Explaining Uncertainty: Exploring the Synergies of Explainable Artificial Intelligence and Uncertainty Quantification

Emily Schiller[1]

[1]XITASO GmbH IT & Software Solutions, Austraße 35, 86153 Augsburg, Germany

Abstract

Despite the transformative potential of machine learning in critical areas such as healthcare, its integration is challenged by a lack of trust, reliability, and transparency. Explainable AI (XAI) aims to make AI predictions interpretable for humans. At the same time, Uncertainty Quantification (UQ) enables the estimation of confidence in predictions — both crucial for responsible AI usage and increasing trust and transparency. The intersection of the two research fields holds significant potential to advance these objectives further; however, research in this area has been limited. This doctoral proposal addresses the need to investigate the intersection of XAI and UQ. My research will emphasize explaining uncertainty estimates in healthcare applications and time series tasks. By developing computationally efficient methods to identify sources of uncertainty in AI predictions, my research seeks to enhance model performance and interpretability. The challenge of scheduling nurses serves as a consistent case study to identify real-world challenges in healthcare applications. By introducing novel techniques to explain uncertainty estimates, this work will explore the synergies of XAI and UQ, contributing to developing more transparent and reliable AI systems and ultimately advancing their integration into high-stakes domains.

Keywords

Explainable AI, Uncertainty Quantification, Time Series, Healthcare

1. Background and Motivation

During the past decade, remarkable progress has been made in the field of deep neural networks, leading to their adoption across various research disciplines, including earth observation, healthcare, and autonomous systems. Nevertheless, their practical use in critical real-world scenarios is still limited, as acceptance and trust among users remain insufficient [1]. The intersection of Explainable AI (XAI) and Uncertainty Quantification (UQ) presents a promising way to enhance the reliability and trustworthiness of Artificial Intelligence (AI) systems, particularly in critical domains such as healthcare. My doctoral research aims to explore this intersection, focusing on explaining uncertainty estimates as a key direction.

1.1. Bridging Explainability and Uncertainty in AI

The field of XAI is currently experiencing significant research activity, with the goal of making AI predictions more interpretable for both end users and data scientists [2]. It covers techniques that are used to turn a non-interpretable model into an explainable one. By offering transparency, XAI not only enhances user trust, but also facilitates the identification and mitigation of biases and errors within AI models. Furthermore, XAI supports regulatory compliance by aligning with the demands for accountability and ethical standards in AI deployment [2].

Alongside XAI, UQ is crucial for ensuring trust, safety, and reliability in AI systems [3, 4]. Predictive uncertainty includes aleatoric uncertainty, arising from data distribution, and epistemic uncertainty, stemming from model limitations like data sparsity. Both types of uncertainty provide valuable insights

3rd World Conference on eXplainable Artificial Intelligence, July 09–11, 2025, Istanbul, Turkey

emily.schiller@xitaso.com (E. Schiller)
0009-0004-7348-3266 (E. Schiller)

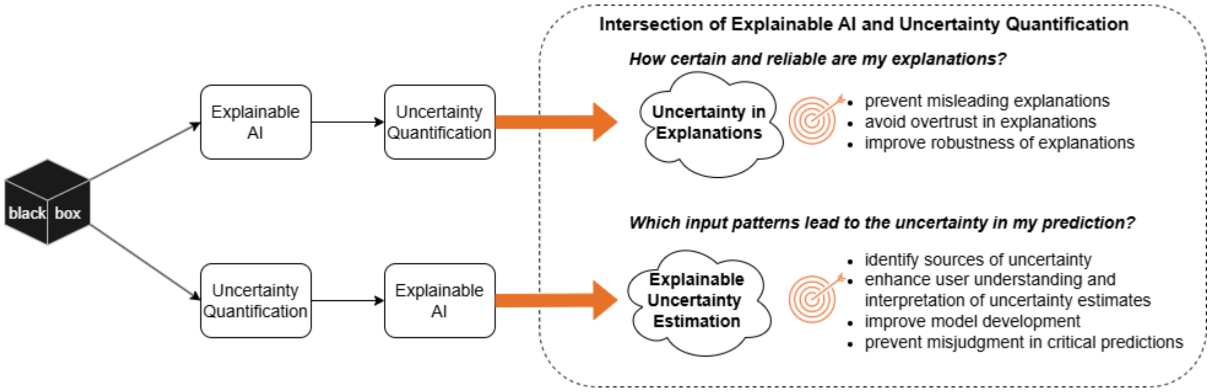

Figure 1: Two ways of combining Explainable AI and Uncertainty Quantification.

into the confidence of a prediction. Moreover, UQ is a valuable tool for improving robustness to unseen data, improving decision-making processes, and promoting transparency [1, 3].

Both XAI and UQ are fields dedicated to enhancing the reliability of AI models, promoting responsible AI usage, and building user trust. Additionally, XAI and UQ enable data scientists to evaluate the AI models, uncover their weaknesses and ultimately improve them. The intersection of the two fields of research holds the potential to further advance these objectives. So far, there has been limited research on the intersection of XAI and UQ, although interest in this area is growing [4, 5, 6], making it a crucial moment to contribute to this emerging field.

More specifically, the intersection of XAI and UQ can be explored in two key directions, also visualized in Figure 1:

1. **Uncertainty in explanations:** *How certain and reliable are my explanations?* Quantifying uncertainty in explanations is critical for assessing their reliability and effectiveness. Explanations help to improve models and increase user acceptance. However, the usefulness of these explanations depends on their quality and reliability. By estimating the uncertainty associated with the explanations, we can better assess their trustworthiness and prevent misleading information.
2. **Explainable uncertainty estimation:** *Which input patterns lead to the uncertainty in my prediction?* Providing explainable uncertainty estimates can assist both users and data scientists in comprehending these measures more effectively. For instance, explaining the sources of uncertainty in predictions can help data scientists enhance their models by identifying areas where training data is sparse. Additionally, studies have shown that users of AI systems often have difficulty interpreting numerical information, such as probabilities and standard deviations [5, 7]. Therefore, I believe that explanations of uncertainty can help users understand these measures better by providing reasons behind the uncertainty, making it more interpretable and ultimately ensuring the responsible use of AI.

Both directions offer significant potential to increase the reliability of Machine Learning (ML) models, user trust, and responsible usage of AI. During my PhD, I plan to focus on the emerging research field of explainable uncertainty estimation. Thus, I will only focus on this direction in the following sections.

1.2. The Role of Explainability and Uncertainty for Healthcare Applications

As a consistent case study, I will examine the challenge of nurse scheduling, which involves assigning nurses to shifts. Efficient and needs-based scheduling of nurses is crucial given the growing nurse shortage in many countries which negatively impacts nurse satisfaction, care quality, and patient satisfaction [8]. ML predictions can support the optimization of nurse scheduling by accurately forecasting patient demand and staffing needs, ensuring adequate coverage and reducing overstaffing or understaffing. Despite the high potential of such AI systems, there remains a low adoption of AI systems in clinical practice. One reason for low adoption is the lack of trust and understanding of AI systems among end users in healthcare. Given the high risks associated with healthcare applications and the often inherent life-or-death consequences, ensuring reliability and trustworthiness of AI systems is essential and a key factor for their adoption [3].

Explainability and UQ are key tools that can help improve these properties [5]. While UQ and XAI methods already allow a deeper understanding of an AI model's predictions, simply quantifying the uncertainty does not always give enough information to make a decision based on the prediction. Users may wonder why the uncertainty arises. Furthermore, AI systems in healthcare have different target user groups, such as patients, nurses, doctors, or administrative staff, requiring different communication of predictive uncertainty. For example, in a cancer diagnostic scenario, a doctor may need quick access to numeric confidence scores to make time-sensitive decisions. In contrast, a patient might need comprehensive explanations to understand their diagnosis and the uncertainty behind the prediction [7]. Explanations of uncertainty can be a solution to provide these types of uncertainty representation for different user groups. For data scientists, it is beneficial to know if the uncertainty comes from noisy samples, a distribution shift, or which features or patterns lead to the model's uncertainty [9].

From a technical perspective, the healthcare domain also presents particular challenges. In healthcare, data are often multimodal and include (mostly multivariate) time series. Tasks related to time series, such as classification, forecasting, and regression, remain relatively underexplored in the context of UQ [1, 3, 10]. In addition, XAI methods have to be designed to tackle the challenges of time series data to capture the prediction's dependency on temporal patterns and trends and capture correlations between features. Also, the respective methods must be scalable in order to process large volumes of multimodal health data and be used in practice.

2. Related Work

Previous work on XAI and UQ has largely explored these fields independently, with each aiming to improve the reliability and trustworthiness of AI models. However, the intersection of these fields has only recently begun to attract research interest [4, 5, 6, 11]. Although there exist conceptual papers [5, 9] that emphasize the importance of closing this gap by researching how to make uncertainty estimates explainable, very little work has been done in this area. In the following, I describe identified related work to explain uncertainty estimates.

Table 1
Comparison of related work on explainable uncertainty estimation

Method	Computational Efficiency	Model Agnostic	High-Quality Explanations	Time Series	User Evaluation
Antorán et al. [6]			x		x
Wang et al. [12]			x		
Iversen et al. [4]	x		x		
Watson et al. [11]	x	x	x		
Bley et al. [13]	(x)	x	x	x	
Depeweg et al. [14]	x	x			
Planned Work	x	x	x	x	x

Table 1 provides a comparison of various methods in the field, highlighting their strengths and limitations according to several criteria. Notably, none of the identified methods have been applied to healthcare applications and therefore are not tailored to the specific needs of stakeholders in this setting.

CLUE [6] is a counterfactual approach for explaining uncertainty estimates in differentiable probabilistic models like Bayesian neural networks. It modifies uncertain inputs through a generative model, generating attribution maps by measuring the differences between the original and modified inputs. Although CLUE effectively identifies how to adjust inputs to increase model confidence, training the generative model adds complexity. UA-Backprop [12] uses gradient information to create uncertainty attribution maps in Bayesian deep learning models, highlighting regions contributing to epistemic and aleatoric uncertainty. However, since the method relies on Bayesian deep learning models, it remains rather computationally inefficient. Iversen et al. [4] present a two-step method for explaining aleatoric uncertainty in neural networks by adding a variance output neuron to pre-trained models and applying feature attribution methods to the variance output. Compared to the other described methods, this approach is computationally efficient and simpler to adopt; however, it is not able to capture epistemic uncertainty. Watson et al. [11] theoretically investigated the application of Shapley Values to uncertainty measures and evaluated their method in contexts such as covariate shift and feature selection. While their method is designed to be computationally efficient, calculating Shapley values can still be computationally intensive, especially for large data sets or complex models, limiting its applicability in practice. The approach recently introduced by Bley et al. [13] estimates uncertainty using ensemble-based methods. To explain these estimates, they generate explanations for each ensemble member and calculate the covariance of these explanations. The computational efficiency of their approach depends greatly on the chosen ensemble-based method and the explanation method. Although their approach is applied to time series data, it is not specifically designed for this datatype.

In summary, while these approaches provide promising results in explaining uncertainty estimates, there are several gaps that remain unaddressed: (a) the lack of methods specifically tailored to time series related ML tasks, (b) the need for computationally efficient solutions that can handle complex models without excessive overhead while ensuring high-quality explanations, and (c) the evaluation of these methods with end users and data scientists.

My study aims to address these challenges by developing methods that are both interpretable and computationally efficient, specifically targeting time series data. Additionally, I plan to evaluate my methods with end users and data scientists. Through my research, I seek to contribute to the development of more reliable and transparent AI systems by providing insights into the underlying factors that contribute to uncertainty in ML models.

3. Research Questions, Hypotheses, and Objectives

The goal of my research is to design methods for explaining uncertainty in ML models, focusing on healthcare applications and time series tasks. By developing scalable and computationally efficient methods that provide insights into the sources of uncertainty in AI predictions, I will contribute to advancing the integration of XAI and UQ. Ultimately, my work aims to provide actionable insights that improve the performance of ML models and foster greater trust in AI systems by providing explanations of uncertainties to different stakeholder groups.

In my PhD research project, I aim to address the following research questions:

- R1: How can existing explainability techniques be effectively combined with UQ in ML models to explain the source of uncertainty in ML predictions?
- R2: What are the challenges and potential solutions for incorporating UQ and explanations into ML models trained on time series data?
- R3: What are the most effective and computationally efficient methods for explaining uncertainty estimates in ML predictions?
- R4: How should ML models and UQ/XAI methods be designed to provide actionable insights and explanations for uncertainty in predictions?

These questions focus on designing methods to explain uncertainty estimates. Once I have developed approaches to explain the uncertainty in ML model predictions, I plan to explore the impact of these explanations on users and their benefits for data scientists in training and refining ML models. The following research questions address these areas:

- R5: How can data scientists use explanations of uncertainty to enhance the performance and robustness of ML models?

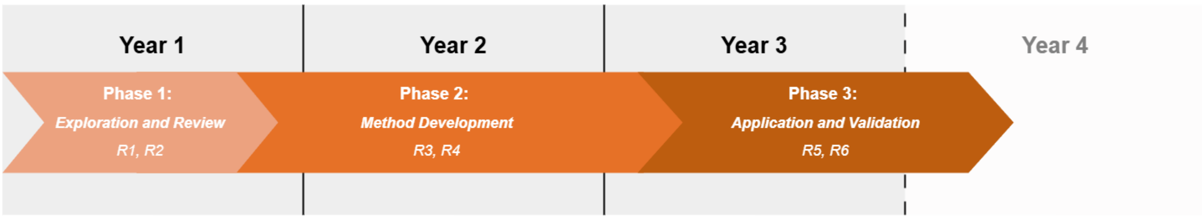

Figure 2: The structure of my PhD research is divided into three phases. The initial phase, lasting approximately 9 months, focuses on exploring and reviewing existing methods to address research questions R1 and R2. The second phase, spanning about 15 months, is dedicated to developing novel methods to explain uncertainties, addressing research questions R3 and R4. The final phase, expected to take around 12 months, involves the validation and application of the developed methods, answering research questions R5 and R6.

- R6: What impact does the communication of explanations of uncertainties have on users of healthcare applications? To what extent does it influence user trust and acceptance?

4. Research Methodology

My doctoral project adopts a quantitative and exploratory research design to explore the intersection of XAI and UQ in ML. As can be seen in Figure 2, my research is structured into three phases, each addressing specific research questions.

Phase 1: *Exploration and Review* The initial phase involves a comprehensive review and experimentation with existing XAI and UQ methods, such as SHAP [15], LIME [16] and Monte Carlo Dropout [17]. This phase aims to evaluate their effectiveness in explaining uncertainty estimates, particularly in time series data. Scalable techniques for UQ such as Monte Carlo Dropout will be prioritized, with Bayesian networks serving as potential alternatives and common techniques such as quantile regression serving as baselines. Furthermore, I will interact with users to identify their needs through workshops and interviews. The insights gained will inform the development of novel methods tailored to enhance the interpretability of uncertainty in ML predictions.

Phase 2: *Method Development* The second phase focuses on developing novel methods to explain uncertainty estimates, tailored specifically to time series-related ML tasks. I plan to explore gradient-based techniques that are architecture-agnostic and do not rely on Bayesian networks, offering a promising direction for achieving computational efficiency. The development process will be iterative, adapting to findings from the initial phase to refine and improve the methods. This approach acknowledges the inherent uncertainty in research results, allowing flexibility and responsiveness to emerging challenges and opportunities.

Phase 3: *Application and Validation* The final phase explores the practical application of uncertainty explanations and evaluates their ability to improve user trust and model performance. This phase involves a comprehensive evaluation and validation process, using both quantitative metrics and qualitative user studies. User studies will be conducted to gather feedback from healthcare professionals, ensuring that the explanations are not only technically sound, but also valuable and interpretable for end users. In addition, I will explore the potential of uncertainty explanations to improve the underlying model. For example, if an explanation reveals sparsity in a specific data region, additional data instances can be strategically introduced to enhance model understanding and confidence. The effectiveness of this data augmentation strategy will be evaluated by comparing uncertainty levels before and after augmentation.

The consistent case study addressing nurse scheduling serves as a real-world example to identify relevant user needs and requirements for the ML models, explainability and uncertainty estimation methods. However, I plan to conduct research that can be generalized to different domains and use cases. To perform the case study, data on patients and nurses' shift plans have been gathered from a university hospital. To complement this, publicly available data sets such as MIMIC IV [18], ETTh1/h2, and ETTm1/m2 [19], and other common time series data sets will be utilized to support time series related ML tasks.

In summary, my research aims to combine theoretical advancements in XAI and UQ with their practical applications in healthcare, thus contributing to academic knowledge and real-world impact.

5. Preliminary Results and Contributions

In my previous research, I have focused on gaining a comprehensive understanding of the healthcare domain, particularly in the context of nurse scheduling. This involved an extensive review and implementation of state-of-the-art models for long-term time series forecasting to predict care capacity and demand. Such predictions are valuable to support nurse scheduling and ensure efficient healthcare delivery through AI. Specifically, I conducted a comparative analysis of several advanced forecasting models, including TSMixer [20] and LightGBM [21] among others. This analysis aimed to evaluate their performance in forecasting nursing staff capacity over a long time horizon. A key aspect of this study was the investigation of whether exogenous variables can enhance the accuracy of these long-term forecasts. We discovered a high increase in performance when including exogenous variables in the model [22]. This adds complexity when designing methods to explain uncertainty estimates, but at the same time offers interesting challenges. The insights gained from this research provide a solid foundation for further exploration, where the high-performing models identified, such as TSMixer and LightGBM, will serve as base models for which I aim to develop methods to explain the uncertainties in their predictions.

To support the development of novel methods to explain uncertainty estimates, I have pre-processed two high-quality data sets obtained from the University Hospital. These high-quality and high-resolution time series data sets will be crucial in training models and evaluating the effectiveness of XAI and UQ methods. The data sets include data with distribution shifts due to the COVID-19 pandemic, providing a unique opportunity to analyze and assess the robustness of the proposed methods.

In addition, I engaged with shift planners, potential end users of AI systems designed to support nurse scheduling. Through these interactions, I gained valuable insights into their specific needs and preferences regarding AI systems. A recurring theme in these discussions was the importance of transparency and understanding the rationale behind AI predictions. This feedback strongly supports the direction of my future work, which will focus on integrating XAI and UQ to address these user concerns. Another important aspect for end users is the reliability of these methods. Since wrong predictions can lead to poor patient care, decreased job satisfaction, and higher absence rates, it is crucial that the ML models are reliable and trustworthy. This also highlights the need for future work on uncertainty estimation and its explainability.

In summary, my preliminary work has laid a strong foundation for advancing the integration of XAI and UQ in healthcare applications, specifically in the context of nurse scheduling.

6. Future Research and Expected Contributions

Given my preliminary work on time series forecasting, my immediate future work will focus on time series forecasting as underlying ML task. However, it is important to emphasize that my research will not be limited to time series forecasting alone. As my research is still in its early stages, there are several key areas that require further investigation to ensure a comprehensive understanding and effective integration of time series forecasting with UQ and XAI. A primary focus will be on the comparison of existing XAI and UQ methods with regard to their suitability for time series forecasting, as well as metrics for their evaluation. This will involve exploring existing methodologies and identifying gaps where novel approaches can be developed. I will explore and implement various UQ methods tailored for time series forecasting. These methods include Monte Carlo Dropout, ensemble techniques, and quantile regression. Furthermore, I plan to add the few methods that already exist to explain uncertainty estimates, described in Section 2. The aim is to establish a robust pipeline that not only incorporates these established methods but also accommodates the development and integration of innovative approaches. This pipeline will serve as a foundation for future research and allow for seamless testing and validation of the combination of UQ and XAI methods as well as novel methods as my research progresses. This approach enables me to quantitatively evaluate and compare existing methods to those that I plan to develop in the future, ensuring a broad and impactful contribution to the field.

Furthermore, I plan to engage with domain experts and end users to ensure that the methods developed are aligned with real-world needs. After having developed methods that accurately explain uncertainty estimates, I plan to investigate how these explanations benefit data scientists in model debugging and performance monitoring, such as detecting distribution shifts. To the best of my knowledge, this has not been investigated in depth yet.

7. Conclusion

In conclusion, the integration of explainability and UQ in ML models holds significant promise for enhancing trust and acceptance, particularly in high-stakes domains such as healthcare. This research aims to bridge the gap between two traditionally separate research fields: Explainable AI and Uncertainty Quantification. My research will explore innovative methods to explain uncertainty estimates, providing actionable insights that can improve decision-making processes and foster transparency. The expected contributions include not only theoretical advancements, but also practical frameworks that can be adopted across various domains, thereby enhancing the overall impact and reach of AI technologies. Ultimately, my work seeks to contribute to the development of more reliable and transparent AI systems, thereby advancing the field of ML and its application in critical real-world scenarios.

Acknowledgments

I express my sincere gratitude to my scientific supervisor, Jan-Philipp Steghöfer, for his invaluable guidance and support throughout my research. This research is sponsored by the Bavarian State Ministry of Economic Affairs, Regional Development and Energy under grant number 41-6618c/570/2-LSM-2203-0010.

Declaration on Generative AI

During the preparation of this work, the author used ChatGPT-4 in order to: Grammar and spelling check, Paraphrase and reword. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya, et al., A review of uncertainty quantification in deep learning: Techniques, applications and challenges, Information fusion 76 (2021) 243–297.

[2] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al., Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai, Information fusion 58 (2020) 82–115.

[3] S. Seoni, V. Jahmunah, M. Salvi, P. D. Barua, F. Molinari, U. R. Acharya, Application of uncertainty quantification to artificial intelligence in healthcare: A review of last decade (2013–2023), Computers in Biology and Medicine (2023) 107441.

[4] P. Iversen, S. Witzke, K. Baum, B. Y. Renard, Identifying drivers of predictive uncertainty using variance feature attribution, arXiv preprint arXiv:2312.07252 (2023).

[5] R. Tomsett, A. Preece, D. Braines, F. Cerutti, S. Chakraborty, M. Srivastava, G. Pearson, L. Kaplan, Rapid trust calibration through interpretable and uncertainty-aware ai, Patterns 1 (2020) 100049.
[6] J. Antorán, U. Bhatt, T. Adel, A. Weller, J. M. Hernández-Lobato, Getting a CLUE: A method for explaining uncertainty estimates, in: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021, OpenReview.net, 2021.
[7] U. Bhatt, J. Antorán, Y. Zhang, Q. V. Liao, P. Sattigeri, R. Fogliato, G. Melançon, R. Krishnan, J. Stanley, O. Tickoo, et al., Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty, in: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, 2021, pp. 401–413.
[8] M. Peters, Time to solve persistent, pernicious and widespread nursing workforce shortages, International nursing review 70 (2023) 247–253.
[9] M. Salvi, S. Seoni, A. Campagner, A. Gertych, U. R. Acharya, F. Molinari, F. Cabitza, Explainability and uncertainty: Two sides of the same coin for enhancing the interpretability of deep learning models in healthcare, International Journal of Medical Informatics (2025) 105846.
[10] K. Stankeviciute, A. M Alaa, M. van der Schaar, Conformal time-series forecasting, Advances in neural information processing systems 34 (2021) 6216–6228.
[11] D. Watson, J. O'Hara, N. Tax, R. Mudd, I. Guy, Explaining predictive uncertainty with information theoretic shapley values, Advances in Neural Information Processing Systems 36 (2023) 7330–7350.
[12] H. Wang, D. Joshi, S. Wang, Q. Ji, Gradient-based uncertainty attribution for explainable bayesian deep learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 12044–12053.
[13] F. Bley, S. Lapuschkin, W. Samek, G. Montavon, Explaining predictive uncertainty by exposing second-order effects, Pattern Recognition 160 (2025) 111171. doi:`https://doi.org/10.1016/j.patcog.2024.111171`.
[14] S. Depeweg, J. M. Hernández-Lobato, S. Udluft, T. A. Runkler, Sensitivity analysis for predictive uncertainty, in: 26th European Symposium on Artificial Neural Networks, ESANN 2018, Bruges, Belgium, April 25-27, 2018, 2018.
[15] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Advances in neural information processing systems 30 (2017).
[16] M. T. Ribeiro, S. Singh, C. Guestrin, " why should i trust you?" explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135–1144.
[17] Y. Gal, Z. Ghahramani, Dropout as a bayesian approximation: Representing model uncertainty in deep learning, in: international conference on machine learning, PMLR, 2016, pp. 1050–1059.
[18] A. Johnson, L. Bulgarelli, T. Pollard, S. Horng, L. A. Celi, R. Mark, MIMIC-IV (2022). doi:`10.13026/RRGF-XW32`.
[19] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, W. Zhang, Informer: Beyond efficient transformer for long sequence time-series forecasting, in: Proceedings of the AAAI conference on artificial intelligence, volume 35, 2021, pp. 11106–11115.
[20] S.-A. Chen, C.-L. Li, S. O. Arik, N. C. Yoder, T. Pfister, TSMixer: An all-MLP architecture for time series forecast-ing, Transactions on Machine Learning Research (2023).
[21] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, Lightgbm: A highly efficient gradient boosting decision tree, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017.
[22] E. Schiller, S. Müller, K. Ebertsch, J.-P. Steghöfer, No data left behind: Exogenous variables in long-term forecasting of nursing staff capacity, in: 2024 IEEE 11th International Conference on Data Science and Advanced Analytics (DSAA), IEEE, 2024, pp. 1–10.