also referred to as its weight, is then assessed. If the combined weight of all the input variables is greater than the activation threshold, an output is generated. Neural networks are commonly comprised of a few layers: the input layer, the hidden layers and the output layer (Figure 2.7). The number of nodes in the input layer reflects the features present in the dataset and the number of nodes in the output layer is the number of classes being predicted. The number of hidden layers and hidden layer nodes used in a neural network is decided by the researcher. As there is no rule-of-thumb when it comes to the number of hidden layers and hidden layer neurons (HLN) used, it is more of an optimisation step in constructing neural networks.

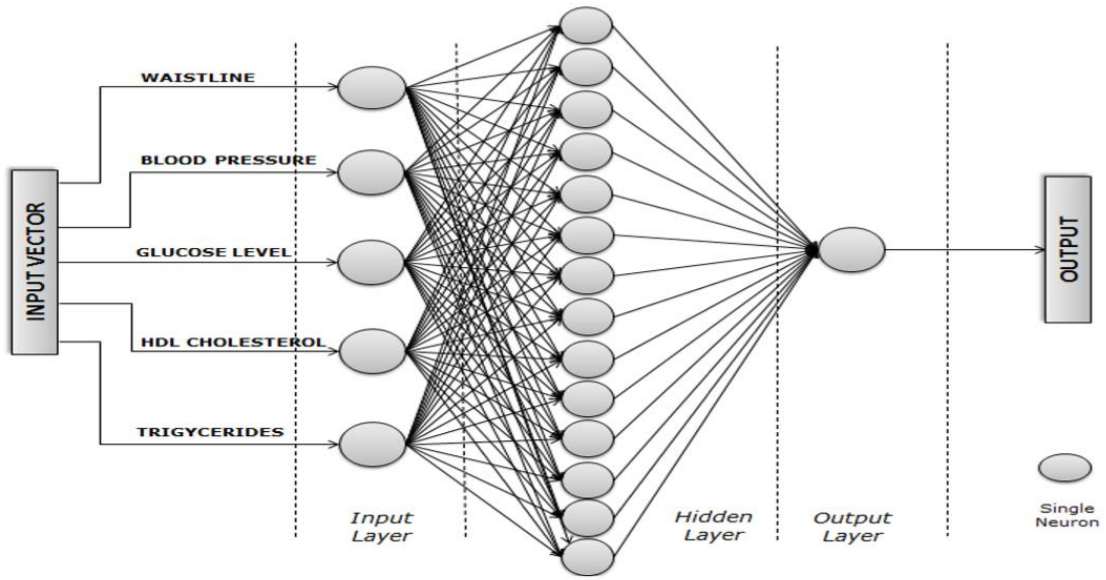


Figure 2.7. Example of the artificial neural network structure, with the input layer, one hidden layer with 14 hidden layer neurons, and the output layer [73]. (Reprinted by permission from Springer Nature Customer Service Centre GmbH: Springer Nature, Springer eBook, Classification of metabolic syndrome patients using implemented expert system, Alić et al., © 2017.)

Alić et al. [73] evaluated how well the MetS criteria was able to predict MetS in a Bosnian and Herzegovinian cohort while also exploring the impact of HLN numbers on the predictive ability of ANN. The study built several networks with 5, 10, 14 and 17 HLN using the same dataset. The study concluded that the network built with 14 HLN was able to classify MetS in 300 individuals most accurately, with 94.3% accuracy, 96% sensitivity and 92.7% specificity. However, as Alić et al. only used 6 input neurons to build the network, there's a high chance that the use of 14 hidden layer neurons would result in overfitting. In addition, the study only reported the findings of the training set rather than the testing set which is likely due to the low performance found in the testing set as a result of overfitting. ANNs have also been regularly used to predict MetS-related chronic diseases, including diabetes. Using a neural network built with 15 hidden layer nodes, Meng et al. [72] was able to classify diabetes in 1,487 individuals with an accuracy of 77.87%, 80.68% sensitivity and 75.13% specificity. Diabetes status was predicted using 12 input variables derived from self-reporting questionnaires with variables such as: age, BMI, family history of diabetes, preference for salty food, coffee consumption, and fish consumption. Compared to the study by Alić et al., the ratio of input layer neurons to

HLN used was much smaller and after comparing the accuracy attained by the training and testing sets there was no evidence of overfitting. The ANN was also optimised using a training set consisting of 70% of the participants and a testing set containing 30% of participants. The performance of the model may have been further improved, however, if biochemical measures, like glucose concentration, had been included, as seen in a study conducted by Hirose et al. [71]. The 6-year incidence of MetS was predicted in 410 Japanese male teachers using 17 health parameters, including 12 haematological measures. The use of relevant biomarkers for the prediction of MetS is likely what contributed most to achieving a sensitivity of 93% and a specificity of 91%. Although the study reported that the leave-some-out cross-validation technique was implemented, the overall structure of the neural network was not shown. As such, it cannot be concluded whether the number of HLN used led to the model overfitting. Biochemical measurements itself can also be predicted using ANN, as demonstrated by Pappada et al. [80]. Using factors like insulin dosages, metered glucose values, nutritional intake, and lifestyle and emotional factors, Pappada et al. predicted 88.6% participants to have normal glucose concentrations, 72.6% to be hyperglycaemic and 2.1% to be hypoglycaemic. The results from the study suggested that the model had performed well, with 92.3% of the predictions being regarded as clinically accepted predictions. Unfortunately, the study was conducted with a sample size of 17, which is considered to be too small to construct a reliable prediction model. Although ANNs are known for having a “black box” nature, and thus the importance of variables cannot be easily determined, the model had the advantage of scaling well to high-dimensional data. Additionally, there are also methods of overcoming this limitation, including the incorporation of a feature selection technique to the model.

2.6.5 Comparing classification algorithms

Prediction models as a whole have been used extensively to predict diseases other than MetS and related chronic diseases, however, the conclusion of which prediction model performs best has never been reached. Deciding which prediction model to use to answer a particular research question is a challenge within itself and thus studies have compared different models by applying them on the same dataset. Some studies listed in Table 2.1 compared the ability of different prediction models to predict MetS and related diseases using the same dataset. Fernández-Navarro et al. also used DTs to predict BMI using various biomarkers and found a higher performance of 71% sensitivity and 86% specificity compared to the 57% sensitivity and 83% specificity achieved by SVM. In addition to the higher predictive ability, DT was able to identify the biomarkers that contributed significantly to a high model performance: serum eicosapentaenoic acid (EPA), palmitic acid, and gut microbes Bifidobacterium and Faecalibacterium. If a researcher was looking to identify the biomarkers that best predict MetS, DTs would be the better option. On the other hand, Hirose et al. used both LR and ANN to predict the 6-year incidence of MetS in a Japanese male cohort. Logistic regression performed very poorly with a sensitivity of 27% compared to ANN which had a sensitivity of 93%. Although LR was able to identify the important biomarkers to help ANN achieve a high performance, it was not able to predict the incidence of MetS well. However, the structure of the ANN that was built was not reported and thus there is a chance that it may have overfitted, allowing it to achieve a substantially high performance. In another study, LR performed slightly better than ANN, with classification accuracies of 76.54% and 72.59%, respectively. The same study also used DTs which had the highest prediction accuracy of 76.97%. While the performance for all three prediction models were relatively high, it may be further improved if biochemical and haematological biomarkers were used to build the models instead of demographic variables. The choice of which classification model to use is very much

dependent on the research question being asked. For researchers looking to identify the important biomarkers for clinical diagnosis, LR and DTs would be the better choice. In other studies, such as image recognition for diseases like breast cancer, the high dimensionality of data involved would be better handled by SVM and ANN. Furthermore, the variables chosen for the construction of models will largely dictate the performance of the classification model. In cases such as MetS development, haematological measurements would explain more variance in the data compared to demographic variables. Deciding the best combination of risk factors to include in model construction may be difficult and thus many studies have combined feature selection techniques with classification models to increase the performance of the base model.

2.7 Feature selection

The input variables used to build a prediction model has a large impact on the overall performance of the model in terms of classification accuracy. The selection of features to include in the model is therefore an essential process in building prediction models. The fundamentality of this process has led to techniques being developed for the purpose of feature selection. There are three different methods of feature selection which can be categorised into either classifier-independent (filter method) or classifier-dependent (wrapper and embedded methods) [81].

Filter methods performs feature selection by ranking the input features and filtering out the less relevant variables without the use of an algorithm, hence classifier-independent. The relevance of variables is measured by the influence it has on the class label which can be determined based on correlation or mutual information (MI) [82]. The correlation between input features and the class label is assessed by Pearson correlation while MI calculates the dependency of

the class label on each input feature. The input features with high correlation with the class label or high MI are then highly ranked and the top features are then used for prediction in an algorithm. The downfall with filter methods, however, is the failure to consider interdependency of the input features within the subset [81]. The correlation or dependency between input features used for prediction suggests that redundant variables are likely to be included. With the inclusion of redundant variables, the subset used for prediction is not at its optimal size, increasing the computational time for prediction models. Furthermore, while being classifier-independent helps filter methods to avoid the issue of the data overfitting to the learning algorithm, it is difficult to find a suitable algorithm to use when it comes time for prediction.

Both wrapper and embedded methods do not have this problem as they are both classifier-dependent. Two common wrapper methods are the sequential selection algorithm and the heuristic search algorithm [82]. With the sequential selection algorithm, the method begins with the variable that provides the highest classification accuracy when used in a prediction model. The variable that provides the second highest classification accuracy is then added on and this continues until the accuracy starts to decrease [82]. The input features with high predictive ability then become a subset that is to be used for prediction models. Meanwhile, the heuristic search approach randomly generates subsets of input features before testing each subset with a classifier to identify which subset provides the highest prediction accuracy. Both wrapper methods incorporate cross-validation to find the ideal subset for prediction [81]. The use of cross-validation, however, means that the same feature will be evaluated multiple times as subsets are not stored for later retrieval, as each training dataset is different. Wrapper methods are therefore prone to overfitting as, with cross-validation, the model may learn the data too well and thus its generalisation capability worsens. The inclusion of cross-validation

in wrapper methods makes it the most computationally expensive out of the three feature selection techniques and hence it is unfeasible for high-dimensional datasets.

To reduce the computation time in wrapper methods, embedded methods incorporates feature selection into the training process of constructing a prediction model [81]. At the same time, embedded methods also overcome the issue of interdependency between input features, as seen in MI of filter methods. To achieve this, embedded methods maximise MI between the selected feature and the class label while also minimising the MI between the selected feature and other input features. The learning algorithm that is typically used for the embedded method is DT. The DT algorithm selects features in each step and bifurcates into smaller subsets. The subset of features with the most child nodes are considered to be more informative than other subsets. While feature selection with concurrent prediction model construction saves computation time, without cross-validation, embedded models cannot achieve the high learning capacity and prediction accuracy that wrapper methods can attain.

The decision of which feature selection technique to use may be a difficult choice to make for researchers. Ultimately the choice depends on the research question being asked, the dataset being used and the resources available, such as computational power. For studies that are using high-dimensional data to identify biological differences between participants with chronic diseases and healthy controls, the wrapper method may be the most suitable. Although its computation time is longer than the both filter and embedded methods, the wrapper method is able to identify the subset of variables with the highest predictive ability through cross-validation. Additionally, as the feature selection and prediction model building processes are different, the construction and training of the prediction model itself will not be affected. Furthermore, while the wrapper method is prone to overfitting, this will also not affect the performance of the prediction model as they are separate processes. Genetic algorithm (GA) is an example of a wrapper method that utilises the heuristic search approach and has been shown

to be reliable in biomarker discovery [83]. Reflecting Charles Darwin's theory of natural selection, GA identifies the best variables in each generation to find the optimal combination for building a prediction model. The five steps in genetic algorithm are:

1. Initial population

To start off, a random population of chromosomes is generated (Figure 2.8). Each chromosome is a potential solution which represents the inclusion or exclusion of certain input features. Every input feature is referred to as a gene, with '0' indicating the inclusion of the feature while '1' signifies its exclusion in the solution.

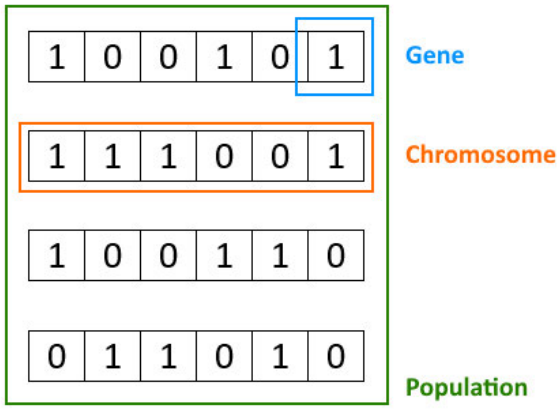


Figure 2.8. Terminology used in genetic algorithm (GA), with genes representing each input feature, chromosomes representing a subset of features and the population showing the collection of different subsets chosen by GA.

2. Fitness function

Each randomly generated chromosome within a population is evaluated using a fitness function. In the case of prediction models, this is typically the classification accuracy that each set of input features, or chromosomes, is able to produce. After calculating the fitness score for each individual, the selection step is initiated. However, if the stopping criterion is reached, the feature selection process ends. The stopping criterion can either be a maximum number of iterations set by the user or if the algorithm has not improved after a predefined number of iterations.

3. Selection

The chromosomes are ranked from highest to lowest fitness values and the better half is kept for further analysis. The selection step reduces the number of chromosomes to N/2, where N is the initial size of the population. The remaining chromosomes which had a high fitness score are then used for recombination.

4. Crossover

During the crossover step, two chromosomes with high fitness values are selected at random and recombined to produce new offspring (Figure 2.9A). Both the parents and offspring are kept and the process continues until the new population reaches the same size as the initial population.

5. Mutation

To maintain diversity within the new population, chromosomes may undergo mutation, whereby genes may be flipped (Figure 2.9B). In this case, a feature that was initially excluded is now included in the solution, resulting in a new chromosome being produced. If the stopping criterion has not been reached following the mutation step, the process will repeat from the selection step until the stopping criterion is satisfied.

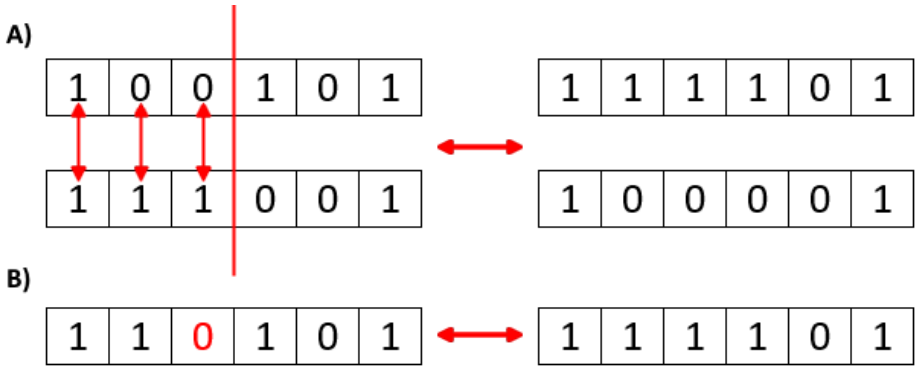


Figure 2.9. Demonstration of the (A) crossover and (B) mutation steps of genetic algorithm.

As the method of combining ANN with GA is relatively new, there are very limited studies that have used this hybrid model to predict MetS. There are, however, studies which have used GA in conjunction with ANN to both identify the important features in diabetes as well as predict diabetes prevalence. Karegowda, Manjunath and Jayaram [83] used the hybrid model to predict diabetes in 392 Pima Indian individuals using eight input variables: plasma glucose, diastolic blood pressure, triceps skinfold thickness, serum insulin, BMI, diabetes pedigree function and age. The study applied GA as a feature selection technique and identified four important features for diabetes prediction: plasma glucose, serum insulin, BMI and age. The comparison in predictive ability of the models using all eight features and using the four features selected by GA found the latter to yield a higher performance, with classification accuracies of 77.7% and 84.7%, respectively. In a more recent study, Mortajez and Jamshidinezhad [84] also used a hybrid ANN and GA model to predict diabetes. The classification accuracy of the hybrid model was 84.5%, which is higher than the 68% accuracy achieved from a simple neural network using a similar dataset. Mortajez & Jamshidinezhad did not report the features selected by GA as the aim of the paper was to compare the hybrid model with other constructed networks. From both these studies, the hybrid models were shown to perform better than a simple neural network. The results from these studies have demonstrated the need for including feature selection techniques in building prediction models to increase the performance of the model.

2.8 Future directions

In 2014-5, over half of Australian adults and over a quarter of Australian children were classified as either overweight or obese [85]. Three years later, in 2017-8, the number of Australian adults reported increased to two-thirds while the number of Australian children

affected remained the same [86]. Obesity affects many different biological systems across the body simultaneously, including the immune system and the gut microbiome. The alteration of the gut microbiome in obesity increases the energy harvested from consumed food, resulting in increased adipocyte size for energy storage. As adipocytes increase in size, they exert a pro-inflammatory influence on the immune system. The complex interactions between different body systems may be one reason why no biomarker profile has been found to identify individuals at the highest risk of obesity-related chronic diseases and who can be targeted for specific preventive or therapeutic intervention. Despite extensive research, most obesity studies have only focussed on one particular area of obesity-related dysregulation. However, as a multifactorial disease, integrated analysis may be the key to producing a biomarker profile across different body systems that may be targeted for the purpose of intervention. Many multivariate analytical methods are now available, uncovering underlying interactions between different body systems which may otherwise be missed in univariate analysis.

Prediction models are a powerful method capable of handling high-dimensional data with very high accuracy. With its ability to not only accurately predict disease but also reveal the variables that contribute significantly to the prediction, classification models have quickly become a popular method of data mining. In addition, there are techniques such as feature selection and ensemble modelling which are able to further improve the performance of individual prediction models, making them an attractive option to researchers. The current study will be using prediction models to simultaneously assess the impact MetS has on haematological measures, gut microbial composition and gene expression levels.

2.9 References

[1] Cornier, M.A., et al., The metabolic syndrome. Endocr Rev. vol. 29, pp. 777-822, 2008.

[2] Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (2001). Executive Summary of The Third Report of The National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, And Treatment of High Blood Cholesterol In Adults (Adult Treatment Panel III). JAMA, 285(19), 2486–2497. 2001.

[3] International Diabetes Federation, The IDF consensus worldwide definition of the Metabolic Syndrome. pp. 1-24, 2006.

[4] Belligoli, A., et al., Characterization of subcutaneous and omental adipose tissue in patients with obesity and with different degrees of glucose impairment. Sci Rep. vol. 9, pp. 1-12, 2019.

[5] Verboven, K., et al., Abdominal subcutaneous and visceral adipocyte size, lipolysis and inflammation relate to insulin resistance in male obese humans. Sci Rep. vol. 8, pp. 1-8, 2018.

[6] Liu, L., et al., Visceral adipose tissue is more strongly associated with insulin resistance than subcutaneous adipose tissue in Chinese subjects with pre-diabetes. Curr Med Res Opin. vol. 34, pp. 123-129, 2018.

[7] Rakotoarivelo, V., et al., Inflammatory cytokine profiles in visceral and subcutaneous adipose tissues of obese patients undergoing bariatric surgery reveal lack of correlation with obesity or diabetes. EBioMedicine. vol. pp. 237-47, 2018.

[8] Neeland, I.J., et al., Associations of visceral and abdominal subcutaneous adipose tissue with markers of cardiac and metabolic risk in obese adults. Obesity (Silver Spring). vol. 21, pp. 439-447, 2013.

[9] Sell, H., Habich, C., and Eckel, J., Adaptive immunity in obesity and insulin resistance. Nature Reviews Endocrinology. vol. 8, pp. 709-716, 2012.

[10] McLaughlin, T., et al., Role of innate and adaptive immunity in obesity-associated metabolic disease. J Clin Invest. vol. 127, pp. 5-13, 2017.

[11] Boutens, L. and Stienstra, R., Adipose tissue macrophages: going off track during obesity. Diabetologia. vol. 59, pp. 879-894, 2016.

[12] Deiuliis, J., et al., Visceral adipose inflammation in obesity is associated with critical alterations in Tregulatory cell numbers. PLoS One. vol. 6, pp. 1-11, 2011.

[13] Strissel, K.J., et al., T-cell recruitment and Th1 polarization in adipose tissue during diet-induced obesity in C57BL/6 mice. Obesity (Silver Spring). vol. 18, pp. 1918-1925, 2010.

[14] Rocha, V.Z., et al., Interferon-gamma, a Th1 cytokine, regulates fat inflammation: a role for adaptive immunity in obesity. Circ Res. vol. 103, pp. 467-476, 2008.

[15] Endo, Y., et al., Obesity drives Th17 cell differentiation by inducing the lipid metabolic kinase, ACC1. Cell Rep. vol. 12, pp. 1042-1055, 2015.

[16] Bertola, A., et al., Identification of adipose tissue dendritic cells correlated with obesity-associated insulin-resistance and inducing Th17 responses in mice and patients. Diabetes. vol. 61, pp. 2238-2247, 2012.

[17] Nishimura, S., et al., CD8+ effector T cells contribute to macrophage recruitment and adipose tissue inflammation in obesity. Nat Med. vol. 15, 2009.

[18] Deng, T., et al., Class II major histocompatibility complex plays an essential role in obesity-induced adipose inflammation. Cell Metab. vol. 17, pp. 411-422, 2013.

[19] Stolarczyk, E., et al., Improved insulin sensitivity despite increased visceral adiposity in mice deficient for the immune cell transcription factor T-bet. Cell Metab. vol. 17, pp. 520-533, 2013.

[20] Chuang, H.-C., et al., HGK/MAP4K4 deficiency induces TRAF2 stabilization and Th17 differentiation leading to insulin resistance. Nat Commun. vol. 5, pp. 1-14, 2014.

[21] McLaughlin, T., et al., T-cell profile in adipose tissue is associated with insulin resistance and systemic inflammation in humans. Arterioscler Thromb Vasc Biol. vol. 34, pp. 2637-2643, 2014.

[22] Seck, A., et al., Th1/Th2 dichotomy in obese women with gestational diabetes and their macrosomic babies. J Diabetes Res. vol. 2018, pp. 1-7, 2018.

[23] van der Weerd, K., et al., Morbidly obese human subjects have increased peripheral blood CD4+ T cells with skewing toward a Treg- and Th2-dominated phenotype. Diabetes. vol. 61, pp. 401-408, 2012.

[24] Lynch, L.A., et al., Are natural killer cells protecting the metabolically healthy obese patient? Obesity (Silver Spring). vol. 17, pp. 601-605, 2009.

[25] Miller, A.M., et al., Interleukin-33 induces protective effects in adipose tissue inflammation during obesity in mice. Circ Res. vol. 107, pp. 650-658, 2010.

[26] Molofsky, A.B., et al., Innate lymphoid type 2 cells sustain visceral adipose tissue eosinophils and alternatively activated macrophages. J Exp Med. vol. 210, pp. 535-549, 2013.

[27] Chng, M.H., et al., Adaptive Immunity and Antigen-Specific Activation in Obesity-Associated Insulin Resistance. Mediators of Inflammation. vol. 2015, pp. 1-15, 2015.

[28] Winer, S., et al., Normalization of obesity-associated insulin resistance through immunotherapy: CD4+ T Cells control glucose homeostasis. Nat Med. vol. 15, pp. 921-929, 2009.

[29] Gao, Z.-g. and Ye, J.-p., Why do anti-inflammatory therapies fail to improve insulin sensitivity? Acta Pharmacol Sin. vol. 33, pp. 182-188, 2012.

[30] Rausch, M., et al., Obesity in C57BL/6J mice is characterized by adipose tissue hypoxia and cytotoxic T-cell infiltration. Int J Obes (Lond). vol. 32, pp. 451-463, 2008.

[31] Boden, G., Obesity and free fatty acids. Endocrinol Metab Clin North Am. vol. 37, pp. 635-46, 2008.

[32] Liu, Z., et al., High-Fat Diet Induces Hepatic Insulin Resistance and Impairment of Synaptic Plasticity. PloS. vol. 10, pp. 1-16, 2015.

[33] Xiao, L., et al., Large adipocytes function as antigen-presenting cells to activate CD4(+) T cells via upregulating MHCII in obesity. Int J Obes (Lond). vol. 40, pp. 112-120, 2016.

[34] Ley, R.E., et al., Evolution of mammals and their gut microbes. Science. vol. 320, pp. 1647-1651, 2008.

[35] Marotza, C.A. and Zarrinpar, A., Treating obesity and metabolic syndrome with fecal microbiota transplantation. Yale J Biol Med. vol. 89, pp. 383-388, 2016.

[36] Flint, H.J., et al., Polysaccharide utilization by gut bacteria: potential for new insights from genomic analysis. Nat Rev Microbiol. vol. 6, pp. 121-131, 2008.

[37] Tilg, H. and Kaser, A., Gut microbiome, obesity, and metabolic dysfunction. J Clin Invest. vol. 121, pp. 2126-2132, 2011.

[38] Turnbaugh, P.J., et al., An obesity-associated gut microbiome with increased capacity for energy harvest. Nature. vol. 444, pp. 1027-31, 2006.

[39] Cani, P.D., et al., Metabolic endotoxemia initiates obesity and insulin resistance. Diabetes. vol. 56, pp. 1761-72, 2007.

[40] Erridge, C., et al., A high-fat meal induces low-grade endotoxemia: evidence of a novel mechanism of postprandial inflammation. Am J Clin Nutr. vol. 86, pp. 1286-1292, 2007.

[41] Warmbrunn, M.V., et al., Gut microbiota: a promising target against cardiometabolic diseases. Expert Rev Endocrinol Metab. vol. 15, pp. 13-27, 2020.

[42] Schwiertz, A., et al., Microbiota and SCFA in lean and overweight healthy subjects. Obesity. vol. 18, pp. 190-195, 2010.

[43] Duncan, S.H., et al., Human colonic microbiota associated with diet, obesity and weight loss. Int J Obes (Lond). vol. 32, pp. 1720-1724, 2008.

[44] Zhang, H., et al., Human gut microbiota in obesity and after gastric bypass. PNAS. vol. 106, pp. 2365-2370, 2009.

[45] Ley, R.E., et al., Obesity alters gut microbial ecology. Proc Natl Acad Sci U S A. vol. 102, pp. 11070-11075, 2005.

[46] Turnbaugh, P.J., et al., Diet-Induced Obesity Is Linked to Marked but Reversible Alterations in the Mouse Distal Gut Microbiome. Cell Host & Microbe. vol. 3, pp. 213-223, 2008.

[47] Most, J., et al., Gut microbiota composition strongly correlates to peripheral insulin sensitivity in obese men but not in women. Benef Microbes. vol. 8, pp. 557-62, 2017.

[48] Koliada, A., et al., Association between body mass index and Firmicutes/Bacteroidetes ratio in an adult Ukrainian population. BMC Microbiol. vol. 17, pp. 1-6, 2017.

[49] Turnbaugh, P.J., et al., An obesity-associated gut microbiome with increased capacity for energy harvest. Nature. vol. 444, pp. 1027-1031, 2006.

[50] Vallianou, N.G., Stratigou, T., and Tsagarakis, S., Microbiome and diabetes: Where are we now? Diabetes Res Clin Pract. vol. 146, pp. 111-118, 2018.

[51] Ghanim, H., et al., Increase in plasma endotoxin concentrations and the expression of Toll-like receptors and suppressor of cytokine signaling-3 in mononuclear cells after a high-fat, high-carbohydrate meal: implications for insulin resistance. Diabetes Care. vol. 32, pp. 2281-2287, 2009.

[52] Cani, P.D., et al., Changes in gut microbiota control metabolic endotoxemia-induced inflammation in high-fat diet-induced obesity and diabetes in mice. Diabetes. vol. 57, pp. 1470-1481, 2008.

[53] Pendyala, S., Walker, J.M., and Holt, P.R., A high-fat diet is associated with endotoxemia that originates from the gut. Gastroenterology. vol. 142, pp. 1100-1101, 2012.

[54] Pussinen, P.J., et al., Endotoxemia is associated with an increased risk of incident diabetes. Diabetes Care. vol. 34, pp. 392-397, 2011.

[55] Batushansky, A., Toubiana, D., and Fait, A., Correlation-based network generation, visualization, and analysis as a powerful tool in biological studies: a case study in cancer cell metabolism. Biomed Res Int. vol. January 2016, pp. 1-9, 2016.

[56] Nishihara, R., et al., Biomarker correlation network in colorectal carcinoma by tumor anatomic location. BMC Bioinformatics. vol. 18, pp. 1-14, 2017.

[57] Haring, R., et al., A network-based approach to visualize prevalence and progression of metabolic syndrome components. PLoS One. vol. 7, pp. 1-7, 2012.

[58] Zhang, W., Xin, L., and Lu, Y., Integrative analysis to identify common genetic markers of metabolic syndrome, dementia, and diabetes. Med Sci Monit. vol. pp. 5885-5891, 2017.

[59] Su, L.-n., et al., Network analysis identifies common genes associated with obesity in six obesity-related diseases. J Zhejiang Univ Sci B. vol. 18, pp. 727-732, 2017.

[60] Al-Thani, M.H., et al., Prevalence and determinants of metabolic syndrome in Qatar: results from a National Health Survey BMJ Open. vol. 6, pp. 1-10, 2016.

[61] Gradidge, P.J.-L., et al., Metabolic and body composition risk factors associated with metabolic syndrome in a cohort of women with a high prevalence of cardiometabolic disease. PLoS One. vol. 11, pp. 1-13, 2016.

[62] Mao, L., et al., Metabolic syndrome in Xinjiang Kazakhs and construction of a risk prediction model for cardiovascular disease risk. PLoS One. vol. 14, pp. 1-14, 2018.

[63] Worachartcheewan, A., et al., Identification of metabolic syndrome using decision tree analysis. Diabetes Res Clin Pract. vol. 90, pp. e15-18, 2010.

[64] AlJarullah, A.A., *Decision tree discovery for the diagnosis of type II diabetes*, in *2011 International Conference on Innovations in Information Technology*. 2011: Abu Dhabi. pp. 303-307.

[65] Karimi-Alavijeh, F., Jalili, S., and Sadeghi, M., Predicting metabolic syndrome using decision tree and support vector machine methods. ARYA Atheroscler. vol. 12, pp. 146-152, 2016.

[66] Chen, M.-S. and Chen, S.-H., A data-driven assessment of the metabolic syndrome criteria for adult health management in Taiwan. Int J Environ Res Public Health. vol. 16, pp. 1-11, 2019.

[67] Kim, T.N., et al., A decision tree-based approach for identifying urban-rural differences in metabolic syndrome risk factors in the adult Korean population. J Endocrinol Invest. vol. 35, pp. 847-852, 2012.

[68] Kumari, V.A. and Chitra, R., Classification of diabetes disease using support vector machine. IJERA. vol. 3, pp. 1797-1801, 2013.

[69] Chung, J.W., et al., *Screening for pre-diabetes using support vector machine model*, in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. 2014: Chicago, IL. pp. 2472-2475.

[70] Fernández-Navarro, T., et al., Exploring the interactions between serum free fatty acids and fecal microbiota in obesity through a machine learning algorithm. Food Res Int. vol. 121, pp. 533-541, 2019.

[71] Hirose, H., et al., Prediction of metabolic syndrome using artificial neural network system based on clinical data including insulin resistance index and serum adiponectin. Comput Biol Med. vol. 41, pp. 1051-1056, 2011.

[72] Meng, X.-H., et al., Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. Kaohsiung J Med Sci. vol. 29, pp. 93-99, 2013.

[73] Alić, B., et al. *Classification of metabolic syndrome patients using implemented expert system*. 2017. Singapore: Springer Singapore.

[74] Disse, E., et al., An artificial neural network to predict resting energy expenditure in obesity. Clin Nutr. vol. 37, pp. 1661-1669, 2018.

[75] Patil, D.D., Wadhai, V.M., and Gokhale, J.A., Evaluation of decision tree pruning algorithms for complexity and classification accuracy International Journal of Computer Applications. vol. 11, pp. 23-30, 2010.

[76] Furey, T.S., et al., Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics. vol. 16, pp. 906-914, 2000.

[77] MathWorks. *Support vector machines for binary classification*. n.d. [cited 2018 30 October]; Available from: https://au.mathworks.com/help/stats/support-vector-machines-for-binary-classification.html.

[78] Jain, R. *Simple tutorial on SVM and parameter tuning in python and R*. HackerEarth 2017 [cited 2019 27 August]; Available from: https://www.hackerearth.com/blog/developers/simple-tutorial-svm-parameter-tuning-python-r.

[79] Barakat, N. *Diagnosis of Metabolic Syndrome: A Diversity Based Hybrid Model*. 2016. Cham: Springer International Publishing.

[80] Pappada, S.M., et al., Neural network-based real-time prediction of glucose in patients with insulin-dependent diabetes. Diabetes Technol Ther. vol. 13, pp. 135-141, 2011.

[81] Liu, C., et al., A new feature selection method based on a validity index of feature subset. Pattern Recognit. vol. 92, pp. 1-8, 2017.

[82] Chandrashekar, G. and Sahin, F., A survey on feature selection methods. Computers & Electrical Engineering. vol. 40, pp. 16-28, 2014.

[83] Karegowda, A.G., Manjunath, A.S., and Jayaram, M.A., Application of genetic algorithm optimized neural network connection weights for medical diagnosis of Pima Indians diabetes IJSC. vol. 2, pp. 15-23, 2011.

[84] Mortajez, S. and Jamshidinezhad, A., An artificial neural network model to diagnosis of type II diabetes. J Res Med Dent Sci. vol. 7, pp. 66-70, 2019.

[85] Australian Bureau of Statistics, *4364.0.55.001 - National Health Survey: First Results, 2014-15* 2015: Canberra, Australia.

[86] Australian Bureau of Statistics, *4364.0.55.001 - National Health Survey: First Results, 2017-18*. 2018: Canberra, Australia.

CHAPTER 3
Discovery of biomarkers in MetS using correlation-based network analysis

3.1 Abstract

Metabolic syndrome (MetS) is a collection of risk factors, including obesity, which increases the risk of developing chronic diseases. Its multifactorial nature, with impacts across different body systems, has made it difficult for researchers to identify biomarker profiles characterising individuals more at risk of developing MetS and related diseases. As univariate analysis only provides a broad comparison between MetS and healthy controls, the current study employed correlation-based network analysis (CNA) to better understand the relationships between biomarkers affected by MetS. The results of the current study build upon a previously published paper which had used preliminary data and can be found in Appendix 3.2. In the current study, correlation networks were constructed using four groups of measurements obtained from 117 healthy weight and 35 obese with MetS individuals: anthropometric measures, haematological measures, gene expression levels and gut microbial composition. The obese with MetS network had a denser network in each of the four variable groups, with higher numbers of correlations found. In addition, the obese with MetS network found correlations between biomarkers across different variable groups, which was not found in the healthy weight network. The relationship between biomarkers from different body systems emphasises the complexity of MetS development. In particular, the CNA revealed molecular interactions that suggest the link between MetS and other conditions, including anaemia of inflammation, while also identifying key hubs in the expression of transcription factor EB (TFEB), lipocalin 2 (LCN2) and cluster of differentiation- (CD-)68. The expression of TFEB is important for regulatory T cell differentiation while neutrophils express both LCN2 and CD68. The frequency of regulatory T cells and neutrophils were considered to be key hubs in the previous study using preliminary

data and thus the results of the current study support that of the previous study. Overall, correlation networks were found to outperform univariate analysis by providing the means to better understand biomarkers involved in disease development which is critical for future intervention studies looking to reduce the incidence of MetS and related diseases.

3.2 Introduction

Obesity is a multifactorial disease that causes the dysregulation of many different body systems, including the immune system [1] and gut microbiome [2]. As one of the risk factors of metabolic syndrome (MetS), obesity is therefore also associated with an increased risk of developing chronic diseases, including type 2 diabetes mellitus (T2DM) and cardiovascular disease (CVD). The complex aetiology of obesity has made the identification of biomarker profiles to classify individuals more at risk of obesity- and MetS-related diseases a challenge for many researchers. Biomarkers across different body systems have similar roles and thus the inactivation of one biomarker may have little to no effect on the overall system. Through simultaneous analysis, the complex interactions between cells and molecules across different body systems may be revealed, allowing researchers to identify pathways that may be targeted together in future research. One method of simultaneous analysis is through correlation-based network analysis (CNA) which allows the visualisation of interactions between biomarkers and is therefore useful when trying to understand disease development. There are many different properties of CNA that are useful in describing the relationships between biomarkers and their importance within a system, including network density, vertex degree, betweenness centrality. The detailed explanation of these different properties can be found in Section 2.5, page 39. These properties were useful in previous MetS studies in identifying hubs of biomarkers that were common in MetS and related diseases. Zhang, Xin & Lu [3] found 10 different genes,

including apolipoprotein E (APOE), that were common across MetS, dementia and diabetes. Similarly, Su et al. [4] identified 31 out of the 2,499 assessed genes to be hub genes for obesity-related diseases. Many of the key genes were also found in the obesity gene set, suggesting the effects of obesity on the development of chronic diseases. The current study applied the same analytical method to detect key hubs of biomarkers across different systems of the body that were affected by obesity and MetS. In obesity research, the understanding of molecular interactions is more beneficial than the broad comparison of biomarker measurements between obese and healthy weight individuals. The use of multivariate analysis, such as CNA, is therefore an important analytical tool that should be used in conjunction with univariate analysis when looking to gain insight into disease pathogenesis.

3.3 Research design

3.3.1 Study design and ethics

Correlation-based networks were constructed using data collected from 152 participants that were classified as either obese with MetS, as per the International Diabetes Federation (IDF) criteria [5], or healthy weight control. All participants involved in the study were aged between 18 and 65 years and provided written informed consent prior to their involvement in the study. As the study aimed to identify biomarker profiles to characterise MetS, participants with medical conditions known to also alter these potential biomarkers, such as diabetes and hypertension, were excluded from the study. Participants with conditions that would affect the immune system, including cancer, Crohn's disease, liver disease, and irritable bowel syndrome, were also excluded. Additionally, participants were excluded if they used any medications or supplements that would affect the measurements two weeks prior to the study. Such medications and supplements included non-steroidal anti-inflammatory drugs (NSAIDs), fish

oil, probiotics and antihypertensives. The participants from the study were recruited under different studies completed by the same research group, the Mucosal Immunology Research Group, and all received ethical approval (2015/229, 2019/257, AHS/12/14/HREC, MED/19/15/HREC, 2017/646, 2014/537 and 2013/868). To reduce the potential bias associated with combining samples from various other studies, the participants were chosen based on the same strict inclusion and exclusion criteria.

3.3.2 Sample collection

Fasting blood samples were collected for analysis of metabolic (lipids, glucose, glycated haemoglobin [HbA1c]) and inflammatory (C-reactive protein [CRP], erythrocyte sedimentation rate [ESR]) measures. In addition, RNA was isolated and gene expression was measured using nCounter PlexSet-48 assays (nCounter® PanCancer Immune Profiling Panel, NanoString Technologies, Washington, USA). The 48 genes selected for the PlexSet were based on the results of an exploratory analysis using an immune profiling panel of 770 genes on 12 healthy weight and 11 obese with MetS individuals. Based on the results, 3 common housekeeping genes were selected based on their high, average and low expression (ribosomal protein lateral stalk subunit P0 [RPLP0], glucose-6-phosphate dehydrogenase [G6PD] and ATP-binding cassette subfamily F member 1 [ABCF1], respectively). From the remaining 45 genes, 39 were selected based on the following criteria:

- A fold change in expression greater than 1.3 between the healthy weight controls and obese with MetS;
- No correlation with any other genes chosen;
- Average expression level of over 50; and
- An expression level range of less than 1,000 within the respective cohorts.

The remaining six genes were chosen if they were common in at least three other gene expression panels curated by NanoString, including obesity, mammalian target of rapamycin kinase (mTOR), insulin signalling and T2DM.

Faecal samples were also collected, and gut microbial compositional sequencing was undertaken via 16S ribosomal ribonucleic acid (rRNA) sequencing and taxonomic classification. To allow a deeper analysis into the gut microbes involved in MetS, the analysis of gut microbes was conducted at the species level. To reduce computational cost, species types were removed from the analysis if they were not detected in at least 80% of participants. The data collected from the samples were separated into four different groups of variables for ease of analysis: anthropometric measure, haematological measure, gene expression level and gut microbial composition.

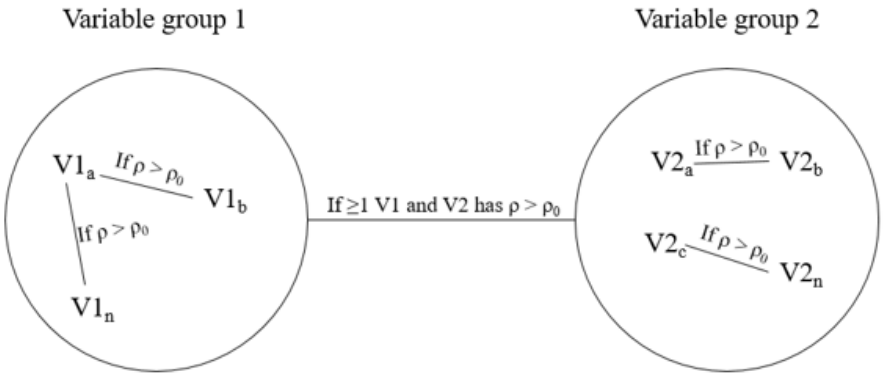
3.3.3 Univariate analysis

A descriptive analysis based on anthropometric and haematological measures for the two cohorts being studied was performed using a Mann-Whitney U test. The choice of using the Mann-Whitney U test, which is a nonparametric alternative to the independent t-test, was due to the non-normal distribution of the majority of measurements. The results of the univariate analysis were presented as median (interquartile range [IQR]) and differences were considered significant if the p-value was less than 0.05.

3.3.4 Correlation-based network analysis

Correlation-based networks were constructed for each of the four variable groups to identify any underlying connections between biomarkers that may not have been revealed through

univariate analysis. The gene expression level and gut microbial data were not normally distributed and as such, Spearman's correlation was used to calculate the relationship between biomarkers. Spearman's correlation is a nonparametric statistic that measures the strength of monotonic relationship between biomarkers. A perfect relationship between two biomarkers is represented by a correlation coefficient (ρ) of |±1| and |±0.7| often suggests a strong correlation. The strength of relationships between biomarkers in the anthropometric and haematological measures calculated by Spearman's correlation were compared to that of its parametric counterpart, Pearson's correlation. There were no significant differences found between the results and thus Spearman's correlation was used for the analysis of all four variable groups. The choice of utilising Spearman's correlation for all variable groups was to keep the consistency, particularly when analysing the relationship between biomarkers across different variable groups. The Spearman's correlation coefficient threshold (ρ0) was set at |±0.7|. Two correlated biomarkers with a correlation coefficient greater than the threshold were considered to have a strong correlation, visually represented by a link connecting the two nodes (Figure 3.1). If strong correlations were found between biomarkers of different variable groups, the correlation was denoted by a single line connecting the two variable groups involved, regardless of the total number of correlations found (Figure 3.1). To avoid cluster and allow ease of interpretability, biomarkers without strong correlations with any other biomarker were not included in the CNA. Due to the small sample size, the Spearman's correlation coefficient threshold was set very high in place of using p-values to define significance. A complete case correlation analysis was conducted, meaning that biomarkers with missing data were excluded from the network analysis. The networks were built using the psych package in R (R Development Core Team, R Foundation for Statistical Computing, Vienna, Austria).



If two biomarkers within a variable group has a ρ value greater than the initial ρ_0 threshold specified, the two biomarkers will be connected by a line. Biomarkers without correlations with another biomarker, or with ρ values smaller than ρ_0, will not appear in the network. If one or more biomarker from one variable group is correlated with one or more biomarker from another variable group with a ρ value greater than ρ_0, a single line will connect the two variable groups, regardless of the actual number of correlations present.

Figure 3.1. Example of a multi-analyte network constructed in the current study [6].

All the variables involved in the correlation analysis were continuous variables. Node degree and betweenness centrality were calculated for each node in the correlation network and network density for each variable group. The node degree is the number of strong correlations a particular node has with other biomarkers. Different node sizes in the network visually demonstrate the degree of each node, with a bigger-sized node representing a greater node degree. Betweenness centrality scores describe the number of shortest paths between any two biomarkers that passes through the node in question. Nodes with higher betweenness centrality scores were more well-connected within the network and were therefore considered to be drivers of the network. As each variable group had different numbers of nodes, there was difficulty in comparing betweenness centrality scores across variable groups. Instead, the essentiality of nodes in a network was determined by a high node degree and high ranking of betweenness centrality score within their respective variable groups. Lastly, the network density, which is the ratio of existing connections to the total number of possible connections in a network, was calculated for each correlation network constructed for each of the four

variable groups. High network density represents a high number of correlations present within a network.

3.4 Results

3.4.1 Descriptive analysis

The univariate analysis on each of the four variable groups (anthropometric measures, haematological measures, gene expression levels and gut microbial composition) was completed separately due to the differences in sample sizes. As the data used for this study was collected from different studies completed by the same research group, some participants did not have certain data measured. For the anthropometric measurements, there were significant differences between the obese with MetS and healthy weight control groups for all variables except for height ($p = 0.411$) (Table 3.1). The differences in these variables were expected between the two studied groups except age. The significant difference in age was therefore accounted for in any of the analyses that followed.

Table 3.1. Anthropometric and haematological measures in healthy weight (n = 117) and obese with MetS (n = 35) individuals.

Variable	Healthy weight (n = 117) Median (IQR)	Obese with MetS (n = 35) Median (IQR)	P-value
Anthropometric Measures			
Age (years)	36 (20)	47 (14)	**<0.001**
Height	172 (15)	171.5 (10.75)	0.411
Weight	67.2 (14.3)	103.2 (21.12)	**<0.001**
BMI (kg/m^2)	22.9 (2.3)	34.5 (6)	**<0.001**
Muscle mass (%)	31.7 (9.12)	25.2 (4.32)	**<0.001**
Fat mass (%)	25.2 (12.55)	35.9 (12.9)	**<0.001**
Fat mass (kg)	17.38 (6.66)	37.69 (14.89)	**<0.001**
FMH (kg/m^2)	5.58 (2.8)	13.27 (5.49)	**<0.001**
Visceral fat	5 (3)	13 (5)	**<0.001**
RMR	1475 (310.5)	1891 (250)	**<0.001**
Waist (cm)	79 (10)	108.5 (15.5)	**<0.001**
Hip (cm)	98 (8)	115.7 (14.75)	**<0.001**
WHR	0.79 (0.09)	0.94 (0.1)	**<0.001**
SBP (mmHg)	118 (14)	136.5 (14.25)	**<0.001**
DBP (mmHg)	76 (9.5)	90 (9.5)	**<0.001**
Haematological measures			
HbA1c (%)	5.2 (0.4)	5.5 (0.4)	**<0.001**
HG (g/L)	139 (16)	143 (19.25)	**0.025**
RCC ($x10^{12}/L$)	4.7 (0.8)	5 (0.7)	**0.002**
HCT	0.42 (0.04)	0.44 (0.06)	**0.011**
PLT ($x10^9/L$)	236 (82.25)	269 (95.75)	**0.028**
WCC ($x10^9/L$)	5.5 (2.2)	6.55 (2.35)	**<0.001**
Neutrophil ($x10^9/L$)	2.8 (1.3)	3.5 (1.15)	**0.007**
Lymphocyte ($x10^9/L$)	1.8 (0.8)	2.4 (1.08)	**<0.001**
Monocyte ($x10^9/L$)	0.4 (0.2)	0.5 (0.2)	**0.025**
Eosinophil ($x10^9/L$)	0.12 (0.16)	0.18 (0.17)	**0.040**
Basophil ($x10^9/L$)	0.05 (0.02)	0.06 (0.02)	**0.035**
ESR (mm/hr)	4 (4)	6 (9.5)	**0.008**
CRP (mg/L)	0.65 (1.45)	2.04 (2.5)	**<0.001**
FPG (mmol/L)	4.8 (0.54)	5.37 (0.97)	**<0.001**
Cholesterol (mmol/L)	4.97 (1.1)	5.4 (1.5)	**0.005**
TG (mmol/L)	0.8 (0.4)	1.88 (0.85)	**<0.001**
HDL-C (mmol/L)	1.58 (0.49)	1.12 (0.24)	**<0.001**
LDL-C (mmol/L)	2.71 (0.96)	3.2 (1.17)	**0.001**

BMI: body mass index; CRP: C-reactive protein; DBP: diastolic blood pressure; ESR: erythrocyte sedimentation rate; FMH: fat mass-to-height ratio; FPG: fasting blood glucose; HbA1c: haemoglobin A1c; HCT: haematocrit; HDL-C: high-density lipoprotein cholesterol; HG: haemoglobin; LDL-C: low-density lipoprotein cholesterol; PLT: platelet; RCC: red blood cell count; RMR: resting metabolic syndrome; SBP: systolic blood pressure; TG: triglycerides; WHR: waist-to-hip ratio

By design of the study, the obese with MetS group had significantly higher body mass index (BMI) ($p < 0.001$), waist circumference (WC) ($p < 0.001$) and blood pressure ($p < 0.001$ for both systolic blood pressure [SBP] and diastolic blood pressure [DBP]) in the anthropometric group. The remaining MetS risk factors were compared using haematological measures along with other metabolic and inflammatory biomarkers measured from fasted blood (Table 3.1). As expected, HbA1c ($p < 0.001$), fasting blood glucose ($p < 0.001$), cholesterol ($p = 0.005$), triglycerides ($p < 0.001$) and low-density lipoprotein cholesterol (LDL-C) ($p = 0.001$) were all significantly higher in the obese with MetS group compared to the healthy weight control group. Additionally, high-density lipoprotein cholesterol (HDL-C) ($p < 0.001$) was significantly lower in the obese with MetS group, which was also an expected outcome. The two major markers of inflammation, CRP and ESR, were also found to be significantly higher in the obese with MetS group compared to the healthy weight control group ($p < 0.001$ and $p = 0.008$, respectively).

The expression levels of 48 different genes were measures for each participant, with 3 housekeeping genes: RPLP0, G6PD and ABCF1. The metabolic effects of the remaining 45 genes, according to current literature, have been compiled in Table 3.2.

Table 3.2. Metabolic effects of genes chosen for comparison between healthy weight and obese with MetS groups.

Gene	Gene name	Metabolic effect	Model
AKT1	AKT Serine/Threonine Kinase 1	A: glucose and lipid metabolism; IA: insulin resistance	Human [7]
AKT3	AKT Serine/Threonine Kinase 3	A: glucose and lipid metabolism; IA: insulin resistance	Human [7]
ATG7	Autophagy Related 7	A: insulin sensitivity; IA: obesity	Mice adipose tissue [8]
CAMP	Cathelicidin Antimicrobial Peptide	A: inflammatory markers, T2DM; IA: HDL-C	Human [9]
CCL3	C-C Motif Chemokine Ligand 3	A: obesity, inflammation, insulin resistance	Human adipose tissue [10]
CD163	CD163 Molecule	A: obesity-related comorbidities (diabetes, NAFLD, MetS)	Human [11]
CD1C	CD1c Molecule	A: HOMA-IR	Human adipose tissue [12]
CD68	CD68 Molecule	A: obesity, inflammation	Human adipose tissue [13]
CD84	CD84 Molecule	A: proliferation of immune cells	Human [14]
CDH1	Cadherin 1	A: obesity, metabolic syndrome, endometrial cancer	Mice [15]
CEACAM3	CEA Cell Adhesion Molecule 3	A: obesity, insulin metabolism in liver	Human [16]
CSF3R	Colony Stimulating Factor 3 Receptor	A: TLR4 and TLR9 activation, oxidative stress	Human adipose tissue [17]
CXCL5	C-X-C Motif Chemokine Ligand 5	A: obesity, hyperglycaemia, impaired islet function	Mice [18]
CXCR6	C-X-C Motif Receptor Ligand 6	A: inflammation, NAFLD, atherosclerosis	Mice [19]
FCAR	Fc Fragment of IgA Receptor	A: inflammation	Human [20]
FCER2	Fc Fragment of IgE Receptor II	A: obesity, asthma, inflammation	Human PBMC [21]
FPR1	Formyl Peptide Receptor 1	A: high-fat feeding, glucose intolerance	Mice [22]
GZMH	Granzyme H	A: diabetes, cytolysis, apoptosis	Human PBMC [23]
GZMM	Granzyme M	A: NK function, apoptosis	Human [24]
HMGB1	High Mobility Group Box 1	A: inflammation, apoptosis	Human adipose tissue [25]
IFIT1	Interferon Induced Protein with Tetratricopeptide Repeats 1	IA: obesity, insulin resistance	Human adipose tissue [26]
IL11RA	Interleukin 11 Receptor Subunit Alpha	A: inflammation, hyperglycaemia, apoptosis	Mice [27]
IL1RN	Interleukin 1 Receptor Antagonist	A: obesity, hypertension	Human adipose tissue [28]
INSR	Insulin Receptor	IA: obesity, diabetes	Human [29]
IRF7	Interferon Regulatory Factor 7	A: obesity, inflammation, T2DM, metabolic abnormalities	Mice [30]
IRS1	Insulin Receptor Substrate 1	IA: hyperinsulinemia, inflammation	Human [31]

ITGAE	Integrin Subunit Alpha E	A: diabetes	Mice [32]
KLRC2	Killer Cell Lectin Like Receptor C2	IA: obesity, inflammation	Human [33]
LCN2	Lipocalin 2	A: obesity, inflammation	Human [34]
LTF	Lactotransferrin	A: HDL-C, insulin signalling; IA: obesity, fasting plasma glucose	Human [35]
MAPK1	Mitogen-Activated Protein Kinase 1	A: inflammation	Human whole blood [36]
MTOR	Mechanistic Target of Rapamycin Kinase	A: obesity, diabetes, apoptosis, autophagy	Human [37]
NRF1	Nuclear Respiratory Factor 1	A: inflammation, lipogenic gene expression, metabolic syndrome	Mice [38]
PIK3CA	Phosphatidylinositol-4,5-Bisphosphate 3-Kinase Catalytic Subunit Alpha	A: insulin signalling	Human [39]
PIK3R1	Phosphoinositide-3-Kinase Regulatory Subunit 1	A: insulin signalling, (mutations – insulin resistance)	Human [40]
PYCARD	PYD And CARD Domain Containing	A: obesity, inflammation	Human adipose tissue [41]
RPS6KA1	Ribosomal Protein S6 Kinase A1	A: fasting plasma glucose, diabetes	Human [42]
S100A12	S100 Calcium Binding Protein A12	A: inflammation, CVD	Human whole blood [43]
SERPING1	Serpin Family G Member 1	IA: inflammation, T2DM	Human whole blood [44]
SLC2A4	Solute Carrier Family 2 Member 4	IA: obesity, insulin resistance, inflammation	Mice [45]
SOCS1	Suppressor of Cytokine Signalling 1	A: insulin resistance, obesity, NAFLD	Human whole blood [46]
TFEB	Transcription Factor EB	A: autophagy; IA: obesity, metabolic disorders	Mice [47]
TNFSF13	TNF Superfamily Member 13	A: inflammation, insulin resistance	Human [48]
TSC1	TSC Complex Subunit 1	IA: cell hypertrophy, hyperinsulinemia	Mice [49]
ULK1	UNC-51 Like Autophagy Activating Kinase 1	IA: obesity	Mice [50]

Following the analysis of gene expression levels of each gene, there were two genes, insulin receptor substrate 1 (IRS1) and solute carrier family 2 member 4 (SLC2A4), which were found to have levels that were equivalent to background noise and were therefore excluded from further analysis.

The final variable group was gut microbial composition, which was analysed at the species level. There was a total of 450 gut microbial species detected across all the participants. After removing the species that were found in less than 80% of the participants, there were 51 species remaining. A list of the remaining gut microbial species, and the phylum to which they belong to, that were used for analysis is shown in Appendix 3.1. The three types of phylum that were prevalent within participants included Bacteroidetes, Firmicutes and Proteobacteria.

3.4.2 Correlation-based networks

For each of the four variable groups, correlation-based networks were constructed for both the healthy weight (Figure 3.2) and obese with MetS (Figure 3.3) groups. Correlations were built for each variable group individually and is visually represented by the different coloured circles. Each node shown represented a biomarker that had a strong correlation with another biomarker within the same variable group, denoted by a link between the two nodes. If biomarkers were correlated across different variable groups, this was shown by a single line connecting the variable groups involved, regardless of the number of correlations found. Overall, the obese with MetS group produced a denser network compared with the healthy network, with 232 and 134 total number of edges found, respectively.

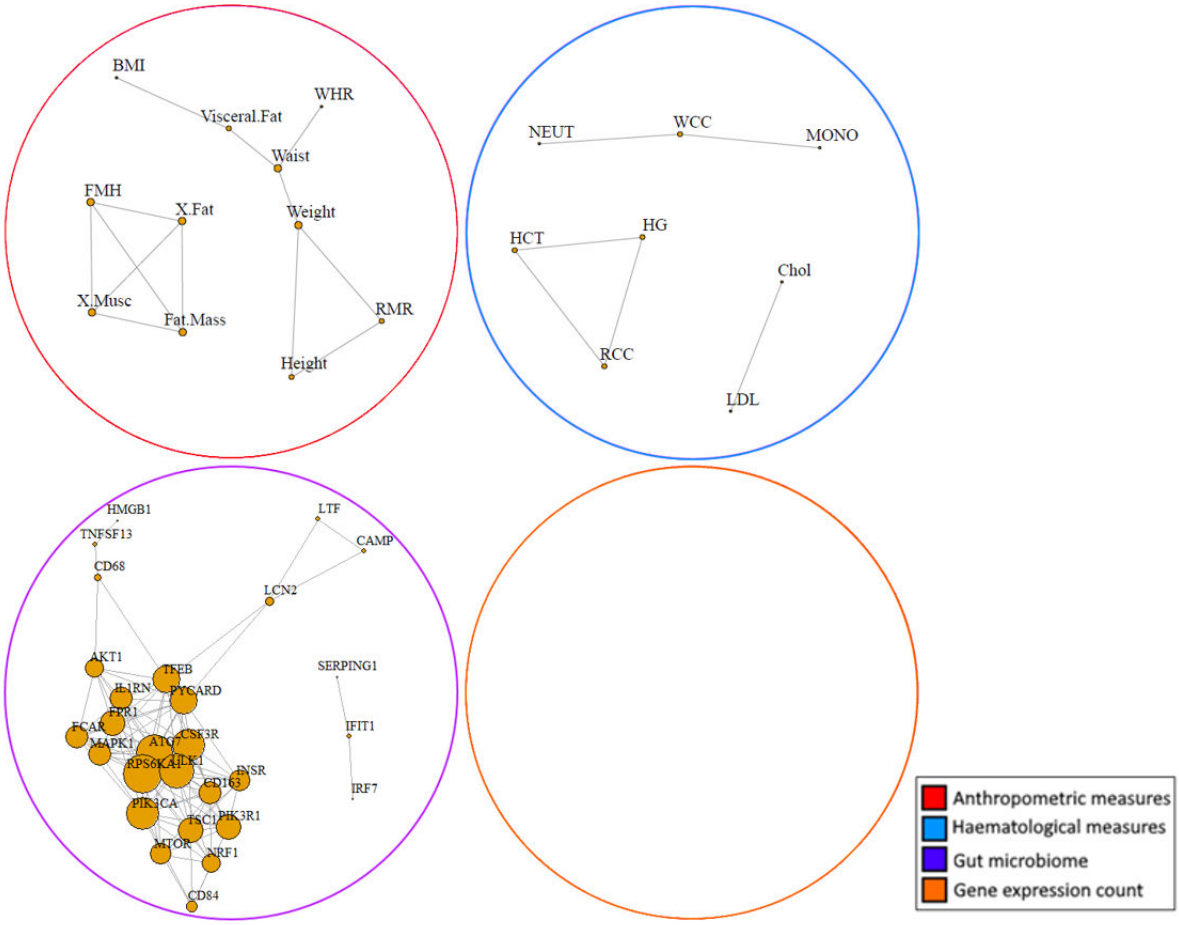


Figure 3.2. Multi-level correlation-based network built using measurements from healthy weight individuals.

AKT1: AKT serine/threonine kinase 1; ATG7: autophagy related 7; BMI: body mass index; CAMP: cathelicidin antimicrobial peptide; Chol: cholesterol; CSF3R: colony stimulating factor 3 receptor; FCAR: Fc fragment of IgA receptor; FMH: fat mass-to-height ratio; FPR1: formyl peptide receptor 1; HCT: haematocrit; HG: haemoglobin; HMGB1: high mobility group box 1; IFIT1: interferon induced protein with tetratricopeptide repeats 1; IL1RN: interleukin-1 receptor antagonist; INSR: insulin receptor; IRF7: interferon regulatory factor 7; LCN2: lipocalin 2; LDL: low-density liproptein; LTF: lactotransferrin; MAPK1: mitogen-activated protein kinase 1; MONO: monocyte; MTOR: mechanistic target of rapamycin kinase; NEUT: neutrophil; NRF1: nuclear respiratory factor 1; PIK3CA: phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha; PIK3R1: phosphoinositide-3-kinase regulatory subunit 1; PYCARD: PYD and CARD domain containing; RCC: red blood cell count; RMR: resting metabolic rate; RPS6KA1: ribosomal protein S6 kinase A1; SERPING1: Serpin family G member 1; TFEB: transcription factor EB; TNFSF13: TNF superfamily member 13; TSC1: TSC complex subunit 1; ULK1: UNC-51 like autophagy activating kinase 1; WCC: white cell count; WHR: waist-to-hip ratio; X.Fat: percentage fat mass; X.Musc: percentage muscle mass

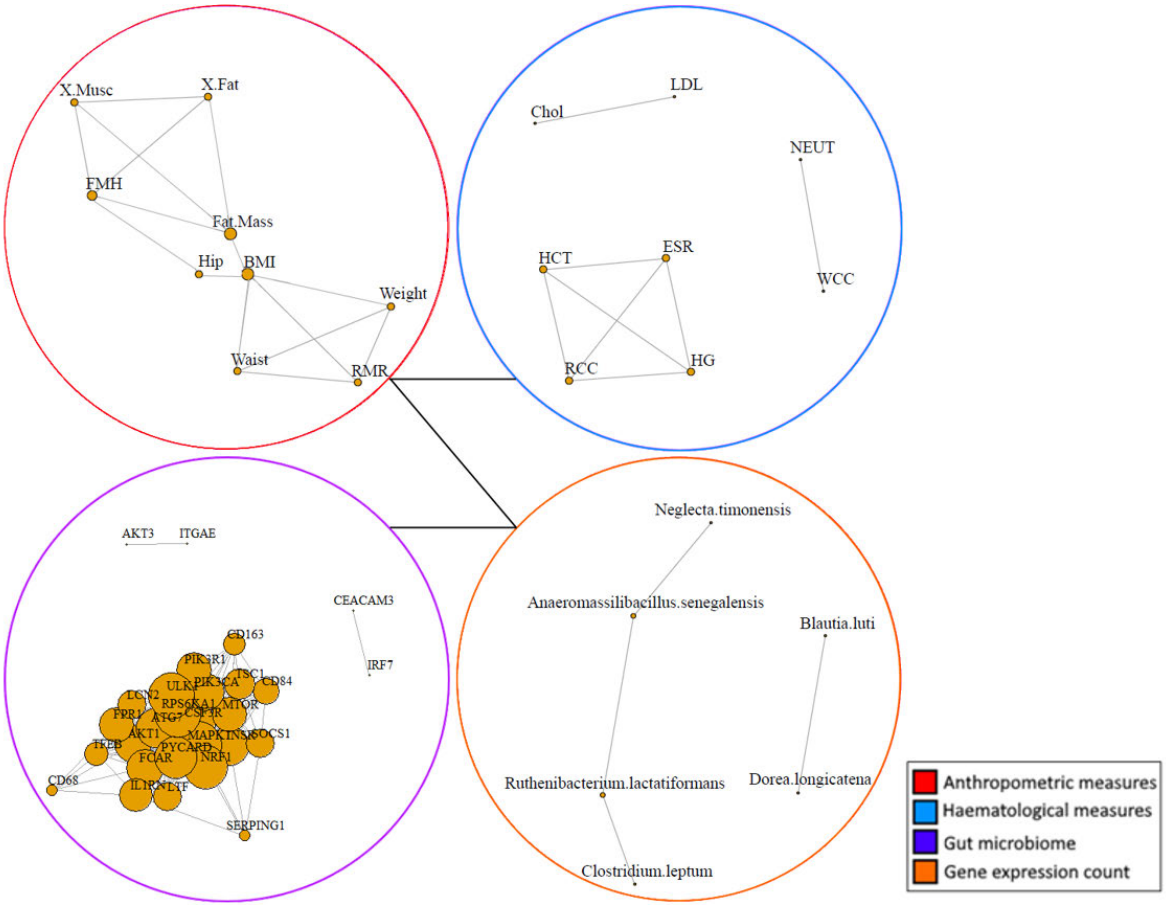


Figure 3.3. Multi-level correlation-based network built using measurements from obese with MetS individuals.

AKT1: AKT serine/threonine kinase 1; AKT3: AKT serine/threonine kinase 3; ATG7: autophagy related 7; BMI: body mass index; CEACAM3: CEA cell adhesion molecule 3; Chol: cholesterol; CSF3R: colony stimulating factor 3 receptor; ESR: erythrocyte sedimentation rate; FCAR: Fc fragment of IgA receptor; FMH: fat mass-to-height ratio; FPR1: formyl peptide receptor 1; HCT: haematocrit; HG: haemoglobin; IL1RN: interleukin-1 receptor antagonist; INSR: insulin receptor; IRF7: interferon regulatory factor 7; ITGAE: integrin subunit alpha E; LCN2: lipocalin 2; LDL: low-density liproptein; LTF: lactotransferrin; MAPK1: mitogen-activated protein kinase 1; MTOR: mechanistic target of rapamycin kinase; NEUT: neutrophil; NRF1: nuclear respiratory factor 1; PIK3CA: phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha; PIK3R1: phosphoinositide-3-kinase regulatory subunit 1; PYCARD: PYD and CARD domain containing; RCC: red blood cell count; RMR: resting metabolic rate; RPS6KA1: ribosomal protein S6 kinase A1; SERPING1: Serpin family G member 1; SOSC1: suppressor of cytokine signalling 1; TFEB: transcription factor EB; TSC1: TSC complex subunit 1; ULK1: UNC-51 like autophagy activating kinase 1; WCC: white cell count; X.Fat: percentage fat mass; X.Musc: percentage muscle mass

In the anthropometric variable group, the healthy weight control network found 13 correlations while the obese with MetS group found 16. The network density was then calculated to be 0.12 and 0.15, respectively. Betweenness centrality scores are a measure of the number of shortest paths between two biomarkers that passes through the biomarker in question. A high betweenness centrality score therefore suggests the biomarker to be a central node, connecting

different hubs of biomarkers. Targeting biomarkers with high betweenness centrality scores may then be considered to cause the biggest change on a network. As there is no guideline for a big measure in betweenness centrality, the current study considered a betweenness centrality score of over 0.1 as significant. For the anthropometric measures variable group, there were 3 biomarkers in the healthy weight control network that had a betweenness centrality score of 0.1. These variables were: WC (betweenness centrality [BC] = 0.24), weight (BC = 0.18) and visceral fat (BC = 0.11). On the other hand, the obese with MetS network only found two variables with a BC score above 0.1: BMI (BC = 0.54) and fat mass (BC = 0.39). Although the obese with MetS network found fewer biomarkers with a BC score above 0.1, the BC scores were much higher than that of the healthy weight network.

The haematological measures variable group on the other hand, saw no biomarkers with high betweenness centrality scores in both networks. However, all the correlations found in the healthy weight network were also present in the obese with MetS network, with the exception of the correlation found between monocytes and white cell count. In addition, there were 3 other correlations in the obese with MetS network, all of which involved ESR which was not found to be correlated with any biomarkers in the healthy weight network. ESR was found to be negatively correlated with haemoglobin measures (correlation coefficient [ρ] = -0.85), red blood cell count (ρ = -0.80) and haematocrit measures (ρ = -0.85).

For gene expression levels, the obese with MetS network found 374 correlations between biomarkers while the healthy weight network only found 240. While the obese with MetS network was more dense, the healthy weight network found three biomarkers with a BC score higher than 0.1 while the obese with MetS network did not have any. The three biomarkers in the healthy weight network were: transcription factor EB (TFEB) (BC = 0.19), lipocalin 2 (LCN2) (BC = 0.13) and cluster of differentiation- (CD-)68 (BC = 0.13). However, TFEB was positively correlated with both LCN2 (ρ = 0.70) and CD68 (ρ = 0.72). In addition, there were

12 other biomarkers that were correlated with TFEB, LCN2 or CD68. Out of these 12 biomarkers, there were 10 biomarkers that were positively correlated with TFEB: AKT serine/threonine kinase 1 (AKT1) ($\rho = 0.82$), autophagy related 7 (ATG7) ($\rho = 0.78$), CD163 ($\rho = 0.71$), colony stimulating factor 3 receptor (CSF3R) ($\rho = 0.72$), formyl peptide receptor 1 (FPR1) ($\rho = 0.85$), interleukin-1 receptor antagonist (IL1RN) ($\rho = 0.81$), mitogen-activated protein kinase 1 (MAPK1) ($\rho = 0.75$), PYD and CARD domain containing (PYCARD) ($\rho = 0.86$), ribosomal protein S6 kinase A1 (RPS6KA1) ($\rho = 0.78$) and UNC-51 like autophagy activating kinase 1 (ULK1) ($\rho = 0.76$). Correlations involving TFEB, LCN2 and CD68 were also found in the obese with MetS network, with LCN2 and CD68 having a higher vertex degree in the obese with MetS network compared to the healthy weight network.

The healthy weight network found no correlations between biomarkers in the gut microbial composition variable group. Conversely, the obese with MetS network found eight correlations involving six different gut microbial species. Each of the six species belonged to the Firmicutes phylum and the correlations were all positive. There were also correlations between biomarkers in the gut microbial composition variable group with biomarkers from both anthropometric measures and gene expression levels in the obese with MetS network. *C. comes* was positively correlated with CD163 ($\rho = 0.75$), insulin receptor (INSR) ($\rho = 0.75$), RPS6KA1 ($\rho = 0.75$), phosphoinositide-3-kinase regulatory subunit 1 (PIK3R1) ($\rho = 0.71$) and TSC complex subunit 1 (TSC1) ($\rho = 0.72$) while *B. faecis* was negatively correlated with MAPK1 ($\rho = -0.80$), PYCARD ($\rho = -0.73$) and RPS6KA1 ($\rho = -0.71$). In addition, *B. thetaiotaomicron* was positively correlated with visceral fat measures ($\rho = 0.82$). Some anthropometric measures from the obese with MetS group were also correlated with haematological measures. Percentage muscle mass was positively correlated with haemoglobin ($\rho = 0.77$), haematocrit ($\rho = 0.75$) and red cell count ($\rho = 0.71$) while negatively correlate with ESR ($\rho = -0.72$). On the other hand, percentage fat mass and fat mass-to-height were both negatively correlated with

haemoglobin (ρ = -0.83 and ρ = -0.77) and haematocrit (ρ = -0.81 and ρ = -0.75). The healthy weight network found no correlation between biomarkers across different variable groups.

3.5 Discussion

Metabolic syndrome is a collection of risk factors that increases the risk of developing chronic diseases. One of the risk factors is obesity, a multifactorial disease that causes the dysregulation of many different systems of the body. Due to the complexity of the human body, univariate analyses are unable to generate biomarker profiles that characterise individuals more at risk of obesity-related diseases. Integrated networks are therefore necessary to understand the intricate interactions between systems and potentially identify central biomarkers that may be targeted in future intervention studies. A study conducted with preliminary data collected from 11 obese with MetS and 12 healthy weight men has been published and can be found in Appendix 3.2. The current study builds upon the previous study by creating correlation-based networks using anthropometric measures, haematological measures, gene expression levels and gut microbial counts for comparison between 117 healthy weight and 35 obese with MetS individuals. The correlations of biomarkers within each variable group as well as between different variable groups may reveal key hubs and central biomarkers that exacerbate obesity and the risk of developing chronic diseases.

By design of the study, the obese with MetS group had significantly higher measures of risk factors that constitute metabolic syndrome, including lipids, cholesterol, blood pressure and fasting plasma glucose. Unexpectedly, the obese with MetS group was significantly older than the healthy weight group ($p < 0.001$); however, there were no correlations found between age and any other biomarkers measured.

Consistent with the results attained with preliminary data, the obese with MetS network was found to be denser than the healthy weight network, with more correlations found between biomarkers. In addition, there were correlations found between biomarkers across different variable groups in the obese with MetS network which were absent in the healthy weight network. Biomarkers from the anthropometric measures were correlated with biomarkers from both the haematological measures and gut microbiome variable groups. There were also correlations between biomarkers from the gut microbial composition with gene expression levels. The large number of correlations between biomarkers both within and across different variable groups demonstrates the complexity of the obese with MetS network. The network complexity confirms the paradigm that obesity and MetS dysregulates different systems of the body simultaneously and thus multivariate analysis is required to better understand the biomarkers involved.

In correlation networks, there are different properties that can be used to describe the relationships between biomarkers. One of these properties is betweenness centrality, which is the measure of the shortest paths between any two nodes that passes through the node in question. A node with a high betweenness centrality would then be the connecting node between many other nodes and would therefore be considered to be a central node. In the anthropometric measures variable group, the central nodes in the obese with MetS cohort were BMI and fat mass. As both these variables are measures of obesity, the centrality of these two nodes within the obese with MetS network came as no surprise.

There were, however, no central nodes that were found in haematological measures, both for the healthy weight network and obese with MetS network. Instead, the obese with MetS network found 8 correlations between biomarkers, 5 of which were also found in the healthy weight network. The 3 addition correlations all involved ESR, suggesting the important role that ESR plays in the development of MetS. The 3 correlations that were found were negative

associations with haemoglobin measures, red blood cell count and haematocrit measures. Reduced levels of haemoglobin and haematocrit, along with red blood cell count directly, are all signs of anaemia. As all three measurements were found to be negatively correlated with ESR, a common measurement of inflammation, the obese with MetS individuals may be subjected to anaemia of inflammation. Obesity is often described as a state of chronic inflammation, with associations to other inflammatory conditions, including anaemia of inflammation [51]. The negative correlation has therefore been able to confirm this finding which would not otherwise have been identified with univariate analysis, reinforcing the importance of multivariate analysis. Furthermore, there was a positive correlation between white blood cells counts and monocytes in the healthy weight network which was not found in the obese with MetS network. Although white blood cells, particularly monocytes, are an indication of pro-inflammation, both measurements were significantly lower in the healthy weight group compared to the obese with MetS group. The finding therefore suggests that the healthy weight group was subject to a lower pro-inflammatory profile compared to the obese with MetS group.

In the gene expression variable group, the healthy weight network found three biomarkers, TFEB, LCN2 and CD68 which had a high betweenness centrality score of above 0.1. However, the two latter genes were both found to be positively correlated with TFEB. At the same time, the 10 of the 12 remaining biomarkers that were found to be correlated with at least one of the three biomarkers were all positively correlated with TFEB. The centrality of LCN2 and CD68 therefore become much less credible. TFEB encodes a transcription factor that regulates the expression of genes involved in autophagy, lipid metabolism, oxidative stress and inflammation [52]. The genes that were found to be positively correlated with TFEB all fit under the categories:

- Autophagy: AKT1, ATG7, ULK1;
- Lipid metabolism: CD163, FPR1, IL1RN, RPS6KA1;
- Oxidative stress: CSF3R; and
- Inflammation: MAPK1, PYCARD.

As TFEB regulates many different pathways that prevents the development of obesity and MetS, its positive correlation with the healthy weight network found in the current study supports the findings of previous literature. In the obese with MetS network, many of the biomarkers that were correlated with TFEB were also found to be correlated with both LCN2 and CD68. Therefore, instead of recognising the genes as central nodes, together with the other biomarkers, they formed a network hub, as evident in Figure 3.3. Overall, the correlation network has demonstrated its ability to reveal relationships between biomarkers which in turn provides an insight into the roles of each biomarker as well as its interactions with other biomarkers.

The most significant finding in the previously published study, which had used preliminary data, was the key hub identified in the obese with MetS network. The key hub involved regulatory T cells, neutrophils and cytotoxic cell frequency. In the current study, the key hubs identified in the gene expression variable group involved TFEB, LCN2 and CD68. As TFEB expression is crucial in regulatory T cell differentiation [53] and neutrophils have been found to express both LCN2 [34] and CD68 [54], the results of the current study therefore support the findings using preliminary data.

There were no correlations between gut microbes in the healthy weight network while the obese with MetS network found positive correlations between six gut microbial species, all of which belonged to the Firmicutes phylum: *B. luti*, *D. longicatena*, *A. senegalensis*, *N. timonensis*, *C. leptum* and *R. lactatiformans*. Out of the six gut microbial species, only two have been reported

in other obesity studies: *B. luti* and *C. leptum*. *B. luti* has been found to be associated with metabolic inflammation and the development of insulin resistance [55] while *C. leptum* was found to be less abundant in obese individuals [56]. Very few studies have specifically compared the differences between obese with MetS and healthy weight individuals of gut microbiomes, particularly on the species level. Although many obesity studies have described obese individuals as having a higher Firmicutes abundance compared to lean individuals [57, 58], there have also been studies that did not support this finding [59]. The differences in results may lie in the bacterial species that were examined and thus it is important to utilise analytical tools such as correlation networks to better understand the relationship between microbial species in diseases. There were also correlations between gut microbes and anthropometric measures, with *B. thetaiotaomicron* found to be positively correlated with visceral fat. As previous literature has suggested the role of *B. thetaiotaomicron* in protection against obesity in mice [60], the increased abundance in response to high measures of visceral fat reported by the current study does not support previous findings.

Finally, there were correlations between biomarkers from the anthropometric and haematological measures variable groups in the obese with MetS network that supported the pathophysiology behind anaemia of inflammation. Percentage fat mass and fat mass-to-height ratio were both negatively correlated with haemoglobin and haematocrit. The finding was supported by the positive correlation between percentage muscle mass with haemoglobin, haematocrit and red cell count as well as negative correlations with ESR. Individuals with high percentage muscle mass would not be affected by the chronic inflammation associated with obesity and are therefore less likely to develop anaemia of inflammation. The complexity of the interactions shown through the correlation of biomarkers across different variable groups demonstrates the need for multivariate analysis when analysing diseases with a multifactorial nature.

While univariate analysis may be able to provide a broad comparison between the cohorts being studied, this information alone is not enough to narrow down on the biomarkers to be focussed on for future research looking to reduce the incidence of obesity and MetS. The results of univariate analysis are also not enough to neither confirm nor deny the findings of previous literature. Conversely, multivariate analysis through correlation-based network was able to provide a snapshot of the differences between the body systems of healthy weight and obese with MetS individuals. Through these networks, the molecular interactions between biomarkers are revealed which were able to demonstrate the involvement of different body systems in the development of obesity and MetS. The gene expression network in particular was able to identify key hubs and confirm the relationship between TFEB and genes associated with autophagy, lipid metabolism, oxidation and inflammation. Additionally, correlations with haematological biomarkers suggested the presence of anaemia of inflammation in obese with MetS individuals. Lastly, while many studies have found high Firmicutes-to-Bacteroidetes ratios in obese with MetS individuals, other studies have reported conflicting results. The correlation networks found that while some species belonging to the Firmicutes phylum were positively correlated with other markers of obesity and MetS, others were not. Future studies should therefore carefully consider the species that is used for comparison between obese with MetS and healthy weight individuals. With these findings, researchers are able to narrow down and create a biomarker profile that is able to classify individuals more at risk of MetS and related diseases. Correlation-network analysis is able to provide detailed information on the body systems affected by obesity without a high computational cost and is also highly interpretable by researchers. The use of CNA should therefore be considered in all areas of research in conjunction with other more complex analytical tools.

The limitations of this study were recognised, including the differences in sample sizes that was used in each variable group, the type of measurements included for analysis and the effects

of ethnical diversity among participants. With a larger sample size, more correlations may have been revealed. To account for the issue in sample sizes, the current study utilised a high correlation coefficient threshold in place of p-values to define significant results. Additionally, the measurements used for analysis were taken from peripheral blood as opposed to biopsies from adipose tissue. As majority of obesity research uses measurements from adipose tissue, the results of this study could not be compared as easily to previous studies. Furthermore, the current study did not consider participant ethnicity during sample collection, which may skew the measurements taken. While majority of the participants were Caucasian, the exact percentage as well as the number of participants from other ethnical backgrounds were not reported. Despite the limitations of the study, there were still many results that were attained that were consistent with previous literature which were not affected by these limitations.

3.6 Appendices

Appendix 3.1. List of the gut microbial species, and the phylum to which they belong, that were used for correlation-based network analysis.

Phylum	Species
Firmicutes	Agathobaculum butyriciproducens
Bacteroidetes	Alistipes onderdonkii
Bacteroidetes	Alistipes putredinis
Firmicutes	Anaerostipes hadrus
Bacteroidetes	Bacteroides stercoris
Bacteroidetes	Bacteroides uniformis
Bacteroidetes	Bacteroides vulgatus
Firmicutes	Blautia luti
Firmicutes	Blautia wexlerae
Firmicutes	Clostridium clostridioforme
Firmicutes	Clostridium leptum
Firmicutes	Clostridium methylpentosum
Firmicutes	Clostridium spiroforme
Firmicutes	Coprococcus catus
Firmicutes	Coprococcus comes
Proteobacteria	Desulfovibrio simplex
Firmicutes	Eubacterium coprostanoligenes
Firmicutes	Eubacterium eligens
Firmicutes	Eubacterium rectale
Firmicutes	Faecalibacterium prausnitzii
Firmicutes	Hespellia porcina
Firmicutes	Intestinimonas butyriciproducens
Firmicutes	Lachnoclostridium pacaense
Firmicutes	Murimonas intestini
Firmicutes	Neglecta timonensis
Firmicutes	Oscillibacter ruminantium
Bacteroidetes	Parabacteroides merdae
Firmicutes	Pseudoflavonifractor phocaeensis
Firmicutes	Romboutsia timonensis
Firmicutes	Ruminococcus bromii
Firmicutes	Ruminococcus torques
Firmicutes	Ruthenibacterium lactatiformans

Appendix 3.2. Previous publication with preliminary data.

Statement of contribution to co-authored published paper

A co-authored published paper was included in this chapter. The bibliographic details of the co-authored published paper, including all authors, are:

Chen, P. Y., Cripps, A. W., West, N. P., Cox, A. J., & Zhang, P. (2019). A correlation-based network for biomarker discovery in obesity with metabolic syndrome. BMC bioinformatics, 20 (Suppl 6), 477. https://doi.org/10.1186/s12859-019-3064-2

My contribution to the publish paper involved:

drafting of the concept and design of the study together with each of the listed co-authors; recruitment of participants and collection of data with my co-authors; processing the data and writing the source codes required for data analysis; reviewing the interpretations for the data analysis with my co-author; drafting the manuscript with my co-authors revising the drafts and approving the final manuscript.

Signed:

Pin-Yen Chen PhD Candidate School of Medical Science Griffith University	Pin-Yen Chen Corresponding author School of Medical Science Griffith University	Prof. Allan Cripps Primary Supervisor School of Medicine Griffith University
Date:		
13/08/2020	13/08/2020	12/08/2020

Chen et al. BMC Bioinformatics 2019, 20(Suppl 6):477
https://doi.org/10.1186/s12859-019-3064-2

BMC Bioinformatics

RESEARCH Open Access

A correlation-based network for biomarker discovery in obesity with metabolic syndrome

Pin-Yen Chen[1,2*], Allan W. Cripps[1,3], Nicholas P. West[1,2], Amanda J. Cox[2] and Ping Zhang[1]

From 2nd International Workshop on Computational Methods for the Immune System Function
Madrid, Spain. 3-6 December 2018

Abstract

Background: Obesity is associated with chronic activation of the immune system and an altered gut microbiome, leading to increased risk of chronic disease development. As yet, no biomarker profile has been found to distinguish individuals at greater risk of obesity-related disease. The aim of this study was to explore a correlation-based network approach to identify existing patterns of immune-microbiome interactions in obesity.

Results: The current study performed correlation-based network analysis on five different datasets obtained from 11 obese with metabolic syndrome (MetS) and 12 healthy weight men. These datasets included: anthropometric measures, metabolic measures, immune cell abundance, serum cytokine concentration, and gut microbial composition. The obese with MetS group had a denser network (total number of edges, $n = 369$) compared to the healthy network ($n = 299$). Within the obese with MetS network, biomarkers from the immune cell abundance group was found to be correlated to biomarkers from all four other datasets. Conversely in the healthy network, immune cell abundance was only correlated with serum cytokine concentration and gut microbial composition. These observations suggest high involvement of immune cells in obese with MetS individuals. There were also three key hubs found among immune cells in the obese with MetS networks involving regulatory T cells, neutrophil and cytotoxic cell abundance. No hubs were present in the healthy network.

Conclusion: These results suggest a more complex interaction of inflammatory markers in obesity, with high connectivity of immune cells in the obese with MetS network compared to the healthy network. Three key hubs were identified in the obese with MetS network, involving Treg, neutrophils and cytotoxic cell abundance. Compared to a t-test, the network approach offered more meaningful results when comparing obese with MetS and healthy weight individuals, demonstrating its superiority in exploratory analysis.

Keywords: Network analysis, Obesity, Metabolic syndrome, Inflammation, Immune system, Gut microbiome, Multidimensional data, Multivariate analysis

* Correspondence: pin-yenfiona.chen@griffithuni.edu.au
[1]Menzies Health Institute Queensland, Griffith University, Gold Coast, Australia
[2]School of Medical Science, Griffith University, Gold Coast, Australia
Full list of author information is available at the end of the article

Background

Obesity is a multifactorial disease that dysregulates many different body systems, including the immune system [1] and the gut microbiota [2], leading to increased risk of chronic diseases, including type 2 diabetes mellitus (T2DM), some cancers, and increased mental health problems [1]. Despite extensive research, no specific biomarker profile is clinically recognised to characterise individuals with a greater risk of developing obesity-related disease [3]. A key reason may be failure to consider the interconnected nature of the immune system, host microbiota and metabolic interactions. Many functional studies have now recognised the need for integrated analysis to overcome the issue of redundancy [4, 5], whereby many biomarkers have similar roles, rendering univariate analysis ineffective. Recent technological advances that allow for multiple biomarker analysis are overcoming the limitations associated with biological complexity to better understand the basis of diseases. However, the interpretation and visualisation of the significant amount of data generated from these methods still poses a challenge [6].

Correlation-based network analysis (CNA) has recently become a popular data-mining method as it allows complexity reduction of multidimensional data while still retaining the majority of information needed for interpretation [6]. CNA provides the means to visualise disease-related perturbations of molecular interactions to provide insight into key underlying mechanisms that drive disease development [7]. In biological network analysis, biomarkers are represented as nodes and the links between them as edges. A number of network properties have been developed to allow interpretation of correlation networks [6], including (a) node degree: the number of other nodes to which a given node is significantly correlated, (b) betweenness centrality: the measure of shortest paths between any two nodes that passes through the node in question, and (c) network density: the ratio of existing edges to the total number of possible edges in a network. Using these properties, researchers can also detect highly connected nodes, also known as hubs. These properties were useful in many obesity studies which used CNA to identify key hubs that differ between obese and healthy weight individuals. Walley et al. used a network approach to compare genes in subcutaneous adipose tissue of obesity-discordant siblings [8]. The study found a third of the transcripts to be differentially expressed between lean and obese siblings, with obesity-associated neuronal growth regulator 1 (NEGR1) acting as a central hub. A later study by Wang et al. [9] also used network analysis to identify significant genes between seven discordant monozygotic twins. From this study, at least eight different hub genes were identified. Both Walley et al. and Wang et al. were able to detect central genes affected by obesity, providing insight into future research looking to target specific biomarkers for obesity treatment. However, these studies are limited by their focus on specific areas of the body. Considering obesity being a multifactorial disease and the functional interdependencies of different systems of the human body, a multi-analyte network should be utilised instead.

Studies examining immune profiles [1] and gut microbial composition [10] in obese individuals have found alterations in favour of pro-inflammatory biomarkers when compared to their lean counterparts. A study by Winer et al. has also found high pro-inflammatory to anti-inflammatory biomarker ratios in obese individuals that exacerbate chronic disease development [11]. However, studies have still struggled to find a profile of biomarkers that distinguish individuals more at risk of obesity-related diseases. Due to the multitude of molecular interactions affected by obesity, a holistic approach is required to identify key biomarkers involved. The aim of this study was to use CNA to compare anthropometric measures, metabolic measures, immune cell abundance, serum cytokine concentrations, and gut microbial composition to identify biomarker profiles that distinguish obese with metabolic syndrome (MetS) from healthy weight individuals.

Results

The molecular interactions associated with obesity were analysed by comparing networks within obese with MetS and healthy weight individuals through CNA. The characteristics of participants from the two distinct groups are described in Table 1. Significant differences were observed in all the key demographic measures, except for age, between the two groups. By design, the obese with MetS group had values outside the healthy range for variables that constitute the criteria for MetS [12]. Two major markers of inflammation, CRP and ESR, were also compared between the two groups, both of which were higher in the obese with MetS group although the difference was not significant for ESR. As obesity has been described as a state of chronic low-grade inflammation [3], the higher levels of inflammatory markers observed in the obese with MetS group was expected. The findings from the exploratory univariate analysis justified the use of other analytical methods to find possible underlying interactions between inflammatory biomarkers.

A multi-level correlation network was built for the two studied groups (Figs. 1 and 2). In the CNA, each node represented a biomarker that had a strong correlation with another biomarker in the same variable group, denoted by a link between the two nodes based on a Pearson correlation analysis. Correlations between biomarkers from different variable groups was visually represented by a single line connecting the two variable groups involved. The obese with MetS group produced a

Table 1 Demographic characteristics and metabolic measures in obese with MetS ($N = 11$) and healthy weight ($N = 12$) males

	Obese with Mets ($n = 11$)	Healthy weight ($n = 12$)	*P*-value*
Demographic variables			
Age (Years)	47.74 ± 8.52	40.98 ± 12.36	0.1
BMI (kg/m2)	35.25 ± 3.57	23.05 ± 1.30	< 0.001
Waist (cm)	177.82 ± 10.31	82.71 ± 5.03	< 0.001
Fat Mass (%)	34.2 ± 2.30	20.48 ± 2.52	< 0.001
Muscle Mass (%)	26.3 ± 2.09	36.27 ± 2.87	< 0.001
Visceral Fat	16.64 ± 3.53	5.75 ± 1.45	< 0.001
Metabolic variables			
MetS	3.55 ± 0.69	0.17 ± 0.39	< 0.001
SBP (mmHg)	144.55 ± 13.37	122 ± 4.78	< 0.001
DBP (mmHg)	96.91 ± 9.98	76.58 ± 6.49	< 0.001
Triglycerides (mmol/L)	2.18 ± 0.50	1.10 ± 0.62	< 0.001
Cholesterol (mmol/L)	5.58 ± 1.01	5.08 ± 1.15	0.24
HDL (mmol/L)	1.13 ± 0.18	1.54 ± 0.34	< 0.001
LDL (mmol/L)	3.46 ± 0.87	3.03 ± 0.88	0.22
HbA1c (%)	5.36 ± 0.43	5.23 ± 0.26	0.42
Glucose (mmol/L)	5.74 ± 0.71	5.20 ± 0.33	0.04
CRP (mg/L)	1.77 ± 0.86	0.95 ± 1.04	0.01
ESR (mm/hr)	6.18 ± 4.62	3.58 ± 0.90	0.27

MetS: scored out of a maximum of five based on presence of five defined metabolic syndrome features
BMI Body mass index, *BP* Blood pressure, *MetS* Metabolic syndrome, *HDL* High-density lipoprotein, *LDL* Low-density lipoprotein, *HbA1c* Haemoglobin A1c, *CRP* C-reactive protein, *ESR* Erythrocyte sedimentation rate
**P* value is based on an unpaired t-test using log-transformed data

much denser network compared to the healthy network, with the total number of edges being 369 and 299, respectively. In addition, the obese with MetS network found correlations between biomarkers within each variable group as well as between each of the five variable groups (Fig. 1). Interestingly, immune cells within the healthy network were not found to be correlated with two of the four other biomarker groups: anthropometric measures and metabolic measures (Fig. 2). The high interconnectivity of immune cells in the obese with MetS network compared to the healthy network suggests immune cells to be highly involved in obesity.

The lack of interaction in the healthy network between three variable groups was compared to correlations within the obese with MetS network. While the healthy network found no correlation between biomarkers in the immune cell abundance group and the anthropometric measures group, the obese with MetS network found age to be negatively correlated with Th2 cell abundance (correlation coefficient [ρ] = − 0.74). Furthermore, there was no correlation between biomarkers from the immune cell abundance group with the metabolic measures group. On the other hand, the obese with MetS network found correlations between systolic blood pressure and mast cell abundance (ρ = 0.71), absolute lymphocyte count and macrophage abundance (ρ = − 0.73), absolute lymphocyte count and neutrophil abundance (ρ = − 0.73), and high-density lipoprotein with T-helper cell abundance (ρ = − 0.78).

The increased involvement of immune cells as obesity develops is also supported by the large number of correlations between immune cell biomarkers in the obese with MetS network compared to the healthy network. The obese with MetS network had higher numbers of correlations, higher network density and more biomarkers with high betweenness centrality scores. As betweenness centrality measures the shortest paths between two nodes that passes through the particular node, it signifies how central the biomarker is within a network. A biomarker with a high betweenness centrality score is therefore considered to be a hub in a network that will cause the biggest change on a network if targeted. The obese with MetS network saw 11 correlations between biomarkers, a network density of 0.09 and 3 biomarkers with a high betweenness centrality score of over 0.1 (Table 2). On the other hand, the healthy network only had 7 correlations between biomarkers, a network density of 0.06 and no biomarkers with a high betweenness centrality score (Table 2). The three key hubs of the obese with MetS network stem from the biomarkers: Treg cells (betweenness centrality [BC] =

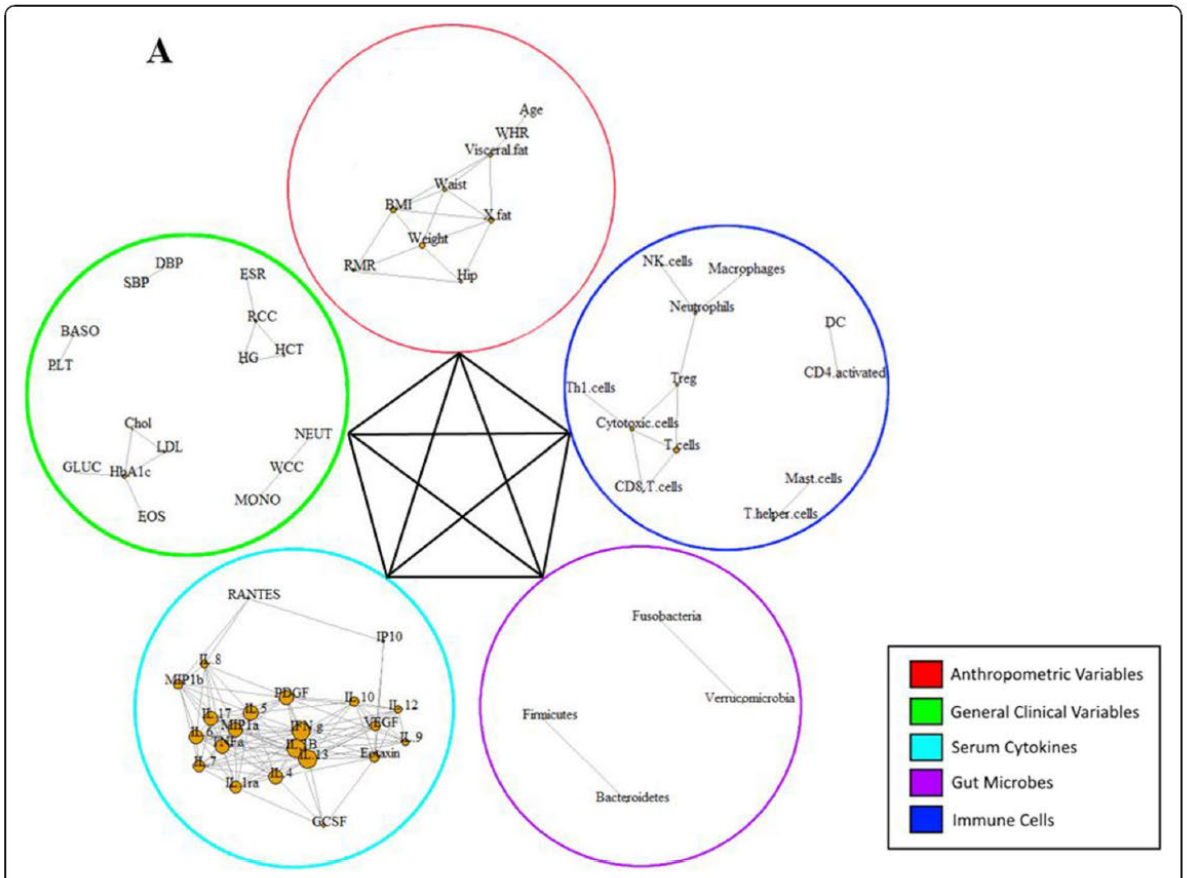


Fig. 1 Multi-level CNA constructed for obese with MetS participants. AUSDRISK: Australian type 2 diabetes risk; WHR: waist-hip ratio; BMI: body mass index; X.fat: percentage fat mass; X.musc: percentage muscle mass; RMR: resting metabolic rate; SBP: systolic blood pressure; DBP: diastolic blood pressure; MetS: metabolic syndrome; Chol: cholesterol; LDL: low-density lipoprotein; HDL: high-density lipoprotein; HCT: haematocrit; RCC: red cell count; ESR: erythrocyte sedimentation rate; PLT: platelet; BASO: basophil; CRP: C-reactive protein; LYMPHO: lymphocyte; NK cells: natural killer cells; DC: dendritic cells; Treg: T-regulatory cells; Th1 cells: T-helper 1 cells; VEGF: vascular endothelial growth factor; IL-: interleukin; IP10: interferon gamma-induced protein 10; PDGF: platelet-derived growth factor; IFN.g: interferon gamma; TNFa: tumour necrosis factor alpha; GCSF: granulocyte-colony stimulating factor; MIP1a: macrophage inflammatory protein 1 alpha; MIP1b: macrophage inflammatory protein 1 beta

0.22), neutrophils (BC = 0.20) and cytotoxic cells (BC = 0.15).

Within the immune cell abundance variable group, Treg cell abundance was correlated with neutrophil abundance (ρ = 0.73), cytotoxic cell abundance (ρ = – 0.73) and T cell abundance (ρ = – 0.74); neutrophil abundance was correlated with macrophage abundance (ρ = 0.80) and NK cell abundance (ρ = 0.74), and cytotoxic cell abundance was correlated with Th1 cell abundance (ρ = 0.78), T cell abundance (ρ = 0.77) and $CD8^+$ T cell abundance (ρ = 0.74). Additionally, Treg cell abundance was correlated with MIP-1β concentration (ρ = 0.71) from the serum cytokine group while neutrophil abundance was correlated with a number of biomarkers from the gut microbial group, including: *Escherichia/Shigella* abundance (ρ = 0.74), *Akkermansia* abundance (ρ = 0.71), *Anaerostipes* abundance (ρ = 0.72), *Blautia* abundance (ρ = 0.78), *Flavonifractor* abundance (ρ = 0.73), and *Holdemania* abundance (ρ = 0.70). These correlations may be considered for intervention studies looking to reduce the prevalence of obesity-related diseases.

An unpaired t-test was also performed on the same dataset (Table 3) and the results were compared to that of the CNA. For the immune cell abundance variable group, the only significant difference found between the healthy weight and obese with MetS groups was in mast cell ($P = 0.02$) and T-helper cell abundance ($P = 0.04$). Mast cell abundance was negatively correlated with T-helper cell abundance, with no correlations with any other immune cells for either of the two biomarkers. Compared to the t-test, the CNA was able to reveal more detailed information on the differences

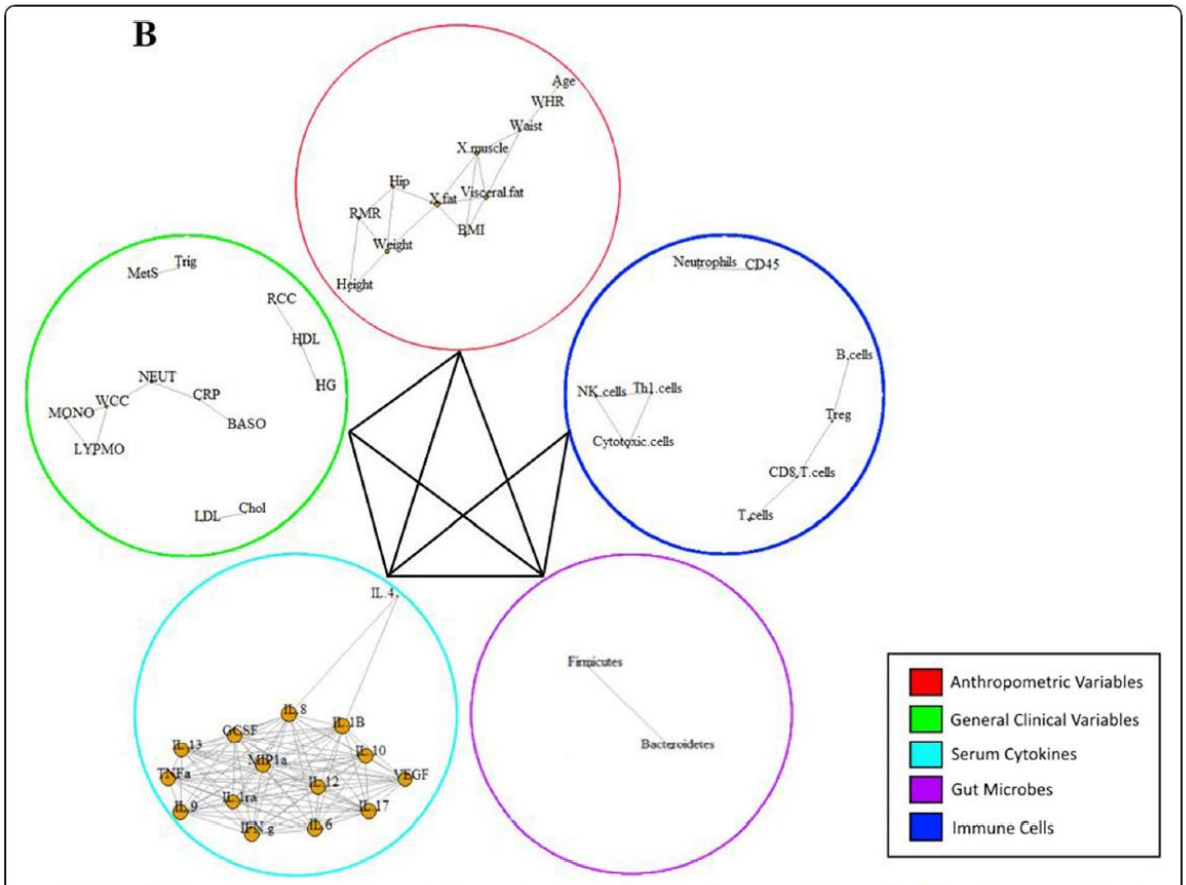


Fig. 2 Multi-level CNA constructed for healthy weight (**b**) participants. AUSDRISK: Australian type 2 diabetes risk; WHR: waist-hip ratio; BMI: body mass index; X.fat: percentage fat mass; X.musc: percentage muscle mass; RMR: resting metabolic rate; SBP: systolic blood pressure; DBP: diastolic blood pressure; MetS: metabolic syndrome; Chol: cholesterol; LDL: low-density lipoprotein; HDL: high-density lipoprotein; HCT: haematocrit; RCC: red cell count; ESR: erythrocyte sedimentation rate; PLT: platelet; BASO: basophil; CRP: C-reactive protein; LYMPHO: lymphocyte; NK cells: natural killer cells; DC: dendritic cells; Treg: T-regulatory cells; Th1 cells: T-helper 1 cells; VEGF: vascular endothelial growth factor; IL-: interleukin; IP10: interferon gamma-induced protein 10; PDGF: platelet-derived growth factor; IFN.g: interferon gamma; TNFa: tumour necrosis factor alpha; GCSF: granulocyte-colony stimulating factor; MIP1a: macrophage inflammatory protein 1 alpha; MIP1b: macrophage inflammatory protein 1 beta

between obese with MetS and healthy weight individuals, demonstrating the importance of using multivariate analysis rather than univariate.

Discussion

Many systems of the body have been reported in the literature as being dysregulated in obesity and subsequently increasing the risk of chronic disease development. Due to the complexity of the human body, integrated networks are necessary to better understand the intricate interactions between biomarkers involved in obesity-related diseases. CNA was performed on various datasets obtained from 11 obese men with MetS and 12 healthy weight men. Datasets included were: anthropometric measures, metabolic measures, immune cell abundance, serum cytokine concentrations, and gut microbial composition. Until recently, functional studies in obesity have had conflicting outcomes due to the issue of redundancy and functional interdependencies between biomarkers across different body systems. The aim of this study was to compare the networks constructed for the two studied groups and identify key biomarker interactions that may characterise obesity and related diseases.

Table 2 Main properties of the obese with MetS (Fig. 1) and healthy weight networks (Fig. 2)

Network	Total number of edges	Network density	Number of hubs
Obese with MetS	11	0.09	3
Healthy weight	7	0.06	0

Table 3 Immune cell abundance measures in obese with MetS ($N = 11$) and healthy weight ($N = 12$) males

Immune cells	Obese with Mets ($n = 11$)	Healthy weight ($n = 12$)	*P*-value*
Mast cells	3.55 ± 0.68	4.22 ± 0.65	0.02
sNK cells	7.34 ± 0.29	7.26 ± 0.43	0.56
CD8 T cells	6.98 ± 0.53	7.07 ± 0.41	0.64
DC	2.02 ± 0.55	1.7 ± 0.63	0.24
Treg	3.8 ± 0.34	3.89 ± 0.5	0.72
CD45	12.44 ± 0.26	12.33 ± 0.24	0.31
Macrophages	8.89 ± 0.25	8.84 ± 0.23	0.62
T cells	8.88 ± 0.17	8.93 ± 0.18	0.49
Neutrophils	11.01 ± 0.37	10.96 ± 0.33	0.73
Cytotoxic cells	9.34 ± 0.57	9.3 ± 0.51	0.89
Th1 cells	5.49 ± 0.49	5.37 ± 0.66	0.56
Normal mucosa	3.36 ± 0.47	3.25 ± 0.39	0.60
T-helper cells	8.22 ± 0.14	8.32 ± 0.07	0.04
B cells	7.04 ± 0.7	7.12 ± 0.69	0.78
Th2 cells	3.36 ± 0.34	3.9 ± 0.96	0.11
CD4 activated	2.24 ± 0.59	2.1 ± 0.53	0.61

*P value is based on an unpaired t-test using log-transformed data

When comparing the networks constructed for each group, the obese with MetS group had a denser overall network than the healthy weight group. The differences in the number of correlations suggest the obese with MetS network displayed a more complex connectivity compared to the healthy weight group. The concept of a more complex network confirms the paradigm that obesity is associated with an alteration of multiple parameters across a broad range of biological systems. The interconnected nature of different body systems calls for the need to utilise integrated analytical approaches to deconstruct the complexity of the biological dysregulation in obesity. Through this approach, biomarkers that may be central for investigation in future studies may be identified. The correlation network analysis used in this study supports the use of cluster-based analysis to better understand obesity-related diseases.

In the obese with MetS network, biomarkers of each individual variable group were found to be correlated with other biomarkers from their own group as well as other variable groups. On the other hand, immune cell biomarkers in the healthy weight network were not shown to be correlated with biomarkers from two other variable groups: anthropometric measures and metabolic measures. The contrast between correlations in the obese with MetS and healthy weight networks suggest immune cells to be heavily perturbed in obesity. Both human and animal studies have reported on obesity-related changes in the immune cell abundance and activity which were linked with the development of chronic diseases [13–16]. The similarity in findings between the current study and previous literature suggests CNA to be a reliable analytical method which can be used in studies looking at diseases with complex aetiology.

Further comparisons between the two networks in relation to immune cell abundance revealed more correlations in the obese with MetS network compared to the healthy weight network, with 11 and 7 correlations, respectively. Within the correlations in the obese with MetS network, there were three biomarkers with high betweenness centrality scores. Betweenness centrality is a measure of the number of shortest paths between two other biomarkers that passes through the biomarker in question. A high betweenness centrality score would therefore suggest the biomarker to be the centre of a key hub within the network. The three central biomarkers were: Treg cell abundance, neutrophil abundance and cytotoxic T cell abundance. The correlations found in our study that constitute these hubs have shown positive correlations between pro-inflammatory biomarkers, such as between neutrophils and macrophages, and negative correlations between pro-inflammatory and anti-inflammatory biomarkers, including Treg cells and cytotoxic cells. These correlations are consistent with the findings from earlier studies which have reported a dysregulation in the immune system of obese individuals, resulting in a high pro-inflammatory-to-anti-inflammatory biomarker ratio [11]. All biomarkers have connections with a number of other biomarkers and therefore the recognition of key hubs is crucial in identifying biomarker profiles that characterise obesity-related diseases.

While correlation networks are particularly useful in discovering correlations between biomarkers and key hubs

of a system, unpaired t-tests reveal very little in comparison. Performed on the same immune cell abundance data, an unpaired t-test between the obese with MetS and healthy weight group only observed significant differences in mast cell and T-helper cell abundances. Both mast cell and T-helper cell abundances were higher in the healthy weight group. In a study by Liu et al., mast cells contributed to obesity by producing pro-inflammatory cytokines [14]. Therefore, mast cell abundance is expected to be higher in the obese with MetS group, inconsistent with the findings from the current study. Additionally, neither mast cell nor T-helper cell abundance were present in any of the three key hubs found in the obese with MetS network, suggesting the findings from the t-test to be uninformative. The clear distinction between the results of the correlation network and t-test is attributable to the inability of linear causality models to account for the complexity of human body systems.

Using correlation networks, the current study also found many interesting relationships, such as a positive correlation between pro-inflammatory neutrophils and anti-inflammatory Tregs. As obese individuals typically have a high pro-inflammatory-to-anti-inflammatory ratio, this finding was unexpected. A possible explanation for this relationship is suggested in a study by Mishalian et al. who observed the ability of neutrophils to recruit Tregs, exacerbating the impairment of the immune system in disease [17]. Without the use of CNA, a finding that is pertinent in better understanding this multi-factorial disease would be missed in a simple t-test. Relationships between biomarkers, such as neutrophils and Tregs, are important in intervention research which may consider targeting both biomarkers for an exacerbated effect.

Both Treg and neutrophil abundances were also correlated with biomarkers outside of the immune cell abundance variable group. Treg cell abundance was positively correlated with serum MIP-1β concentration, consistent with the findings of Patterson et al., whereby stimulated Tregs produced MIP-1β to assist with T cell migration [18]. Our study also found neutrophils to be associated with a number of gut microbes which is also consistent with earlier studies [19–22]. Neutrophil abundance was positively correlated with gut microbes belonging to neutrophil-associated microbiomes [23]: Firmicutes (*Anaerostipes*, *Blautia*, *Flavonifractor* and *Holdemania*) and Proteobacteria (*Escherichia/Shigella*) phyla. The correlations between biomarkers from different variable groups demonstrates the complexity of interactions between physiological systems and the importance of utilising multi-analyte networks when analysing diseases with complex aetiology.

The differences in results obtained in univariate and multivariate analysis highlights the biggest advantage to using CNA in high-throughput studies. Multivariate analysis allows researchers to consider underlying connections between biomarkers, both within the same or across different variable groups. A simple comparison of biomarker levels between groups does not have the ability to recognise key hubs within a network which may be targeted for future intervention studies. Multivariate analysis has the means to overcome the limitation of redundancy among biomarkers which has limited the ability of functional research to identify key biomarkers in obesity-related disease. Other advantages to using multivariate CNA includes its ease of use and interpretability. The use of correlation networks should therefore be considered for exploratory analysis, rather than unpaired t-test, prior to the use of more complex analytical tools.

The limitations of this study have also been recognised, in particular the small sample size that was used. As a pilot study, the current work was exploratory and utilised high correlation coefficient cut-offs rather than *p*-values to define important results. Another limitation is the small number of molecular markers included in the analysis. While many obesity studies examined markers within adipose tissue, the current study performed analysis on peripheral blood to examine systemic rather than peripheral immune dysregulation. Additionally, the current study did not consider the effects of participant ethnicity in genetic analysis which may result in false positive findings. However, from the known participant ethnicities, 70% were Caucasian, 0.04% were Hispanic and the remaining were unknown. Despite these limitations, the study was still able to gather a multitude of results that supports further research with larger sample sizes and datasets.

Conclusion

Our study found that obesity with MetS is associated with a more densely connected and therefore complex interaction between inflammatory, gut microbial and metabolism in comparison to that observed in healthy weight individuals. Further analysis revealed immune cells to be highly involved in obesity, with three key hubs in the obese with MetS network that consisted of Treg, neutrophils and cytotoxic cell abundance. The results from the network analysis were much more informative compared to a t-test, suggesting it to be a better choice as an exploratory analytical tool. Our findings demonstrate the need for integrated analysis of multidimensional data to identify specific and multiple interactions between biomarkers that may be targeted for treatment strategies.

Methods

Study design and ethics

A correlation-based network analysis was performed on anthropometric measures, metabolic measures, immune gene expression, serum cytokine concentrations and gut microbial composition collected from 12 healthy weight

men and 11 obese men with MetS, defined as per the Adult Treatment Panel III criteria [12] (Three or more of the following risk factors: (1) abdominal obesity: ≥30 kg/m^2 BMI or > 94 cm waist circumference; (2) high blood pressure: ≥130/≥85 mmHg; (3) high triglycerides: ≥1.7 mmol/L; (4) low HDL cholesterol: ≤1 mmol/L; (5) high fasting plasma glucose: ≥6.1 mmol/L or ≥ 6.5% HbA1c). All participants were aged between 18 and 65 years without a history of medical conditions known to affect the immune system, including: cancer, Crohn's disease, liver disease, and irritable bowel syndrome. Additionally, participants were excluded if they used any immune-modulating medications or supplements, such as: non-steroidal anti-inflammatory drugs (NSAIDs), fish oil and probiotics. Ethics for this study was approved by the Griffith University Human Research Ethics Committee (MED 18.15.HREC) and all participants provided written informed consent prior to their involvement in the study.

Sample collection and analysis

Fasting blood samples were collected for analysis of metabolic (lipids, glucose, glycated haemoglobin [HbA1c]) and inflammatory (C-reactive protein [CRP], erythrocyte sedimentation rate [ESR], circulating cytokines) measures. In addition, RNA was isolated and analysed using an immune profiling panel of 770 genes (nCounter® PanCancer Immune Profiling Panel, NanoString Technologies, Washington, USA) to estimate the abundance of different immune cells, including mast cells, neutrophils and different T cell subsets. Faecal samples were also collected and microbial compositional sequencing was undertaken via 16S rRNA sequencing and taxonomic classification.

Correlation-based network analysis

To compare key demographic measures of obese with MetS and healthy weight participants, an unpaired t-test was used and measures were expressed as mean ± standard deviation. Differences in measures were considered significant if the p-value was less than 0.05. The dataset was split into five different variable groups: anthropometric measures, metabolic measures, immune cell abundance, serum cytokine concentrations, and gut microbial composition.

Correlation networks were constructed by firstly calculating the Pearson correlation coefficient (ρ) for each biomarker with all other biomarkers in the five variable groups. A Pearson correlation coefficient threshold was set at $|\pm 0.7|$. Two biomarkers with a correlation coefficient greater than the threshold will be considered as having a strong correlation, visually represented by a link between the two nodes (Fig. 3). Biomarkers that had a strong correlation with another biomarker appeared in the CNA. Strong correlations between biomarkers of different variable groups were indicated by a single line connecting the two variable groups involved, regardless of the total number of correlations found. Nodes of biomarkers without strong correlations with any other biomarker were not included in the CNA. Due to the small sample size, the Pearson correlation coefficient threshold required for a correlation to be considered significant was set very high rather than using a significance level. A complete case correlation analysis was conducted, meaning that biomarkers with missing data were excluded from the network analysis.

All the variables involved in the correlation analysis were continuous variables. Node degree and betweenness centrality was calculated for each node in the correlation network and network density was computed for each

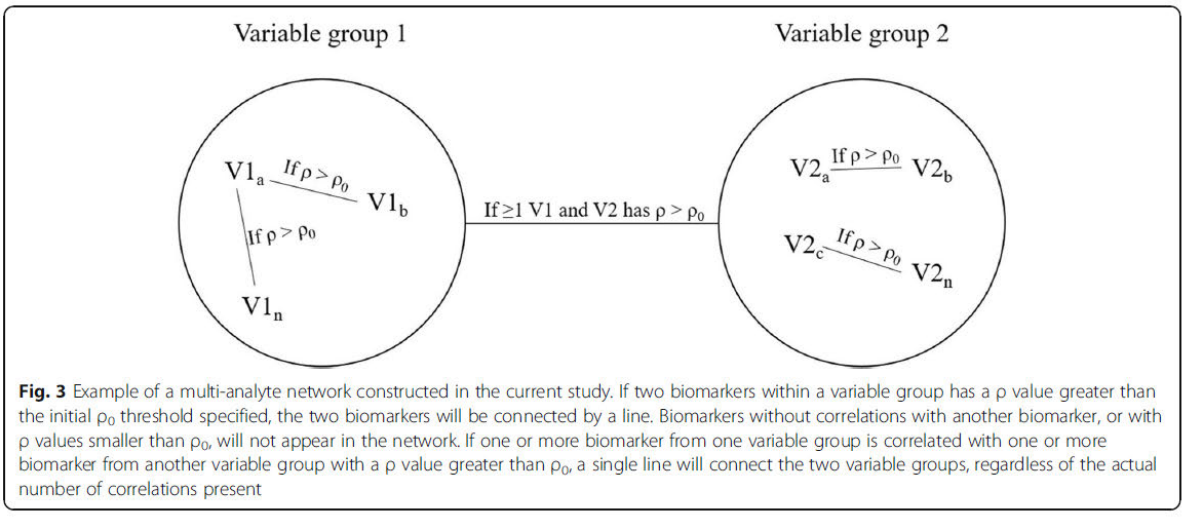


Fig. 3 Example of a multi-analyte network constructed in the current study. If two biomarkers within a variable group has a ρ value greater than the initial ρ_0 threshold specified, the two biomarkers will be connected by a line. Biomarkers without correlations with another biomarker, or with ρ values smaller than ρ_0, will not appear in the network. If one or more biomarker from one variable group is correlated with one or more biomarker from another variable group with a ρ value greater than ρ_0, a single line will connect the two variable groups, regardless of the actual number of correlations present

variable group. The node degree is the number of strong correlations a particular node has with other biomarkers. Different node sizes in the network visually demonstrate the degree of each node, with a bigger sized node representing a greater node degree. Betweenness centrality scores describe the number of shortest paths between any two biomarkers that passes through the node in question. Nodes with higher betweenness centrality scores are more well-connected within the network and therefore are considered to be drivers of the network. As each variable group has different numbers of nodes, it is difficult to compare betweenness centrality scores across variable groups. Instead, the essentiality of nodes in a network was determined by a high node degree and a high ranking of betweenness centrality score within their respective variable groups. The variable groups from the obese with MetS and healthy networks were compared based on network density, which is the ratio of existing connections to the total number of possible connections within a network. The higher the network density, the more connections there are in the network.

All the statistical analyses and network analyses were carried out with custom R (R Development Core Team, R Foundation for Statistical Computing, Vienna, Austria) scripts. To avoid computational issues that may occur with a large sample size, the code was developed into multiple modules. Each type of analysis, for example Pearson correlation coefficient calculation or network visualisation, had its own module.

Abbreviations

BC: Betweenness centrality; BMI: Body mass index; CNA: Correlation-based network analysis; CRP: C-reactive protein; ESR: Erythrocyte sedimentation rate; HbA1c: Haemoglobin A1c; HDL: High-density lipoprotein; MetS: Metabolic syndrome; MIP-1β: Macrophage inflammatory protein 1 beta; NEGR1: Neuronal growth regulator 1; NK: Natural killer; NSAID: Non-steroidal anti-inflammatory drugs; RNA: Ribonucleic acid; rRNA: Ribosomal ribonucleic acid; T2DM: Type 2 diabetes mellitus

Acknowledgments

The authors wish to acknowledge the subjects of the study for their participation. This work has been partly presented previously at the 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM).

About this supplement

This article has been published as part of BMC Bioinformatics Volume 20 Supplement 6, 2019: Towards computational modeling on immune system function. The full contents of the supplement are available online at https://bmcbioinformatics.biomedcentral.com/articles/supplements/volume-20-supplement-6.

Author's contributions

PY, PZ, AWC, NPW, and AJC contributed to the concept and design of the study. PY, NPW, and AJC performed participant recruitment and data collection. PY and PZ designed the data analysis. PY processed the data, wrote the source codes for data analysis, and drafted the manuscript. PZ assisted with interpretation of the analysis. PZ, AWC, NPW, and AJC contributed to revising drafts of the paper. PZ, AWC, and NPW provided supervision. All authors read and approved the final manuscript.

Funding

This project was supported by Griffith Health Institute/Gold Coast Hospital Foundation. Scholarship support for PYC was provided by a Griffith University Heath Group Postgraduate Research Scholarship. Salary support for PZ and AJC was provided by the Griffith University Area of Strategic Investment in Chronic Disease Prevention. The content is solely the responsibility of the authors and does not necessarily represent the official views of Griffith Health Institute/Gold Coast Hospital Foundation. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. Publication costs are funded by Menzies Health Institute Queensland, Griffith University, Australia.

Availability of data and materials

Data are available upon request from the Menzies Health Institute Queensland for researchers who meet the criteria for access to confidential data.

Ethics approval and consent to participate

Written informed consent was obtained from all participants. This study was approved by the Griffith University Human Research Ethics Committee (GU Ref No: MED/19/15/HREC).

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Menzies Health Institute Queensland, Griffith University, Gold Coast, Australia. [2]School of Medical Science, Griffith University, Gold Coast, Australia. [3]School of Medicine, Griffith University, Gold Coast, Australia.

Received: 9 July 2019 Accepted: 29 August 2019
Published: 12 December 2019

References

1. Sell H, Habich C, Eckel J. Adaptive immunity in obesity and insulin resistance. Nat Rev Endocrinol. 2012;8:709–16.
2. Bäckhed F, et al. Host-bacterial mutualism in the human intestine. Science. 2005;307(5717):1915–20.
3. Cox AJ, West NP, Cripps AW. Obesity, inflammation, and the gut microbiota. Lancet Diabetes Endocrinol. 2015;3:207–15.
4. Ideker T, et al. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. Science. 2001;292:929–34.
5. Tian Q, et al. Integrated genomic and proteomic analyses of gene expression in mammalian cells. Mol Cell Proteomics. 2004;3:960–9.
6. Batushansky A, Toubiana D, Fait A. Correlation-based network generation, visualization, and analysis as a powerful tool in biological studies: a case study in cancer cell metabolism. Biomed Res Int. 2016;2016(8313272):1–9.
7. Nishihara R, et al. Biomarker correlation network in colorectal carcinoma by tumor anatomic location. BMC Bioinf. 2017;18(304):1–14.
8. Walley AJ, et al. Differential coexpression analysis of obesity-associated networks in human subcutaneous adipose tissue. Int J Obes. 2012;36(1):137–47.
9. Wang W, et al. Weighted gene co-expression network analysis of expression data of monozygotic twins identifies specific modules and hub genes related to BMI. BMC Genomics. 2017;18(1):1–17.
10. Ley RE, et al. Microbial ecology: human gut microbes associated with obesity. Nature. 2006;444:1022–3.
11. Winer S, et al. Normalization of obesity-associated insulin resistance through immunotherapy: CD4+ T cells control glucose homeostasis. Nat Med. 2009;15(8):921–9.
12. Grundy SM, et al. Definition of metabolic syndrome. Arterioscler Thromb Vasc Biol. 2004;24(2):e13–8.
13. Fujisaka S, et al. Regulatory mechanisms for adipose tissue M1 and M2 macrophages in diet-induced obese mice. Diabetes. 2009;58(11):2574–82.
14. Liu J, et al. Genetic deficiency and pharmacological stabilization of mast cells reduce diet-induced obesity and diabetes in mice. Nat Med. 2009;15(8):940–5.

15. Bertola A, et al. Identification of adipose tissue dendritic cells correlated with obesity-associated insulin-resistance and inducing Th17 responses in mice and patients. Diabetes. 2012;61(9):2238–47.
16. Talukdar S, et al. Neutrophils mediate insulin resistance in mice fed a high-fat diet through secreted elastase. Nat Med. 2012;18:1407–12.
17. Mishalian I, et al. Neutrophils recruit regulatory T-cells into tumors via secretion of CCL17--a new mechanism of impaired antitumor immunity. Int J Cancer. 2014;135:1178–86.
18. Patterson SJ, et al. T regulatory cell chemokine production mediates pathogenic T cell attraction and suppression. J Clin Invest. 2016;126(3):1039–51.
19. Roy U, et al. Distinct microbial communities trigger colitis development upon intestinal barrier damage via innate or adaptive immune cells. Cell Rep. 2017;21(4):994–1008.
20. Larsen N, et al. Gut microbiota in human adults with type 2 diabetes differs from non-diabetic adults. PLoS One. 2010;5(2):1–10.
21. Qin J, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature. 2012;490:55–60.
22. Sen T, et al. Diet-driven microbiota dysbiosis is associated with vagal remodeling and obesity. Physiol Behav. 2017;173:305–17.
23. Li Q, et al. Identification and characterization of blood and neutrophil-associated microbiomes in patients with severe acute pancreatitis using next-generation sequencing. Front Cell Infect Microbiol. 2018;8(5):1–16.

Publisher's Note

3.7 References

[1] Ellulu, M.S., et al., Obesity and inflammation: the linking mechanism and the complications. Arch Med Sci. vol. 13, pp. 851-63, 2015.

[2] Warmbrunn, M.V., et al., Gut microbiota: a promising target against cardiometabolic diseases. Expert Rev Endocrinol Metab. vol. 15, pp. 13-27, 2020.

[3] Zhang, W., Xin, L., and Lu, Y., Integrative analysis to identify common genetic markers of metabolic syndrome, dementia, and diabetes. Med Sci Monit. vol. pp. 5885-5891, 2017.

[4] Su, L.-n., et al., Network analysis identifies common genes associated with obesity in six obesity-related diseases. J Zhejiang Univ Sci B. vol. 18, pp. 727-732, 2017.

[5] International Diabetes Federation, *The IDF consensus worldwide definition of the Metabolic Syndrome*. 2006. pp. 1-24.

[6] Chen, P.-Y., et al., A correlation-based network for biomarker discovery in obesity with metabolic syndrome. BMC Bioinformatics. vol. 20, pp. 1-10, 2019.

[7] Huang, X., Liu, G., and Su, Z., The PI3K/AKT pathway in obesity and type 2 diabetes. Int J Biol Sci. vol. 14, pp. 1483-96, 2018.

[8] Menikdiwela, K.R., et al., Autophagy in metabolic syndrome: breaking the wheel by targeting the renin-angiotensin system. Cell Death Dis. vol. 11, 2020.

[9] Li, Y.-X., Li, B.-Z., and Yan, D.-Z., Upregulated expression of human cathelicidin LL-37 in hypercholesterolemia and its relationship with serum lipid levels. Mol Cell Biochem. vol. 449, pp. 73-9, 2018.

[10] Tourniaire, F., et al., Chemokine Expression In Inflamed Adipose Tissue Is Mainly Mediated By NF-κB. PLoS One. vol. 8, 2013.

[11] Hu, T.-Y., et al., Soluble CD163-associated dietary patterns and the risk of metabolic syndrome. Nutrients. vol. 11, 2019.

[12] Cho, K.W., et al., Adipose tissue dendritic cells are independent contributors to obesity-induced inflammation and insulin resistance. J Immunol. vol. 197, pp. 3650-61, 2016.

[13] Huang, C.-W., et al., Role of n-3 polyunsaturated fatty acids in ameliorating the obesity-induced metabolic syndrome in animal models and humans. Int J Mol Sci. vol. 17, 2016.

[14] Das, S.K., Ma, L., and Sharma, N., Adipose tissue gene expression and metabolic health of obese adults. Int J Obes (Lond). vol. 39, pp. 869-73, 2015.

[15] Neff, R., et al., Functional characterization of recurrent FOXA2 mutations seen in endometrial cancers. Int J Cancer. vol. 143, pp. 2955-61, 2018.

[16] Hong, S.W., et al., Understanding the molecular aspects of oriental obesity pattern differentiation using DNA microarray. J Transl Med. vol. 13, 2015.

[17] Naruse, K., et al., Involvement of visceral adipose tissue in immunological modulation of inflammatory cascade in preeclampsia. Mediators Inflamm. vol., 2015.

[18] Nunemaker, C.S., et al., Increased serum CXCL1 and CXCL5 are linked to obesity, hyperglycemia, and impaired islet function. J Endocrinol. vol. 222, pp. 267-76, 2014.

[19] Ma, K.L., et al., Activation of the CXCL16/CXCR6 pathway promotes lipid deposition in fatty livers of apolipoprotein E knockout mice and HepG2 cells. Am J Transl Res. vol. 10, pp. 1802-16, 2018.

[20] Rajakumar, K., et al., Gene expression and cardiometabolic phenotypes of vitamin D-deficient overweight and obese black children. Nutrients. vol. 11, 2019.

[21] Rastogi, D., Suzuki, M., and Greally, J.M., Differential epigenome-wide DNA methylation patterns in childhood obesity-associated asthma. Sci Rep. vol. 3, 2013.

[22] Wollam, J., et al., Microbiota-produced N-formyl peptide fMLF promotes obesity-induced glucose intolerance. Diabetes. vol. 68, pp. 1415-26, 2019.

[23] Gao, L., et al., Comparative analysis of mRNA expression profiles in Type 1 and Type 2 diabetes mellitus. Epigenomics. vol. 11, pp. 685-99, 2019.

[24] Aho, V., et al., Partial sleep restriction activates immune response-related gene expression pathways: Experimental and epidemiological studies in humans. PLoS One. vol. 8, 2013.

[25] Zhang, J., et al., HMGB1, an innate alarmin, plays a critical role in chronic inflammation of adipose tissue in obesity. Mol Cell Endocrinol. vol. 454, pp. 103-11, 2017.

[26] Wieser, V., et al., Adipose type I interferon signalling protects against metabolic dysfunction. Gut. vol. 67, pp. 157-65, 2016.

[27] Widjaja, A.A., et al., Inhibiting interleukin 11 signaling reduces hepatocyte death and liver fibrosis, inflammation, and steatosis in mouse models of nonalcoholic steatohepatitis. Gastroenterology. vol. 157, pp. 777-92, 2019.

[28] Yang, S.-A., Exonic polymorphism (rs315952, Ser133Ser) of interleukin 1 receptor antagonist (IL1RN) is related to overweigh/obese with hypertension. J Exerc Rehabil. vol. 10, pp. 322-6, 2014.

[29] Cignarelli, A., et al., Insulin and insulin receptors in adipose tissue development. Int J Mol Sci. vol. 20, 2019.

[30] Hirotsu, Y., et al., Transcription factor NF-E2-related factor 1 impairs glucose metabolism in mice. Genes Cells. vol. 19, pp. 650-65, 2014.

[31] Guo, S., Insulin signaling, resistance, and the metabolic syndrome: Insights from mouse models to disease mechanisms. J Endocrinol. vol. 220, pp. 1-23, 2014.

[32] Barrie, E.S., et al., Role of ITGAE in the development of autoimmune diabetes in non-obese diabetic mice. J Endocrinol. vol. 224, pp. 235-43, 2015.

[33] Jung, U.J., et al., Differences in metabolic biomarkers in the blood and gene expression profiles of peripheral blood mononuclear cells among normal weight, mildly obese and moderately obese subjects. Br J Nutr. vol. 116, pp. 1022-32, 2016.

[34] Moschen, A.R., et al., Lipocalin-2: A master mediator of intestinal and metabolic inflammation. Trends Endocrinol Metab. vol. 28, pp. 388-97, 2017.

[35] Marcil, V., et al., Cardiometabolic risk factors and lactoferrin: polymorphisms and plasma levels in French-Canadian children. Pediatr Res. vol. 82, pp. 741-8, 2017.

[36] Miranda, D.N., et al., Increases in insulin sensitivity among obese youth are associated with gene expression changes in whole blood. Obesity (Silver Spring). vol. 22, pp. 1337-44, 2014.

[37] Das, A., et al., mTOR signaling in cardiometabolic disease, cancer, and aging. Oxid Med Cell Longev. vol., 2017.

[38] More, V.R., et al., Keap1 Knockdown increases markers of metabolic syndrome after long-term high fat diet feeding. Free Radic Biol Med. vol. pp. 85-94, 2013.

[39] Karczewska-Kupczewska, M., et al., Serum and adipose tissue chemerin is differentially related to insulin sensitivity. Endocr Connect. vol. 9, pp. 360-9, 2020.

[40] McCurdy, C.E. and Klemm, D.J., Adipose tissue insulin sensitivity and macrophage recruitment. Adipocyte. vol. 2, pp. 135-42, 2013.

[41] Yin, Z., et al., Transcriptome analysis of human adipocytes implicates the NOD-like receptor pathway in obesity-induced adipose inflammation. Mol Cell Endocrinol. vol. 394, pp. 80-7, 2014.

[42] Hebbar, P., et al., Genome-wide association study identifies novel risk variants from RPS6KA1, CADPS, VARS, and DHX58 for fasting plasma glucose in Arab population. Sci Rep. vol. 10, 2020.

[43] Shah, R.D., et al., Expression of calgranulin genes S100A8, S100A9 and S100A12 is modulated by n-3 PUFA during inflammation in adipose tissue and mononuclear cells. PLoS One. vol. 12, 2017.

[44] Doumatey, A.P., et al., Proinflammatory and lipid biomarkers mediate metabolically healthy obesity: A proteomics study. Obesity (Silver Spring). vol. 24, pp. 1257-65, 2016.

[45] Poletto, A.C., et al., Reduced Slc2a4/GLUT4 expression in subcutaneous adipose tissue of monosodium glutamate obese mice is recovered after atorvastatin treatment. Diabetol Metab Syndr. vol. 7, 2015.

[46] Kempinska-Podhorodecka, A., et al., The association between SOCS1−1656G>a polymorphism, insulin resistance and obesity in nonalcoholic fatty liver disease (NAFLD) patients. J Clin Med. vol. 8, 2019.

[47] Bala, S. and Szabo, G., TFEB, a master regulator of lysosome biogenesis and autophagy, is a new player in alcoholic liver disease. Dig Med Res. vol. 1, 2018.

[48] Zhang, Y., et al., QTL-based association analyses reveal novel genes influencing pleiotropy of Metabolic Syndrome (MetS). Obesity (Silver Spring). vol. 21, pp. 2099-111, 2013.

[49] Mao, Z. and Zhang, W., Role of mTOR in glucose and lipid metabolism. Int J Mol Sci. vol. 19, 2018.

[50] Sciarretta, S., Volpe, M., and Sadoshima, J., Is reactivation of autophagy a possible therapeutic solution for obesity and metabolic syndrome? Autophagy. vol. 8, pp. 1252-4, 2012.

[51] Nemeth, E. and Ganz, T., Anemia of inflammation. Hematol Oncol Clin North Am. vol. 28, pp. 671-81, 2014.

[52] Brady, O.A., Martina, J.A., and Puertollano, R., Emerging roles for TFEB in the immune response and inflammation. Autophagy. vol. 14, pp. 181-9, 2018.

[53] Bettinger, P. and Roman, C., The induction of FoxP3 in naive T cells is dependent upon the transcription factors TFE3 and TFEB. J Immunol vol. 188, pp. 163-4, 2012.

[54] Amanzada, A., et al., Identification of CD68(+) neutrophil granulocytes in in vitro model of acute inflammation and inflammatory bowel disease. Int J Clin Exp Pathol. vol. 6, pp. 561-70, 2013.

[55] Benítez-Páez, A., et al., Depletion of Blautia Species in the Microbiota of Obese Children Relates to Intestinal Inflammation and Metabolic Phenotype Worsening. mSystems. vol. 5, 2020.

[56] Clarke, S.F., et al., The gut microbiota and its relationship to diet and obesity. Gut Microbes. vol. 3, pp. 186-202, 2012.

[57] Most, J., et al., Gut microbiota composition strongly correlates to peripheral insulin sensitivity in obese men but not in women. Benef Microbes. vol. 8, pp. 557-62, 2017.

[58] Koliada, A., et al., Association between body mass index and Firmicutes/Bacteroidetes ratio in an adult Ukrainian population. BMC Microbiol. vol. 17, pp. 1-6, 2017.

[59] Yun, Y., et al., Comparative analysis of gut microbiota associated with body mass index in a large Korean cohort. BMC Microbiol. vol. 17, pp. 151, 2017.

[60] Greenhill, C., Gut microbiome and serum metabolome changes. Nat Rev Endocrinol. vol. 13, pp. 501, 2017.

CHAPTER 4
Application of classification models for the prediction of MetS

4.1 Abstract

The area of research involved in the development of disease has recently taken an interest in classification models for their ability to handle and understand high dimensional data. There are a range of classification models that have been used for the prediction of diseases, each with their own advantages and disadvantages. As metabolic syndrome (MetS) is a relatively new area of research compared to the chronic diseases it is linked with, including diabetes and cardiovascular disease (CVD), the best choice of classification model for its prediction has yet to be determined. The current study applied four different classification models for the prediction of MetS: logistic regression (LR), decision tree (DT), support vector machine (SVM) and artificial neural network (ANN). The prediction of MetS was undertaken with three variable groups measured from 152 participants, including haematological measures, gene expression levels and gut microbial counts. Overall, ANN was found to have the highest predictive ability for MetS using both haematological measures and gut microbial counts. However, it was outperformed by SVM when predicting MetS using gene expression levels. Although SVM and ANN were found to predict MetS most accurately, the results of both LR and DT were both still very high. In addition to the high classification accuracies, LR and DT also have the ability to identify key biomarkers in MetS development, making them an invaluable choice for obtaining clinically significant outcomes. In particular, the most important haematological measure that was associated with MetS development was found to be triglycerides, appearing in all LR and DT models constructed by the current study. Studies looking to obtain a clinical outcome should therefore consider the use of LR and DT over SVM

and ANN. The study has demonstrated the ability of each prediction model to achieve high predictive ability through the use of relevant biomarkers and proper optimisation.

4.2 Introduction

Classification models can be used to predict diseases in individuals through discerning hard-to-detect patterns within large data. A range of classification models exist, including logistic regression (LR), decision trees (DT), support vector machines (SVM) and artificial neural network (ANN). Each prediction model has its own strengths in dealing with particular types of data and the most appropriate prediction model to be used depends on the research question being asked. Section 2.6, page 41 provides a detailed comparison of the four aforementioned prediction models in regard to their use for the prediction of metabolic syndrome (MetS) and related diseases. The most important factor in constructing a model with high predictive ability lies in the relevancy of the variables included. For a condition such as MetS, classification models built with haematological and biochemical measurements are likely to yield a higher prediction accuracy compared to those built on data, such as education level. The choice in the type of prediction model used for research is also crucial. The most appropriate type of prediction model used not only depends on the research question being asked but also in how well the researcher understands the model being used. While LR and DT are easier to understand and interpret, SVM and ANN require researchers to understand the architecture of the model and how to tune its hyperparameters to achieve a high prediction accuracy. Support vector machines can be built using different kernel functions, each separating the data in a different way. The choice of which kernel function to use therefore has a significant impact on the performance of the model. Additionally, ANNs have a complex architecture comprised of multiple layers, each requiring its own optimisation. The tuning of its hyperparameters

therefore require researchers to have an extensive knowledge of its architecture to achieve a high predictive ability. Despite being computationally complex, both SVM and ANN have the ability to handle high-dimensional data and are often considered to be more powerful algorithms than LR and DT. On the other hand, LR and DT are both able to produce clinically significant data, namely the identification of variables that are more likely to contribute to the development of diseases. As such, both models are a popular choice for clinical research looking to identify important biomarkers that may be targeted for preventative or therapeutic intervention. Overall, the type of classification model that should be used depends on the research question being asked and how well the researcher understands the model.

4.3 Research design

4.3.1 Study design

Current literature has yet to identify which biomarkers best predict individuals to be more at risk of MetS and related diseases, or which prediction model performs the best for MetS prediction. The current study has therefore decided to compare the performance of LR, DTs, SVM and ANN in predicting MetS status using data collected from the same 152 participants in Section 3.3.1, page 74. The prediction models were built using three of the four variable groups measured in Section 3.3.2, page 75. As participants were separated into either the obese with MetS or healthy weight control groups based on their anthropometric measures, using these measurements to predict MetS was considered to be redundant. Each prediction model was used to predict MetS using each of the variable groups separately. Prior to the construction of prediction models, Spearman's correlation was used to remove strongly correlated variables for the purpose of reducing computational cost and improving the prediction accuracy of

prediction models. The methods for the correlation analysis mirrored what was used in Section 3.3.4, page 76.

4.3.2 Cross-validation

To evaluate the performance of each prediction model, 10-fold cross-validation was used. Cross-validation allows the models to be tested on unseen data that were not used to build the model, avoiding bias. The participants were separated into 10 training and 10 testing sets. Each training set was used to build a prediction model. The performance of the model was assessed by how well it correctly classified individuals as either healthy weight control ('0') or obese with MetS ('1'). The model was used to predict individuals in both the training and testing sets to calculate whether the models were either overtrained or undertrained. If the classification accuracy was at least 20% higher in the training set compared to the testing set, the model was considered to be overtrained. An overtrained model is one that has been moulded to fit the training set and thus prediction of any other data would be unreliable. On the contrary, a model that predicted the testing set with at least 10% higher accuracy than the training set was undertrained.

To create the training and testing sets, the full dataset was split with 90% of the participants in the training set and the remaining 10% in the testing set. In total, 10 different training and testing sets were created using sample without replacement, ensuring that no participant was found in more than one testing set. Figure 4.1 provides an example of 5-fold cross-validation, whereby the whole dataset is split into a training-to-testing set ratio of 5-to-1. The figure demonstrated how the data is split based on the number of folds chosen and how each participant will never appear in more than one testing set. The ratio of healthy weight controls to obese with MetS individuals was the same across each training and testing set. Each of the

10 training sets were used to construct an individual prediction model, resulting in a total of 10 models built for each type of prediction model.

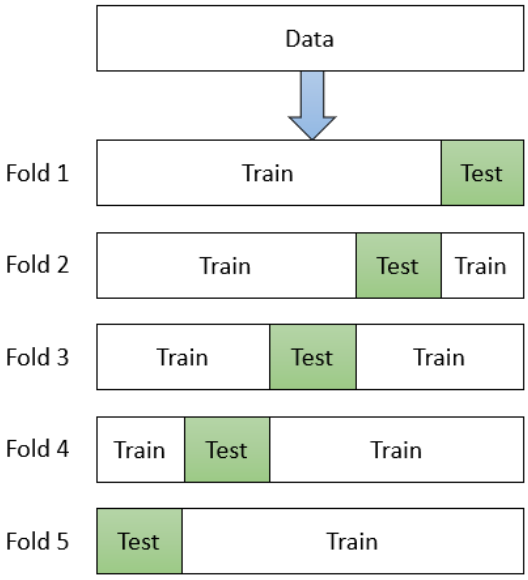


Figure 4.1. Example of k-fold cross-validation.

The overall performance of each prediction model type was then calculated by finding the average performance of the 10 models built. The performance of the model was evaluated using classification accuracy, sensitivity, specificity and AUC values. The output for each prediction model is a probability prediction between '0' and '1' for each participant. The cut-off to class the participant as either '0' or '1' was then decided to achieve the highest classification accuracy, sensitivity and specificity values. The sensitivity and specificity values were calculated using the *caret* package in R. Figure 4.2 provides a visual representation of a confusion matrix and how it can be used to calculate each statistical parameter by hand. The confusion matrix consists of four numbers that represent the performance of a model in classifying samples. The four numbers of the matrix show the true positive (TP), false positive (FP), true negative (TN) and false negative (FN).

	Predicted: Obese with MetS	Predicted: Healthy
Actual: Obese with MetS	True Positive (TP)	False Negative (FN)
Actual: Healthy	False Positive (FP)	True Negative (TN)

Figure 4.2. Example layout of confusion matrix used to calculate the different values that describe the performance of prediction models.

Using these four values, the classification accuracy, sensitivity and specificity criteria were calculated using the following formulae:

$$Classification\ accuracy = \frac{TP + TN}{TP + TN + FN + FP}$$

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{FP + TN}$$

Furthermore, using the sensitivity and specificity values at different thresholds from the output of the classification model, the receiver operating characteristic (ROC) curve can be constructed (Figure 4.3). The ROC curve is a graph that represents the performance of classification models at all classification thresholds. From the ROC curve, the area under the ROC curve, or the AUC value, can be calculated by finding the total area of the three polygons shown in Figure 4.3. The AUC represents the probability that the model will rank a randomly chosen positive instance higher than a randomly chosen negative instance. In the current study, the AUC value was calculated with the *ROC* package.

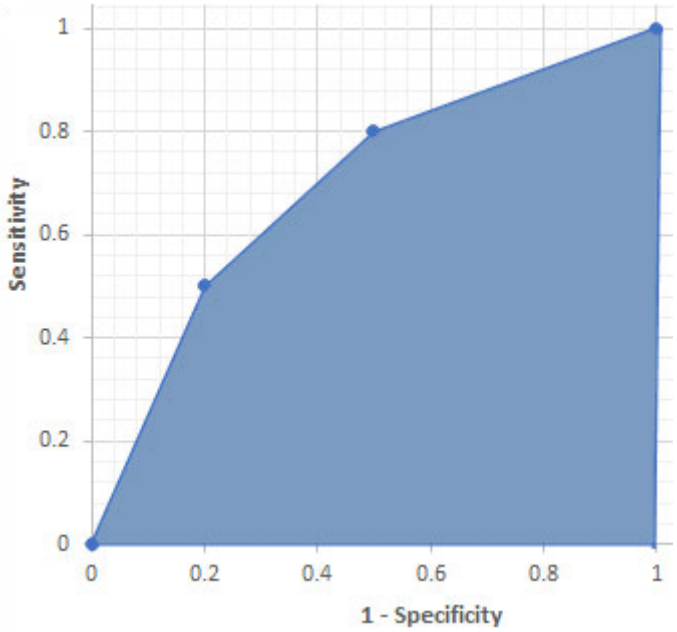


Figure 4.3. Example of a ROC curve constructed from plotting sensitivity and specificity values calculated at different thresholds.

4.3.3 Logistic regression

Multivariate logistic regression models allow the analysis of nonlinear relationships in data and was therefore used in the current study to predict MetS using haematological measures, gene expression levels and gut microbial compositions. In addition to its predictive ability, regression models are also able to identify features that contributed significantly to class label prediction. The current study used the forward stepwise technique, in which variables were added to the model one by one to identify which variables played a significant role in the accurate classification of MetS. The variables that were able to improve the predictive ability of the model were kept and thus considered important in the prediction of MetS. The use of the forward stepwise technique allows the model to handle large numbers of input variables. With cross-validation, 10 final models were created for each variable group (haematological measures, gene expression level and gut microbial composition). The variables that appeared most frequently across the 10 models built were considered to contribute the most to MetS

prediction. In addition, performances of the forward stepwise models were compared to that of full models, in which all variables were used in the construction of the models, to assess the validity of using the forward stepwise technique. The models were built using the three following codes:

$$fullModel = lm(Cohort \sim ., data = trainingSet) \quad (1)$$

$$startModel = lm(Cohort \sim 1, data = trainingSet) \quad (2)$$

$$\text{stepwiseModel} = \text{step}(\text{startModel}, \text{direction} = \text{"}forward\text{"}, scope = formula(fullModel)) \quad (3)$$

All models were built using the *lm* function in R. Code 1 created the full regression model with all the variables available. Code 2 was used to build models with only the coefficient, creating a base model to which variables were added one by one. The variables were added using the stepwise model in Code 3.

4.3.4 Decision tree

Like logistic regression, decision trees are also able to reveal the input variables that contribute to the prediction of class variables. Decision tree construction begins with deciding the root node, which is the variable that provides the largest information gain when separating class variables. The data is then split based on the next best variables until the terminal node is reached and a prediction is made. As the decision tree bifurcates by the variables that provide the largest information gain, the chosen variables are expected to contribute significantly to the prediction of class variables. The current study used the *rpart* package in R to build decision trees for the purpose of MetS prediction. Initially, each decision tree was allowed to grow to the biggest size possible using the *rpart* parameters *minsplit* = 2 and *minbucket* = 1 (Code 4).

$$fullTree = rpart(Cohort \sim ., data = trainingSet, minsplit = 2, minbucket = 1) \quad (4)$$

minsplit allows users to specify the minimum number of observations required in each decision node in order for a split to be made. *minbucket* specifies the minimum number of observations required in each terminal node. After the construction of the trees, the complexity parameter (CP) table was used to decide how to prune each tree, an essential step in preventing overfitting. As a rule of thumb, the trees were pruned at the lowest level in which the sum of the relative error and standard error was lower than the cross-validation error. Table 4.1 provides an example of a CP table generated by the *rpart* package in R. The tree in this example would have been pruned at tree level 3, the lowest level in which the sum of the relative error ("rel_error" = 0.290323) and standard error ("xstd" = 0.123629) was lower than the cross-validation error ("xerror" = 0.548387).

Table 4.1. Example of a complexity parameter table.

Tree level	CP	nsplit	rel_error	xerror	xstd
1	0.516129	0	1	1	0.15575
2	0.193548	1	0.483871	0.580645	0.126623
3	**0.032258**	**2**	**0.290323**	**0.548387**	**0.123629**
4	0.019355	7	0.129032	0.612903	0.129483
5	0.016129	12	0.032258	0.806452	0.144263
6	0.01	14	0	0.806452	0.144263

CP: complexity parameter; nsplit: number of splits; rel_error: relative error; xerror: cross-validation error; xstd: standard error

Code 5 shows an example of how pruned trees were constructed using the results from Table 4.1. Both the full trees and their respective pruned trees were used to predict individuals as either healthy weight controls or obese with MetS. The performances of both trees were then compared to determine whether the pruning technique utilised was successful in eliminating any cases of overfitting.

$$\text{prunedTree=rpart(Cohort} \sim \text{., data = trainingSet, minsplit=2, minbucket=1, cp=0.032258)} \quad (5)$$

In addition, another set of DTs were also created using the R package *rpart* and optimised using the package *e1071*, whereby a grid search approach was used on the parameters *minsplit* and CP. For each tree, a range of *minsplit* (2, 5, 10 and 20) and CP values (0.01, 0.02, 0.05, 0.07 and 0.1) were inputted (Code 6), from which the algorithm constructed and tuned each tree to create the optimal tree model.

$$tunedTree = tune.rpart(Cohort \sim ., data = trainingSet, minsplit = c(2, 5, 10, 20), cp = c(0.01, 0.02, 0.05, 0.07, 0.1))\$best.model \quad (6)$$

The results of each tuned tree, optimised using *e1071*, was then compared to that of the manually pruned tree, using *rpart*, to determine which method was able to best predict MetS. Implementation of the grid search approach means that every time the code was run, a different tree would be created. As such, 50 trees were constructed for each training set and the best performing tree was kept (Figure 4.4). The variables that were deemed important in MetS prediction for all 10 training sets were also noted. As with logistic regression, the most important variables were the ones that appeared most frequently among the 10 models created through cross-validation.

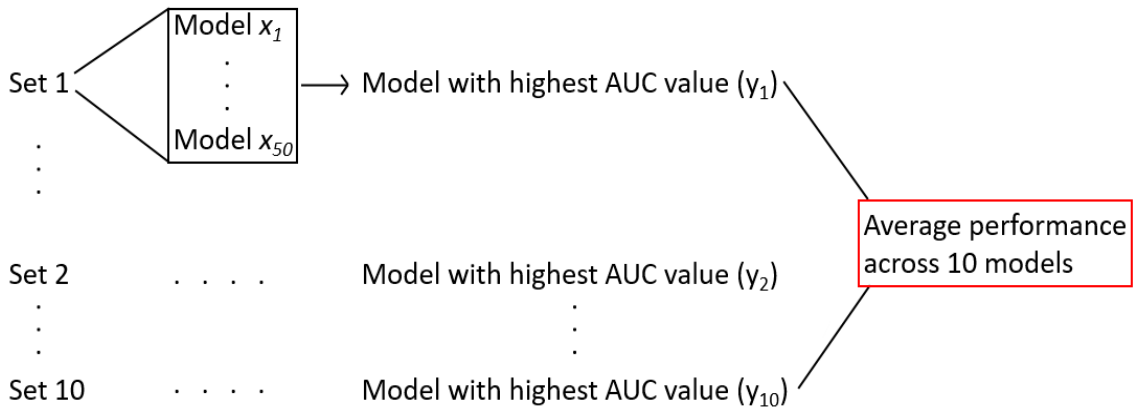


Figure 4.4. Visual demonstration of how the best model for the 10 training sets were chosen for decision tree, support vector machine and artificial neural network.

4.3.5 Support vector machine

Support vector machine is another type of prediction model that can be used with many different types of data, including high-dimensional data. To separate data, four different types of kernel functions can be applied: linear, polynomial, radial basis function (RBF) or sigmoid. Each kernel function has its own range of shape parameters that must be optimised, making SVMs more computationally complex compared to LR and DTs. The current study built a model using each of the four kernel functions and compared them to determine the most appropriate kernel function for the data being analysed. Each SVM model was built and optimised using the R package *e1071*. As with DTs, *e1071* was used to apply a grid search approach to identify the best value to use for each shape parameter specific for each kernel function type (Table 2.2, Section 50, page 51). The grid search was applied on the following range of values for each shape parameter:

- Cost: 0.1, 1, 10, 100;
- Gamma: 0.1, 1, 10;
- Degree: 1, 2, 3, 5, 7, 10; and
- Coefficient: 0.1, 1, 10.

Using the grid search function of the *e1071* package ensured the optimal SVM model was constructed each time. Code 7 shows an example of the code that was used to construct the SVM, with the range of numbers for each shape parameter included.

$$tune.svm(Cohort \sim ., data = trainingSet, kernel = \text{"specifiedKernel"},\\ \text{cost} = 10^{(-1:2)}, \text{gamma} = \text{c}(0.1,1,10), \text{degree} = \text{c}(1,2,3,5,7,10),\\ \text{coef} = \text{c}(0.1,1,10)) \quad (7)$$

The most appropriate kernel function for the data used in the current study was the one that had the highest average performance across the 10 testing sets. The best performing kernel function

was then used to construct 50 different models for each of the 10 training sets to identify the best performing model (Figure 4.4).

4.3.6 Artificial neural network

Another complex and computationally expensive prediction model is the artificial neural network. Due to its "black box" nature, ANNs are difficult to understand, let alone optimise. However, its ability to handle complex and nonlinear data well has made it a popular choice among researchers. ANNs are loosely modelled after a human brain as they interpret input data, recognise a pattern and produce an output. The structure of an ANN consists of three types of layers: the input layer, hidden layer and output layer (Figure 4.5). While the overall structure of ANN sounds simple, there are many factors that must be considered when constructing the model. Some of these factors include the number of hidden layers as well as the number of neurons within each hidden layer, the activation function used, and the method used to train the neural network. For simplicity, the current study only used one hidden layer and applied a grid search to identify the optimal number of neurons to use within the hidden layer. There are no specific guidelines to the number of hidden layer neurons (HLNs) that should be used. Instead, many users follow the rules-of-thumbs formulated by Heaton [1]:

- Between the input layer size and the output layer size;
- Two-thirds of the input layer size plus the output layer size; and
- Less than half the size of the input layer size.

The current study incorporated all three rules and decided to apply a grid search to identify the number of HLNs that provide the best prediction of MetS. The output layer for each model was 1, which was the predicted class output of either '0' or '1'. The number of input layers depended on the variable group used. There were 12 significantly different haematological

measures between the healthy weight and obese with MetS groups, 5 differentially expressed genes and 8 significantly different gut microbial counts. None of the biomarkers from each variable group were correlated with each other. Based on the rules of thumbs, the range of HLN numbers upon which the grid search approach was applied, were 3 to 10 for haematological measures, 3 to 13 for gene expression levels and 3 to 35 for gut microbial counts. For each number in the range of HLNs, 50 different neural networks were constructed for each training set and the best was kept for comparison (Figure 4.4). The number of HLNs that produced the best average performance across the 10 training sets was then deemed the most appropriate. The *neuralnet* package in R was used for building all neural networks. The method applied when training the algorithms was resilient backpropagation with weight backtracking (rprop+). For all methods, neural networks are built by initially allocating random weights to each input neuron which determines the influence each variable has on the prediction model. The weighted sum of the inputs is then adjusted by the bias, which has the same role as the intercept in a linear equation:

$$Y = \Sigma(weight * input) + bias$$

Activation functions allow ANNs to operate on a nonlinear level to handle complex and nonlinear data. The current study used the most common activation function, logistic (also referred to as sigmoid).

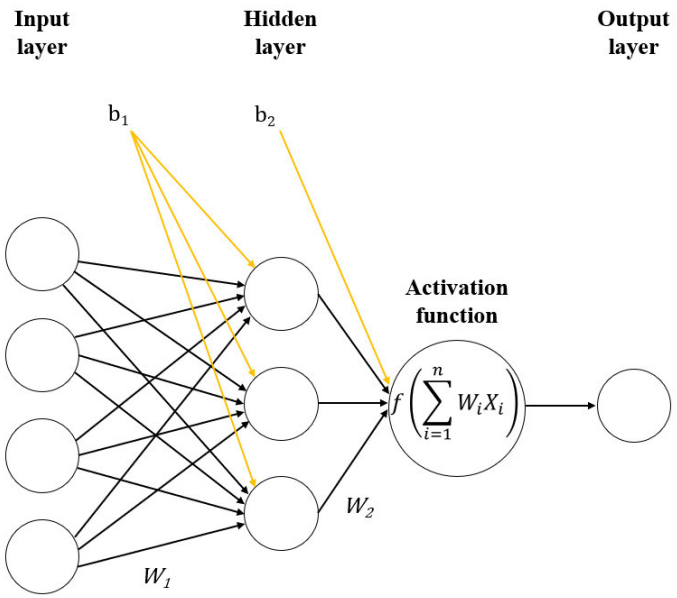


Figure 4.5. Example of neural network layout using 1 input layer with 4 input layer neurons, 1 hidden layer with 3 hidden layer neurons and 1 output layer of size 1. The activation function applied was logistic function.

The error of the model is then calculated by finding the difference between the predicted output and the actual output. Using the error, the loss function can then be calculated which provides an indication of the effect that the error had on the prediction accuracy. The loss function then dictates the direction in which the weights of neurons must be altered to improve the performance of the model. The model has improved if the change in weight moves the loss function towards the optimum. Figure 4.6 shows an example of the loss function calculation. The local optimum is the optimal within a neighbouring set of possible solutions while the global optimum is the optimal solution among all possible solutions. Typically, the loss function will be calculated with the attempt of reaching the closest, or local optimum, which may turn out to be the global optimum. rprop+ works by calculating the derivative, or slope, which determines whether the loss function moved towards or away from the local optimum. If the derivative is positive, the weight was reduced, moving the loss function towards the left. On the other hand, a negative derivative will lead to an in increase in weight, moving the loss function towards the right. The learning rate of the model determines the number of steps the loss function will take towards the local optimum. A lower learning rate indicates smaller steps

being taken, at the cost of being computationally expensive. While the learning rate is typically determined by the user in other training methods, rprop+ is able to identify the best learning rate to use through calculating the derivative. If the derivative changes from positive to negative, rprop+ will recognise that the loss function has overstepped the local optimum. Adjustments will then be made to decrease the step size while simultaneously increasing the weight of input neurons. At the same time, if the steps taken are too small, leading to an increased computational time, rprop+ will also adjust the learning rate accordingly. rprop+ was used to build each ANN while applying a manual grid search approach to identify the optimal number of HLNs to use. The optimal number of HLNs was determined by the performance of the model when predicting MetS.

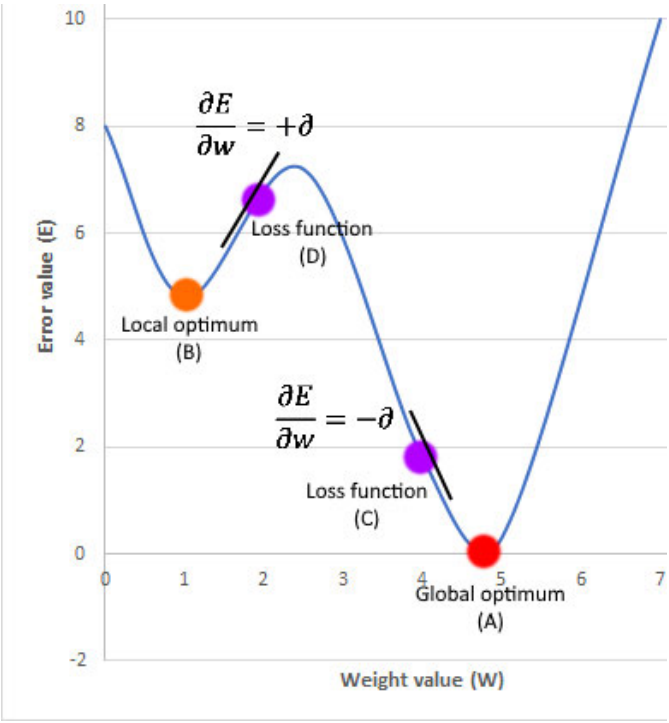


Figure 4.6. Calculation of the loss function in rprop+ to move towards the global (A) or local (B) optima. Loss function (C) has overstepped the global optimum, resulting in a negative derivative while local function (D) has a positive derivative and the step size must now be adjusted for it to move towards the local optimum.

4.4 Results

The current study evaluated the performance of each type of prediction model (LR, DT, SVM and ANN) in predicting MetS using 10-fold cross-validation. The full dataset used for each variable group was split into 10 training and 10 testing datasets, with 90% of participants in each training set and the remaining 10% of participants in each testing set. The testing sets were formed through random sampling without replacement and as such, each participant only appeared in one testing set. Table 4.2 shows the number of healthy weight and obese with MetS participants that appeared in the training and testing sets for each variable group.

Table 4.2. Number of healthy weight and obese with MetS participants in each training and testing sets of the three variable groups.

	Training set		Testing set	
Variable group	Healthy weight (‘0’)	Obese with MetS (‘1’)	Healthy weight (‘0’)	Obese with MetS (‘1’)
Haematological measures	94	31	10	3
Gene expression levels	63	28	6	3
Gut microbial count	81	26	9	2

The current study had a small sample size, resulting in a very small number of participants in each testing set. As such, any misclassification in the testing set would have a significant impact on the performance of the model, particularly the AUC value. The AUC value was therefore calculated by combining the predicted probabilities of the training and testing sets. Rather than reporting an independent AUC value for the training and testing sets, each model only had one AUC value calculated.

The biomarkers that were including in the construction of each prediction model was determined by the lack of strong correlation with any other biomarker. The correlations between biomarkers were assessed using measurements from both the healthy weight and obese with MetS cohorts. This method differed from Section 3.3.4, page 76, whereby the

correlations were assessed in the two groups independently. Strong correlations were found between biomarkers in the haematological measures and gene expression levels, shown in Appendix 4.1 and Appendix 4.2, respectively. No correlations were found between biomarkers in the gut microbial compositions group. After removing the highly correlated variables, the remaining variables that were kept for prediction model construction are shown in Appendix 4.3. As age was found to be significantly different between the two studied cohorts, it was included as a biomarker in the haematological measure variable group to assess its importance in predicting MetS.

In logistic regression, the forward stepwise technique was used which reduced the number of variables used to build the model to only include the variables that increased the performance of the model. In doing so, the computational cost of the model is reduced significantly while also preventing the issue of overfitting. The performances of the resulting models were then compared to the full models built with all variables (Table 4.3). The results of the training set from the forward stepwise model were comparable with that of the full model while the testing set achieved a much higher performance compared to the full model. Additionally, the AUC values for MetS prediction using all three variable groups were all higher in the stepwise model compared to the full model. As such, the models built with the forward stepwise technique were used for further analysis.

Table 4.3. Comparison of the performance by full logistic regression models and models optimised by the forward stepwise technique.

	Full model			Stepwise model		
Variable group	Training Accuracy	Testing Accuracy	AUC	Training Accuracy	Testing Accuracy	AUC
Haematological measures	0.944	0.900	0.982	0.930	0.923	0.978
Gene expression	0.905	0.789	0.927	0.889	0.833	0.917
Gut microbiome	0.921	0.691	0.938	0.834	0.800	0.901

Decision trees are also very prone to overfitting and thus pruning is a crucial step in the construction of DTs. Two different pruning methods were implemented by the current study, manual pruning and pruning though tuning the parameters using the *e1071* package. The results of the trees pruned by the two different methods were compared to determine which one should be used for the final analysis (Table 4.4). Overall, the trees pruned by tuning the parameters attained a higher prediction accuracy in the training set as well as higher AUC values.

Table 4.4. Comparison of the performance of manually pruned trees with that of trees pruned using the grid search approach.

	Pruned Tree			Tuned Tree		
Variable group	Training Accuracy	Testing Accuracy	AUC	Training Accuracy	Testing Accuracy	AUC
Haematological measures	0.931	0.869	0.906	0.960	0.838	0.958
Gene expression	0.811	0.653	0.724	0.902	0.689	0.863
Gut microbiome	0.819	0.736	0.668	0.929	0.727	0.895

Other than reducing the risk of overfitting, pruning also reduces the high computational cost often associated with full-sized DTs. The trees that were tuned using the grid search approach were compared to the full-sized DTs to determine whether the performance was improved (Table 4.5). It is clear that the full-sized DTs were overfitted as the classification accuracy of the training set were all much higher than that of the testing set for all three variable groups. On the other hand, the differences in the classification accuracy between the training and testing sets were reduced in the tuned trees. Additionally, the AUC values were much higher in the tuned trees compared to the full-sized trees and thus tuned trees were used in place of full-sized DTs for the final prediction of MetS.

Table 4.5. Comparison of the performance by full-sized trees and pruned trees.

	Full Tree			Tuned Tree		
Variable group	Training Accuracy	Testing Accuracy	AUC	Training Accuracy	Testing Accuracy	AUC
Haematological measures	1.000	0.872	0.980	0.960	0.838	0.958
Gene expression	1.000	0.654	0.967	0.902	0.689	0.863
Gut microbiome	1.000	0.691	0.965	0.929	0.727	0.895

The pruning of the DTs using the tuning method applied a grid search approach to identify the *minsplit* and CP values that would produce the best model for predicting MetS. The construction of each model may then result in a different result each time, despite being built using the same training data. As such, 50 different trees were built for each training set and the best performing tree was kept for comparison (Figure 4.4). In the end, there were 10 best performing trees for each variable group. For all three variable groups, the CP value that was most commonly used across the 10 training sets was 0.01 while the most common *minsplit* value was 2. The best performing decision tree among the 10 that were constructed in all three variable groups also used a CP value of 0.01. The *minsplit* value used for the best performing trees for each variable group were: 20 for haematological measures, 5 for gene expression levels and 2 for gut microbial counts.

There are four common kernel functions (linear, polynomial, RBF and sigmoid) that can be used with SVM, each splitting the data in different ways. The current study implemented all four kernel functions and compared the results to identify which was most appropriate for use in predicting MetS using the available data. Each kernel function has its own shape parameters that need to be specified. As with decision trees, a grid search was applied to optimise each SVM model with the shape parameter values that will create a model with the best predictive ability. The averaged performance of each kernel function built with 10-fold cross-validation was compared for each variable group (Table 4.6). The RBF kernel predicted MetS using haematological measures and gene expression levels with the highest AUC values of 0.996 and 0.991, respectively. Using gut microbiome, however, the AUC value was 0.694, which is much lower than that of both linear and polynomial kernels. Nevertheless, RBF was able to achieve the highest sensitivity value in the training set of 0.565. As such, RBF was deemed to be the most appropriate kernel function to use for the prediction of MetS. For RBF kernel functions, the shape parameters that need to be optimised are cost and gamma. In all three variable groups,

the best performing SVM models used a gamma value of 0.1. The cost value used was 0.1 for gene expression levels and gut microbial counts and 10 for haematological measures. Similar to the construction of DTs, as a grid search approach was applied, each SVM model built may be different. As a result, there were also 50 SVM models that were built for each training set and the best performing model was kept for comparison.

The current study built neural networks with 1 input layer, 1 hidden layer and 1 output layer. While the number of neurons in both the input and output layers were set, the best number of neurons to use within the hidden layer needed to be identified. As the number of HLNs used increases, so does the risk of overfitting. The current study applied 3 different rules-of-thumbs devised by Heaton [1] to deduce a range of suitable HLN numbers that can be used. The number of variables present in each of the 3 variable groups were: 14 for haematological measures, 19 for gene expression levels and 51 for gut microbial counts. As such, the ranges for HLN used were 3 to 10, 3 to 13 and 3 to 35, respectively. As each neural network is built with a random assignment of weight to each input variable, each neural network built using the same data may be different. Consequently, the current study built 50 different neural networks using each training set for every number of HLN within the range. The neural network that had the highest predictive ability was then used for comparison. As such, for each number in the range of HLN, there were 10 neural networks, one from each training set, that were built. The averages of the 10 networks were used for comparison and to determine the most appropriate number of HLN to use. The averages of the models were compared by AUC values shown in Table 4.7. In the haematological measures variable group, the AUC values for each HLN size were all very high, above 0.9. The number of HLN chosen for the construction of ANNs was 6 as it had a high AUC value of 0.998 as well as high sensitivity values in both the training and testing sets of 0.997 and 0.867, respectively. For analysis with gene expression levels, the HLN size chosen

was 12. Although HLN sizes 11 and 13 both had higher AUC values, using 12 HLNs gave a higher sensitivity value in both the training and testing sets. Finally, the best performing neural network using gut microbial counts was constructed with an HLN size of 22. At this size, the AUC value was calculated to be 0.967 while the sensitivity value in the testing set was 0.750. The sensitivity values of ANNs built with greater HLN sizes does not increase higher than 0.750 and thus an HLN size of 22 was chosen to prevent potential overfitting from using too many HLNs.

As the current study utilised 10-fold cross-validation, each prediction model type had 10 final models remaining for comparison for each of the three variable groups. The predicted probability threshold for each model was altered to produce the highest classification accuracy, sensitivity and specificity values. The average of the best performances achieved by each model is shown in Table 4.8. Overall, the performances of each prediction model type across all 3 variables groups were very high, with the lowest classification accuracy being 0.689. In both the haematological measures and gut microbial composition variable groups, ANN was found to predict MetS with the highest predictive ability, with AUC values of 0.998 and 0.967, respectively. The classification accuracy and sensitivity values attained by ANN were also the highest for both variable groups in the testing set. For gene expression levels, however, the best prediction model for predicting MetS was SVM, with an AUC value of 0.967. However, both LR and ANN achieved a higher sensitivity value than SVM using gene expression levels.

Table 4.6. Comparison of the 4 kernel functions used to build SVM models using the 3 different variable groups.

		Training Set			Testing Set			
Variable group	Kernel function	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	AUC
	Linear	0.946	0.858	0.974	0.923	0.833	0.950	0.984
	Polynomial	0.961	0.887	0.985	0.908	0.733	0.960	0.989
	RBF	0.975	0.932	0.989	0.915	0.767	0.960	0.996
Haematological measures	Sigmoid	0.876	0.700	0.934	0.869	0.700	0.920	0.938
	Linear	0.863	0.650	0.957	0.789	0.567	0.900	0.916
	Polynomial	0.853	0.618	0.957	0.778	0.567	0.883	0.928
	RBF	0.990	0.968	1.000	0.678	0.233	0.900	0.991
Gene expression	Sigmoid	0.735	0.382	0.892	0.744	0.400	0.917	0.761
	Linear	0.771	0.058	1.000	0.773	0.000	0.944	0.928
	Polynomial	0.793	0.154	0.998	0.836	0.150	0.989	0.913
	RBF	0.894	0.565	1.000	0.809	0.000	0.989	0.694
Gut microbiome	Sigmoid	0.745	0.012	0.980	0.791	0.000	0.967	0.494

Table 4.7. Comparison of the averaged best performing neural networks using different hidden layer neuron sizes.

Variable group	Hidden layer neuron size	Training Set			Testing Set			AUC
		Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	
	3	0.979	0.958	0.986	0.931	0.767	0.980	0.992
	4	0.972	0.945	0.981	0.946	0.933	0.950	0.994
	5	0.973	0.955	0.979	0.915	0.800	0.950	0.996
	6	**0.991**	**0.997**	**0.989**	**0.931**	**0.867**	**0.950**	**0.998**
	7	0.990	0.984	0.991	0.962	0.867	0.990	0.998
	8	0.989	0.977	0.993	0.946	0.833	0.980	0.999
	9	0.992	0.990	0.993	0.931	0.833	0.960	0.999
Haematological measures	10	0.990	0.977	0.995	0.946	0.867	0.970	0.999
	3	0.720	0.207	0.948	0.756	0.300	0.983	0.693
	4	0.713	0.179	0.951	0.767	0.367	0.967	0.725
	5	0.732	0.211	0.963	0.744	0.233	1.000	0.738
	6	0.725	0.211	0.954	0.744	0.333	0.950	0.754
	7	0.725	0.211	0.954	0.778	0.367	0.983	0.762
	8	0.732	0.250	0.946	0.789	0.467	0.950	0.751
	9	0.748	0.300	0.948	0.744	0.300	0.967	0.767
	10	0.758	0.404	0.916	0.800	0.500	0.950	0.780
	11	0.754	0.396	0.913	0.789	0.400	0.983	0.806
	12	**0.798**	**0.475**	**0.941**	**0.778**	**0.500**	**0.917**	**0.804**
	13	0.757	0.386	0.922	0.800	0.467	0.967	0.820
Gene expression	14	0.786	0.375	0.968	0.811	0.467	0.983	0.804
	3	0.787	0.227	0.967	0.864	0.400	0.967	0.776
	4	0.780	0.273	0.943	0.864	0.450	0.956	0.793
	5	0.790	0.300	0.947	0.909	0.550	0.989	0.829
	6	0.808	0.396	0.941	0.855	0.500	0.933	0.843
	7	0.841	0.523	0.943	0.836	0.700	0.867	0.866
	8	0.807	0.365	0.948	0.827	0.350	0.933	0.872
	9	0.837	0.492	0.948	0.909	0.650	0.967	0.892

	10	0.843	0.508	0.951	0.891	0.650	0.944	0.888
	11	0.868	0.592	0.957	0.855	0.450	0.944	0.908
	12	0.853	0.565	0.946	0.891	0.850	0.900	0.911
	13	0.895	0.700	0.958	0.873	0.700	0.911	0.923
	14	0.881	0.700	0.940	0.873	0.700	0.911	0.928
	15	0.893	0.688	0.959	0.891	0.650	0.944	0.922
	16	0.923	0.762	0.975	0.909	0.600	0.978	0.942
	17	0.919	0.746	0.974	0.927	0.800	0.956	0.939
	18	0.922	0.777	0.969	0.873	0.700	0.911	0.950
	19	0.927	0.808	0.965	0.882	0.700	0.922	0.953
	20	0.941	0.831	0.977	0.891	0.700	0.933	0.959
	21	0.961	0.896	0.981	0.864	0.550	0.933	0.963
	22	**0.951**	**0.869**	**0.978**	**0.900**	**0.750**	**0.933**	**0.967**
	23	0.949	0.854	0.979	0.873	0.500	0.956	0.965
	24	0.945	0.865	0.970	0.900	0.650	0.956	0.966
	25	0.961	0.888	0.984	0.882	0.750	0.911	0.971
	26	0.961	0.900	0.980	0.873	0.600	0.933	0.979
	27	0.963	0.904	0.981	0.882	0.750	0.911	0.978
	28	0.970	0.908	0.990	0.873	0.700	0.911	0.977
	29	0.963	0.892	0.985	0.882	0.600	0.944	0.983
	30	0.971	0.938	0.981	0.936	0.750	0.978	0.984
	31	0.976	0.912	0.996	0.873	0.750	0.900	0.982
	32	0.982	0.950	0.993	0.882	0.650	0.933	0.987
	33	0.972	0.935	0.984	0.873	0.550	0.944	0.985
	34	0.976	0.912	0.996	0.891	0.750	0.922	0.986
Gut microbiome	35	0.970	0.908	0.990	0.882	0.700	0.922	0.983

Table 4.8. The averaged performance of the 10 best predicted training and testing sets for each prediction model.

		Training Set			Testing Set			
Variable group	Model	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	AUC
	LR	0.930	0.868	0.951	0.923	0.833	0.950	0.978
	DT	0.960	0.913	0.976	0.838	0.633	0.900	0.958
	SVM	0.992	0.994	0.991	0.915	0.867	0.930	0.979
Haematological measures	**ANN**	**0.990**	**0.997**	**0.988**	**0.962**	**0.967**	**0.960**	**0.998**
	LR	0.889	0.764	0.944	0.833	0.733	0.883	0.917
	DT	0.902	0.786	0.954	0.689	0.567	0.750	0.863
	SVM	**0.996**	**0.989**	**0.998**	**0.811**	**0.600**	**0.917**	**0.967**
Gene expression	ANN	0.781	0.661	0.835	0.733	0.633	0.783	0.804
	LR	0.834	0.727	0.868	0.800	0.600	0.844	0.901
	DT	0.929	0.838	0.958	0.727	0.450	0.789	0.895
	SVM	0.992	0.965	1.000	0.827	0.200	0.967	0.945
Gut microbiome	**ANN**	**0.961**	**0.927**	**0.972**	**0.845**	**0.700**	**0.878**	**0.967**

Other than predicting MetS status, two of the prediction models used, LR and DTs, were also able to identify the variables that contributed the most to accurate MetS prediction. In the haematological measures variable group, erythrocyte sedimentation rate (ESR), fasting plasma glucose (FPG), high-density lipoprotein cholesterol (HDL-C), platelets (PLT) and triglycerides (TG) appeared in all 10 models constructed by logistic regression (Table 4.9). The variables C-reactive protein (CRP), age, haemoglobin (HG), glycated haemoglobin A1c (HbA1c), cholesterol appeared in 9, 6, 6, 5 and 1 models, respectively. The best performing LR model from the 10 training sets constructed the following formula:

$$y = -1.91 + 0.16 \times TG - 0.23 \times HDL + 0.001 \times PLT + 0.02 \times ESR + 0.1 \times FPG + 0.01 \times HG + 0.12 \times HbA1c \quad (8)$$

The formula demonstrates that while TG, PLT, ESR, FPG, HG and HbA1c would all increase the odds of being obese with MetS, the odds would decrease with a high HDL-C measurement.

Table 4.9. Important haematological variables used to build the best performing logistic regression models.

Training set	Age	Cholesterol	CRP	ESR	FPG	HbA1c	HDL-C	HG	PLT	TG
1	1	0	1	1	1	0	1	0	1	1
2	1	0	1	1	1	1	1	1	1	1
3	1	0	1	1	1	1	1	1	1	1
4	1	0	1	1	1	0	1	1	1	1
5	1	0	1	1	1	0	1	0	1	1
6	0	1	1	1	1	1	1	0	1	1
7	1	0	1	1	1	0	1	1	1	1
8	0	0	1	1	1	1	1	1	1	1
9	0	0	0	1	1	1	1	1	1	1
10	0	0	1	1	1	0	1	0	1	1
Sum	**6**	**1**	**9**	**10**	**10**	**5**	**10**	**6**	**10**	**10**

Triglycerides also appeared in all 10 models created by DTs (Table 4.10) and was also the root node in all 10 models, signifying its importance. Despite being considered significant in all 10 LR models, HDL and PLT were only used in 8 and 6 DTs, respectively. Additionally, ESR and

FPG, which were also found to be important by all 10 LR models, were not used in the construction of any DT. Other variables that appeared in LR and not DT included cholesterol and HbA1c. On the other hand, eosinophils and lymphocytes, which were not deemed important by LR models, appeared as nodes in DTs. Furthermore, the significance of age in the prediction of MetS was greater in DTs compared to LR, appearing as a node in 8 different DTs compared to 6 LR models.

Table 4.10. Important haematological variables identified by the best performing decision trees.

Training set	Age	CRP	Eosinophils	HDL-C	HG	Lymphocytes	PLT	TG	WCC
1	1	0	0	1	0	0	0	1	0
2	0	0	0	1	0	0	1	1	0
3	1	1	0	1	1	0	1	1	0
4	0	1	1	1	0	0	0	1	0
5	1	1	0	1	1	0	1	1	1
6	1	1	1	1	0	0	1	1	1
7	1	0	0	1	1	1	1	1	0
8	1	1	0	0	0	0	0	1	0
9	1	1	0	0	0	0	0	1	0
10	1	1	0	1	0	0	1	1	0
Sum	**8**	**7**	**2**	**8**	**3**	**1**	**6**	**10**	**2**

Decision trees are able to identify the variables that are likely to contribute significantly to the development of MetS as well as the level of measurement for each variable at which the studied cohorts should be divided. Figure 4.7 provides a visual representation of the best performing DT out of the ten training sets. The DT shows that participants with a TG level less than 1.2 mmol/L, HDL-C of at least 0.96 mmol/L and a PLT count of less than 310 x 10^9/L and were classified as a healthy weight control. However, if the HG level is less than 159 g/L, the participants were classified as obese with MetS. On the other hand, 22% of participants had a TG level higher than 1.2 mmol/L, were over 33 years of age with a CRP level of over 0.69 mg/L, and an HDL-C level of less than 1.6 mmol/L and were subsequently predicted as obese with MetS.

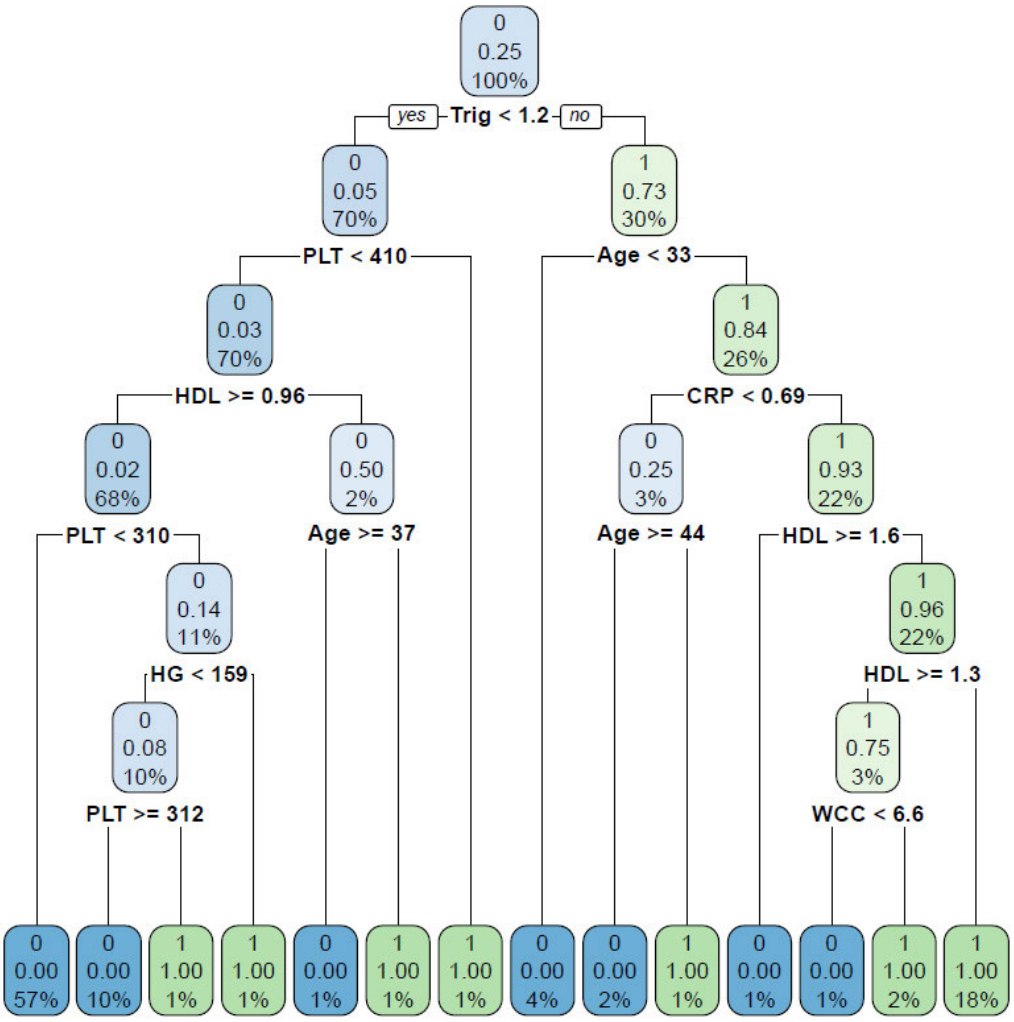


Figure 4.7. Best performing decision tree constructed using haematological measures.

Logistic regression identified the expression of AKT serine/threonine kinase 3 (AKT3), cathelicidin antimicrobial peptide (CAMP), C-C motif chemokine ligand (CCL-)3, C-X-C motif chemokine receptor 6 (CXCR6) and Fc fragment of IgE receptor II (FCER2) to play significant roles in the development of MetS, appearing in all 10 LR models. Other important genes were integrin subunit alpha E (ITGAE) and killer cell lectin-like receptor 2 (KLRC2), appearing in 9 models each, interferon-induced protein with tetratricopeptide repeats 1 (IFIT1) which appeared in 8 models and granzyme H (GZMH) was found in 7 models.

The best performing LR model construction from gene expression levels was:

$$y = -4.34 + 0.70 \times AKT3 + 0.17 \times CAMP + 0.17 \times FCER2 - 0.17 \times CXCR6 - 0.11 \times CCL3 + 0.20 \times IL11RA - 0.17 \times KLRC2 \quad (9)$$

The formula indicates that while AKT3, CAMP, FCER2 and interleukin-11 receptor subunit alpha (IL11RA) expression increases the odds of developing obese with MetS, the expression of CXCR6, CCL3 and KLRC2 reduces these odds. Table 4.11 shows the immune genes that were selected by the 10 constructed LR models as having expressions that were important to the prediction of MetS.

Table 4.11. Important genes used to build the best performing logistic regression models. The table has been split for editorial purposes.

Training set	AKT3	CAMP	CCL3	CXCL5	CXCR6	FCER2	GZMH
1	1	1	1	0	1	1	1
2	1	1	1	0	1	1	1
3	1	1	1	0	1	1	1
4	1	1	1	0	1	1	0
5	1	1	1	1	1	1	0
6	1	1	1	0	1	1	1
7	1	1	1	0	1	1	0
8	1	1	1	0	1	1	1
9	1	1	1	0	1	1	1
10	1	1	1	0	1	1	1
Sum	**10**	**10**	**10**	**1**	**10**	**10**	**7**

Table 4.11 (Cont.)

Training set	HMGB1	IFIT1	IL11RA	ITGAE	KLRC2	SOCS1
1	0	0	1	0	1	1
2	0	1	1	0	1	0
3	0	1	1	0	1	0
4	1	0	1	0	1	1
5	0	1	0	1	0	0
6	0	1	1	0	1	0
7	0	1	1	0	1	0
8	0	1	1	0	1	0
9	0	1	1	0	1	0
10	0	1	1	0	1	0
Sum	**1**	**8**	**9**	**1**	**9**	**2**

All of the variables that were considered to be significant to MetS prediction by LR also appeared as nodes in DTs. There were also five other biomarkers unique to DTs: AKT serine/threonine kinase 1 (AKT1), cluster of differentiation- (CD-)1C, cadherin 1 (CDH1), CEA cell adhesion molecule 3 (CEACAM3) and granzyme M (GZMM). The only biomarker that appeared in all 10 DTs was AKT3, which was also the root node for 6 different DTs. The use of AKT3 in all 10 LR and DT models demonstrates its importance in the development of MetS. For the remaining four DTs, the root node was CAMP. Although CAMP was considered significant by all 10 LR models, it only appeared in 7 DTs. Furthermore, despite the data being split by AKT1 in 9 different DTs, it was not considered a root node. Table 4.12 lists all the immune genes that were found by the 10 constructed DT as having expressions that were important for the prediction of MetS.

Table 4.12. Important genes identified by the best performing decision trees. The table has been split for editorial purposes.

Training set	AKT1	AKT3	CAMP	CCL3	CD1C	CDH1	CEACAM3	CXCR6
1	0	1	1	0	0	0	0	0
2	1	1	0	0	1	0	0	1
3	1	1	1	0	1	1	0	0
4	1	1	1	1	1	0	0	1
5	1	1	0	0	0	1	0	1
6	1	1	1	1	1	0	0	0
7	1	1	1	0	1	1	1	0
8	1	1	1	0	0	1	0	0
9	1	1	0	1	1	0	0	0
10	1	1	1	0	1	0	0	0
Sum	**9**	**10**	**7**	**3**	**7**	**4**	**1**	**3**

Table 4.12 (Cont.)

Training set	FCER2	GZMH	GZMM	HMGB1	IFIT1	IL11RA	ITGAE	KLRC2	SOCS1
1	0	0	0	0	0	0	0	0	0
2	1	0	0	0	1	0	0	0	0
3	0	0	1	0	0	0	1	0	0
4	1	0	0	0	0	0	1	1	0
5	1	1	0	1	0	0	0	0	1
6	0	0	0	0	0	1	0	0	0
7	1	0	1	0	0	0	0	1	1
8	1	0	0	1	1	1	0	0	0
9	1	0	0	0	0	0	0	0	0
10	0	0	0	0	0	1	0	0	1
Sum	**6**	**1**	**2**	**2**	**2**	**3**	**2**	**2**	**3**

The best performing tree constructed using gene expression data is shown in Figure 4.8. The tree identified AKT3 as the root node and found that 47% of participants had low AKT3, FCER2, CAMP and high ITGAE and KLRC2 expression and were classified as a healthy weight control. At the same time, an individual with high expression of AKT3, AKT1 and low expression of CD1C, CCL3 and CXCR6 were predicted to be obese with MetS.

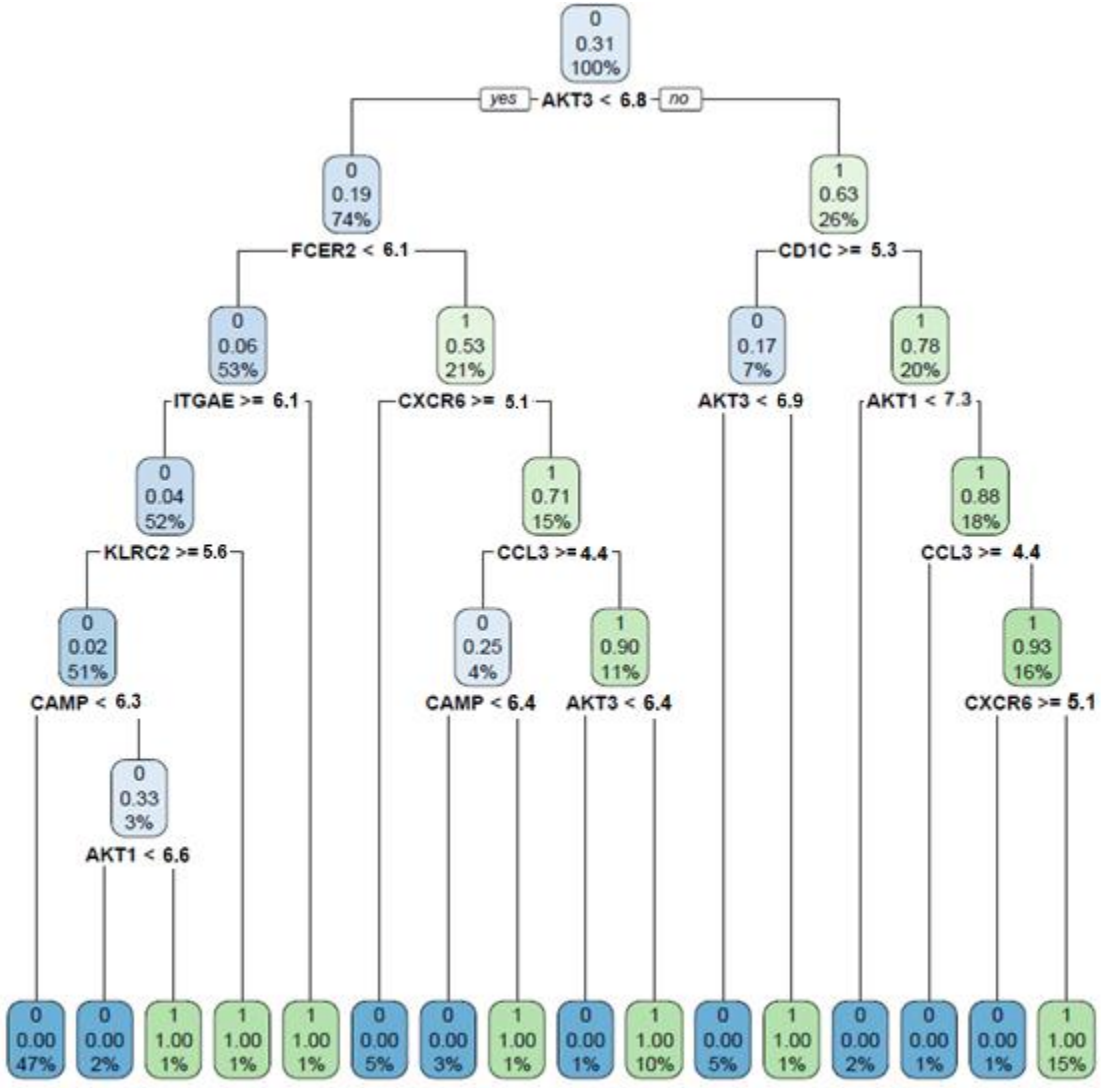


Figure 4.8. Best performing decision tree constructed using gene expression level data.

Due to the high number of gut microbial species that were included in the construction of classification models, for editorial purposes, Table 4.13 and Table 4.14 only show the biomarkers that appeared in at least 5 models. The gut microbial species that appeared in all 10 LR models was *F. prausnitzii*. The model with the highest predictive ability was:

$$\begin{aligned} y \;=\; & 3.04e-01 + 9.33e-05 \times E.rectale - 3.799e-05 \times F.prausnitzii - 1.96e \\ & - 04 \times O.ruminantium + 1.04e-04 \times A.putredinis + 5.71e \\ & - 04 \times B.luti - 3.90e-03 \times M.intestini - 6.48e \\ & - 04 \times A.hadrus + 3.27e-04 \times R.timonensis + 7.62e \\ & - 05 \times C.methylpentosum - 1.10e-03 \times R.faecis \end{aligned} \tag{10}$$

Interestingly, species counts belonging to Firmicutes and Bacteroidetes phyla were both associated with MetS development.

Table 4.13. Important gut microbial species used to build the best performing logistic regression models. The table has been split for editorial purposes.

Training set	Anaerostipes hadrus	Blautia luti	Blautia wexlerae	Clostridium methylpentosum	Eubacterium rectale	Faecalibacterium prausnitzii
1	1	1	0	0	1	1
2	0	1	1	1	1	1
3	0	0	0	0	0	1
4	1	1	1	1	1	1
5	1	1	0	1	1	1
6	0	1	1	1	1	1
7	1	1	1	1	0	1
8	1	1	1	1	0	1
9	1	1	0	1	1	1
10	1	1	1	1	1	1
Sum	**7**	**9**	**6**	**8**	**7**	**10**

Table 4.13 (Cont.)

Training set	Fusicatenibacter saccharivorans	Murimonas intestini	Oscillibacter ruminantium	Romboutsia timonensis	Ruminococcus bromii	Ruminococcus faecis
1	1	1	0	1	1	1
2	0	0	1	1	0	1
3	1	1	1	0	0	1
4	1	1	1	1	1	0
5	1	1	1	1	1	1
6	0	1	1	1	1	1
7	1	1	1	1	1	1
8	1	1	1	1	1	1
9	0	1	1	1	0	1
10	1	1	1	1	1	1
Sum	**7**	**9**	**9**	**9**	**7**	**9**

In the DTs, only three gut microbial species were used as nodes in more than five models: *B. luti*, *F. prausnitzii* and *I. butyriciproducens* (Table 4.14). There were, however, five different species that were used as the root node for the DTs. *I. butyriciproducens* was the root node for 5 DTs, *B. luti* was the root node for 2 trees and *C. comes*, *R. torques* and *R. timonensis* were each the root node for 1 tree.

Table 4.14. Important gut microbial species identified by the best performing decision trees.

Training set	Blautia luti	Faecalibacterium prausnitzii	Intestinimonas butyriciproducens
1	1	0	1
2	1	1	1
3	1	1	1
4	1	1	1
5	1	1	1
6	0	0	0
7	1	0	1
8	1	1	1
9	1	1	1
10	1	0	1
Sum	**9**	**6**	**9**

The best performing tree found a low *B. luti* and *P. merdae* count as well as a high *I. butyriciproducens* count would result in an individual being predicted as a healthy weight control. Additionally, individuals with the same gut microbial profile with the addition of high *F. prausnitzii* count were also predicted to be a healthy weight control. On the other hand, 20% of all participants had a low *B. luti* count and were predicted to be obese with MetS. Other factors that led to an obese with MetS prediction were low *P. phocaeensis*, *M. intestine* and *C. clostridioforme* counts.

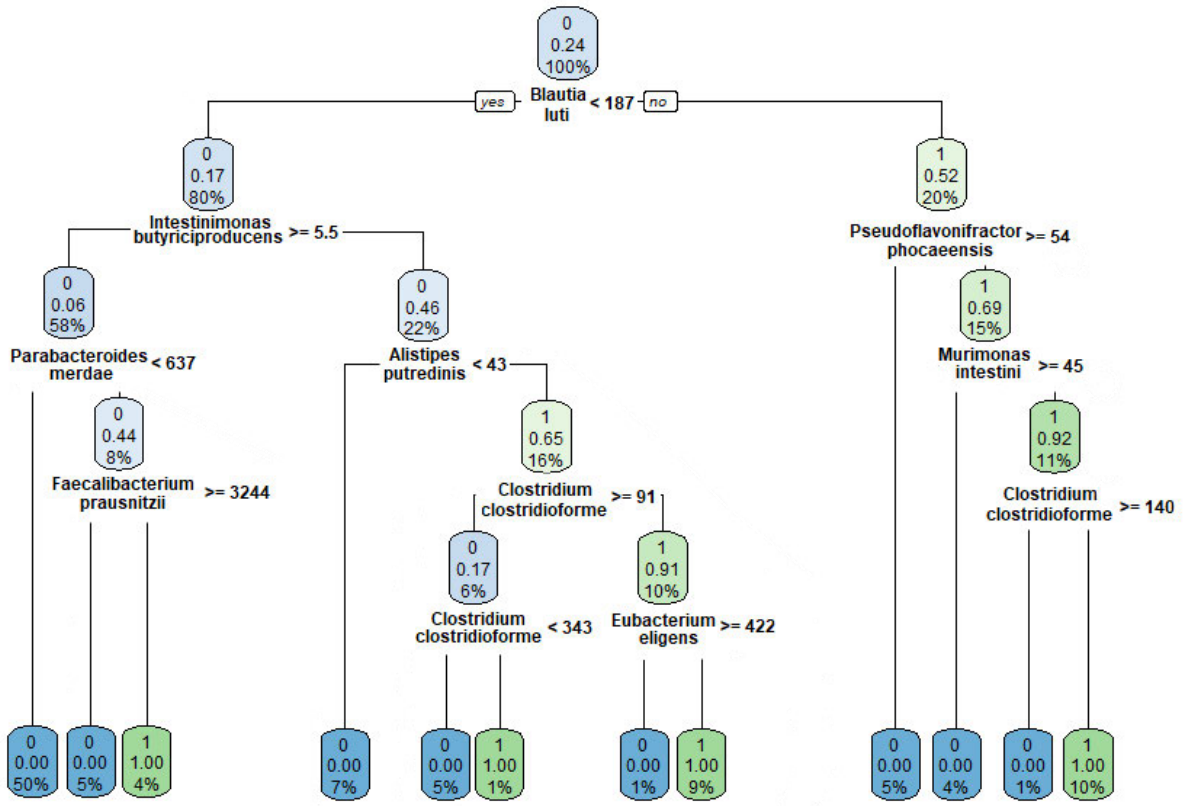


Figure 4.9. Best performing decision tree constructed using gut microbial composition data.

4.5 Discussion

The use of prediction models in research for the purpose of predicting diseases in individuals has increased over the years. Many researchers have come to realise the benefits of multi-analyte methods, with their ability to effectively handle high dimensional data with high accuracy. However, there are a range of prediction models available for use and the challenge lies in deciding which model is most suitable for certain research questions. Studies in MetS have also implemented the use of prediction models but have yet to conclude the most appropriate type of model to use when predicting MetS and related diseases. The current study therefore predicted MetS using four different types of prediction models (LR, DT, SVM and ANN) constructed with haematological measures, gene expression levels and gut microbial composition. The performances of the four models were compared to determine the most appropriate model for predicting individuals more at risk for MetS-related disease development.

For each the four different prediction models, the performance was evaluated using 10-fold cross-validation. The full dataset was divided into 10 training sets, from which each model was constructed, and 10 testing sets, to evaluate the model's predictive ability. Each of the 3 variable groups therefore had their own set of 10 training and testing sets. The use of 10-fold cross-validation is important as it ensures that the predictive ability of each model is assessed with unseen data that was not used to build the model. Using the same data to both build and assess the predictive ability of a model is likely to create bias due to overfitting, whereby the model has learned the data too well. The most appropriate type of prediction model was selected based on the average performance across the 10 training and testing sets. Additionally, LR and DT are both able to produce clinically significant results by identifying the variables that are likely to have contributed the most to MetS development. As such, the variables that were considered to be important across the 10 models produced will be compared to previous literature to determine whether the results of the study supported the findings of previous studies.

For logistic regression, each training set was used to build a model using the forward stepwise technique. Models were also built with the full set of variables to assess the effectiveness of the forward stepwise technique. It was found that the results of the forward stepwise models were comparable with that of the full models. The results suggest the forward stepwise technique to not only be a viable option, but a better choice than building full models when using logistic regression. As the forward stepwise technique typically uses less variables than the full model and thus the prediction accuracy is expected to be lower. However, the results of the training set in the current study have demonstrated that this is not the case. In fact, with a lower number of variables used when building the model, the risk of overfitting as well as the computational cost required is reduced. The forward stepwise technique was able to achieve a comparable classification accuracy and AUC value in the training set, while performing better

than the full model in the testing set. The AUC value increased from 0.906 to 0.978 using haematological measures, from 0.653 to 0.917 with gene expression levels and from 0.736 to 0.901 using gut microbial counts. Further assessments involving LR were therefore all performed using the forward stepwise technique.

MetS status was also predicted through the construction of decision trees. In general, decision trees are very prone to overfitting as they select the nodes that provide the largest information gain, that is the node that provides the best split of participants. To avoid overfitting, decision trees are often pruned used CP values. The current study employed two different methods of pruning: manual pruning by CP and tuning through a grid search approach for the CP and *minsplit* values using the *e1071* package in R. The results of the two pruning methods were compared and the tuning method outperformed the manual pruning method, with AUC values and classification accuracies in the training set using all three variable groups. The trees pruned by the tuning method were therefore used for any further analysis. The predictive abilities of the tuned trees were compared to that of the full-sized trees to determine whether the issue of overfitting was avoided with pruning. The results of MetS prediction using the full-sized trees on both training and testing data indicated without a doubt that the models were overfitted. A clear indication of overfitting is when the classification accuracy in the training set is significantly higher than that of the testing set, which was evident in the results of the current study. Overfitted models fit too closely with the data on which they were built from and thus are unable to accurately predict unseen data. On the other hand, the pruned trees were able to prevent overfitting by closing the gap between the classification accuracies of the training and testing sets for all three variable groups. Using gene expression levels and gut microbial counts, the pruned trees had a higher classification accuracy than the full-sized trees for the testing sets. However, the accuracy using haematological measures was lower in the pruned trees which may be attributable to the variables that were used. As the criteria for MetS mostly consists of

haematological measures, the use of more variables within this variable group will allow a more accurate prediction. Nevertheless, the differences between the classification accuracies between the full-sized trees and pruned trees were very small, 0.869 and 0.838. As such, the pruned trees are still the better choice for MetS prediction as it avoids the issue of overfitting that is evident with full-sized trees. It is clear that by pruning the trees, the issue of overfitting was avoided and in most of the cases, the predictive abilities of the models were increased while simultaneously reducing computational cost.

Both LR and DT have the ability to identify the variables that contributed the most to an accurate prediction of MetS in a clinical setting. Logistic regression identifies important variables through the forward stepwise technique, whereby only the variables that increased the performance of the model are used in the final model. In DTs, branches bifurcate by the variables that provide the greatest information gain and thus these variables are considered significant in predicting the outcome.

For haematological measures, the top variables recognised by LR were ESR, FPG, HDL-C, PLT and TG, all of which appeared in every model constructed. In DTs however, only TG measurement was considered to be of high importance, also appearing in all 10 decision trees. Triglycerides was also the root node for all 10 trees, further reinforcing its importance in MetS prediction. On the other hand, the significance of HDL-C and PLT reduced significantly, appearing only in 8 and 6 DTs, respectively. Additionally, ESR, FPG, cholesterol and HbA1c were not used in the construction of any DTs. The absence of these variables is very interesting as obesity with MetS is associated with a state of chronic low-grade inflammation and thus ESR is expected to be an important factor in the prediction of MetS. Furthermore, FPG, cholesterol and HbA1c are all risk factors of MetS development and are therefore expected to contribute significantly to an accurate prediction of MetS. Despite the exclusion of these variables, however, the DT was able to produce a high classification accuracy of 83.8%, though

this was lower than the 92.3% from LR. The best performing LR model found TG, PLT, ESR, FPG, HG and HbA1c to have a positive contribution to the MetS prediction model, while HDL-C had a negative contribution, as expected. Interestingly, age was not considered as a factor in the best predicting LR model, although it was considered important in 6 other LR models. The best performing DT utilised TG, PLT, HDL-C, age, HG, CRP and white cell count (WCC) as nodes by which data was split. The tree found 57% of participants in the study with low TG, PLT and high HDL-C and predicted them to be healthy weight controls. There were also a few participants, 10% of the dataset, who had low HG but a slightly higher PLT count of at least 312 and were also predicted to be healthy weight controls. The same DT also found age to contribute to MetS development, with 26% of participants aged 33 or over with a high TG level being predicted to be obese with MetS.

Logistic regression identified the expression of 13 out of the total 19 analysed genes to play significant roles in the development of MetS. The expression of five genes in particular were considered to be very important, as they were used for the prediction of MetS in all 10 training sets: AKT3, CAMP, CCL3, CXCR6 and FCER2. On the other hand, AKT3 expression was the only biomarker that appeared in all 10 DTs. It was also used as the root node in 6 DTs, with the other 4 DTs using CAMP. All the genes, except C-X-C motif chemokine ligand- (CXCL-)5, that were considered important by LR were also used as nodes in the prediction of MetS using DTs. The expression of an additional five other genes were also considered important in DTs: AKT1, CD1C, CDH1, CEACAM3 and GZMM. The best performing LR identified the expression of AKT3, FCER2, CAMP and IL11RA to increase the odds of developing MetS while CXCR6, CCL3 and KLRC2 expression reduces these odds. Supporting these results, the best performing DT also predicted individuals with high expression of AKT3, FCER2 and CAMP to be obese with MetS. At the same time, participants with high expression of KLRC2, CXCR6 and CCL3 were predicted as healthy weight controls. FCER2 and CAMP expression

are both associated with obesity and inflammation [2, 3] and thus the prediction of individuals with high expression of these genes as obese with MetS is consistent with previous literature. However, AKT3 is associated with glucose and lipid metabolism [4] with evidence of its expression leading to the protection against insulin resistance. Additionally, both CXCR6 and CCL3 expression are associated with inflammation [5, 6] and as such, the expression of both these genes as well as lower AKT3 expression in the healthy weight control group was unexpected and not supported by previous literature. Conversely, both KLRC2 expression is inversely associated with obesity and inflammation [7, 8] and thus the high expression of KLRC2 leading to the prediction of healthy weight controls was anticipated.

Lastly, prediction of MetS using gut microbial counts with LR models found *F. prausnitzii* counts to be an important factor as it was used in all 10 LR models. Logistic regression found high counts of *E. rectale, A. putredinis, B. luti, R. timonensis* and *C. methylpentosum* to increase the risk of developing MetS. Conversely, high counts of *F. prausnitzii*, *O. ruminantium*, *M. intestini*, *A. hadrus* and *R. faecis* reduced the risk of developing MetS. Microbial species belonging to the Firmicutes and Bacteroidetes phyla were both associated with the increased risk of developing obesity with MetS. Although obesity and MetS has been largely characterised by a higher Firmicutes-to-Bacteroidetes ratio compared to healthy weight controls [9] it may be largely dependent on the species being studied. The counts of each microbial species used in the current study have not all been reported in obese with MetS studies in previous literature. As such, it is likely that the high Firmicutes-to-Bacteroidetes ratio found in obese with MetS individuals may only be true for specific gut microbial species. *F. prausnitzii* appeared in 6 DTs but was not the root node for any of the trees. *I. butyriciproducens* counts, on the other hand, was the root node for 5 DTs and appeared in 9 DTs in total, despite only being considered important by 4 LR models. Interestingly, *B. luti* counts were also considered important by 9 DTs but was only the root node for 2 DTs. The

best performing DT also supported the findings of LR models. Low *B. luti, A. putredinis* along with high *I. butyriciproducens* counts were found in 7% of participants who were classified as healthy weight controls. Similarly, 5% of participants with low *B. luti* and *P. merdae* in conjunction with high *I. butyriciproducens* and *F. prausnitzii* counts and also were predicted to be healthy weight controls. On the other hand, 15% of participants who had high *B. luti* counts and low *P. phocaeensis* and *M. intestini* counts were predicted to be obese with MetS. Despite belonging to the Firmicutes phylum, high counts of *I. butyriciproducens, F. prausnitzii* and *M. intestini* were all associated with healthy weight controls. The findings suggest that researchers should look closely at the gut microbiomes at the species level to confirm whether or not the gut microbial composition of obese with MetS individuals can be broadly described as having a high Firmicutes-to-Bacteroidetes ratio.

The association between gut microbial species and obesity with MetS was also investigated by another published study, shown in Appendix 4.7. The study combined principal component analysis (PCA) with genetic algorithm (GA) to identify the best combination of gut microbes that best predict obesity with MetS. Corresponding to the results from LR, the best performing principal component found high *bromii* counts and low *prausnitzii* and *faecis* to be associated with obesity and MetS. The findings of DTs also reported high *bromii* counts and low *prausnitzii* to be related to obesity and MetS. The results of the current study therefore support the conclusions of the previous literature.

Another type of classification model that was used for the prediction of MetS was SVM, which predicts class outputs by firstly plotting the datapoints onto an n-dimensional space, where n is the number of biomarkers. The hyperplane is then identified to separate the datapoints into the two studied cohorts, healthy weight control and obese with MetS. The separation of the data may differ depending on the kernel function that is used. There are four commonly used kernel functions when building SVM: linear, polynomial, RBF and sigmoid. The current study

compared the predictive ability of models built with each of the four kernel functions to decide which was most appropriate for predicting MetS with the available data. In most cases, the kernel functions linear, polynomial and RBF were considered to be the top performing functions. However, due to the nonlinearity of the data that was used for MetS prediction, the linear kernel function was not used any further. The RBF function had a higher AUC value compared to the polynomial function when predicting MetS using haematological measures and gene expression levels but a lower AUC value when using gut microbial counts. However, the sensitivity value of RBF was much higher than that of the polynomial function in the training set when using gut microbial counts to predict MetS. The results for polynomial and RBF were very close but as RBF is known to handle and represent complex data better without risking saturation of the model, RBF was chosen as the final kernel function. The RBF kernel has the ability to map and approximate almost any nonlinear function through the fine-tuning of its optimisation parameters and thus its high predictive ability came as no surprise. The two shape parameters that must be specified when using the RBF kernel function are cost and gamma. As a grid search was applied to identify the best value of both cost and gamma to use when constructing SVM models, each model built may be different. Consequently, 50 SVM models were built for each training set and the best performing models across the 10 training sets were kept. The threshold for predicted probability was also altered for the 10 remaining models to maximise the performance. The models that best predicted MetS for each of the 10 testing sets across the 3 variable groups had an averaged accuracy of 0.915 when using haematological measures, 0.811 with gene expression levels and 0.827 with gut microbial composition. The averaged AUC values for each variable group were 0.979, 0.967 and 0.945, respectively. Despite having high AUC values for all three variable groups, the sensitivity value when predicting MetS using gut microbial counts in the testing set was very low. The low sensitivity of SVM may be due to its inability to handle gut microbial counts well.

The fourth prediction model type used, ANN, also required optimisation of its own through identifying the best number of HLN to use when predicting MetS. Using three different rules-of-thumbs, a range of appropriate HLN numbers for each variable group was specified. Due to the random allocation of weight to each input variable during the construction of every neural network, every network built will differ from the previous models. As such, for every number in the HLN size range, 50 neural networks were built for every training set. The average of the best performing model across the 10 training sets for each variable group was used to determine which HLN size has the highest predictive ability without overfitting. For each variable group, the HLN size that was chosen was the smallest HLN size with the highest predictive ability overall. For haematological measures, an HLN size of 6 was found to be most appropriate when predicting MetS. While the AUC values for all HLN sizes used were all very high, above 0.99, models constructed with 6 HLNs were found to have the highest sensitivity values in both the training and testing sets. While a higher HLN size would have provided greater predictive ability, the performance increase was not significant enough to warrant the risk of overfitting. Neural networks built using gene expression levels with 12 HLNs produced a very high AUC value of 0.804, surpassed by only two other HLN sizes. However, both the other HLN sizes had lower sensitivity values in both the training and testing sets. As such, 12 HLNs were found to be the most appropriate for predicting MetS using gene expression levels. Lastly, an HLN size of 22 was found to be optimal for building ANNs to predict MetS using gut microbial counts. The smallest number of HLNs at which a sensitivity value of 0.750 in the testing set was reached for the first time was 22 HLNs. Although the highest sensitivity value reached altogether using gut microbial counts was 0.850, at HLN size 12, the overall performance of the model was not as high. At higher HLN sizes, the performance does not increase much higher and thus an HLN size of 22 was used for further analysis.

Following the optimisation of each prediction model, the predicted probability threshold was adjusted to maximise the performance through the classification accuracy, sensitivity and specificity values. The highest performance of the 10 models constructed for each prediction model type and variable group was then averaged. The resulting outcome was then used to compare the four different types of prediction models used and determine which one was able to predict MetS with the highest accuracy. While ANN achieved the highest values for MetS prediction using haematological measures, the other three prediction models also yielded very high AUC values of at least 0.95. The high predictive ability of all four models was expected, as the variables used for the prediction included measurements that comprise the MetS criteria. In some studies [10], fasting plasma glucose did not contribute significantly to an accurate prediction of MetS, while others [11] found it to be the core predictor. As such, the current study included all the MetS measurements to determine whether they would be considered in the optimal combination of biomarkers for MetS prediction. Other than haematological measures, ANN also had the highest performance for MetS prediction using gut microbial data. Although SVM outperformed ANN when predicting MetS using gene expression levels, ANN still achieved a high AUC value of 0.804. Consequently, ANN was regarded as the prediction model that had the highest predictive ability for MetS as a whole.

Studies using classification models to predict MetS using genes and gut microbiome data is very limited and thus the results of the current study could not be compared to that of previous literature. The results of predicting MetS using haematological measures could however be compared. The current study predicted MetS using logistic regression and achieved a sensitivity value of 0.833 and specificity value of 0.95 which was higher than the values from the studies by Al-Thani et al. (57.8 sensitivity and 58.4% specificity) [12] and Mao et al. (80.49% sensitivity and 81.71% specificity) [13]. The low performance of LR models in the study by Al-Thani et al. is likely due to the parameters that were used for prediction. Al-Thani

et al. used waist circumference (WC), waist-to-hip ratio (WHR) and body mass index (BMI) to predict at least two other risk factors of MetS. As MetS is comprised mostly of haematological factors, Al-Thani et al. may have yielded higher prediction values if haematological variables were used, as demonstrated by the current study. Similarly, while Mao et al. used more suitable variables for the prediction of MetS, the replacement of variables such as fructosamine and albumin with measurements used by the current study, such as triglycerides and HbA1c, may increase the predictive ability of the models.

On the other hand, while the current study predicted MetS using decision tree with 96% accuracy in the training set and 83.8% in the testing set, it was lower than the 99.86% accuracy reached by Worachartcheewan et al. [14]. The higher classification accuracy in MetS prediction by Worachartcheewan et al. may be a result of using more biomarkers than the current study to build the decision tree. Worachartcheewan et al. used 20 different haematological and biochemical parameters while the current study only used 14 variables in the haematological measures variable group. However, although 10-fold cross-validation was used by Worachartcheewan et al., there was no mention of pruning of the tree and thus it cannot be concluded whether the decision trees built were subject to overfitting. If comparisons were made based on full-sized decision trees, the current study would have a higher classification accuracy, despite being overfitted. The results of the current study can also be compared to that of AlJarullah [11], who predicted the MetS-related disease, diabetes, using DTs and attained an accuracy of 78.18%. AlJarullah only used 5 health parameters (age, BMI, plasma glucose, diabetes pedigree function and number of times pregnant) suggesting the effect that both the number of variables used, as well as the relevancy to the disease, on the performance of the model built.

Support vector machines were used by Karimi-Alavijeh, Salili & Sadeghi [15] to predict MetS with an accuracy of 75.7%. Despite the authors acknowledging the importance of selecting the

right kernel function and the impact it may have on the results, there was no justification in the choice of polynomial as the final kernel function. Without knowing how each kernel function was optimised, it is difficult to comment on the choice of kernel function and why other functions were not a better choice. At the same time, although Karimi-Alavijeh, Salili & Sadeghi found polynomial to produce the highest classification accuracy, it was still lower than the results of the current study, even when compared to the polynomial kernel function. While the current study found RBF to be the best performing kernel in predicting MetS, the results from the polynomial kernel function were also very high, with 90.8% classification accuracy. In another study, SVM with the RBF kernel was used by Kumari & Chitra to predict diabetes. However, the classification accuracy was only 75.5% compared to the 91.5% from the current study. A closer look into the design of the study revealed that although training and testing sets were utilised, the size of the testing set was 260 while the training set only included 200 participants. As the models are built using the training set and evaluated using the testing set, the size of the training set should always exceed that of the testing set.

The prediction of MetS was also undertaken with ANN models in a study by Alić et al. [16]. As with the current study, Alić et al. compared the performance of each model using different HLN sizes to determine which one was most suitable. After comparing networks built with 5, 10, 14 and 17 HLN, the study found an HLN size of 14 to produce the highest classification accuracy of 94.3%. However, Alić et al. built the neural network with only 6 input variables and thus an HLN size of 14 is likely to result in overfitting. A model that has been overfitted typically has a much lower performance in the training set compared to the testing set, which then explains why Alić et al. only reported the results of the training set. To allow a fair comparison, the results of Alić et al. were compared to the training set of the current study which had a classification accuracy of 99.1%, higher than what was achieved by Alić et al. Another study by Meng et al. [17] built neural networks with 15 HLN to predict diabetes in

1,487 individuals. The classification accuracy of the model used by Meng et al. was 73.23%, which was also lower than that of the current study, with an accuracy of 93.8% in the testing set. Similar to Alić et al., the number of HLN used by Meng et al. was greater than the number of input variables that were used to build the model. There were only 12 input variables used and therefore an HLN size of 15 is likely to result in an overfitted model. In addition, the variables used by Meng et al. included preference for salty food, coffee consumption and fish consumption. By including more blood parameters, the accuracy of diabetes prediction is likely to increase significantly.

Many recent studies have started to recognise the benefits of using classification models to predict diseases, including the ability to handle high dimensional data as well as the high predictive ability, as demonstrated by the current study. However, deciding the best classification model to use for predicting diseases such as MetS has stunted the research progress. In response to this issue, the current study has compared the performance of four different types of classification models in predicting MetS using three groups of variables. Overall, it was found that each model was able to predict MetS with very high accuracy particularly ANN and SVM. However, the most appropriate prediction model to use in research does not only depend on the predictive ability of the model but also the research question that is being asked. While ANN and SVM both produced demonstrated high predictive abilities, neither model is able to produce a result that is easily interpretable in the clinical setting. As such, in studies that may be looking for biomarkers to target for intervention, LR and DT would be the better choice.

There were two notable limitations in this study. First, the type of data that was used for the prediction of MetS. While the current study used haematological measures, gene expression levels and gut microbial counts for the prediction of MetS, there are other measurements that may be more suitable. As many studies in obesity and MetS have utilised data collected from

adipose tissue and derived many clinically significant results, future research should also include these measurements. The second limitation of the study was the sample size of the study. Due to the small number of obese with MetS participants in the dataset, the testing sets only had either two or three obese with MetS participants. As such, any misclassification would heavily affect the predictive values, particularly the sensitivity and AUC values. To minimise the effect of the small sample size, the current study decided to calculate AUC by combining the predicted probabilities of both the training and testing sets. Even with these limitations, the models used in the study still had high predictive ability and the study was able to draw a conclusion for the type of prediction models that are most suitable for different research questions using specific types of data.

4.6 Appendices

Appendix 4.1. Correlations between biomarkers from the combined haematological measures variable group.

Biomarker	Correlated biomarker	Correlation coefficient
HG	RCC	0.799973
HG	HCT	0.922215
RCC	HCT	0.816959
WCC	Neutrophils	0.86965
WCC	Monocytes	0.717338
Cholesterol	LDL-C	0.87219

Appendix 4.2. Correlations between biomarkers from the combined gene expression variable group.

Biomarker	Correlated biomarker	Correlation coefficient
AKT1	ATG7	0.78075
ATG7	CD163	0.753144
AKT1	CD68	0.787639
ATG7	CD84	0.702957
CD163	CD84	0.774307
AKT1	CSF3R	0.717828
ATG7	CSF3R	0.83556
CD163	CSF3R	0.796133
CD84	CSF3R	0.724847
AKT1	FCAR	0.75754
ATG7	FCAR	0.809685
CSF3R	FCAR	0.777678
AKT1	FPR1	0.797708
ATG7	FPR1	0.831323
CSF3R	FPR1	0.863066
FCAR	FPR1	0.84144
AKT1	IL1RN	0.753771
ATG7	IL1RN	0.791287
CSF3R	IL1RN	0.789727
FCAR	IL1RN	0.863954
FPR1	IL1RN	0.899598
ATG7	INSR	0.784588
CD163	INSR	0.769776
CD84	INSR	0.728708
CSF3R	INSR	0.723873
IFIT1	IRF7	0.76102
ATG7	LCN2	0.741818
CAMP	LCN2	0.79784
FPR1	LCN2	0.703306
ATG7	LTF	0.709376
CAMP	LTF	0.795221
LCN2	LTF	0.921307
ATG7	MAPK1	0.825827
CSF3R	MAPK1	0.807753
FCAR	MAPK1	0.765101
FPR1	MAPK1	0.806229
IL1RN	MAPK1	0.763708
INSR	MAPK1	0.741347
LCN2	MAPK1	0.725005
ATG7	MTOR	0.809937
CD163	MTOR	0.799542
CD84	MTOR	0.79581
CSF3R	MTOR	0.747387
INSR	MTOR	0.763154
MAPK1	MTOR	0.709763

ATG7	NRF1	0.730109
CD163	NRF1	0.730562
CD84	NRF1	0.758839
CSF3R	NRF1	0.701209
FPR1	NRF1	0.715431
INSR	NRF1	0.762192
MAPK1	NRF1	0.708944
MTOR	NRF1	0.759201
AKT1	PIK3CA	0.711191
ATG7	PIK3CA	0.842016
CD163	PIK3CA	0.791177
CD84	PIK3CA	0.828892
CSF3R	PIK3CA	0.86183
FCAR	PIK3CA	0.758152
FPR1	PIK3CA	0.753099
INSR	PIK3CA	0.794297
MAPK1	PIK3CA	0.825275
MTOR	PIK3CA	0.836028
NRF1	PIK3CA	0.761979
ATG7	PIK3R1	0.786799
CD163	PIK3R1	0.807294
CD84	PIK3R1	0.879489
CSF3R	PIK3R1	0.81661
INSR	PIK3R1	0.824792
MAPK1	PIK3R1	0.775662
MTOR	PIK3R1	0.879256
NRF1	PIK3R1	0.757557
PIK3CA	PIK3R1	0.941014
AKT1	PYCARD	0.7591
ATG7	PYCARD	0.822982
CSF3R	PYCARD	0.857126
FCAR	PYCARD	0.844993
FPR1	PYCARD	0.902322
IL1RN	PYCARD	0.889133
INSR	PYCARD	0.719493
LCN2	PYCARD	0.716616
MAPK1	PYCARD	0.8421
PIK3CA	PYCARD	0.717636
AKT1	RPS6KA1	0.767825
ATG7	RPS6KA1	0.838452
CD163	RPS6KA1	0.852107
CD84	RPS6KA1	0.781217
CSF3R	RPS6KA1	0.942886
FCAR	RPS6KA1	0.782094
FPR1	RPS6KA1	0.86021
IL1RN	RPS6KA1	0.767141
INSR	RPS6KA1	0.794729
MAPK1	RPS6KA1	0.816226

MTOR	RPS6KA1	0.797516
NRF1	RPS6KA1	0.770788
PIK3CA	RPS6KA1	0.900594
PIK3R1	RPS6KA1	0.87502
PYCARD	RPS6KA1	0.844825
IL1RN	SERPING1	0.718689
AKT1	TFEB	0.831107
ATG7	TFEB	0.815542
CD68	TFEB	0.707927
CSF3R	TFEB	0.779214
FCAR	TFEB	0.735302
FPR1	TFEB	0.880288
IL1RN	TFEB	0.84108
MAPK1	TFEB	0.783258
PYCARD	TFEB	0.879484
RPS6KA1	TFEB	0.802772
CD68	TNFSF13	0.803732
ATG7	TSC1	0.769635
CD163	TSC1	0.855447
CD84	TSC1	0.849219
CSF3R	TSC1	0.806145
INSR	TSC1	0.794741
MAPK1	TSC1	0.726452
MTOR	TSC1	0.879033
NRF1	TSC1	0.858187
PIK3CA	TSC1	0.840411
PIK3R1	TSC1	0.879117
RPS6KA1	TSC1	0.867343
AKT1	ULK1	0.731125
ATG7	ULK1	0.846697
CD163	ULK1	0.751452
CD84	ULK1	0.742189
CSF3R	ULK1	0.933789
FCAR	ULK1	0.788839
FPR1	ULK1	0.876916
IL1RN	ULK1	0.793171
INSR	ULK1	0.789388
LCN2	ULK1	0.70423
MAPK1	ULK1	0.824687
MTOR	ULK1	0.763756
NRF1	ULK1	0.741114
PIK3CA	ULK1	0.859874
PIK3R1	ULK1	0.84198
PYCARD	ULK1	0.85883
RPS6KA1	ULK1	0.93805
TFEB	ULK1	0.79724
TSC1	ULK1	0.813346

Appendix 4.3. List of remaining variables from each group after removing highly correlated variables.

Haematological measures	Gene expression levels	Gut microbial counts
Age	AKT1	Agathobaculum butyriciproducens
Basophils	AKT3	Alistipes onderdonkii
Cholesterol	CAMP	Alistipes putredinis
CRP	CCL3	Anaeromassilibacillus senegalensis
EOS	CD1C	Anaerostipes hadrus
ESR	CDH1	Bacteroides stercoris
FPG	CEACAM3	Bacteroides thetaiotaomicron
HbA1c	CXCL5	Bacteroides uniformis
HDL-C	CXCR6	Bacteroides vulgatus
HG	FCER2	Blautia faecis
Lymphocytes	GZMH	Blautia luti
PLT	GZMM	Blautia wexlerae
TG	HMGB1	Clostridium clostridioforme
WCC	IFIT1	Clostridium leptum
	IL11RA	Clostridium methylpentosum
	ITGAE	Clostridium spiroforme
	KLRC2	Clostridium xylanolyticum
	S100A12	Coprococcus catus
	SOCS1	Coprococcus comes
		Desulfovibrio simplex
		Dorea formicigenerans
		Dorea longicatena
		Eubacterium coprostanoligenes
		Eubacterium eligens
		Eubacterium ramulus
		Eubacterium rectale
		Faecalibacterium prausnitzii
		Flavonifractor plautii
		Flintibacter butyricus
		Fusicatenibacter saccharivorans
		Gemmiger formicilis
		Hespellia porcina
		Ihubacter massiliensis
		Intestinimonas butyriciproducens
		Lachnoclostridium pacaense
		Murimonas intestini
		Neglecta timonensis
		Odoribacter splanchnicus
		Oscillibacter ruminantium
		Oscillibacter valericigenes
		Parabacteroides distasonis
		Parabacteroides merdae
		Pseudoflavonifractor phocaeensis
		Romboutsia timonensis
		Roseburia inulinivorans

	Ruminococcus bromii
	Ruminococcus champanellensis
	Ruminococcus faecis
	Ruminococcus torques
	Ruthenibacterium lactatiformans
	Sporobacter termitidis

AKT1: AKT serine/threonine kinase 1; AKT3: AKT serine/threonine kinase 3; CAMP: cathelicidin antimicrobial peptide; CCL3: C-C motif chemokine ligand 3; CD1C: CD1C molecule; CDH1: cadherin 1; CEACAM3: CEA cell adhesion molecule 3; CRP: C-reactive protein; CXCL5: C-X-C motif chemokine ligand 5; CXCR6: C-X-C motif chemokine receptor 6; ESR: erythrocyte sedimentation rate; FCER2: Fc fragment of IgE receptor II; FPG: fasting plasma glucose; GZMH: granzyme H; GZMM: granzyme M; HbA1c: haemoglobin A1c; HDL-C: high-density lipoprotein cholesterol; HG: haemoglobin; HMGB1: high mobility group box 1; IFIT1: interferon induced protein with tetratricopeptide repeats 1; IL11RA: interleukin 11 receptor subunit alpha; ITGAE: integrin subunit alpha E; KLRC2: killer cell lectin like receptor C2; PLT: platelets; S100A12: S100 calcium binding protein A12; SOCS1: suppressor of cytokine signalling 1; TG: triglycerides; WCC: white blood cell count

Appendix 4.4. List of the gut microbial species, and the phylum to which they belong, that were used for the construction of classification models.

Phylum	Species
Firmicutes	Agathobaculum butyriciproducens
Bacteroidetes	Alistipes onderdonkii
Bacteroidetes	Alistipes putredinis
Firmicutes	Anaeromassilibacillus senegalensis
Firmicutes	Anaerostipes hadrus
Bacteroidetes	Bacteroides stercoris
Bacteroidetes	Bacteroides thetaiotaomicron
Bacteroidetes	Bacteroides uniformis
Bacteroidetes	Bacteroides vulgatus
Firmicutes	Blautia faecis
Firmicutes	Blautia luti
Firmicutes	Blautia wexlerae
Firmicutes	Clostridium clostridioforme
Firmicutes	Clostridium leptum
Firmicutes	Clostridium methylpentosum
Firmicutes	Clostridium spiroforme
Firmicutes	Clostridium xylanolyticum
Firmicutes	Coprococcus catus
Firmicutes	Coprococcus comes
Proteobacteria	Desulfovibrio simplex
Firmicutes	Dorea formicigenerans
Firmicutes	Dorea longicatena
Firmicutes	Eubacterium coprostanoligenes
Firmicutes	Eubacterium eligens
Firmicutes	Eubacterium ramulus
Firmicutes	Eubacterium rectale
Firmicutes	Faecalibacterium prausnitzii
Firmicutes	Flavonifractor plautii
Firmicutes	Flintibacter butyricus
Firmicutes	Fusicatenibacter saccharivorans
Firmicutes	Gemmiger formicilis
Firmicutes	Hespellia porcina
Firmicutes	Ihubacter massiliensis
Firmicutes	Intestinimonas butyriciproducens
Firmicutes	Lachnoclostridium pacaense
Firmicutes	Murimonas intestini
Firmicutes	Neglecta timonensis
Bacteroidetes	Odoribacter splanchnicus
Firmicutes	Oscillibacter ruminantium
Firmicutes	Oscillibacter valericigenes
Bacteroidetes	Parabacteroides distasonis
Bacteroidetes	Parabacteroides merdae
Firmicutes	Pseudoflavonifractor phocaeensis
Firmicutes	Romboutsia timonensis

Firmicutes	Roseburia inulinivorans
Firmicutes	Ruminococcus bromii
Firmicutes	Ruminococcus champanellensis
Firmicutes	Ruminococcus faecis
Firmicutes	Ruminococcus torques
Firmicutes	Ruthenibacterium lactatiformans
Firmicutes	Sporobacter termitidis

Appendix 4.5. Complete list of the important gut microbial species identified by the best performing logistic regression models. Table has been split for editorial purposes.

Training set	Agathobaculum butyriciproducens	Alistipes onderdonkii	Alistipes putredinis	Anaeromassilibacillus senegalensis	Anaerostipes hadrus
1	1	0	0	1	1
2	0	0	1	0	0
3	0	0	0	0	0
4	0	1	1	0	1
5	0	1	0	0	1
6	0	0	1	0	0
7	0	1	0	0	1
8	0	1	0	0	1
9	1	0	1	0	1
10	0	0	0	0	1
Sum	**2**	**4**	**4**	**1**	**7**

Appendix 4.5 (Cont.)

Training set	Bacteroides uniformis	Bacteroides vulgatus	Blautia faecis	Blautia luti	Blautia wexlerae	Clostridium clostridioforme
1	0	0	0	1	0	0
2	0	1	0	1	1	1
3	0	0	0	0	0	0
4	0	1	1	1	1	0
5	0	0	0	1	0	0
6	1	0	0	1	1	0
7	1	0	0	1	1	0
8	0	0	0	1	1	0
9	1	0	1	1	0	0
10	0	0	0	1	1	0
Sum	**3**	**2**	**2**	**9**	**6**	**1**

Appendix 4.5 (Cont.)

Training set	Flavonifractor plautii	Fusicatenibacter saccharivorans	Intestinimonas butyriciproducens	Lachnoclostridium pacaense	Murimonas intestini
1	0	1	1	0	1
2	0	0	1	0	0
3	0	1	0	1	1
4	1	1	1	1	1
5	0	1	0	0	1
6	0	0	0	0	1
7	0	1	0	0	1
8	0	1	1	0	1
9	0	0	0	0	1
10	0	1	0	0	1
Sum	**1**	**7**	**4**	**2**	**9**

Appendix 4.5 (Cont.)

Training set	Oscillibacter ruminantium	Parabacteroides merdae	Pseudoflavonifractor phocaeensis	Romboutsia timonensis
1	0	0	0	1
2	1	0	1	1
3	1	0	0	0
4	1	0	1	1
5	1	1	0	1
6	1	0	0	1
7	1	0	1	1
8	1	0	1	1
9	1	0	0	1
10	1	1	0	1
Sum	**9**	**2**	**4**	**9**

Appendix 4.5 (Cont.)

Training set	Ruminococcus bromii	Ruminococcus faecis	Ruminococcus torques	Ruthenibacterium lactatiformans
1	1	1	0	0
2	0	1	0	0
3	0	1	0	0
4	1	0	1	1
5	1	1	0	1
6	1	1	0	0
7	1	1	1	0
8	1	1	0	0
9	0	1	0	0
10	1	1	0	0
Sum	**7**	**9**	**2**	**2**

Appendix 4.6. Complete list of the important gut microbial species identified by the best performing decision trees.

Training set	Agathobaculum butyriciproducens	Alistipes onderdonkii	Alistipes putredinis	Anaerostipes hadrus	Bacteroides stercoris	Bacteroides uniformis
1	0	0	0	0	0	0
2	0	0	0	0	0	0
3	0	0	0	0	1	0
4	0	1	0	0	0	1
5	0	1	0	0	0	0
6	1	0	0	0	0	0
7	0	0	1	0	1	0
8	0	0	1	0	0	0
9	0	1	1	1	0	0
10	0	0	1	0	0	0
Sum	**1**	**3**	**4**	**1**	**2**	**1**

Appendix 4.6 (Cont.)

Training set	Bacteroides vulgatus	Blautia luti	Blautia wexlerae	Clostridium clostridioforme	Clostridium leptum	Clostridium methylpentosum
1	1	1	0	0	1	0
2	0	1	1	0	0	0
3	1	1	0	1	0	0
4	1	1	0	1	1	0
5	0	1	0	0	1	0
6	0	0	1	0	0	1
7	0	1	0	1	0	0
8	0	1	0	1	0	0
9	0	1	0	0	0	0
10	0	1	0	0	0	0
Sum	**3**	**9**	**2**	**4**	**3**	**1**

Appendix 4.6 (Cont.)

Training set	Clostridium spiroforme	Coprococcus catus	Coprococcus comes	Desulfovibrio simplex	Eubacterium coprostanoligenes	Eubacterium eligens
1	0	0	0	0	0	1
2	0	0	0	0	0	0
3	1	1	1	0	0	0
4	1	0	0	0	0	0
5	1	0	0	0	0	0
6	0	0	1	1	0	1
7	0	0	0	0	1	0
8	0	0	0	0	0	1
9	0	0	0	0	0	0
10	0	0	0	0	0	1
Sum	**3**	**1**	**2**	**1**	**1**	**4**

Appendix 4.6 (Cont.)

Training set	Eubacterium rectale	Faecalibacterium prausnitzii	Hespellia porcina	Intestinimonas butyriciproducens	Lachnoclostridium pacaense
1	0	0	0	1	1
2	0	1	0	1	0
3	0	1	0	1	1
4	0	1	0	1	1
5	1	1	0	1	0
6	0	0	1	0	0
7	0	0	0	1	0
8	0	1	0	1	0
9	1	1	0	1	0
10	0	0	0	1	0
Sum	**2**	**6**	**1**	**9**	**3**

Appendix 4.6 (Cont.)

Training set	Murimonas intestini	Neglecta timonensis	Oscillibacter ruminantium	Parabacteroides merdae	Pseudoflavonifractor phocaeensis
1	0	0	0	0	0
2	0	0	0	0	0
3	1	0	1	0	0
4	0	0	0	1	0
5	1	0	0	1	1
6	0	0	1	0	0
7	0	0	0	1	0
8	1	0	0	1	1
9	0	0	0	0	0
10	0	1	0	0	1
Sum	**3**	**1**	**2**	**4**	**3**

Appendix 4.6 (Cont.)

Training set	Romboutsia timonensis	Ruminococcus bromii	Ruminococcus torques	Ruthenibacterium lactatiformans
1	0	1	0	0
2	0	0	0	0
3	0	1	0	0
4	0	0	0	0
5	0	0	0	1
6	0	0	1	0
7	0	0	0	0
8	0	0	0	0
9	0	0	0	0
10	1	1	0	0
Sum	**1**	**3**	**1**	**1**

Statement of contribution to co-authored published paper

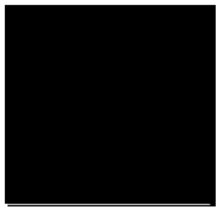
A co-authored published paper was included in this chapter. The bibliographic details of the co-authored published paper, including all authors, are:

Zhang, P., West, N. P., **Chen, P. Y.**, Thang, M., Price, G., Cripps, A. W., & Cox, A. J. (2019). Selection of microbial biomarkers with genetic algorithm and principal component analysis. BMC bioinformatics, 20(Suppl 6), 413. https://doi.org/10.1186/s12859-019-3001-4

My contribution to the publish paper involved:

recruitment of participants and collection of data with my co-authors; reviewing the interpretations for the data analysis with my co-author; revising the drafts with my co-authors and approving the final manuscript.

Signed:



Pin-Yen Chen PhD Candidate School of Medical Science Griffith University	Dr. Ping Zhang Corresponding author Menzies Health Institute Queensland Griffith University	Prof. Allan Cripps Primary Supervisor School of Medicine Griffith University

Date:

13/08/2020	12/08/2020	12/08/2020

Zhang et al. BMC Bioinformatics 2019, 20(Suppl 6):413
https://doi.org/10.1186/s12859-019-3001-4

BMC Bioinformatics

RESEARCH Open Access

Selection of microbial biomarkers with genetic algorithm and principal component analysis

Ping Zhang[1*], Nicholas P. West[1,2], Pin-Yen Chen[1], Mike W. C. Thang[3], Gareth Price[3], Allan W. Cripps[1,4] and Amanda J. Cox[2]

From 2nd International Workshop on Computational Methods for the Immune System Function
Madrid, Spain. 3-6 December 2018

Abstract

Background: Principal components analysis (PCA) is often used to find characteristic patterns associated with certain diseases by reducing variable numbers before a predictive model is built, particularly when some variables are correlated. Usually, the first two or three components from PCA are used to determine whether individuals can be clustered into two classification groups based on pre-determined criteria: control and disease group. However, a combination of other components may exist which better distinguish diseased individuals from healthy controls. Genetic algorithms (GAs) can be useful and efficient for searching the best combination of variables to build a prediction model. This study aimed to develop a prediction model that combines PCA and a genetic algorithm (GA) for identifying sets of bacterial species associated with obesity and metabolic syndrome (Mets).

Results: The prediction models built using the combination of principal components (PCs) selected by GA were compared to the models built using the top PCs that explained the most variance in the sample and to models built with selected original variables. The advantages of combining PCA with GA were demonstrated.

Conclusions: The proposed algorithm overcomes the limitation of PCA for data analysis. It offers a new way to build prediction models that may improve the prediction accuracy. The variables included in the PCs that were selected by GA can be combined with flexibility for potential clinical applications. The algorithm can be useful for many biological studies where high dimensional data are collected with highly correlated variables.

Keywords: PCA, Genetic algorithm, Obesity, Biomarker

Background

Association between the human gut microbiome and a diverse range of health issues has been reported in a number of studies [1, 2]. Knight and colleagues [3] reviewed the methodological approach in microbiome studies, including: experimental design, choice of molecular analysis technology, methods for data analysis, and the integration of multiple -omics data sets. Different methods for surveying microbial communities include 16S ribosomal RNA, and metagenomic and metatranscriptomic sequencing. Next-step data analyses are needed to search for overall patterns in microbiome variation. The association between obesity and the gut microbiome from the phylum level to the species level has been studied and various results have been reported [4–6].

Several well-known sequence data analysis pipelines for microbiota study have been published, for example Quantitative Insights into Microbial Ecology (QIIME) [7], MetaGenome Rapid Annotation using Subsystem Technology (MG-RAST) [8] and mothur [9]. These packages include the functions of sequence alignment, operational taxonomic unit (OTU) identification, taxonomy classification, and alpha and beta diversity

* Correspondence: p.zhang@griffith.edu.au
[1]Menzies Health Institute QLD, Griffith University, Gold Coast, Australia
Full list of author information is available at the end of the article

calculation. They have been widely used for different biological and medical research purposes, such as associating gut microbiome diversity with diseases [10–13]. It is important to recognise that due to some possible pitfalls in sample processing, the abundance of specific bacterial species and overall community composition can be distorted, thus hampering the analysis and threatening the validity of the research findings [14]. In addition, a key limitation of using 16S rRNA gene analysis for genus and species level classification is that related bacterial species may be indistinguishable due to near identical 16S rRNA gene sequences [15]. The potential for different data analysis approaches to produce different outcomes has also been recognised. Plummer et al. [15] compared three pipelines commonly used for 16S rRNA gene analysis: QIIME, MG-RAST and mothur. Favourably, their results showed that the three pipelines produced comparable results for analysis of faecal samples, in terms of alpha diversity and usability. Although a difference was observed between the pipelines in terms of taxonomic classification of genera from the Enterobacteriaceae family, the three pipelines detected the same phylum in similar abundances. D'Argenio et al. [16] also compared QIIME and MG-RAST, and observed a statistically significant difference between these two bioinformatics pipelines with regards to beta diversity measures.

Despite the effort from researchers to develop high quality analytical pipelines, it is recognised that the complexity and variability of the human microbiome can be sensitive to various environmental factors [17]. Improvement of analytical pipelines has been complicated by the limitation of available sample material and the relatively high cost of the sequence analysis necessary for microbiome profiling. As a result, most microbiome studies have used limited sample sizes, raising questions regarding the accuracy of their findings. In addition to efforts to improve the accuracy of OTU detection and taxonomic classification, especially at the genus and species levels, researchers have been studying ways to characterise diseases based on microbial composition. Rather than simply associating diseases and individual microbial features, such as a phylum or species, studies have started looking at defining microbial signatures for specific diseases. This includes the application of computational modelling and variable selection techniques. For example, Rivera-Pinto et al. [18] presented a greedy stepwise algorithm for selection of microbial signatures that preserves the principles of compositional data analysis. Sze and Schloss [19] performed a meta-analysis on associations between specific microbiome-based markers and obesity, concluding that although there was support for a relationship between human faecal microbial communities and obesity status, this association was relatively weak and its detection is confounded by large interpersonal variation and insufficient sample sizes. The same study also tested random forest models for classifying individuals as obese on the basis of microbiome composition and did not find obvious patterns that could separate the obese and healthy groups. Random forest models were also used by Peters et al. [20] to identify taxonomic signatures of obesity. These models were evaluated with Receiver Operator Characteristic (ROC) curves and the area under the curve (AUC) value produced by the optimal model, which included 49 OTUs, was 0.81. When the repeated cross-validation was performed, the AUC value decreased to 0.65. Other machine learning methods used for microbiome studies have been reviewed by Knights et al. [21].

With the potential for large numbers of microbial species to be identified in human faecal samples and the high correlation between many of the species detected, principal components analysis (PCA) is often used. Studies use PCA to find characteristic patterns associated with certain diseases by reducing variable numbers based on their correlation with a principal component (PC), before a predictive model is built. The first two or three principle components account for the greatest proportion of the variance in the dataset. Usually, these components are then used to determine whether individuals can be clustered into one of two classification groups, control or diseased, based on pre-determined criteria. However, we have asked the following questions: (i) Is it possible that the proportion of variance captured by the first two or three PCs is unrelated to the disease groups, and that the variance explained by other components is able to better distinguish disease individuals from healthy controls? (ii) Are there different groups of bacterial species associated with individual obesity?

With these questions in mind, we developed a prediction method that combines PCA and a genetic algorithm (GA) for microbial biomarkers identification. We applied this approach to faecal microbial data collected from our obesity study, to identify potential sets of bacterial species that may be associated with obesity with metabolic syndromes (MetS). The preliminary work has been presented in the 2018 IEEE International Conference on Bioinformatics and Biomedicine [22].

Methods

Principal components analysis

PCA is often used as a tool in exploratory data analysis for variable dimensionality reduction prior to building predictive models. It can be used to reduce a large number of predictor variables to a few PCs, particularly in datasets that are noisy or have strongly correlated explanatory variables. The PCs can then be used to build predictive models. The PCs are the linear combinations

of the original variables that account for variance in the data. PCA can be performed using either eigenvalue decomposition of a data covariance matrix or singular value decomposition of a data matrix. The coefficients corresponding to each variable in the linear combinations indicate the relative weight of the variable in the component. The larger the absolute value of the coefficient, the more important the corresponding variable is in calculating the component. To make the coefficient value for each variable comparable, the data should be normalized to have the same unit of measurement before PCA is used.

Genetic algorithms

GA is a search heuristic to find optimal solutions by mimicking Charles Darwin's theory of natural evolution--fittest individuals are selected for reproduction for the next generation. In GA, the potential solutions compete and mate with each other to produce increasingly fitter individuals over multiple generations.

GAs can be useful and efficient when searching for the best combination of variables to achieve the best outcome (e.g. accuracy of prediction). GAs have been developed and applied for biomarker profile identification in a range of settings such as Alzheimer's disease progression and breast cancer diagnosis [23–25]. The GAs have also been modified and improved to adapt to different computational environments and for different applications [26, 27]. Carter et al. [28] applied GA to their study to select vaginal microbiome features associated with bacterial vaginosis. However, the actual features were not reported, as authors explained that evaluation was needed from both microbial and clinical perspectives in the future.

In this study, GA will be used to find the best subset of principal components produced from a PCA using gut microbial species data.

Proposed method

The method described here uses normalized OTU abundance with taxonomy assigned across the sample as the input for PCA. The OTUs can be identified by any of the sequence analysis pipelines mentioned above or other software packages, such as "DADA2" [29] in R (https://cran.r-project.org/). GA is then applied for selection of the set of components created from the PCA that best predict individuals as obese or healthy weight. The scores of selected PCs calculated for each individual are used as the input for building a classification model. ROC curve analysis is used to evaluate the classification models and is used as the fitness function for the GA. The method is shown diagrammatically in Fig. 1. In this research, logistic regression (LR) is used for building the classification models and more details about how to implement the GA can be found in reference [24].

Experiments and results

In this study, faecal samples from 22 obese and 105 healthy-weight subjects were collected and sequenced using a 16S-based approach. The obese sample here was designed as those with body mass index (BMI) over 30 and with MetS [10]. The healthy-weight subjects included 39 recreational individuals and 66 athletes who were involved in rugby, football soccer, judo, rowing, triathlon or weightlifting. For sequencing analysis, paired-end reads were merged using the PEAR software (v0.9.6) [30]. Contaminant human reads were removed by mapping to the hg19 human genome using BWA software package (v0.7.12) [31] and the remaining reads were searched against the Greengenes 16S taxonomy database (GG v13.5) [32] using sequence analysis tool VSEARCH (v1.9.7) [33] to generate a single OTU raw count/abundance table for all 127 subjects. Amongst the 127 subjects 68,590 OTUs were identified (at all taxonomic levels), which mapped at the level of species to 163 observations, from Greengenes total reportable content of 3093 species. Species with low diversity across the cohort were filtered from future analysis, this was achieved by removing the species with zero abundance in 80% of both healthy and obese subjects. This excluded 126 species (77.3%) of the data leaving 37 species for further analysis. The abundance values of each of these species were normalized to the range of [0, 1] (highest abundance across the individuals as 1 and the lowest as 0) before applying the proposed method which combines PCA and GA for identifying obese from healthy subjects. The results were compared with those produced without GA and with those produced by using GA to select

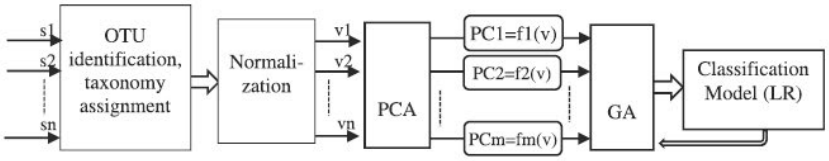


Fig. 1 A diagram of proposed method. s1, s2...sn are the 16S rRNA sequences for this study (can be from other sequencing). v1 to vn are the abundance (normalized) of each species detected in each individual. m = number of PCs created by PCA, *n* = number of individuals included in the sample. PCA is used to produce PC scores for each individual, and GA is used to select the best subset of PCs to distinguish obesity from healthy cases

combinations of bacterial species for heathy and obese classification without PCA in the model.

GA models to select combination of PCs for classification

For experiments, we performed PCA across three circumstances using: the whole dataset, obese only sample, and healthy weight only sample. This approach is based on the possibility that for different populations, the correlation between species might differ. The function "prcomp" from the "stats" package in R [34] was used to create the PCs and calculate the scores for each individual. These scores were then used to build the classification model with GA to select the best components for identifying obese from healthy subjects. The algorithm used by "prcomp" for creating the PCs can be found in reference [35]. Essentially, the PC calculation is performed by a singular value decomposition of the data matrix. If there are n observations with p variables, then the number of distinct PCs is $min(n,p)$.

GA was completed with the fitness function of the cross-validated AUC value created from the logistic regression model. More explanation about AUC can be found in Johnson et al. [24]. Constraints for GA were set to include 1 to 6 PCs in the classification model. Ten-times repeated five-fold cross-validation was used for testing the classification model with selected PCs. With each data set (all, healthy or obese), GA was run 100 times repeatedly. The PC sets that were selected the most in the repeated runs were chosen as the final result. From the results (Table 1) it can be seen that the selection from GA was quite consistent with slight variation from each run.

The PCA constructed from the whole data set and healthy-weight subjects both created 37 principal components (PC1 to PC37) while the PCA from obese subjects created 22 components (PC1 to PC22). Table 1 lists the sets of PCs selected by GA and the cross-validated AUC produced from each prediction model built with the selected PC(s). The symbols "+" or "-" following the PC numbers indicate whether the coefficient of this PC is positive or negative in the corresponding classification model. Positive coefficient means that an increased score of this PC will increase the probability of the individual being characterised as obese. For example, PC1+ represents that the first PC created from the species abundance data will have a positive contribution to obesity with MetS.

Table 2 lists the top five species that have the highest contribution to each PC selected by GA. The symbols "+" or "-" following the species names indicate whether it has positive or negative contribution to the corresponding PC. For example, Prausnitzii- within column Comp1 represents that Prausnitzii has negative correlation with Comp1 (PC1 for Whole, PC14 for obese, and PC1 for Healthy). That suggests that increased Prausnitzii abundance will decrease the Comp1 value. As Comp1 has a positive correlation with being overweight, it can be speculated that increased Prausnitzii abundance leads to decreased likihood of being obese.

From the results presented in Table 1 and Table 2 each of the species were analysed and categorized into two groups; positive (indicated with an asterisk (*) in Table 2) or negative correlations with the probability of having healthy body mass. The combination of having any one of the microbial species from each column can be a set of species that can have a high impact on health. For example, based on the results from the first set of the experiments which ran PCA on the whole dataset, either "Prausnitzii, Faecis, Eutactus, Lenta, Eggerthii and Zeae", "Formicigenerans, Faecis, Eutactus, Lenta, Eggerthii and Zeae" or "Prausnitzii, Formicigenerans, Faecis, Eutactus, Lenta, Eggerthii and Zeae" can be a combination to have a

Table 1 GA selected PCs and the classification model performance (ROC)

Data for creating PCA	Result	Model _6 PCs	Model _5 PCs	Model _4 PCs	Model _3 PCs	Model _2 PCs	Model _1 PC
All	PCs selected	PC1+, PC2–, PC7+, PC11+, PC15–, PC27–	PC1+, PC2–, PC7+, PC11+, PC27–	PC1+, PC2–, PC7+, PC27–	PC1+, PC2–, PC7+	PC1+, PC7+ (or PC2–)	PC1+
	AUC (CV)	0.87	0.85	0.84	0.81	0.77	0.69
Obesity	PCs selected	PC2–, PC4–, PC14+, PC16–PC18+, PC19–	PC2–, PC4–, PC14+, PC18+ PC19–	PC2–, PC4–, PC14+, PC18+	PC2–, PC14+, PC18+	PC14+, PC18+	PC14+
	AUC (CV)	0.92	0.92	0.90	0.87	0.84	0.80
Healthy	PCs selected	PC1+,PC3+, PC5–,PC23+, PC28–,PC34+	PC1+,PC3+, PC23+,PC28–PC34+	PC1+,PC23+, PC28–, PC34+	PC1+,PC23+, PC34+	PC1+,PC34+	PC1+
	AUC (CV)	0.92	0.90	0.88	0.87	0.83	0.72

+ Positive correlation coefficient in the model
– Negative correlation coefficient in the model

Table 2 Top species included in the GA selected 1, 2, 3, 4, 5 or 6 PCs produced with different data sets

Dataset for creating PCA	High contribution variables (high coefficients in the corresponding PC) included in the most selected components					
	Comp1	*Comp2*	*Comp3*	*Comp4*	*Comp5*	*Comp6*
Whole (PC1, PC7, PC2, PC27, PC11, PC15)	Prausnitzii–[a]	Gnavus+	Eutactus+[a]	Moorei–	Eggerthii–[a]	Zeae+[a]
	Eutactus–[a]	Faecis–[a]	Prausnitzii+[a]	Obeum–	Dispar–[a]	Gnavus–
	Formicigenerans–[a]	Copri+	Aerofaciens–	Lenta+[a]	Adolescentis+	Stutzeri+[a]
	Catus–[a]	Muciniphila–[a]	Catus–	Animalis–	Mucilaginosa–[a]	Bromii+[a]
	Faecis–[a]	Adolescentis–[a]	Adolescentis–	Torques–	Aerofaciens+	Fragilis+[a]
Obesity (PC14, PC18, PC2, PC4, PC19, PC16)	Eutactus–[a]	Uniformis+	Dolichum–	Producta–	Caccae+[a]	Formicigenerans+[a]
	Bromii+	Catus–[a]	Lenta–	Prausnitzii+[a]	Parainfluenzae+[a]	Bromii–
	Adolescents–[a]	Dispar+	Aerofaciens+[a]	Aerofaciens–	Formicigenerans+[a]	Distasonis–
	Formicigenerans+	Faecis+	Producta–	Fragilis–	Adolescentis–	Eutactus+[a]
	Producta–[a]	Distasonis–[a]	Gnavus–	Faecis+[a]	Dispar–	Perfringens+[a]
Healthy (PC1, PC34, PC23, PC28, PC3, PC5)	Prausnitzii–[a]	Stutzeri–[a]	Callidus–[a]	Ovatus–	Copri+	Copri+[a]
	Eutactus–[a]	Zeae+	Moorei+	Longum+[a]	Muciniphila–[a]	Muciniphila+[a]
	Catus–[a]	Gnavus+	Formigenes+	Distasonis+[a]	Formigenes–[a]	Prausnitzii–
	Formicigenerans–[a]	Dispar+	Prausnitzii+	Fragilis–	Catus+	Formigenes+[a]
	Faecis–[a]	Lenta–[a]	Catus–[a]	Aerofaciens–	Biforme+	Eutactus+[a]

Comp1, Comp2, Comp3, Comp4, Comp5 and Comp6 represent the 6 PCs selected by GA. For experiment with whole dataset they are PC1, PC7, PC2, PC27, PC11 and PC15 respectively; for experiment with obesity sample, they are PC14, PC18, PC2, PC4, PC19 and PC16; for experiment with healthy sample, they are PC1, PC34, PC23, PC28, PC3 and PC5

[a]Species has a positive correlation with the probability of having healthy body mass

+ Positive correlation with the corresponding PC

– Negative correlation with the corresponding PC

potential benefit on health. On the other hand, high values for Gnavus, Catus, Moorei and Aerofaciens together are associated with high probability with of being obese.

A final classification model was built with each set of PCs selected by GA or first 1 to 6 PCs (which explain the most variance of the data) from the PCA. Again, the PCs were calculated from the whole dataset, healthy-weight dataset or obese dataset. The AUCs produced from the GA-selected PCs were quite obviously higher than the ones from the top PCs of PCA. Figure 2 shows

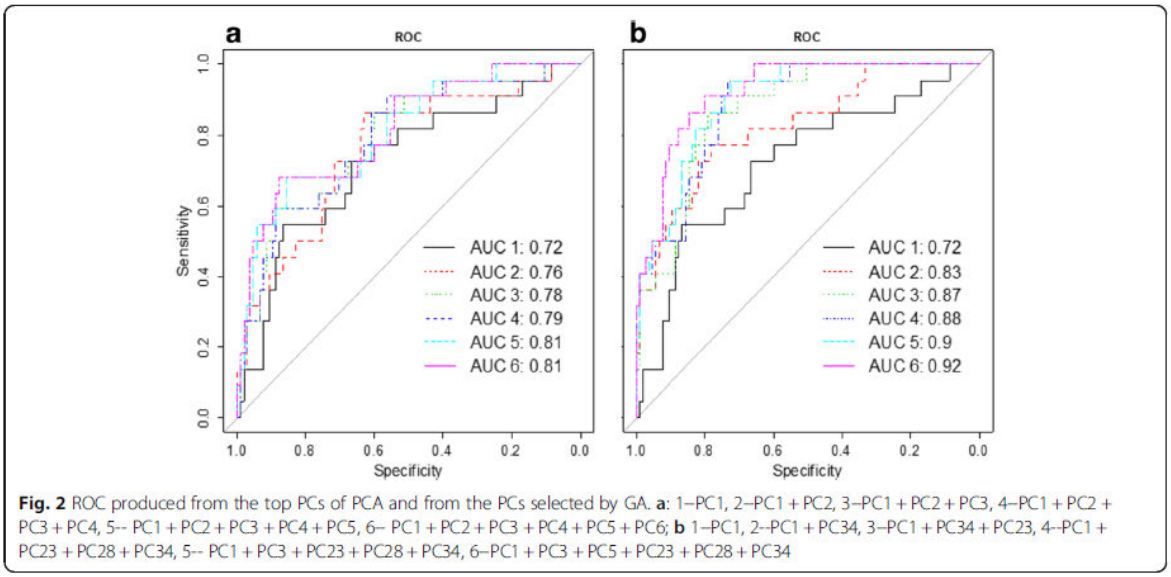


Fig. 2 ROC produced from the top PCs of PCA and from the PCs selected by GA. **a**: 1–PC1, 2–PC1 + PC2, 3–PC1 + PC2 + PC3, 4–PC1 + PC2 + PC3 + PC4, 5-- PC1 + PC2 + PC3 + PC4 + PC5, 6– PC1 + PC2 + PC3 + PC4 + PC5 + PC6; **b** 1–PC1, 2--PC1 + PC34, 3–PC1 + PC34 + PC23, 4--PC1 + PC23 + PC28 + PC34, 5-- PC1 + PC3 + PC23 + PC28 + PC34, 6–PC1 + PC3 + PC5 + PC23 + PC28 + PC34

the ROCs created from the models built with the selected PCs and the first PCs of the PCA. The PCs in the graph were calculated with the healthy-weight dataset (when compared with the result from the PCs calculated from whole dataset and obese dataset, the first PCs from the healthy data produced the highest AUC values).

GA models to select sets of species for classification

To compare the results from the model that combined PCA and GA and from the model where GA was applied directly for selection of the combination of the bacterial species, GA was implemented in conjunction with logistic regression using the species abundance directly as the input for classification.

For experiments, the number of species (number of input variables for logistic regression) was restricted to maximum six, which was the same as the maximum number of PCs used in the earlier experiments. Table 3 shows the combinations of the bacterial species selected by 100 repeated runs of GA, which achieved the highest AUC values. It can be seen that some of the species were commonly selected in different sets of the selections. The selection frequency of each species from the 100 repeated GA runs was calculated and a frequency chart showing the top 10 most selected species was drawn in Fig. 3. Eutactus and Gnavus appeared in the final selection of almost every run of the GA (96 out of 100 runs and 95 out of 100 runs). Muciniphila, Distasonis and Prausnitzii were also selected frequently (> 50% frequency) in the repeated GA runs. These highly selected bacterial species appeared to have relatively high contribution to the selected PCs shown in the previous section.

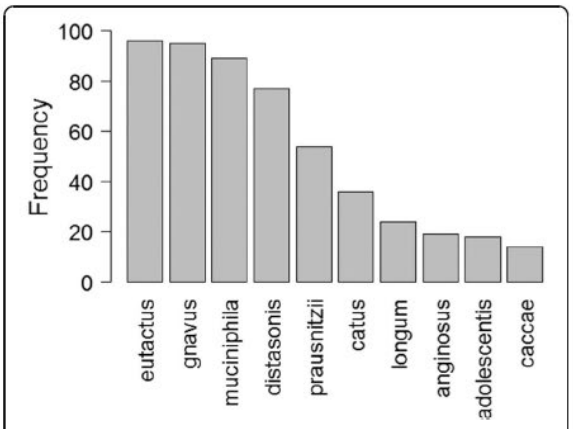


Fig. 3 Frequencies of the species selected from the multiple runs of GA. The GA was run with logistic regression as the classification model and species abundance as the input. GA was used to select the combination of the species for classification of obese individuals from the health group. The number on top of each bar is how many times out of 100 GA runs the corresponding species was selected

Discussion

In this study, a computational method that combines PCA and GA has been proposed to produce accurate prediction result and to find sets of features (variables) that contribute the most to the prediction models. The model was applied to identify sets of bacterial species associated with high body mass. Due to the high correlation between many species of the gut bacteria, constructing PCA before the GA can improve the efficacy of GA for selecting multiple sets of microbial species associated with obesity and MetS. The result from this study showed that the prediction models built with the PCs selected by GA produced much higher AUC values than the models built with the top PCs that explained the greatest proportion of the variance in the sample. The results were also compared with those produced from the GA selected models with bacterial species abundance values as the input directly, and it showed its advantages.

In the microbiome study, the results produced from the described method depends on the accuracy of the sequencing analysis. The microbial species identified here was based on the sample of 22 obese subjects and 105 healthy-weight subjects. Assuming this result was validated in multiple datasets with bigger sample sizes, the results from Table 1 and Table 2 can suggest a few combinations of microbial species groups that are

Table 3 Sets of species selected by GA using the species abundance as the input variables of logistic regression models

GA Selected Species						AUC
Adolescentis	Catus	Eutactus	Gnavus	Muciniphila	Prausnitzii	0.87
Adolescentis	Distasonis	Eutactus	Gnavus	Muciniphila	Prausnitzii	0.87
Aerofaciens	Distasonis	Eutactus	Gnavus	Longum	Muciniphila	0.88
Anginosus	Distasonis	Eutactus	Gnavus	Muciniphila	Prausnitzii	0.87
Catus	Distasonis	Eutactus	Gnavus	Muciniphila	Prausnitzii	0.88
Catus	Eutactus	Gnavus	Longum	Muciniphila	Prausnitzii	0.86
Distasonis	Eutactus	Gnavus	Longum	Muciniphila	Prausnitzii	0.88

GA Selected Species lists the set of species selected by GA, each row one set. AUC is the area under the ROC curve produced by the corresponding logistic regression model with the selected set of species. The result was cross validated with the same cross validation set up as the earlier experiments

beneficial to health. Some of the species in the combinations can be replaced by equivalent alternative species that are suggested by the algorithm, which gives flexibility for further intervention. As described in the previous sections, the bacterial species detected can be different when applying different sequencing analysis and taxonomy classifications. To validate the findings from this study, the presented algorithm should be run with the outcomes from metagenomics sequencing and with other sequencing analysis pipelines. Different reference databases (e.g. NCBI) can also be used for taxonomy classification of the OTUs identified.

Conclusion

This study demonstrated the value of applying GA for selection of subsets of PCs from PCA to improve the performance of prediction models. The features included in the PCs that were selected by GA can be combined with flexibility for potential clinical applications. With the flexible options of combining the features included in the PCs selected by the GA, different interventions can be recommended for different patients, which contributes to the practice of personalised medicine. The proposed algorithm was designed in a general way and was tested in a study comparing obese individuals with MetS and healthy-weight subjects. It can be applied for any other classification or biomarker identification study. The model takes into account correlations of the variables (bacteria species in this study) and the advantages of GA for feature selection. It overcomes the limitations of the ways in which PCAs are currently used for prediction modelling. The algorithm can be useful for many biological studies where high dimensional data are collected with strongly correlated variables.

Abbreviations

AUC: Area under the ROC; BMI: Body mass index; GA: Genetic algorithm; MetS: Obesity with metabolic syndromes; OTU: Operational taxonomic unit; PC: Principal component; PCA: Principal component analysis; ROC: Receiver operator characteristic curve

Acknowledgements

The authors would like to thank the facility support from Queensland Facility for Advanced Bioinformatics (https://qfab.org/) and the participants of this study for their value contributions. Salary support for PZ and AJC was provided by the Griffith University Area of Strategic Investment in Chronic Disease Prevention. The microbial compositional profiling data used was generated as part of projects funded by the Australian Institute of Sport and the Gold Coast Hospital Foundation. Part of this work has previously been presented at conference (IEEE International Conference on Bioinformatics and Biomedicine (BIBM)).

About this supplement

This article has been published as part of BMC Bioinformatics Volume 20 Supplement 6, 2019: Towards computational modeling on immune system function. The full contents of the supplement are available online at https://bmcbioinformatics.biomedcentral.com/articles/supplements/volume-20-supplement-6.

Authors' contributions

PZ designed the methodology, performed the data analysis and computational experiments, and drafted the paper. MT and GP contributed to the 16S sequencing analysis and OTU identification. AWC, NPW and AJC designed the obesity study and contributed to participant recruitment and data collection. PY assisted for data interpretation and revised the draft of the manuscript. All authors read the paper, made comments, and agreed with the content. All authors read and approved the final manuscript.

Funding

This project was supported by Griffith Health Institute/Gold Coast Hospital Foundation. The content is solely the responsibility of the authors and does not necessarily represent the official views of Griffith Health Institute/Gold Coast Hospital Foundation. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. Publication costs are funded by Menzies Health Institute QLD, Griffith University, Australia.

Availability of data and materials

Data are available upon request from the Menzies Health Institute Queensland for researchers who meet the criteria for access to confidential data.

Ethics approval and consent to participate

This study was approved by Bond University Human Research Ethics Committee (0000015530) and Griffith University Human Research Ethics Committee (GU 2016/213 and GU 2015/229). Consent in writing was obtained from all participants who provided samples.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Menzies Health Institute QLD, Griffith University, Gold Coast, Australia. [2]School of Medical Science, Griffith University, Gold Coast, Australia. [3]QFAB Bioinformatics, Institute for Molecular Bioscience, University of Queensland, Brisbane, Australia. [4]School of Medicine, Griffith University, Gold Coast, Australia.

Received: 12 June 2019 Accepted: 18 July 2019
Published: 12 December 2019

References

1. Jackson MA, Verdi S, Maxan ME, et al. Gut microbiota associations with common diseases and prescription medications in a population-based cohort. Nat Commun. 2018;9(1):2655.
2. Gilbert JA, Blaser MJ, Caporaso JG, et al. Current understanding of the human microbiome. Nat Med. 2018;24:392–400.
3. Knight R, Vrbanac A, Taylor BC, et al. Best practices for analysing microbiomes. Nat Rev Microbiol. 2018;16(7):410–22.
4. Ottosson F, Brunkwall L, Ericson U, et al. Connection between BMI-related plasma metabolite profile and gut microbiota. J Clin Endocrinol Metab. 2018;103(4):1491–501.
5. Million M, Lagier JC, Yahav D, et al. Gut bacterial microbiota and obesity. Clin Microbiol Infect. 2013;19(4):305–13.
6. Chakraborti CK. New-found link between microbiota and obesity. World J Gastrointest Pathophysiol. 2015;6(4):110–9.
7. Caporaso JG, Kuczynski J, Stombaugh J, et al. QIIME allows analysis of high-throughput community sequencing data. Nat Methods. 2010. https://doi.org/10.1038/nmeth.f.303.
8. Keengan KP, Glass EM, Meyer F. MG-RAST, a metagenomics Service for Analysis of microbial community structure and function. Methods Mol Biol. 2016;1399:207–33. https://doi.org/10.1007/978-1-4939-3369-3_13.
9. Schloss PD, Westcott SL, Ryabin T, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol. 2009;75(23):7537–41.

10. Han GG, Lee JY, Jin JD, et al. Evaluating the association between body weight and the intestinal microbiota of weaned piglets via 16S rRNA sequencing. Vet Microbiol. 2016;196:55–62.
11. Clemente J, Ursell L, Parfrey L, et al. The impact of the gut microbiota on human health: an integrative view. Cell. 2012;148(6):1258–70.
12. Spencer M, Hamp T, Reid R, et al. Association between composition of the human gastrointestinal microbiome and development of fatty liver with choline deficiency. Gastroenterology. 2011;140(3):976–86. https://doi.org/10.1053/j.gastro.2010.11.049.
13. Zhong L, Shanahan ER, Raj A, et al. Dyspepsia and the microbiome: time to focus on the small intestine. Gut. 2016. https://doi.org/10.1136/gutjnl-2016-312574.
14. Brooks JP, Edwards DJ, Harwich MD, et al. The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies. BMC Microbiol. 2015;15:66. https://doi.org/10.1186/s12866-015-0351-6.
15. Plummer E, Twin J, Bulach DM, et al. A comparison of three bioinformatics pipelines for the analysis of preterm gut microbiota using 16S rRNA gene sequencing data. J Proteomics Bioinformatics. 2015;8:283–91. https://doi.org/10.4172/jpb.1000381.
16. D'Argenio V, Casaburi G, Precone V, et al. Comparative metagenomic analysis of human gut microbiome composition using two different bioinformatic pipelines. Biomed Res Int. 2014;2014:325340. https://doi.org/10.1155/2014/325340.
17. Huttenhower C, Gevers D, Knight R, et al. Structure, function and diversity of the healthy human microbiome. Nature. 2012;486:207–14.
18. Rivera-Pinto J, Egozcue JJ, Pawlowsky-Glahn V, et al. Balances: a new perspective for microbiome analysis. mSystems. 2018;3(4). https://doi.org/10.1128/mSystems.00053-18.
19. Sze M, Schloss P. Looking for a signal in the noise: revisiting obesity and the microbiome. mBio. 2016;7(4):e01018-16. https://doi.org/10.1128/mBio.01018-16.
20. Peters BA, Shapiro JA, Church TR, et al. A taxonomic signature of obesity in a large study of American adults. Sci Rep. 2018;8:9749. https://doi.org/10.1038/s41598-018-28126-1.
21. Knights D, Costello EK, Knight R. Supervised classification of human microbiota. FEMS Microbiol Rev. 2011;35:343–59.
22. Zhang P, West N, Chen P, Cripps A, Cox A. Combination of principal component analysis and genetic algorithm for microbial biomarker identification in obesity. Madrid: Proceedings of IEEE International Conference on Bioinformatics and Biomedicine (BIBM); 2018.
23. Zhang P, Verma B, Kumar K. Neural vs. statistical classifier in conjunction with genetic algorithm based feature selection. Pattern Recogn Lett. 2003;26(7):909–19.
24. Johnson P, Vandewater L, Wilson L, et al. Genetic algorithm with logistic regression for prediction of progression to Alzheimer's disease. BMC Bioinformatics. 2015;15:S11.
25. Zhang P, Kumar K, Verma B. A hybrid classifier for mass classification with different kinds of features in mammography. LNCS. 2005;3614:316–9.
26. Khan M, Mendes A, Zhang P, et al. Evolving multi-dimensional wavelet neural networks for classification using Cartesian genetic programming. Neurocomputing. 2017;247:39–58.
27. Vandewater L, Brusic V, Wilson W, et al. An adaptive genetic algorithm for selection of blood-based biomarkers for prediction of Alzheimer's disease progression. BMC Bioinformatics. 2015;16(18):S1.
28. Carter J, Beck D, Williams H, et al. GA-based selection of vaginal microbiome features associated with bacterial vaginosis. Genet Evol Comput Conf. 2014;2014:265–8.
29. Callahan B, McMurdie P, Rosen M, et al. DADA2: high-resolution sample inference from Illumina amplicon data. Nat Methods. 2016;13:581–3. https://doi.org/10.1038/nmeth.3869.
30. Zhang J, Kobert K, Flouri T, et al. PEAR: a fast and accurate Illumina paired-end reAdmergeR. Bioinformatics. 2014;30:614–20.
31. Li H, Durbin R. Fast and accurate short read alignment with burrows-wheeler transform. Bioinformatics. 2009;25:1754–60.
32. DeSantis T, Hugenholtz P, Larsen N, et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol. 2006;72(7):5069–72.
33. Rognes T, Flouri T, Nichols B, et al. VSEARCH: a versatile open source tool for metagenomics. Peer J. 2016;4:e2584.
34. R Core Team. R: a language and environment for statistical computing. Vienna: R Foundation for statistical computing; 2018. URL https://www.R-project.org/ (Accessed on 20 Jul 2018)
35. Mardia KV, Kent JT, Bibby JM. Mutivariate analysis. London: Academic; 1979.

Publisher's Note

4.7 References

[1] Heaton, J., *Introduction to Neural Networks with Java*, ed. M. McKinnis. 2008: Heaton Research, Inc.

[2] Rastogi, D., Suzuki, M., and Greally, J.M., Differential epigenome-wide DNA methylation patterns in childhood obesity-associated asthma. Sci Rep. vol. 3, 2013.

[3] Li, Y.-X., Li, B.-Z., and Yan, D.-Z., Upregulated expression of human cathelicidin LL-37 in hypercholesterolemia and its relationship with serum lipid levels. Mol Cell Biochem. vol. 449, pp. 73-9, 2018.

[4] Huang, X., Liu, G., and Su, Z., The PI3K/AKT pathway in obesity and type 2 diabetes. Int J Biol Sci. vol. 14, pp. 1483-96, 2018.

[5] Ma, K.L., et al., Activation of the CXCL16/CXCR6 pathway promotes lipid deposition in fatty livers of apolipoprotein E knockout mice and HepG2 cells. Am J Transl Res. vol. 10, pp. 1802-16, 2018.

[6] Tourniaire, F., et al., Chemokine Expression In Inflamed Adipose Tissue Is Mainly Mediated By NF-κB. PLoS One. vol. 8, 2013.

[7] Jung, U.J., et al., Differences in metabolic biomarkers in the blood and gene expression profiles of peripheral blood mononuclear cells among normal weight, mildly obese and moderately obese subjects. Br J Nutr. vol. 116, pp. 1022-32, 2016.

[8] Wieser, V., et al., Adipose type I interferon signalling protects against metabolic dysfunction. Gut. vol. 67, pp. 157-65, 2016.

[9] Koliada, A., et al., Association between body mass index and Firmicutes/Bacteroidetes ratio in an adult Ukrainian population. BMC Microbiol. vol. 17, pp. 1-6, 2017.

[10] Chen, M.-S. and Chen, S.-H., A data-driven assessment of the metabolic syndrome criteria for adult health management in Taiwan. Int J Environ Res Public Health. vol. 16, pp. 1-11, 2019.

[11] AlJarullah, A.A., *Decision tree discovery for the diagnosis of type II diabetes*, in *2011 International Conference on Innovations in Information Technology*. 2011: Abu Dhabi. pp. 303-307.

[12] Al-Thani, M.H., et al., Prevalence and determinants of metabolic syndrome in Qatar: results from a National Health Survey BMJ Open. vol. 6, pp. 1-10, 2016.

[13] Mao, L., et al., Metabolic syndrome in Xinjiang Kazakhs and construction of a risk prediction model for cardiovascular disease risk. PLoS One. vol. 14, pp. 1-14, 2018.

[14] Worachartcheewan, A., et al., Identification of metabolic syndrome using decision tree analysis. Diabetes Res Clin Pract. vol. 90, pp. e15-18, 2010.

[15] Karimi-Alavijeh, F., Jalili, S., and Sadeghi, M., Predicting metabolic syndrome using decision tree and support vector machine methods. ARYA Atheroscler. vol. 12, pp. 146-152, 2016.

[16] Alić, B., et al. *Classification of metabolic syndrome patients using implemented expert system*. 2017. Singapore: Springer Singapore.

[17] Meng, X.-H., et al., Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. Kaohsiung J Med Sci. vol. 29, pp. 93-99, 2013.

CHAPTER 5
Combination of artificial neural network with genetic algorithm for the prediction of MetS

5.1 Abstract

Genetic algorithm (GA) is a popular feature selection method that is useful in many studies, particularly for the prediction of metabolic syndrome (MetS) which is a multifactorial disease. Studies that have used classification models for the prediction of diseases have often found the performance of the models to be hindered by the biomarkers used for the construction of the models. The ability of GA to identify the best combination of biomarkers for predicting diseases is therefore very advantageous in this space. The current study combined GA with artificial neural network (ANN) to determine whether the hybrid model was able to improve the performance of four individual classification models: logistic regression (LR), decision tree (DT), support vector machine (SVM) and ANN. The variables that were used for MetS prediction were haematological measures, gene expression levels and gut microbial count. The results found the hybrid GA with ANN model to achieve extremely high area under the curve (AUC) values of 0.992 and 0.972 when predicting MetS using haematological measures and gut microbial counts, respectively. The hybrid model was also able to improve the performance of the independent ANN model when predicting MetS using gene expression levels. Additionally, the ability of the hybrid model to identify the best combination of biomarkers for MetS prediction makes it an important tool for consideration in clinical research looking to reduce the incidence of MetS.

5.2 Introduction

Metabolic syndrome (MetS) is a multifactorial disease that affects many different biomarkers across multiple body systems, leading to an increased risk of chronic diseases, including type 2 diabetes mellitus (T2DM) and cardiovascular disease (CVD). Research in MetS has, until recently, been hindered by the complexity of molecular interactions, both genetic and environmental. While recent technological advances have allowed researchers to simultaneously analyse biomarkers from different body systems, they also pose a challenge on their own. Classification models have become more popular for MetS prediction, however, its predictive ability also depends on the parameters used for its construction. Previous studies have discovered the importance of using the right parameters for the prediction of diseases. A comparison of the health parameters used by different studies for the construction of prediction models and their performance as a result can be found in Section 2.6, page 41. To assist the process of choosing the most appropriate biomarkers, researchers have also looked into the use of feature selection techniques, including genetic algorithm (GA). Charles Darwin's theory of evolution is the basis of GA as the best variables in each generation are used to find the optimal combination of biomarkers for the prediction of diseases. Many studies have used GA to identify the best combination of biomarkers to use as an input for prediction models, such as artificial neural network (ANN). Due to MetS research and the use of combining GA with ANN being relatively new, no studies were found to have used this hybrid model for the prediction of MetS. On the other hand, studies have used the combined method for predicting the MetS-related disease, diabetes. Karegowda, Manjunath and Jayaram [1] used the model to predict diabetes prevalence in 392 Pima Indian individuals. The study used eight biomarkers: plasma glucose, diastolic blood pressure, triceps skinfold thickness, serum insulin, body mass index (BMI), diabetes pedigree function and age. Using GA, the best combination of features to achieve the highest prediction accuracy included four of the eight features: plasma glucose,

serum insulin, BMI and age. While the use of fewer features may be expected to yield a lower classification accuracy, the hybrid model achieved an accuracy of 84.7%, which was an improvement from the 77.7% achieved using all eight biomarkers. In a more recent study, the hybrid model was also used for the prediction of diabetes. Mortajez and Jamshidinezhad [2] also reported an increase in classification accuracy from 68% using a simple neural network to 84.5% from the hybrid GA and ANN model. However, Mortajez and Jamshidinezhad did not report the features selected by the model. Both studies demonstrated the ability of feature selection techniques to improve the prediction of diabetes prevalence. The use of the hybrid GA and ANN model is therefore expected to also achieve a high accuracy in MetS prediction, improving the performance of individual prediction models.

5.3 Research design

5.3.1 Study design

Genetic algorithm was used as a feature selection technique to identify the best combination of biomarkers for the prediction of MetS. The biomarkers that were used were measurements taken from 152 participants, classified as either healthy weight controls or obese with MetS individuals. The measurements that were taken from each participant were split into three different variable groups: haematological measures, gene expression levels and gut microbial composition. Section 3.3.2, page 75 provides a detailed explanation of how the measurements were obtained.

5.3.2 Genetic algorithm

An overview of the five different steps of GA can be found in Section 2.7, page 58. The current study generated an initial population with a size of 50 chromosomes, each representing a

combination of biomarkers. A chromosome is a string of binary values, with '0' meaning the particular variable was not included in the combination, while '1' meant the variable was included. The chromosomes were randomly generated with a fixed maximum number of '1's that were allowed. The maximum number for each variable group was the sample size of said variable group divided by 10 and rounded down. The remaining number of variables then equated to the number of '0' used for that particular chromosome. The fitness value for each chromosome was evaluated using a custom fitness function. Logistic regression (LR) was used to evaluate the fitness function by using each chromosome as an input. The calculated area under the curve (AUC) value from the LR model was then used as the fitness function. Genetic algorithms were then built using the *GA* package in R. The values for the parameters set were crossover rate of 0.9, mutation rate of 0.1, maximum iteration of 100 and elitism selection of 1. The chromosome with the highest fitness value identified by GA was then used as the input for ANN.

5.3.3 Genetic algorithm with artificial neural network

The method for ANN was as described in Section 4.3.6, page 124 with changes to the number of variables and hidden layer neurons (HLN) used. As GA is a feature selection technique, all the variables from the full dataset was used for consideration when identifying the best combination of variables for MetS prediction. Consequently, biomarkers were not removed if they were strongly correlated with another variable. The total number of biomarkers for each variable group were then: 19 for haematological measures, 43 for gene expression levels and 51 gut microbial species. The rules-of-thumbs established by Heaton [3] were again used as a guideline to the number of HLNs that should be used to prevent overfitting. The range of HLNs used for each variable group were 3 to 14 for haematological measures, 3 to 30 for gene expression levels and 3 to 35 for gut microbial species.

5.4 Results

Every number within the HLN size range calculated for all three variable groups were used to construct ANN models. The result for each HLN size is shown in Table 5.1. As with the results of the independent ANN model, the optimal HLN size was the one with a high AUC value and has a great predictive ability overall. The optimal HNL size for haematological measures was 12 as it had a high AUC value of 0.992 and the highest sensitivity values of 0.984 and 0.9 in the training and testing sets, respectively. In a similar fashion, an HLN size of 28 was chosen for gene expression levels as it had the highest sensitivity value in the testing set, with a value of 0.567. Lastly, the optimal HLN size that was decided for gut microbial counts was 20 as the AUC value was high, 0.972, and the sensitivity value was 0.85. Although the sensitivity value of 0.85 was also reached by the HLN size 19, there was a considerable difference in both the AUC and training set sensitivity values compared to HLN size 20. On the other hand, despite HLN size 35 having the highest testing set sensitivity value of 0.9, the differences in the performance of the two models were not great enough to risk overfitting. The chosen optimal HLN sizes for each variable group were then used for further analysis.

Table 5.1. Comparison of the averaged best performing hybrid genetic algorithm with neural networks using different hidden layer neuron sizes.

Variable group	Hidden layer neuron size	Training set			Testing set			AUC
		Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	
Haematological measures	3	0.960	0.913	0.976	0.892	0.833	0.910	0.972
	4	0.969	0.935	0.980	0.885	0.733	0.930	0.979
	5	0.978	0.948	0.988	0.900	0.733	0.950	0.984
	6	0.970	0.935	0.981	0.923	0.933	0.920	0.986
	7	0.986	0.981	0.988	0.877	0.767	0.910	0.989
	8	0.987	0.971	0.993	0.900	0.733	0.950	0.989
	9	0.986	0.961	0.995	0.923	0.833	0.950	0.991
	10	0.993	0.984	0.996	0.862	0.767	0.890	0.991
	11	0.995	0.984	0.999	0.892	0.733	0.940	0.991
	12	**0.990**	**0.984**	**0.993**	**0.946**	**0.900**	**0.960**	**0.992**
	13	0.990	0.981	0.993	0.885	0.800	0.910	0.994
	14	0.991	0.981	0.995	0.877	0.800	0.900	0.994
	3	0.719	0.268	0.919	0.733	0.367	0.917	0.675
	4	0.707	0.171	0.944	0.733	0.233	0.983	0.693
	5	0.718	0.243	0.929	0.733	0.333	0.933	0.686
	6	0.722	0.304	0.908	0.767	0.467	0.917	0.708
	7	0.733	0.257	0.944	0.756	0.300	0.983	0.721
	8	0.744	0.282	0.949	0.744	0.367	0.933	0.738
	9	0.744	0.336	0.925	0.778	0.367	0.983	0.759
	10	0.735	0.211	0.968	0.744	0.233	1.000	0.748
	11	0.740	0.325	0.924	0.744	0.367	0.933	0.757
	12	0.755	0.357	0.932	0.756	0.367	0.950	0.764
	13	0.780	0.418	0.941	0.700	0.300	0.900	0.793
	14	0.764	0.329	0.957	0.744	0.300	0.967	0.785
	15	0.768	0.386	0.938	0.800	0.433	0.983	0.782
	16	0.775	0.346	0.965	0.744	0.333	0.950	0.793
	17	0.792	0.507	0.919	0.789	0.467	0.950	0.820

	18	0.776	0.407	0.940	0.800	0.500	0.950	0.819
	19	0.785	0.468	0.925	0.744	0.467	0.883	0.800
	20	0.781	0.404	0.949	0.778	0.467	0.933	0.811
	21	0.782	0.471	0.921	0.789	0.433	0.967	0.825
	22	0.786	0.429	0.944	0.744	0.367	0.933	0.826
	23	0.779	0.393	0.951	0.733	0.300	0.950	0.835
	24	0.775	0.379	0.951	0.800	0.433	0.983	0.818
	25	0.808	0.486	0.951	0.767	0.433	0.933	0.838
	26	0.788	0.450	0.938	0.722	0.400	0.883	0.822
	27	0.812	0.518	0.943	0.744	0.433	0.900	0.849
	28	**0.807**	**0.525**	**0.932**	**0.811**	**0.567**	**0.933**	0.848
	29	0.795	0.446	0.949	0.778	0.533	0.900	0.845
Gene expression	30	0.816	0.554	0.933	0.811	0.533	0.950	0.870
	3	0.786	0.319	0.936	0.809	0.300	0.922	0.775
	4	0.814	0.392	0.949	0.800	0.300	0.911	0.816
	5	0.811	0.419	0.937	0.827	0.500	0.900	0.843
	6	0.814	0.400	0.947	0.882	0.550	0.956	0.857
	7	0.838	0.481	0.953	0.855	0.500	0.933	0.863
	8	0.839	0.527	0.940	0.845	0.550	0.911	0.886
	9	0.840	0.496	0.951	0.864	0.600	0.922	0.889
	10	0.872	0.646	0.944	0.891	0.600	0.956	0.907
	11	0.883	0.638	0.962	0.909	0.800	0.933	0.918
	12	0.883	0.635	0.963	0.900	0.700	0.944	0.921
	13	0.903	0.735	0.957	0.818	0.650	0.856	0.935
	14	0.888	0.677	0.956	0.873	0.700	0.911	0.935
	15	0.906	0.719	0.965	0.827	0.650	0.867	0.941
	16	0.918	0.773	0.964	0.873	0.700	0.911	0.947
	17	0.920	0.750	0.974	0.882	0.700	0.922	0.954
	18	0.933	0.785	0.980	0.873	0.700	0.911	0.956
	19	0.923	0.754	0.978	0.936	0.850	0.956	0.965
	20	**0.941**	**0.835**	**0.975**	**0.900**	**0.850**	**0.911**	**0.972**

	21	0.943	0.804	0.988	0.882	0.700	0.922	0.967
	22	0.936	0.819	0.974	0.864	0.550	0.933	0.969
	23	0.951	0.831	0.990	0.836	0.600	0.889	0.976
	24	0.967	0.908	0.986	0.800	0.750	0.811	0.974
	25	0.959	0.892	0.980	0.864	0.600	0.922	0.978
	26	0.972	0.931	0.985	0.882	0.700	0.922	0.979
	27	0.963	0.888	0.986	0.809	0.650	0.844	0.978
	28	0.964	0.892	0.988	0.909	0.800	0.933	0.984
	29	0.969	0.904	0.990	0.873	0.750	0.900	0.985
	30	0.971	0.900	0.994	0.855	0.800	0.867	0.985
	31	0.965	0.904	0.985	0.909	0.800	0.933	0.986
	32	0.966	0.888	0.991	0.864	0.850	0.867	0.988
	33	0.979	0.931	0.995	0.855	0.650	0.900	0.986
	34	0.980	0.942	0.993	0.891	0.500	0.978	0.987
Gut microbiome	35	0.973	0.904	0.995	0.909	0.900	0.911	0.990

The models built using the respective HLN sizes chosen for each variable group were then used for the prediction of MetS and the predicted probability threshold was adjusted to maximise the performance of each model. The average of the best models for each training and testing set were then used to compare the performance of the hybrid GA with ANN models with that of the individual classification models: LR, decision tree (DT), support vector machine (SVM) and ANN (Table 5.2). When using haematological measures to predict MetS, the hybrid GA with ANN model was found to have the highest classification accuracy of 99.7% in the training set. However, the AUC value of 0.992 was only the second highest, with the best performing model being ANN, with an AUC value of 0.998. Nevertheless, the difference between the performance of the hybrid model and the independent ANN model was very small. On the other hand, the hybrid model predicted MetS with much higher accuracy compared to the independent ANN model when using gene expression levels. While the hybrid model yielded an AUC value of 0.834, the ANN model only achieved an AUC value of 0.804. However, both models had the lowest AUC values compared to the three other independent models: LR, DT and SVM. Lastly, using gut microbial counts, the hybrid model was able to outperform all four independent classification models in MetS prediction, with an AUC value of 0.972. Additionally, while the other models yielded a relatively low sensitivity value in the testing set, the hybrid model was able to attain a sensitivity value of 0.950. Overall, it can be said that the hybrid model performed very well in MetS prediction against the other four independent classification models.

Table 5.2. The averaged performance of the 10 training and testing sets for each individual classification model and hybrid genetic algorithm with artificial neural network.

		Training Set			Testing Set			
Variable group	Model	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	AUC
	LR	0.930	0.868	0.951	0.923	0.833	0.950	0.978
	DT	0.960	0.913	0.976	0.838	0.633	0.900	0.958
	SVM	0.992	0.994	0.991	0.915	0.867	0.930	0.979
	ANN	**0.990**	**0.997**	**0.988**	**0.962**	**0.967**	**0.960**	**0.998**
Haematological measures	GA	0.997	0.990	0.999	0.900	0.867	0.910	0.992
	LR	0.889	0.764	0.944	0.833	0.733	0.883	0.917
	DT	0.902	0.786	0.954	0.689	0.567	0.750	0.863
	SVM	**0.996**	**0.989**	**0.998**	**0.811**	**0.600**	**0.917**	**0.967**
	ANN	0.781	0.661	0.835	0.733	0.633	0.783	0.804
Gene expression	GA	0.833	0.586	0.943	0.767	0.533	0.883	0.834
	LR	0.834	0.727	0.868	0.800	0.600	0.844	0.901
	DT	0.929	0.838	0.958	0.727	0.450	0.789	0.895
	SVM	0.992	0.965	1.000	0.827	0.200	0.967	0.945
	ANN	0.961	0.927	0.972	0.845	0.700	0.878	0.967
Gut microbiome	**GA**	**0.937**	**0.923**	**0.942**	**0.855**	**0.950**	**0.833**	**0.972**

The optimal combination of biomarkers identified by GA for each of the 10 training sets in all three variable groups can be found in Appendix 5.1, 5.2, 5.3. The combination with the highest fitness function value in haematological measures utilised the biomarkers: age, glycated haemoglobin A1c (HbA1c), haemoglobin (HG), white cell count (WCC), neutrophils, erythrocyte sedimentation rate (ESR), C-reactive protein (CRP), cholesterol and triglycerides (TG). Genetic algorithm also found the combined expressions of AKT serine/threonine kinase 1 (AKT1), C-C motif chemokine ligand- (CCL-)3, cluster of differentiation- (CD-)163, CEA cell adhesion molecule 3 (CEACAM3), C-X-C motif chemokine ligand- (CXCL-)5, C-X-C motif chemokine receptor 6 (CXCR6), Fc fragment of IgA receptor (FCAR), formyl peptide receptor 1 (FPR1), granzyme H (GZMH), interferon-induced protein with tetratricopeptide repeats 1 (IFIT1), insulin receptor (INSR), interferon regulatory factor 7 (IRF7), mitogen-activated protein kinase 1 (MAPK1), mechanistic target of rapamycin kinase (MTOR), phosphoinositide-3-kinase regulatory subunit 1 (PIK3R1), PYD and CARD domain containing (PYCARD), ribosomal protein S6 kinase A1 (RPS6KA1), S100 calcium-binding protein A12 (S100A12), serpin family G member 1 (SERPING1), transcription factor EB (TFEB), TNF superfamily member 13 (TNFSF13) and UNC-51 like autophagy activating kinase 1 (ULK1) to have the highest fitness function value. Finally, using gut microbial species, the combination that achieved the highest fitness function value included *A. putredinis*, *B. uniformis*, *B. luti*, *C. methylpentosum*, *C. spiroforme*, *C. catus*, *C. comes*, *E. coprostanoligenes*, *E. ramulus*, *E. rectale*, *F. butyricus*, *F. saccharivorans*, *G. formicilis*, *H. porcina*, *I. massiliensis*, *M. intestini*, *N. timonensis*, *O. ruminantium*, *P. distasonis*, *P. merdae*, *P. phocaeensis*, *R. inulinivorans*, *R. champanellensis* and *R. lactatiformans*.

The biomarkers that were identified by GA as part of the optimal combination were then compared with the biomarkers that were considered significant for MetS prediction, shown in Section 4.4, page 128. HbA1c and ESR were also important in LR while age, WCC and CRP

were also used as nodes in DT. Additionally, both HG and TG were considered important in both LR and DT. For gene expression levels, IFIT1 expression was also significant in the LR model while AKT1 was important in the DT. Both CCL3 and CXCR6 were used in LR, DT and GA. Lastly, the gut microbial species that were also found to be important by LR were *B. uniformis*, *C. methylpentosum*, *E. rectale* and *O. ruminantium*, while DT found *P. merdae* and *P. phocaeensis* important. Three microbial species were labelled as significant to MetS prediction in all three models: *A. putredinis*, *B. luti* and *M. intestini*.

5.5 Discussion

The biomarkers that are used to construct classification models have a large impact on the overall performance of the model. Users must decide on which biomarkers to use depending on the relevancy to the disease being predicted. Care must also be taken as to not include too many biomarkers to prevent the risk of overfitting, reducing computational cost and potentially improve the performance of the model. The choice of which biomarkers to omit, however, may pose a challenge and thus many researchers have implemented feature selection techniques to classification models. A popular choice of feature selection methods is GA, which is commonly paired with ANN. The current study used the hybrid GA with ANN model to predict MetS using three variable groups: haematological measures, gene expression levels and gut microbial counts. The results of the hybrid model were then compared to that of individual prediction models to determine whether the performance was improved, thus making GA a viable choice of feature selection.

The predictive ability of the hybrid GA and ANN model was compared to that of the four independent models. The performance of all four models in predicting MetS using haematological measures were all very high, with AUC values over 0.95. The hybrid model was able to achieve the second highest AUC value of 0.992. Although the AUC value of ANN

was higher, the difference was immaterial. On the other hand, the AUC value of the hybrid model was higher than that of ANN when using gene expression levels for MetS prediction. However, both models performed worse than the other three independent models: LR, DT and SVM. As the hybrid model relies heavily on ANN, the difficulty that ANN models have when dealing with gene expression level data is likely to have impacted the performance of the hybrid model. Overall, each model can be seen to have problems predicting MetS using gene expression level data as the sensitivity level is concerning low, with the highest value being 0.733. On the other hand, ANN handles gut microbial data well as the sensitivity values for independent ANN model and the hybrid model were 0.7 and 0.95, respectively. As a whole, the hybrid model was able to predict MetS using different types of measurements with high accuracy. The performance of the hybrid model was therefore a competitive option and should be considered in future research.

As GA is a feature selection technique, the biomarkers that it included in the optimal combination for MetS prediction were compared to that of the best performing LR and DT models. The biomarkers that appeared in all three models for haematological measures were HG and TG. For genes, the expression of CCL3 and CXCR6 were also found by all three models to be significant in MetS prediction. Both genes are associated with inflammation and thus, as obesity and MetS are described as a state of chronic low-grade inflammation, the importance of these two genes in MetS prediction was expected. Finally, the three gut microbial species that were important were *A. putredinis* of the Bacteroidetes phylum, and *B. luti* and *M. intestini*, both belonging to the Firmicutes phylum. As both phyla are extensively studied in obesity and MetS research, the importance of these three gut microbial species is consistent with current literature.

The gut microbial species that were identified as important for the development of MetS also supported the findings of the previously published paper detailed in Appendix 4.7, page 173.

The published paper used a combination of principal component analysis (PCA) and GA to identify the gut microbial species profile that is associated with increased obesity and MetS risk. As with the findings of the paper, the current study also found high *bromii* counts and low *prausnitzii, formicigenerans, catus* and *faecis* counts to be associated with an increased risk of obesity and MetS.

There was one noteworthy limitation in this study, which was the use of ANN and LR in the construction of the hybrid models. The hybrid GA with ANN model has been commonly used in research and has been found to increase the predictive ability compared to independent prediction models. However, its performance relies heavily on how well ANN is able to deal with the data used. The current study found that when predicting MetS using variable groups that ANN did not perform particularly well in, the performance of the hybrid model also took a toll and was lower than all three other independent classification models. Additionally, the use of classification accuracies from prediction models as part of the fitness function to identify the optimal combination of biomarkers for predicting MetS means that the function is biased to the prediction model used. Future research may consider the use of other methods, such as other filter methods which was described in Section 2.7, page 58. Alternatively, independent fitness functions within the GA itself can also be used to identify the best combination of factors associated with MetS development. In spite of this limitation, however, the hybrid GA with ANN model achieved a very high performance when predicting MetS and should therefore be used in future research with small changes to the methodology.

5.6 Appendices

Appendix 5.1. Complete list of the important haematological measures and the generated fitness values identified by genetic algorithm. The table was split for editorial purposes.

Training set	Age	Basophils	Cholesterol	CRP	Eosinophils	ESR	FPG	HbA1c	Fitness value
1	0	0	1	0	1	0	0	1	0.885
2	1	1	1	1	1	0	0	1	0.894
3	1	1	0	1	1	0	1	0	0.863
4	1	1	0	1	1	0	0	0	0.875
5	0	0	1	1	0	0	1	1	0.877
6	1	0	1	1	0	1	0	1	0.897
7	0	1	0	1	0	1	0	0	0.864
8	0	0	1	0	1	0	0	0	0.880
9	0	1	0	0	1	0	0	0	0.878
10	0	0	0	1	0	1	1	0	0.857
Sum	**4**	**5**	**5**	**7**	**6**	**3**	**3**	**4**	

Appendix 5.1 (Cont.)

Training set	HCT	HDL-C	HG	LDL	Lymphocytes	Monocytes	Neutrophils	PLT	Fitness value
1	1	0	1	0	1	0	1	1	0.885
2	0	1	1	0	0	1	1	0	0.894
3	0	0	0	0	1	1	1	0	0.863
4	0	1	0	0	0	1	1	0	0.875
5	1	1	1	1	1	1	0	1	0.877
6	0	0	1	0	0	0	1	0	0.897
7	1	1	1	1	0	0	1	0	0.864
8	1	1	1	1	1	1	1	0	0.880
9	1	0	1	0	1	1	1	0	0.878
10	0	1	1	0	1	1	1	1	0.857
Sum	**5**	**6**	**8**	**3**	**6**	**7**	**9**	**3**	

Appendix 5.1 (Cont.)

Training set	RCC	TG	WCC	Fitness value
1	0	0	0	0.885
2	0	1	0	0.894
3	1	0	0	0.863
4	0	0	0	0.875
5	1	0	1	0.877
6	0	1	1	0.897
7	0	0	1	0.864
8	0	1	1	0.880
9	0	1	0	0.878
10	0	1	0	0.857
Sum	**2**	**5**	**4**	

Appendix 5.2. Complete list of the important genes and the generated fitness values identified by genetic algorithm. The table was split for editorial purposes.

Training set	AKT1	AKT3	ATG7	CAMP	CCL3	CD163	CD1C	CD68	CD84	Fitness value
1	1	1	1	0	1	0	0	1	0	0.873
2	1	0	0	0	1	1	0	0	0	0.889
3	0	1	1	0	0	1	0	0	0	0.867
4	1	1	0	1	1	0	0	0	1	0.875
5	1	1	0	1	1	1	1	0	1	0.884
6	1	1	0	0	1	0	0	1	1	0.888
7	1	1	1	0	0	0	1	0	1	0.883
8	1	1	0	0	1	1	1	0	1	0.867
9	0	1	0	0	1	1	0	1	1	0.863
10	1	0	0	1	0	0	1	1	0	0.889
Sum	**8**	**8**	**3**	**3**	**7**	**5**	**4**	**4**	**6**	

Appendix 5.2 (Cont.)

Training set	CDH1	CEACAM3	CSF3R	CXCL5	CXCR6	FCAR	FCER2	FPR1	Fitness value
1	0	1	0	1	1	1	0	1	0.873
2	0	1	0	1	1	1	0	1	0.889
3	0	0	1	1	0	1	1	1	0.867
4	1	0	0	1	0	1	0	0	0.875
5	0	1	0	0	1	1	1	1	0.884
6	1	0	0	1	0	0	1	0	0.888
7	0	0	1	1	0	1	0	1	0.883
8	0	1	0	0	1	1	1	1	0.867
9	1	0	0	0	0	0	1	1	0.863
10	0	1	0	0	0	0	0	1	0.889
Sum	**3**	**5**	**2**	**6**	**4**	**7**	**5**	**8**	

Appendix 5.2 (Cont.)

Training set	GZMH	GZMM	HMGB1	IFIT1	IL11RA	IL1RN	INSR	IRF7	Fitness value
1	1	0	0	1	0	0	0	1	0.873
2	1	0	0	1	0	0	1	1	0.889
3	0	1	0	1	0	1	1	1	0.867
4	0	0	0	0	0	0	1	1	0.875
5	1	1	0	0	1	0	1	1	0.884
6	0	0	1	0	0	0	1	1	0.888
7	0	0	1	0	1	1	0	0	0.883
8	1	1	1	1	1	1	1	0	0.867
9	1	0	0	0	1	0	1	1	0.863
10	0	0	1	1	1	1	0	1	0.889
Sum	**5**	**3**	**4**	**5**	**5**	**4**	**7**	**8**	

Appendix 5.2 (Cont.)

Training set	ITGAE	KLRC2	LCN2	LTF	MAPK1	MTOR	NRF1	Fitness value
1	0	1	1	1	0	0	0	0.873
2	0	0	0	0	1	1	0	0.889
3	0	1	1	0	1	0	1	0.867
4	0	1	0	0	1	0	1	0.875
5	0	0	0	0	0	1	0	0.884
6	1	1	0	0	1	0	1	0.888
7	0	0	0	0	0	0	1	0.883
8	0	0	0	0	1	0	1	0.867
9	1	1	1	1	1	0	1	0.863
10	0	0	0	0	1	0	1	0.889
Sum	**2**	**5**	**3**	**2**	**7**	**2**	**7**	

Appendix 5.2 (Cont.)

Training set	PIK3CA	PIK3R1	PYCARD	RPS6KA1	S100A12	SERPING1	SOCS1	Fitness value
1	1	1	0	1	0	1	0	0.873
2	0	1	1	1	1	1	0	0.889
3	0	0	1	0	0	0	1	0.867
4	0	0	0	0	0	1	0	0.875
5	0	0	1	0	1	0	0	0.884
6	1	1	1	1	0	0	0	0.888
7	0	0	0	1	1	1	0	0.883
8	0	0	1	0	0	0	0	0.867
9	1	0	1	1	1	1	1	0.863
10	0	1	1	1	0	1	1	0.889
Sum	**3**	**4**	**7**	**6**	**4**	**6**	**3**	

Appendix 5.2 (Cont.)

Training set	TFEB	TNFSF13	TSC1	ULK1	TNFSF13	Fitness value
1	0	1	1	1	1	0.873
2	1	1	0	1	1	0.889
3	1	0	1	0	0	0.867
4	0	1	0	1	1	0.875
5	0	0	0	1	0	0.884
6	0	0	0	1	0	0.888
7	0	0	0	1	0	0.883
8	0	0	0	1	0	0.867
9	1	1	0	0	1	0.863
10	1	0	0	1	0	0.889
Sum	**4**	**4**	**2**	**8**	**4**	

Appendix 5.3. Complete list of the important gut microbial species and the generated fitness values identified by genetic algorithm. The table was split for editorial purposes.

Training set	Agathobaculum butyriciproducens	Alistipes onderdonkii	Alistipes putredinis	Anaeromassilibacillus senegalensis	Anaerostipes hadrus	Fitness value
1	0	1	0	0	0	0.809
2	0	0	1	0	0	0.828
3	0	0	0	1	0	0.809
4	0	0	0	1	0	0.800
5	1	0	0	0	0	0.808
6	0	1	1	0	0	0.820
7	0	1	1	0	0	0.802
8	1	0	1	0	1	0.781
9	0	1	1	1	0	0.792
10	0	0	0	1	0	0.800
Sum	**2**	**4**	**5**	**4**	**1**	

Appendix 5.3 (Cont.)

Training set	Bacteroides stercoris	Bacteroides thetaiotaomicron	Bacteroides uniformis	Bacteroides vulgatus	Blautia faecis	Blautia luti	Fitness value
1	1	0	0	0	1	0	0.809
2	0	0	1	0	0	1	0.828
3	0	0	1	0	1	0	0.809
4	0	1	1	0	1	0	0.800
5	0	1	0	0	1	0	0.808
6	1	1	0	0	0	0	0.820
7	1	0	0	0	1	1	0.802
8	1	1	1	1	1	0	0.781
9	1	1	1	0	1	1	0.792
10	0	1	1	0	1	0	0.800
Sum	**5**	**6**	**6**	**1**	**8**	**3**	

Appendix 5.3 (Cont.)

Training set	Blautia wexlerae	Clostridium clostridioforme	Clostridium leptum	Clostridium methylpentosum	Clostridium spiroforme	Fitness value
1	1	0	1	1	0	0.809
2	0	0	0	1	1	0.828
3	1	0	0	0	1	0.809
4	1	1	0	1	0	0.800
5	1	1	0	1	1	0.808
6	1	0	1	0	1	0.820
7	1	1	1	0	0	0.802
8	1	0	1	0	0	0.781
9	0	0	1	0	1	0.792
10	0	1	1	0	1	0.800
Sum	**7**	**4**	**6**	**4**	**6**	

Appendix 5.3 (Cont.)

Training set	Clostridium xylanolyticum	Coprococcus catus	Coprococcus comes	Desulfovibrio simplex	Dorea formicigenerans	Fitness value
1	1	1	0	1	0	0.809
2	0	1	1	0	0	0.828
3	0	0	0	0	1	0.809
4	0	1	0	1	0	0.800
5	0	1	1	0	0	0.808
6	1	0	0	1	1	0.820
7	0	1	1	1	0	0.802
8	0	0	1	1	1	0.781
9	1	1	0	0	1	0.792
10	0	1	1	0	1	0.800
Sum	**3**	**7**	**5**	**5**	**5**	

Appendix 5.3 (Cont.)

Training set	Dorea longicatena	Eubacterium coprostanoligenes	Eubacterium eligens	Eubacterium ramulus	Eubacterium rectale	Fitness value
1	0	1	0	1	1	0.809
2	0	1	0	1	1	0.828
3	0	0	1	1	1	0.809
4	0	1	1	1	0	0.800
5	0	1	0	0	1	0.808
6	1	0	0	1	0	0.820
7	1	0	1	0	1	0.802
8	1	0	0	0	0	0.781
9	1	0	1	1	0	0.792
10	0	1	0	0	0	0.800
Sum	**4**	**5**	**4**	**6**	**5**	

Appendix 5.3 (Cont.)

Training set	Faecalibacterium prausnitzii	Flavonifractor plautii	Flintibacter butyricus	Fusicatenibacter saccharivorans	Gemmiger formicilis	Fitness value
1	0	1	1	1	0	0.809
2	0	0	1	1	1	0.828
3	1	0	1	1	0	0.809
4	0	0	1	1	0	0.800
5	1	1	0	1	0	0.808
6	1	0	0	1	1	0.820
7	1	1	0	1	0	0.802
8	1	1	1	0	0	0.781
9	0	0	1	0	1	0.792
10	0	1	0	1	0	0.800
Sum	**5**	**5**	**6**	**8**	**3**	

Appendix 5.3 (Cont.)

Training set	Hespellia porcina	Ihubacter massiliensis	Intestinimonas butyriciproducens	Lachnoclostridium pacaense	Murimonas intestini	Fitness value
1	0	0	1	0	1	0.809
2	1	1	0	0	1	0.828
3	0	0	0	0	0	0.809
4	1	0	0	0	1	0.800
5	0	1	1	1	1	0.808
6	1	1	0	1	1	0.820
7	1	0	1	0	1	0.802
8	1	1	1	1	1	0.781
9	1	1	1	0	0	0.792
10	1	1	0	1	1	0.800
Sum	**7**	**6**	**5**	**4**	**8**	

Appendix 5.3 (Cont.)

Training set	Neglecta timonensis	Odoribacter splanchnicus	Oscillibacter ruminantium	Oscillibacter valericigenes	Parabacteroides distasonis	Fitness value
1	1	0	0	1	1	0.809
2	1	0	1	0	1	0.828
3	0	0	1	0	0	0.809
4	1	1	1	1	1	0.800
5	0	1	0	0	1	0.808
6	0	1	0	0	0	0.820
7	1	0	0	1	0	0.802
8	1	1	1	0	0	0.781
9	0	1	0	0	1	0.792
10	0	0	1	1	1	0.800
Sum	**5**	**5**	**5**	**4**	**6**	

Appendix 5.3 (Cont.)

Training set	Parabacteroides merdae	Pseudoflavonifractor phocaeensis	Romboutsia timonensis	Roseburia inulinivorans	Fitness value
1	1	1	0	1	0.809
2	1	1	0	1	0.828
3	1	0	1	1	0.809
4	1	1	1	0	0.800
5	1	0	0	0	0.808
6	0	0	0	1	0.820
7	0	1	1	0	0.802
8	0	1	1	0	0.781
9	0	1	0	1	0.792
10	1	1	0	1	0.800
Sum	**6**	**7**	**4**	**6**	

Appendix 5.3 (Cont.)

Training set	Ruminococcus bromii	Ruminococcus champanellensis	Ruminococcus faecis	Fitness value
1	1	0	1	0.809
2	0	1	0	0.828
3	1	1	1	0.809
4	1	1	1	0.800
5	1	0	1	0.808
6	0	0	0	0.820
7	0	0	0	0.802
8	1	0	0	0.781
9	0	0	0	0.792
10	0	0	1	0.800
Sum	**5**	**3**	**5**	

Appendix 5.3 (Cont.)

Training set	Ruminococcus torques	Ruthenibacterium lactatiformans	Sporobacter termitidis	Fitness value
1	0	0	0	0.809
2	0	1	0	0.828
3	1	0	0	0.809
4	1	0	1	0.800
5	1	0	0	0.808
6	1	1	1	0.820
7	1	0	1	0.802
8	1	0	0	0.781
9	1	1	1	0.792
10	0	1	1	0.800
Sum	**7**	**4**	**5**	

5.7 References

[1] Karegowda, A.G., Manjunath, A.S., and Jayaram, M.A., Application of genetic algorithm optimized neural network connection weights for medical diagnosis of Pima Indians diabetes IJSC. vol. 2, pp. 15-23, 2011.

[2] Mortajez, S. and Jamshidinezhad, A., An artificial neural network model to diagnosis of type II diabetes. J Res Med Dent Sci. vol. 7, pp. 66-70, 2019.

[3] Heaton, J., *Introduction to Neural Networks with Java*, ed. M. McKinnis. 2008: Heaton Research, Inc.

CHAPTER 6
Improving MetS prediction through the use of weighted majority voting

6.1 Abstract

Weighted majority voting combines the prediction of independent classification models to improve the accuracy of prediction. Although the most appropriate prediction model for metabolic syndrome (MetS) classification has yet to be identified, weighted majority voting provides an alternative which may yield even higher classification accuracy. The current study used the predicted MetS output of logistic regression (LR), decision trees (DT), support vector machine (SVM) and artificial neural network (ANN) as inputs for weighted majority voting. The current study predicted MetS using haematological measures, gene expression levels and gut microbial composition. Weighted majority voting was able to achieve the highest area under the curve (AUC) value for the prediction of MetS using gut microbial counts as well as the second highest AUC when using haematological measures and gene expression levels. The performance of the voting method was hindered by the low performance of some base learners, namely DTs, which had a low sensitivity value in the testing set of all variable groups. Nevertheless, weighted majority voting was still able to attain a high predictive ability and thus its use in future research should be considered.

6.2 Introduction

Each type of prediction model, also referred to as base learners, has its own advantages and disadvantages when it comes to the different types of data being used and the research question looking to be answered. The type of prediction model chosen by the researchers has a large impact on the overall predictions made. The choice of which prediction model to use has

therefore become a challenge due to the weight that it holds. To overcome this challenge, there are methods such as ensemble modelling, which are able to increase the classification accuracy of individual prediction models by combining the predicted output. There are also many different ensemble modelling techniques which may be implemented, including majority voting. By compiling the predicted classes from either different types of base learners or repeated predictions from one single base learner, majority voting will identify the most commonly predicted class output which then becomes the final output. Studies in metabolic syndrome (MetS) have also implemented majority voting and compared the results with that of various base learners to determine whether majority voting increases the classification accuracy of individual base learners. A study conducted by Barakat [1] compared the accuracy of majority voting with support vector machine (SVM), C5.0 decision tree (DT) algorithm, classification and regression tree (CART) and JRIP algorithm in predicting MetS. Majority voting was only able to achieve the same classification accuracy as the highest performing base learner, SVM, of 97.3%. However, the area under the curve (AUC) value improved from 0.976, yielded by SVM, to 0.988. More recently, Liao et al. [2] compared majority voting to the base learner logistic regression. The study found a slight increase in the performance, with classification accuracy improving from 72.2% to 72.4% and an increase in AUC from 0.791 to 0.801. In both studies, it was found that although the differences were immaterial, there was still an improvement in the prediction of MetS and thus the use of ensemble modelling techniques, such as majority voting, is recommended for future research in this field.

6.3 Research design

6.3.1 Study design

Weighted majority voting was used in an attempt to increase the classification accuracy of four individual base learners: logistic regression (LR), decision tree, support vector machine and artificial neural network (ANN). Each of the four base learners were initially used to predict MetS status in 152 participants, classified as either obese with MetS or healthy weight controls. There were three different groups of biomarker measurements which were used for the prediction of MetS: haematological measures, gene expression levels and gut microbial composition. The detailed explanation of how these measurements were acquired can be found in Section 3.3.2, page 75. Additionally, the methodology for each base learner used to obtain the prediction of MetS status in each of the 152 participants has been detailed in Section 4.3, page 115. In this section, the splitting of the full dataset into 10 training and testing sets for the purpose of 10-fold cross-validation was explained. The predictions of each participant in every training and testing set for all four prediction models were then extracted as inputs for the majority voting technique.

6.3.2 Weighted majority voting

Participants in each training and testing set were predicted as either '0' for healthy weight control or '1' for obese with MetS by each of the four prediction models. Consequently, there were four predicted outputs for each participant in every training and testing set. The predictions made by each of the four predictions were then used as the input for the weighted majority voting technique. If a participant was predicted to be in a particular class output by at least three of the prediction models, this became the final prediction by weighted majority voting. However, if there was an even split of predictions for the two class outputs, that is two

prediction models predicted the participant to be ‘0’ while the other two predicted the participant to be ‘1’, the weight of each model was considered. The weight of each model reflected the final average classification accuracy for each variable group (Table 4.8, Section 4.4, page 128). The final output for the majority voting was determined by the higher combined weight of the two models that predicted the same class output.

6.4 Results

The results of the weighted majority voting were compared to the final performance for each of the four base learners to determine whether the ensemble modelling technique was able to improve predictive ability (Table 6.1). When predicting MetS using gut microbial counts, weighted majority voting yielded an AUC value of 0.97, outperforming all four base learners. The weighted majority voting method was also able to yield the second highest AUC for both haematological measures and gene expression levels.

The results of the hybrid genetic algorithm (GA) with ANN model were also included as an input for the weighted majority voting technique to determine whether the performance of the voting method could be further improved (Table 6.2). While the AUC value increased for the voting method when predicting MetS using gut microbial counts, the AUC value decreased for both haematological measures and gene expression levels.

Table 6.1. The averaged performance of the 10 training and testing sets for each base learner and majority voting.

		Training Set			Testing Set			
Variable group	Model	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	AUC
	LR	0.930	0.868	0.951	0.923	0.833	0.950	0.978
	DT	0.960	0.913	0.976	0.838	0.633	0.900	0.958
	SVM	0.992	0.994	0.991	0.915	0.867	0.930	0.979
	ANN	**0.990**	**0.997**	**0.988**	**0.962**	**0.967**	**0.960**	**0.998**
Haematological measures	Voting	0.993	1.000	0.990	0.946	0.900	0.960	0.989
	LR	0.889	0.764	0.944	0.833	0.733	0.883	0.917
	DT	0.902	0.786	0.954	0.689	0.567	0.750	0.863
	SVM	**0.996**	**0.989**	**0.998**	**0.811**	**0.600**	**0.917**	**0.967**
	ANN	0.781	0.661	0.835	0.733	0.633	0.783	0.804
Gene expression	Voting	0.951	0.886	0.979	0.844	**0.700**	0.917	0.921
	LR	0.834	0.727	0.868	0.800	0.600	0.844	0.901
	DT	0.929	0.838	0.958	0.727	0.450	0.789	0.895
	SVM	0.992	0.965	1.000	0.827	0.200	0.967	0.945
	ANN	**0.961**	**0.927**	**0.972**	**0.845**	**0.700**	**0.878**	**0.967**
Gut microbiome	Voting	0.992	0.977	0.996	0.891	0.550	0.967	0.970

Table 6.2. The performance of the weighted majority voting method after the inclusion of the hybrid model.

		Training Set			Testing Set			
Variable group	Model	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	AUC
	LR	0.930	0.868	0.951	0.923	0.833	0.950	0.978
	DT	0.960	0.913	0.976	0.838	0.633	0.900	0.958
	SVM	0.992	0.994	0.991	0.915	0.867	0.930	0.979
	ANN	**0.990**	**0.997**	**0.988**	**0.962**	**0.967**	**0.960**	**0.998**
	GA	0.997	0.990	0.999	0.900	0.867	0.910	0.992
Haematological measures	Voting	0.991	0.994	0.990	0.938	0.833	0.970	0.984
	LR	0.889	0.764	0.944	0.833	0.733	0.883	0.917
	DT	0.902	0.786	0.954	0.689	0.567	0.750	0.863
	SVM	**0.996**	**0.989**	**0.998**	**0.811**	**0.600**	**0.917**	**0.967**
	ANN	0.781	0.661	0.835	0.733	0.633	0.783	0.804
	GA	0.833	0.586	0.943	0.767	0.533	0.883	0.834
Gene expression	Voting	0.947	0.861	0.986	0.844	0.667	0.933	0.912
	LR	0.834	0.727	0.868	0.800	0.600	0.844	0.901
	DT	0.929	0.838	0.958	0.727	0.450	0.789	0.895
	SVM	0.992	0.965	1.000	0.827	0.200	0.967	0.945
	ANN	0.961	0.927	0.972	0.845	0.700	0.878	0.967
	GA	**0.937**	**0.923**	**0.942**	**0.855**	**0.950**	**0.833**	**0.972**
Gut microbiome	Voting	0.992	0.977	0.996	0.918	0.750	0.956	0.976

6.5 Discussion

Different prediction model types have their own strengths and weaknesses when it comes to dealing with specific types of data. The most suitable prediction model to use therefore depends on the research question being asked. There are, however, ways to improve on specific prediction model types, or base learners, which may improve generalisation ability to provide a more accurate prediction. One common method is through weighted majority voting, which combines the outcomes of different prediction models to produce an even higher predictive ability. The base learners that are used can either be the same prediction model or different types. The current study used the outcomes of four different base learners (LR, DT, SVM and ANN) to predict MetS using weighted majority voting.

The ability of the weighted majority voting to predict MetS using the three variable groups (haematological measures, gene expression levels and gut microbial count) were compared to the predictive ability of each individual base learner. For the three variable groups, weighted majority voting predicted MetS with AUC values of 0.989 using haematological measures, 0.921 using gene expression levels and 0.970 with gut microbial counts. The weighted majority voting method was found to outperform the best model, ANN, for the prediction of MetS using gut microbial counts, which had an AUC value of 0.967. However, ANN still had a higher sensitivity value of 0.7 in the testing set, compared to the 0.550 yielded by weighted majority voting. Predictions using the other two variable groups, haematological measures and gene expression levels, found weighted majority voting to have the second highest AUC values. Additionally, the performance of the weighted majority voting was very competitive with the best performing prediction models. Using haematological measures, the voting method had higher performance than ANN in the training set, and using gene expression levels, the voting method had a higher classification accuracy and sensitivity value in the testing set compared to SVM.

For each variable group, there was always at least one base learner that had a below average performance which is likely what prevented the voting method from becoming the best performing model. When DTs were used to predict MetS using haematological measures, the sensitivity value in the testing set was very low, at 0.633. The misclassification of obese with MetS participants is likely what hindered the performance of the voting method. The pattern of low sensitivity values of base learners affecting the performance of the weighted majority voting method was also evident with gene expression levels and gut microbial counts. In spite of these shortcomings, the weighted majority voting method was still able to attain a high predictive ability for the prediction of MetS.

The hybrid model of GA with ANN constructed in Section 5.3.3, page 187 was also added as a base learner for the weighted majority voting method to determine whether the predictive ability could be further improved. As the hybrid model had the highest performance when predicting MetS using gut microbial counts, its inclusion as a base learner was expected to increase the AUC of the weighted majority voting method. As expected, the AUC value of the voting method increased, from 0.970 to 0.976. On the other hand, as the hybrid model did not perform particularly well in predicting MetS using haematological measures and gene expression levels, the AUC value of the voting method subsequently dropped. Despite the reduction of the AUC value, however, the weighted majority voting method was still able to yield a high performance with AUC values of 0.984 for haematological measures and 0.912 for gene expression levels.

The current study was limited by the performance of the individual base learners used to predict MetS. Although DT had relatively high AUC values overall, the sensitivity values in the testing set when using DT were always the very low. The low sensitivity values indicate the inability of DT to accurately identify obese with MetS participants using the provided measurements. As the performance of the weighted majority voting method depends on each base learner

included, the voting method would subsequently achieve a higher performance. In lieu of this limitation, however, weighted majority voting was still able to improve the prediction of MetS for all base learners, except the best performing learner. The use of weighted majority voting for the purpose of improving the performance of base learners in MetS prediction is therefore valid.

6.6 References

[1] Barakat, N. *Diagnosis of Metabolic Syndrome: A Diversity Based Hybrid Model*. 2016. Cham: Springer International Publishing.

[2] Liao, X., et al., Application of machine learning to identify clustering of cardiometabolic risk factors in U.S. adults. Diabetes Technol Ther. vol. 21, pp. 245-53, 2019.

CHAPTER 7
Discussion

The identification of biomarker profiles that characterise individuals who are more at risk of developing metabolic syndrome (MetS) is an important first step to enable early intervention and reduce the incidence of MetS and related diseases. Furthermore, biomarker profiles may provide potential targets for the development of treatment strategies to improve disease outcomes. Metabolic syndrome is a collection of cardiometabolic risk factors, comprising of abdominal obesity, hypertension, hyperglycaemia and dyslipidaemia [1]. The exacerbation of these risk factors leads to an increased risk of developing chronic diseases, namely type 2 diabetes mellitus (T2DM) and cardiovascular disease (CVD). To reduce the incidence of both MetS and related diseases, it is important to better understand the dysregulation of biomarkers across different body systems, including the immune system [2] and gut microbiome [3], that lead to the development of MetS. Despite originating from different body systems, many biomarkers have been recognised to play similar roles overall within the body. As such, it is important to utilise simultaneous analysis to identify the associations between biomarkers and better understand the complex network of the human body. The current study used correlation-based network analysis as well as an array of classification models for the purpose of better understanding MetS and the dysregulation of biomarkers involved in its development.

Correlation-based network analysis (CNA) was used to compare the networks built for obese with MetS participants and healthy weight controls. The biomarkers that were used to build these networks fall into four different groups of variables: anthropometric measures, haematological measures, gene expression levels and gut microbial counts. As expected, the obese with MetS network was a lot denser compared to the healthy weight network. In addition, the obese with MetS network found correlations between biomarkers across the different variable groups while the healthy weight network saw none. The associations of biomarkers

from different variable groups demonstrates the complexity of MetS, with the involvement of different body systems. Through the use of CNA, key hubs were also identified in the obese with MetS network, particularly in the gene expression levels variable group. The expression of three genes, transcription factor EB (TFEB), lipocalin 2 (LCN2), and cluster of differentiation- (CD-)68, were found to be important in the development of MetS. The results of the current study were compared to that of a preliminary study which used a much smaller dataset. In the preliminary study, the frequency of two cells, regulatory T cell and neutrophils, were recognised as key hubs. As the expression of TFEB is involved in regulatory T cell differentiation and LCN2 and CD68 are both expressed by neutrophils, the findings of the current study were consistent with that of the preliminary study.

The current study also used four different classification models for the prediction of MetS: logistic regression (LR), decision tree (DT), support vector machine (SVM) and artificial neural network (ANN). Two of the four chosen classification models, LR and DT, were also used to identify important variables for the development of MetS, as the results produced by the two models are both easily interpretable and clinically significant. Despite the expression of TFEB, LCN2, and CD68 being recognised as key hubs by CNA, none of these genes were used as inputs for the classification models due to strong correlations with the expression of other genes. However, the expression of TFEB and CD68 was strongly correlated with AKT serine/threonine kinase 1 (AKT1) expression while LCN2 expression was strongly correlated to the expression of cathelicidin antimicrobial peptide (CAMP). Both AKT1 and CAMP expression were recognised to have significant associations with MetS development, thus the findings from LR and DT were consistent with that of CNA. Additionally, classification models also found the expression of Fc fragment of IgE receptor II (FCER2), CAMP and interleukin-11 receptor subunit alpha (IL11RA) to increase the risk of developing MetS while C-X-C motif chemokine receptor 6 (CXCR6), C-C motif chemokine ligand- (CCL-)3 and killer cell lectin-

like receptor 2 (KLRC2) expression reduces the risk. Based on previous literature, FCER2 and CAMP expression is associated with obesity and inflammation [4, 5] while KLRC2 expression is inversely associated with inflammation [6, 7]. The findings of the current study involving FCER2, CAMP and KLRC2 expression is therefore consistent with previous literature. On the other hand, AKT serine/threonine kinase 3 (AKT3) expression is associated with glucose and lipid metabolism [8], with evidence of high expression leading to protection against insulin resistance. In the current study, high AKT3 expression was associated with increased obese with MetS risk, which is inconsistent with the findings of previous studies. Lastly, the expression of CXCR6 and CCL3 was linked to a lower risk of developing obese with MetS. As the high expression of both these genes are often accompanied by inflammation [9, 10], this finding did not support current literature.

Haematological measures and gut microbial counts were also used as inputs for the classification models to predict MetS. Logistic regression and DT both consistently found triglyceride measurements to be an important indicator of MetS development. As high triglyceride is a cardiometabolic risk factor of MetS, its importance in MetS development was expected. Other haematological measures that were found to increase the risk of MetS development were high levels of platelets, erythrocyte sedimentation rate (ESR), fasting plasma glucose (FPG), haemoglobin (HG) and glycated A1c (HbA1c) and low high-density lipoprotein cholesterol (HDL-C). Again, the results were anticipated as high FPG, high HbA1c and low HDL-C are all risk factors of MetS and high levels of platelets, ESR and haemoglobin are indicators of inflammation, a state which MetS is often characterised by. Despite LR and DT models being more suitable to obtain clinically meaningful results, the models do not involve causality analysis. As such, the biomarkers identified are interpreted to be associated with MetS and may indicate contribution to its development. The results from the models can

only suggest further investigations into the link between the identified biomarkers and development of MetS rather than produce conclusive findings in clinical settings.

Although the results of CNA only found correlations between species from the Firmicutes phylum in the obese with MetS network, the classification models found species from both the Firmicutes and Bacteroidetes phyla to be associated with an increased MetS risk. Previous literature has typically found obesity and MetS to be associated with a higher Firmicutes-to-Bacteroidetes ratio compared to healthy weight controls. However, there are also studies that have reported higher proportions of both the Firmicutes and Bacteroidetes phyla in the obese group compared to the normal weight group [11]. Most gut microbial species report the findings at the phylum level and thus very sparse information is available regarding the effects of specific gut microbial species on MetS development. The results of the current study therefore suggest, as well as the conflicting findings from previous studies, suggests the importance of investigating gut microbes at the species level to better understand the relationship between certain gut microbes and MetS.

Despite being able to identify important variables for MetS development in a clinical setting, neither LR nor DT were able to achieve higher predictive values compared to the two other classification models used, SVM and ANN. The prediction of MetS using haematological measures and gut microbial counts found ANN to be the best model while SVM had the highest performance using gene expression levels. However, ANN still achieved a high area under the curve (AUC) value of 0.804 using gene expression levels and thus it was considered to be the most appropriate classification model to use, overall. The inability of ANN to identify the clinically significant variables in MetS development was a downside of using this classification model. To overcome this issue, the current study implemented genetic algorithm (GA) as a feature selection technique. The hybrid GA with ANN model was able to improve the performance of the independent ANN model when predicting MetS using gene expression

levels and gut microbial counts. Using haematological measures, however, the predictive ability of ANN was extremely high, with an AUC value of 0.998. At the same time, the hybrid model was able to achieve an AUC value of 0.992, deeming it the best choice for MetS prediction. The optimal combination of biomarkers for the best prediction of MetS that was identified by GA was also consistent with the findings of both LR and DT. Triglyceride measures were again included as part of the optimal combination of haematological measures, alongside HbA1c, HG, WCC, ESR and CRP. The expression of genes, including CCL3 and CXCR6, which were considered clinically significant by LR and DT also appeared in the optimal combination identified by GA. Furthermore, the expression of TFEB, a key hub identified by CNA, was also recognised as important in MetS development. Finally, three bacterial species in particular were deemed to be clinically significant by LR, DT and GA: *A. putredinis*, *B. luti* and *M. intestini*.

Weighted majority voting was also used in the current study to determine whether the performances of individual classification models, or base learners, could be further improved by combining the final predicted outcomes. Prediction of MetS using gut microbial counts found weighted majority voting to yield the highest AUC value of 0.976. The voting method also had the second highest AUC value when using haematological measures and gene expression levels, with AUC values of 0.984 and 0.912, respectively. However, the weighted majority voting method is highly dependent on the performances of each base learner used. As some base learners, particularly DT, face more difficulty dealing with specific types of data compared to other models, the performance of the weighted majority voting method may be reduced. The weighted majority voting method may therefore not always be the best choice when looking to achieve high classification accuracies in MetS prediction. As such, the classification model that is most suitable for MetS prediction using the measurements collected by the current study is the hybrid GA with ANN model.

Correlation-network analysis and classification models both provide the means to simultaneously analyse different body systems to better understand interactions, which may not be revealed through univariate analysis. The findings of the current study contributed to obesity and MetS research by identifying the biomarkers from different body systems that were affected by obesity and MetS development. The current study also demonstrated the importance of understanding which type of classification model is more suitable for specific types of data. Future work will include the external validation of prediction models built in the current study, using public datasets with new measurements upon which the models were not built. Additionally, new prediction models will be built using different types of measurements, namely biomarkers from adipose tissue, which the majority of studies in obesity and MetS have focused on. Furthermore, longitudinal studies may help with the early detection of risk factors associated with the development of MetS and related diseases.

7.1 References

[1] Expert Panel on Detection, E. and Adults, T.o.H.B.C.i., Executive summary of the third report of the National Cholesterol Education Program (NCEP). JAMA. vol. 285, pp. 2486-2497, 2001.

[2] Ellulu, M.S., et al., Obesity and inflammation: the linking mechanism and the complications. Arch Med Sci. vol. 13, pp. 851-63, 2015.

[3] Warmbrunn, M.V., et al., Gut microbiota: a promising target against cardiometabolic diseases. Expert Rev Endocrinol Metab. vol. 15, pp. 13-27, 2020.

[4] Rastogi, D., Suzuki, M., and Greally, J.M., Differential epigenome-wide DNA methylation patterns in childhood obesity-associated asthma. Sci Rep. vol. 3, 2013.

[5] Li, Y.-X., Li, B.-Z., and Yan, D.-Z., Upregulated expression of human cathelicidin LL-37 in hypercholesterolemia and its relationship with serum lipid levels. Mol Cell Biochem. vol. 449, pp. 73-9, 2018.

[6] Jung, U.J., et al., Differences in metabolic biomarkers in the blood and gene expression profiles of peripheral blood mononuclear cells among normal weight, mildly obese and moderately obese subjects. Br J Nutr. vol. 116, pp. 1022-32, 2016.

[7] Wieser, V., et al., Adipose type I interferon signalling protects against metabolic dysfunction. Gut. vol. 67, pp. 157-65, 2016.

[8] Huang, X., Liu, G., and Su, Z., The PI3K/AKT pathway in obesity and type 2 diabetes. Int J Biol Sci. vol. 14, pp. 1483-96, 2018.

[9] Ma, K.L., et al., Activation of the CXCL16/CXCR6 pathway promotes lipid deposition in fatty livers of apolipoprotein E knockout mice and HepG2 cells. Am J Transl Res. vol. 10, pp. 1802-16, 2018.

[10] Tourniaire, F., et al., Chemokine Expression In Inflamed Adipose Tissue Is Mainly Mediated By NF-κB. PLoS One. vol. 8, 2013.

[11] Ismail, N.A., et al., Frequency of Firmicutes and Bacteroidetes in gut microbiota in obese and normal weight Egyptian children and adults. Arch Med Sci. vol. 7, pp. 501-7, 2011.