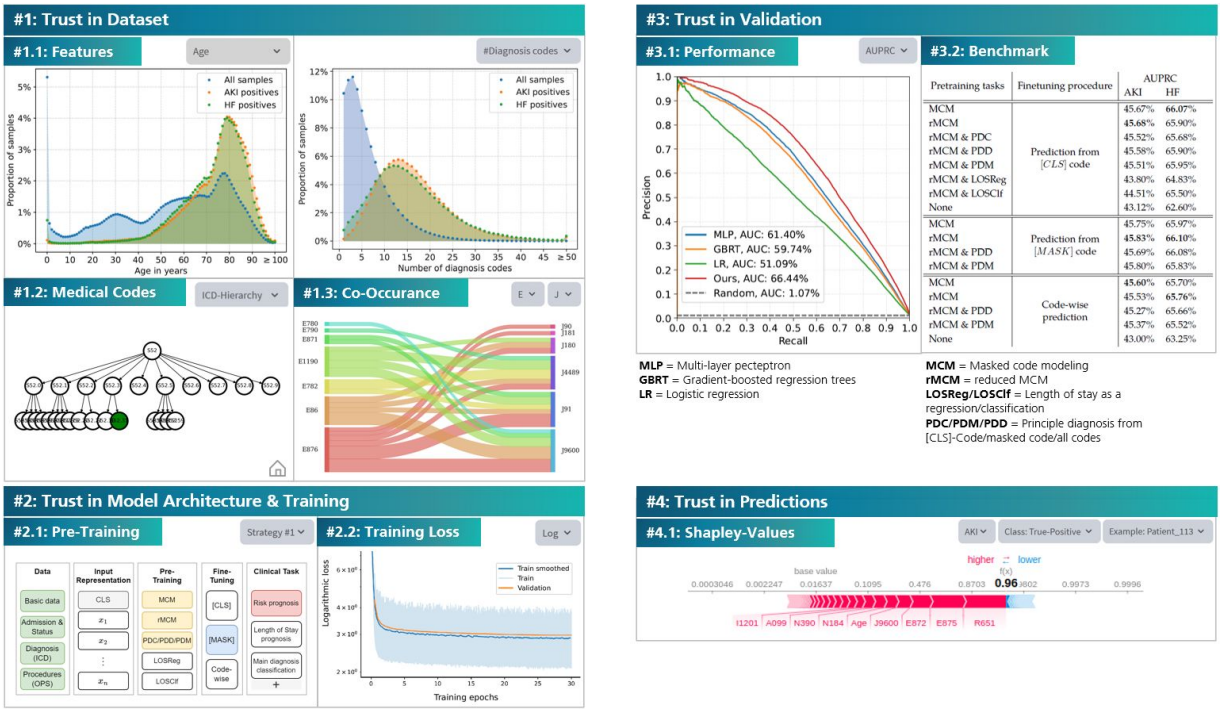
Multi-Task Transformer Visualization to build Trust for Clinical Outcome Prediction

Dario Antweiler*
Fraunhofer IAIS, Fraunhofer Center for Machine Learning

Florian Gallusser†
Fraunhofer IAIS

Georg Fuchs‡
Fraunhofer IAIS



Pretraining tasks	Finetuning procedure	AUPRC AKI	AUPRC HF
MCM	Prediction from [CLS] code	45.67%	66.07%
rMCM		45.68%	65.90%
rMCM & PDC		45.52%	65.68%
rMCM & PDD		45.58%	65.90%
rMCM & PDM		45.51%	65.95%
rMCM & LOSReg		43.80%	64.83%
rMCM & LOSCIf		44.51%	65.50%
None		43.12%	62.60%
MCM	Prediction from [MASK] code	45.75%	65.97%
rMCM		45.83%	66.10%
rMCM & PDD		45.69%	66.08%
rMCM & PDM		45.80%	65.83%
MCM	Code-wise prediction	45.60%	65.70%
rMCM		45.53%	65.76%
rMCM & PDD		45.27%	65.66%
rMCM & PDM		45.37%	65.52%
None		43.00%	63.25%

MLP = Multi-layer pecteptron
GBRT = Gradient-boosted regression trees
LR = Logistic regression

MCM = Masked code modeling
rMCM = reduced MCM
LOSReg/LOSCIf = Length of stay as a regression/classification
PDC/PDM/PDD = Principle diagnosis from [CLS]-Code/masked code/all codes

Figure 1: Overview of the proposed visual analytics system that fosters trust into clinical transformer models, consisting of multiple interactive views: ① Trust in dataset with feature distribution plots and coordinated hierarchical medical code visualization & co-occurance diagrams ② Trust in model architecture & training with architecture diagram and training loss graph ③ Trust in validation with precision-recall/ROC curves & baseline benchmarks and ④ Trust in prediction with Shapley-values to display feature importance for individual predictions.

Abstract

Clinical decision support systems based on machine learning are a rising application in healthcare. Early detection of deteriorating conditions provide the opportunity for medical intervention in hospital patients. Recent approaches increasingly rely on Large Language Models such as BERT, because patient data is often in the form of structured temporal data. These models are notoriously hard to interpret and therefore to trust, while precisely trust is an essential principle for technology in healthcare. We develop a visual analytics system to inspect, compare, and explain pre-trained transformer models for a given clinical outcome prediction task. The work is developed on the basis of a large hospital patient dataset and prediction tasks for acute kidney injury and heart failure. Discussion with healthcare professionals confirms that our system can lead to a faster decision process and improved modeling results.

*e-mail: dario.antweiler@iais.fraunhofer.de
†e-mail: flo.gallusser@gmail.com
‡e-mail: georg.fuchs@iais.fraunhofer.de

1 Introduction

Predicting clinical risk for patients in hospital wards is a highly sought after goal for healthcare information systems [13]. Conditions like acute kidney injury (AKI) or heart failure (HF) affect more than half of adult patients in the intensive care unit [16]. In recent years, machine learning based Clinical Decision Support Systems (CDSS) have seen a rising interest and a promising track record of partial success in clinical use [32]. Designing, developing, implementing, and using these systems is a complex interplay of technological, social, economical, and regulatory challenges. The development phase includes the consolidation of practical requirements with technical feasibility, and should incorporate both computer scientists and healthcare professionals. Typical challenges from the viewpoint of clinicians include i) the need for high accuracy to combat alert fatigue [8], ii) the ability to examine and understand the decision process of a system [33], iii) the usage of medical codes stemming from complex hierarchies not intended for that purpose, and iv) the black box nature of Large Language Models (LLMs).

Multi-task transformer models are of special interest, because they can learn to achieve multiple objectives at once using shared representations. In healthcare applications specifically, objectives can include predicting risk, length of stay, selecting the principal diagnosis, or identifying missing medical codes as well as other

tasks that may arise in the future [18]. This model class often involves multi-billion parameter-size neural network architectures, which are inherently hard to understand for humans. We observe that i) and ii) call for an appropriate balance between trust and knowledge on limits towards a newly developed model, while iii) and iv) increase the difficulty of achieving that. Visual analytics systems enable users to explore datasets, models, and outcomes of machine learning models and have been shown to increase trust in them [9]. In healthcare specifically, previous work has shown that visual analytics enables domain experts to identify better performing models based on high-dimensional data [20]. Therefore, we aim to answer the following research questions:

Q_1: How can visual analytics support healthcare professionals and data scientists to build trust in multitask transformer models?
Q_2: How can we visualize the dependencies between data, model, and output within complex medical hierarchical terminologies?

We introduce a visual analytics dashboard which addresses these questions. It includes four separate but interlinked views, which guide users through all major machine learning model development phases and are tailored specifically towards transformer models as well as hierarchical medical data. Our proposed system is applicable to different use cases, as many current healthcare analytics applications rely heavily on structured temporal information to predict clinical outcomes and need to deal with hierarchical medical data. We display an overview of our proposed system in Fig. 1. To summarize, our main contributions are:

- We introduce an interactive visual dashboard to support healthcare professionals and data scientists to assess multi-task transformer models.
- We design a trust-building approach along the development steps of a clinical outcome prediction model.
- We describe the utilization of our framework to the real-world use case of predicting AKI and HF from medical code sets.
- We conduct qualitative interviews with health professionals as to clinical applicability and discuss our lessons learned.

The rest of the paper is structured as follows: review of related publications (Sec. 2), characterization of the dataset used in development (3.1), specification of the model (3.2), identification of appropriate users and tasks (3.3) as well as a detailed description of the proposed system (4.1). An exemplary workflow showcases the usage in practice (4.2). We conclude with final remarks and an outlook of future research (5).

2 Related Work

Our paper focuses on the intersection of clinical decision support, visual analytics, and trustworthy machine learning, specifically the class of transformer models. We divide existing approaches into the following three areas of interest.

Transformer Model Visualization Since the birth of the Transformer architecture in 2017 [36] and its subsequent success story, multiple attempts at visualizing the model and the corresponding outputs have been made due to the nature of its complexity. For the application area of NLP, a summarizing survey can be found at [7]. Existing approaches can be roughly grouped into architecture-agnostic [23] or dedicated to a specific component of the transformer architecture, e.g., the attention mechanism [12, 15, 37] or the contextual token representations [34, 40]. Additionally, the literature includes systems targeting the comparison of multiple transformer models [38] or dedicated to certain data types [28]. Some of them have been successfully integrated into visual analytics tools [1, 21, 22].

Visual Analytics for Clinical Outcome Prediction Within the large body of work on healthcare-related visual analytics [25], we identified approaches targeted at risk prediction visualization. They include graphical representations of simple statistical models [6, 26], disease-specific approaches, e.g., for cancer [14, 24], or infectious diseases [3, 29] and bespoke systems for specific model classes, e.g., rule sets [2]. Our approach is most closely related to *VBridge* [11], which employs a hierarchical visualization to connect data, model, and explanations for clinical machine learning models.

Trust & Explainability in Visual Analytics Trust is related but not identical to explainability in machine learning [35]. Building trust is a complex sociological and psychological challenge, that can change dynamically and be highly personal [5]. As such, visual analytics systems cannot entirely solve, but rather support this process. Trust can be increased through transparency, robustness, and fairness of a system [4]. Tangible measures within visual analytics include displaying uncertainty [19], declaring data provenance [27], or letting the user explore surrogate models [10, 39]. Recent surveys give an overview of present approaches [9, 17].

The body of work presented is lacking several major requirements: Firstly, it is not suited for the hierarchical feature sets prevalent in medical applications. Secondly, it is insufficient in fostering trust along the entire data science development process comprised of data understanding, architecture selection, model training, and validation. In contrast, our approach combines multiple customized views to support users through the trust-building process.

3 Context

3.1 Data

Our real-world dataset consists of individual inpatient visits between 2011 and 2020 in multiple German hospitals. After cleaning, it contains 24m samples that each include age, sex, attending department, length of stay, admission year, principal diagnosis, and, most importantly, two sets of medical codes describing diagnosis and procedures respectively. The former set is encoded as diagnosis codes from ICD-GM-10[1] with 13.376 unique codes across all visits and an average of 6.54 codes per visit. The latter set is encoded as operation and procedure codes from OPS[2], a German classification system with 27.960 unique codes and an average of 3.04 codes per visit. Additionally, it includes timestamps that date the code assignments and patients' admissions.

As the dataset covers a highly diverse set of patients ranging from 0 to 120 years of age, including both sexes and covering all hospital departments, the medical codes are not distributed evenly, e.g., 15.7% of all samples have no associated procedure code at all and 40.7% of procedures are performed on the day of admission to the hospital. Similarly, 73.3% of all samples have a length of stay (LOS) of less than seven, while 6.7% stay more than 14 nights in hospital. The departments most common in our dataset are internal medicine, general surgery, and gynecology/obstetrics, which together cover 46% of all samples.

We preprocess the data for the classification tasks by creating labels based on the definition of AKI and HF provided by a collaborating healthcare professional. The label definitions are rule sets that each consist of the inclusion or exclusion of specific medical codes. After computing both labels, we create two separate training datasets where medical codes are excluded if they are part of the label definition for each task respectively. The resulting datasets contain 549.944 (2.3%) positive samples for AKI and 269.476 (1.1%) positive samples for HF with 67.728 (0.3%) positive for both.

The dataset is of corporate nature and was provided anonymized for scientific analysis within this research project.

[1] International Statistical Classification of Diseases, German Modification https://www.bfarm.de/EN/Code-systems/Classifications/ICD/ICD-10-GM/_node.html

[2] Operationen- und Prozedurenschlüssel https://www.bfarm.de/EN/Code-systems/Classifications/OPS-ICHI/OPS/_node.html

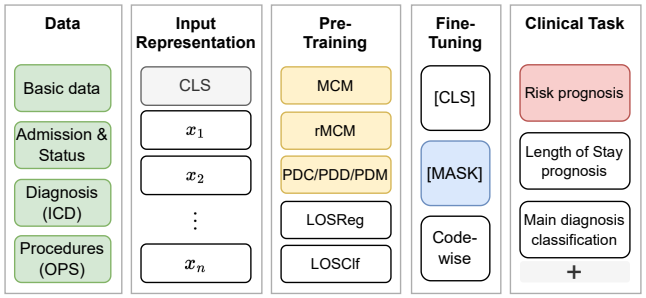


Figure 2: Detailed view of the model architecture diagram within our system (② in Fig. 1). It displays the flow of data from input data via representation to pre-training and fine-tuning strategies within the transformer-based model to finally the clinical tasks which can be extended to address future problems.

3.2 Multi-Task Transformer Model

As outlined in Sect. 1, multi-task transformer models are of special interest, because they can learn to achieve multiple objectives corresponding to functionally related domain tasks at once using shared representations. Instead of training multiple highly-specific models that are dedicated to a single task, a multi-task model can easily be adapted to novel tasks. This reduces training time and computational resources required.

We use a BERT-based architecture to model our task of binary classification based on patient data regarding a clinical outcome, e.g., AKI (cf. Fig. 2). The available data per patient described above is used to construct a single input sequence to the model. We introduce the artificial tokens *[MASK]* and *[CLS]* depending on the selected pre-training task and fine-tuning strategy as suggested in the literature. Pre-training tasks include *masked code modeling* (MCM) as well as prediction of principle diagnosis (PD) and length of stay (LOS). Our final multi-task model is trained with a weighted average of all individual losses. After pre-training, we fine-tune the model to our binary classification task. The architecture is implemented in PyTorch, trained with mixed precision and smart batching on hardware consisting of Intel(R) Xeon(R) Silver 4116 CPU @ 2.10GHz, 377GB RAM, and two NVIDIA Tesla V100 GPUs. We use hyperparameters suggested for the original BERT model and do an 80/10/10% train-validation-test split. Our best performing model consists of 117m parameters and obtains an Area under Precision Recall Curve (AUPRC) score of 0.46% for AKI and 0.66% for HF prediction respectively. Surprisingly, a smaller model of the same architecture featuring 10× less trainable parameters achieves similar performance (cf. Tab. 1). The larger model outperforms baseline models including gradient boosted regression trees (GBRT), multi-layer perceptron (MLP), and logistic regression (LR). It serves as the use case for the development of our visual analytics system.

3.3 Users & Tasks

We are developing this approach for a user group consisting of both data scientists and healthcare professionals working together. It should facilitate cooperation and communication between the members of the group towards the creation of a CDSS. We aim to support the building of trust towards a model and lay the basis for an informed discussion. Based on lessons learned from multiple healthcare analytics projects and interviews conducted with domain experts, we gathered a set of tasks:

- **T$_1$**: Inspect hierarchical medical code datasets represented within their respective classification systems
- **T$_2$**: Consult on different pre-training and fine-tuning strategies within the transformer architecture
- **T$_3$**: Compare different model architectures and hyperparameter settings regarding clinical outcome prediction quality
- **T$_4$**: Verify factors of influence for individual predictions

It should be noted that currently employed tools do not support "cross-lingual" communication between medical and computer science domains. During discussions with prospective users, we identified the following requirements:

- **R$_1$**: Use of widely familiar visualizations that are accessible both to data scientists and healthcare professionals
- **R$_2$**: Use of suitable terms and definitions for domain experts and model developers

We will refer back to these users and tasks throughout the rest of the paper to demonstrate how we address them.

4 Proposed Visual Analytics System

We developed our dashboard on the basis of the aforementioned tasks and requirements. It is implemented as a web-interface to increase accessibility via any browser. We use python in the backend, the Streamlit library as a frontend, which together can be distributed as a single Docker image to minimize implementation efforts. Our goal is to foster trust into a trained transformer model by combining multiple trust aspects including i) increasing transparency ii) surveying for fairness, and iii) ensuring robustness.

4.1 Dashboard overview

The dashboard is divided into four segments (Fig. 1, #1–#4) that each address a specific process step along the model development.

① Trust in Dataset (T_1)

The quality of training data is a key factor for prediction quality. Bias and imbalances in the data will impact the fairness of a system and may also decrease its robustness and reliability in medical decision-making due to epistemic uncertainty [30].

In order to facilitate trust into the underlying training dataset, we provide a set of complementary views (#1). Histogram plots (R_1) show the distribution of user-selected features (#1.1). Feature distributions are split along the outcome classes to showcase differences, i.e., between AKI positive and the general patient cohort. Users including healthcare professionals can assess whether the distribution of input data contains any inherent bias and conforms to the local conditions of a healthcare facility [32]. E.g., the age distribution severely differs depending on the location and has an influence on model congruence.

Core data features are the hierarchical ICD and OPS codes. Both hierarchies are very familiar to health professionals (R_2), however this is typically constrained to their specific medical department. The chapters directly correspond to related groups of diagnosis or procedures. Medical coding is affected by official guidelines, local customs, as well as personal preference. Thus, it is relevant to discern common from unusual code combinations, especially regarding risk prediction. The two complementary views (#1.2, #1.3) support users in exploring possible causal relationships between codes, even within unfamiliar chapters.

A tree-based view (#1.2) allows to explore both medical code hierarchies, as selected by the drop-down in the top right. Interaction is simplified in that it is limited to only two types of actions, either selecting a lower level node to expand the hierarchy at this point, or using the "home" button at the lower right to reset the view. This is an effective way to traverse the seven or five hierarchy levels of ICD or OPS respectively (R_1).

The current selection in the hierarchy view also updates the co-occurence sankey diagram of medical codes (#1.3). Each segment represents a single medical code and two codes are connected via link whenever patient stays contain both of them. The weight of the link is determined by the relative count of co-occurances, normalized

Table 1: Comparison of transformer models with different sizes to predict acute kidney injury (AKI) or heart failure (HF) for hospital patients. Selected pre-training and fine-tuning strategies are masked code modeling (MCM) and code-wise prediction. Training time includes pre-training and fine-tuning. Hyperparameters refer to embedding size M, attention heads H, and transformer encoder layers L.

Transformer size	AUPRC		Train time	#params
M, H, L	AKI	HF		
768, 12, 12	**46.3%**	**66.4%**	35h+10h	117.6m
192, 8, 8	45.6%	65.7%	7h+2h	11.9m

to the current selection. The user can further filter by chapters or subchapters of interest from the pre-selection done via view #1.2 using the dropdowns in the top right.

② Trust in Model Architecture & Training (T_2)

For any machine learning application, it is crucial to assess whether a particular architecture class is suitable among available choices. This typically also includes considering tradeoffs between model performance and explainability, with implications on robustness and transparency facets of model trust. While transformers are currently one of the most popular choices in the machine learning community, many healthcare professionals have yet to encounter them in practice. We visualize the selected model architecture and training strategy via a flow-chart style diagram (#2.1). This follows the objective to at least partially illuminate the interaction between the main components of a transformer model, the input representation, encoder, and fine-tuning tasks.

Alongside, we include a logarithmic line chart showing train and test loss across training epochs (#2.2). This targets two goals: Checking convergence of the training scheme as well as assessing the model's generalizability through the difference between train and validation loss.

③ Trust in Validation (T_3)

Datasets in healthcare are often highly imbalanced, e.g., in our use case only 2.3% of patients are positively labeled for AKI. This not only influences model training, but also model validation. Chosen evaluation metrics must be capable of uncovering potential impacts on robustness and fairness of model predictions. Metrics like precision or recall should always be assessed in combination with another. Although the area under the receiver operating characteristic (AUROC) does satisfy this requirement, it is unsuitable as a singular metric for situations with uneven class distributions, as is in many medical use cases. The precision-recall curve on the other hand better distinguishes models for unbalanced datasets, albeit it requires the calibration of the threshold parameter. Therefore, we employ a combination of multiple evaluation metrics available to the user (#3.1). Through switching between the precision-recall curve and the ROC-curve, the model performance can be investigated thoroughly. In healthcare machine learning, there is often a trade-off between model complexity and performance. To this end, our proposed system supports the comparison of the current transformer model with a fixed set of established baselines.

Juxtaposed is a table comparing combinations of pre-training and fine-tuning strategies (#3.2) with top-performing entries printed in bold. We opted for a structured depiction, as it is easily comprehensible (R_1), superior to more complex visualizations when portraying slight differences, and simple to extend. Note, that the left panel shows the best-performing model, while the table on the right displays values attained by smaller models due to computational overhead.

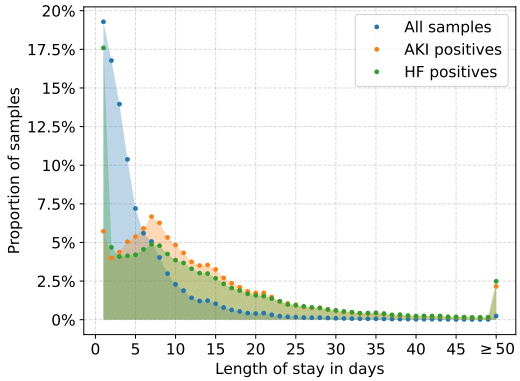


Figure 3: Detail view from our proposed system: Histogram plots showing the distribution of lengths of stay for hospital patients in the cohort ①. Cases that are positive for either acute kidney injury (AKI) or heart failure (HF) show a distinctive pattern, correlating with longer stays, especially for patients with ≥ 50 medical codes.

④ Trust in Outcome Predictions (T_4)

Finally, after inspecting the architecture, training, and evaluation of an entire model, the remaining trust uncertainty concerns individual predictions. Patients and clinicians alike expect models to explain their outputs in order to explore decision paths, alternatives, and reasoning. We employ a SHAPley value visualization [31] to inform users about feature contributions when considering an individual patient (#4.1). Users can select a prediction class (true/false, positive/negative) and within the class from a set of exemplary patients to explore relationships between medical codes and model predictions. From discussions with healthcare professionals in multiple projects, we observed a high accessibility of SHAPley plots due to their simplicity (R_1). Co-occuring codes can be cross-referred with view #1.2 to check for atypical combinations across chapters. Additionally, the influence of demographic features like age, sex as well as categorical features like admission type and medical department is visualized.

4.2 Workflow

In the following, we want to detail an exemplary workflow from clinical practice. We consulted healthcare professionals (HCPs) and data scientists through structured interviews to discuss our proposal, expectations of users, and to gain further insights into the practical usage. The workflow is explicitly motivated by and detailed along the real-world dataset and transformer model described in Sections 3.1 and 3.2.

① Initially, the users wanted to inspect the training dataset. One mentioned question asked for a comparison of the local patient cohort to the model's training data to identify systematic biases, such as a shifted age distribution towards younger patients or a higher mean disease severity typically present in urban hospitals. In our instance, healthcare professionals would like to explore the distribution of length of stay, comparing AKI patients with the general cohort. After selecting "Length of stay" as the relevant feature, the user reconstructs two facts from the visualization (cf. Fig. 3): Existence of both AKI and HF correlates positively with length of stay and patients with a large (≥ 50) number of medical codes are especially affected. This resulted in an increase in reported trust into the model, as this confirmed pre-existing assumptions. Afterward, the users moved to views #1.2 and #1.3. Through the tree visualization, they navigated to the chapters "E" (*endocrine, nutritional, and metabolic diseases*) and "J" (*diseases of the respiratory system*) of the ICD hierarchy. Within the sankey diagram the users explored co-occurances and amongst others identified a strong connection between codes "E87.6" (*hypokalemia*) and "J91" (*pleural effusion*),

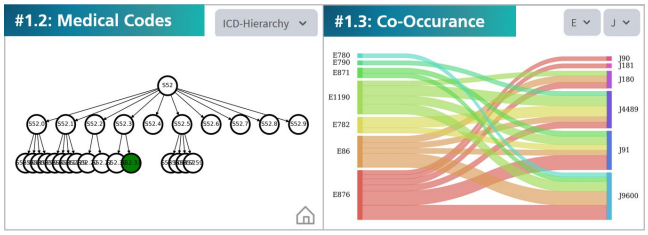


Figure 4: Detail view from our proposed system: The tree visualization for the ICD hierarchy next to the sankey diagram displaying the co-occurance of codes during patient stays ①

both well-known comorbidities, further increasing trust (cf. Fig. 4).

② With the second component, the attention shifts towards the model architecture. During the interviews, it became clear that an overview of the main model components was necessary. Within the view #2.1, users can retrace the choices that were needed to select a specific model architecture. From user feedback, we gathered that two insights were particularly helpful: First, the fact that all available patient data, including admission dates and diagnosis/procedure codes were combined into a single representation and secondly, that more clinical tasks are viable for future modeling. In combination with the evaluation of performance in the following section, this boosted trust into the architecture choice, as it became clear, that this flexible framework was able to solve multiple challenges at once. Moving to #2.2, users performed a quick glance to assess that the training loss converged steadily. During an exchange between HCP and data scientist, the graph was used to assert that generalizability was achieved, i.e., similar loss on validation compared to training data. Both views were more easily approached by data scientists as anticipated, yet HCPs could apprehend all major conclusions during the interdisciplinary discussion due to the clear visuals and labeling.

③ Next, users explored the benchmark between the newly proposed model against alternatives, namely established machine learning algorithms within view #3.1. During the interviews, we observed that users checked the displayed curves for significant (e.g., transformer vs. LR) or insignificant differences (e.g., MLP vs. GBRT). In addition, HCPs used the view to assess whether the model could replace already existing deployed systems in terms of specific precision and recall combinations. It needs to be both robust enough to identify most patients at risk and precise enough to limit alert fatigue in clinical use. When viewing the AUROC curve by using the dropdown selector, users were bewildered for a moment. It suggests, at first glance, that all compared models performed nearly perfectly. The confusion was fully dissolved after the switch to the AUPRC curve and in exchange with the data scientists, as it is better suited for highly-imbalanced data and clearly shows qualitative differences. This is a cautionary example for possible overtrust into a model, which our visual analytics system averts. To the right the view #3.2 was distinguished by both HCPs and data scientists for simplicity and using the same visual ordering as in #2.1. In this specific case, the choice of the pre-training strategy has only a minor impact on model performance and should therefore play a secondary role in selecting a model architecture.

④ Last, the user group turns to the view targeting individual model predictions. They quickly recognize the separation into features that increase and those that decrease the predication probability via the colored arrow segments. The focus naturally shifted towards high SHAPley values, indicating larger relevance. Furthermore, data scientists analyzed the view towards spurious correlations the model might have picked up, as one user had experienced in a different use case previously. With the limited number of examples, the lack of such correlations, increased the trust of the model. HCP users additionally asked for a colorization of features by hierarchy chapter within ICD. We plan to integrate toggling as a feature in an upcoming version.

4.3 Discussion

Our approach includes the visualization of aspects regarding dataset, model architecture, pre-training tasks, model training, validation, and prediction, thereby assisting users along the entire development process. Each process step addresses different but interlocking facets of trust, categorized into transparency, fairness, and robustness.

Interdisciplinary teams often face challenges in terms of different vocabularies, expectations, and conventions. Our design strives to bridge the gap between HCPs and data scientists in particular by using established, accessible visualizations and interactions to facilitate communication and mutual understanding. For example, the model architecture view proved to be highly valuable to convey to HCPs the essentials, role, and impact of various architecture and fine-tuning variations of transformer models by the data scientist. A lack of this common basis of understanding has proven to be a roadblock in several of our previous projects. Overcoming this roadblock increases the effectiveness of interdisciplinary discussions regarding ③ trust in validation and ④ trust in outcome predictions.

The utility of our approach and system have been checked during an iterative design process involving HCPs. Unsurprisingly, findings from interim interviews reinforced the notion that complex hierarchical data sets such as ICD and OPS must be made accessible to HCP from multiple perspectives informed by domain conventions. Specifically, we found that HCPs regularly expect CDSS to provide the ability to confirm their prior domain knowledge and personal empirical experience is captured in the model, to the point a positive, if episodic, confirmation is a prerequisite for their willingness to follow through with the process of CDSS evaluation/adaptation.

One concrete finding of the described use case was that applying optimization through our approach resulted in a model that performed less than 1% worse than the original model 10x larger. After confirming they could indeed trust the smaller model with its performance still sufficient, HCPs stated it would be the preferred choice in clinical practice due to ease of deployment on restricted (and thus often capability limited) hardware approved for medical use. This further underlines the utility of our approach.

Our approach does have limitations, including the lack of laboratory or vital patient data, as well as missing timestamps for ICD codes. Medical codes are an imperfect representation of reality and cannot capture the complexity of a clinical picture. Auxiliary data is therefore desirable to close the gap between research and practice. We strive to include additional features such as medication or pre-existing comorbidities in a prospective extension to our system.

5 Conclusion and Future Work

In this paper, we contributed a novel visual analytics system to enable healthcare professionals, with support from data scientists, to inspect and evaluate CDSS based on multi-task transformer model architectures with the overarching goal of facilitating trust building.

As further future work, we would like to augment our system with additional views depicting prediction uncertainty and systematic weaknesses of the model. Lastly, recent developments on all-purpose clinical predictive engines should be evaluated against use-case specific approaches not only regarding quality of prediction but also explainability and robustness [18].

Acknowledgments

This work was partially done within the SmartHospital.NRW project, funded by the Ministry of Economic Affairs, Industry, Climate Action and Energy of the State of North Rhine-Westphalia, Germany. In addition, this work was funded by the Fraunhofer Research Center for Machine Learning within the Fraunhofer Cluster of Excellence Cognitive Internet Technologies.

REFERENCES

[1] J. Alammar. Ecco: An open source library for the explainability of transformer language models. ACL, 2021. doi: 10.18653/v1/2021.acl-demo.30

[2] D. Antweiler and G. Fuchs. Visualizing rule-based classifiers for clinical risk prognosis. In *VIS*, 2022. doi: 10.1109/vis54862.2022.00020

[3] D. Antweiler, D. Sessler, M. Rossknecht, B. Abb, S. Ginzel, and J. Kohlhammer. Uncovering chains of infections through spatio-temporal and visual analysis of COVID-19 contact traces. *C&G*, 106:1–8, Aug. 2022. doi: 10.1016/j.cag.2022.05.013

[4] O. Asan, A. E. Bayrak, and A. Choudhury. Artificial intelligence and human trust in healthcare: Focus on clinicians. *JMIR*, 22(6):e15154, 2020. doi: 10.2196/15154

[5] E. Beauxis-Aussalet, M. Behrisch, R. Borgo, D. H. Chau, C. Collins, D. Ebert, M. El-Assady, A. Endert, D. A. Keim, J. Kohlhammer, D. Oelke, J. Peltonen, M. Riveiro, T. Schreck, H. Strobelt, and J. J. van Wijk. The role of interactive visualization in fostering trust in AI. *IEEE CG&A*, 41(6):7–12, Nov. 2021. doi: 10.1109/mcg.2021.3107875

[6] V. V. Belle and B. V. Calster. Visualizing risk prediction models. *PLOS ONE*, 10(7):e0132614, 2015. doi: 10.1371/journal.pone.0132614

[7] A. M. P. Brasoveanu and R. Andonie. Visualizing transformers for NLP: A brief survey. In *IEEE IV*, Sept. 2020. doi: 10.1109/iv51561.2020.00051

[8] B. Cánovas-Segura, A. Morales, J. M. Juarez, and M. Campos. Meaningful time-related aspects of alerts in clinical decision support systems. a unified framework. *JBI*, 143:104397, July 2023. doi: 10.1016/j.jbi.2023.104397

[9] A. Chatzimparmpas, R. M. Martins, I. Jusufi, K. Kucher, F. Rossi, and A. Kerren. The state of the art in enhancing trust in machine learning models with the use of visualizations. *CGF*, 39(3):713–756, 2020. doi: 10.1111/cgf.14034

[10] A. Chatzimparmpas, R. M. Martins, A. C. Telea, and A. Kerren. Deforestvis: Behavior analysis of machine learning models with surrogate decision stumps, 2023. doi: 10.48550/arxiv.2304.00133

[11] F. Cheng, D. Liu, F. Du, Y. Lin, A. Zytek, H. Li, H. Qu, and K. Veeramachaneni. VBridge: Connecting the dots between features and data to explain healthcare models. *IEEE TVCG*, 28(1):378–388, Jan. 2022. doi: 10.1109/tvcg.2021.3114836

[12] A. Coenen, E. Reif, A. Yuan, B. Kim, A. Pearce, F. Viégas, and M. Wattenberg. Visualizing and measuring the geometry of bert, 2019. doi: 10.48550/arxiv.1906.02715

[13] A. Couffinhal. Tackling Wasteful Spending on Health, 2017.

[14] M. Cypko, J. Wojdziak, M. Stoehr, B. Kirchner, B. Preim, A. Dietz, H. U. Lemke, and S. Oeltze-Jafra. Visual verification of cancer staging for therapy decision support. *CGF*, 2017. doi: 10.1111/cgf.13172

[15] B. Hoover, H. Strobelt, and S. Gehrmann. exBERT: A visual analysis tool to explore learned representations in transformer models. In *ACL 2020*. ACL, 2020. doi: 10.18653/v1/2020.acl-demos.22

[16] E. A. J. Hoste, J. A. Kellum, N. M. Selby, A. Zarbock, P. M. Palevsky, S. M. Bagshaw, S. L. Goldstein, J. Cerdá, and L. S. Chawla. Global epidemiology and outcomes of acute kidney injury. *Nature Rev. Neph.*, 14(10):607–625, Aug. 2018. doi: 10.1038/s41581-018-0052-0

[17] Y. Jia, J. McDermid, T. Lawton, and I. Habli. The role of explainability in assuring safety of machine learning in healthcare. *IEEE TETC*, 10(4):1746–1760, 2022. doi: 10.1109/tetc.2022.3171314

[18] L. Y. Jiang, X. C. Liu, N. P. Nejatian, M. Nasir-Moin, D. Wang, A. Abidin, K. Eaton, H. A. Riina, I. Laufer, P. Punjabi, M. Miceli, N. C. Kim, C. Orillac, Z. Schnurman, C. Livia, H. Weiss, D. Kurland, S. Neifert, Y. Dastagirzada, D. Kondziolka, A. T. M. Cheung, G. Yang, M. Cao, M. Flores, A. B. Costa, Y. Aphinyanaphongs, K. Cho, and E. K. Oermann. Health system-scale language models are all-purpose prediction engines. *Nature*, 2023. doi: 10.1038/s41586-023-06160-y

[19] C. Kinkeldey, T. Korjakow, and J. J. Benjamin. Towards supporting interpretability of clustering results with uncertainty visualization. *TrustVis at EuroVis*, 2019. doi: 10.2312/TRVIS.20191183

[20] J. Krause, A. Perer, and E. Bertini. INFUSE: Interactive feature selection for predictive modeling of high dimensional data. *IEEE TVCG*, 20(12):1614–1623, Dec. 2014. doi: 10.1109/tvcg.2014.2346482

[21] V. Lal, A. Ma, E. Aflalo, P. Howard, A. Simoes, D. Korat, O. Pereg, G. Singer, and M. Wasserblat. InterpreT: An interactive visualization tool for interpreting transformers. In *ACL 2021 European Chapter*. ACL, 2021. doi: 10.18653/v1/2021.eacl-demos.17

[22] R. Li, W. Xiao, L. Wang, H. Jang, and G. Carenini. T3-vis: visual analytic for training and fine-tuning transformers in NLP. In *EMNLP 2021*. ACL, 2021. doi: 10.18653/v1/2021.emnlp-demo.26

[23] S. Liu, T. Li, Z. Li, V. Srikumar, V. Pascucci, and P.-T. Bremer. Visual interrogation of attention-based models for natural language inference and machine comprehension. In *EMNLP*. ACL, 2018. doi: 10.18653/v1/d18-2007

[24] G. E. Marai, C. Ma, A. T. Burks, F. Pellolio, G. Canahuate, D. M. Vock, A. S. R. Mohamed, and C. D. Fuller. Precision risk analysis of cancer therapy with interactive nomograms and survival plots. *IEEE TVCG*, 25(4):1732–1745, 2019. doi: 10.1109/tvcg.2018.2817557

[25] J. Ooge, G. Stiglic, and K. Verbert. Explaining artificial intelligence with visual analytics in healthcare. *WIREs Data Mining and Knowledge Discovery*, 12(1), 2021. doi: 10.1002/widm.1427

[26] R. Osuala and O. Arandjelovic. Visualization of patient specific disease risk prediction. In *2017 EMBS BHI*. IEEE, 2017. doi: 10.1109/bhi.2017.7897250

[27] E. D. Ragan, A. Endert, J. Sanyal, and J. Chen. Characterizing provenance in visualization and data analysis: An organizational framework of provenance types and purposes. *IEEE TVCG*, 22(1):31–40, Jan. 2016. doi: 10.1109/tvcg.2015.2467551

[28] L. Schwenke and M. Atzmueller. Show me what you're looking for. *FLAIRS*, 34(1), Apr. 2021. doi: 10.32473/flairs.v34i1.128399

[29] S. Schäfer, T. Baumgartl, A. Wulff, A. Kuijper, M. Marschollek, S. Scheithauer, and T. von Landesberger. Interactive Visualization of Machine Learning Model Results Predicting Infection Risk. In M. Krone, S. Lenti, and J. Schmidt, eds., *EuroVis 2022*. The Eurographics Association, 2022. doi: 10.2312/evp.20221113

[30] R. Senge, S. Bösner, K. Dembczyński, J. Haasenritter, O. Hirsch, N. Donner-Banzhoff, and E. Hüllermeier. Reliable classification: Learning classifiers that distinguish aleatoric and epistemic uncertainty. *Inf. Sci.*, 255:16–29, 2014. doi: 10.1016/j.ins.2013.07.030

[31] L. S. Shapley. 17. a value for n-person games. In *Contributions to the Theory of Games (AM-28), Volume II*, pp. 307–318. Princeton University Press, 1953. doi: 10.1515/9781400881970-018

[32] H. Sun, K. Depraetere, L. Meesseman, P. C. Silva, R. Szymanowsky, J. Fliegenschmidt, N. Hulde, V. von Dossow, M. Vanbiervliet, J. D. Baerdemaeker, D. M. Roccaro-Waldmeyer, J. Stieg, M. D. Hidalgo, and F.-M. Dahlweid. Machine learning–based prediction models for different clinical risks in different hospitals: Evaluation of live performance. *JMIR*, 24(6):e34295, 2022. doi: 10.2196/34295

[33] S. Tonekaboni, S. Joshi, M. D. McCradden, and A. Goldenberg. What clinicians want: Contextualizing explainable machine learning for clinical end use, 2019. doi: 10.48550/arxiv.1905.05134

[34] B. van Aken, B. Winter, A. Löser, and F. A. Gers. Visbert: Hidden-state visualizations for transformers. 2020. doi: 10.48550/arXiv.2011.04507

[35] S. van den Elzen, G. Andrienko, N. Andrienko, B. D. Fisher, R. M. Martins, J. Peltonen, A. C. Telea, and M. Verleysen. The flow of trust: A visualization framework to externalize, explore, and explain trust in ML applications. *IEEE CG&A*, 43(2):78–88, Mar. 2023. doi: 10.1109/mcg.2023.3237286

[36] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need, 2023.

[37] J. Vig. A multiscale visualization of attention in the transformer model. In *ACL*, pp. 37–42, 2019. doi: 10.18653/v1/P19-3007

[38] J.-L. Wu, P.-C. Chang, C. Wang, and K.-C. Wang. Aticvis: A visual analytics system for asymmetric transformer models interpretation and comparison. *Applied Sciences*, 13(3), 2023. doi: 10.3390/app13031595

[39] J. Yuan, B. Barr, K. Overton, and E. Bertini. Visual exploration of machine learning model behavior with hierarchical surrogate rule sets, 2022. doi: 10.48550/arxiv.2201.07724

[40] Z. Yun, Y. Chen, B. Olshausen, and Y. LeCun. Transformer visualization via dictionary learning: contextualized embedding as a linear superposition of transformer factors. In *DeeLIO 2021*. ACL, 2021. doi: 10.18653/v1/2021.deelio-1.1