BE-498P MAJOR TECHNICAL PROJECT -1

Metabolomics and Machine learning approaches for data analysis of Lung Cancer patient and find potential Biomarkers

Submitted by

Parul Goyal

Roll No (B20016)

Work done in supervision of

Dr. Shyam k Masakapalli

Submitted in partial fulfillment of the requirements for the award of

Bachelor of Technology

in

Bioengineering

School of Biosciences and Bioengineering
Indian Institute of Technology Mandi
Kamand - 175075 Himachal Pradesh, INDIA.

Declaration

I hereby declare that the entire work embodied in this Post Graduate Project titled as "**Metabolomics and Machine learning approaches for data analysis of Lung Cancer patient and find potential Biomarkers**" is the result of investigation carried out by me in the School of Bioscience and Bioengineering (SBB), Indian Institute of Technology, Mandi in supervision of **Dr. Shyam K Masakapalli,** for the award for the degree of Bachelor of Technology (B.Tech) in Bioengineering. It is a bonafide record of the research work carried out by me from August 2023 to November 2023 and that it has not been submitted elsewhere for any Degree of Diploma. In keeping with the general practice, due acknowledgements have been made whatever the work described is based on finding of the other investigation. I undertake that no part of this work is plagiarized.

Name: Parul Goyal **Place: Mandi**
Enrollment: B20016 **Date: 4-12-2023**
School of Bioscience and Bioengineering
Indian Institute of Technology, Mandi
Himachal Pradesh - 175075

CERTIFICATE

This is to certify that the project entitled “Metabolomics and Machine learning approaches for data analysis of Lung Cancer patients and find potential Biomarkers”
is a bonafide work done by

Parul Goyal (B20016)
in supervision of
Dr. Shyam K Masakapalli

in partial fulfillment of the requirements for the award of
Bachelor of Technology in Bioengineering
By
Indian Institute of Technology Mandi
Himachal Pradesh – 175075, INDIA

Project Guide(s)
Dr. shyam K Masakapalli

Coordinator SBB Under-Graduate Committee
Dr. Sumit Murab

School Chair SBB
Dr. Shyam Kumar Masakapalli

Acknowledgement

I want to express my gratitude to my Mentor , Dr. Shyam Kumar Masakapalli , for their encouragement, insightful suggestions, and mentorship. I'd also like to extend my thanks to Trayambak Basak sir , for granting me this wonderful opportunity to be part of this project.

Also, I would like to express my appreciation to my friends, for their feedback and support, especially to Phd scholars who have supported me to the completion of this project. Your input was invaluable in helping me to develop and refine my ideas. Your assistance, guidance, and encouragement have been invaluable. Thank you for helping me to complete this project.I am confident that my learning and personal growth have been enriched as a result.

Table of contents -

Section 1

Introduction

One of the main causes of cancer-related fatalities globally is lung cancer. Blood profiles with therapeutic value for diagnosis and therapy monitoring may be produced by metabolic changes in tumor cells combined with systemic markers of the host response to tumor development. Because lung cancer is a major worldwide health concern, improved methods for early detection and individualized treatment are required. The thorough investigation of tiny molecules connected to biological functions, or "metabolomics," has become a potent instrument for comprehending changes in metabolism in a number of illnesses, including lung cancer. The integration of metabolomics with machine learning is a promising strategy to uncover possible biomarkers in lung cancer patients, especially when combined with machine learning's skills, which are particularly useful in extracting patterns from complex datasets.

Large-scale metabolomics datasets can be efficiently analyzed using machine learning methods, such as random forests, support vector machines, and neural networks, to find minute patterns suggestive of disease states. To improve the precision and dependability of biomarker identification, these algorithms use genomic, proteomic, and clinical data while learning from a variety of data sources. Incorporating machine learning not only makes it easier to find possible biomarkers, but it also makes it possible to create prediction models for lung cancer patients' early diagnosis and prognosis evaluation.

Motivation and origin

My motivation for this project is to use metabolomics and machine learning together to solve significant challenges related to disease classification. One can learn a great deal about the biochemical changes linked to different physiological and pathological situations by examining metabolomics data. Patient outcomes can be greatly enhanced by early detection of stressful or tumor circumstances, which enables prompt and focused therapies. The need for accurate and reliable classification is the reason that contributed to the logical growth of machine learning into metabolomics.. Machine learning allows us to navigate the complexities of high-dimensional metabolomics data, providing a powerful tool for pattern recognition and biomarker discovery.

Objectives of work

Identification of potential biomarkers crucial for early detection and personalized treatment. Validation involves comparing our algorithm results with online tools, ensuring reliable biomarker discovery for enhanced clinical outcomes.

The project's goal is to build a machine learning model that evaluates how well different machine learning algorithms techniques classify metabolomics data.

Section 2 : Literature survey

Following my analysis of numerous articles and research papers, I've included the following conclusions:

Review of Machine Learning Methods in Metabolomic Research: This thorough analysis focuses on the use of machine learning methods (RF, SVM, and NNs) in metabolomic research, with a particular emphasis on the categorization of cancer in order to identify biomarkers. This review can be used by lung cancer researchers to choose suitable machine learning algorithms for their investigations.

Novel Approaches for Heart Failure with Lower Ejection Fraction (HFrEF):

A promising model is provided by the combination of machine learning and untargeted metabolomics for the HFrEF diagnostic workup. This strategy can be used to lung cancer by investigating metabolite panels as possible markers for diagnosis. Using knowledge from cardiovascular studies could lead to new ways of diagnosing lung cancer.

Predicting Gallstone Disease (GSD) with Machine Learning: Using machine and deep learning to predict GSD provides a methodological approach that takes into account variables like the gender and integrates clinical and metabolomic data. Applying these methods to the study of diseases may make it easier to find possible biomarkers. Investigating the integration of multi-omics data could improve our knowledge of the etiology of disease.

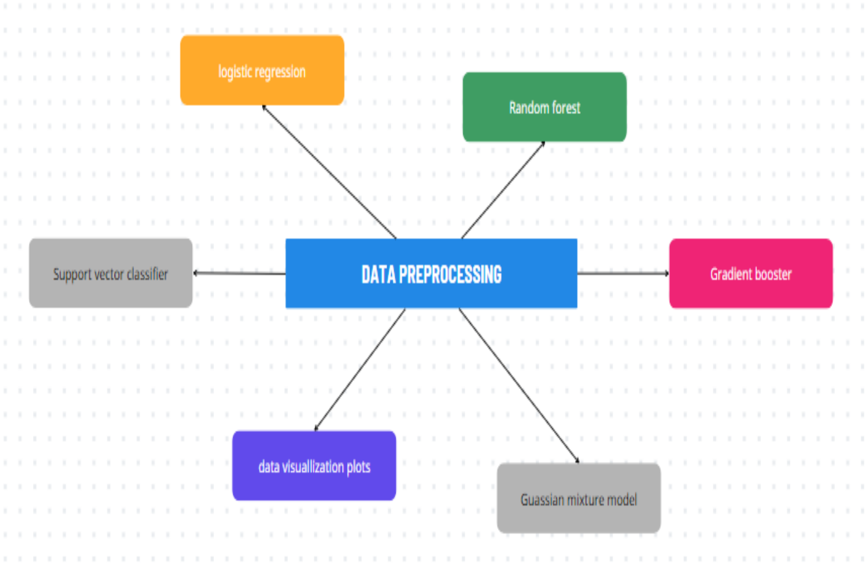
In summary , Integrating and modifying metabolomics and machine learning techniques within the framework of lung cancer investigation might yield significant understanding of the metabolic characteristics of the disease and facilitate the identification of novel diagnostic and indicative biomarkers. When planning their investigations, researchers had to take into account the advantages and disadvantages of various machine learning methods and evaluate their results within the larger framework of lung cancer biology.

Section 3 : Experimental design and Methods

Synthetic Dataset :

I've generated a synthetic dataset for my MTP project which contains 20 metabolites concentrations for 50 normal and 50 stressed patients data.

Experimental Design -



Methodology and Algorithms -

I used Machine learning algorithms to give predictions for finding the patient status if it falls in a normal state or stressed state. Here are some algorithms I used for predictions : **Visualize the distributio**n of each metabolite (plots) for normal and cancerous patients, **Random Forest classifier**, **Gradient boosting classifier feature** importance , **Support Vector** Machine classifier , **Logistic Regression** , **Gaussian Mixture Model**.

Random forest feature importance :

Feature importance represents the contribution of each feature in the model's decision making. Importance is calculated based on how much each feature reduces impurity (incorrect predictions). Advantages - high accuracy , handle missing dataset, large datasets with many features.

Gradient boosting classifier-feature importance :

It's a sequential method of misclassification. The final prediction is the sum of the prediction from all trees. For feature impotence it is based on how frequently a feature is used for splitting and how much each split improves models performance.

Gaussian mixture model :

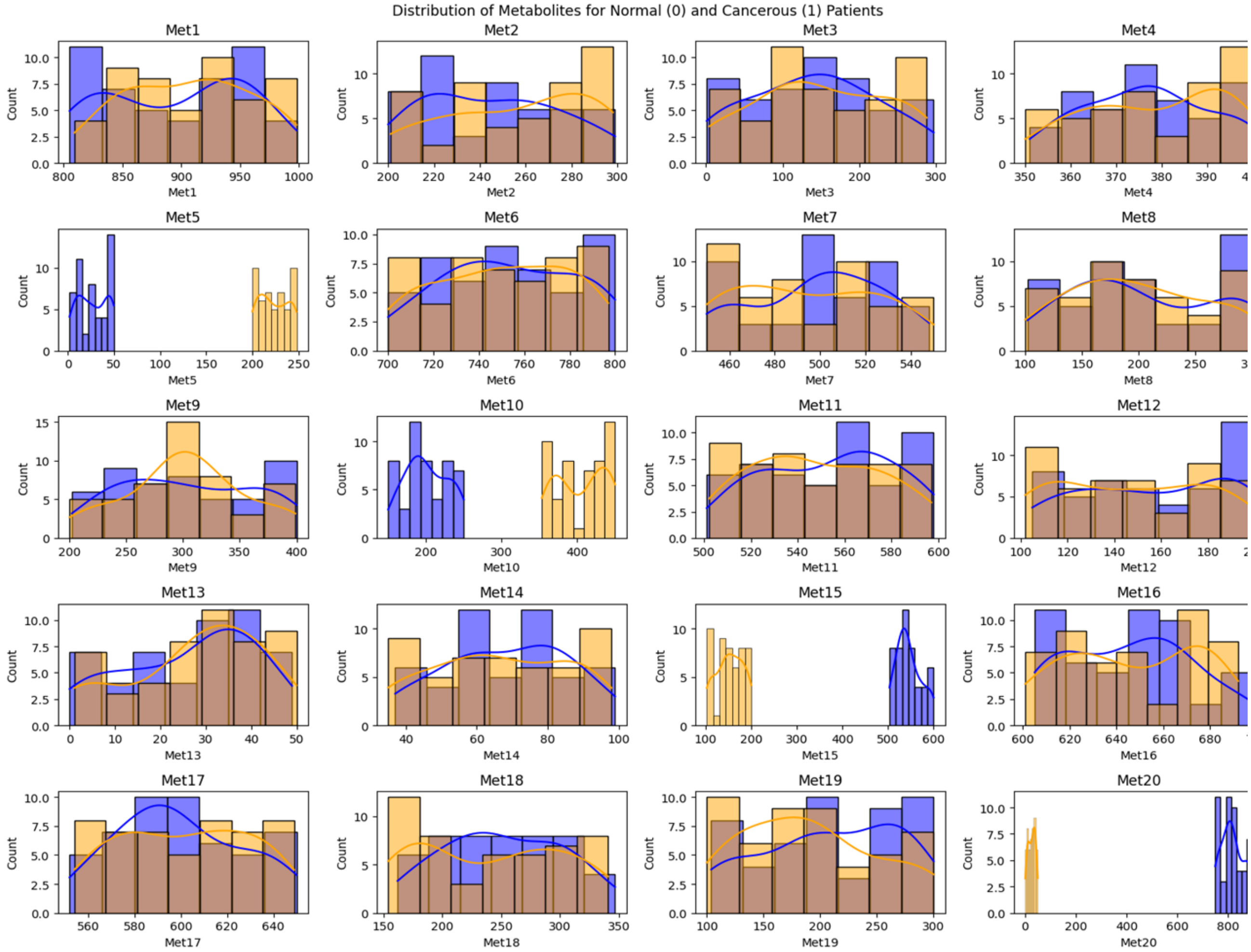
Probabilistic model used for clustering and density estimation. Each gaussian distribution in the mixture represents a cluster in the data.GMM results can be visualized by plotting the data points along with the estimated-gaussian distribution.

Support vector classifier :

It is a supervised machine learning algorithm used for classification. It works by finding the hyperplane that best separates different classes in the feature space.the hyperplane is the linear decision boundary. The goal of SVC is to find the optimal hyperplane that maximizes the margin between classes in the feature space.

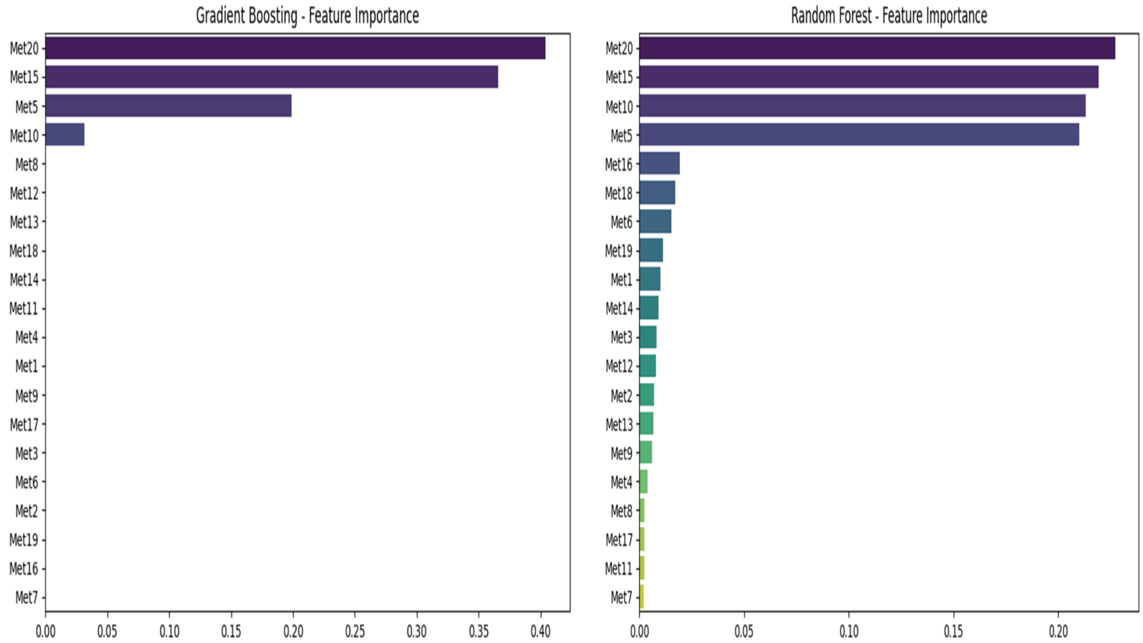
Section 4 : Results and Discussion

1. Visualize the distribution of each metabolite for normal and stressed patients



Discussion : Plotting the distribution of each metabolite allows us to infer that Met5, Met10, Met15, and Met20 are more significant in distinguishing between the patients in the stressed and normal groups. The orange bars in these graphs represent the group of stressed patients, and the blue bars represent the normal patient data. Potential biomarkers include Met5, Met10, Met15, and Met20.

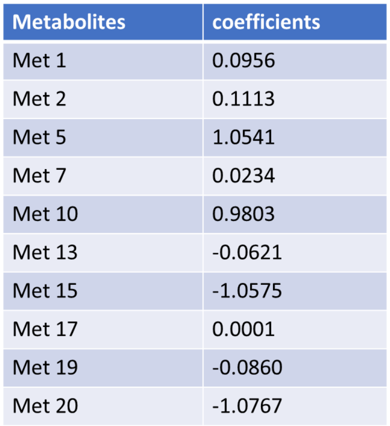
2. Results from gradient boosting classifier and random forest classifier


Gradient Boosting - Feature Importance

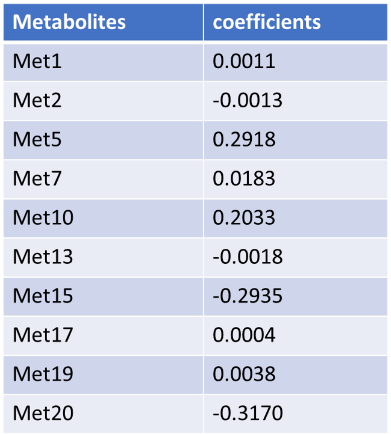
Random Forest - Feature Importance

Discussion : both random forest and gradient boosting are ensemble methods , these feature importance scores based on how often a feature is used to make decisions across multiple trees in the ensemble. Feature importance representing the ranking of metabolites.metabolite 20 , 15 , 10 5 has higher ranking , they are potential biomarkers according to the model. Higher ranking of features suggests that removing or altering the feature would have a more significant impact on the model’s predictive performance.

3. Results from logistic regression and support vector classifier respectively-



Metabolites	coefficients
Met 1	0.0956
Met 2	0.1113
Met 5	1.0541
Met 7	0.0234
Met 10	0.9803
Met 13	-0.0621
Met 15	-1.0575
Met 17	0.0001
Met 19	-0.0860
Met 20	-1.0767



Metabolites	coefficients
Met1	0.0011
Met2	-0.0013
Met5	0.2918
Met7	0.0183
Met10	0.2033
Met13	-0.0018
Met15	-0.2935
Met17	0.0004
Met19	0.0038
Met20	-0.3170

Table 1 : Logistic regression classifier

Table 2 : Support vector classifier

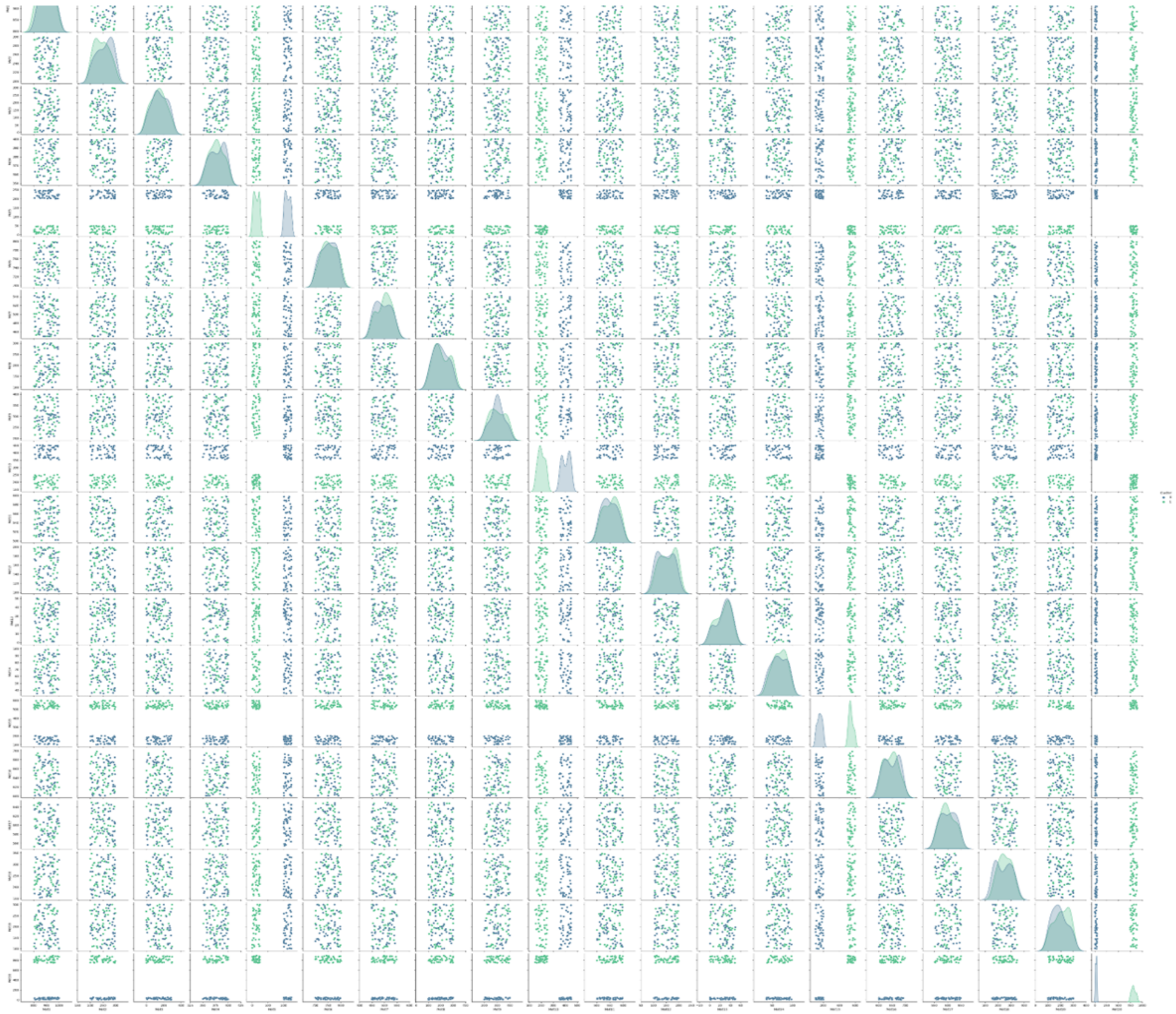
Discussion : **Table 1** - logistic regression - the coefficients are the weights assigned to each feature(metabolite) in logistic regression model. These weights indicate the strength and direction of the relationship between each metabolite and the binary outcome (if normal or stressed).

The negative and positive sign indicates the direction of association and magnitude represents the strength of that association. Positive coefficients such as met 2, met 5 , met 7 met 10 suggest that an increase in concentration of that metabolite is associated with an increased likelihood of being in the stressed group. Negative coefficients indicate that an increase in the concentration of that metabolite is associated with a decreased likelihood of being in the stressed group. Larger magnitudes suggest a stronger impact on the prediction. met 5 , met 10 , met 15 , met 20 seem to have a notable impact on the classification. Met 20 with a negative coefficient could be considered as a potential biomarker inversely associated with stress.

Table 2 - Support vector classifier - the coefficients are the weights assigned to each feature(metabolite) in support vector classifier (svc model). These coefficients represent the hyperplane that best separates the two classes (normal and stressed) in the feature space.

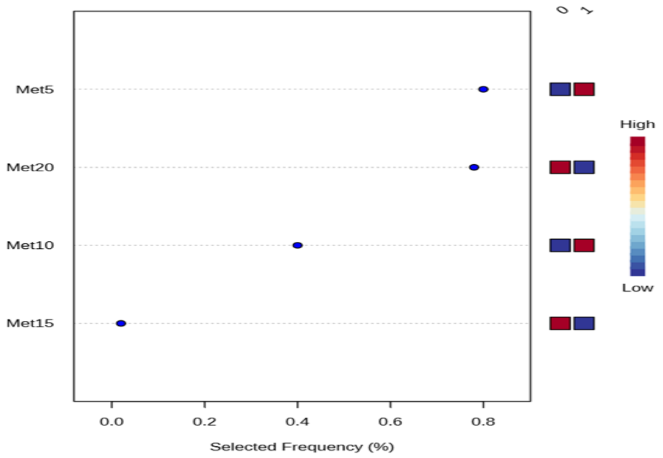
The sign of coefficients indicates the direction of the association. Positive coefficients suggest that an increase in the concentration of that metabolite contributes to a point being classified as stressed, while negative coefficients suggest the opposite. The coefficients contribute to defining the hyperplane. Met5, Met10, Met15, and Met20 have relatively larger magnitudes, suggesting they play a more significant role in defining the separation between normal and stressed groups.

4. Results from Gussian mixture model



Discussion : GMM provides clusters for each data point. The green cluster represents normal patients and the blue cluster represents stressed patients distribution. met 5 , met 10 , met 15 and met 20 are showing significance in data and separates it for normal and stressed(2 clusters).

Comparison of results with online tools



(metaboanalyst tool)

Figure : Plot of the most important features of a selected model ranked from most to least important.

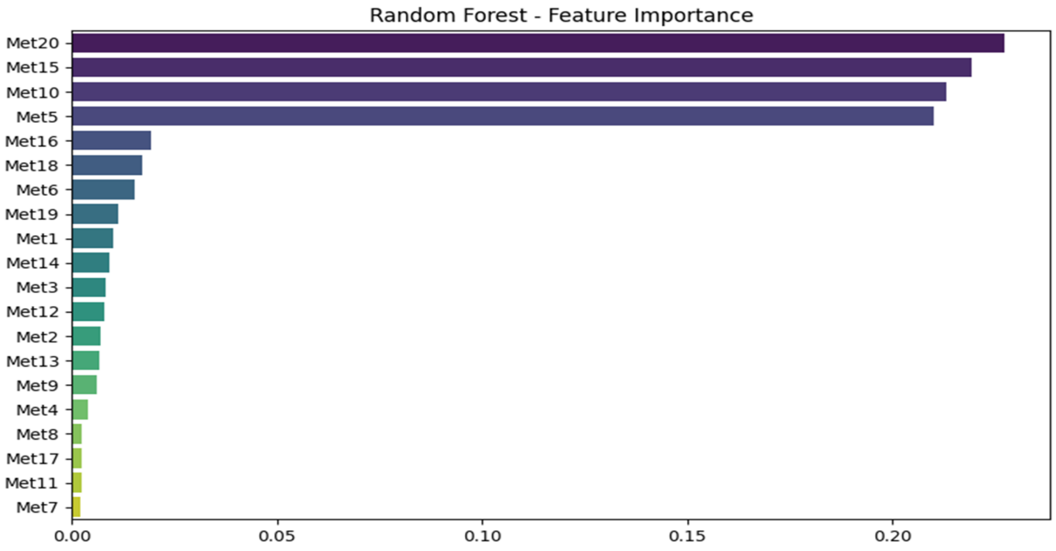


Figure : Random forest - feature importance Machine learning algorithm result

Discussion : The outcomes of the machine learning algorithm model and the readily available web tools are the same, according to my comparison. I looked for a possible biomarker using the METABOANALYST program. Metabolite 5, 10, 15, and 20 are the biomarkers for the data of both normal and stressed patients, according to both methodologies.

Section 5 : Conclusion

1. According to the results I can say that it is applicable on real dataset and will give relevant results.
2. The field of biomarker discovery has seen an increase in the use of machine learning approaches like feature selection.
3. The integration of metabolomics and machine learning provides important new understandings of lung cancer.
4. The discovery of potential biomarkers improves customized therapy and early detection.
5. Integrating 'omics' data yields an in-depth understanding of the biological conditions.
6. Demonstrates how interdisciplinary methods can lead to better therapeutic results.

6. Future work and Implications

1. Conducting longitudinal studies to track changes in metabolite profiles over time.
2. I’ll try to modify this model/technique used for larger datasets that will give better accurate predictions based on the data.
3. Can combine proteomics, genomics, and metabolomics approaches with ML for an extensive knowledge of biological systems.
4. Develop a model for available dataset to predict medical conditions and identify diseases.

7. Literature References

1. Aboud, Orwa, et al. “Application of Machine Learning to Metabolomic Profile Characterization in Glioblastoma Patients Undergoing Concurrent Chemoradiation.” *Metabolites*, vol. 13, no. 2, 17 Feb. 2023, pp. 299–299, www.ncbi.nlm.nih.gov/pmc/articles/PMC9961856/, https://doi.org/10.3390/metabo13020299. Accessed 4 Dec. 2023.
2. Azari, Hanieh, et al. “Machine Learning Algorithms Reveal Potential MiRNAs Biomarkers in Gastric Cancer.” *Scientific Reports*, vol. 13, no. 1, 15 Apr. 2023, www.ncbi.nlm.nih.gov/pmc/articles/PMC10105697/, https://doi.org/10.1038/s41598-023-32332-x. Accessed 4 Dec. 2023.
3. Cardoso, Marcella R., et al. “Metabolomics by NMR Combined with Machine Learning to Predict Neoadjuvant Chemotherapy Response for Breast Cancer.” *Cancers*, vol. 14, no. 20, 15 Oct. 2022, p. 5055, pubmed.ncbi.nlm.nih.gov/36291837/, https://doi.org/10.3390/cancers14205055. Accessed 4 Dec. 2023.
4. Echle, Amelie, et al. “Deep Learning in Cancer Pathology: A New Generation of Clinical Biomarkers.” *British Journal of Cancer*, vol. 124, no. 4, 18 Nov. 2020, pp. 686–696, https://doi.org/10.1038/s41416-020-01122-x.
5. Galal, Aya, et al. “Applications of Machine Learning in Metabolomics: Disease Modeling and Classification.” *Frontiers in Genetics*, vol. 13, 24 Nov. 2022, https://doi.org/10.3389/fgene.2022.1017340.
6. Marcinkiewicz-Siemion, M., et al. “Machine-Learning Facilitates Selection of a Novel Diagnostic Panel of Metabolites for the Detection of Heart Failure.” *Scientific Reports*,

vol. 10, no. 1, 10 Jan. 2020, p. 130, www.nature.com/articles/s41598-019-56889-8, https://doi.org/10.1038/s41598-019-56889-8.

7. Salem, Nourah M, et al. “Machine and Deep Learning Identified Metabolites and Clinical Features Associated with Gallstone Disease.” *Computer Methods and Programs in Biomedicine Update*, vol. 3, 1 Jan. 2023, p. 100106, www.sciencedirect.com/science/article/pii/S2666990023000150?via%3Dihub, https://doi.org/10.1016/j.cmpbup.2023.100106.
8. Sardar, Rahila, et al. “Machine Learning Assisted Prediction of Prognostic Biomarkers Associated with COVID-19, Using Clinical and Proteomics Data.” *Frontiers in Genetics*, vol. 12, 20 May 2021, https://doi.org/10.3389/fgene.2021.636441. Accessed 11 Oct. 2022.
9. “Several Protein Biomarkers Protect against Disease Development.” *ScienceDaily*, www.sciencedaily.com/releases/2021/12/211208143849.htm.
10. Wu, Li-Da, et al. “Analysis of Potential Genetic Biomarkers Using Machine Learning Methods and Immune Infiltration Regulatory Mechanisms Underlying Atrial Fibrillation.” *BMC Medical Genomics*, vol. 15, no. 1, 19 Mar. 2022, https://doi.org/10.1186/s12920-022-01212-0. Accessed 12 Mar. 2023.
11. Zhang, Xiaokang, et al. “Machine Learning Approaches for Biomarker Discovery Using Gene Expression Data.” *PubMed*, Exon Publications, 2021, www.ncbi.nlm.nih.gov/books/NBK569564/.