FeMAI: Föderierte Mikrobiom AI für die menschliche Gesundheit

Hinweis: Das diesem Bericht zugrundeliegende Vorhaben wurde mit Mitteln des Bundesministeriums für Forschung, Technologie und Raumfahrt unter dem Förderkennzeichen 01IS21079 (alt)/16IS21079 (neu) gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt bei den Autoren.

Teil I (Kurzbericht)

1. Ursprüngliche Aufgabenstellung sowie den wissenschaftlichen und technischen Stand, an den angeknüpft wurde

Das Projekt FeMAI zielte darauf ab, Methoden des föderierten maschinellen Lernens für die Analyse mikrobieller Daten zu entwickeln und in einer sicheren, datenschutzkonformen Umgebung bereitzustellen. Die Entstehung von kolorektalen Karzinomen (CRC) wird neben Faktoren wie Ernährung, Lebensstil und Komorbiditäten auch von einem dysbiotischen Darm-Mikrobiom beeinflusst (Dong and Gupta 2019). Durch die Quantifizierung dieses Mikrobioms lassen sich Aussagen über die Entstehung und Progression von CRC treffen (Zwezerijnen-Jiwa et al. 2023). Technisch wurde auf etablierten Analyseverfahren für Sequenzdaten, auf Ontologien für medizinische Datenintegration sowie auf föderierten Lernansätzen aufgebaut, die im Rahmen des FeatureCloud-Systems (Matschinske et al. 2021) bereits grundgelegt waren. Diese bestehenden Werkzeuge und Standards wurden erweitert, kombiniert und für den speziellen Anwendungsfall der Mikrobiomforschung adaptiert.

2. Ablauf des Vorhabens

Das deutsch-französische FeMAI Konsortium bestehend aus der Universität Hamburg (UHH), gefördert vom BMFTR, und dem französischen Partner INRAE MetaGenoPolis Institut, gefördert durch die französische Agence nationale de la recherche (ANR), vereinte alle notwendigen Fachkompetenzen für die erfolgreiche Umsetzung des Projekts: Mikrobiom-Daten Analyse und Datenbanken (INRAE) und föderiertes Lernen (UHH). Diese Verteilung der Fachkompetenzen spiegelt sich auch im Arbeitsplan wider, der aus 5 Arbeitspaketen besteht. Das Vorhaben war durch eine intensive Kollaboration der beiden Projektpartner gekennzeichnet.

Die Arbeitspakete gliederten sich entlang der Datenakquise, Datenintegration, Analyse, Evaluation und Verwertung:

Arbeitspaket 1: Aufbau eines umfassenden Mikrobiom-Datensatzes aus öffentlichen Quellen, ergänzt durch medizinische Metadaten, und deren Integration in eine föderierte Datenbankstruktur **Arbeitspaket 2:** Konzeption und Implementierung eines modularen Datenbanksystems sowie Entwicklung begleitender Tools wie Datenimport-App, Containerisierung der Analyseumgebungen und standardisierte Präprozessierungspipelines **Arbeitspaket 3:** Entwicklung, Evaluation und Optimierung föderierter Analyseverfahren, u. a. Batch-Effect-Korrektur, Pseudo-Count Normalisierung, Random-Forest-Modelle (One-Shot, Multi-Shot) sowie Dimensionsreduktionsmethoden **Arbeitspaket 4:** Systematische Evaluation föderierter Methoden im Vergleich zu zentralisierten und lokalen Methoden anhand von CRC- und Typ 2 Diabetes-Kohorten. Ergänzend: Absicherung der Privatsphäre und Cybersicherheit durch kontinuierliches Monitoring, Watchtower-Dienste und Snyk **Arbeitspaket 5:** Koordination, regelmäßiger wissenschaftlicher Austausch, Dissemination durch Beiträge auf internationalen Konferenzen und Öffentlichkeitsarbeit

Alle Arbeitspakete wurden in enger Zusammenarbeit zwischen UHH und INRAE bearbeitet und jeder Partner war für jeweils konkrete Aufgaben innerhalb der Arbeitspakete zuständig.

3. Wesentliche Ergebnisse sowie ggf. die Zusammenarbeit mit anderen Forschungseinrichtungen

Das Projekt hat wesentliche Beiträge zur Entwicklung und Validierung föderierter Lernverfahren in der Mikrobiomforschung geleistet. Ergebnisse sind unter anderem:

- Aufbau und Integration eines standardisierten Mikrobiom-Datensatzes mit 1722 qualitätsgeprüften Proben aus 14 Studien (Barbet et al. 2022)
- Entwicklung einer robusten Präprozessierungspipeline (fastp (Chen et al. 2018), Bowtie2 (Langmead and Salzberg 2012), Meteor (Ghozlane et al. 2025), CroCoDeEL (Goulet et al. 2024)), validiert durch MetaPhlAn4 (Blanco-Míguez et al. 2023)
- Implementierung föderierter Analyseverfahren, darunter One-Shot und Multi-Shot Random Forests (Qi 2012), die vergleichbare Performanz zu zentralisierten Ansätzen erreichen und zugleich robuster als lokale Modelle sind
- Etablierung einer föderierten Analysepipeline, die über die FeatureCloud-Plattform (Matschinske et al. 2021) öffentlich zugänglich ist (https://featurecloud.ai/workflow/53)
- Systematische Evaluation föderierter vs. traditioneller Modelle, die den Vorteil föderierter Methoden insbesondere bei heterogenen und kleinen Kohorten belegt
- Implementierung von Cybersicherheitsmaßnahmen (z. B. Watchtower (https://github.com/containrrr/watchtower), Snyk (https://snyk.io/)), die eine sichere Nutzung der föderierten Infrastruktur gewährleisten
- Dissemination durch internationale Konferenzbeiträge (PMNET 2022, IHMC 2022, eNMD 2023, GCB 2024) sowie Präsentationen vor einem breiten Publikum (Science City Hamburg Bahrenfeld 2024)

4. Referenzen

Barbet, Pauline, Mathieu Almeida, Niklas Probul, Jan Baumbach, Nicolas Pons, Florian Plaza Onate, and Emmanuelle Le Chatelier. 2022. “Taxonomic Profiles, Functional Profiles and Manually Curated Metadata of Human Fecal Metagenomes from Public Projects Coming from Colorectal Cancer Studies.” Recherche Data Gouv. https://doi.org/10.57745/7IVO3E.

Blanco-Míguez, Aitor, Francesco Beghini, Fabio Cumbo, Lauren J. McIver, Kelsey N. Thompson, Moreno Zolfo, Paolo Manghi, et al. 2023. “Extending and Improving Metagenomic Taxonomic Profiling with Uncharacterized Species Using MetaPhlAn 4.” *Nature Biotechnology* 41 (11): 1633–44.

Chen, Shifu, Yanqing Zhou, Yaru Chen, and Jia Gu. 2018. “Fastp: An Ultra-Fast All-in-One FASTQ Preprocessor.” *Bioinformatics (Oxford, England)* 34 (17): i884–90.

Dong, Tien S., and Arpana Gupta. 2019. “Influence of Early Life, Diet, and the Environment on the Microbiome.” *Clinical Gastroenterology and Hepatology: The Official Clinical Practice Journal of the American Gastroenterological Association* 17 (2): 231–42.

Ghozlane, Amine, Florence Thirion, Florian Plaza Oñate, Franck Gauthier, Emmanuelle Le Chatelier, Anita Annamalé, Mathieu Almeida, Stanislav Ehrlich, and Nicolas Pons. 2025. “Accurate Profiling of Microbial Communities for Shotgun Metagenomic Sequencing with Meteor2.” *Research Square*. https://doi.org/10.21203/rs.3.rs-6122276/v1.

Goulet, Lindsay, Florian Plaza Oñate, Edi Prifti, Eugeni Belda, Emmanuelle Le Chatelier, and Guillaume Gautreau. 2024. *CroCoDeEL: CROss-Sample COntamination DEtection and Estimation of Its Levels* (version 1.0.5).

Langmead, Ben, and Steven L. Salzberg. 2012. “Fast Gapped-Read Alignment with Bowtie 2.” *Nature Methods* 9 (4): 357–59.

Matschinske, Julian, Julian Späth, Reza Nasirigerdeh, Reihaneh Torkzadehmahani, Anne Hartebrodt, Balázs Orbán, Sándor Fejér, et al. 2021. “The FeatureCloud AI Store for Federated Learning in Biomedicine and Beyond.” *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/2105.05734.

Qi, Yanjun. 2012. “Random Forest for Bioinformatics.” In *Ensemble Machine Learning*, 307–23. New York, NY: Springer New York.

Zwezerijnen-Jiwa, Florine H., Hugo Sivov, Petra Paizs, Konstantina Zafeiropoulou, and James Kinross. 2023. “A Systematic Review of Microbiome-Derived Biomarkers for Early Colorectal Cancer Detection.” *Neoplasia* 36 (February):100868.