Machine Learning for Osteoarthritis (OA)

Classification and Metabolic Marker Identification

By
Thanjida Akhter
ID: 201691489

This report presented in partial fulfillment of the requirements for the degree of Master of Science in computer science

Department of Computer Science (CS)
Memorial University of Newfoundland(MUN)
St. John's, Newfoundland, Canada

September 2018

ABSTRACT

Machine Learning explores the study and construction of algorithms that can learn from and make predictions on data and analyze feature importance. Osteoarthritis (OA) is a heterogeneous disease with various pathogenic factors and a most common form of arthritis. OA has a significant metabolic background. If we can find the Important metabolites, then we can predict the disease. Three Machine Learning algorithms were used in this project to detect the most important metabolites related to OA, including Random Forest, Gradient Boosting Machine (GBM) and K-Nearest Neighbors (k-NN). Sixteen metabolites were identified as the most relevant to the disease of OA by all three algorithms.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my gratitude to Almighty Allah for the successful completion of my project work. I would like to take the opportunity to the thanks the Memorial University of Newfoundland for providing me with such a graceful opportunity to become a part of this institution. It has been a privilege for me to pursue the degree of Master of Science in Computer Science here.

I would like to thank my supervisor Dr. Ting Hu, Assistant Professor of Department of Computer Science at the Memorial University of Newfoundland helped me in various ways during the period of my Project. Her office was always open whenever I faced any challenge or had a question about my project.

Also, I must express my very profound gratitude to my mother, sister, brother and to my husband, for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of MS. This accomplishment would not have been possible without them.

TABLE OF CONTENTS

CONTENTS PAGE NO

CHAPTER

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

In this chapter, a brief overview of Machine Learning is provided in section 1.1, Osteoarthritis (OA) in section 1.2 and Metabolism in section 1.3. Motivation and summary of the project is presented in section 1.4 and 1.5.

1.1 Machine Learning

Machine learning uses experimental data to optimize clustering or the classification of samples or features to develop, augment or verify models that can be used to predict the behaviors or properties of systems [1]. It is expected that Machine Learning will provide actionable knowledge from a variety of big data including metabolomics data as well as the results of metabolism models. A variety of Machine Learning methods have been applied in bioinformatics and metabolism analyses including self-organizing maps, support vector machines, kernel machines, Bayesian networks and fuzzy logic. To a lesser extent, Machine Learning has also been utilized to take advantage of the increasing availability of genomics and metabolomics data for the optimization of metabolic network models and their analysis. In this context, Machine Learning has aided in the development of metabolic networks, the calculation of parameters for stoichiometric and kinetic models and the analysis of major features of these models for the optimal application of bioreactors. Examples of this very interesting, albeit highly complex, application of Machine Learning for metabolism modeling will be the primary focus of this review, presenting several different types of applications for model optimization, parameter determination or system analysis using various models as well as the utilization of several different types of machine learning technologies. Several basic terms used in Machine Learning showed in Table 1.1.1

Table 1.1.1 - Definition of several basic terms used in machine learning

Unsupervised learning	Unsupervised learning algorithms identify patterns from data without any user input. Example: Data-driven sample separation or feature grouping based only on data.
Supervised learning	The computer is presented with example inputs and their desired outputs.
Semi-supervised learning	The computer is given only an incomplete training signal.
Classification	A process of mapping input data to a discrete output; for example, a sample label.
Clustering	In clustering, a set of inputs is divided into groups. Unlike in classification, the groups are not known beforehand, making this typically an unsupervised task.
Regression	A process of mapping input data to a continuous output variable; for example, the quantity or the size.

1.2 History of Osteoarthritis (OA)

Osteoarthritis (OA) is a heterogeneous disease with various pathogenic factors and consists of different phenotypes that continually evolve, eventually leading to common clinical and radiographic manifestations [8]. Figure 1.2.1 shows a healthy joint and an osteoarthritis patient's joint. This is unlikely to change in the next few years, reinforcing the relevance of applying targeted interventions at an early stage of the disease or even before the development of symptoms [9]. As a result, clinicians have a challenging task in identifying those at the highest risk of developing severe OA in a timely manner. However, the prediction of OA risk is still in its infancy [3]. Hence, there is an urgent need to identify biomarkers that can be used to identify patients at high risk, to detect early changes in the disease, and to find clinical surrogates that can be used to evaluate the response to lifestyle modifications and disease-modifying drugs for OA [10].

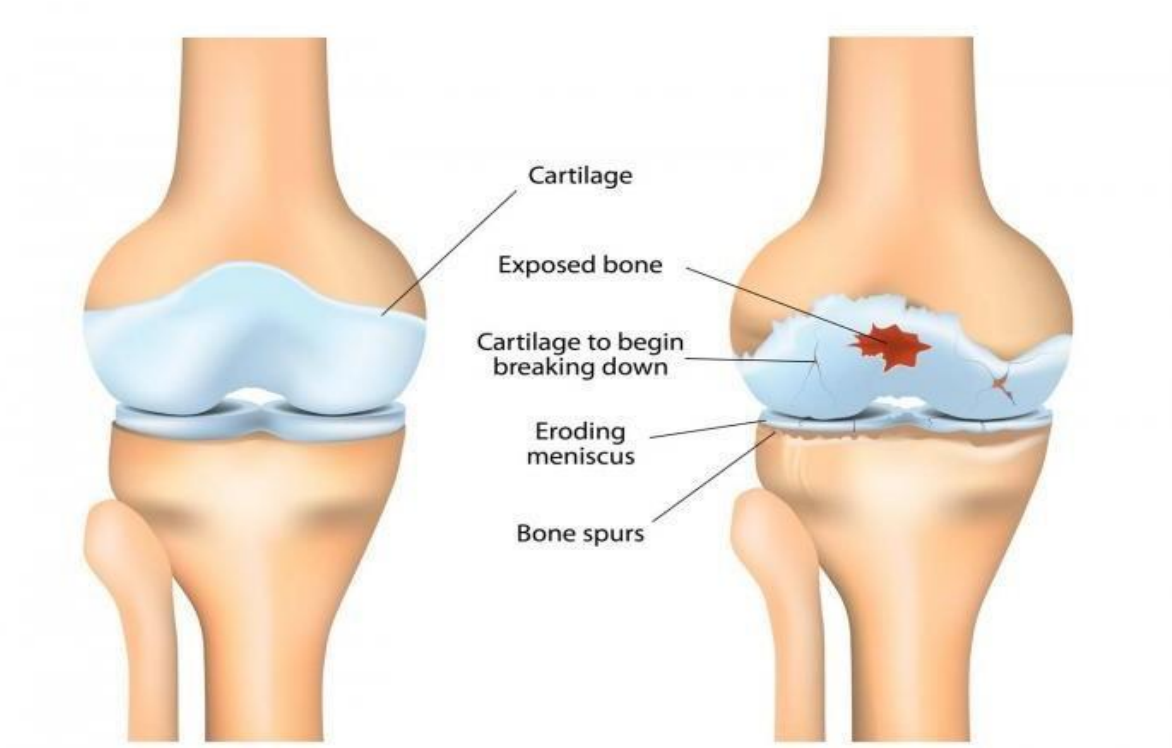

Figure 1.2.1- The image on the left side is a "Healthy joint" and the one on the right side is one with "Osteoarthritis"

1.3 Metabolism

Metabolism is a basic process in every biological system that provides energy, building blocks for cell development and adaptation as well as regulators for a variety of biological processes [1]. It is becoming widely understood and accepted that studying metabolism can give important insights into physiological processes, providing valuable data for disease diagnostics, toxicological studies and treatment follow-up or optimization in a variety of applications from clinical to environmental or agricultural sciences. Recent analysis has shown that, in the domain of precision medicine, metabolomics supplies highly complementary information for next-generation sequencing, providing a very good method for distinguishing between genes that are disease-causing or are benign mutations, thus helping in disease risk assessment and the customization of drug therapies [1]. In addition, the accurate modeling of metabolism, which is made possible through improved computer

technology and the availability of large amounts of biological data, is becoming increasingly important for several highly diverse areas including bioreactor growth, drug target determination and optimization, testing, and environmental bioremediation. Many of these applications produce large amounts of data requiring different types of analysis and allowing the derivation of diverse knowledge including sample or biomolecule clustering or classification, the selection of major features and components as well as the optimization of model parameters. Machine learning has been used for all these tasks [1].

1.4 Motivation

OA is the most common form of arthritis, affecting about 10% of the world's population aged 60 years or older [3]. It is a common source of joint pain and disability [3] and imposes a substantial socioeconomic burden on society with a cost estimate between 1.0% and 2.5% of the gross domestic product [14]. Recent studies have indicated that OA is a metabolic disease linked to several components of metabolic syndrome, such as hypertension, type-2 diabetes and dyslipidemia [11-12]. Presently, available treatments are directed at controlling the pain associated with OA, with joint replacement being used for the end-stage of the disease. Metabolites represent intermediate and end products of various cellular processes, whose levels can be regarded because of biological systems' responses to genotypic and environmental influences [2]. If we can identify the important metabolites, we can better predict the disease and its treatment can be started earlier.

1.5 Summary

In this project, the three machine learning algorithms used are Random Forest, Gradient Boosting Machine (GBM) and K-Nearest Neighbors (k-NN), and these can identify the important metabolites of osteoarthritis (OA). First, we show the accuracy of the three algorithms in predicting disease and healthy samples. We use cross-validation for more accurate results and then show the top eleven metabolites by each algorithm, with four testing data sets. The three algorithms have some common metabolites. Finally, twenty metabolites were identified as the most relevant to OA by all three algorithms.

CHAPTER 2
METHODS

This chapter focuses on the project study design. Section 2.1 presents application platform and all algorithms in section 2.2. All library files are included in this project listed in section 2.3. In section 2.4 we have described methods for Data analysis.

2.1 Application Platform

This project was completed using Python 3.7.0 (32-bit) in Anaconda Prompt and the operating system was Windows 10 Pro 32-bit. For the live code, equations, visualizations, narrative text, data cleaning and transformation, numerical simulation, statistical modeling and data visualization, we used "Jupyter Notebook".

2.2 Algorithms

There are three algorithms used in this project. These algorithms are as follows:

2.2.1 Random Forest

2.2.2 Gradient Boosting Machine (GBM)

2.2.3 K-Nearest Neighbors(k-NN)

These three algorithms are machine learning methods for classification, regression, and other tasks that operate by constructing a multitude of decision trees at the training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. k-NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally, and all computation is deferred until classification. The k-NN algorithm is among the simplest of all machine learning algorithms.

2.3 Library

For this project, sklearn was very important. Scikit-learn or sklearn is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting and k-means and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy. Panda is used for data manipulation and analysis. Matplotlib is a numerical mathematics extension for NumPy. It provides plots into applications such as creating bar charts. NumPy is used for adding support for large, multi-dimensional arrays and matrices.

The entire Python library was free. Following the library, I used in my project.

2.3.1 Sklearn

2.3.2 Pandas

2.3.3 Numpy

2.3.4 Matplotlib

2.3.5 Matplotlib_venn

2.4 Data analysis

OA samples were in CSV file format, and there were 389 samples. There were both healthy and diseased samples. Every sample has 159 metabolites statues values.

Cross-validation is a technique used to evaluate predictive models by partitioning the original sample into a training set to train the model and a test set to evaluate it.

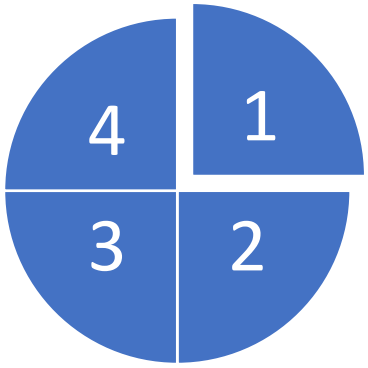

Figure 2.4.1 - Cross-validation example for the total data set

In the project, the original sample is randomly partitioned into 4 equal-sized subsamples. Of the 4 subsamples, a single subsample is retained as the validation data for the testing data set, and the remaining 3 subsamples are used as the training data set. The cross-validation process is then repeated 4 times, with each of the 4 subsamples used once as the validation data. The advantage of this method is that all observations are used for both training and validation, and each observation is used for validation once. As a percentage, the training data set is 75% and the testing data set is 25%. Using cross-validation, the accuracy of the algorithms is shown in Table 3.1.1.

CHAPTER 3

RESULT

Results related to the project are presented in this chapter. Section 3.1, described about the accuracy of three algorithms with different testing data sets, and in section 3.2 the important metabolites ranked by each algorithm and the most important metabolites of all three algorithms are in section 3.3.

3.1 Accuracy of Algorithms

The three machine learning algorithms used in this project randomly take 75% as training data and the remaining 25% is used as test data. Table 3.1.1 represents the accuracy or prediction of finding a diseased or a healthy sample.

Table 3.1.1 - The accuracy of algorithms in identifying diseased or healthy samples.

Test set	Algorithm Name	Train Accuracy (%)	Test Accuracy (%)
1	Random Forest	100%	92.78%
	Gradient Boosting	100%	92.10%
	KNN Model	92.86%	86.60%
2	Random Forest	100%	89.73%
	Gradient Boosting	100%	91.44%
	KNN Model	87.63%	83.22%
3	Random Forest	98.97%	94.52%
	Gradient Boosting	100%	90.41%
	KNN Model	92.78%	85.62%
4	Random Forest	100%	90.41%
	Gradient Boosting	100%	89.04%
	KNN Model	88.66%	81.16%

Using cross-validation for more accurate prediction also ensures that no data sample shall be eliminated from the data set. From Table 3.1, we can see that the Random Forest and Gradient Boosting Machine (GBM) training accuracy is nearly 100%, and that both algorithms' predictions are almost the same. The KNN algorithm is different from the other two algorithms, and that is why the KNN predictions are different from those of the other two algorithms.

3.2 Important metabolites rank from each algorithm

From Table 3.1, we can see that our training accuracy with the test accuracy prediction result is always 80% or above. Now, using the same cross-validation methods, we identified the important metabolites. In the important metabolites bar graph, the vertical represents the percentage of how important each of the metabolites is for the disease, and the horizontal shows the important metabolites' ranks.

Figure 3.2.1 represents the important metabolites ranked by the "Random Forest" algorithm with four different testing data sets. In different testing periods, the one metabolite which is always ranked at the top by the "Random Forest" algorithm is "Arg", and the percentage of this metabolite is between ten (10) and fifteen (15). There was a total of eighteen (18) metabolites from the "Random Forest" algorithm related to disease as shown in Table 3.3.1. in row one.

Figure 3.2.2 shows the important metabolites ranked by the "GBM" algorithm with four different testing data sets. In different testing periods, the one metabolite's percentage which was ranked much higher than from the other algorithms is "Ac.Orn", and the percentage of this metabolite is between ten (10) and thirty (30). There was a total of twenty-eight (28) metabolites from the "GBM" algorithm related to the disease as shown in Table 3.3.1. in row two.

Figure 3.2.3 represents the important metabolites ranked by the "KNN" algorithm with four different testing data sets. The five metabolites which always come to the top from the "KNN" algorithm are "Arg", "Ac.Orn", "Orn", "C18.1" and "C18.2", and their percentages were between seven (7) and ten (10). There was a total of twenty-two (22) metabolites from the "KNN" algorithm related to the disease as shown in Table 3.3.1. in row three.

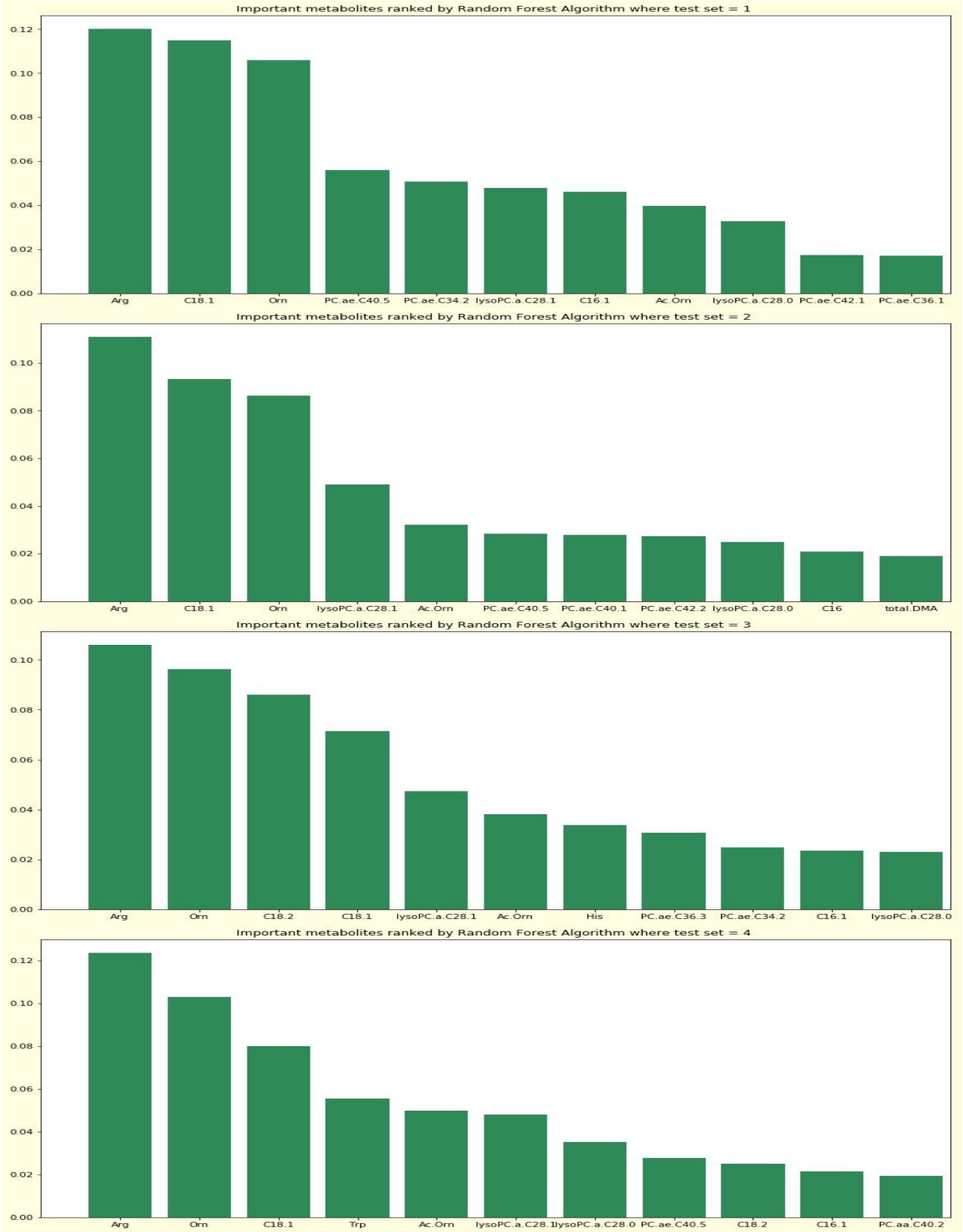

Figure 3.2.1 – Important metabolites ranked by the "Random Forest" Algorithm

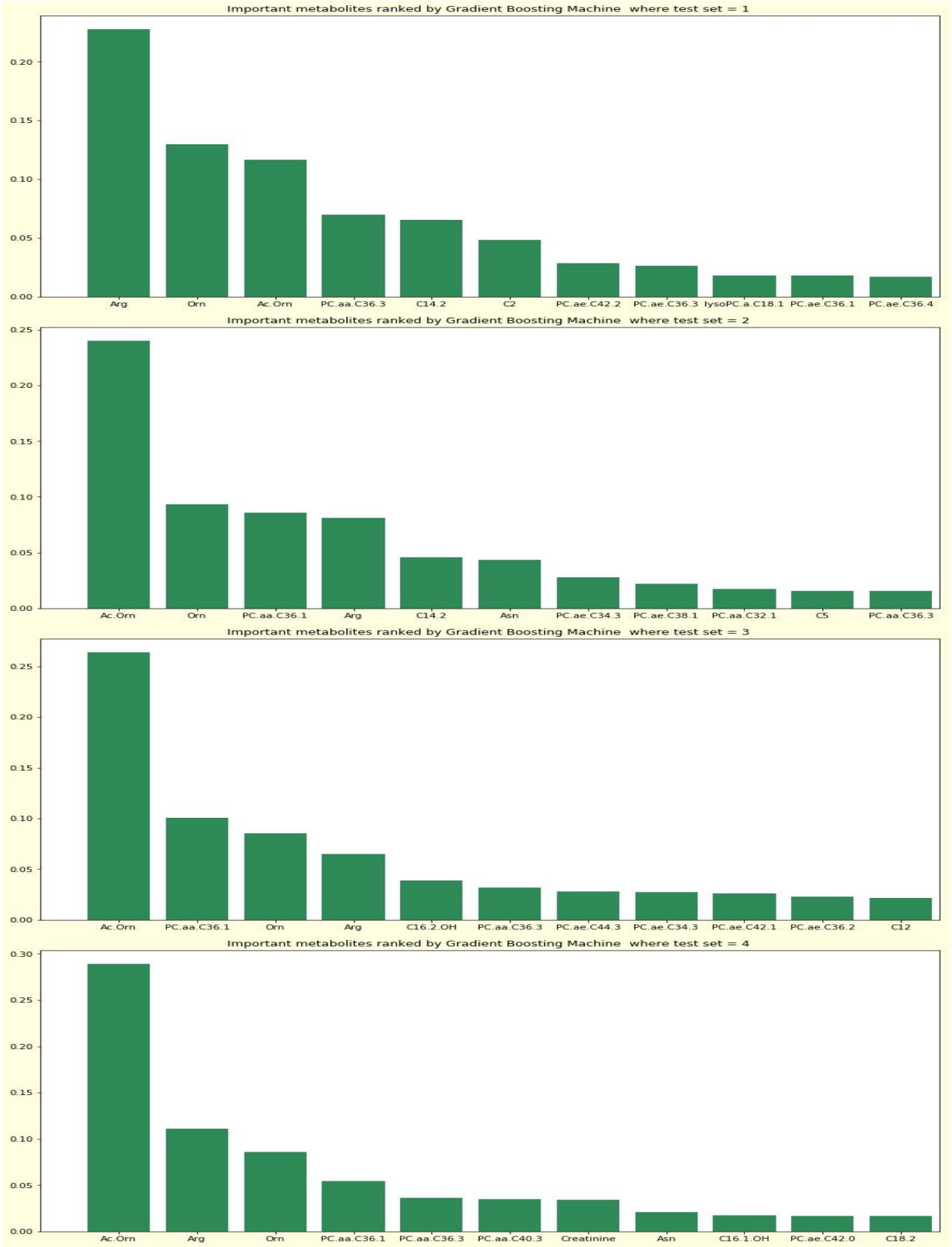

Figure 3.2.2 – Important metabolites ranked by the "GBM" Algorithm

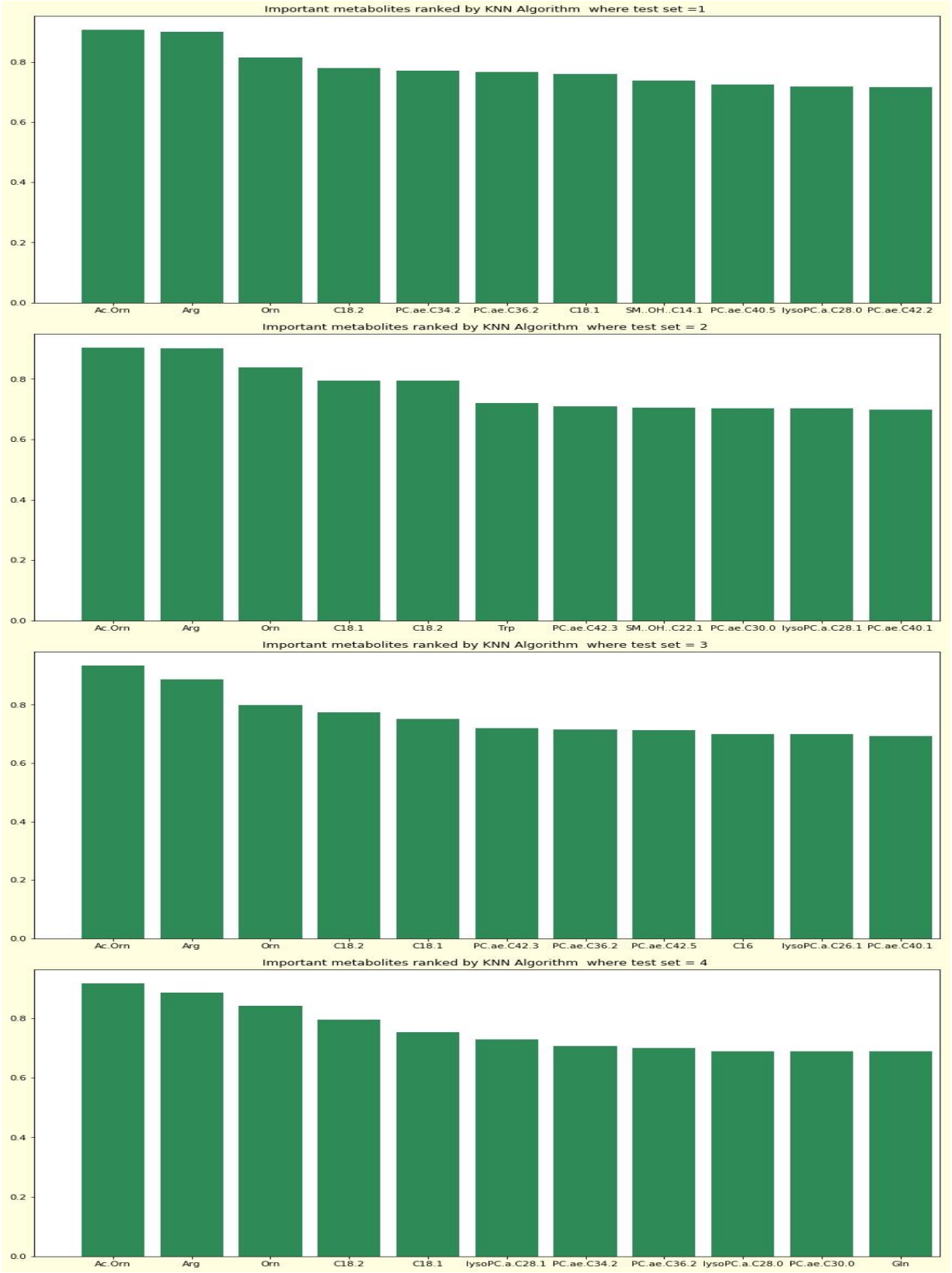

Figure 3.2.3 – Important metabolites ranked by the "KNN" Algorithm

3.4 Important metabolites list from three algorithms

Table 3.3.1- Important metabolites list from three algorithms

	Random Forest	**GBM**	**KNN**
2	Ac.Orn	Ac.Orn	Ac.Orn
3	Arg	Arg	Arg
4	C16	Asn	C16
5	C16.1	C12	C18.1
6	C18.1	C14.2	C18.2
7	C18.2	C16.1	Gln
8	His	C16.1.OH	lysoPC.a.C26.1
9	lysoPC.a.C28.0	C16.2.OH	lysoPC.a.C28.0
10	lysoPC.a.C28.1	C18.2	lysoPC.a.C28.1
11	Orn	C2	Orn
12	PC.aa.C40.2	C5	PC.ae.C30.0
13	PC.ae.C34.2	Creatinine	PC.ae.C34.2
14	PC.ae.C36.1	lysoPC.a.C18.1	PC.ae.C36.2
15	PC.ae.C36.3	Orn	PC.ae.C40.1
16	PC.ae.C40.1	PC.aa.C32.1	PC.ae.C40.5
17	PC.ae.C40.4	PC.aa.C36.1	PC.ae.C42.2
18	PC.ae.C40.5	PC.aa.C36.3	PC.ae.C42.3
19		PC.aa.C40.3	PC.ae.C42.5
20		PC.ae.C34.3	SM..OH.C14.1
21		PC.ae.C36.1	SM..OH.C22.1
22		PC.ae.C36.2	Tri
23		PC.ae.C36.3	
24		PC.ae.C38.1	
25		PC.ae.C42.0	
26		PC.ae.C42.1	
27		PC.ae.C42.2	
28		PC.ae.C44.3	

CHAPTER 4

DISCUSSION

In this chapter, Vann diagram representations of metabolites by all algorithms are presented in section 4.1. In section 4.2 all common metabolites from three algorithms were analyzed. Finally, top metabolites from those three algorithms are presented in section 4.3.

4.1 Venn Diagram

Table 3.3.1 shows all the important metabolites identified by the three different algorithms with four different data test sets. All three algorithms have some common metabolites. If we put all the metabolites in a Venn diagram, then we can find the important metabolites from the three algorithms.

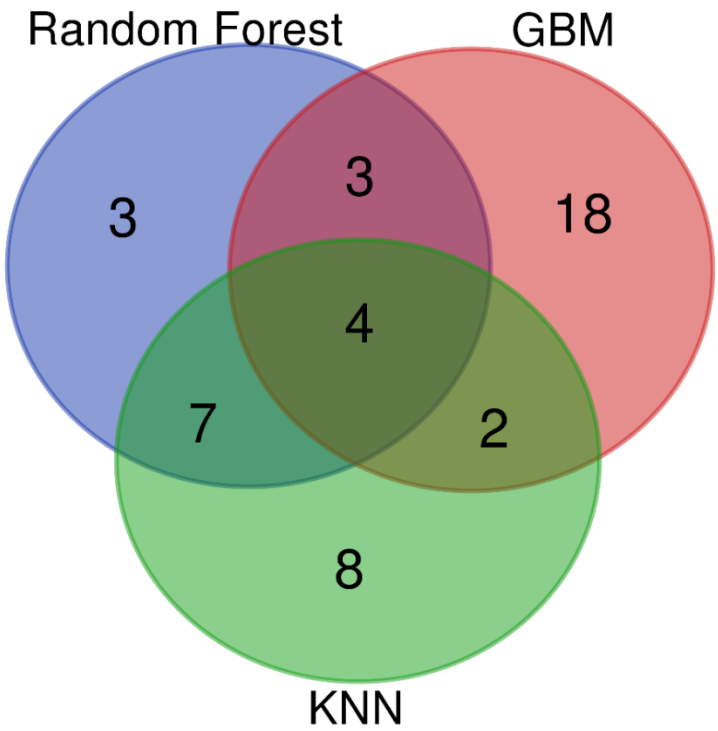

Figure 4.1.1- Metabolites presentation in a Venn diagram

4.2 Common Metabolites from all Algorithms

Figure 4.1.1 shows a Venn diagram for all the important metabolites from the three algorithms. From the Venn diagram, we can see that the three algorithms have four (4) common metabolites. Those four metabolites are “Arg”, “Ac.Orn”, “C18.2”and “Orn”.

Table 4.2.1- All the top metabolites from the three algorithms

Algorithm Names	**Total**	**Metabolites**
Random Forest, GBM & KNN	**4**	Arg, Ac.Orn, C18.2, Orn
Random Forest & GBM	**3**	PC.ae.C36.1, C16.1, PC.ae.C36.3
Random Forest & KNN	**7**	lysoPC.a.C28.0, PC.ae.C40.1, C16 PC.ae.C40.5, C18.1, lysoPC.a.C28.1, PC.ae.C34.2
GBM & KNN	**2**	PC.ae.C42.2, PC.ae.C36.2
Random Forest	**3**	PC.ae.C40.4, His, PC.aa.C40.2
GBM	**18**	C16.1.OH, PC.ae.C44.3, Asn, PC.ae.C34.3, Creatinine, C2, C5, C14.2, PC.aa.C36.1, PC.ae.C42.0, PC.aa.C36.3, PC.ae.C42.1, C12, lysoPC.a.C18.1, PC.aa.C40.3, C16.2.OH, PC.aa.C32.1, PC.ae.C38.1,
KNN	**8**	PC.ae.C42.5, Gln, SM..OH.C14.1, Tri, PC.ae.C42.3, SM..OH.C22.1, lysoPC.a.C26.1, PC.ae.C30.0

The Random Forest and KNN algorithms have seven (7) common metabolites, the Random Forest and GBM have three (3) common metabolites and the KNN and GBM algorithms have two (2) common metabolites. Of the total of forty-five (45) metabolites found from the three algorithms shown in Table 4.2.1.

4.3 Top Sixteen Metabolites

The Random Forest algorithm had three (3), KNN had eight (8) and GBM had eighteen (18) metabolites, which are not a match with any other algorithm metabolites. The three algorithms have sixteen (16) metabolites in common. The names of these sixteen (16) metabolites are provided in Table 4.3.1.

Table 4.3.1- Top Sixteen (16) common metabolites from all three algorithms

1. Arg	7. Orn	12. PC.ae.C36.2
2. Ac.Orn	8. lysoPC.a.C28.0	13. PC.ae.C36.3
3. C16	9. lysoPC.a.C28.1	14. PC.ae.C40.1
4. C16.1	10. PC.ae.C34.2	15. PC.ae.C40.5
5. C18.1	11. PC.ae.C36.1	16. PC.ae.C42.2
6. C18.2		

CHAPTER 5
CONCLUSION AND FUTURE WORK

5.1 Conclusion

This is the first study using Machine Learning to identify important metabolites of OA patients. If we can study OA patients at an early stage, then these patients will recover more easily, and the patients' treatment will be easier and less expensive. This may also mean that they do not require surgery during the final stage.

5.2 Future Work

In the future, I would like to extend this project and build a web application where all the algorithms could be generalized with their data sets. If we supply a new sample, then the web application can provide information about whether it is a diseased or a healthy sample.

CHAPTER 6

REFERENCE

1. Cuperlovic-Culf, Miroslava. "Machine Learning Methods for Analysis of Metabolic Data and Metabolic Pathway Modeling." Metabolites 8.1 (2018): 4.
2. Zhang, Weidong, et al. "Classification of osteoarthritis phenotypes by metabolomics analysis." BMJ open 4.11 (2014): e006286.
3. Zhang, W., et al. "Metabolomic analysis of human plasma reveals that arginine is depleted in knee osteoarthritis patients." Osteoarthritis and cartilage 24.5 (2016): 827-834.
4. Zhang, Weidong, et al. "Relationship between blood plasma and synovial fluid metabolite concentrations in patients with osteoarthritis." The Journal of rheumatology (2015): jrheum-141252.
5. Peek, N., C. Combi, and A. Tucker. "Biomedical data mining." Methods of Information in Medicine 48.03 (2009): 225-228.
6. Michell, T.M. Machine Learning; McGraw-Hill: New York, NY, USA, 1997.
7. Samuel, A.L. Some studies in machine learning using the game of checkers. IBM J. Res. Dev. 1959, 3, 210–229. [CrossRef]
8. Castañeda S, Roman-Blas JA, Largo R, et al. Osteoarthritis: a progressive disease with changing phenotypes. Rheumatology (Oxford) 2014;53:1–3
9. . Glyn-Jones S, Palmer AJ, Agricola R, Price AJ, Vincent TL, Weinans H, et al. Osteoarthritis. Lancet 2015;386:376e87, http://dx.doi.org/10.1016/S0140-6736(14)60802-3

10. . Blanco FJ. Osteoarthritis year in review 2014: we need more biochemical biomarkers in qualification phase. Osteoarthritis Cartilage 2014;22:2025e32, http://dx.doi.org/10.1016/j.joca.2014.09.009.
11. Katz JD, Agrawal S, Velasquez M. Getting to the heart of the matter: osteoarthritis takes its place as part of the metabolic syndrome. Curr Opin Rheumatol 2010;22:512–19.
12. Bijlsma JW, Berenbaum F, Lafeber FP. Osteoarthritis: an update with relevance for clinical practice. Lancet 2011;377:2115–26.
13. Zhuo Q, Yang W, Chen J, et al. Metabolic syndrome meets osteoarthritis. Nat Rev Rheumatol 2012;8:729–37
14. https://www.disability-benefits-help.org/working-ability/osteoarthritis
15. http://scikit-learn.org/stable/