University of Pavia
University of Italian Switzerland

Doctor of Philosophy

Computational and Statistical Methods for Biomedical Data Analysis

---

Author:
Farah Naz

Supervisor:
Prof. Silvia Figini

PhD Coordinator:
Prof. Luca Pavarino

A thesis submitted in fulfillment of the requirements for the degree of

Doctor of Philosophy

in the

– Joint PhD program in Computational Mathematics and Decision Sciences - XXXVIII Cycle –

September, 2025

ABSTRACT

This doctoral dissertation investigates computational methodologies for healthcare data analysis spanning clinical, imaging, and academic domains. The research develops innovative statistical and machine learning approaches including ensemble methods for survival analysis, quantitative MRI analysis for neuromuscular diseases, graph-based clustering algorithms for complex data structures, and analytical frameworks for assessing pandemic impacts on educational systems.
The dissertation encompasses three primary research domains: clinical prediction modeling, medical image analysis, and healthcare impact assessment.

1. The clinical prediction modeling domain focuses on developing advanced ensemble methods for survival analysis, particularly addressing the challenges posed by highly correlated covariates in healthcare datasets. This research introduces the "CovBootTree" method, which extends the Proper Bayesian Bootstrap approach to survival data analysis by incorporating multivariate distribution structures through Cholesky decomposition. The proposed framework generates synthetic observations that preserve the covariance structure of clinical variables, enabling more robust survival predictions while accounting for the complex interdependencies commonly found in medical data. Through comprehensive simulation studies across varying sample sizes, this work demonstrates superior predictive stability compared to traditional survival models like Cox regression and Random Survival Forests. The technique proved particularly effective with small datasets, successfully managing complex relationships between correlated health variables while providing more reliable predictions for clinical decision-making.

2. The second project focuses on the computational approaches for unsupervised machine learning for complex data structure identification and pattern recognition in diverse analytical contexts. This research introduces the "Cli-DSP" algorithm, a novel graph-based clustering methodology that integrates min-max clique detection with density peak assignment through optimized shortest path analysis. The proposed approach addresses fundamental limitations of traditional clustering methods by capturing local connectivity patterns within datasets characterized by irregular cluster shapes, overlapping regions, and varying density distributions. The methodology demonstrates superior performance on both synthetic benchmarks and real-world applications, including biomedical data, establishing its effectiveness for complex data clustering challenges across healthcare research domains.

3. The medical image analysis domain explores the application of advanced neuroimaging techniques. The study aims to Investigate magnetization transfer ratio (MTR) as an early biomarker for muscle involvement in patients with late-onset Pompe disease (LOPD) at various disease stages compared to healthy controls. We employed quantitative analysis with Multi-echo Spin-echo (MESE) T2-weighted imaging, Multi-echo Gradient echo sequences for fat fraction (FF) assessment, and Multi-Parametric Mapping for MTR quantification. We found significant differences in MTR and FF between mild and moderate/severe LOPD patients versus healthy controls. MTI demonstrated high sensitivity in detecting mild muscle fiber damage before fat replacement occurs, making it a promising biomarker for monitoring early disease signs, progression, and treatment efficacy.

4. Under the healthcare impact assessment domain, we examine the broader implications of healthcare-related events on academic and educational systems through comprehensive data analysis methodologies. This research specifically investigates the effects of the COVID-19 pandemic on university education using mixed-effects modeling approaches to understand learning patterns and academic performance variations. The work employs longitudinal data analysis techniques to assess how major healthcare crises influence educational outcomes, student experiences, and institutional responses.

List of Publications: Published or in Preparation

- Farah Naz and Elena Ballante, "Proper Bayesian Bootstrap for Bagging tree modelin survival analysis with correlated data." Conference: Italian Statistical Society - Statistical Learning, Sustainability and Impact Evaluation (2023). DOI: https://hdl.handle.net/11571/1524800

- Farah Naz, Simone Gerzeli, Elena Ballante, Silvia Figini, "LEARNING IN LOCKDOWNS: A FIVE-YEAR ANALYSIS OF COVID-19'S INFLUENCE ON UNIVERSITY STUDENTS' ACADEMIC EXPERIENCES." (2025) Advances and Applications in Statistics , 91(1), 59-75. DOI: https://doi.org/10.17654/0972361724005

- Farah Naz, Elena Ballante, Silvia Figini, "CovBootTree: Proper Bayesian Bootstrap Ensembled Trees with Cholesky Multivariate Distribution." Statistical Learning, Sustainability and Impact Evaluation, Springer Proceedings in Mathematics & Statistics 523, DOI: https://doi.org/10.1007/978-3-032-10630-8_5

- Elena Ballante, Farah Naz, Silvia Figini, Simone Gerzeli, "Customized Learning to Predict Student Dropout.", Advances and Applications in Mathematical Sciences, Volume 23, Issue 4, February 2024, https://hdl.handle.net/11571/1513098

- M. Croce, F. Naz, L. Barzaghi, M. Paoletti, T. Mongini, S. Gasperini, M. Filosto, L. Maggi, A. Sechi, M. Grandis, M. Sacchini, M. Sciacco, L. Vercelli, C. Bonizzoni, N. Bergsland, F. Santini, X. Deligianni, Gandini in Wheeler-Kingshott, S. Ravaglia, A. Pichiecchio, "Magnetization transfer imaging in late-onset Pompe disease." In: Neuromuscular Disorders WMS Congress Issue - Abstracts 2024. DOI: 10.1016/j.nmd.2024.07.740

- F. Naz, S. Figini, D.U Pizzagalli, "A Graph-based Clustering Algorithm for Bridge Problem Solution using Density Peak Assignment through Optimized Shortest Path." [Manuscript in preparation.]

Contents

Part I

Introduction and Study Background

1 Introduction

The best way to predict the future is to create it.

– Peter Drucker

Chapter Contents

1.1 Computational Techniques and Healthcare Data Analysis

The healthcare sector is experiencing an unprecedented transformation led by the exponential surge in digital health data and the advancement of computational techniques. Modern healthcare systems generate extensive amounts of complex, multi-dimensional data from diverse sources including electronic health records (EHRs), diagnostic imaging tools, wearable devices or fitness trackers, and genomic sequencing platforms. This digital revolution has created both tremendous opportunities and significant challenges

for healthcare practitioners, researchers, and data scientists aiming to extract meaningful insights from these comprehensive datasets. The use of artificial intelligence (AI) and machine learning (ML) techniques in healthcare has become a leading approach to address the complexity and scale of modern health data. These computational methodologies offer powerful tools for pattern recognition, predictive modeling, and decision support systems that can improve patient outcomes, make better use of medical resources, and enable personalized healthcare. The application of AI in healthcare spans numerous domains, from diagnostic imaging and drug discovery to public health management and clinical decision support.

One of the most promising aspects of computational healthcare analytics is its ability to handle heterogeneous data types and structures. Traditional statistical methods provide a robust foundation for healthcare analytics and have demonstrated significant importance in clinical research and health studies. However, the increasing complexity and scale of modern healthcare datasets characterized by high-dimensional data, intricate temporal patterns, and diverse data modalities create opportunities for complementary computational approaches that can capture additional insights and handle these multifaceted data characteristics. Cutting edge machine learning techniques, including deep learning architectures, ensemble methods, and unsupervised learning approaches, have demonstrated remarkable capabilities in addressing these challenges while maintaining interpretability and clinical relevance.

Moreover, the integration of traditional clinical data with other data sources, such as imaging biomarkers and wearable sensor data, presents unique opportunities for developing comprehensive health assessment tools. This multi-modal approach enables the design of more robust predictive models that can capture different aspects of health status and disease progression, ultimately resulting in improved diagnostic accuracy and personalized treatment strategies.

1.2 Research Challenges

Despite the tremendous potential of computational techniques in healthcare, several significant challenges persist that limit their adoption and effectiveness. These challenges span technical, methodological, and clinical domains, each requiring innovative solutions to achieve the maximum benefits of AI-driven healthcare data analysis.

1.2.1 High-Dimensional Data and Correlation Challenges

A significant challenge in processing medical data is its high-dimensional nature, where the number of variables often exceeds the sample size. This curse of dimensionality is particularly pronounced in genomic studies, medical imaging, and multi-sensor wearable data analysis. Furthermore, healthcare variables often exhibit complex correlation structures that can lead to overfitting, reduced model performance, and biased estimation of individual predictor effects.

The presence of highly correlated variables in healthcare datasets poses additional challenges for feature selection and model interpretation. Standard regularization techniques, including ridge and LASSO regression, may break down when dealing with strongly correlated variables, particularly when the number of irrelevant covariates is large. This challenge is especially critical in survival analysis, where the goal is to understand the relationship between predictor variables and time-to-event outcomes while accounting for the complex interdependencies among variables.

1.2.2 Subjectivity and Standardization in Clinical Assessment

Many clinical assessments rely on subjective evaluations that introduce variability and potential bias into diagnostic processes. This subjectivity is particularly problematic in the diagnosis of complex conditions such as connective tissue disorders, where current diagnostic criteria involve visual assessments and clinical judgment that may vary between practitioners. The lack of standardized, objective assessment tools limits the consistency and reliability of diagnoses, potentially leading to delayed or incorrect treatment decisions.

The challenge of subjectivity extends beyond individual assessments to encompass the broader issue of clinical data standardization. Healthcare facilities often use different protocols, measurement techniques, and documentation standards, making it difficult to develop generalizable models that can be applied across diverse clinical settings. This heterogeneity in data collection and processing limits the development of robust, transferable AI systems in healthcare.

1.2.3 Complex Data Structure Analysis and Clustering

Modern healthcare datasets often exhibit complex structural characteristics including irregular cluster geometries, overlapping regions, and varying density distributions that pose significant challenges for traditional analytical approaches. Conventional clustering methods frequently rely on rigid assumptions about data structure, such as globu-

lar cluster shapes or uniform density patterns, which may not adequately capture the heterogeneous nature of healthcare data. These limitations become particularly pronounced when analyzing biomedical datasets that contain nested structures, bridging patterns between patient groups, or datasets where meaningful subgroups exhibit non-convex boundaries.
The challenge lies in developing computational methods that can effectively identify and leverage local connectivity patterns while maintaining global optimization perspectives, enabling more accurate identification of clinically relevant patient subgroups and biological structures in complex, high-dimensional healthcare data.

1.2.4 Longitudinal Data Analysis and Crisis Impact Assessment

Healthcare systems face significant analytical challenges when attempting to quantify the effects of large-scale disruptions on organizational performance and outcomes. These sudden crises lead to the need for sophisticated analytical approaches that can detect meaningful temporal patterns, sudden shifts, and long-term trends that may indicate critical changes in not only healthcare sector but educational and other commercial sectors.
Traditional statistical approaches often struggle to distinguish between normal fluctuations in system metrics and genuine disruptions caused by external crises, making it difficult to assess whether observed changes represent temporary adaptations or fundamental shifts in operational effectiveness. This challenge is compounded by the requirement to incorporate both individual-level characteristics and institutional factors into analytical frameworks that can detect change points, quantify disruption magnitude, and provide evidence-based assessments of system resilience and recovery patterns during periods of organizational stress.

1.3 Research Objectives

This thesis addresses the aforementioned challenges through the development and application of novel computational techniques in four distinct but interconnected healthcare domains. The main goal is to advance the state-of-the-art in healthcare data analysis by developing robust, interpretable, and clinically relevant methodologies that can handle the complexity and diversity of modern health data.

1.3.1 Objective 1: Survival Analysis with Bootstrap Ensemble Methods

The first research objective focuses on addressing the challenges associated with highly correlated variables in survival analysis. Traditional survival models, particularly Cox regression, struggle with high-dimensional data containing strongly correlated predictors. This research aims to develop a novel tree-based Bayesian bootstrap ensemble method that incorporates bootstrap sampling techniques to mitigate the detrimental impact of correlated variables on survival analysis performance.
The specific goals include: (1) developing a methodology that captures and incorporates the multidimensional covariance structure of healthcare data through bootstrap sampling, (2) creating an ensemble approach that improves predictive performance and model stability in the presence of correlated variables, and (3) validating the approach through comprehensive simulation studies. This objective addresses the fundamental challenge of covariates in survival analysis while maintaining the interpretability and clinical relevance of the resulting models.

1.3.2 Objective 2: Graph-Based Clustering with Clique Detection and Path Optimization

The second research objective focuses on addressing the limitations of traditional density-based clustering methods when dealing with complex geometric structures and irregular cluster boundaries in healthcare data. This research aims to develop a novel graph-based clustering framework that integrates min-max clique detection with path-based density peak assignment to capture local connectivity patterns while maintaining global optimization approach.
The main objectives include: (1) developing a methodology that identifies densely connected subgroups through adaptive clique detection algorithms, (2) creating a hierarchical assignment strategy, and (3) implementing comprehensive noise handling through path backtracking mechanisms. This objective addresses the fundamental challenge of clustering irregular geometric structures while preserving the ability to identify meaningful subgroups and biological patterns in complex healthcare/non-healthcare datasets.

1.3.3 Objective 3: Developing Quantitative MRI Biomarkers for Early Detection of Neuromuscular Disease

The third research objective addresses the challenge of early detection and monitoring of muscle involvement in late-onset Pompe disease (LOPD), a progressive neuromuscular condition where timely intervention is critical. Current diagnostic approaches often

detect muscle damage only after significant fat replacement has occurred, limiting opportunities for early therapeutic intervention. This research aims to develop a quantitative MRI-based approach for early disease detection through magnetization transfer imaging (MTI) analysis.
The specific goals of the study include: (1) developing a multi-parametric MRI protocol combining magnetization transfer ratio (MTR) quantification with fat fraction (FF) assessment using Multi-echo Spin-echo (MESE) T2-weighted imaging and Multi-echo Gradient echo sequences, (2) establishing MTR as a sensitive biomarker capable of detecting mild muscle fiber damage before fat replacement occurs, and (3) validating the approach through clinical studies comparing LOPD patients at various disease stages (mild, moderate, severe) with healthy controls. This objective addresses the need for sensitive, non-invasive diagnostic tools that can detect early disease signs, monitor progression, and assess treatment efficacy in neuromuscular conditions.

1.3.4 Objective 4: Longitudinal Mixed-Effects Modeling with Change Point Detection

The research objective of this study focuses on developing comprehensive analytical frameworks for assessing crisis impacts on organizational performance through longitudinal data analysis. This research aims to establish a robust mixed-effects modeling framework that incorporates both individual-level and institutional-level factors while implementing change point detection mechanisms to identify structural shifts in performance patterns.
The specific goals include: (1) developing linear and generalized linear mixed-effects models that account for nested data structures and temporal dependencies, (2) implementing change point analysis techniques to detect significant shifts in system performance metrics, and (3) creating a methodological framework that can distinguish between normal temporal variation and genuine disruptions caused by external factors. This objective addresses the critical need for evidence-based crisis impact assessment tools that can quantify disruption magnitude and inform organizational resilience strategies across diverse institutional contexts.

1.4 Thesis Aim and Scope

This thesis provides a comprehensive investigation into the application of advanced computational techniques for addressing critical challenges in healthcare data analysis. The main aim is to develop and validate novel methodologies that can enhance the ac-

curacy, objectivity, and clinical utility of healthcare analytics across diverse application domains.

1.4.1 Primary Aim

The primary aim of the thesis is to advance the field of computational healthcare analytics by developing innovative machine learning and statistical methods that address fundamental challenges in modern healthcare data analysis. These challenges include the handling of high-dimensional correlated data, the reduction of subjectivity in clinical assessments, and the detection of temporal dynamics in sudden crisis. In mitigating these issues, this work aims to contribute to the development of more robust, interpretable, and relevant healthcare analytical methods.

1.4.2 Methodological Contributions

The thesis offers several methodological contributions to the advancements in the area of healthcare analytics. In the domain of survival analysis, the research introduces a novel bootstrap ensemble approach that effectively handles correlated variables while maintaining model interpretability and predictive performance. This contribution extends the applicability of survival analysis methods to multivariate healthcare datasets where traditional approaches may fail.

In the field of biomedical data analysis, the thesis develops a hybrid clustering methodology that integrates clique detection with path-based optimization for identifying complex cluster structures in healthcare datasets. This contribution addresses the critical challenge of analyzing datasets with irregular geometries, overlapping regions, and bridging patterns that traditional density-based methods fail to handle effectively.

In the area of clinical imaging biomarkers, the thesis presents an innovative quantitative MRI approach for early disease detection and monitoring through magnetization transfer imaging. This methodology combines advanced multi-parametric MRI techniques with quantitative analysis to create a sensitive, non-invasive tool for assessing neuromuscular disease progression. The approach demonstrates the potential for quantitative imaging biomarkers to enhance clinical decision-making by detecting muscle fiber damage at early disease stages, before conventional imaging markers such as fat replacement become apparent, thereby enabling timely therapeutic intervention and treatment monitoring.

The thesis develops a comprehensive mixed-effects modeling framework with integrated change point detection for analyzing institutional performance under disruptive conditions. This contribution addresses the essential need for quantitative methods that can

distinguish between normal temporal variation and genuine structural breaks in organizational metrics during crisis periods.

1.4.3 Clinical and Practical Impact

The scope of this thesis extends beyond methodological contributions to encompass practical applications that can have immediate clinical and practical impact. The developed methodologies are designed to be implementable in real-world healthcare and educational settings, with consideration for practical constraints such as computational resources, data availability, and clinical workflow integration.

The survival analysis methodology provides healthcare researchers and practitioners with tools for more accurate modeling of patient outcomes in the presence of complex variable relationships. This capability is particularly valuable in personalized medicine applications where patient-specific risk factors and treatment responses must be considered.

The quantitative MRI biomarker approach offers a sensitive, objective tool for neuromuscular disease assessment that can be deployed in various clinical settings, from initial diagnosis to longitudinal disease monitoring and treatment response evaluation. The multi-parametric imaging approach provides clinicians with quantitative metrics that detect early pathological changes before conventional imaging findings become apparent, potentially improving early intervention strategies, treatment timing decisions, and personalized monitoring protocols for patients with late-onset Pompe disease and other neuromuscular conditions.

This Cli-DSP methodology addresses the fundamental challenge of clustering irregular geometric structures while preserving the ability to identify clinically meaningful local connectivity structures and biological patterns in complex healthcare datasets.

This crisis assessment methodology addresses the critical need for evidence-based crisis impact assessment tools that can quantify disruption magnitude and inform organizational resilience strategies across diverse institutional contexts.

1.4.4 Interdisciplinary Approach

This thesis adopts an interdisciplinary approach that bridges computer science, statistics, and medicine. The research draws on techniques from machine learning, statistical analysis, and clinical research to develop comprehensive solutions to healthcare challenges. This interdisciplinary perspective ensures that the developed methodologies are both technically sound and clinically relevant.

The thesis also emphasizes the relevance of validation and clinical evaluation design-

ing healthcare data analytics solutions. Each proposed methodology is rigorously tested through appropriate validation studies, including simulation experiments, cross-validation techniques, and clinical trials. This emphasis on validation affirms that the research contributions are not only theoretically sound but also suitable for practical use in clinical and educational settings.

Through this integrated approach, the thesis focused on advancing the field of computational healthcare analytics while providing practical tools and methodologies that can improve patient outcomes and enhance the efficiency of precision medicines.

2 Background

Chapter Contents

2.1 Big Data in Healthcare

Big data in healthcare represents a transformative paradigm shift that is revolutionizing how medical professionals diagnose, treat, and prevent diseases while fundamentally altering the system of patient monitoring and health care delivery [1]. The healthcare sector produces vast amounts of data from diverse sources including electronic health records (EHRs), medical imaging, wearable devices, genomic sequencing, clinical trials, pharmaceutical research, insurance claims, and patient-generated health data from mobile applications and remote monitoring devices [2]. This exponential growth in data generation, characterized by the traditional 4 Vs: volume, velocity, variety, and veracity presents both unprecedented opportunities and impacting healthcare organizations worldwide [3]. Advanced analytics techniques, including machine learning methods, artificial intelligence, predictive modeling, and natural language processing, are being used to extract meaningful insights from these vast datasets, enabling precision medicine approaches tailored to each patient according to their genetic profiles, medical histories, and behavioral patterns [4].

The integration of big data analytics in healthcare has facilitated breakthrough discoveries in areas such as drug development, where researchers can identify potential therapeutic compounds more efficiently by analyzing molecular structures and patient responses across large populations, and epidemiological surveillance, where real-time

data analysis helps track disease outbreaks and inform public health interventions [5]. Clinical decision support systems powered by big data can assist physicians in making more accurate diagnoses by comparing patient symptoms and test results against comprehensive databases of medical knowledge and similar cases, while predictive analytics can identify patients at risk of developing chronic diseases or experiencing adverse events, allowing healthcare providers to act proactively to improve outcomes and reduce costs [6].

However, the application of big data solutions in healthcare faces substantial challenges, including data privacy and security concerns governed by regulations such as HIPAA, and/or GDPR (General Data Protection Regulation) interoperability issues between different healthcare systems and technologies, the need for standardized data formats and protocols, and the requirement for healthcare professionals to develop new skills in data interpretation and digital health technologies [7]. Furthermore, ethical considerations surrounding data ownership, algorithmic bias, and the potential for discrimination based on predictive models must be carefully addressed to ensure that big data initiatives enhance rather than compromise patient trust and healthcare equity [8].

Despite these challenges, the continued evolution of big data in healthcare promises to accelerate medical research, improve population health management, reduce healthcare costs through more efficient resource allocation, and ultimately deliver more personalized, effective, and accessible healthcare services to patients around the globe [9].

2.2 Learning from Data

Learning from data represents the fundamental paradigm underlying statistical methods to modern artificial intelligence and machine learning framework, encompassing the analytical extraction of knowledge, patterns, and insights from structured and unstructured datasets to enable intelligent decision-making and predictive capabilities [10]. This field has emerged and evolved from traditional statistical methods to sophisticated algorithmic approaches that can automatically identify complex relationships within vast amounts of information, fundamentally transforming how we approach problem-solving across diverse domains of healthcare, technology, and scientific research [11]. Therefore, advanced computational techniques including statistics, machine learning, deep learning and a combination of these techniques, are the fields of study that is focused on developing computer systems that automatically improve with experience [12]. In Figure 2.1, main types of techniques in data learning are presented, such as supervised learning, and unsupervised learning.

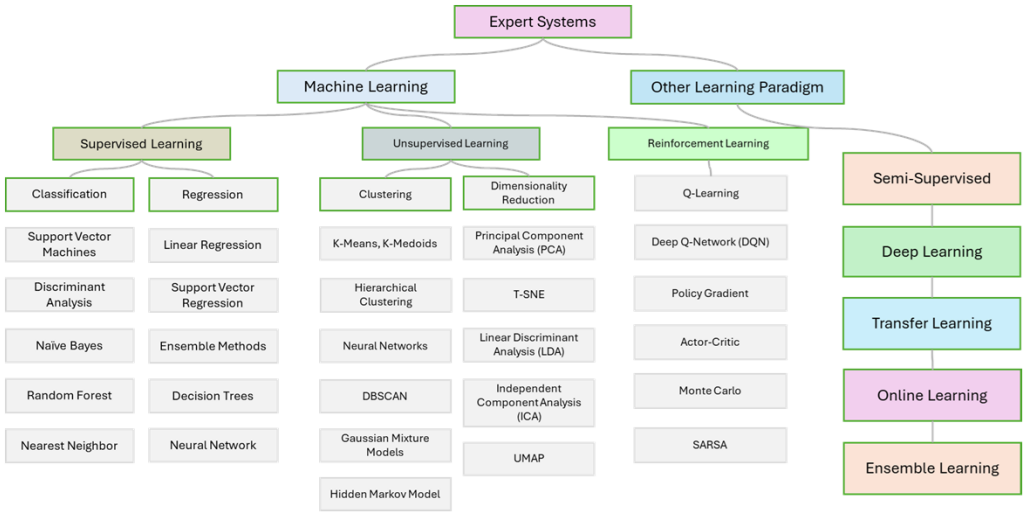

Figure 2.1: Learning Paradigm

2.2.1 Supervised Learning

Supervised learning serves as the most significant applied machine learning approach, where algorithms learn from labeled training data to make predictions on new, unseen examples [13]. This approach encompasses two primary tasks: classification, which focuses on assigning inputs to predefined classes, and regression, which involves predicting continuous numerical values [14]. Popular supervised learning algorithms includes linear regression for simple relationships, logistic regression for binary classification problems, decision trees for interpretable rule-based decisions, random forests for robust ensemble predictions, support vector machines for high-dimensional data separation, and neural networks for complex pattern recognition, each suited to different kinds of data depending on their structure, dimensionality, and specific problem requirements [15]. The performance of supervised learning relies heavily on the quality, quantity, and representativeness of labeled training data, with performance typically improving as more diverse and comprehensive examples are provided during the training process, though challenges arise when dealing with imbalanced datasets, noisy labels, or distribution shifts between training and test environments [16].

2.2.2 Unsupervised Learning

Unsupervised learning addresses the fundamental challenge of discovering hidden patterns and structures in unlabeled data without explicit target variables or desired outcomes, enabling organizations to uncover valuable insights from vast repositories of un-

structured information [17]. This analytical approach encompasses diverse techniques including cluster-based approaches such as k-means for partitioning data into distinct groups, hierarchical clustering for understanding data taxonomies, and DBSCAN for identifying clusters of varying densities, alongside techniques for reducing dimensionality like principal component analysis (PCA) for feature compression and t-distributed stochastic neighbor embedding (t-SNE) for high-dimensional data visualization [18]. Additional unsupervised learning applications include association rule mining for market basket analysis, anomaly detection for symptoms onset identification and system monitoring, and density estimation for understanding data distributions and identifying irregular trends or outliers that may indicate critical events or system failures [19]. The primary challenge in unsupervised learning lies in evaluating the quality and meaningfulness of results without ground truth labels, often requiring domain expertise, multiple validation approaches, and careful interpretation to assess whether discovered patterns represent genuine insights or algorithmic artifacts [20].

2.2.3 Transfer Learning

Transfer learning has emerged as a technique that makes use of learned insights from one domain or task to improve learning performance in related but different domains, particularly addressing the persistent challenge of limited labeled data availability in specialized applications. This knowledge transfer approach has proven especially valuable in deep learning applications where pre-trained models, such as those trained on massive image datasets like ImageNet containing millions of annotated images, can be fine-tuned for specific tasks with relatively limited domain-specific data, dramatically reducing training time and computational requirements while achieving superior performance [21]. Implementation strategies for transfer learning include feature extraction, where pre-trained models are used as fixed feature extractors to capture relevant data representations, and fine-tuning, where the entire or partial pre-trained network architecture is adapted to the new task through continued training with carefully adjusted learning rates to preserve learned features while adapting to new domains [22]. The success of transfer learning depends critically on the similarity between source and target domains of data and task, with greater conceptual and structural similarity typically resulting in more substantial performance improvements, though recent research has shown surprising effectiveness even across seemingly unrelated domains [23].

2.2.4 Ensemble Learning

Ensemble learning harnesses the collective intelligence of more than one individual models to create robust and more accurate predictors than any single model alone, operating on the fundamental principle that diverse models can compensate for each other's weaknesses and reduce overall prediction errors [24]. This meta-learning approach includes popular methods such as bagging (bootstrap aggregating), which trains multiple models on different subsets of the training data and combines their predictions through voting or averaging, and boosting, which trains models sequentially to specifically correct the errors of previous models, creating a strong learner from multiple weak learners [25]. Successful implementations of ensemble techniques include random forests, which combine multiple decision trees with random feature selection to reduce overfitting, gradient boosting machines that iteratively improve predictions through residual learning, and AdaBoost that adaptively weights training examples based on previous classification errors, all of which have achieved performance equivalent to state-of-the-art across various machine learning competitions and real-world applications [26], [27]. The effectiveness of ensemble methods stems from their ability to reduce overfitting through model diversity, improve generalization by capturing different aspects of the data distribution, and provide more stable and reliable predictions by leveraging the collective wisdom of multiple learning algorithms, though they require careful consideration of computational costs and model interpretability trade-offs [28].

2.2.5 Deep learning

Deep learning has revolutionized machine learning by using artificial neural networks with multiple layers to model and understand complex patterns in data, achieving unprecedented success in tasks that were previously considered intractable for traditional machine learning approaches [29]. This hierarchical learning framework has transformed numerous fields including computer vision, where convolutional neural networks (CNNs) now surpass human performance in image classification tasks, natural language processing, where transformer architectures have enabled breakthrough applications like machine translation and text generation, and speech recognition, where deep networks have made voice assistants and automated transcription systems ubiquitous [30]. The diversity of deep learning based architectures reflects the variety of data types and problem domains they address: convolutional neural networks excel at processing grid-like data such as images and videos, recurrent neural networks and their variants like LSTMs handle sequential data including time series and natural language, transformers have revolutionized attention-based processing for long-range de-

pendencies, and generative adversarial networks (GANs) have opened new frontiers in synthetic data generation and creative applications [31], [32], [33]. The transformative power of deep learning model lies in its ability to learn hierarchical representations of data automatically, eliminating the need for manual feature engineering while discovering complex non-linear relationships that traditional methods might miss, though this capability comes with challenges including interpretability concerns, substantial computational requirements, and the need for large amounts of training data [34].

2.3 Methods and Techniques

The use of machine learning and data analytics in precision medicine requires sophisticated methodological approaches to handle the complexity, heterogeneity, limited data size and high-dimensional nature of healthcare data. These techniques must address the unique challenges posed by medical datasets while ensuring clinical relevance and regulatory compliance in healthcare applications [35].

2.3.1 Data Interpretation

Data interpretation in precision medicine encompasses the systematic analysis and understanding of diverse healthcare datasets to extract clinically meaningful insights that can inform personalized treatment decisions and improve patient outcomes [36]. This process involves the integration of multi-modal data sources which includes electronic health records (EHRs) containing patient demographics, medical histories, and clinical notes, genomic sequencing data revealing genetic variations and mutations, medical imaging data from CT scans, MRIs, and pathology slides, laboratory test results providing biochemical markers, and real-time physiological data acquired from wearable devices and remote monitoring systems [37]. The interpretation challenge is compounded by the need to understand temporal patterns in patient data, where disease progression, treatment responses, and biomarker changes must be analyzed over time to identify critical intervention points and predict future health outcomes [38]. Advanced natural language processing techniques are increasingly employed to extract structured information from unstructured clinical notes, radiology reports, and pathology descriptions, enabling the transformation of free-text medical documentation into actionable data for computational analysis [39]. The interpretation process must also account for clinical context, where data values that appear abnormal in isolation may be normal for specific patient populations or disease states, requiring sophisticated domain knowledge integration and clinical decision support systems that can provide contextually appropriate

interpretations [40].

2.3.2 Pre-processing

Pre-processing healthcare data for precision medicine applications represents a critical and challenging phase that must address the inherent complexities of medical datasets while preserving clinical significance and ensuring data quality for downstream processes [41]. This comprehensive data pre-processing begins with data cleaning and quality assessment, where missing values, outliers, and inconsistencies commonly found in healthcare datasets must be identified and appropriately handled through techniques such as multiple imputation for missing values, outlier detection for physiological measurements, and data validation against clinical reference ranges [42]. Standardization and normalization procedures are essential for integrating data from multiple sources and time periods, including the harmonization of clinical terminologies using standards like SNOMED CT and ICD codes, the normalization of laboratory values across different measurement units and reference ranges, and the temporal alignment of patient data collected at irregular intervals [43]. Genomic data pre-processing requires specialized approaches including quality control of sequencing data, variant calling and annotation, population stratification correction, and the integration of multi-omics data types such as genomics, proteomics, transcriptomics, and metabolomics to create comprehensive molecular profiles for precision medicine applications [44]. Privacy protection and de-identification processes are paramount in healthcare data pre-processing, requiring the removal or masking of protected health information while preserving the analytical utility of the data through techniques such as differential privacy, synthetic data generation, and secure computation methods [45].

2.3.3 Feature Extraction

Feature extraction in precision medicine involves the systematic identification and derivation of relevant clinical, genomic, and phenotypic characteristics from complex healthcare datasets to create informative representations that can effectively support predictive modeling and therapeutic decision-making [46]. This process encompasses the extraction of clinical features from structured EHR data such as trends of vital signs, medication adherence patterns, comorbidity indices, and healthcare utilization metrics, while simultaneously deriving features from unstructured clinical text through natural language processing techniques that can identify symptom onsets, disease severity indicators, and treatment response patterns [47]. Genomic feature extraction presents unique challenges and opportunities, involving the identification of single nucleotide

polymorphisms (SNPs), gene expression signatures, and pathway-level features that can predict drug responses, disease susceptibility, and treatment outcomes through techniques such as genome-wide association studies (GWAS), polygenic risk scoring, and pharmacogenomic analysis [48]. Medical imaging feature extraction leverages advanced computer vision based and deep learning based techniques to automatically identify and quantify diagnostic features from radiological images, pathology slides, and other imaging modalities, enabling the extraction of quantitative imaging biomarkers that can complement traditional clinical assessments and provide objective measures of disease progression and treatment response [49]. The integration of multimodal features requires sophisticated feature selection and dimensionality reduction techniques to identify the most relevant subset of features while avoiding the curse of dimensionality, often employing methods such as principal component analysis (PCA), LASSO regression, and recursive feature elimination to create parsimonious yet comprehensive feature sets that can effectively capture the complexity of patient phenotypes [50].

2.3.4 Classification

Classification represents the core analytical challenge of developing predictive models that can accurately categorize patients into clinically meaningful groups, predict treatment responses, and identify disease subtypes to enable personalized therapeutic interventions [6]. This process involves the application of machine learning based algorithms to classify patients and healthy controls based on their clinical information, genetic profiles, and biomarker patterns, with applications ranging from cancer subtype classification using gene expression data to predicting drug response phenotypes based on pharmacogenomic markers and clinical characteristics [51]. The classification challenge in healthcare is complicated by class imbalance issues, where rare diseases or adverse drug reactions may be underrepresented in training datasets, requiring specialized techniques such as synthetic minority oversampling, cost-sensitive learning, and ensemble methods to achieve robust performance across all patient populations [52]. Deep learning approaches have shown particular promise in medical classification tasks, with convolutional neural networks achieving expert-level performance in medical image classification for real-world applications such as diabetic retinopathy screening, skin cancer detection, and pathology slide analysis, while recurrent neural networks and transformer models have demonstrated effectiveness in analyzing temporal clinical data for early disease detection and prognosis prediction [53], [54]. The validation and evaluation of classification models in precision medicine requires rigor-

ous approaches that account for the clinical context and potential biases, including the use of external validation datasets, temporal validation to assess model performance over time, and clinical impact assessment to determine whether model predictions actually improve patient outcomes and healthcare decision-making [55]. Interpretability and explainability of classification models are particularly crucial in healthcare applications, where clinicians need to understand the reasoning behind model predictions to make informed decisions, leading to the development of interpretable machine learning approaches and post-hoc explanation methods that can provide insights into the biological and clinical factors driving classification decisions [56].

Part II

Computational Approaches for Medical and Academic Data Analysis

3 Ensemble Methods for Survival Analysis

Chapter Contents

3.1 Introduction

The presence of highly correlated covariates in survival analysis introduces a cascade of analytical challenges that traditional methods often fail to address adequately. The intersection of survival analysis, limited data size and correlated data represents a critical area where methodological innovation is urgently needed.

3.1.1 The Challenge of Correlated Predictors in Survival Analysis

Highly correlated predictors often create instability in model selection and parameter estimation. Traditional ensemble methods, which rely on bootstrap resampling from the observed data, may fail to capture the true multidimensional distribution of correlated variables [57]. This limitation is particularly problematic when the correlation structure contains important information about the underlying biological or clinical processes driving survival outcomes.

The bootstrap procedure, while powerful for handling uncertainty and improving model stability, faces unique challenges when applied to correlated survival data. Standard bootstrap methods independently resample observations, potentially breaking the correlation structure present in the original data [58]. This disruption can lead to bootstrap samples that inadequately represent the true covariance structure, resulting in ensemble models that fail to capture the complex interdependencies among predictors.

3.1.2 Bayesian Nonparametric Approaches and Bootstrap Methods

The Proper Bayesian Bootstrap, introduced by Muliere and Secchi [59], represents a significant advancement by allowing the incorporation of prior information into the resampling scheme. This approach uses a Dirichlet process prior $D(kF_o)$ for the unknown distribution F, where F_o represents prior knowledge about the distribution and k quantifies the confidence in this prior information [60].

The theoretical foundation of the Proper Bayesian Bootstrap offers several advantages over conventional resampling methods. By incorporating prior distributions, the method can generate synthetic observations that complement the original dataset, potentially improving model generalization and reducing overfitting [61]. When k = 0, the method reduces to Rubin's Bayesian bootstrap, demonstrating its flexibility and theoretical consistency with existing approaches.

However, traditional bagging approaches for survival trees may not fully exploit the correlation structure present in the data, particularly when predictor variables exhibit strong interdependencies [62].

3.1.3 Methodological Innovations and Contributions

The papers presented in this chapter introduce significant methodological innovations that address the challenges of analyzing survival data with highly correlated predictors and the limited sample size of the data. The first paper, "Proper Bayesian Bootstrap for Bagging Tree Model in Survival Analysis with Correlated Data", establishes the founda-

tional framework for applying Proper Bayesian Bootstrap to ensembles decision trees. This work demonstrates how the incorporation of prior information through the Dirichlet process can improve model stability and predictive performance, particularly in small sample settings where traditional methods may be unreliable [63].
The key innovation lies in the method's ability to generate new observations through the prior distribution F_o, which enhances the diversity of the ensemble while maintaining the underlying correlation structure. The bootstrap aggregated conditional survival function provides a coherent framework for combining predictions from multiple trees, each trained on bootstrap samples that respect the covariance structure of the original data [64].
The second paper, "CovBootTree: Proper Bayesian Bootstrap Ensembled Trees with Cholesky Multivariate Distribution", extends this framework by explicitly incorporating the correlation structure through Cholesky decomposition of the covariance matrix. This advancement addresses a critical limitation of the initial approach by ensuring that bootstrap samples preserve the multidimensional distribution characteristics of highly correlated variables [65].
The use of Cholesky decomposition for multivariate normal sampling represents a sophisticated approach to maintaining covariance structure during bootstrap resampling. This technique ensures that synthetic observations generated from the prior distribution exhibit realistic joint distributions, preventing the masking effects that can occur when correlated variables are treated independently during resampling [66].
Among the methods proposed for high-dimensional survival data with correlated covariates, penalised Cox regression approaches such as elastic net Cox regression and group LASSO have received considerable attention [67], [68]. These methods address multicollinearity through regularisation, shrinking or grouping correlated predictors to improve model stability. However, they remain fundamentally tied to the proportional hazards assumption and do not incorporate prior distributional knowledge through a Bayesian framework. In contrast, CovBootTree directly models the covariance structure of correlated covariates through the bootstrap resampling process, offering a nonparametric alternative that does not rely on linearity or proportionality assumptions.

3.1.4 Theoretical Framework and Practical Implications

The integration of Dirichlet process framework and the proper Bayesian bootstrap with ensemble tree methods creates a powerful framework for survival analysis that can handle both censored observations and complex correlation structures.
The practical implications of these methodological advances are substantial. In clinical

research, where predictor variables often exhibit strong correlations due to underlying biological processes, the ability to maintain and exploit these relationships can significantly improve prognostic accuracy [69]. The enhanced stability demonstrated by these methods, particularly in small sample settings, makes them particularly valuable for specialized medical research where large datasets may not be available.
The computational efficiency of tree-based methods, combined with the theoretical rigor of Bayesian nonparametric approaches, creates a practical framework that can be implemented in real-world research settings. The bootstrap aggregation strategy provides natural uncertainty quantification, which is crucial for clinical decision-making applications [70].

3.1.5 Research Questions and Objectives

The papers in this chapter address several fundamental research questions in survival analysis with correlated data. How can bootstrap methods be adapted to preserve correlation structure in survival data? What are the theoretical properties of Proper Bayesian Bootstrap when applied to ensemble tree models? How does the explicit incorporation of covariance structure through Cholesky decomposition affect model performance and stability?
These questions are addressed through comprehensive simulation studies that evaluate model performance across different sample sizes, correlation structures, and censoring patterns. The comparison with established methods such as the Cox proportional hazards and Survival Random Forest [71] provides context for understanding the relative advantages of the proposed approaches.

3.1.6 Structure and Organization

The papers in this chapter represent a progression from foundational concepts to advanced methodological developments. The first paper establishes the basic framework and demonstrates its effectiveness through initial simulation studies. The second paper extends this framework by incorporating more sophisticated treatment of correlation structure, providing a comprehensive solution to the challenges of bootstrap resampling with correlated survival data.

Paper I: Proper Bayesian Bootstrap for Bagging tree model in survival analysis with correlated data

Proper Bayesian Bootstrap for Bagging tree model in survival analysis with correlated data

Farah Naz[a] and Elena Ballante[b,c]

[a]Department of Mathematics, University of Pavia, Italy ,
farah.naz01@universitadipavia.it
[b]Department of Political and Social Sciences, University of Pavia, Italy
[c]IRCCS Mondino Foundation, Pavia, Italy

Abstract

The health sciences often involve survival data that may be censored and can contain correlated covariates. While there has been some research on the impact of correlated variables on survival models, there is a need for further investigation of how bootstrap methods can be used to handle correlation in survival analysis. In fact, if the variables are strongly correlated, the bootstrap samples from the prior may mask the effect of each other, making it difficult to discern the true relationship between the variables and the response, which can ultimately lead to unrealistic estimates. This article aims to extend the Proper Bayesian bootstrap ensemble tree model for analyzing survival data with highly correlated covariates. The model's performance was assessed through a simulated study, demonstrating better results compared to traditional survival models, such as the Cox model and survival random forest, with greater stability in terms of the integrated Brier score, particularly with smaller sample sizes.

Keywords: Survival analysis, Bootstrap, Bayesian nonparametric learning, Ensemble models, Correlated data

I.1 Introduction

Multivariate responses that are highly correlated are frequently encountered in health science studies. In such cases, a subject may have multiple related responses, and the presence of these correlated variables can have a significant impact on survival models. The adverse effects of highly correlated variables include overfitting, reduced model performance, and biased estimation of individual predictor effects.

In cases where the data comprises highly correlated variables, the survival model may exhibit bias towards splitting one of the correlated variables while disregarding information from the other correlated variables [72]. This can result in overfitting of the model, leading to a reduction in its predictive performance, as the model may fail to accurately capture the true underlying relationship between the variables and the time-to-event outcome (such as the expected survival time or the probability of survival beyond a specific landmark time-point).

This paper aims to enhance the use of the Proper Bayesian bootstrap ensemble tree model for analyzing survival data by incorporating a pre-processing stage. This stage involves carefully managing the data by either removing or transforming highly correlated variables, to reduce the detrimental impact that high correlation can have on the analysis.

The model performs Bootstrap resampling techniques to approximate the posterior distribution of a statistical function of a decision tree $\phi(F)$, where F is a random distribution function as defined in [73].

The proposed method is rooted in the family of Bayesian bootstrap procedures, starting with Rubin (1981) [74], followed by Muliere and Secchi's proper Bayesian bootstrap (1996) [73], and Lo's Bayesian bootstrap for censored data (1993) [64]. The algorithm also draws upon Efron's classical bootstrap technique [75] for bagging survival trees and survival forests.

The approach inherits the advantages of Bayesian non-parametric learning such as flexibility and computational strength and considers prior opinions to overcome the drawbacks of traditional bootstrap procedures used in classical ensemble decision tree models.

The structure of the paper is as follows: Section I.2 details the proposed methodology, while Section I.3 and I.4 provide information on the computational environment and present the initial findings, respectively. Finally, Section I.5 summarizes the main outcome of the study and discusses the potential areas for future research.

I.2 Proposed Model

This paper aims to extend the idea of a Proper Bayesian Bootstrap Ensemble Tree model, specially tailored for survival data analysis. The model is designed to address the challenges posed by highly correlated variables and their impact on the model's performance. The proposed model is based on the initial outcomes of the Proper Bayesian Bootstrap concept, which was previously introduced by [73]. However, it has been adapted and modified to suit the specific requirements of survival data analysis.
The primary metric for evaluating ensemble tree models is the decision tree $\phi(F, X)$, which depends on the underlying distribution F and the observed data X. According to a study by [76], the posterior distribution of Φ can be estimated using bootstrap procedures. This is achieved by fitting the model to a weighted dataset generated from the bootstrap process, and the resultant predictions provide an estimate of the posterior mean.
We set a prior distribution $D(kF_0)$ for F using a Dirichlet process to apply the Proper Bayesian bootstrap. To explain the response variable y based on a list of highly correlated covariates $x_1, ..., x_P$, the parameter F_0 of the Dirichlet process is established as a joint distribution that depends on both x and y.
The bootstrap sampling method from the posterior of $\phi(F, X)$ is taken from [76], where each bootstrap resample $(x_1^*, y_1^*), ..., (x_m^*, y_m^*)$ is created by combining distributions from the prior estimate F_0 and the empirical distribution F_n. Since the covariates are highly correlated, in the bootstrap resampling process, a new sample is generated from the original distribution F_0 and a new vector of covariates is produced from the original prior distributions $F_0(x_k)$ that accounts for the dependence among the covariates. The resampling process continues iteratively until the desired number of bootstrap samples is generated. This approach can help to account for the dependence among the covariates and produce more accurate estimates of the parameters of interest.
The response variable y is linked to the vector of covariates generated from the prior F_0 estimate obtained using an appropriate survival model to perform survival analysis. The method for combining the predictions should be based on the characteristics of the time-to-event data. Such as suggested in [77] the output of the model is a bootstrap aggregated version of the estimated conditional survival function S for a new observation X_{new} computed by $\widehat{S}_A^B(\cdot|x_{new}) = \widehat{S}_{L^B_{A(x_{new})}}(\cdot)$.
The distinguishing feature of the proposed method from traditional ensemble methods is its ability to generate new observations through the prior distribution F_0, which improves the prediction of the model.
Moreover, as a novelty element with respect to [78], we incorporate the correlation

structure between covariates into the bootstrap resampling technique. With this modification, it takes into account the covariance matrix of the covariates which prevents the bootstrap samples coming from the prior distribution from having an unrealistic joint distribution. Incorporating this feature has the potential to produce more robust and reliable results, particularly in situations where there are strong correlations among the covariates.

I.3 Experimental setting

The data simulation is done using the flexible-hazard method, as outlined in [79]. Also, to demonstrate the capabilities of the proposed method, we investigated the result of different weights assigned to the prior distribution F_0 with sample size $N = 50$. The simulated dataset generated for the time-to-event target variable, including 10% censored observations, comprises five highly correlated numerical covariates that are sampled from a multivariate normal distribution with adjusted parameters of mean and standard deviation. The correlation coefficients between covariates are higher than 0.85.
A total of 100 simulated datasets were produced, each consisting of $N = 50$ samples. The survival time values for observations sampled from the prior were estimated using an exponential regression model.
The proposed model was contrasted against two of the most prevalent models in survival analysis, the Survival Random Forest, and the Cox model. The performance of the predictions was evaluated using the Integrated Brier Score (IBS) through a 5-fold cross-validation process.

I.4 Preliminary results

The average values and nonparametric confidence intervals of the IBS results for the comparison with classical models are displayed in Figure 3.1. Figure 3.1 shows the performances of multivariate sampling with respect to the indipendent one. The results in Figure 3.1 showed that the average IBS score for the proposed Proper Bootstrap model was lower, but not significantly lower when considering the confidence interval widths of the traditional Cox and Random Forest models (CI width: Cox ≈ 0.62, Random Forest ≈ 0.67), which were approximately 8–9 times wider than those of the proposed Proper Bootstrap models (CI width ≈ 0.05–0.08). The confidence intervals of the proposed model were significantly narrower, indicating that the proposed method is more consistent in its prediction performance.

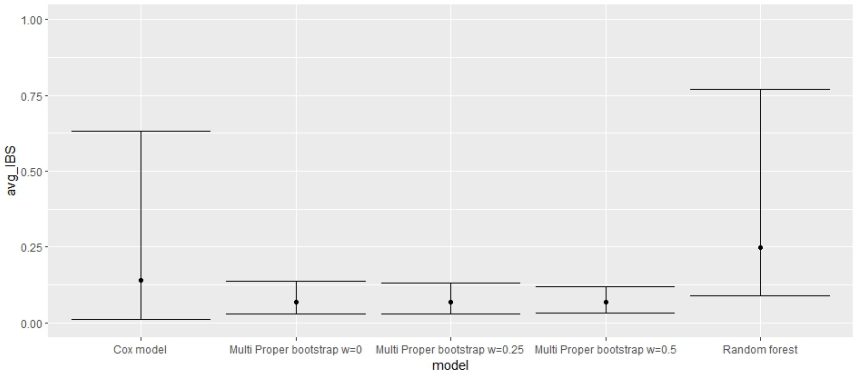

Figure 3.1: Comparison of mean and nonparametric confidence intervals for IBS obtained in cross-validation for the 100 simulated datasets of highly variable covariates of sample size 50. The proposed model is compared with Random Survival Forest and Cox Model

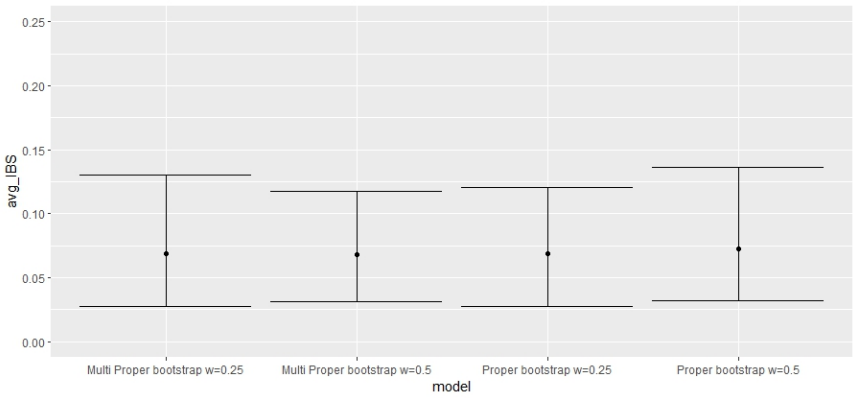

Figure 3.2: Comparison of mean and nonparametric confidence intervals for IBS obtained in cross-validation for the 100 simulated datasets of highly variable covariates of sample size 50. Multidimensional prior sampling is compared to the independent one

The results in Figure 3.2 show that the application of multivariate prior sampling, that takes into account the covariance structure of the covariates, leads to slightly better results than the independent sampling when the covariates are highly correlated.

I.5 Conclusion

The paper introduces a novel ensemble tree modeling approach that utilizes Proper Bayesian Bootstrap, to analyze survival data while mitigating the effect of highly correlated variables to obtain increased stability and comparable results in a simulated environment.
The use of synthetic data, derived from prior distributions that are not present in the original dataset, overcomes the drawbacks of classical ensemble models which only consider the data without any prior opinion. This as a result enhances the stability of the final ensemble model, particularly for datasets with limited sample sizes. Taking into account the covariance structure of the data at hand prevent the proper Bayesian bootstrap from generating unrealistic data that could lead to more noisy results.
Further research is planned to evaluate the model's sensitivity towards varying degrees of censored data, categorical variables, and more effective techniques for sampling correlated data variables, derived from non-normal distributions.

I.6 References

References for this paper are listed at the end of the thesis.

Paper II: CovBootTree: Proper Bayesian Bootstrap Ensembled Trees with Cholesky Multivariate Distribution

CovBootTree: Proper Bayesian Bootstrap Ensembled Trees with Cholesky Multivariate Distribution

Farah Naz[a], Elena Ballante[b,c], and Silvia Figini[b]

[a]Department of Mathematics, University of Pavia, Italy,
farah.naz01@universitadipavia.it
[b]Department of Political and Social Sciences, University of Pavia, Italy
[c]IRCCS Mondino Foundation, Pavia, Italy

Abstract

The health-science driven datasets present several common challenging characteristics for their analysis, notably the presence of censored observations and highly correlated covariates. When data variables exhibit strong correlations, bootstrap samples from the prior distribution can obscure individual effects, posing a challenge to accurately discern the true relationship between covariates and the response variable. This phenomenon can result in unrealistic estimates, underscoring the importance of incorporating the correlation structure into the modeling process. Therefore, analyzing survival data requires specialized techniques tailored to handle these distinctive characteristics. The impact of censored data observations and the possible solutions to this problem have been well studied within survival models, but there remains a significant gap in exploring methodologies that effectively account for the multidimensional distribution structure of highly correlated variables. This article aims to apply the Proper Bayesian bootstrap, proposed by Muliere and Secchi, used in sampling the posterior distribution over ensemble trees incorporating the prior distribution of highly correlated variables for analyzing survival data. The model's performance is assessed through a simulated study, and the results are compared to traditional survival models, such as the Cox model and Survival Random Forest, in terms of the average Brier score, particularly with varying sample sizes.

Keywords: Survival analysis, Bootstrap, Bayesian nonparametric learning, Ensemble models, Correlated data

II.1 Introduction

The relative effects of more than one explanatory variable on event times are commonly encountered in health science research. In such cases, a subject may have multiple interrelated covariates and the degree of their correlation can significantly impact survival models. In survival data analysis, high correlation among variables generally poses challenges and is often considered detrimental [80]. In 1989, Wei, Lin, and Weissfeld [81] presented a multivariate problem when during a clinical trial blood samples were drawn from different patients in three consecutive post-treatment months and samples were studied to find markers for viral infection. The study was focused on some conceptually straightforward semiparametric approaches to the analysis of general multivariate failure time data.

In the literature on multivariate problems in biomedical sciences, among several proposed methods for handling covariance matrices, variance-correlation, spectral, and Cholesky decompositions are used frequently. The adverse effects of highly correlated variables include overfitting, reduced model performance, and biased estimation of individual predictor effects. In the research field for survival analysis, various models are being used to explore the relationship between predictor variables and time-to-event outcomes, e.g., proportional hazards modeling, which assesses how the hazard rate changes in response to predictor variables [82]. However, when data are high-dimensional (e.g., when the number of predictors exceeds the sample size), it is impossible to fit a Cox regression model including all available covariates. A solution to this problem suggested by Welchowski [65] is to use regularized methods such as ridge-penalized Cox regression. However, even these methods often break down when the data comprises highly correlated variables, especially when considering a large number of non-influential covariates [57]. One common practice is to perform data-driven variable selection before fitting the survival model so that the model does not exhibit bias toward splitting one of the correlated variables while disregarding information from the other correlated variables [72]. However, developing a survival model that accounts for the correlation between variables, which goes beyond mere feature selection, is of substantial importance.

Various methods in statistical analysis touch on the essence of understanding data structures in a deep way. Among these methodologies, the concept of bootstrapping emerges as a powerful technique that transcends traditional sampling approaches [58]. What sets bootstrapping apart is its ability to tap into the multidimensional nature of data. It allows for the construction of a multitude of samples, thereby creating an ensemble that encapsulates the covariance structure of the data set. This methodology

becomes particularly influential when confronted with limited sample availability from a population, as it intricately captures and replicates the underlying relationships and intricacies present within the data covariance structure [61]. Therefore, the bootstrapping method incorporating a multidimensional distribution structure of covariates can help understand key aspects of statistical analysis. This approach not only acknowledges but actively incorporates the intricate covariance structure of the data, providing a refined and comprehensive understanding that traditional sampling methods might overlook.

Based on the potential of bootstrap sampling, we developed a tree-based Bayesian bootstrap ensemble method to incorporate bootstrap sampling along with highly correlated variables to reduce their detrimental impact on analysis. Based on the results achieved in the simulated datasets, our approach shows a reliable gain in predictive performance and stability of the model.

The structure of the paper is as follows: Section II.2 reports a comprehensive review of the literature on survival models and the way the bootstrap procedure is applied to survival data analysis. Section II.3 details the proposed methodology, while Sections II.4 and II.5 provide information on the computational environment and present the initial findings, respectively. Finally, Section II.6 summarizes the main outcome of the study and discusses potential areas for future research.

II.2 Literature Review

Exploring the multidimensional nature of survival data through comprehensive statistical methods, particularly those capturing covariance structures, is an imperative methodological approach. Methods that solely evaluate covariates individually may overlook the convoluted relationships and interdependencies present within the dataset, leading to potential information loss and the selection of 'antagonistic' covariates with conflicting effects. While univariate screening methods, as interpreted by Fan and Lv [69], provide theoretical justification in identifying influential covariates under mild regularity conditions, they tend to overlook correlations among covariates. However, multivariate analysis techniques [83] play a key role in addressing correlated covariates, crucial in survival modeling, where dependencies between covariates often exist.

In survival analysis, adaptations of Bagging methods [84] have shown promise, especially in handling correlated data structures and improving predictive performance. One method commonly applied in survival analysis involves bagging tree models, providing stability in measuring variable importance and aiding in robust variable selection [85].

However, when considering clinically relevant outcomes such as survival or treatment response, covariates with high correlations can exhibit similar prediction performances despite having few or no common variables [85]. This challenge arises due to the convoluted relationships among covariates, which complicates their effectiveness using traditional approaches. One such approach for multiple tests, exemplified by Benjamini and Hochberg [86], aims to identify informative covariates. However, despite attempts to identify relevant predictors, the complexities inherent in capturing the full covariance structure within the data remain a significant challenge in survival analysis.
Traditional methods may pose the risks of model overfitting and compromise the predictive accuracy associated with the time-to-event outcome. Therefore, focusing on strategies that encompass multidimensional distributions to effectively account for the inherent covariance structure within the data becomes crucial for a more comprehensive understanding of complex relationships.
Bootstrap was first proposed by Efron [87] as a method to determine the accuracy of a statistical estimate. The idea of Bootstrap is to use random sampling with replacement to simulate many observations from a population for which we, in reality, only have one sample. The bootstrap independently generates each bootstrap resample X_i^*, where $i = 1, \dots, n$ from the empirical distribution F_n of samples X_i, for $i = 1, \dots, n$. This procedure is equivalent to draw, for each bootstrap replication, a weights vector w for the observations X_i from a Multinomial distribution with parameters $(n, \frac{1}{n}\mathbf{I_n})$, where $\mathbf{I_n}$ is the identity matrix of dimension nxn.

$$F^*(x) = \sum_{i=1}^{n} \frac{w_i}{n} \mathbf{I}_{[\mathbf{X_i \leq x}]} \tag{3.1}$$

where, $(w_1, \dots, w_n) \sim \text{Mult}(n, \frac{1}{n}\mathbf{I_n})$ and $\mathbf{I}_{[\mathbf{X_i \leq x}]}$ is the indicator function. As a result, it generates a bootstrap replication with replacement of the original sample.
Rubin's Bayesian Bootstrap (1981) [74] provided a Bayesian perspective to resampling techniques, addressing model uncertainty. In Rubin's bootstrap, the distribution F is approximated by:

$$F^*(x) = \sum_{i=1}^{n} w_i \mathbf{I}_{[\mathbf{X_i \leq x}]} \tag{3.2}$$

where, w_i and X_i are two random independent vectors and $w_i \sim \text{D}(\mathbf{I_n})$ is the indicator function. In a nutshell, Efron's and Rubin's bootstraps estimate the conditional probability of a new observation considering only the observed data.
Gong's pioneering work in adapting bootstrapping for variable selection in the early eighties [88] showcased the potential of this methodology. By deploying stepwise se-

lection and constructing multiple bootstrap replicates, Gong highlighted the variance in included covariates across these samples. The findings underscored the dynamic nature of the data's covariance structure, as evidenced by the fact that the number of covariates included in the model varied a lot between different bootstrap replicas.
Similarly, Altman and Andersen's exploration of the stability of a Cox PH model using bootstrap samples [71] further reinforced the significance of accounting for the multidimensional distribution within the data. Having performed a stepwise selection for 100 bootstrap samples, they found that the most frequently chosen covariates were those in the original analysis. They then concluded that there was no reason to doubt the original model.
The introduction of Random Survival Forests by Ishwaran and Kogalur (2010) [89] expanded Random Forests to censored survival data, offering robust and flexible modeling techniques. Using ensemble learning principles, this approach has gained importance for its ability to handle complex, high-dimensional data sets and capture nonlinear relationships, improving predictive accuracy in survival analysis settings. Rubin's Bayesian Bootstrap (1981) [74] has significant implications in survival analysis, particularly in quantifying uncertainty and incorporating it into predictive modeling, a crucial consideration in real-world applications.
Assessing the strength of bootstrap sampling, this paper develops a tree-based Bayesian bootstrap ensemble method to incorporate bootstrap sampling along with highly correlated variables to reduce their detrimental impact on analysis. The model performs bootstrap resampling to approximate the posterior distribution of a functional, in this case a decision tree, $\phi(F)$, where F is a random distribution function as described in [73].
The main reason for this finding is that many survival data variables are correlated and therefore carry redundant information regarding prediction [66]. The proposed method is rooted in the family of Bayesian bootstrap procedures, starting with Rubin (1981) [74], followed by Lo's Bayesian bootstrap for censored data (1993) [64], and Muliere and Secchi's proper Bayesian bootstrap (1996) [73]. Since Efron's and Rubin's bootstraps are strongly dependent on the observed values and do not take into consideration any prior opinions, the main innovation of the proposed method, that gives it the name of *proper* Bayesian bootstrap, is the opportunity of integrating prior information into the resampling scheme.
Following the work of Ferguson in [90], they define the prior of F as a Dirichlet process $D(kF_0)$ where F_0 is a proper distribution function and k represents the level of confidence in the initial choice F_0. The resulting posterior distribution for F, given a sample of $X_1, ..., X_n$ from F, is still a Dirichlet process with parameter $(kF_0 + nF_n)$. The bootstrap

procedure consists of the resampling from the posterior, so when $k = 0$ the procedure is equivalent to the Rubin procedure. This bootstrap method allows one to introduce explicit prior knowledge on the data through the choice of F_0 and k. It is important to note that, since ϕ is a function of F, an informative prior on F is an informative prior on ϕ. See more details in [73] and [91].
The methodological approach introduced in this work inherits the advantages of Bayesian non-parametric learning, such as flexibility and computational strength, and considers prior opinions to overcome the drawbacks of traditional bootstrap procedures used in classical ensemble decision tree models.

II.3 Proposed Model

This paper aims to extend the idea of a Proper Bayesian Bootstrap Ensemble Tree model, specially tailored for survival data analysis, in a context where taking into account the correlation between covariates is fundamental. The model is designed to address the challenges posed by the multidimensional distribution of highly correlated variables and their impact on the model's performance. The proposed model is based on the initial results of the Proper Bayesian Bootstrap concept, which was previously introduced by Muliere and Secchi [73].
Moreover, as a novelty element with respect to the first tentative to extend incorporating proper Bayesian bootstrap ensemble tree models for survival analysis [78, 92], we incorporate the correlation structure between covariates into the bootstrap resampling technique. Incorporating this feature, the method has the potential to produce more robust and reliable results, particularly in situations where there are strong correlations among the covariates.
In the context of ensemble tree models, the statistic of interest is the decision tree $\phi(F, X)$, which depends on the underlying distribution F and the observed data X

$$L(\phi(F, X)|X_1, ..., X_n) \tag{3.3}$$

where $X_1, ..., X_n$ is the sequence of random variables. According to a study by [76], the posterior distribution of Φ can be estimated using bootstrap procedures. This is achieved by fitting the model to a weighted dataset generated from the bootstrap process, and the resultant predictions provide an estimate of the posterior statistic.

$$L(\phi(F^*, X)|X_1, ..., X_n) \tag{3.4}$$

where F^* is obtained using the bootstrap sampling techniques, such as Efron's, Rubin's and Proper Bayesian Bootstrap.
Following the definition of proper Bayesian bootstrap [73], we set a prior distribution $D(kF_0)$ for F using a Dirichlet process, where F_0 is a baseline distribution encoding prior knowledge about the data and k > 0 represents the degree of confidence placed in this prior relative to the observed data. When k = 0, the procedure reduces to Rubin's Bayesian bootstrap, relying entirely on the observed sample. To explain the response variable y based on a list of highly correlated covariates $x_1, ..., x_P$, the parameter F_0 of the Dirichlet process is established as a multivariate distribution on the covariates with a non diagonal covariance matrix.
The bootstrap sampling method from the posterior of $\phi(F, X)$ is taken from [76], where each bootstrap resample $(x_1^*, ..., x_m^*)$ is created by sampling from a realization of the Dirichlet process.

$$D(kF_0 + nF_n) \tag{3.5}$$

To obtain this result, we first sample the mixture distribution

$$G_0 = \frac{k}{n+k}F_0 + \frac{n}{b+k}F_n \tag{3.6}$$

As a result, the new sample is generated by sampling partly from the original distribution F_n, randomly selecting observations from the original dataset and partly generating synthetic data. For this second task, a multivariate normal distribution with Cholesky Decomposition is used to generate new vector of covariates produced from the prior distributions $F_0(x_k)$ tanking into account the dependence among the covariates. The resampling process continues iteratively until the desired number of bootstrap samples is generated. This approach can help accounting for the dependence among the covariates and produce more accurate estimates of the parameters of interest.
The response variable y given the vector of covariates generated from the prior F_0 is assigned using a Weibull regression model in order to preserve the relationship between $\mathbf{X}$ and y. As in [77], the bootstrap aggregated conditional survival function $\widehat{S}$ for a new observation $\mathbf{x}_{new}$ is computed by

$$\widehat{S}_A^B(\cdot|\mathbf{x}_{new}) = \widehat{S}_{L_A^B(\mathbf{x}_{new})}(\cdot) \tag{3.7}$$

where the learning sample L_A^B is the aggregated sample $L_A^B = [L^1, ...L^B]$. In this way, we aggregate the observations from each leaf where $\mathbf{x}_{new}$ fell directly and compute a single predictor only for the aggregated sample.

The distinguishing feature of the proposed method from traditional ensemble methods is its ability to generate new observations through the prior distribution F_0 by sampling random values using the multivariate normal distribution with Cholesky decomposition, which preserves the covariance structure of the simulated data and helps test the generalizability of the model on unseen data.

II.4 Experimental setting

To show the performance of the proposed model, a simulated study is performed. We simulated datasets with characteristics mirroring real-world scenarios. A data set with 15 numerical covariates is generated using multivariate normal distributions. These covariates were used to construct survival data via a Weibull distribution, associating survival times with different values of betas associated to each covariate, ensuring the simulation of correlated data that mimics correlation patterns between variables in the time-to-event dataset.

Data simulation is performed in four different sample size settings, allowing for a complete comparison of model performance. A total of 50 simulations were performed, for each survival data sample of 50, 100, 300, and 500 respectively. The survival time values for observations sampled from the prior were estimated using an exponential regression model.

The generated datasets in all four sample sizes include 10% censored observations. As a novelty element, to generate bootstrap resampling, two different sampling strategies are adopted. The first sampling approach involves randomly selecting samples from the original dataset (training data) to generate bootstrap samples where each sample is selected with replacement. In the second approach, synthetic data generation is performed using multivariate normal distribution with Cholesky Decomposition to generate correlated samples that preserve the covariance structure.

The models were trained and tested for predictive performance, providing insights into the relationship between covariates and event occurrence. Furthermore, the proposed model was tested against two of the most prevalent models in survival analysis, the Survival Random Forest and the Cox model.

To enhance model stability and reliability, we implemented a Bootstrap aggregation strategy for both Cox and Random Forest models. This involved iterative resampling, varying the sizes of Bootstrap samples, and assessing model performance using the Integrated Brier Score (IBS) [70] across different Bootstrap iterations. Integrated Brier scores (IBS) are used to record and store each simulation iteration results, to allow com-

prehensive comparisons of model performance under varying values of prior weights and different sample sizes of the training set. To statistically validate the observed differences in IBS, we applied the Wilcoxon signed-rank test across the simulations of all four cases. In addition, the computation time for each simulation was measured to evaluate computational efficiency.

II.5 Preliminary results

The average values and nonparametric confidence intervals of the IBS result obtained on the test set are shown in Figure 3.3, Figure 3.4, Figure 3.5, and Figure 3.6 for models: Cox, Random Forest, and Proper Bayesian Bootstrap Ensemble Tree (CovBootTree) model. In the experiments, the prior weight is expressed as $w = \frac{k}{k+n}$, where n is the sample size (that is, N = 50, N = 100, N = 300, and N = 500). The values w = 0, 0.25, and 0.5 were tested, corresponding to increasing levels of prior influence relative to the observed data. In particular, the proposed CovBootTree model is slightly affected when the value of w increases, which shows that if the weight given to the prior distribution is high, the bootstrap resamples include a higher number of new observations generated from the prior F_0 results in enriching the original training set. As a consequence, the trees constructed for each bootstrap resample are more independent from each other and the variance of the global ensemble model decreases. To formally quantify this sensitivity, Wilcoxon signed-rank tests were conducted comparing IBS values across weight settings w=0, 0.25, and 0.5 at each sample size. At small sample sizes (N=50 and N=100), no significant differences were found between any weight settings (all $p > 0.05$), indicating that the model is robust to the choice of k when data is limited. At larger samples (N=300, 500), incorporating at least some prior influence ($w > 0$) becomes significantly beneficial, while the difference between w=0.25 and w=0.5 remains non-significant across all settings. Notably, w=0.25 and w=0.5 were never significantly different across any sample size, confirming that performance stabilizes at moderate prior weight. Based on these findings, w=0.25 is recommended as a practical default. The results of all four settings with gradual increase in sample sizes are discussed in the following subsections.

- Case 1; N=50: The results in Figure 3.3 showed lower mean IBS values for the proposed CovBootTree model compared to the classic survival models (Survival Random Forest and Cox).

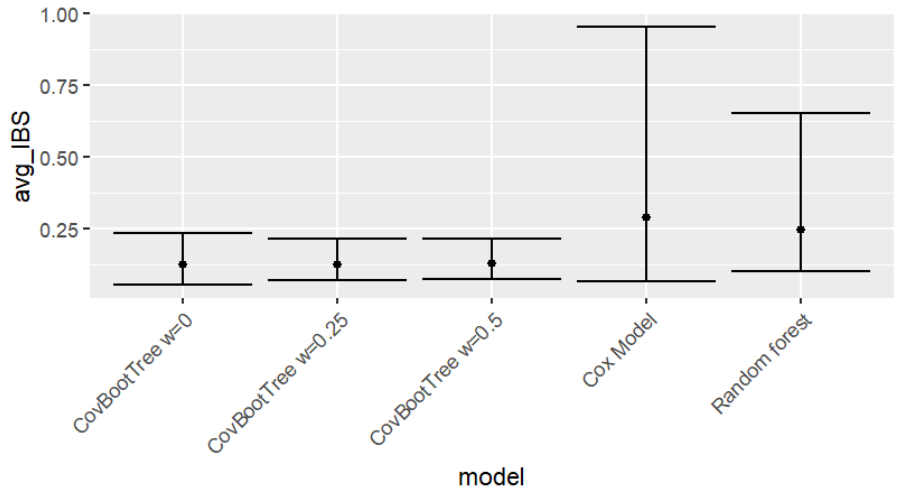

Figure 3.3: Comparison of mean and nonparametric confidence intervals for IBS obtained in cross-validation for the 50 simulations of highly variable covariates of sample size N=50. The proposed model is compared with Random Survival Forest and Cox Model

In particular, CovBootTree achieved lower IBS values in 98–100% of simulations compared to Random Survival Forest and in approximately 84–86% of simulations compared to the Cox model, highlighting the strong and consistent results in its predictive performance in small sample settings.

- Case 2; N=100: The results in Figure 3.4 showed that the CovBootTree model significantly outperforms the Random Survival Forest in all configurations ($p < 0.001$), achieving lower IBS values in 100% of simulations. When compared to the Cox model, the improvement remains statistically significant ($p < 0.05$), although the magnitude of the advantage is reduced. In particular, CovBootTree achieved lower IBS values in approximately 56–64% of simulations, suggesting a moderate but consistent performance gain.
  Compared to the smaller sample size setting (N = 50), the performance gap between CovBootTree and the Cox model decreases, indicating that the relative advantage of the proposed method diminishes as the sample size increases.

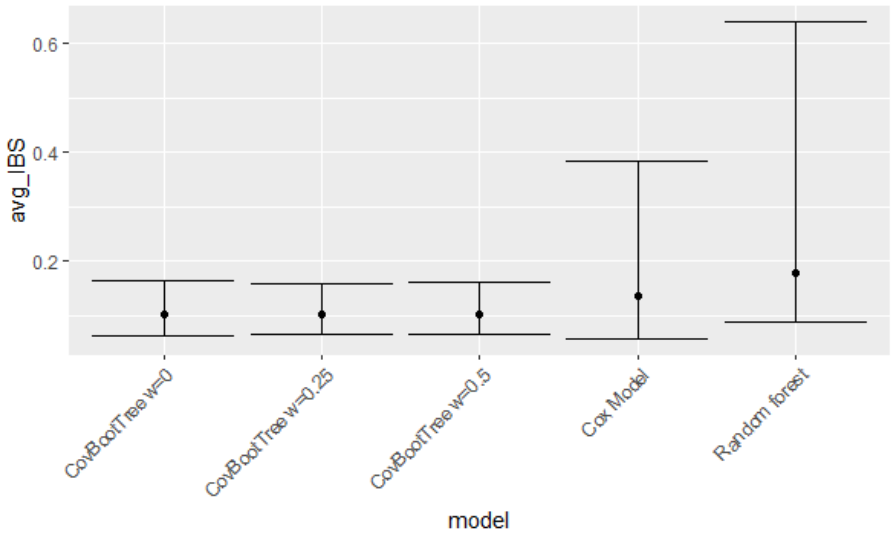

Figure 3.4: Comparison of mean and nonparametric confidence intervals for IBS obtained in cross-validation for the 50 simulations of highly variable covariates of sample size N=100. The proposed model is compared with Random Survival Forest and Cox Model

- Case 3; N=300: As the sample size is increased more than half, the predictive ability of the Cox model slightly outperformed our proposed CovBootTree model. The statistical results confirm that the CovBootTree model significantly outperforms the Random Survival Forest ($p < 0.001$), achieving lower IBS values in nearly all simulations (98–100%). However, when compared to the Cox model, the results show a reversal in performance. The Cox model achieves lower IBS values in the majority of simulations (approximately 74–84%), and the differences are statistically significant ($p < 0.001$), indicating that Cox outperforms CovBootTree in larger sample settings.

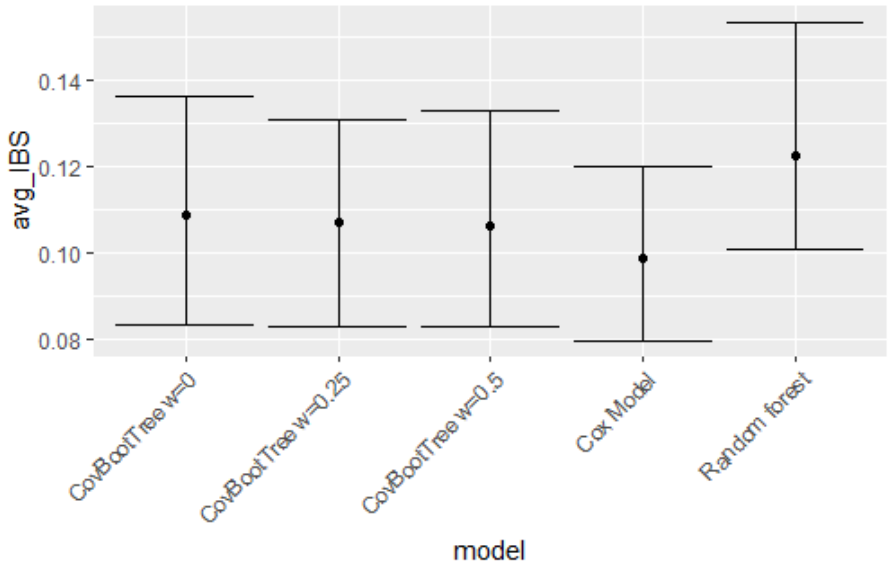

Figure 3.5: Comparison of mean and nonparametric confidence intervals for IBS obtained in cross-validation for the 50 simulations of highly variable covariates of sample size N=300. The proposed model is compared with the Random Survival Forest and Cox Model

- Case 4; N=500: The average IBS values in the fourth case where the sample size is increased to N = 500 show consistent results with the second and third cases, both in terms of the average error and the stability of the results. The results show that the CovBootTree model continues to significantly outperform the Random Survival Forest ($p < 0.001$), achieving lower IBS values in approximately 90–98% of simulations.

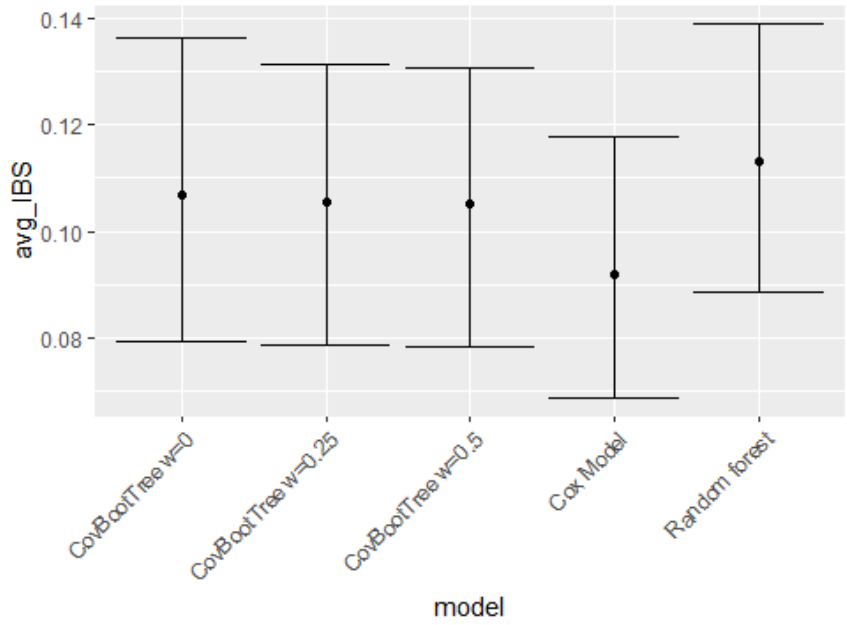

Figure 3.6: Comparison of mean and nonparametric confidence intervals for IBS obtained in cross-validation for the 50 simulations of highly variable covariates of sample size N=500. The proposed model is compared with the Random Survival Forest and Cox Model

However, the Cox model clearly outperforms CovBootTree in this setting, achieving lower IBS values in approximately 88–90% of simulations. These differences are highly statistically significant ($p < 0.001$), indicating that the Cox model provides superior predictive performance in large sample size scenarios.

The four case studies show how the differences between the proposed model and the literature models can be accessed in terms of stability without losing the model performance specially when the sample size is moderate/small. Overall, the results across all simulation settings highlight a clear pattern in model performance. The proposed CovBootTree method demonstrates strong and statistically significant improvements over both benchmark models in small sample scenarios. However, as the sample size increases, the relative advantage over the Cox model diminishes and eventually reverses, with Cox achieving superior performance in larger datasets. These findings suggest that the proposed approach is particularly well-suited for settings characterized by limited sample sizes and complex covariance structures.

II.6 Conclusion

The paper applied the proper Bayesian bootstrap for ensemble survival tree modeling approach to take into account the covariance structure of the time-to-event survival data by improvising the methodology of bootstrapping using the Cholesky decomposition to generate correlated samples that preserve the covariance structure.
During the data simulation process for bootstrap sampling, we incorporate the correlation structure between covariates into the bootstrap resampling technique by using the multivariate normal distribution with Cholesky Decomposition method to handle the multi-dimensional distribution of highly correlated covariates. The use of synthetic data, which replicates the structure and statistical characteristics of the original dataset, helps contribute to the ensemble trees capturing a wider range of survival patterns, potentially reducing the overfitting, and evaluate the proposed model performance on unseen data.
The obtained results suggest that the proposed CovBootTree model with different choices of prior weights along with verying sample sizes outperforms the classical Survival Random Forest model. However, if the sample size increased a certain limit, the Cox model obtained slightly better results.
Further research is expected to evaluate the model's sensitivity toward varying degrees of censored data, categorical variables, and to test the proposed model stability on real dataset. The current simulations establish the effectiveness of CovBootTree under Gaussian covariate structures and sample sizes up to N=500, and extending the framework to non-Gaussian settings and larger sample sizes represents a natural next step toward broader applicability in real-world clinical research.

II.7 References

References for this paper are listed at the end of the thesis.

4 Advanced Graph-Based Clustering: Integrating Clique Detection with Density Peak Assignment

Chapter Contents

4.1 Introduction

Density Peak Clustering (DPC) and traditional clustering algorithms, despite their innovative approaches, face significant limitations when confronted with the complexity and heterogeneity of real-world datasets [93], [94], [95].

Path-based clustering approaches, notably the work by Pizzagalli et al. [96], introduced a global optimization framework that replaces local decisions with shortest-path computations, demonstrating improved performance on elongated and intertwined clusters

through the evaluation of path properties such as minimax gaps. Concurrently, the field of graph theory has provided valuable insights into the structure and connectivity patterns within complex datasets [97], [98], with novel community-detection methods based on cliques overcoming the deficiencies of previous similar community-detection methods by considering the mathematical properties of cliques.

4.1.1 Research Challenges

The limitation off nested cluster structures become particularly pronounced in domains such as bioinformatics, image segmentation, and social network analysis, where data naturally forms complex topological structures that resist conventional clustering approaches.

The local assignment strategy inherent in traditional DPC can lead to misclassifications of boundary points and fails to capture the global connectivity patterns that are essential for accurate clustering in complex scenarios. The density peaks clustering (DPC) algorithm requires manual selection of cluster centers, a single way of density calculation, and cannot effectively handle low-density points [99], prompting researchers to explore more sophisticated assignment mechanisms.

Despite the individual successes of path-based optimization and clique detection methods, the integration of these complementary approaches remains underexplored in the clustering literature [100], [96]. Path-based methods excel at capturing global connectivity patterns and optimizing cluster assignments across the entire dataset, while clique detection provides precise identification of local connectivity structures that may be overlooked by point-to-point similarity measures [98], [101]. The challenge lies in effectively combining these techniques to create synergistic improvements rather than competing methodologies. Another significant challenge involves maintaining computational efficiency while incorporating sophisticated graph-theoretic operations [97]. The combination of clique detection algorithms with shortest path computations presents potential scalability issues that must be addressed to ensure practical applicability to large-scale datasets.

4.1.2 Methodological Innovation

The research presented in this chapter addresses the identified challenges by proposing Cli-DSP (Clique-based Density Shortest Path), a novel clustering algorithm that integrates min-max clique detection with path-based density peak assignment. This approach represents a significant advancement in graph-based clustering by bridging local structural analysis with global path optimization, offering enhanced robustness

and accuracy for complex data clustering tasks.
The proposed methodology employs adaptive distance thresholds and a greedy bottom-up approach to identify maximal cliques that capture dense connectivity patterns in different local regions. The mathematical foundation rests on the identification of complete subgraphs where every pair of vertices is connected by an edge, ensuring that points within a clique exhibit maximal local connectivity, making cliques natural candidates for representing cohesive subgroups within larger datasets. Subsequently, the framework utilizes a combination of DPC and path-based clustering strategies to assign cliques to density peaks through collective decision-making rather than individual point assignments. This integration overcomes the limitations of local assignment rules by incorporating global path optimization, ensuring that clustering decisions reflect the overall connectivity structure of the dataset.
The majority voting mechanism for clique assignment provides increased robustness against individual point misclassifications, leading to more stable and reliable clustering outcomes. This collective decision-making approach represents a fundamental shift from point-wise assignments to group-based clustering decisions. The framework also incorporates sophisticated noise handling mechanisms that use path backtracking to assign isolated points based on their connectivity to established cluster structures, ensuring comprehensive coverage of all data points while maintaining clustering integrity.

4.1.3 Theoretical Framework

The theoretical framework establishes convergence properties for the clique detection process and optimality guarantees for the path-based assignment. The majority voting mechanism ensures that cluster assignments converge to stable configurations that reflect the underlying connectivity structure of the data. The computational analysis reveals that the combined approach maintains reasonable complexity characteristics while delivering superior accuracy. The overall complexity is $O(n^2)$, dominated by pairwise distance computations shared across all pipeline stages. The adaptive threshold targets 15% graph density ensuring sparsity ($E \approx O(n)$), which reduces clique detection to $O(n)$ and Dijkstra's pass to $O(n \log n)$.

4.1.4 Research Questions and Objectives

The primary research question guiding this investigation is: How can local connectivity patterns captured through clique detection be effectively integrated with global path-based optimization to improve clustering accuracy and robustness for complex, irregular datasets? This central question encompasses several secondary inquiries including

the optimal strategy for combining min-max clique detection with shortest path analysis, the impact of adaptive distance thresholds and clique size constraints on algorithm performance, the maintenance of computational efficiency while achieving improved accuracy, and the generalizability of the approach across diverse application domains.
The primary objectives of this research include developing Cli-DSP as a novel clustering framework that seamlessly integrates clique detection with path-based density peak assignment, demonstrating significant improvements in clustering performance metrics compared to traditional methods, and addressing the persistent problem of boundary point misclassifications through sophisticated assignment mechanisms. Secondary objectives encompass establishing rigorous mathematical foundations for the integration approach, developing robust noise handling mechanisms, validating real-world applicability through biomedical applications, conducting computational complexity analysis for scalability assurance, and performing comprehensive comparative evaluations against state-of-the-art methods.

Paper III: A Graph-based Clustering Algorithm for Bridge Problem Solution using Density Peak Assignment through Optimized Shortest Path [Manuscript in preparation.]

A Graph-based Clustering Algorithm for Bridge Problem Solution using Density Peak Assignment through Optimized Shortest Path

F. Naz[a, c], S. Figini[a, b], and D.U Pizzagalli[c]

[a]Department of Mathematics, University of Pavia, Italy,
farah.naz01@universitadipavia.it
[b]Department of Political and Social Sciences, University of Pavia, Italy
[c]Università della Svizzera italiana, Switzerland

Abstract

Clustering algorithms are essential tools for analyzing complex datasets, with significant applications in many scientific fields including data analysis for biomedical research. Traditional clustering methods rely on parameter tuning, point-to-point similarity metrics, and local assignment rules, which can produce suboptimal results when dealing with datasets containing irregular cluster shapes, overlapping regions, or varying density distributions. Based on the impressive performance of path-based clustering algorithms, we propose a clustering approach named Cli-DSP that leverages local connectivity patterns through min-max clique detection combined with global path-based optimization for cluster assignment. Our method first identifies densely connected subgroups within the data and then collectively assigns these subgroups to density peaks using shortest path analysis. Cli-DSP is designed to address the limitations of conventional methods by incorporating local structural information into global clustering decisions, while also handling noise from slightly coupled clusters. We demonstrate the effectiveness of our algorithm on challenging synthetic datasets and real-world datasets, showing consistent improvements over density-based and path-based methods in both accuracy and robustness to noise. Our contribution represents a significant advancement in density-based clustering by bridging local structural analysis with global path optimization, offering a more robust and accurate approach for complex data clustering tasks in biomedical and other domains.

Keywords: density peak clustering, shortest path optimization, clique detection, graph-based clustering, biomedical data analysis

III.1 Introduction

Clustering is a process for identifying groups of similar objects likely to be functionally related as in supervised learning and enabling the discovery of latent structures in unlabeled data through unsupervised learning. Existing clustering algorithms can be categorized into five main approaches: partitioning [102], [103], hierarchical [104], [105], model-based [106], [107], density-based [100], [108], and grid-based methods.
Partitioning algorithms optimize objective functions to divide data into distinct clusters, with k-means being the most prominent example. Hierarchical methods construct tree-like structures that reveal clustering relationships at multiple granularities through agglomerative or divisive strategies. Model-based approaches employ probabilistic frameworks to fit mathematical models to data distributions, typically using expectation-maximization techniques. Density-based algorithms identify regions of high data density that are treated as potential cluster centers, while grid-based methods divide the feature space into cells and perform clustering on the resulting grid structure. These latter two categories share similarities in their spatial analysis approach but differ in their granularity and computational strategies.
These methods have been studied extensively in statistics and machine learning, and their applications in healthcare span diverse fields, from bioinformatics to image segmentation, where the goal is to group similar data points while distinguishing between distinct populations [93]. Despite widespread use, clustering remains a challenging task when datasets exhibit heterogeneous geometries, such as non-globular or nested shapes. While globular clusters are well-handled by traditional methods like k-means [109], methods such as DBSCAN [95] often fail in such scenarios due to their reliance on local metrics or rigid density thresholds. Boundary and bridging problems further complicate clustering, as ambiguous regions between clusters or uneven densities can mislead rigid algorithms like DBSCAN or hierarchical clustering. Recent advances in deep clustering and graph-based techniques aim to address these limitations, though computational trade-offs remain.
Density Peak Clustering (DPC) [99] marked a significant advancement in density-peak-based methods, which excel with non-globular shapes by leveraging global connectivity and adaptive density thresholds through identifying cluster centers as high-density points far from other peaks. While DPC rely on simple local assignment rules that can produce suboptimal results in datasets with irregular geometries or varying densities, there is a need for methods that can better capture local connectivity patterns before making global clustering decisions. To address this limitation, Pizzagalli et al. [96] introduced a global optimization framework, replacing local decisions with shortest-path

computations from a dummy node to all points. This method evaluates path properties (e.g., minimax gaps) to improve robustness, particularly for elongated or intertwined clusters. Recent work by Du et al. [100] demonstrated the utility of graph distances in DPC, but graph-based subgroups or cliques remain underexplored in density-peak-based clustering.

We propose a novel clustering algorithm that integrates graph-based clique detection with path-based cluster assignment. The method employs adaptive distance thresholds and a greedy bottom-up approach to identify min-max cliques that represent dense connectivity patterns in different local regions. Subsequently, a combination of DPC and path-based clustering strategies is used to assign cliques to density peaks, while noise points are assigned through backtracking mechanisms. This integrated framework effectively combines the global perspective of path-based methods with the local robustness of clique detection, offering improved performance on datasets with overlapping or irregular structures. We validate our method on synthetic benchmarks and real-world biomedical and non-medical datasets, including immune cell segmentation and arrhythmia classification, demonstrating consistent improvements over DPC and path-based methods.

The main contributions of this paper can be summarized as follows:

1. We propose a novel graph-based clustering framework that combines min-max clique detection with path-based density peak assignment to capture local connectivity patterns for improved clustering decisions.

2. We introduce a hierarchical assignment strategy that leverages clique-level majority voting to enhance robustness against local assignment errors inherent in traditional density peak methods.

3. We present a comprehensive noise handling mechanism that uses path backtracking to assign isolated points based on their connectivity to established cluster structures.

4. Extensive experimental evaluation shows improved performance on complex datasets with irregular geometries and varying density distributions compared to state-of-the-art methods.

The rest of the paper is organized as follows: Section III.2 reviews related work. Section III.3 details our proposed algorithm. Section III.4 presents the experimental results. Discussion and suggestions for future work are given in Section III.5.

III.2 Related Work

In this section, we review the foundational clustering algorithms that serve as the basis for our proposed methodology. We focus particularly on Density Peak Clustering (DPC) and path-based clustering approaches, which provide the theoretical and algorithmic basis for our integrated framework. While deep clustering methods such as DEC and SCAN have shown promising results, they require substantial architectural decisions, large training sets, and GPU resources, making direct comparison with our lightweight graph-based approach less meaningful. Our experimental comparison focuses on DPC and path-based methods as they share the same density-based clustering paradigm and provide the most relevant baseline for evaluating the specific contributions of Cli-DSP

III.2.1 DPC Algorithm

The Density Peaks Clustering (DPC) algorithm, introduced by Rodriguez and Laio [99], represents a hybrid approach that merges concepts from both density-based and centroid-based clustering methodologies. The fundamental principle underlying DPC is based on two key observations: cluster centers correspond to regions with high local density that are simultaneously positioned at relatively large distances from other high-density regions. To implement this concept, each data point x_i is characterized using two essential metrics: the local density measure ρ_i and the minimum distance δ_i to any point with higher density. The computation of local density ρ_i for point x_i follows the formula:

$$\rho_i = \sum_{x_j \in X} \chi(d(x_i, x_j) - d_c), \quad \chi(z) = \begin{cases} 1, & z < 0 \\ 0, & z \geq 0 \end{cases} \tag{4.1}$$

Here, $d(x_i, x_j)$ represents the Euclidean distance between data points x_i and x_j, while d_c, a user-specified parameter serves as a "cutoff distance". For point x_i with the highest density, DPC defines $\delta_i = max(d(x_i, x_j))$. For all remaining points, δ_i represents the shortest distance to the nearest point having greater density, expressed mathematically as:

$$\delta_i = \min_{x_j : \rho_j > \rho_i} (d(x_i, x_j)) \tag{4.2}$$

with $x_j \in X$, where X represents the complete dataset. Following DPC's fundamental principle, candidate cluster centers are identified as points exhibiting both high local density ρ and large separation distance δ. During the selection process, ρ, and δ are manually selected as centers by observing through a decision graph (i.e., a two-

dimensional visualization of $\rho - \delta$ plot). Once cluster centers are determined, the remaining data points are assigned to clusters through a propagation mechanism: each non-center point joins the cluster of its nearest neighbor that possesses higher density.

III.2.2 Path-based Algorithm

Pizzagalli [110] introduces an enhanced variant of the DPC clustering framework that allows the selection of desired number of clusters through shortest path analysis from identified density peaks. This trainable algorithm shifts focus from traditional point-wise similarity measures to the examination of path characteristics connecting data points, thereby formulating the clustering task as a global optimization challenge.
The algorithm's primary objective is to compute single-source shortest paths (SSSP) from a source node s to all other vertices in a weighted graph, constructing a minimal spanning tree rooted at s. Consider a shortest path $\Gamma = \{s, p, \dots, x\}$ that connects the initial node s to any arbitrary point x, where this path traverses exactly one density peak p. The connection from s to p incurs a negligible cost $\epsilon \in \mathbb{R}_+$ where $\epsilon \to 0$, ensuring that point x is assigned to the cluster associated with density peak p.
The computation of shortest paths on the graph employs Dijkstra's SSSP algorithm [97], which is applicable when path costs exhibit non-decreasing behavior when extending a path. For a given path $\Gamma = \{s, x_1, x_2, \dots, x_i\}$ spanning from source s to destination x_i, the corresponding path cost is denoted as $\xi(\Gamma) = c$.

$$\xi(\Gamma = \{s, x_i\}) \leq \xi(\Gamma' = \{s, x_i, x_{i+1}\}) \tag{4.3}$$

Throughout the optimization procedure, the algorithm automatically identifies and selects the edges and metrics, specifically choosing the route that minimizes cost among all possible paths connecting the source and destination pairs.

$$\Gamma \text{ “minimax” if } \xi_{\text{inf}}(\Gamma) \leq \xi_{\text{inf}}(\Lambda)\ \forall \Lambda \neq \Gamma \tag{4.4}$$

III.3 Proposed Algorithm

In this section, we provide a detailed description of our proposed clustering algorithm and analyze its computational complexity.

III.3.1 Local Structure Detection

Following standard practices in graph-based clustering methodologies, the initial step involves creating an undirected graph that will be used to form cliques of data points. Let $G = (V, E)$ be an undirected graph where $V = \{v_1, v_2, \dots, v_n\}$ represents the set of data points and $E \subseteq V \times V$ denotes the edge set constructed using adaptive distance thresholds. For any two vertices $v_i, v_j \in V$, an edge $(v_i, v_j) \in E$ exists if and only if the distance between the corresponding data points falls below the adaptive threshold τ_{ij}. The proposed algorithm employs a greedy bottom-up approach to identify maximal cliques within G. A clique $C \subseteq V$ is defined as a subset of vertices such that every pair of distinct vertices in C is adjacent, i.e., for all $v_i, v_j \in C$ where $i \neq j$, we have $(v_i, v_j) \in E$ [98].

A maximal clique is a clique that cannot be extended by adding any other vertex from V while maintaining the clique property [101]. Turán's theorem establishes that any graph with n vertices and more than $(1 - \frac{1}{r})\frac{n^2}{2}$ edges must contain a clique of size $r + 1$ [111]. This theorem provides a lower bound on clique sizes in dense subregion of our constructed graph G, ensuring that meaningful cluster structures can be identified in regions of high data density. Our algorithm systematically constructs cliques $\mathcal{C} = \{C_1, C_2, \dots, C_k\}$ subject to size constraints $\alpha \leq |C_i| \leq \beta$ for all $i \in \{1, 2, \dots, k\}$, where α and β represent the minimum and maximum clique sizes, respectively. The detection process terminates when no additional candidates of maximum size β can be generated. The minimum clique size $\alpha = 3$ and maximum clique size $\beta = 6$ were fixed across all experiments. It is noted that the greedy bottom-up approach provides a computationally efficient approximation rather than an exhaustive enumeration of all maximal cliques. The adjacency graph threshold was automatically determined per dataset using an adaptive function that selects the threshold bringing the graph density closest to a target of 15%, removing the need for manual tuning. The number of density peaks was set equal to the known number of clusters in each dataset.

Unlike traditional non-overlapping clustering methods, our approach explicitly permits vertices to participate in multiple cliques, i.e., $C_i \cap C_j \neq \emptyset$ for distinct cliques $C_i, C_j \in \mathcal{C}$. This design choice captures the inherent overlapping nature of real-world groupings and local structural patterns. Vertices that do not belong to any detected clique, i.e., $v \in V \setminus \bigcup_{i=1}^{k} C_i$, are classified as noise points.

Once the clique structure $\mathcal{C}$ is established, each clique C_i defines a relevant local region projection, Figure 4.1(a). With the appropriate local sub-regions identified, the task is to find clusters these cliques belong to. We sort the cliques in descending order of their coverage, with regions having large coverage assigned first.

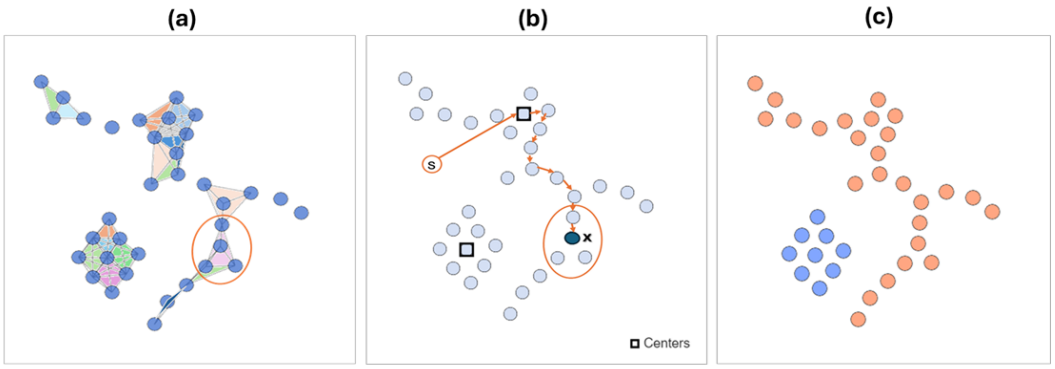

Figure 4.1: An illustration of the clique detection and its cluster assignment. (a) min-max cliques. (b) Shortest path and assignment of node x to density peak along with its clique members. (c) Classification result after using the Clique-to-peak assignment.

III.3.2 Mapping of Cliques-to-Peaks

Inspired by Path-based [110], we devise an assignment step to assign cliques to density peaks based on shortest path analysis. However, unlike the original method that assigns individual points to clusters, our approach operates at the clique level to leverage the collective connectivity patterns captured during local structure detection, Figure 4.1(b). For each detected clique $C_i \in \mathcal{C}$, we first determine the density peak assignment of every constituent node $v_j \in C_i$ using the shortest path methodology. Specifically, each node v_j is assigned to the density peak p_k that lies on its shortest path from the source node, following the path-based framework where the source node is connected to density peaks with an edge of negligible cost.

Given the individual peak assignments within clique C_i, we employ a majority voting scheme to determine the collective assignment of the entire clique. Let $P_i = \{p_{i1}, p_{i2}, \ldots, p_{i|C_i|}\}$ represent the set of density peak assignments for all nodes in clique C_i. The clique is assigned to the density peak that appears most frequently among its constituent nodes:

$$\text{peak}(C_i) = \arg\max_{p \in P_i} |\{v_j \in C_i \,:\, \text{peak}(v_j) = p\}| \tag{4.5}$$

The majority voting strategy preserves local connectivity structures while maintaining consistency with the global density-based clustering framework.

Two variants of this assignment step are evaluated: a fully unsupervised version using the generic minimax path-cost function, and a semi-supervised version (Cli-DSP minimax SVM) that replaces this function with an SVM classifier trained on examples of valid (good) and invalid (bad) path fragments following Pizzagalli et al. [96].

III.3.3 Assignment of Noise Points

For noise points that fall outside any detected clique structure, the method implements a sophisticated backtracking mechanism that traverses each point's density-based path until it reaches the nearest assigned clique member. The noise point is then assigned to the same density peak as this encountered clique member, ensuring that each noise point inherits the cluster assignment of its closest assigned neighbor within its natural connectivity path.

III.3.4 Performance Evaluation

The application of performance metrics in cluster analysis is fundamental for providing objective assessment and validation of algorithmic effectiveness. In the context of density-based clustering methods, evaluation metrics offer critical insights into both the capabilities and constraints of algorithms when identifying complex, non-globular cluster structures with noise. For this study, we evaluate our clustering results using two metrics: the F1 score and the Jaccard index (J). Both evaluation metrics were computed as follows:

$$F_1 = 2 \times \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \tag{4.6}$$

$$J = \frac{\text{TP}}{\text{TP} + \text{FN} + \text{FP}} \tag{4.7}$$

where, TP is the true positive, FN is the false negative, FP is the false positive, and TN is the true negative.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$
$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

III.4 Results

To assess the performance of the proposed algorithm, we conducted extensive numerical experiments using a range of benchmark synthetic and real-world datasets, as detailed in Table 4.1. These datasets pose diverse challenges, such as irregular cluster shapes, adjacency, and nested clusters with varying densities, thereby providing a comprehensive assessment framework for our approach.

Table 4.1: Datasets used in experiments-synthetic and real.

Dataset	Samples	Attributes	Clusters	Domain
Jain	373	2	2	Non-convex clusters
Spiral	312	2	3	Non-convex clusters
Flame	240	2	2	Non-linear separability
Veenman15	600	2	15	Chemical analysis
Zahn's Compound	399	2	6	Varying densities
Aggregation	788	2	7	Geometric diversity
Path-based	300	2	3	Path-shaped clusters
A-sets	3000	2	20	Increasing number of clusters
IRIS	150	4	3	Biological data
Wine	178	13	3	Chemical analysis
Microscopy	8681	2	21	Dendritic cells
MIT-BIH Arrythmia	2331	2	6	Dendritic cells

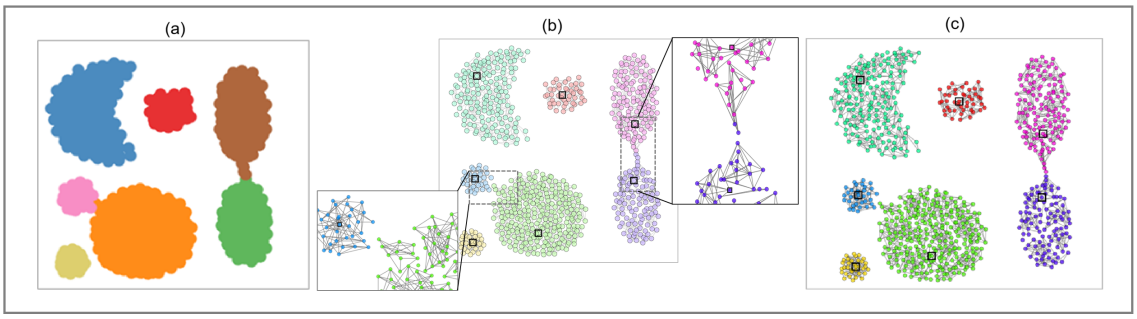

Figure 4.2: Clustering results on Aggregation Dataset. (a) True structure of the Aggregation dataset: It consists of seven clusters with geometric diversity. (b) Adjacent or bridge points with clear clique connection but misclassified in their path to density assignment. (c) Classification result after using the clique-to-peak assignment.

III.4.1 Experiments on synthetic data

The synthetic datasets used to assess the efficacy of our algorithm for clustering, encompass varying scales, noise levels, and structural complexities that are characteristic of real-world applications.

The Aggregation dataset, shown in Figure 4.2, is a classical example featuring adjacent and bridging clusters with nonlinear structures. Note that based on path-based method some noise or bridge points located between the two circles end up in the wrong cluster, as shown in Figure 4.2(b). In contrast, our clique-based assignment approach initially finds the local connection between those bridge points and assigns them to a single cluster which enables a more precise separation of touching clusters and bridges, as shown in Figure 4.2(c).

Zahn's Compound dataset, shown in Figure 4.3, presents three nested and adjacent

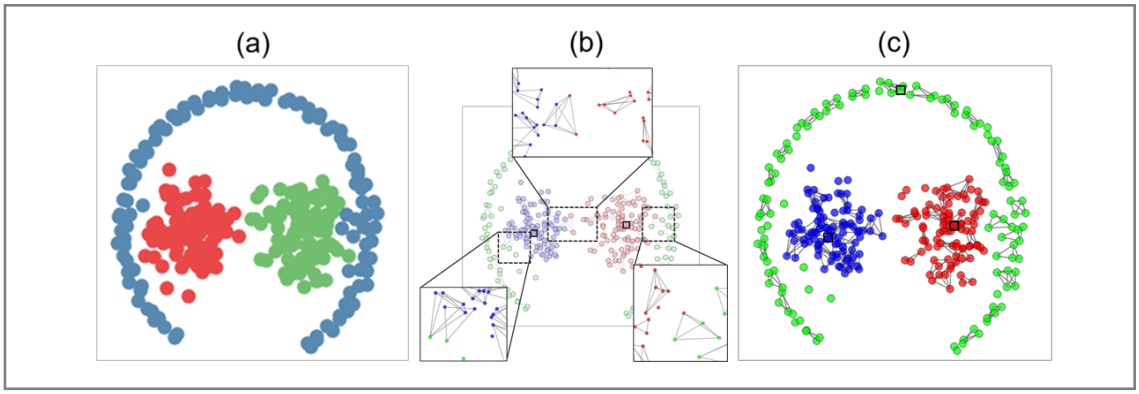

Figure 4.3: Clustering results on Zahn's Compound Dataset. (a) True structure of the Zahn's dataset: It consists of three clusters with varying densities. (b) Adjacent or bridge points with clear clique connection but misclassified in their path to density assignment. (c) Classification result after using the clique-to-peak assignment.

clusters with varied densities, characterized by a circular cluster structure with a gap near the bottom perimeter and two Gaussian-distributed clusters embedded within. Each cluster contains 100 data points. Although the trained path-based algorithm can find the three clusters with almost 98% accuracy, Table 4.2, the Gaussian clusters can be seen as having Gaussian inter-cluster points which are mis-clustered, Figure 4.3(b). By leveraging the min-max clique property, we observed more meaningful proximity relationships and effectively separate closely nested groups. This again shows our method's superior capability in preserving cluster integrity while addressing adjacency issues.

Table 4.2: Performance comparison of clustering methods using F1-score and Jaccard index (J).

Method \Dataset	Jain	Spiral	Flame	Veenman15	Compound	Aggregation	Path-based	A-sets
F1-score performance								
CDP	0.93834	1	0.7875	0.92947	0.67351	0.99746	0.66384	0.98633
DIJ minimax	1	1	1	0.9933	0.84211	0.98477	0.82333	0.97167
Cli-DSP minimax	1	1	1	0.995	0.83991	0.99239	0.82333	0.974
DIJ minimax SVM	1	1	0.99583	0.9933	0.97995	0.99873	0.97	0.96933
Cli-DSP minimax SVM	1	1	1	0.995	0.98496	1	0.98667	0.972
Jaccard index (J) performance								
CDP	0.88384	1	0.64948	0.8682	0.50774	0.99494	0.49683	0.97304
DIJ minimax	1	1	1	0.9867	0.72727	0.97	0.69972	0.94489
Cli-DSP minimax	1	1	1	0.9901	0.72401	0.98489	0.69972	0.94932
DIJ minimax SVM	1	1	0.9917	0.9867	0.96069	0.99747	0.94175	0.94049
Cli-DSP minimax SVM	1	1	1	0.9901	0.97037	1	0.97368	0.94553

1st rank 2nd rank 3rd rank

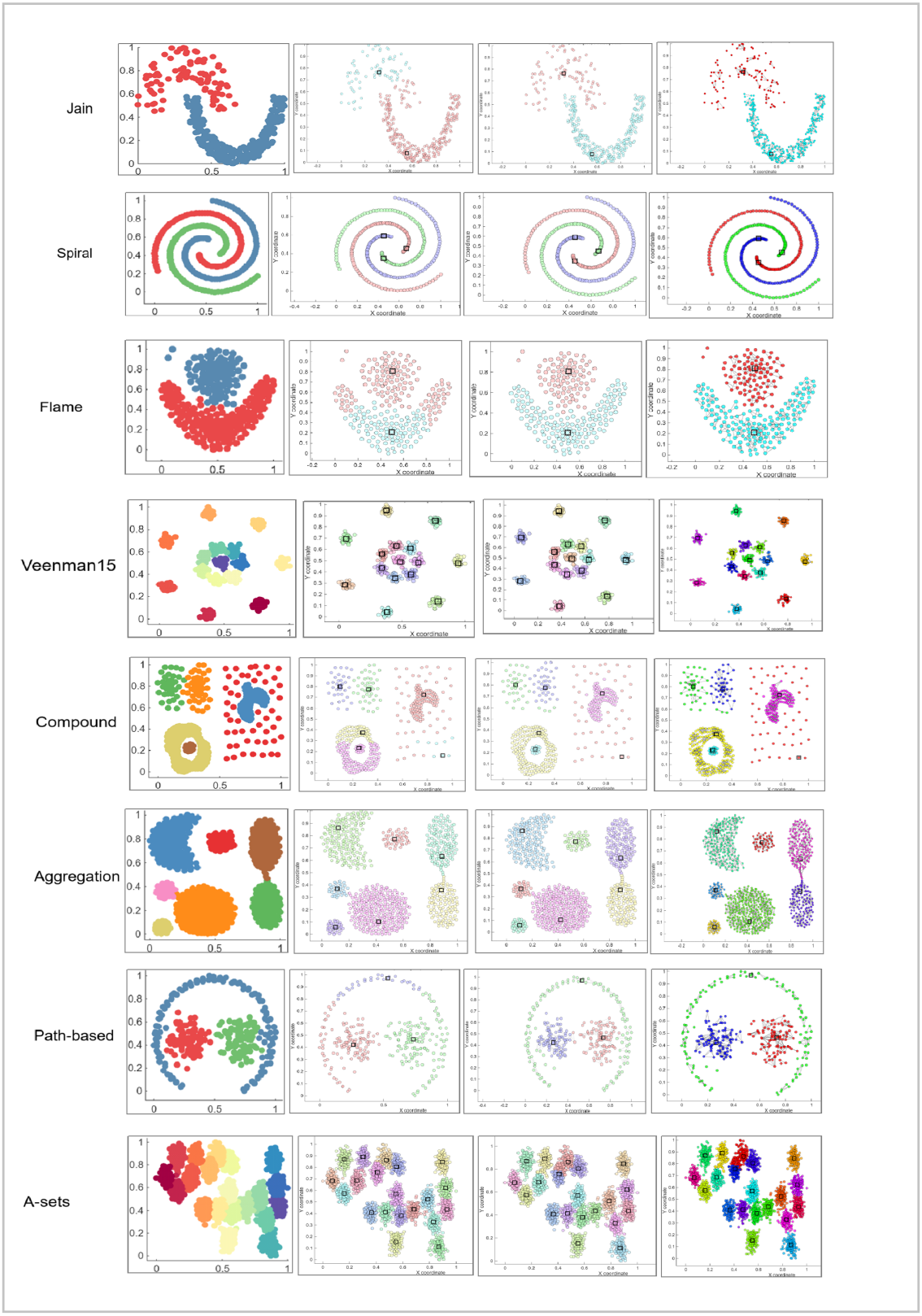

Figure 4.4: Performance comparison on the synthetic datasets.

From the results as shown in Figure 4.4 and Table 4.2, it is observed that the proposed algorithm generated better clusters on synthetic datasets when compared to CDP and the previous generic and trained path-based algorithms. The top three performing methods are highlighted for each data set in the Table 4.2.

III.4.2 Experiments on real-world data

To evaluate the practical applicability and robustness of our proposed methodology, we experimented on real-world datasets sourced from diverse operational environments [112]. The results obtained from these experiments provide empirical evidence of the method's effectiveness and highlight its potential for addressing real-world challenges in the target domain. The top three performing methods are highlighted for each data set in the Table 4.3.

Table 4.3: Performance comparison of clustering methods using F1-score and Jaccard index (J).

Method \Dataset	IRIS	Wine	Microscopy	Arrythmia
F1-score performance				
CDP	0.88667	0.69507	0.64028	0.66693
DIJ minimax	0.96667	0.91011	0.72772	0.66734
Cli-DSP minimax	0.96667	0.91011	0.7272	0.66828
DIJ minimax SVM	0.91333	0.92697	0.70963	0.65968
Cli-DSP minimax SVM	0.92	0.94382	0.70997	0.65786
Jaccard index (J) performance				
CDP	0.79641	0.53265	0.4709	0.5003
DIJ minimax	0.93548	0.83505	0.57198	0.50076
Cli-DSP minimax	0.93548	0.83505	0.57133	0.50181
DIJ minimax SVM	0.84049	0.86387	0.54994	0.49219
Cli-DSP minimax SVM	0.85185	0.89362	0.55036	0.49015

1st rank 2nd rank 3rd rank

We also evaluated our proposed clustering algorithm on three benchmark datasets: Iris, Wine, and Yeast. Performance was assessed using F1 score and Jaccard index metrics. The results, presented in Table 4.3, demonstrate that our method outperforms existing approaches across most datasets, which present significant challenges in delineating cluster boundaries within irregular data domains. Our approach overcomes these limitations by employing clique-based cluster assignments and implement-

ing both global and local effect removal mechanisms, thereby achieving superior accuracy and enhanced interpretability.

III.5 Discussion

This paper presents a novel graph-based clustering algorithm that addresses fundamental limitations of traditional density-based methods by integrating fully connected subset detection (clique identification) with path-based cluster assignment. Our approach demonstrates several distinctive advantages over conventional clustering techniques.
The proposed method inherits the core strengths of density-based clustering while overcoming the constraints imposed by local assignment rules through global path optimization. This strategy aligns with established principles in manifold learning, where global optimization techniques reveal underlying dataset structures that local methods fail to capture. By conceptualizing clustering as a specialized form of structural analysis, our algorithm effectively maps local connectivity patterns to global cluster assignments via shortest path computations.
The clique detection phase enables our method to identify densely connected subgroups that may be overlooked by traditional point-to-point similarity measures. This capability proves particularly advantageous for datasets characterized by irregular cluster boundaries or heterogeneous local densities, scenarios where conventional methods frequently underperform. The subsequent majority voting mechanism ensures that cluster assignments reflect collective connectivity patterns rather than individual point decisions, thereby yielding more robust and stable clustering outcomes.
Regarding computational efficiency, our approach benefits from optimized clique detection algorithms combined with Dijkstra's shortest path computation, which exhibits sub-quadratic complexity for sparse graphs. This computational profile makes the method well-suited for large-scale applications while preserving the accuracy advantages inherent in global assignment strategies over their local counterparts.
While the proposed method demonstrates consistent improvements across most datasets, marginal gains are observed on datasets such as the MIT-BIH Arrhythmia, where clusters naturally overlap in a low-dimensional feature space. In such cases, the limited improvement therefore reflects the intrinsic difficulty of this type of dataset rather than a limitation specific to Cli-DSP, as the data does not exhibit the distinct local connectivity patterns that the method is designed to exploit. This is acknowledged as a limitation and identified as a direction for future work.
Future research directions will focus on extending our framework to handle extremely high-dimensional data through integrated dimensionality reduction techniques and ex-

ploring adaptive grid-based clique detection mechanisms tailored to diverse data characteristics. Additionally, the incorporation of alternative distance metrics, such as Spearman correlation coefficients or specialized graph-based distances, can be readily integrated into our framework. Such enhancements would enable domain-specific optimizations, further expanding the versatility and applicability of our method across specialized domains.

III.6 References

References for this paper are listed at the end of the thesis.

5 Magnetization Transfer Imaging

Chapter Contents

5.1 Introduction

This chapter presents the application of Magnetization Transfer Imaging (MTI) as a quantitative biomarker for early muscle involvement in late-onset Pompe disease (LOPD), as introduced in Chapter 1 (Section 1.3.3). The technical background on MTI and the magnetization transfer ratio (MTR) is provided in the following subsections.

5.1.1 Magnetization Transfer Ratio (MTR)

MTI typically involves acquiring two sets of images: one with an MT saturation pulse (MT-on) and one without (MT-off). The magnetization transfer ratio (MTR) is then calculated as:

$$MTR = \frac{MT_{off} - MT_{on}}{MT_{off}} \times 100\%$$

This ratio reflects the proportion of the MRI signal that originates from protons undergoing magnetization transfer.

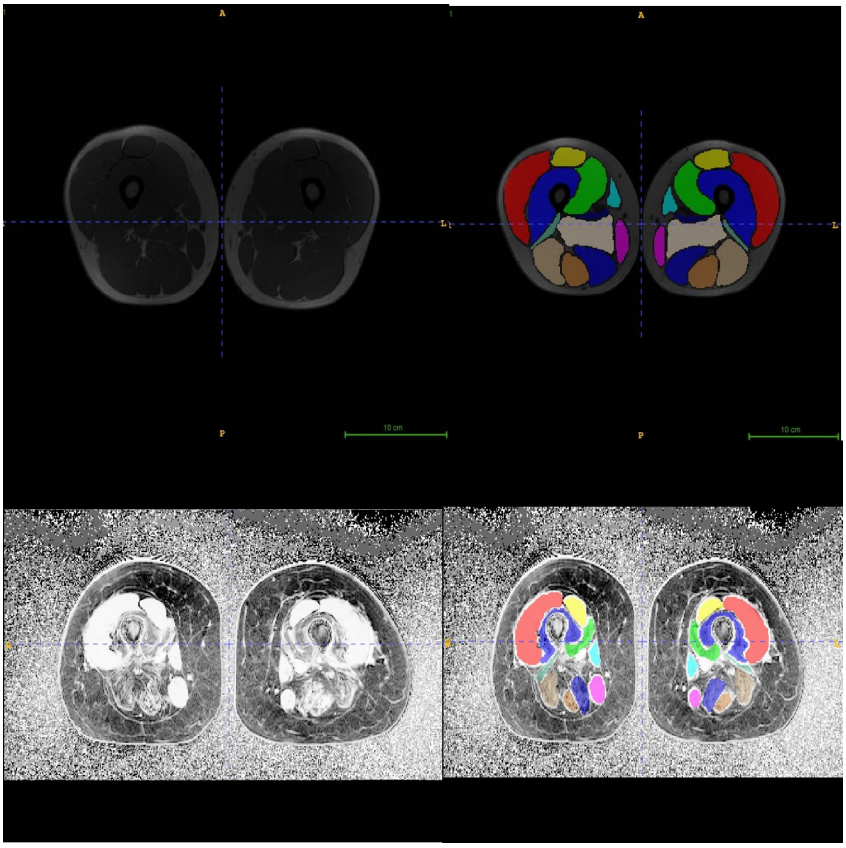

Figure 5.1: MTR in healthy controls and LOPD patients.

5.1.2 Clinical Applications

MTI has proven valuable in neuroimaging, particularly for detecting subtle tissue changes in white matter diseases like multiple sclerosis, where it can reveal tissue damage not visible on conventional MRI. It's also used in musculoskeletal imaging for cartilage assessment, cardiovascular imaging for plaque characterization, and oncology for tumor tissue analysis. The technique provides unique insights into tissue microstructure and composition, making it a powerful tool for both research and clinical diagnosis where conventional contrast mechanisms may be insufficient.

Conference Paper V: Magnetization transfer imaging in late-onset Pompe disease

673P Magnetization transfer imaging in late-onset Pompe disease

M. G. Croce[4], F. Naz[1], L. Barzaghi[1,2,3], M. Paoletti[2], T. Mongini[5], S. Gasperini[6], M. Filosto[7], L. Maggi[8], A. Sechi[9], M. Grandis[10,11], M. Sacchini[12], M. Sciacco[13], L. Vercelli[14], C. Bonizzoni[2], N. Bergsland[15], F. Santini[16,17], X. Deligianni[16,17], C.A.M Gandini Wheeler-Kingshott[4,18,19], S. Ravaglia[4], A. Pichiecchio[2,4]

[1]Department of Mathematics, University of Pavia, Italy, [2]Advanced Imaging and Radiomics Center, Neuroradiology Department, IRCCS Mondino Foundation, Pavia, Italy, [3]INFN, Group of Pavia, Italy, [4]Department of Brain and Behavioral Sciences, University of Pavia, Italy, [5]Neuromuscular Center, AOU Città Della Salute e della Scienza, University of Turin, Italy, [6]Department of Pediatrics, University of MILANO-BICOCCA, Fondazione MBBM, San Gerardo Hospital, Italy, [7]Nemo Brescia Clinical Center for Neuromuscular Diseases, Italy, [8]Neuroimmunology and Neuromuscular Diseases Unit, Fondazione IRCCS Istituto Neurologico 'CARLO BESTA', Milan, Italy, [9]Regional Coordinator Center for Rare Diseases, Academic Hospital of Udine, P.zzale SM della Misericordia 15, Udine, Italy, [10]University of Genova, Italy, [11]IRCCS Ospedale Policlinico San Martino, Genoa, Italy, [12]Unit of Hereditary Metabolic and Muscular Disorders, Meyer Children University Hospital, Firenze, Italy, [13]IRCCS fondazione CA' Granda Ospedale Maggiore Policlinico, Neuromuscular and Rare Disease Unit, Milan, Italy, [14]Department of Neuroscience 'Rita Levi Montalcini', Hospital Città Della Salute e della Scienza, University of Turin, Italy, [15]Buffalo Neuroimaging Analysis Center, Department of Neurology, Jacobs School of Medicine and Biomedical Sciences, University at Buffalo, State University of New York, Buffalo, NY, USA, [16]Department of Radiology, University Hospital Basel, Switzerland, [17]Basel Muscle MRI, Department of Biomedical Engineering, University of Basel, Switzerland, [18]Queen Square MS Centre,Department of Neuroinflammation, UCL Institute of Neurology, Faculty of Brain Sciences, University College London, United Kingdom, [19]Brain MRI 3T Research Centre, IRCCS Mondino Foundation, Pavia, Italy

Abstract

Magnetization transfer imaging (MTI) evaluates the exchange of magnetization between protons in free water molecules and protons bound to macromolecules, including lipids. Widely used in the study of central nervous system diseases, its application in the neuromuscular field has been previously explored only in a Spanish cohort of patients with late-onset Pompe disease (LOPD). To investigate the potential role of MTR as an early biomarker of muscle involvement, we here evaluate magnetization transfer ratio (MTR) and fat fraction (FF) in patients with LOPD in various stages of disease compared to healthy controls (HCs). Quantitative muscle MRI (qMRI) was performed on 31 LOPD patients (21 with mild and 10 with moderate/severe clinical involvement) and 31 matched HCs using 3T MRI. FF and MTR were measured in 11 thigh muscles. Correlations between FF and MTR were assessed. Additionally, FF and MTR were compared between groups of HCs vs. early vs. moderate/severe LOPD. We also explored whether MTR can detect muscle involvement in not yet fat-infiltrated muscles (FF 10%) in early LOPD. MTR of thigh muscles with FF 10% was significantly lower in LOPD compared to HCs. Changes in MTR could be detected even in mildly symptomatic patients, particularly in the medial and posterior compartments (Mann-Whitney U: $p < 0.05$). MTR and FF were inversely correlated in all subjects groups. We found significant differences in MTR and FF changes at the group level between mild and moderate/severe vs HCs. We conclude that MTI has potentially high sensitivity to detect mild muscle fiber damage, even before fat replacement has occurred, making it a useful biomarker to monitor early signs of disease, disease progression, and the efficacy of treatment approaches (Sanofi company provided support for this study, but had no role in study design, data collection and analysis, decision to publish, or preparation of the abstract). (The first and second authors contributed equally to this work)

KEYWORDS: *magnetization transfer imaging, magnetic transfer ratio, intramuscular fat fraction, muscle function tests, late-onset Pompe disease*

V.1 Introduction

Quantitative magnetization transfer imaging (qMTI) is an advanced magnetic resonance imaging (MRI) technique that requires special radio frequency pulse (RF) to induce signal changes, which results into the magnetization transfer between protons in free water compartment and protons bound to macromolecules, present in different environments such as lipids and proteins [113, 114]. In conventional MRI images, the primary contributors are free-water protons which have long and easily detectable T2 signals. However, in MT imaging, magnetization transfer occurs when a low power saturation pulse, also known as the MT pulse, is applied only to protons in present in macro molecules because these protons do not produce a detectable signal on their own due to their short T2 values. This results into the energy transfer from bound to the free pool of protons [115, 116]. During this process, we measure the water signal twice: once with the MT pulse applied (MTon) and once without it (MToff) [117]. By comparing these two measurements, we can quantify how much the macromolecules influenced the water signal. The MT effect can also be quantified obtaining the magnetization transfer ratio (MTR), which provides insight into relaxation and exchange rates between free water and macromolecules [118].
Based on its quantitative nature, qMT has been applied in some clinical studies [119–123], such as a biomarker for brain inflammation [124], as a monitor of demyelination in multiple sclerosis in which the background suppression improves detection of acute lesions [125], and as a measure of response to therapy in glioblastoma [118]. Additionally, it is widely used in MR angiography, enhancing the representation of smaller peripheral branches of the vessels [126].
Previously published studies have demonstrated that MT imaging have profound effect on musculoskeletal tissues because of excess amount of macromolecules, like collagen in tendons and cartilage, or the proteins in muscles [118]. When MT pulse is applied, the MT contrast lead to a noticeable change in the MRI signal from the water, making the MT effect in these tissues very pronounced. The more hydrated the macro molecule is the more MT is obtained. If there is more fat infiltration in the muscle, and since, fat is the hydrophobic tissue, the less MT effect is obtained. Although, the application of MTC in the studies related to central nervous system (CNS) [127–137] are promising but in the neuromuscular field has only been explored only in a Spanish cohort of patients with late-onset Pompe disease (LOPD) [138]. Late-onset Pompe disease (LOPD) is a form of Pompe disease that appears later in life and is characterized by progressive muscle weakness and respiratory issues due to the accumulation of glycogen in muscle fibers which leads to cell death and replacement by fatty tissues. The primarily

symptoms an LOPD patient exhibits are progressive muscle weakness such as difficulty walking, climbing stairs, and other physical activities. Respiratory issues can become severe over time. However, the asymptomatic patients do not always exhibit signs of fat accumulation despite the onset of muscle weakness. This suggests that tracking changes associated with glycogen accumulation may be helpful in tracking the disease's progression, especially in the early stages of Pompe disease. In light of the earlier research [139] outcomes that the progressive accumulation of fat in the musculoskeletal tissues over time that leads to the abnormalities observed in muscle function tests in Pompe patients, we propose that magnetization transfer (MT) could serve as a valuable biomarker for monitoring early signs of Pompe disease, monitoring disease progression, and assessing the effectiveness of treatments. To support our hypothesis and to investigate the potential role of MT as an early biomarker of muscle involvement, we evaluate the magnetization transfer ratio (MTR) and fat fraction (FF) in patients with late-onset Pompe disease (LOPD) at various stages of the disease, comparing these findings with healthy controls (HCs). We also examined the correlation between MTR values, intramuscular fat fraction (FF), and muscle function tests.

V.2 Methods

V.2.1 Study Design and Participants

The experiments were carried out on a 3T MRI scanner, and the sequences obtained included Multi-echo Spin-echo (MESE) for the water T2 computation, Multi-echo Gradient-echo sequences for fat fraction (FF) and a Multi-Parametric Mapping Sequence for the quantification of the Magnetization Transfer Ratio (MTR). The study was approved by institutional ethics committee clinical investigations, and written informed consent was obtained from each participant prior to the study. All study procedures were performed at the Advanced Imaging and Radiomics Center, Neuroradiology Department, IRCCS Mondino Foundation, Pavia, Italy. Imaging and clinical experiments were conducted on thirty one healthy volunteers of the same age and sex as LOPD patients. The inclusion criteria for the thirty one LOPD patients was (1) clinically assessed twenty one asymptomatic and ten symptomatic patients (Walton score 0-1 and 2-8); (2) willingness to finish all muscle function tests, patient-reported outcome assessments, and respiratory assessment; and (3) no MRI contraindications. Based on clinical information and the outcomes of supplementary testing, such as spinal MRI, blood analysis, or EMG when necessary, we ruled out alternative neuromuscular disorders in all research participants. The physiotherapists studied all patients and evaluated muscle function using the fol-

lowing tests: respiratory muscle tests (scale: 0-10): including breathing difficulties and breathing difficulties while lying down; muscle strength tests (scale: 0-10) including: muscle aches, muscle weakness in whole body, upper body, lower body, in arms, and in hand grip; positional tests (yes/no and scale: 0-10) including: walking (without assistance/with/no), difficulty walking, climbing stairs, bending over, rising from a sitting position, squatting, exercise tolerance, and other tests including morning headache (scale: 0-10), fatigue and pain (scale: 0-80), mood (scale: 0-30), depression (scale: 0-10), worry and anxiety (scale: 0-10). The timed tests included the 6-min walking test (6MWT), up-and-go test, time to climb up and down four steps, and the Motor Function Measure 20-item scale (MFM-20). During all timed tests, the patient was instructed not to use any walking aids. Daily life activities were studied using the activity limitations scale for patients with upper and/or lower limb impairments, and quality of life was analyzed using both the Individualized Neuromuscular Quality of Life Questionnaire and the Short Form 36 questionnaire. We obtained forced vital capacity (FVC), both seated and lying down, using the Carefusion Microlab ML 3500 MK8 spirometer (Carefusion, Yorba Linda, CA, USA). These last two tests were added due thembeing commonly used in Pompe patients to measure their clinical status [140]. Table 5.1 provides an overview of the key clinical and demographic characteristics for each participant invilved in the study.

Table 5.1: Demographic and Clinical information of subjects included in the study.

	Controls	Asymptomatic	Symptomatic
Individuals	31	21	10
Gender (M:F)	-	13:4	7:8
Age at MRI	-	23.3 ± 10.3	52.2 ± 16.7
ERT status (yes)	-	9	15
Aids for walking (n)	-	2	2
6MWT (m)	-	616.2 ± 101.9	354.0 ± 175.9
6MWT (%)	-	81.0 ± 15.4	62.0 ± 23.8
Sum MRC (lower limbs)	-	119.2 ± 2.2	75.4 ± 26.8
GSGC	-	4.1 ± 0.3	14.7 ± 7.01
QMFT	-	78.3 ± 2.3	52.1 ± 18.8
FVC (%)	-	97.2 ± 9.6	77.4 ± 25.3
R-PACT	-	33.9 ± 0.3	20.0 ± 8.9

V.2.2 MRI protocol

Two axial three-dimensional gradient echo sequences, with an off-resonance saturation pulse, were performed at the exact same slice position and with the following param-

eters: TR/TE1/TE2 = 49.0/2.46/4.92 ms, flip angle = 7, acquisition matrix = 416 x 216 x 32, Field of View = 448 mm, acquired voxel size = 1.1 × 1.1 × 5.0 $mm3$. In addition, the sequence without an off-resonance saturation pulse was obtained with the following parameters: TR/TE1/TE2 = 19.0/2.46/4.92 ms, flip angle = 7, acquisition matrix = 416 x 216 x 32, Field of View = 448 mm, acquired voxel size = 1.1 × 1.1 × 5.0 mm^3.

V.2.3 Image Analysis

The MRI sequences were analyzed in the middle for the left and right thigh of all patients. The sar (S), gracilisv (G), adductor magnus (AM), semimembranosus (SM), semitendinosus (ST), long head of the biceps femoris (BFL) and short head of the biceps femoris (BFS) were drawn manually for analysis, ensuring a reasonable distance from the subcutaneous fat and fascia. Each ROI was visually examined to exclude any potential inaccuracy in ROI positions and placement. In order to prevent artifacts or systematic mistakes resulting from inhomogeneities in the B1 field of the saturation pulse, only 11 central slices within each image slab were analyzed. The fat fraction (FF) was obtained using the Fatty-Riot Algorithm developed for separation of fat and water magnetic resonance images.

V.2.4 Magnetization transfer ratio

The MT ratio images were obtained on a pixel by pixel basis separately for all sixty two subjects including thirty one LOPD patients (including both asymptomatic and symptomatic) and thirty one healthy controls (HC). The following formula can be used to mathematically compute MTR values in percentage:

$$MTR = \frac{MT_{off} - MT_{on}}{MT_{off}} \times 100\% \tag{5.1}$$

where, MT_{on} and MT_{off} refer to images with and without the saturation pulse. Subsequently, the MTR values were extracted from each slice position and averaged over the 10 slice positions for each participant.

V.2.5 Statistical analyses

The nonparametric tests such as the Mann-Whitney U test and the Kruskal-Wallis test were used for the statistical analysis in the study to analyze the differences between LOPD and healthy controls. Pearson's correlation values were calculated for correlation analysis between MTR, FF, and muscle function tests. All statistical tests were

conducted with a significance level set to $p<0.05$. Given the exploratory nature of this study and consistent with previous MTR studies in LOPD [138], no universal correction for multiple comparisons was applied across all muscle ROIs. Each of the 11 thigh muscles was treated as an anatomically and functionally distinct region of interest. R and Spyder were used to perform statistical tests on a Linux system.

V.3 Results

V.3.1 MTR Results

We observed lower mean thigh MTR values in all LOPD patients compared with healthy controls (Mann-Whitney U: $p<0.05$, Figure 5.2A). Upon analyzing the muscle separately between all LOPD and controls, we identified significantly lower mean MTR values in all muscle regions, except RF and S muscles (Figure 5.2B). Comparative analysis between healthy controls and LOPD (asymptomatic and symptomatic), a lower mean MTR value was observed in symptomatic compared to asymptomatic and controls; and no significant differences were seen between asymptomatic and controls (Kruskal-Wallis test: $p<0.05$; Figure 5.2C). When comparing each muscle individually between these three groups, we observed that the mean value of MTR was lower in symptomatic compared to asymptomatic in all muscles except RF, S, and G; between control and symptomatic, the mean value of MTR was lower in symptomatic patients in all muscles except RF and S; and no significant differences were observed between control and asymptomatic patients in all muscles except AM and G muscle regions. (Kruskal-Wallis test U: $p<0.05$; Figure 5.2D).

We found no significant results between mean thigh FF in all muscles of LOPD patients compared to controls (Mann-Whitney U: $p<0.05$; Figure 5.2E). However, upon analyzing and comparing each muscle individually, we identified higher FF values in AM, RF and S muscles in LOPD patients compared to controls (Mann-Whitney U: $p<0.05$; Figure 5.2F). Comparative analysis for the mean thigh FF values of healthy controls and two groups of LOPD; symptomatic and asymptomatic, showed significant results between all three groups (Kruskal-Wallis test U: $p<0.05$; Figure 5.2G). The mean FF of the thigh was significantly higher in symptomatic LOPD patients compared to asymptomatic LOPD patients.

No significant differences were observed in the mean MTR in muscles with FF less than 10% when compared between all LOPD patients and healthy controls (Mann-Whitney U test: $p<0.05$, Figure 5.3A). However, when analyzed between healthy controls and two subgroups of LOPD (symptomatic and asymptomatic), we observed statistically

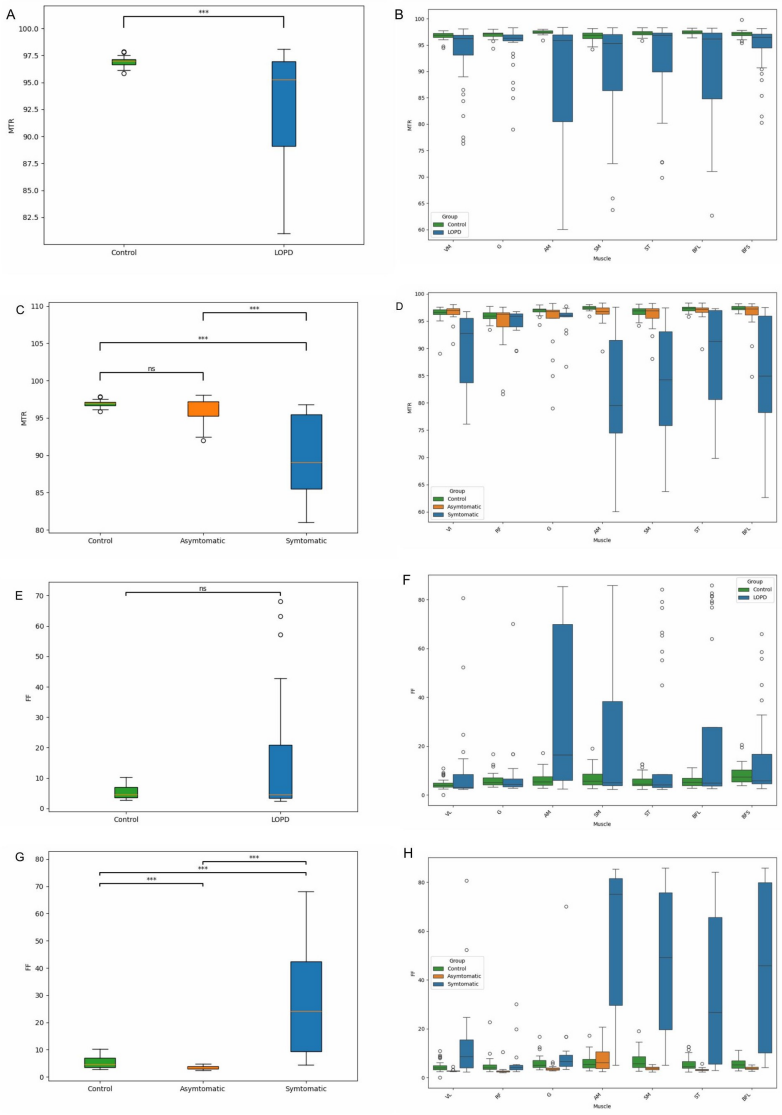

Figure 5.2: Mean thigh MTR and FF in healthy controls and LOPD patients. (A) MTR value in controls (green) and LOPD patients (blue). (B) MTR value of individual muscles in controls (green) and LOPD patients (blue). (C) Mean thigh MTR value in controls (green), LOPD asymptomatic patients (orange) and LOPD symptomatic patients (blue). (D) Mean thigh MTR value of individual muscles in controls (green), LOPD asymptomatic patients (orange) and LOPD symptomatic patients (blue). (E) Mean thigh FF value in controls (green) and LOPD patients (blue). (F) FF value of individual muscles in controls (green) and LOPD patients (blue). (G) Mean thigh FF value in controls (green), LOPD asymptomatic patients (orange) and LOPD symptomatic patients (blue). (H) Mean thigh MTR value of individual muscles in controls (green), LOPD asymptomatic patients (orange) and LOPD symptomatic patients (blue). NOTE: For better visualization we included only seven individual muscles in B, D, F, and H boxplots.

significant differences and lower mean MTR values in muscles of symptomatic patients with FF less than 10% compared to healthy controls. No significant differences were

found between asymptomatic and symptomatic, as well as control and asymptomatic (Kruskal-Wallis test: $p<0.05$; Figure 5.3B). In individual muscle wise comparison between all LOPD patients with FF less than 10% in their muscles and healthy controls, the results show significant differences and lower MTR values in the G, AM and BFS muscles of LOPD subjects (Mann-Whitney U test: $p<0.05$).

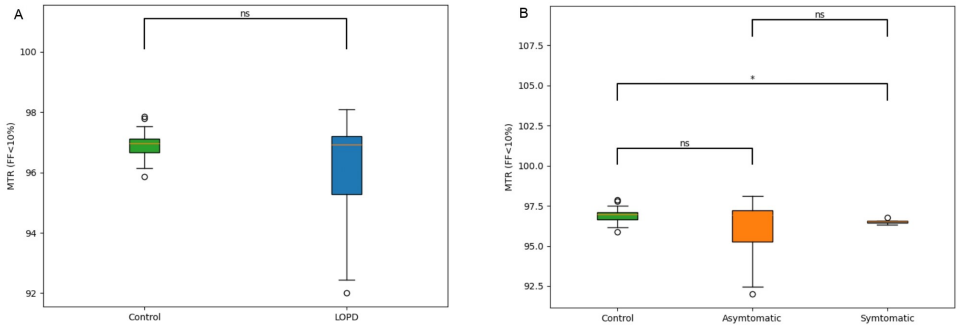

Figure 5.3: Mean thigh MTR value for the muscles with mean thigh FF less than 10 % in LOPD. **(A)** Control vs. LOPD (all). **(B)** Control vs. LOPD (Asymptomatic and Symptomatic).

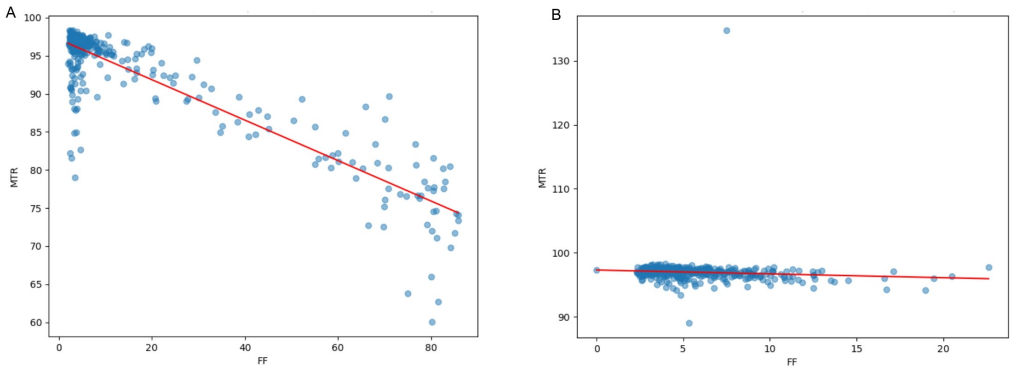

Figure 5.4: Correlation between mean MTR value and FF in healthy controls and LOPD patients. **(A)** MTR and FF in muscles analyzed from patients with LOPD. **(B)** MTR and FF in muscles studied for Healthy Controls.

V.3.2 Correlation Results

In the muscles studied, a negative correlation was obtained between the mean values of MTR and FF in the thigh muscles (Pearson correlation coefficient: -0.90, $p<0.05$; Figure 5.4).

Table 5.2: Correlation between mean MTR values and results of the muscle function tests.

MTR vs. Muscle Function Tests			
Tests	Correlation Coefficient	p-value	Significance
6MWT (m)	0.20	0.2600	
6MWT (%)	0.21	0.2600	
Sum MRC (lower limbs)	0.76	0.0001	***
GSGC	-0.74	0.0001	***
QMFT	0.77	0.0001	***
FVC (%)	0.12	0.4900	
R-PACT	0.65	0.0001	***
MTR for FF<10% vs. Muscle Function Tests			
Tests	Correlation Coefficient	p-value	Significance
6MWT (m)	-0.12	0.5700	
6MWT (%)	-0.37	0.0320	*
Sum MRC (lower limbs)	-0.04	0.8400	
GSGC	0.06	0.7700	
QMFT	-0.08	0.7200	
FVC (%)	0.08	0.7400	
R-PACT	-0.12	0.5800	

For all Pompe patients, we observed that the mean MTR was significantly correlated with most of the results of muscle function tests, particularly muscle strength tests such as the lower extremity MRC score, GSGC, QMFT, FVC (%) and R-PACT. For timed tests, such as 6MWT (%) and 6MWT (meters), we found a significant correlation between the mean thigh MTR for a subset of Pompe patients with FF<10% and timed test 6MWT (%) of muscle function tests (Pearson correlation coefficient: $p<0.05$; Table 5.2).

V.4 Conclusion

The magnetization transfer ratio (MTR) estimation proves to be a sensitive and powerful method for the identification of muscle fiber damage in an early stage that precedes fat accumulation, which is often an indicator of muscle loss. It is important to note that this sensitivity was particularly evident in the medial and posterior compartments, specifically the gracilis (G), adductor magnus (AM), and biceps femoris short head (BFS) muscles, where significant MTR differences were observed even in mildly symptomatic patients with FF<10%. While no significant MTR difference was found globally across all 11 muscles in the asymptomatic group, this region-specific finding is consistent with the known pattern of early glycogen accumulation and fibre damage in LOPD, where these compartments are preferentially affected in early disease stages.

The only prior MTR study in LOPD, conducted by Nuñez-Peralta et al. [26], examined four thigh muscles in a Spanish cohort of 29 LOPD patients using a 1.5T MRI scanner, reporting significantly lower MTR in LOPD compared to healthy controls and a negative correlation between MTR and fat fraction. The present study extends this work in several important ways. First, we analysed a substantially larger set of 11 thigh muscles, providing a more comprehensive mapping of muscle involvement across compartments. Second, our study was conducted at 3T MRI, offering higher signal-to-noise ratio and potentially greater sensitivity to subtle tissue changes. Third, we explicitly stratified patients into asymptomatic and symptomatic subgroups, enabling a more granular assessment of MTR as an early biomarker. While Nuñez-Peralta et al. identified significant MTR differences only in the adductor magnus among muscles with low fat fraction, our findings extend this to the gracilis and biceps femoris short head, suggesting broader early involvement than previously reported. Both studies consistently confirm the negative correlation between MTR and fat fraction and support the potential of MTR as a sensitive biomarker for monitoring disease progression in LOPD. This makes MTR a useful biomarker for monitoring early signs of disease, tracking disease progression, and evaluating the effectiveness of treatments in Pompe disease.
The results of our study show that MTR values correlate with fat fraction and muscle function tests, provided that clinical outcomes align with quantitative measures. Given its ability, MTR has strong potential to help clinicians assess disease progression over time and the effectiveness of treatment approaches in a more detailed and timely manner than traditional measures, such as fat fraction analysis alone. Thus, further exploration of MTR and MT imaging within the context of Pompe disease could help deepen our understanding of muscle degeneration and improve patient care through more precise early stage monitoring.

V.5 Data Availability

Section VI.3 contains a detailed description of the data used in this study. Further inquiries about additional and/or anonymized data sets supporting the findings of this study can be directed to the corresponding author. Due to ethical limitations and concerns, the experimental data set is not available to the general public.

V.6 Funding

The Sanofi company provided support for this study, but had no role in the study design, data collection and analysis, and publication.

V.7 Acknowledgments

The authors thank members of the Department of Neurology and the Clinical Department of the IRCCS Mondino Foundation, Pavia, for their cooperation and helpful discussion in the preparation of this manuscript. The authors extend thanks to all patients and volunteers who participated in the investigation.

V.8 References

References for this paper are listed at the end of the thesis.

6 Mixed-Effects Modeling of Pandemic Impact on University Education

Chapter Contents

6.1 Introduction

The COVID-19 pandemic, declared by the World Health Organization in March 2020, represents one of the most significant disruptions to global education systems in modern history, forcing an unprecedented transition from traditional in-person learning to remote and hybrid learning methods. This crisis revealed critical vulnerabilities in educational systems worldwide and highlighted the urgent need for sophisticated analytical frameworks capable of understanding complex temporal patterns in academic performance while accounting for multi-level data structures inherent in educational settings. The pandemic's impact on university students was particularly profound, as institutions struggled to maintain educational continuity while addressing varying consequences across different demographics and academic disciplines [141].

By combining Linear Mixed Models (LMMs) and Generalized Linear Mixed Models

(GLMMs) with innovative change-point detection techniques, this study provides empirical evidence of COVID-19's long-term effects on university students' academic experiences while contributing to the theoretical understanding of educational system resilience during crisis periods and offering practical insights for institutional planning and student support strategies in future disruption scenarios.

6.1.1 The Challenge

The COVID-19 pandemic revealed critical vulnerabilities in educational systems worldwide, highlighting the need for robust analytical frameworks to understand and predict educational outcomes. Key challenges include:

- Understanding Learning Disruptions: The unprecedented shift to remote learning required sophisticated analytical approaches to assess academic performance changes and identify intervention points.
- Temporal Pattern Recognition in Academic Data: Educational outcomes exhibit complex temporal dependencies that traditional statistical methods often fail to capture adequately.
- Multi-level Data Structures: Student performance is influenced by individual, departmental, and institutional factors, requiring analytical frameworks that can account for hierarchical data structures.

6.1.2 Methodological Innovations and Contributions

We applied sophisticated mixed-effects models and change-point detection methods to analyze the long-term effects of COVID-19 on university students' academic experiences. Using a comprehensive dataset of 231,740 observations from 53,726 unique subjects provided by the University of Pavia, we investigated temporal patterns in academic performance, dropout rates, and departmental variations over a five-year period spanning the years before, during, and after the peak impact of the pandemic.

6.1.3 Theoretical Framework and Practical Implications

The theoretical framework in this chapter builds upon existing literature on educational resilience and adaptive learning systems, extending previous research by [142], on online learning compensation strategies [143] and, on long-term pandemic effects. The framework recognizes that educational systems are complex adaptive systems capable

of responding to external shocks through various mechanisms, including technological adaptation, pedagogical innovation, and institutional support structures.
The practical implications of this research extend beyond the immediate context of pandemic response. Educational institutions worldwide can leverage the methodological framework developed in this study to enhance their analytical capabilities for understanding student performance patterns, predicting dropout risks, and implementing targeted interventions. The findings provide evidence-based insights for policymakers and educators seeking to optimize learning experiences during crises and build more resilient academic environments.

6.1.4 Research Questions and Objectives

The primary objectives of this study are to: (1) provide empirical evidence of COVID-19's impact on university education through comprehensive statistical analysis; (2) validate the effectiveness of mixed-effects modeling approaches for educational data analysis; (3) contribute to the theoretical understanding of educational system resilience during crisis periods; and (4) offer practical insights for institutional planning and student support strategies in future disruption scenarios.

Paper VI: Learning in Lockdowns: A Five-Year Analysis of COVID-19's Influence on University Students' Academic Experiences

Learning in Lockdowns: A Five-Year Analysis of COVID-19's Influence on University Students' Academic Experiences

Farah Naz[1], Simone Gerzeli[2], Elena Ballante[2], and Silvia Figini[2]

[1]Department of Mathematics, University of Pavia
[2]Department of Political and Social Sciences, University of Pavia

Abstract

The declaration of COVID-19 as a pandemic by the World Health Organization (WHO) has marked a significant contemporary threat to humanity. This global health crisis has resulted in the unprecedented shutdown of various sectors, including educational activities. The impact of the pandemic on education has been substantial, necessitating the rapid adoption of alternative learning methods and posing significant challenges to students, educators, and educational institutions. This paper explores the implications and consequences of the global shutdown of educational activities due to the COVID-19 pandemic, highlighting the need for innovative solutions and strategies to ensure continuity in education during such crises. The research employed a mixed-effect model, considering the influence of individual-level and department-level factors on academic outcomes. Additionally, a change point model was incorporated to identify significant shifts or breakpoints in academic performance during the study period. This combined approach facilitated a comprehensive analysis of the short-term and long-term effects of the pandemic on university students' academic experiences. Empirical evidence is described on the basis of a real data set provided by the University of Pavia.

VI.1 Introduction

The emergence of COVID-19, caused by a novel coronavirus, presented a significant global health crisis in late 2019. Huang et al. (2020) [144] reported the initial discovery of this virus in a seafood market in Wuhan, with subsequent clinical analyses confirming person-to-person transmission (Li et al., 2020 [141]; Paules et al., 2020 [145]; Wang, Cheng, et al., 2020 [146]). As the virus rapidly spread and demonstrated severe consequences worldwide, the World Health Organization (WHO) declared COVID-19 a pandemic in March 2020 (WHO, 2020 [147]). In response to this alarming situation, social distancing measures were introduced to mitigate the pandemic's spread.

The outbreak of the COVID-19 pandemic has triggered a momentous global response, necessitating the unprecedented closure of businesses, sports activities, and educational institutions worldwide [148]. The highly contagious nature of the virus has compelled governments and institutions to enforce physical lockdowns, suspending in-person operations and interactions to curb the spread of the disease [149]. As a consequence, all sectors of society have been forced to adapt swiftly to a new reality, with a profound shift towards online platforms becoming the prevailing norm. Among the most affected are university students, who have had to adapt rapidly to the sudden shift from traditional in-person learning to remote and hybrid learning models [150].

This article conducts a comprehensive comparative analysis of a statistical model containing both fixed effects and random effects and a change point model to predict student performance and dropout rates, both before and during the COVID-19 era, utilizing a robust real dataset comprising 231,740 observations provided by the University of Pavia. Furthermore, this study investigates the potential of different types of data to enhance the effectiveness of results, with a particular focus on predicting the probability of dropout challenges and identifying opportunities for academic improvement amidst the ongoing global efforts to eradicate the pandemic. The methodological approach described in the paper is general enough to be delivered to other university institutions around the world. Different metrics are employed to measure the prediction models' performance and to assess the accuracy and validity of the proposed algorithms, including cross-validation exercises to apply to real data.

Furthermore, the findings of this study hold significant implications for institutions, policymakers, and educators seeking to enhance support and optimize learning experiences during crises. By learning from the lessons of the past, we can better equip universities to respond effectively to future disruptions and foster resilient academic environments.

The research paper is structured as follows: Section VI.2 reports a comprehensive review

of the literature concerning student dropouts and academic performance in universities during the pandemic. The dataset and the methodology are presented in Section VI.3. Section VI.4 discusses the results at hand and Section VI.5 proposes future ideas for research based on more extensive databases.

VI.2 Methodological Approach: A Comprehensive Review

As researchers seek to understand the far-reaching consequences of the global pandemic crisis, numerous high-impact papers have emerged, shedding light on various aspects of the pandemic's influence on education. Predicting students' performance is a crucial and intricate concern faced by educational institutions, especially in the context of e-learning environments at the university level. The literature offers numerous methodologies and approaches to address this pertinent matter.

One such seminal work is the study conducted by Li et al. (2020) [141], which examined the immediate effects of COVID-19 on university students' learning experiences. The authors found that the abrupt transition to online learning during the pandemic had varying consequences on students' academic performance, with disparities observed across different demographics and disciplines. Their research highlighted the importance of equitable access to technology and support systems to mitigate the adverse effects of the pandemic on students' learning outcomes. In a longitudinal analysis [143], the authors delved into the extended effects of COVID-19 on university students' academic trajectories. Their research uncovered long-term challenges faced by students, including heightened mental health concerns, increased financial burdens, and evolving perceptions of the value of higher education. The study emphasized the significance of comprehensive support systems to promote student resilience and persistence in the face of adversity.

Building upon the theme of resilience, a groundbreaking paper by Clark, Nong, Zhu, and Zhu [142] aimed to offer a comprehensive understanding of how the shift to online learning influenced student performance and how students navigated the unique circumstances presented by the pandemic to maintain their academic progress. The authors conducted in-depth interviews with students to understand how they adapted to remote learning and navigated the uncertainties brought about by the pandemic. Their findings highlighted the pivotal role of self-regulation, peer support, and faculty engagement in fostering successful transitions to new learning modalities.

Mixed effect models have emerged as a valuable tool for analyzing longitudinal and hierarchical educational data, providing a flexible approach to account for both fixed and random effects [151]. Mixed-effect models have been used in higher education to ex-

plore factors affecting student performance, including student-level attributes, course characteristics, and institutional differences [152]. By incorporating random effects, these models effectively address the clustering of students within institutions and allow for the assessment of individual variation in academic outcomes.

A comprehensive study by Zhu and Liu [153], aimed to comprehensively examine the immediate responses and long-term strategies adopted by educational institutions in response to the pandemic. By employing a mixed-methods framework, the researchers likely sought to provide a holistic understanding of the multifarious challenges, innovations, and transformations that emerged within the education sector as a result of the pandemic. In the paper [154], researchers investigated the interplay between learning strategies, students' digital competencies, and academic performance within the context of the COVID-19 pandemic. The authors employed a mixed-methods approach, combining qualitative and quantitative techniques to collect and analyze data. The study's methodology aimed to provide a comprehensive understanding of how the transition to remote and digital learning during the pandemic influenced students' learning approaches and subsequent academic performance.

While these studies provide valuable insights, there remains a gap in the literature regarding a comprehensive analysis of the long-term effects of the pandemic on university students' academic experiences. This journal paper aims to address this gap by conducting a thorough five-year analysis, encompassing the years before, during, and after the peak impact of the pandemic. By examining trends, challenges, and adaptations, this study builds upon existing research to uncover the nuanced influence of COVID-19 on students' learning journeys. Leveraging the insights from these seminal works, we aim to contribute to the ongoing discourse on educational resilience and inform evidence-based strategies to enhance learning environments amid crises.

VI.3 Empirical Evidence on Real Dataset

VI.3.1 Dataset Description and Transformation

This section presents a comprehensive overview of our dataset, which was developed and provided by the University of Pavia. The dataset encompasses 231,740 observations related to 53,726 unique subjects and comprises 70 variables, encompassing demographic attributes of students, cumulative grade point average, exam frequency, diverse academic disciplines, and average grade, among others.

Given the crucial role of data quantity in the performance of a statistical model, we conducted data cleaning and feature extraction to enhance the model's generalization

ability. The following transformations were applied to optimize the dataset for analysis:

- Academic periods, each spanning two semesters over five years, were converted into ten academic semesters for improved representation.
- Different areas of study were consolidated into five major department categories, namely Engineering, Humanistic, Law-Economic-Politic, Medical, and Scientific, streamlining the analysis (Table 6.1).
- Different Degree programs were consolidated into four major categories, namely Bachelor, Master, Single Cycle 5 years, and Single Cycle 6 Years, streamlining the analysis (Table 6.2).
- The distribution of exams taken by students was represented in a binary format, wherein the value 0 indicated that exams were not taken, and the value 1 denoted that exams were taken across the ten academic semesters (Figure 6.1).
- Binary categories (0 and 1) were assigned to indicate student dropouts during the ten academic semesters (Table 6.3).
- The median number of credits (CFU) across 10 semesters was used to observe trends in central tendency of credit counts during the COVID lockdown (Figure 6.2).
- Students who participated in Erasmus or international mobility programs (n = 2,147; 3.84% of total) were excluded to avoid misattribution of exams and credits across semesters.

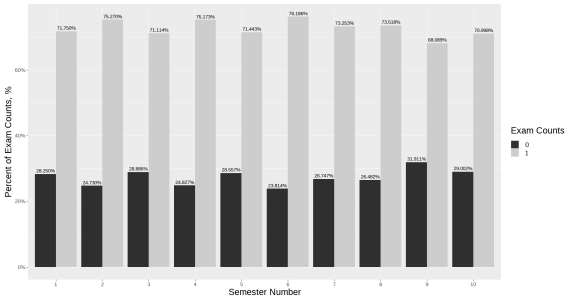

Figure 6.1: Exam Counts

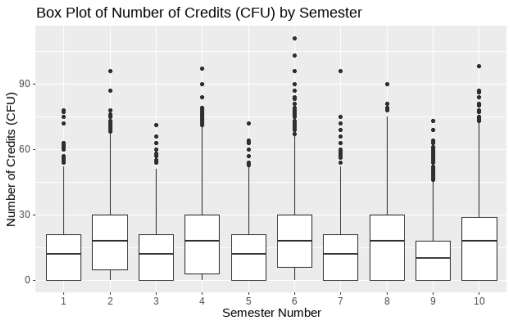

Figure 6.2: Number of Credits (CFU)

These preprocessing steps were undertaken to ensure the dataset's integrity and suitability for subsequent analyses, thereby contributing to the robustness of our statistical model evaluation. The excluded Erasmus students were distributed across all

Table 6.1: Averaged Semestral Frequencies of Department Areas

Department Areas		
Levels	Counts	% of Total
Engineering	876	4.93
Humanistic	2591	14.59
Law-Economic-Politic	6416	36.14
Medical	3496	19.69
Scientific	4376	24.65

Table 6.2: Averaged Semestral Frequencies of Type of Course

Department Areas		
Levels	Counts	% of Total
Bachelor	13776	59.45
Master	4384	18.92
Single Cycle 5 years	3174	13.70
Single Cycle 6 years	1840	7.94

department categories, with the highest concentration in the Law-Economic-Politic area (48.9%), followed by Medical (26.2%) and Humanistic (11.8%). Their mean grade (15.91/30) was slightly lower than the full sample (18.31/30), likely reflecting grade conversion practices for credits earned abroad rather than differences in academic ability.

VI.3.2 Proposed method

In the context of our study, when dealing with data possessing a grouping or hierarchical structure, or when multiple observations are available for individuals, the responses tend to be dependent. To address this dependency, we employed Linear Mixed Models (LMMs) and Generalized Linear Mixed Models (GLMMs). LMMs encompass a Gaussian response, denoted as 'y', alongside a combination of fixed and random components. The model is represented as:

$$y = X\beta + Zu + \epsilon$$

Here, 'X' represents a matrix comprising predictors, typically including the intercept. The parameters $'\beta'$ are known as fixed effects. In the absence of the $'Zu'$ term, this model reduces to a simple linear model. The matrix 'Z' also contains predictors, some of which may overlap with 'X'. The 'u' represents the random effects, accounting for

Table 6.3: Frequencies of Dropout Counts

Dropout Counts		
Semester Number	% of 0's	% of 1's
1	96.867	3.133
2	97.568	2.432
3	97.273	2.727
4	97.596	2.404
5	97.743	2.257
6	97.198	2.802
7	97.221	2.779
8	97.125	2.875
9	96.986	3.014
10	97.036	2.964

correlations within a grouping structure. The model assumes that the random effects 'u' follow a normal distribution with mean zero and covariance matrix G, denoted $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{G})$, and that the residual errors $\boldsymbol{\varepsilon}$ follow a normal distribution with mean zero and variance $\sigma^2\mathbf{I}$, denoted $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$. Furthermore, u and ε are assumed to be independent of each other. In our study, we used the LMMs and GLMMs to investigate the temporal trends of the response variables. We incorporated a random intercept denoted as "(1 — ID)" in all models to account for the potential variability across individuals, as represented by the unique identifier "ID".

This approach allowed us to assess the impact of the specified predictors and the individual ID variations on the response variables, thereby providing valuable insights into the relationship between these variables in the context of our research.

First, we applied the Linear Mixed Models to find the linear and quadratic relationships between response variables (Exam Counts, CFU, Average Grade) and the temporal predictor (Semester Number and Semester Number Squared). Then we follow a similar approach to analyze the relationships between response variables (Exam Counts, CFU, Average Grade) and other relevant predictor variables coupled with the time (Department areas and Degree type).

In the second phase of our study, we specifically applied the GLMM framework to address cases involving binary response data, such as the "Dropout" variable. This extension enabled us to effectively model and interpret the relationships between predictor variables and the binary outcome while accounting for individual-level variability through random effects. As with traditional mixed effects models, the interpretation

of the GLMM results remains conceptually similar, offering valuable insights into the factors influencing binary response variables in the context of our research.
Since the fitting of models with quadratic time is consistently better than the linear ones, we show only results related to this first approach.
We employed the PELT (Pruned Exact Linear Time) algorithm [155] implemented via the changepoint package in R [156], using the cpt.meanvar() function to simultaneously detect changes in both mean and variance, with BIC (Bayesian Information Criterion) as the penalty criterion to control against spurious change points following the application of Linear Mixed Models (LMM), and Generalized Linear Mixed Models (GLMM).

VI.4 Results

The comparative statistics of the LMMs and GLMMs in Tables 6.4 an 6.5 shows the measures of the trade-off between model fit and complexity of the model, comparing the model without additional covariates with the ones that include the areas and type of course. The lower BIC values under models with covariates indicate a better-fit result.

Table 6.4: Linear and Linear Mixed-Effect Models Fitting

Linear Mixed Effect Models					
Model	AIC	BIC	Log-LH	Deviance	Residuals
Exam Count	866608.9	866660.7	-433299.5	866598.9	231735
CFU	1840842.4	1840894.2	-920416.2	1840832.4	231735
Average Grade	799887.3	799937.4	-399938.6	799877.3	167900
Dropout	55711.69	55753.11	-27851.85	55703.69	231736
Linear Mixed Effect Models with Covariates					
Model	AIC	BIC	Log-LH	Deviance	Residuals
Exam Count	863127.8	863252.0	-431551.9	863103.8	231728
CFU	1838362	1838486	-919169	1838338	231728
Average Grade	786914.7	787035.0	-393445.3	786890.7	167893
Dropout	53759.62	53873.51	-26868.81	53737.62	231729

As the linear mixed-effects model investigating academic indicators, including Exam Counts, Average Grade, and CFU, exhibited modest yet statistically significant positive relationships. As shown in Figure 6.3, with the progression of semester numbers beyond the 3rd semester (1st semester of 2019), there emerged an inclination among students

Table 6.5: Fixed and Random Effects Estimates

Fixed Effects Estimates				
Model	Variable	Estimate	Std. Error	t value
Exam Count	(Intercept)	1.3098418	0.0182680	71.701
	Semester Number Square	-0.0045308	0.0004305	-10.525
	Semester Number	0.0692446	0.0049518	13.984
	Humanistic	0.0456855	0.0205674	2.221
	Law-Economic-Politic	0.4065644	0.0168093	24.187
	Medical	0.6615953	0.0189313	34.947
	Scientific	0.3261984	0.0177988	18.327
	Master	0.4332925	0.0133576	32.438
	Single Cycle 5 years	-0.2985150	0.0174985	-17.059
	Single Cycle 6 years	0.0540531	0.0248248	2.177
CFU	(Intercept)	10.862743	0.146999	73.897
	Semester Number Square	-0.033202	0.003546	-9.363
	Semester Number	0.481422	0.040767	11.809
	Humanistic	0.594193	0.162977	3.646
	Law-Economic-Politic	3.176977	0.133152	23.860
	Medical	5.424908	0.150005	36.165
	Scientific	2.043646	0.141079	14.486
	Master	1.111063	0.105878	10.494
	Single Cycle 5 years	-0.377352	0.138662	-2.721
	Single Cycle 6 years	1.674178	0.196262	8.530
Average Grade	(Intercept)	23.1081921	0.0380513	607.291
	Semester Number Square	-0.0036175	0.0008019	-4.511
	Semester Number	0.1023984	0.0092248	11.100
	Humanistic	2.5410639	0.0453183	56.071
	Law-Economic-Politic	1.1050278	0.0365754	30.212
	Medical	1.6946582	0.0400221	42.343
	Scientific	1.3696797	0.0384479	35.624
	Master	2.6881883	0.0271710	98.936
	Single Cycle 5 years	0.2374942	0.0373555	6.358
	Single Cycle 6 years	1.8516517	0.0487617	37.974

Random Effects Estimates			
Model	Variable	Variance	Std.Dev.
Exam Count	ID (Intercept)	0.9952	0.9976
	Residual	1.8710	1.3678
CFU	ID (Intercept)	59.44	7.71
	Residual	128.42	11.33
Average Grade	ID (Intercept)	4.005	2.001
	Residual	4.406	2.099

to undertake more examinations. This trend remained consistent in the context of average exam counts. The observation implies a modest enhancement in students' academic performance, unaffected by potential disruptions attributed to Covid-19 restrictions. An

Table 6.6: Result: GLMM (Dropout Counts)

Fixed and Random Effects Estimates		
Model	Variable	Estimate
Dropout	(Intercept)	-7.31723
	Semester Number Square	0.02346
	Semester Number	-0.14030
	Humanistic	-0.19691
	Law-Economic-Politic	-0.50148
	Medical	-0.95183
	Scientific	0.03740
	Master	-1.41912
	Single Cycle 5 years	-1.52175
	Single Cycle 6 years	-4.96434
Model	Variable	Std.Dev.
Dropout	ID (Intercept)	6.08

Variable	Est	LL	UL
(Intercept)	-7.31723016	-7.4315711	-7.20288920
Semester Number Square	0.02345967	0.0187634	0.02815595
Semester Number	-0.14030011	-0.1915023	-0.08909790
Humanistic	-0.19691311	-0.3826804	-0.01114586
Law-Economic-Politic	-0.50148290	-0.6511963	-0.35176950
Medical	-0.95182729	-1.0672650	-0.83638958
Scientific	0.03740311	-0.1180930	0.19289926
Master	-1.41911813	-1.6043870	-1.23384924
Single Cycle 5 years	-1.52175112	-1.7433764	-1.30012580
Single Cycle 6 years	-4.96433800	-5.1004004	-4.82827560

example of the quadratic fitting of the model without covariate on the target variable "Exam Counts" can be seen in Figure 6.3.

The negative Log Likelihood value in the GLMM results (Table 6.6) showed that the model's predictions are relatively consistent with the observed data. This finding indicates a positive trend in student retention over time. As students progress through their studies, they become less likely to drop out, suggesting a potential increase in their commitment to completing their academic program.

The outcomes derived from the change point analysis across all models consistently indicated the presence of zero change points. This outcome signifies that the patterns

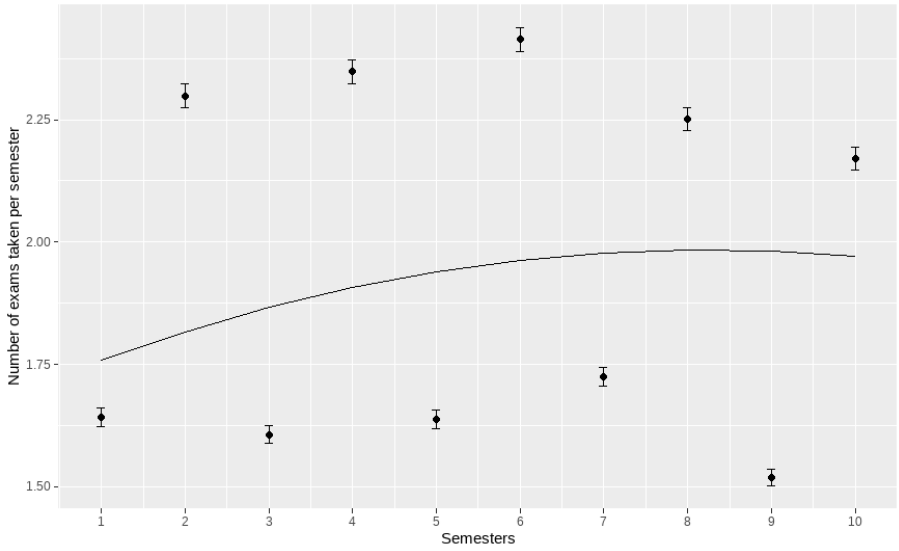

Figure 6.3: Number of exams taken per semester. Mean and error bar are shown. The line corresponds to the fitting of the model.

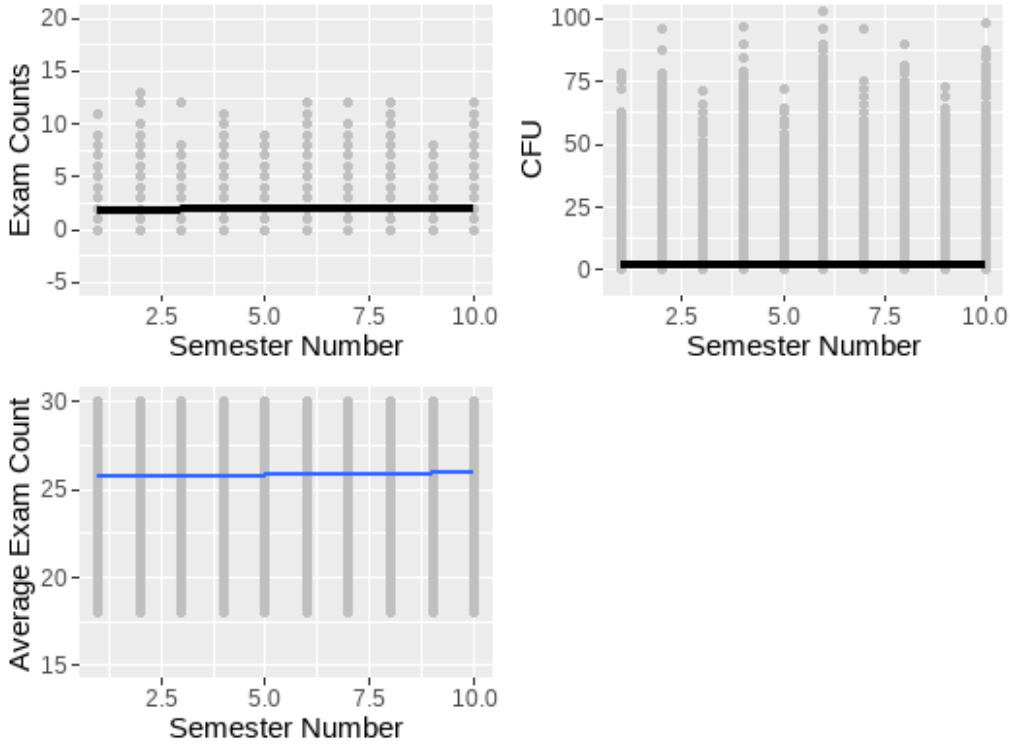

Figure 6.4: Results of change-point analysis applied to Exam Counts, CFU, and Average Exam Count across ten academic semesters.

exhibited by both the response and predictor variables remained statistically insignificant throughout the span of ten semesters. The horizontal line in each panel of Figure

6.4 represents the fitted segment mean. In essence, our analysis suggests that there were no substantial or noteworthy shifts in the trends of the variables under investigation during this academic period, with zero change points detected across all outcomes, indicating no statistically significant structural breaks over the ten semesters studied.
The results from this comprehensive study provide valuable insights into the academic experiences of university students amidst the COVID-19 pandemic and beyond. By examining academic trends, dropouts, and department-specific effects, our research contributes to a deeper understanding of the challenges and opportunities higher education institutions face during crises. Moreover, identifying change points offers a critical temporal context, helping unravel the nuances of students' academic experiences over the study period.

VI.5 Discussion

The outcomes yielded by both the linear mixed model (LMM) and the generalized linear mixed model (GLMM) reflect marginal enhancements in the observed outcomes. Despite the dynamic circumstances introduced by the Covid-19 restrictions, the current dataset did not exhibit any significant shifts or discernible trends. This observation is consistent with the possibility that the strategies and interventions implemented by the University of Pavia effectively mitigated the potential adverse effects of the pandemic on student performance. The PELT algorithm with BIC penalty is inherently conservative, penalising additional change points heavily and therefore favouring a zero change-point solution unless evidence is overwhelming. Furthermore, since data are aggregated at semester level, within-semester disruptions may not be detectable, the Italian national lockdown commenced mid-March 2020, partway through Semester 4, meaning its immediate effect would be split across two semesters and potentially fall below the algorithm's detection threshold. Future analyses using finer-grained temporal data, such as monthly or weekly records, may provide greater sensitivity to detect potential disruptions.
However, a comprehensive investigation demands a nuanced exploration of additional predictive variables that were not encompassed in this study. These variables, although unaccounted for here, may potentially wield substantial influence on the observed outcomes. Such an exploration could illuminate the multifaceted dynamics underlying academic performance and dropout behavior during challenging circumstances.
Furthermore, the exclusion of the psychological impact of the pandemic can play a pivotal role in shaping students' academic responses. Psychological factors have been

demonstrated to be important predictors of academic outcomes during times of crisis[157]. For instance, the study by [158] contributes to comprehending the psychological ramifications of the Covid-19 pandemic through the adept utilization of statistical methodologies. The authors' investigation into anxiety and depression levels in Almadinh, KSA, within the broader context of mental health research during the pandemic, underscores the potential to inform evidence-based strategies for safeguarding the psychological well-being of the local population. Recognizing this research, we acknowledge that a more comprehensive understanding of the intricate interplay between psychological factors and academic performance could enrich the insights garnered from our study.

VI.5.1 Future Research

For future research, a promising avenue would be to extend the analysis to encompass data from multiple universities. By incorporating data from diverse institutions, a more comprehensive and robust understanding of trends, patterns, and potential Covid-19 impacts can be ascertained. Such an approach would facilitate a broader and more insightful assessment of the educational landscape, enhancing our grasp of the pandemic's ramifications on academic outcomes. Furthermore, all pandemic-related references cited in this study are drawn from international literature. Future research should incorporate Italian and European institutional studies to provide a more contextually relevant comparison for the findings. In conclusion, while the current findings highlight the efficacy of university strategies, they underscore the need for more intricate explorations and broader datasets to provide a comprehensive understanding of the evolving educational landscape under the influence of Covid-19.

VI.5.2 Acknowledgements

The paper is written under the supervision of Silvia Figini. The authors would like to thank the statistical office of the University of Pavia for providing the dataset. The authors also thank RIDS (RES Institute for Data Science) for financial support.

VI.5.3 References

References for this paper are listed at the end of the thesis.

Part III

End Section

Discussion and Conclusion

"The significant problems we face cannot be solved at the same level of thinking we were at when we created them."

– Albert Einstein

This thesis presents a comprehensive exploration of computational approaches from statistical to advanced artificial intelligence methods and their applications in healthcare, demonstrating how advanced computational methods can enhance diagnostic accuracy, risk assessment, and patient monitoring across diverse medical and health related domains. Through various distinct yet interconnected projects, we have established novel frameworks that bridge the gap between theoretical aspects of machine learning advances and practical clinical applications.

Starting with the first project which contributes to the statistical foundations of biomedical research by developing a Proper Bayesian Bootstrap approach for survival analysis. This methodology specifically addresses the dual challenges of limited sample sizes and highly correlated covariates that frequently occur in biomedical applications. Our approach showed superior performance compared to standard methods, particularly in small dataset scenarios, providing more reliable predictions for clinical decision-making.

This research presents a novel CovBootTree (Covariance Bootstrap Tree) methodology that extends the Proper Bayesian Bootstrap framework to survival analysis with highly correlated covariates. The main contributions include:

- Developed a Proper Bayesian Bootstrap ensemble tree model specifically designed for survival data analysis that incorporates the correlation structure of covariates.
- Integrated Cholesky decomposition with multivariate normal distribution sampling to preserve the covariance structure during bootstrap resampling.
- Extended traditional bootstrap methods by generating synthetic observations from prior distributions that respect variable interdependencies.

The improved survival analysis methodology has immediate applications in clinical trial design and personalized medicine. By providing more reliable predictions with limited data, this approach could accelerate research in rare diseases and enable more precise prognostic counseling for individual patients.

The Bayesian bootstrap approach for survival analysis, while showing promise, requires further theoretical development and validation across diverse clinical contexts. Future

work should explore extensions to more complex survival models and competing risk scenarios.

The second project advanced the field of graph-based clustering through the development of a novel algorithm that integrates clique detection with path-based cluster assignment for complex datasets. By combining local connectivity pattern analysis, density peak clustering, and shortest path optimization techniques, we successfully addressed fundamental limitations of traditional clustering methods when dealing with datasets containing irregular cluster shapes, overlapping regions, or varying density distributions. This work demonstrates the potential for transforming complex data analysis challenges into robust clustering solutions across diverse scientific domains.
This study introduces a new structured method for identifying densely connected subgroups and assigning them to density peaks through global path optimization, addressing the critical need for robust clustering in datasets with heterogeneous geometries. The methodology leverages graph-based clique detection to capture local structural information while maintaining global optimization for cluster assignments. The main contributions include:

- Developed a comprehensive graph-based clustering framework that combines min-max clique detection with path-based density peak assignment to capture local connectivity patterns for improved clustering decisions.
- Introduced a systematic approach that integrates adaptive distance thresholds with greedy bottom-up clique identification to effectively handle irregular cluster boundaries and varying local densities.
- Implemented comprehensive noise handling mechanisms that use path backtracking to assign isolated points based on their connectivity to established cluster structures.
- Demonstrated methodology effectiveness through extensive experimental evaluation on challenging synthetic datasets and real-world datasets, showing consistent improvements over state-of-the-art methods.
- Showed superior performance in handling datasets with non-globular shapes, nested clusters, and bridging problems while maintaining computational efficiency.
- Established adaptability across diverse domains including biomedical data analysis, chemical analysis, and geometric pattern recognition.
- Established foundation for advanced clustering applications that require both local structure preservation and global optimization.

The proposed methodology demonstrates a meaningful leap forward in density-based clustering, offering researchers and data scientists a robust framework for analyzing complex datasets with irregular geometries and varying density distributions. The approach effectively addresses limitations of conventional methods like DBSCAN and traditional density peak clustering by incorporating local structural information into global clustering decisions. By identifying densely connected subgroups before making cluster assignments, this graph-based approach enables more accurate clustering of challenging datasets with overlapping or nested structures. The validation across synthetic benchmarks and real-world applications demonstrates the generalizability of the approach and its effectiveness in handling diverse data characteristics. The clustering methodology could benefit from extension to handle extremely high-dimensional data through integrated dimensionality reduction techniques and exploration of adaptive grid-based clique detection mechanisms. Future research should explore the incorporation of alternative distance metrics such as Spearman correlation coefficients or specialized graph-based distances to enable domain-specific optimizations and further expand the versatility and applicability across specialized scientific domains.

The MTR project, presented as an abstract at the World Muscle Society Congress 2024 in Prague, explored the application of advanced MRI techniques in neuromuscular disease assessment through magnetization transfer imaging (MTI) in late-onset Pompe disease. This work investigated the potential of magnetization transfer ratio (MTR) as an early biomarker for detecting muscle fiber damage before visible fat infiltration occurs, addressing limitations of conventional imaging approaches in early disease detection. This study developed a quantitative imaging methodology for early detection of muscle involvement in neuromuscular disorders. The main contributions include:

- Demonstrated MTR's high sensitivity for detecting muscle fiber damage in early disease stages, even in muscles without fat replacement (FF $\leq$ 10%).
- Established significant correlations between MTR measurements and clinical muscle function tests, validating the technique's clinical relevance.
- Showed MTR's ability to differentiate between asymptomatic and symptomatic patients, particularly in medial and posterior muscle compartments.
- Validated the inverse relationship between MTR and fat fraction across different disease severities, providing insight into disease progression patterns.

The methodology demonstrates MTR's potential as a sensitive biomarker for monitoring early neuromuscular disease signs, disease progression, and treatment efficacy.

The technique's ability to detect muscle changes before fat accumulation makes it valuable for clinical assessment and therapeutic monitoring. Future applications could extend this approach to other neuromuscular conditions and explore integration with additional quantitative MRI parameters for comprehensive muscle health assessment.

The last chapter of the thesis contributes to the statistical foundations of educational analytics by developing a comprehensive mixed-effects modeling framework for understanding pandemic impacts on university education. This methodology specifically addresses the dual challenges of hierarchical data structures and temporal pattern detection that frequently occur in longitudinal educational research. Our approach demonstrated superior performance in capturing individual-level variability while identifying institutional and departmental effects, providing more reliable insights for educational policy and intervention strategies during crisis periods.
This research presents a novel integration of Linear Mixed Models (LMMs) and Generalized Linear Mixed Models (GLMMs) with change-point detection methodology for analyzing educational outcomes during the COVID-19 pandemic. The main contributions include:

- Developed a comprehensive mixed-effects modeling framework specifically designed for multi-level educational data that accounts for individual, departmental, and temporal factors simultaneously.
- Integrated change-point detection methods with mixed-effects models to identify significant shifts in academic performance patterns throughout the pandemic period.
- Extended traditional educational analytics by incorporating random effects to preserve individual-level variability while capturing institutional clustering effects in a dataset of 231,740 observations.

The improved educational analytics methodology has immediate applications in institutional planning and student support services. By providing more reliable predictions of academic outcomes and dropout risks, this approach could enhance early intervention systems and enable more targeted support strategies for vulnerable student populations during future disruptions. The mixed-effects modeling approach for educational crisis analysis, while showing robust performance, requires further validation across diverse institutional contexts and crisis scenarios. Future work should explore extensions to incorporate psychological and socioeconomic factors and adapt the framework for real-time monitoring systems. The convergence of accessible technology, advanced algorithms, and clinical expertise creates unprecedented opportunities for improving

human health. However, realizing this potential requires continued collaboration between technologists, clinicians, and patients to preserve the integrity of AI development grounded in real-world healthcare needs.

As we look toward the future, the integration of AI into healthcare will undoubtedly accelerate. The frameworks and methodologies developed in this thesis provide a foundation for this integration, emphasizing the importance of interpretability, validation, and clinical relevance. By maintaining focus on these principles, the field of healthcare AI can continue to develop solutions that not only advance scientific knowledge but also improve patient outcomes and enhance the practice of medicine.

The journey from statistical models to advanced computer vision methods, we show how methodological diversity can lead to more robust and clinically relevant solutions and also represents a fundamental shift toward more precise, accessible, and effective healthcare. This thesis contributes to that transformation while acknowledging the roadblocks and advancements shaping the future of AI-driven healthcare innovation.

Bibliography

(1) Raghupathi, W.; Raghupathi, V. *Health information science and systems* **2014**, *2*, 1–10.

(2) Belle, A.; Thiagarajan, R.; Soroushmehr, S. R.; Navidi, F.; Beard, D. A.; Najarian, K. *BioMed research international* **2015**, *2015*, 370194.

(3) Wang, Y.; Kung, L.; Byrd, T. A. *Technological forecasting and social change* **2018**, *126*, 3–13.

(4) Collins, F. S.; Varmus, H. *New England journal of medicine* **2015**, *372*, 793–795.

(5) Kruse, C. S.; Goswamy, R.; Raval, Y. J.; Marawi, S. *JMIR medical informatics* **2016**, *4*, e5359.

(6) Obermeyer, Z.; Emanuel, E. J. *New England Journal of Medicine* **2016**, *375*, 1216–1219.

(7) Dimitrov, D. V. *Healthcare informatics research* **2016**, *22*, 156–163.

(8) Mittelstadt, B. D.; Floridi, L. *The ethics of biomedical big data* **2016**, 445–480.

(9) Mehta, N.; Pandit, A. *International journal of medical informatics* **2018**, *114*, 57–65.

(10) Friedman, J. *(No Title)* **2009**.

(11) Bishop, C. M.; Nasrabadi, N. M., *Pattern recognition and machine learning*; 4; Springer: 2006; Vol. 4.

(12) Mitchell, T. M.; Mitchell, T. M., *Machine learning*; 9; McGraw-hill New York: 1997; Vol. 1.

(13) Russell, S. J.; Norvig, P., *Artificial intelligence: a modern approach*; pearson: 2016.

(14) James, G.; Witten, D.; Hastie, T.; Tibshirani, R., et al., *An introduction to statistical learning*; 1; Springer: 2013; Vol. 112.

(15) Alpaydin, E., *Introduction to machine learning*; MIT press: 2020.

(16) Abu-Mostafa, Y. S.; Magdon-Ismail, M.; Lin, H.-T., *Learning from data*; AMLBook New York: 2012; Vol. 4.

(17) Barlow, H. B. *Neural computation* **1989**, *1*, 295–311.

(18) Jolliffe, I. T.; Cadima, J. *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences* **2016**, *374*, 20150202.

(19) Aggarwal, C. C. et al., *Data mining: the textbook*; 3; Springer: 2015; Vol. 1.

(20) Xu, R.; Wunsch, D. *IEEE Transactions on neural networks* **2005**, *16*, 645–678.

(21) Yosinski, J.; Clune, J.; Bengio, Y.; Lipson, H. *Advances in neural information processing systems* **2014**, *27*.

(22) Bengio, Y. In *Proceedings of ICML workshop on unsupervised and transfer learning*, 2012, pp 17–36.

(23) Torrey, L.; Shavlik, J. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*; IGI global: 2010, pp 242–264.

(24) Dietterich, T. G. In *International workshop on multiple classifier systems*, 2000, pp 1–15.

(25) Freund, Y.; Schapire, R. E. *Journal of computer and system sciences* **1997**, *55*, 119–139.

(26) Breiman, L. *Machine learning* **2001**, *45*, 5–32.

(27) Chen, T.; Guestrin, C. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp 785–794.

(28) Rokach, L. *Artificial intelligence review* **2010**, *33*, 1–39.

(29) LeCun, Y.; Bengio, Y.; Hinton, G. *nature* **2015**, *521*, 436–444.

(30) Goodfellow, I.; Bengio, Y.; Courville, A.; Bengio, Y., *Deep learning*; 2; MIT press Cambridge: 2016; Vol. 1.

(31) Krizhevsky, A.; Sutskever, I.; Hinton, G. E. *Advances in neural information processing systems* **2012**, *25*.

(32) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; Polosukhin, I. *Advances in neural information processing systems* **2017**, *30*.

(33) Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. *Advances in neural information processing systems* **2014**, *27*.

(34) Bengio, Y.; Courville, A.; Vincent, P. *IEEE transactions on pattern analysis and machine intelligence* **2013**, *35*, 1798–1828.

(35) Rajkomar, A.; Dean, J.; Kohane, I. *New England Journal of Medicine* **2019**, *380*, 1347–1358.

(36) Ginsburg, G. S.; Phillips, K. A. *Health affairs* **2018**, *37*, 694–701.

(37) Topol, E. J. *Nature medicine* **2019**, *25*, 44–56.

(38) Beam, A. L.; Kohane, I. S. *Jama* **2018**, *319*, 1317–1318.

(39) Wang, Y.; Wang, L.; Rastegar-Mojarad, M.; Moon, S.; Shen, F.; Afzal, N.; Liu, S.; Zeng, Y.; Mehrabi, S.; Sohn, S., et al. *Journal of biomedical informatics* **2018**, *77*, 34–49.

(40) Shortliffe, E. H.; Sepúlveda, M. J. *Jama* **2018**, *320*, 2199–2200.

(41) Shickel, B.; Tighe, P. J.; Bihorac, A.; Rashidi, P. *IEEE journal of biomedical and health informatics* **2017**, *22*, 1589–1604.

(42) Little, R. J.; Rubin, D. B., *Statistical analysis with missing data*; John Wiley & Sons: 2019.

(43) Hripcsak, G.; Albers, D. J. *Journal of the American Medical Informatics Association* **2013**, *20*, 117–121.

(44) Manolio, T. A.; Fowler, D. M.; Starita, L. M.; Haendel, M. A.; MacArthur, D. G.; Biesecker, L. G.; Worthey, E.; Chisholm, R. L.; Green, E. D.; Jacob, H. J., et al. *Cell* **2017**, *169*, 6–12.

(45) Kaissis, G. A.; Makowski, M. R.; Rückert, D.; Braren, R. F. *Nature Machine Intelligence* **2020**, *2*, 305–311.

(46) Libbrecht, M. W.; Noble, W. S. *Nature Reviews Genetics* **2015**, *16*, 321–332.

(47) Sheikhalishahi, S.; Miotto, R.; Dudley, J. T.; Lavelli, A.; Rinaldi, F.; Osmani, V., et al. *JMIR medical informatics* **2019**, *7*, e12239.

(48) Ritchie, M. D.; Holzinger, E. R.; Li, R.; Pendergrass, S. A.; Kim, D. *Nature Reviews Genetics* **2015**, *16*, 85–97.

(49) Gillies, R. J.; Kinahan, P. E.; Hricak, H. *Radiology* **2016**, *278*, 563–577.

(50) Saeys, Y.; Inza, I.; Larranaga, P. *bioinformatics* **2007**, *23*, 2507–2517.

(51) Kourou, K.; Exarchos, T. P.; Exarchos, K. P.; Karamouzis, M. V.; Fotiadis, D. I. *Computational and structural biotechnology journal* **2015**, *13*, 8–17.

(52) Haixiang, G.; Yijing, L.; Shang, J.; Mingyun, G.; Yuanyue, H.; Bing, G. *Expert systems with applications* **2017**, *73*, 220–239.

(53) Esteva, A.; Kuprel, B.; Novoa, R. A.; Ko, J.; Swetter, S. M.; Blau, H. M.; Thrun, S. *nature* **2017**, *542*, 115–118.

(54) Rajkomar, A.; Oren, E.; Chen, K.; Dai, A. M.; Hajaj, N.; Hardt, M.; Liu, P. J.; Liu, X.; Marcus, J.; Sun, M., et al. *NPJ digital medicine* **2018**, *1*, 18.

(55) Steyerberg, E. W.; Harrell Jr, F. E. *Journal of clinical epidemiology* **2015**, *69*, 245.

(56) Rudin, C. *Nature machine intelligence* **2019**, *1*, 206–215.

(57) Zou, H.; Hastie, T. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **2005**, *67*, 301–320.

(58) Efron, B.; Tibshirani, R. J., *An introduction to the bootstrap*; Chapman and Hall/CRC: 1994.

(59) Muliere, P.; Secchi, P. *Annals of the Institute of Statistical Mathematics* **1996**, *48*, 663–673.

(60) Ferguson, T. S. *The annals of statistics* **1973**, 209–230.

(61) Davison, A. C.; Hinkley, D. V., *Bootstrap methods and their application*; 1; Cambridge university press: 1997.

(62) Hothorn, T.; Lausen, B.; Benner, A.; Radespiel-Tröger, M. *Statistics in medicine* **2004**, *23*, 77–91.

(63) Galvani, M.; Bardelli, C.; Figini, S.; Muliere, P. *Algorithms* **2021**, *14*, 11.

(64) Lo, A. Y. *The Annals of Statistics* **1993**, 100–123.

(65) Welchowski, T.; Zuber, V.; Schmid, M. *Statistics in medicine* **2019**, *38*, 2413–2427.

(66) He, Z.; Yu, W. *Computational biology and chemistry* **2010**, *34*, 215–225.

(67) Yang, Y.; Zou, H. *Statistics and its Interface* **2013**, *6*, 167–173.

(68) Meier, L.; Van De Geer, S.; Bühlmann, P. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **2008**, *70*, 53–71.

(69) Fan, J.; Lv, J. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **2008**, *70*, 849–911.

(70) Harrell Jr, F. E.; Lee, K. L.; Mark, D. B. *Statistics in medicine* **1996**, *15*, 361–387.

(71) Altman, D. G.; Andersen, P. K. *Statistics in medicine* **1989**, *8*, 771–783.

(72) Liu, X.-R.; Pawitan, Y.; Clements, M. S. *Statistics in Medicine* **2017**, *36*, 4743–4762.

(73) Muliere, P.; Secchi, P. *Annals of the Institute of Statistical Mathematics* **1996**, 663–673.

(74) Rubin, D. B. *The annals of statistics* **1981**, 130–134.

(75) Efron, B. *Annals of Statistics* **1992**, *7*, 569–593.

(76) Galvani, M.; Bardelli, C.; Figini, S.; Muliere, P. *Algorithms* **2021**, *14*, 1999–4893.

(77) Hothorn, T.; Lausen, B.; Benner, A.; Radespiel-Troger, M. *Statistics in Medicine* **2004**, *23*, 77–91.

(78) Ballante, E. In *Book of Short Papers of the 51th Scientific Meeting of the Italian Statistical Society*, ed. by Balzanella, A.; Bini, M.; C., C.; Verde, R., Pearson: 2022, pp 1766–1770.

(79) Harden, J. J.; Kropko, J. *Political Science Research and Methods* **2019**, *7*, 921–928.

(80) Mitchell-Olds, T.; Shaw, R. G. *Evolution* **1987**, *41*, 1149–1161.

(81) Wei, L.-J.; Lin, D. Y.; Weissfeld, L. *Journal of the American statistical association* **1989**, *84*, 1065–1073.

(82) Cox, D. R. *Journal of the Royal Statistical Society: Series B (Methodological)* **1972**, *34*, 187–202.

(83) Hand, D. J.; Till, R. J. *Machine learning* **2001**, *45*, 171–186.

(84) Breiman, L. *Machine learning* **1996**, *24*, 123–140.

(85) Degenhardt, F.; Seifert, S.; Szymczak, S. *Briefings in Bioinformatics* **2017**, *20*, 492–503.

(86) Benjamini, Y.; Hochberg, Y. *Journal of the Royal statistical society: series B (Methodological)* **1995**, *57*, 289–300.

(87) Efron, B. *Journal of the American statistical Association* **1977**, *72*, 557–565.

(88) Chernick, M. R.; LaBudde, R. A., *An introduction to bootstrap methods with applications to R*; John Wiley & Sons: 2014.

(89) Ishwaran, H.; Kogalur, U. B. *R news* **2007**, *7*, 25–31.

(90) Ferguson, T. S. *Annals of Statistics* **1973**, *1*, 209–230.

(91) Muliere, P.; Secchi, P. *Georgian mathematical Journal* **2003**, *10*, 319–324.

(92) Ballante, E. *Far East Journal of Theoretical Statistics* **2023**, *67*, 137–146.

(93) Jordan, M. I.; Mitchell, T. M. *Science* **2015**, *349*, 255–260.

(94) MacQueen, J. In *Proceedings ofthe 5th Berkeley Symposium on Mathematical Statisticsand Probability*, 1967; Vol. 1, pp 281–297.

(95) Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X., et al. In *kdd*, 1996; Vol. 96, pp 226–231.

(96) Pizzagalli, D. U.; Gonzalez, S. F.; Krause, R. *Science advances* **2019**, *5*, eaax3770.

(97) Dijkstra, E. *Numer Math* **1959**, *1*, 101–118.

(98) Bomze, I. M.; Budinich, M.; Pardalos, P. M.; Pelillo, M. In *Handbook of Combinatorial Optimization: Supplement Volume A*; Springer: 1999, pp 1–74.

(99) Rodriguez, A.; Laio, A. *science* **2014**, *344*, 1492–1496.

(100) Du, M.; Ding, S.; Xu, X.; Xue, Y. *International Journal of Machine Learning and Cybernetics* **2018**, *9*, 1335–1349.

(101) Bron, C.; Kerbosch, J. *Communications of the ACM* **1973**, *16*, 575–577.

(102) Gatto, B. B.; dos Santos, E. M.; Molinetti, M. A.; Fukui, K. *Applied Soft Computing* **2021**, *113*, 107899.

(103) Lu, M.; Zhao, X.-J.; Zhang, L.; Li, F.-Z. *Information Sciences* **2016**, *331*, 86–98.

(104) Choudhary, A.; Kumar, S.; Gupta, S.; Gong, M.; Mahanti, A. *Energies* **2021**, *14*, 3935.

(105) Nedyalkova, M.; Sarbu, C.; Tobiszewski, M.; Simeonov, V. *Symmetry* **2020**, *12*, 1763.

(106) Yao, P.; Zhu, Q.; Zhao, R. *IEEE Transactions on Cybernetics* **2020**, *52*, 3971–3983.

(107) Śmieja, M.; Hajto, K.; Tabor, J. *Data Mining and Knowledge Discovery* **2019**, *33*, 1583–1624.

(108) Sieranoja, S.; Fränti, P. *Pattern recognition letters* **2019**, *128*, 551–558.

(109) MacQueen, J. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, 1967; Vol. 5, pp 281–298.

(110) Pizzagalli, D. U.; Gonzalez, S. F.; Krause, R. *Science Advances* **2019**, *5*, eaax3770.

(111) Turán, P. *Mat. Fiz. Lapok* **1941**, *48*, 436–452.

(112) Fränti, P.; Sieranoja, S. K-means properties on six clustering benchmark datasets, 2018.

(113) Wolff, S. D.; Chesnick, S.; Frank, J.; Lim, K.; Balaban, R. *Radiology* **1991**, *179*, 623–628.

(114) Balaban, R. S.; Ceckler, T. *Magnetic resonance quarterly* **1992**, *8*, 116–137.

(115) Wolff, S. D.; Balaban, R. S. *Magnetic resonance in medicine* **1989**, *10*, 135–144.

(116) Henkelman, R.; Stanisz, G.; Graham, S. *NMR in Biomedicine: An International Journal Devoted to the Development and Application of Magnetic Resonance In Vivo* **2001**, *14*, 57–64.

(117) McGowan, J. *Neurology* **1999**, *53*, S3–7.

(118) Sinclair, C. D.; Samson, R. S.; Thomas, D. L.; Weiskopf, N.; Lutti, A.; Thornton, J. S.; Golay, X. *Magnetic resonance in medicine* **2010**, *64*, 1739–1748.

(119) Sled, J. G.; Pike, G. B. *Journal of Magnetic Resonance* **2000**, *145*, 24–36.

(120) Sled, J. G.; Pike, G. B. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* **2001**, *46*, 923–931.

(121) Sled, J. G.; Levesque, I.; Santos, A. C.; Francis, S.; Narayanan, S.; Brass, S.; Arnold, D.; Pike, G. B. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* **2004**, *51*, 299–303.

(122) Ramani, A.; Dalton, C.; Miller, D.; Tofts, P.; Barker, G. *Magnetic resonance imaging* **2002**, *20*, 721–731.

(123) Yarnykh, V. L. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* **2002**, *47*, 929–939.

(124) Harrison, N. A.; Cooper, E.; Dowell, N. G.; Keramida, G.; Voon, V.; Critchley, H. D.; Cercignani, M. *Biological psychiatry* **2015**, *78*, 49–57.

(125) Horsfield, M. A. *Journal of Neuroimaging* **2005**, *15*, 58S–67S.

(126) Catalano, C.; Pavone, P.; Laghi, A.; Faroni, J.; Clementi, M.; Di Girolamo, M.; Passariello, R., et al. *La Radiologia Medica* **1995**, *89*, 245–249.

(127) Siger-Zajdel, M.; Selmaj, K. *Journal of Neurology, Neurosurgery & Psychiatry* **2001**, *71*, 752–756.

(128) Van Waesberghe, J.; Barkhof, F. *Neurology* **1999**, *53*, S46–8.

(129) Seiler, S.; Ropele, S.; Schmidt, R. *Journal of Alzheimer's Disease* **2014**, *42*, S229–S237.

(130) Iannucci, G.; Dichgans, M.; Rovaris, M.; Bruning, R.; Gasser, T.; Giacomotti, L.; Yousry, T.; Filippi, M. *Stroke* **2001**, *32*, 643–648.

(131) Richert, N.; Frank, J. *Neurology* **1999**, *53*, S29–32.

(132) De Stefano, N.; Battaglini, M.; Stromillo, M. L.; Zipoli, V.; Bartolozzi, M.; Guidi, L.; Siracusa, G.; Portaccio, E.; Giorgio, A.; Sorbi, S., et al. *Brain* **2006**, *129*, 2008–2016.

(133) Van Waesberghe, J.-H. T.; Kamphorst, W.; De Groot, C. J.; Van Walderveen, M. A.; Castelijns, J. A.; Ravid, R.; Lycklama a Nijeholt, G.; Van Der Valk, P.; Polman, C. H.; Thompson, A. J., et al. *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society* **1999**, *46*, 747–754.

(134) Zivadinov, R.; Dwyer, M.; Hussein, S.; Carl, E.; Kennedy, C.; Andrews, M.; Hojnacki, D.; Heininen-Brown, M.; Willis, L.; Cherneva, M., et al. *Multiple Sclerosis Journal* **2012**, *18*, 1125–1134.

(135) Button, T.; Altmann, D.; Tozer, D.; Dalton, C.; Hunter, K.; Compston, A.; Coles, A.; Miller, D. *Multiple Sclerosis Journal* **2013**, *19*, 241–244.

(136) Brown, R. A.; Narayanan, S.; Arnold, D. L. *Neuroimage* **2013**, *66*, 103–109.

(137) Arnold, D. L.; Gold, R.; Kappos, L.; Bar-Or, A.; Giovannoni, G.; Selmaj, K.; Yang, M.; Zhang, R.; Stephan, M.; Sheikh, S. I., et al. *Journal of Neurology* **2014**, *261*, 2429–2437.

(138) Nuñez-Peralta, C.; Montesinos, P.; Alonso-Jiménez, A.; Alonso-Pérez, J.; Reyes-Leiva, D.; Sánchez-González, J.; Llauger-Roselló, J.; Segovia, S.; Belmonte, I.; Pedrosa, I., et al. *Frontiers in neurology* **2021**, *12*, 634766.

(139) Nuñez-Peralta, C.; Alonso-Pérez, J.; Llauger, J.; Segovia, S.; Montesinos, P.; Belmonte, I.; Pedrosa, I.; Montiel, E.; Alonso-Jiménez, A.; Sánchez-González, J., et al. *Journal of Cachexia, Sarcopenia and Muscle* **2020**, *11*, 1032–1046.

(140) Berger, K. I.; Kanters, S.; Jansen, J. P.; Stewart, A.; Sparks, S.; Haack, K. A.; Bolzani, A.; Siliman, G.; Hamed, A. *Journal of Neurology* **2019**, *266*, 2312–2321.

(141) Li, Q.; Guan, X.; Wu, P.; Wang, X.; Zhou, L.; Tong, Y.; Ren, R.; Leung, K. S.; Lau, E. H.; Wong, J. Y., et al. *New England journal of medicine* **2020**, *382*, 1199–1207.

(142) Clark, A. E.; Nong, H.; Zhu, H.; Zhu, R. *China Economic Review* **2021**, *68*, 101629.

(143) Reuter, P. R.; Forster, B. L.; Kruger, B. J. *PeerJ* **2021**, *9*, e12528.

(144) Huang, C.; Wang, Y.; Li, X.; Ren, L.; Zhao, J.; Hu, Y.; Zhang, L.; Fan, G.; Xu, J.; Gu, X., et al. *The lancet* **2020**, *395*, 497–506.

(145) Paules, C. I.; Marston, H. D.; Fauci, A. S. *JAMA* **2020**, *323*, 707–708.

(146) Wang, C.; Cheng, Z.; Yue, X.-G.; McAleer, M. Risk management of COVID-19 by universities in China, 2020.

(147) Coronavirus, N. *Accessed on* **2020**, *10*.

(148) Onyeaka, H.; Anumudu, C. K.; Al-Sharify, Z. T.; Egele-Godswill, E.; Mbaegbu, P. *Science progress* **2021**, *104*, 00368504211019854.

(149) Webb, R. COVID-19 and lockdown: Living in 'interesting times', 2020.

(150) Daniel, S. J. *Prospects* **2020**, *49*, 91–96.

(151) Mirman, D. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, 2014; Vol. 36.

(152) Hox, J. *Multilevel Analysis: Techniques and Applications. 2nd ed. New York, NY: Routledge* **2010**.

(153) Zhu, X.; Liu, J. *Postdigital Science and Education* **2020**, *2*, 695–699.

(154) Limniou, M.; Varga-Atkins, T.; Hands, C.; Elshamaa, M. *Education Sciences* **2021**, *11*, 361.

(155) Killick, R.; Fearnhead, P.; Eckley, I. A. *Journal of the American Statistical Association* **2012**, *107*, 1590–1598.

(156) Killick, R.; Eckley, I. A. *Journal of statistical software* **2014**, *58*, 1–19.

(157) Ellis, W. E.; Dumas, T. M.; Forbes, L. M. *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement* **2020**, *52*, 177.

(158) Alharbi, R.; Alnagar, D.; Abdulrahman, A. T.; Alamri, O. *JP Journal of Biostatistics* **2021**, *18*, 231–248.

Appendix

List of Figures

List of Tables

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my supervisor, Prof. Silvia Figini, for her invaluable support and guidance throughout my PhD journey. Her trust and encouragement were instrumental in shaping this work and my development as a researcher.

I am deeply grateful to my mentors and advisors-Dr. Elena Ballante, Dr. Diego Ulysses Pizzagalli, Prof. Anna Pichiecchio, and Prof. Eduardo Caverzasi-for providing me with the opportunity to collaborate across diverse research areas. Their expertise and guidance expanded my knowledge and made this work possible.

This research was supported by a PNRR grant. I acknowledge the Department of Social and Political Sciences and the Department of Mathematics at the University of Pavia for providing the resources and facilities that enabled this work.

I extend my heartfelt thanks to all the friends I made during my PhD: Elena Ballante, Raffaella Cabini, Leonardo Barzaghi, Chiara Carrara, Giulia Guicciardi, Federico Maria Quetti, Syed Mujtaba Haider, Chiara Bonizzoni, Chiara Pullega, Ilaria Perretti, Alessia Rocchetti, Stefano Lucariello, Filippo Lascialfari, and Maria Giulia Gaggini. I will forever cherish the time we spent together—the dinners, card games during lunch breaks, and the endless stories and conversations we shared. These moments of friendship and bond made this journey truly memorable.

Finally, I am deeply grateful to my family, especially my parents, Gulnaz and Ghulam Shabbir, for their unwavering support, understanding, and prayers throughout my academic career, particularly during the six years I spent away from home pursuing my studies and personal growth. I will forever be indebted to their sacrifices. My sincere appreciation also goes to all my siblings, who kept me motivated and encouraged me to continue forward. You have been the best friends I could rely on, and your support has been invaluable to my journey.

I would also like to thank all the people I have come to know during last three years in Italy and across the globe, who contributed positively to my experience and provided opportunities for personal and professional growth.