T.R.

GEBZE TECHNICAL UNIVERSITY

FACULTY OF ENGINEERING

DEPARTMENT OF COMPUTER ENGINEERING

HUMAN GENDER IDENTIFICATION AND AGE ESTIMATION USING METABOLOMICS DATA

ZAFER ALTAY

SUPERVISOR
DOÇ. DR. HABIL KALKAN

GEBZE
2022

T.R.
GEBZE TECHNICAL UNIVERSITY
FACULTY OF ENGINEERING
COMPUTER ENGINEERING DEPARTMENT

HUMAN GENDER IDENTIFICATION AND AGE ESTIMATION USING METABOLOMICS DATA

ZAFER ALTAY

SUPERVISOR
DOÇ. DR. HABIL KALKAN

2022
GEBZE

GEBZE TECHNICAL UNIVERSITY	GRADUATION PROJECT JURY APPROVAL FORM

This study has been accepted as an Undergraduate Graduation Project in the Department of Computer Engineering on 03/03/2022 by the following jury.

JURY

Member (Supervisor)	:	Doç. Dr. Habil KALKAN
Member	:	Prof. Dr. Yusuf Sinan AKGÜL
Member	:	M.Sc Başak BULUZ

ABSTRACT

In this project, models were developed that tries to predict the gender and age of people.Models use metabolomic data to make these predictions.There are two models in the project , It was used separate models for gender and age. If two different models are given the same inputs, one will predict age while the other will predict gender.An input consists of metabolomic data.Given an input, we output an age estimate and a gender estimate from the models.

The main purpose of this project is to make inferences about whether the amount of human metabolites in the body changes depending on the age and gender of the people.

In this project, after collecting the metabolomics in order to make a good prediction, many models were trained with deep learning techniques to find the most suitable system for the project. After that ,all model performances were compared and the best estimator was chosen.

ACKNOWLEDGEMENT

Firstly, I would like to thank my supervisor,Doç. Dr. Habil KALKAN, for his expertise and for taking the time to discuss my ideas and to give me insight and constant support and help throughout the duration of the project.

Then, I would also like to thank my family, for always believing in me, my friends,for being there when I needed it the most.

Zafer ALTAY

LIST OF SYMBOLS AND ABBREVIATIONS

Symbol or Abbreviation	:	**Explanation**
NMR	:	Nuclear Magnetic Resonance
GC	:	Gas Chromatography
CB	:	Clinical Biochemistry
CNN	:	Convolutional Neural Network
ML	:	Machine Learning

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

The amount of human metabolites varies depending on many things such as age, gender, drugs used, body temperature, stress,etc.[1]. Age and gender have a great influence on metabolites. Of course, we consider situations where people are healthy[2].

There are many methods of collecting and analyzing metabolites. Using these methods, metabolites are first collected. It is then analyzed. Of course, there is a set of participant for the studies. These participants are generally selected according to the target of the research area. These are some techniques such as liquid chromatography (LC), gas chromatography (GC), Nuclear Magnetic Resonance (NMR), gas chromatography-mass spectrometry (GC-MS), two-dimensional gas chromatography (GC×GC) [3].

In this study, a previously created dataset was used due to the lack of these metabolomic techniques. In this dataset, data collected from uric acid by NMR, GC, Clinical Biochemistry methods were used. Each person's 528 metabolites were looked at.

The aim of the project is to determine whether there is a correlation between sex-metabolite and age-metabolite. Therefore, separate deep learning models were trained for both age and gender. Each model was compared and the model that gave the best result was found and inference was made.

2. REQUIREMENTS

Materials needed for the project:

2.1. Environments

Necessary environmental materials:

- **Python 3+**
- **Metabolomics Dataset**

2.2. Libraries

Necessary libraries:

- **Pandas**
- **NumPy**
- **sklearn**
- **tensorflow**
- **sklearn**

3. METABOLOMICS AND METHODS

3.1. Metabolomics

Substances formed as a result of grinding of substances in tissues or organs, intermediates formed by molecules in reactions occurring one after another, and substances formed in enzymatic reactions of metabolism are called metabolites.
Metabolites have functions such as generating energy, stimulating and inhibiting enzymes, acting as catalysts and interacting with other organisms.[4][5]

Metabolomics, on the other hand, is the study of metabolites occurring at these various times. Metabolomics is used in studies that serve many different purposes. These studies include stress tests, diabetes tests, cholesterol level, etc. It is also used in areas such as early diagnosis and diagnosis of disease.[3][6][7]

3.2. Methods

Many techniques are used for metabolomic studies but here, only the techniques in the dataset are mentioned.

3.2.1. NMR

NMR is the technique that uses the magnetic properties of atomic nuclei. It determines the chemical properties of the atoms in it. It provides detailed information about the dynamics of the atom, the reaction state, and the chemical environment of the molecule. It provides access to the details of the electronic structure of an atom in a molecule.
NMR is good for analyzing well-resolved molecules. For example, because of this, organic chemistry is also used to verify substance identity. NMR's time scale is long, so it is not suitable for observing fast events. Nmr is not very sensitive so although it shows most substances, it is not very sufficient to identify these substances.

Nmr device is not available everywhere because it is a large and expensive device (Figure 3.1 Figure 3.2) [7] .

Figure 3.1: A 900 MHz NMR instrument with a Tesla magnet[7]

3.2.2. GC

It is a common type of chromatography used for the separation and analysis of compounds that can evaporate without decomposition. It is used to test the purity of a particular substance or to separate a mixture of different ingredients. Additionally, it can be used to quantify these ingredients and help identify a compound (Figure 3.3,Figure 3.4).

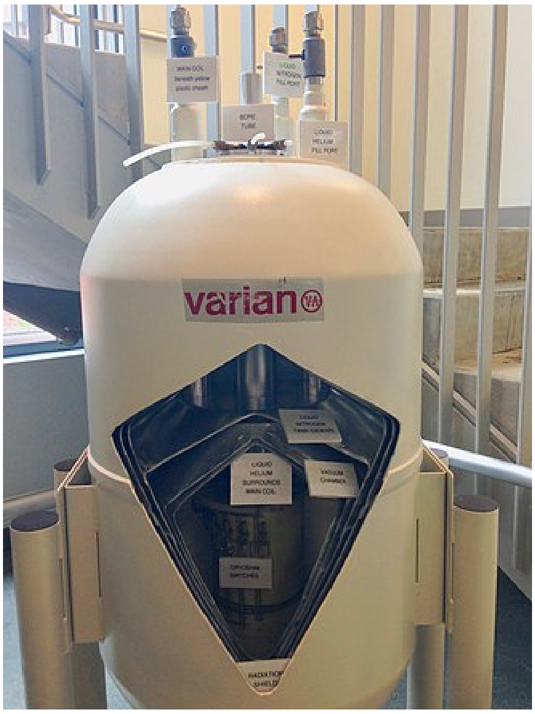
MAIN COIL beneath yellow plastic sheath
LIQUID NITROGEN FILL PORT
BORE TUBE
LIQUID HELIUM FILL PORT
varian
LIQUID HELIUM SURROUNDS MAIN COIL
VACUUM CHAMBER
CRYOSHIM SWITCHES
RADIATION SHIELD

Figure 3.2: NMR section showing the structure of the radiation shield, vacuum cage, liquid nitrogen channel, liquid helium channel, and cryogenic setting liners[7]

Figure 3.3: Gaschromatograph[8]

Figure 3.4: GasChromatographyLaborator[8]

4. MATERIALS

4.1. Dataset

The dataset was taken from other studies because the appropriate conditions were not available [9] .

This dataset contains data from 301 healthy adult humans. 172 of them are men and 129 of them are women. Ages of the participants are between 18-80 years old. In these data, which were collected by paying attention to the health status of the participants, people with chronic diseases, smokers, and infectious diseases were not included [9] .

The main reason for trying to prefer healthy people is that most of the metabolites are affected and changed by activity. Chronically ill drug users or smokers, for example, can affect a metabolite and impair the accuracy of the study.Therefore, it was tried to find the most suitable people to increase the success and accuracy of the study.

"Participants were tested for three days. Participants were extensively tested. Many parameters have been determined. These include many things including height, age, weight, blood pressure and urine biochemistry. "Samples for metabolic analyzes were collected on study day 2 " [9] .

Studies have been done for urine and plasma in the data set. The study was conducted on urine data. This data contains 528 metabolites for each participant collected from NMR, GC, Clinical Biochemistry.Of these, 197 are data obtained by NMR, 324 by GC, and 8 by Clinical Bioshemistry. Gender and age data are available for each patient. Gender and age data are available for each participant. These data are data that are tried to be estimated. It is used as an output. Metabolites were used as inputs.

5. DATA ANALYSIS

Age and gender estimation were made separately in the project. CNN was used. Many models and trials were made. Models were trained with early and late fusion techniques.

In the model, the data for test and train are separated from each other. Train data is used while training for models. Test data is used only to control model.

5.1. Early Fusion

In early fusion, studies are done with all data combined.
In other words, in this study, NMR, GC, Clinical Biochemistry are given as an input to a single model without dividing and separating it.So here, all 528 metabolites are completely an input.
And it was done separately for each technic.
For better understanding, Figures are placed, even if they are not fully explained (Figure 5.1).

INPUT
MODEL
OUTPUT

Figure 5.1: Late Fusion

5.2. Late Fusion

Late fusion is an approach that prevents the model from separating from a single architecture. A new perspective has been given to the model and the final result has been obtained with it.

In late fusion, In the model, the inputs are seperated. In other words, the data was created taken separately in NMR, GC, Clinical Biochemistry studies in a different model and then combined the models.

That is, NMR-GC-Clinical Biochemistry was advanced separately and the models were combined before the final step and became a single model.
For better understanding, Figures are placed, even if they are not fully explained (Figure: 5.2).

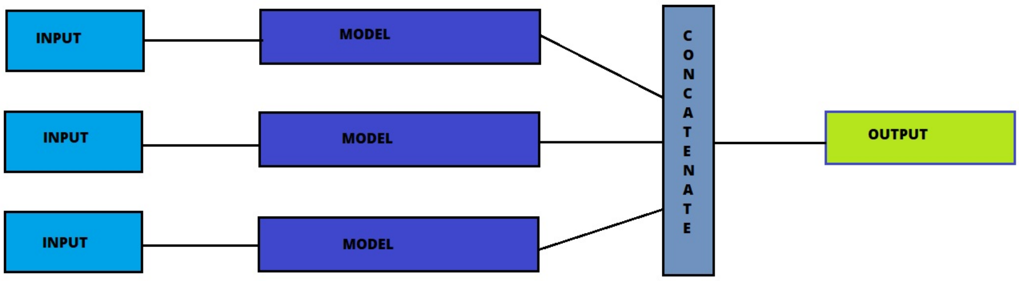

Figure 5.2: Late Fusion

Here, each separate model represents a separate analysis method. 4 models tested on this section. These models are:
-NMR*GC*Clinical Biochemistry
-NMR*GC
-NMR*Clinical Biochemistry
-GC*Clinical Biochemistry

The reason for trying extra models in the study is to find a model that produces better predictions by comparing all of them.

6. GENDER IDENTIFICATION EARLY FUSION RESULTS

Gender Identification's early fusion results are shared below (Table 6.1). The All Techniques model was gave the best results among the different tried-and-tested models. In addition, the ROC curve is shared for each model(All Tecniques: 6.1 , NMR: 6.2 , GC: 6.3 , CB: 6.4).

Table 6.1: Early Fusion for Gender Identification with all techniques

Names	Precision	Recall	F1 Score
All Techniques	0.82	0.77	0.75
NMR	0.71	0.71	0.71
GC	0.69	0.69	0.69
CB	0.61	0.60	0.60

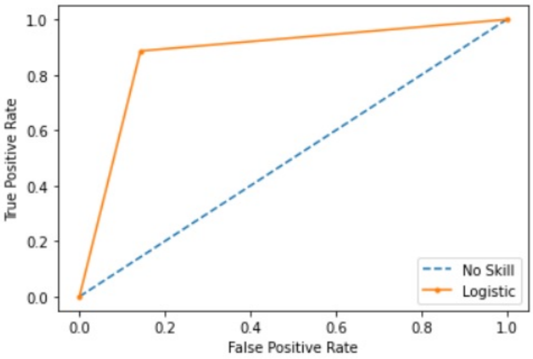

Figure 6.1: Early Fusion ROC Curve for Gender Identification with all techniques

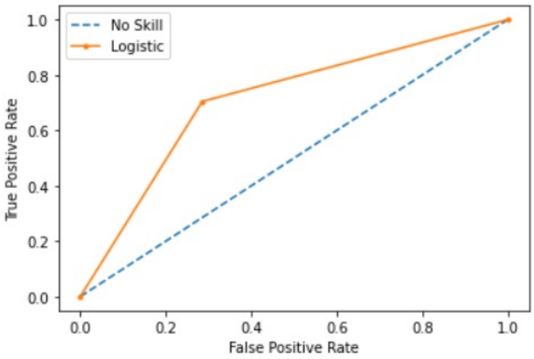

Figure 6.2: Early Fusion ROC Curve for Gender Identification with all techniques

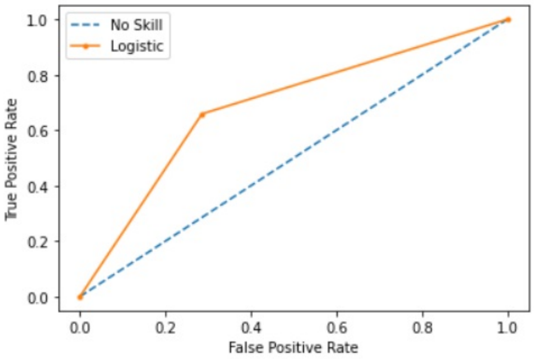

Figure 6.3: Early Fusion ROC Curve for Gender Identification with GC

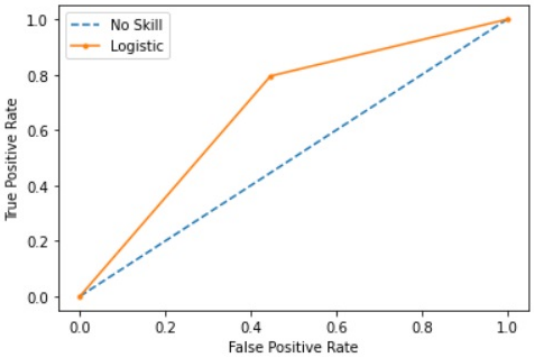

Figure 6.4: Early Fusion ROC Curve for Gender Identification with Clinical Biochemistry

7. GENDER IDENTIFICATION LATE FUSION RESULTS

Gender Identification's late fusion results are shared below (Table 7.1). The NMR-GC model was gave the best results among the different tried-and-tested models. In addition, the ROC curve is shared for each model(All Tecniques: 7.1 , NMR-GC: 7.2 , NMR-CB: 7.3 , GC-CB: 7.4).

Table 7.1: Late Fusion for Gender Identification with all techniques

Names	Precision	Recall	F1 Score
All Techniques	0.78	0.77	0.77
NMR-GC	0.81	0.80	0.81
NMR-CB	0.78	0.77	0.77
GC-CB	0.67	0.67	0.67

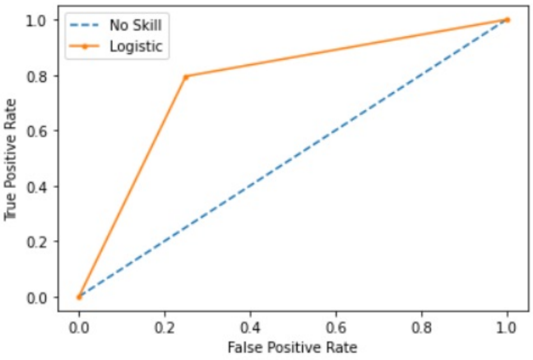

Figure 7.1: Late Fusion ROC Curve for Gender Identification with all techniques

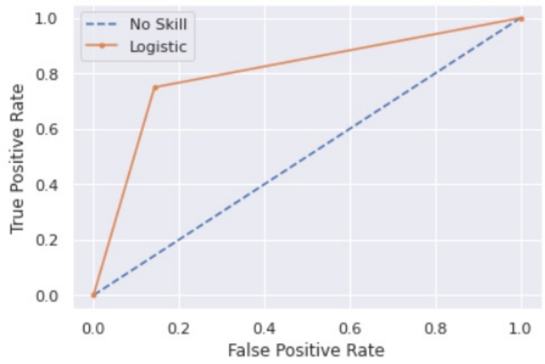

Figure 7.2: Late Fusion ROC Curve for Gender Identification with NMR and GC

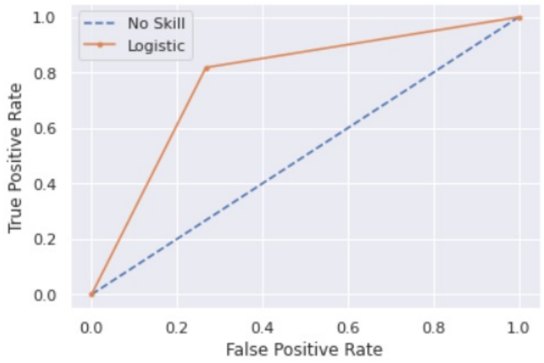

Figure 7.3: Late Fusion ROC Curve for Gender Identification with NMR and Clinical Biochemistry

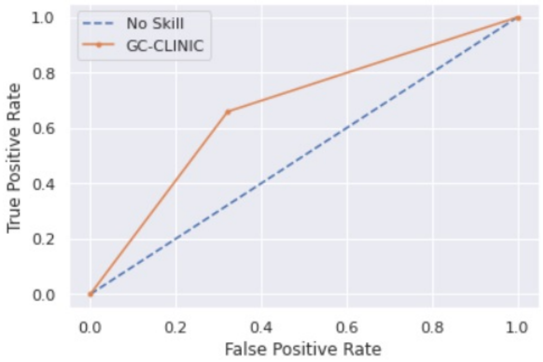

Figure 7.4: Early Fusion ROC Curve for Gender Identification with GC and Clinical Biochemistry

8. AGE ESTIMATION EARLY/LATE FUSION RESULTS

The results of the study are listed in the tables below. Different tables were made for Early Fusion (Table: 8.1) and Late Fusion(Table 8.2). According to the results of the tried models, the best estimation can be made using the GC technique modeled using early fusion.

Table 8.1: Early Fusion for Gender Identification with all results

Names	All Techniques	NMR	GC	Clinical Biochesmitry
Error Rate	12.21%	16.22%	11.97%	15.21%

Table 8.2: Late Fusion for Gender Identification with all results

Names	All Techniques	NMR-GC	NMR-CB	GC-CB
Error Rate	16.07%	18.9%	18.13%	13.95%

9. CONCLUSION

9.1. Conclusion

Study results showed that gender and age had an effect on metabolites. In this study, urinary results were examined and it shows that the sex and age of people can be estimated by analyzing their urine. It has been tried to increase the performance by trying different combinations. And this showed that without the need for all combinations to be together, the results of only some techniques are also sufficient for this estimation. In general, since all of them give successful results, gender and age estimation can be made using the techniques in this study, whether separately or in combination. It has been concluded that metabolites of healthy people can be used as markers for gender and age.

9.2. Development Proposal

In this study, the amount of elements of the dataset can be increased for more successful results. The accuracy of the model can be increased if the number of people is increased and there are more examples. More successful results can be obtained with different machine learning techniques.

BIBLIOGRAPHY

[1] "Interview from milliyet gazetesi." (), [Online]. Available: https://www.milliyet.com.tr/pembenar/metabolizma-hizini-etkileyen-faktorler-2068771#:~:text=BAZAL%20METABOL%C4%B0ZMA%20HIZINI%20ETK%C4%B0LEYEN%20ETMENLER,durumlar%20bazal%20metabolizma%20h%C4%B1z%C4%B1n%C4%B1%20etkilemektedir..

[2] "Molecular phenotyping of a uk population: Defining the human serum metabolome,"

[3] "Metabolomik." (), [Online]. Available: http://docs.neu.edu.tr/staff/serdar.susever/10%20%20metabolomik%20[Uyumluluk%20Modu]_7.pdf.

[4] "Metabolit." (), [Online]. Available: https://www.arkeolojikhaber.com/haber-metabolit-metabolitler-21343/.

[5] "What is metabolite." (), [Online]. Available: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/metabolite.

[6] "Metabolomik nedir." (), [Online]. Available: https://www.bilimvetekno.com/metabolomik-nedir/.

[7] "Metabolomics." (), [Online]. Available: https://en.wikipedia.org/wiki/Metabolomics.

[8] "Gc." (), [Online]. Available: https://tr.wikipedia.org/wiki/Gaz_kromatografisi.

[9] "Metabolite patterns predicting sex and age in participants of the karlsruhe metabolomics and nutrition (karmen) study," [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5558977/#.