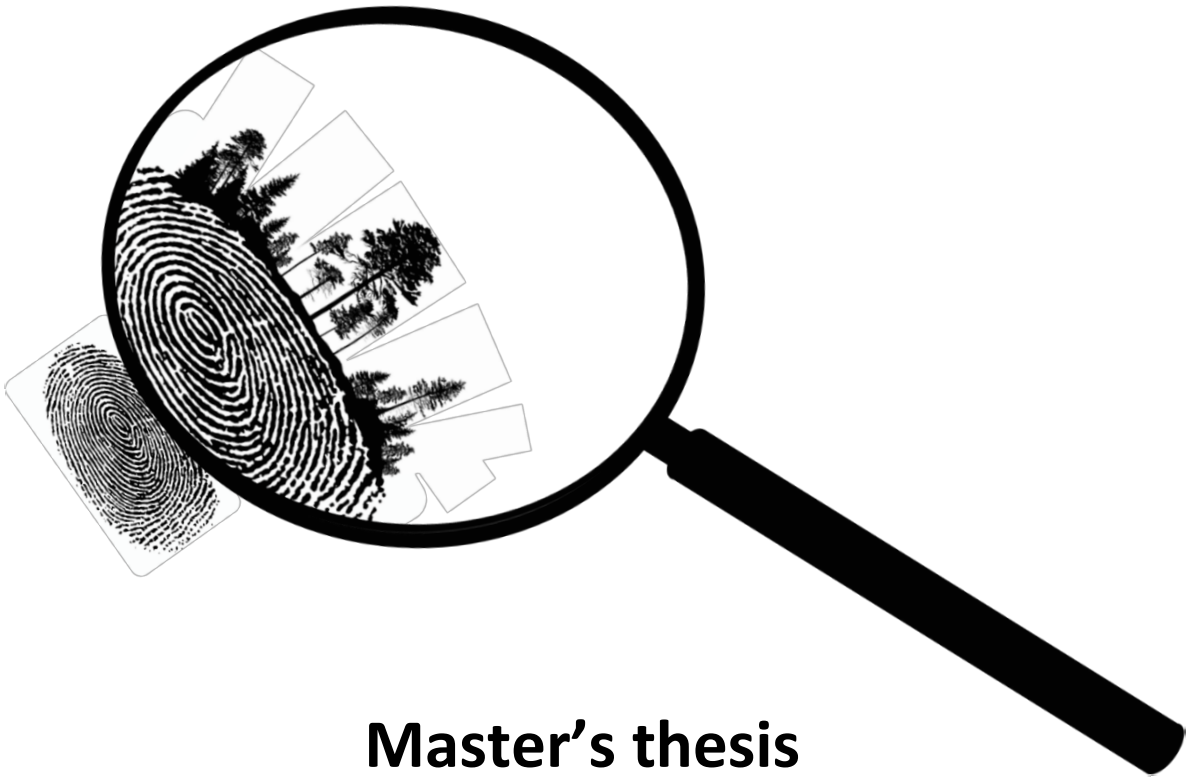
# Inference of Fingerprint Age

## using LC-MS Data and Machine Learning



## Master's thesis

## June 2020

**Stine Andersen, 201208577**

Supervisor: Palle Villesen, Associate Professor
Bioinformatics Research Centre,
Aarhus University

BiRC

AARHUS
UNIVERSITY

# Contents


Preface ..... 2
Abstract ..... 2
Introduction ..... 3
  Fingerprint Composition ..... 3
    The Ageing of Residue Components ..... 5
  Lipidomics Approach: LC-MS ..... 6
    Data Processing using XCMS ..... 8
  Dataset ..... 8
Aim of the Project ..... 10
Analytical Approach: Statistical and Machine Learning ..... 10
Results ..... 12
  Normalization ..... 12
  Evaluation of Regression Methods ..... 13
  Evaluation of Classification Methods ..... 15
    Classification with Ranger ..... 16
  Class-specific Regression Models ..... 18
  Ensemble Approach ..... 19
  Targeted Sequential Approach ..... 22
    Feature Selection ..... 23
Discussion ..... 25
Conclusion and Future Perspectives ..... 28
Methods ..... 29
  Normalization ..... 29
  Modelling ..... 30
    Regression ..... 31
    Classification ..... 31
    Ensemble Approach ..... 32
    Sequential Approach ..... 34
    Feature Selection ..... 34
Code Availability ..... 34
Acknowledgements ..... 34
References ..... 35
Appendix ..... 38

# Preface

This thesis is written as a completion of the master's degree in Bioinformatics at Aarhus University. The work presented here has been carried out at the Bioinformatics Research Centre in February to June 2020 under the supervision of Associate Professor, Ph.D. Palle Villesen. The data analyzed has been produced at the Department of Forensic Medicine at Aarhus University by Assistant Professor, Ph.D. Kirstine Lykke Nielsen in the years 2017 to 2019.

# Abstract

Fingerprints are, and has long been, a valuable source of forensic evidence. However, if doubts arise regarding the time of print deposition, it can be hard to determine the relevance of a suspect with certainty; in these cases, a reliable method for estimation of fingerprint age would be extremely useful. Such a method has been pursued through many studies, using a variety of approaches, but no reliable solution has yet been published. One way to approach the problem is to exploit the changes that happen to the composition of the print residue over time. Particularly, the organic print components, such as proteins and lipids, have been shown to display a time-dependent development in abundance due to degradation processes, which might be utilized in a predictive model.

This thesis has sought to explore the prospects of training a predictive model on lipidomics data produced by an LC-MS setup, using Statistical and Machine Learning methods. For this purpose, several different Random Forest models for both classification and regression have been trained. Most importantly, the analysis revealed the potential of class-specific regression models, and much work went into exploring how they might be applied in a finished model. Suggestions included an advanced ensemble approach, which performed with mixed results, and a simple sequential approach that runs classification and regression in series. Ultimately, the idea of a general, full-scale model was discarded in favor of a more reliable targeted model, which performed regression targeted on the first week, paired with full-scale classification model. For this approach, results also indicated that a comparable model could be trained on a selection of only 50 features. Overall, the results show promise for a future expansion of the project.