KI-SIGS

KI-Space für intelligente Gesundheitssysteme

Programm / Ausschreibung	IKT der Zukunft, IKT der Zukunft, IKT der Zukunft - Vorbereitung Horizon Europe	**Status**	abgeschlossen
Projektstart	01.10.2020	**Projektende**	30.09.2023
Zeitraum	2020 - 2023	**Projektlaufzeit**	36 Monate
Keywords	KI; Erklärbarkeit; Vertraulichkeit; Transparenz; Gesundheitswesen;		

Projektbeschreibung

Das Projekt widmet sich dem brisanten Thema der Vertraulichkeit von KI Modellen im Gesundheitswesen, um zu verhindern, dass sensible Patientendaten durchsickern oder (mittels anderer KI Modelle) rekonstruiert werden können. Dabei sollen aber die KI-Modelle dennoch erklärbar bleiben. Da Erklärbarkeit nach Transparenz verlangt, Transparenz allerdings Vertraulichkeit unterminieren kann, gilt es, diesen Zusammenhang mathematisch zu modellieren und die Effekte zu quantifizieren.

Ziel des Projekts ist daher die Entwicklung eines neuartigen informations-theoretischen Frameworks zur Untersuchung und Optimierung des Gegensatzes zwischen Privatsphäre und Erklärbarkeit von Algorithmen für maschinelles Lernen im Zusammenhang mit Gesundheitsdaten.

Die Innovation besteht in der Entwicklung eines informationstheoretischen Maßes, um beide Aspekte, Datenschutz im Sinne von Vertraulichkeit und Erklärbarkeit, zu quantifizieren. Dabei werden relevante Themen für den Einsatz von KI Modellen im Gesundheitswesen behandelt, die über den Stand der Technik hinausgehen: 1) Wie kann die informationelle Privatsphäre optimiert werden, wenn die statistische Verteilung der Daten unbekannt ist? 2) Wie kann die Erklärbarkeit eines maschinellen/tiefen Lernmodells quantifiziert werden? 3) Wie kann das Zusammenspiel zwischen Privatsphäre, Erklärbarkeit und Nutzen untersucht werden? 4) Wie untersucht man die Auswirkung der Wahrung der Privatsphäre auf die Übertragbarkeit von Wissen aus einem Bereich in einen anderen verwandten Bereich?

Als Ergebnis strebt das Projekt ein Software-Framework an, das den Entwurf und die Analyse von KI-Modellen hinsichtlich Wahrung von Vertraulichkeit unterstützt und so die Grundlage für neue Standards und Best-Practices im Umgang mit Gesundheitsdaten und darauf beruhender KI Modelle schafft.

Abstract

This project addresses the critical issue of the privacy of AI models in healthcare to prevent sensitive patient data from leaking or (using other AI models) being reconstructed. However, the AI models should still remain explainable. Since explainability requires transparency, but transparency can undermine privacy, it is necessary to model this relationship

mathematically and quantify the effects.

Therefore, the aim of the project is to develop a novel information-theoretical framework to investigate and optimize the dichotomy between privacy and explainability of machine learning algorithms in the context of health data.

The innovation consists in the development of an information-theoretical measure to quantify both aspects, data protection in terms of privacy and explainability. Relevant topics for the use of AI models in health care are addressed that go beyond the state of the art: 1) How can informational privacy be optimized if the statistical distribution of data is unknown? 2) How can the explainability of a machine/deep learning model be quantified? 3) How can the interaction between privacy, explainability and utility be investigated? 4) How to investigate the impact of privacy on the transferability of knowledge from one domain to another related domain?

The result will be a software framework that supports the design and analysis of AI models with regard to privacy and explainability, thus creating the basis for new standards and best practices in the handling of healthcare data and AI models based on them.

Projektpartner

- Software Competence Center Hagenberg GmbH