Reflective Essay Assignment

Lena Stempfle

23rd August 2023

1 Research Description: Machine learning for Healthcare

My research revolves around the application of machine learning (ML) for healthcare, with a specific focus on addressing healthcare challenges through the development and application of ML methods and algorithms. The central aim of my research projects is to construct ML models capable of predicting diagnoses and supporting decision-making processes in the treatment of various medical conditions. A key aspect of my research involves addressing scenarios where inputs to the ML models are missing, either during the model's training phase or at the time of making predictions. Inadequate handling of missing values can introduce bias or render models impractical for real-world use without imputing the values of unobserved variables. However, traditional imputation methods often fall short and can be challenging to interpret, particularly when dealing with complex imputation functions [17, 18]. To be able to make highly accurate predictions in situations characterized by incomplete data during testing, I mainly focus on utilizing interpretable models that facilitate human understanding of the model's predictions. Interpretability is especially crucial in healthcare contexts where critical decisions are made. Our findings (as outlined in [23]) demonstrate that the proposed interpretable models can effectively perform predictions even in the presence of missing values, eliminating the need for dependency on imputation techniques. In more recent work, we showed how to learn replacement variables in interpretable rule-based models when variables are sparsely observed. Currently, I am collaborating closely with the dermatology department at Sahlgrenska Hospital in Gothenburg. Our joint effort aims to predict melanoma using data sourced from Swedish registries. In a separate project, I have partnered with Traumabase, an organization comprising fifteen French trauma centers. Their extensive decentralized database contains comprehensive information on over 14,000 trauma cases, spanning from admission to ICU discharge. To delve further into the realm of interpretable ML for healthcare, it's important to highlight how these models provide not only accurate predictions but also insights into the factors influencing those predictions. A basic understanding of software engineering concepts and methods is an indispensable prerequisite in this research field.

Note, that I have not participated in the lectures in person and therefore, the content of this assignment is solely based on the written material.

2 Two concepts from Robert's lectures

Hidden Technical Debt in Machine Learning Systems The paper "Hidden Technical Debt in Machine Learning Systems" by Sculley et al. [21] explores the concept of hidden technical debt in the context of ML systems, which refers to the extra work and

complexity that arises as a result of design and implementation decisions made during the development process. Traditional technical debt arises from shortcuts taken during software development to achieve rapid development or meet deadlines, and the authors argue that ML projects are also prone to accumulating such debt due to factors like data dependencies, model decay, and changing requirements.

Developing software in health care settings shows several occasions to produce hidden technical debts. Healthcare data is often complex and messy, and preprocessing this data is a critical step in building accurate models. Ad hoc data preprocessing practices in healthcare can lead to hidden technical debt as data quality issues may not be apparent initially but can impact model performance and patient safety later on. Rigorous testing and validation are paramount in healthcare ML to ensure that models are accurate, safe, and reliable. Ignoring comprehensive testing can result in hidden technical debt, as incorrect or biased predictions can have serious consequences in a healthcare setting.

Other strategies proposed in the paper such as clear coding standards, and collaboration, are directly applicable to healthcare ML projects. These practices can help ensure that the models remain effective, safe, and maintainable throughout their lifecycle.

Behavioral Software Engineering Behavioral Software Engineering (BSE) is an approach in software development that places a strong emphasis on understanding and addressing human behavior and cognitive processes throughout the software development lifecycle. It recognizes that software systems are ultimately designed for people to use and interact with, so it takes into consideration human factors, user behavior, and psychological aspects to create software that is not only functional but also user-friendly and effective [20].

The research field of explainable AI strives to provide explanations to humans so that decisions made by AI systems are comprehensible and understandable. This transparency is of great importance, especially in safety-critical applications such as medicine or autonomous vehicles [5]. A point made in work by Ghassemi et al. [8] is that people trust advice when it is explained, mainly they over-trust systems where a person believes the robot can perform a function that it cannot, e.g. health or person expects that the system will mitigate the risk. Models providing more "transparency", e.g. by providing explanations hamper people's ability to detect when a model makes serious mistakes.

Such behavior is described as cognitive bias as part of BSE on an individual level, leading to incorrect beliefs about the completeness and correctness of answers or solutions [3]. Cognitive biases may affect human understanding of interpretable ML models as well, in particular of logical rules discovered from data [12]. Since my research focuses on providing small interpretable models, this aspect is important for me to keep in mind for the future.

3 Two concepts from the guest lecture by Fabio Calefato

Software Engineering-specific Sentiment Analysis Document polarity, also known as document-level sentiment analysis, involves determining the overall semantic orientation (positive, negative, neutral) or emotional tone expressed in an entire document. In healthcare, this could be applied to patient reviews, medical notes, or social media discussions related to medical treatments or healthcare experiences. Analyzing the document-level sentiment can provide insights into patient satisfaction, opinions about treatments, and the overall sentiment associated with healthcare services. Word polarity, or word-level sentiment analysis, focuses on determining the sentiment of individual words or text fragments.

Each word is assigned a polarity label indicating whether it is positive, negative, or neutral in meaning. This analysis can be used to understand the sentiment associated with specific medical terms, drugs, symptoms, or treatments. Word-level sentiment analysis can be particularly useful for understanding the emotional connotations of medical terminology and identifying potential bias or misinformation in patient discussions.

As mentioned in the material, limitations include indirect indicators of sentiment which therefore give misleading results. Recent work in my field, applying sentiment analysis to clinal notes, showed that compared with white patients, black patients had 2.54 times the odds of having at least one negative descriptor in the history and physical notes [24]. The same work, raises concerns about stigmatizing language in electronic health records (EHR) and its potential to exacerbate racial and ethnic healthcare disparities.

Automated Machine Learning (AutoML) AutoML refers to the use of automated tools and techniques to streamline and simplify the process of designing, training, and deploying ML models. In the context of sentiment analysis in software engineering for healthcare, AutoML offers a way to leverage ML without requiring extensive expertise in software engineering computer science, or natural language processing knowledge. For sentiment analysis tasks, such as text cleaning, tokenization, and removing stopwords AutoML tools can automatically handle data preprocessing steps. This is especially beneficial when dealing with unstructured text data from patient reviews, clinical notes, or social media discussions related to healthcare. Relevant features can be automatically extracted from text data using n-grams or word embeddings.

AutoML automates various tasks; however, engineers must customize sentiment analysis models to account for the complexity of medical terminology and nuances in patient feedback, reducing the risk of misinterpretations. This potential for impactful change in healthcare software development underscores AutoML's role in making ML accessible to non-experts. Nonetheless, integration success relies on considering healthcare context, data privacy, and ethics, ensuring accurate sentiment analysis within healthcare applications.

4 Choose 2 Topics

Human-Computer Interaction Human-Computer Interaction (HCI) is a multidisciplinary field that focuses on the design, evaluation, and optimization of interactive systems, where "systems" encompass a wide range of digital technologies such as software applications, websites, mobile apps, and even hardware devices [16]. The primary goal of HCI is to create user-centered interfaces and experiences that facilitate effective and efficient interactions between humans and computers. HCI takes into account user needs, preferences, and behaviors, as well as the context in which interactions occur, to ensure that technology is user-friendly, intuitive, and aligned with human capabilities [16].

In high-stake areas such as healthcare, HCI can help support decision-making, foster communication and improve time efficiency [19]. Specifically, in healthcare, the interaction of AI/ML technologies and experts can be used to improve medical diagnosis, treatment, and patient engagement. Note, in areas with sensitive applications even greater safety requirements are needed. As we have seen models like ChatGPT can also hallucinate, and be capable of the spread of misinformation [1].

When engaging patients in their own health management, AI-powered interfaces can provide personalized recommendations and insights to patients, empowering them to take

proactive measures to improve their well-being [19]. One way humans communicate with machines is through voice control. Amazon's Alexa and Apple's Siri are common examples of these speech-operated agents, which many people already use on a daily basis. However, language contains two types of information: textual and emotional. While the textual part can be already understood grammatically well by large language models, e.g. ChatGPT, to achieve a harmonious human-computer interaction experience, the computer should automatically recognize the emotional content of voice signals. Recent work shows speech obtains lots of information, conveying information about the speaker's inner condition and their aim and desire. To classify different emotions in a multiclass classification model, a multilayer perceptron (MLP) estimator was used on the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) open-source data set [2]. Aside from speech, data visualization is an effective way to communicate complex medical data and AI-driven insights to healthcare practitioners, enabling better diagnosis and treatment planning. My own research field of interpretable ML has a related goal to HCI, where the predictions made by ML models are interpretable and comprehensible to healthcare professionals. For instance, a work where you search more efficiently by minimizing the number of unnecessary trials to find an effective medical treatment for patients [10] or traditionally used risk scores are learned by data and presented to clinicians [26].

A recent HCI example from industry is the Care-O-bot 4 which is a state-of-the-art service robot developed by Fraunhofer Institute for Manufacturing Engineering and Automation IPA in Germany [9]. It is designed to assist with various tasks in healthcare, household, and service environments and offers personalized assistance. Key features of Care-O-bot 4 include its humanoid design, enabling it to navigate and interact with human environments. Equipped with a range of sensors, cameras, and touch-sensitive surfaces, it can recognize and respond to human gestures, voice commands, and environmental cues. Additionally, there is the recently presented Med-PaLM M (MPM) that is based on PaLM-E, Google's robot model that combines language and vision. The model can be used for triage of patients, retrieval of knowledge, summarization of key findings, and diagnosis assistance [25].

Regulations and Compliance When working with ML using health care data, there are numerous challenges and risks in conforming to current regulations and compliance. Throughout the process from data collection until model deployment, one must comply with a range of regulatory requirements such as the Health Insurance Portability and Accountability Act (HIPAA), and General Data Protection Regulation (GDPR) ensuring patient data is protected and not exposed to unauthorized parties penteado2022regulation. One example, is that Articles 17 and 19 of the GDPR state the 'right to be forgotten', meaning everyone has the right to have their data erased, without undue delay, by the data controller, if one of the following grounds applies: Where personal data are no longer necessary in relation to the purpose for which it was collected or processed [7].

One would make use of this right in situations when they had painful, embarrassing, life-affecting old disciplinary reports, news articles, criminal convictions, images, videos, religious beliefs, negative posts on social media, divorces, and general information about them that they will never forget but which they wanted to not be reminded of by other people [6].

In order to be able to put such a legal provision into practice a relatively new and challenging field of research has emerged. The research area of *Machine unlearning (MU)* is concerned with developing techniques for removing sensitive or irrelevant data from trained

models [13]. MU is the process of removing specific data points or features from a trained ML model without affecting its performance [22, 11]. The goal of MU is to ensure that trained models are free from biases and sensitive information that could lead to negative outcomes [27]. Originally Cao et al. [4] introduced an approach to efficiently remove data traces by converting learning algorithms into a summation form which can help counter data pollution attacks. Until today, diverse approaches and techniques are under development for data unlearning. These span from regularization methods to model inversion techniques. Additionally, other studies update model weights for unlearning purposes using the entire training data, a subset of it, or stored metadata from training [14].

Nonetheless, challenges persist in this realm. Scaling up to larger datasets, the selective unlearning of specific data subsets, and the repercussions of unlearning on model performance remain unresolved. Straightforward techniques like complete model retraining or check-pointing can incur substantial computational and storage costs [15]. Furthermore, situations may arise where training-related data is unavailable for unlearning purposes.

5 Future trends and directions of Software Engineering

In the field of Software Engineering for ML in healthcare, there are several future trends and directions that are emerging. As healthcare systems continue to generate vast amounts of data from various sources, interoperability, and data integration will become critical. Software engineers will need to develop robust systems that can seamlessly integrate data from EHRs, medical devices, wearables, and other sources. This will involve designing and implementing standards-based interfaces and data exchange protocols. With sensitive patient data being used by software, ensuring privacy and security will be paramount. By focusing on robust security measures to protect patient data and implementing data anonymization techniques, healthcare software users can comply with regulatory requirements such as HIPAA or GDPR.

To support end-users, e.g. healthcare professionals, in their daily work, ML models used in healthcare should be interpretable and provide explanations for their predictions or recommendations. This helps to increase the confidence of healthcare professionals and enables them to understand and validate the results produced by ML algorithms. Many interpretable ML methods enable troubleshooting and integration of clinical expertise. Domain-specific expertise can be brought into the development process through collaboration between software developers and healthcare professionals. In addition, software must be developed in the future that can process and analyze streaming data in real-time to enable immediate warnings and interventions when needed. This requires the use of technologies such as edge computing, stream processing frameworks, and a scalable infrastructure.

In summary, the outlook for the relevance of software engineering to the development of ML methods in healthcare is promising over the next 5-10 years. The existence of supporting tools, such as AutoML, will make it easier for more people to build their own software, at least to create quick baselines on which to build and create better models. However, I believe that in critical applications such as healthcare and justice, working closely with professionals to integrate knowledge, comply with regulatory requirements, and meet user-specific needs still requires a high degree of software engineering skills.

References

[1] Hussam Alkaissi and Samy I McFarlane. 'Artificial hallucinations in ChatGPT: implications in scientific writing'. In: *Cureus* 15.2 (2023).

[2] Abeer Ali Alnuaim et al. 'Human-computer interaction for recognizing speech emotions using multilayer perceptron classifier'. In: *Journal of Healthcare Engineering* 2022 (2022).

[3] Astrid Bertrand et al. 'How cognitive biases affect XAI-assisted decision-making: A systematic review'. In: *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. 2022, pp. 78–91.

[4] Yinzhi Cao and Junfeng Yang. 'Towards Making Systems Forget with Machine Unlearning'. In: *2015 IEEE Symposium on Security and Privacy*. 2015, pp. 463–480. DOI: 10.1109/SP.2015.35.

[5] Finale Doshi-Velez and Been Kim. *Towards A Rigorous Science of Interpretable Machine Learning*. 2017. arXiv: 1702.08608 [stat.ML].

[6] Thorsten Eisenhofer et al. 'Verifiable and provably secure machine unlearning'. In: *arXiv preprint arXiv:2210.09126* (2022).

[7] European Parliament and Council of the European Union. *Regulation (EU) 2016/679 of the European Parliament and of the Council*. of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). 4th May 2016. URL: https://data.europa.eu/eli/reg/2016/679/oj (visited on 13/04/2023).

[8] Marzyeh Ghassemi, Luke Oakden-Rayner and Andrew L Beam. 'The false hope of current approaches to explainable artificial intelligence in health care'. In: *The Lancet Digital Health* 3.11 (2021), e745–e750.

[9] Birgit Graf and Jonathan Eckstein. 'Service Robots and Automation for the Disabled and Nursing Home Care'. In: *Springer Handbook of Automation*. Springer, 2023, pp. 1331–1347.

[10] Samuel Håkansson et al. 'Learning to search efficiently for causally near-optimal treatments'. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 1333–1344.

[11] Zachary Izzo et al. 'Approximate data deletion from machine learning models'. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2021, pp. 2008–2016.

[12] Tomáš Kliegr, Štěpán Bahnık and Johannes Fürnkranz. 'A review of possible effects of cognitive biases on interpretation of rule-based machine learning models'. In: *Artificial Intelligence* 295 (2021), p. 103458.

[13] Meghdad Kurmanji, Peter Triantafillou and Eleni Triantafillou. 'Towards Unbounded Machine Unlearning'. In: *arXiv preprint arXiv:2302.09880* (2023).

[14] Thanh Tam Nguyen et al. 'A survey of machine unlearning'. In: *arXiv preprint arXiv:2209.02299* (2022).

[15] Bruno Elias Penteado et al. 'The Regulation of Artificial Intelligence in Healthcare: An Exploratory Study'. In: *Proceedings of the 2022 Computers and People Research Conference*. 2022, pp. 1–5.

[16] Fuji Ren and Yanwei Bao. 'A review on human-computer interaction and intelligent robots'. In: *International Journal of Information Technology & Decision Making* 19.01 (2020), pp. 5–47.

[17] Donald B Rubin. 'Inference and missing data'. In: *Biometrika* 63.3 (1976), pp. 581–592.

[18] Donald B Rubin. 'Multiple imputation after 18+ years'. In: *Journal of the American statistical Association* 91.434 (1996), pp. 473–489.

[19] François Sainfort et al. 'Human–Computer Interaction in Healthcare'. In: *Human-Computer Interaction*. CRC Press, 2009, pp. 155–172.

[20] Mary Sánchez-Gordón and Ricardo Colomo-Palacios. 'Taking the emotional pulse of software engineering—A systematic literature review of empirical studies'. In: *Information and Software Technology* 115 (2019), pp. 23–43.

[21] David Sculley et al. 'Hidden technical debt in machine learning systems'. In: *Advances in neural information processing systems* 28 (2015).

[22] Thanveer Basha Shaik et al. 'Exploring the Landscape of Machine Unlearning: A Comprehensive Survey and Taxonomy.' In: *CoRR* (2023).

[23] Lena Stempfle and Fredrik Johansson. 'Sharing pattern submodels for prediction with missing values'. In: *arXiv preprint arXiv:2206.11161* (2022).

[24] Michael Sun et al. 'Negative Patient Descriptors: Documenting Racial Bias In The Electronic Health Record: Study examines racial bias in the patient descriptors used in the electronic health record.' In: *Health Affairs* 41.2 (2022), pp. 203–211.

[25] Tao Tu et al. 'Towards generalist biomedical ai'. In: *arXiv preprint arXiv:2307.14334* (2023).

[26] Berk Ustun and Cynthia Rudin. 'Learning Optimized Risk Scores.' In: *J. Mach. Learn. Res.* 20.150 (2019), pp. 1–75.

[27] Haibo Zhang et al. 'A review on machine unlearning'. In: *SN Computer Science* 4.4 (2023), p. 337.