# Statistical Machine Learning and Data Fusion Methodologies: Applications in Healthcare

The increasing availability of healthcare data from diverse sources, such as large biobanks, electronic healthcare records, medical tests, and wearable sensors, has paved the way for the development of novel machine learning (ML) models. These models aim to capture the complexity of human health and disease, thereby enhancing healthcare data analysis. This dissertation addresses three major topics within this domain, presenting innovative solutions for analyzing multi-modal mixed-type data, federated learning for functional regression, and privacy-preserving telemedicine.

The first topic introduces a Multi-modal Mixed-type Structural Equation Model (M2-SEM) with structured sparsity for subgroup discovery from heterogeneous healthcare data. This model effectively handles both continuous and categorical data modalities through a novel Gauss-Hermite-enabled Expectation-Majorization-Minimization (GHEMM) algorithm. Extensive simulation studies and applications to cardiometabolic risk factors demonstrate the model's ability to identify at-risk subgroups, highlighting its potential for enabling targeted health interventions and improving population health management.

The second topic focuses on Federated Function-on-Function Regression with an efficient Gradient Boosting algorithm (fed-GB-LSA). This approach ensures privacy preserving telemedicine by allowing collaborative model training across multiple data sources without sharing sensitive data. The GB-based algorithm facilitates the sparse selection of functional and non-functional features, providing an efficient estimation method. Its application to the telemonitoring of Obstructive Sleep Apnea (OSA) showcases the model's capability to maintain performance comparable to global models while preserving patient privacy, thereby supporting remote health monitoring and personalized treatment plans.

The third topic extends the research to Vertical Federated Learning (VFL) with Differential Privacy for function-on-function regression models. By integrating differential privacy into the federated gradient boosting process, we address the critical trade-off between model performance and privacy protection. Empirical results from simulation studies and a case study on OSA validate the method's robustness and practical relevance, demonstrating its applicability in privacy-sensitive healthcare environments where data security and patient confidentiality are paramount.

Overall, this dissertation significantly advances the field of healthcare data analysis by developing innovative machine learning models and algorithms that address the complexities of multi-modal mixed-type and functional health data. These methodologies ensure data privacy and computational efficiency, laying a strong foundation for future research and development. The findings and approaches proposed here contribute to improving health outcomes and advancing personalized medicine, ultimately enhancing healthcare delivery and patient care.