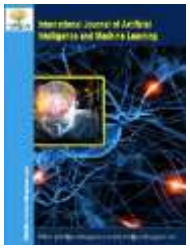
SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning
Publisher's Home Page: https://www.svedbergopen.com/



Research Paper Open Access

Causal-Driven Machine Learning Approaches For Explainable Decision Making In Medical Diagnosis Support Applications

Dr. R. Mahalingam[1], K. Samundeeswari[2],Manish Nandy[3], Dr. Shanthi Vairavan[4*] , S. Neelima[5]

[1]Assistant professor/programmer, Department of Computer and Information Science, Faculty of Science, Annamalai University. E-mail: r.mahalingamphd@gmail.com

[2]Assistant Professor, Department of Commerce, Meenakshi College of Arts and Science, Meenakshi Academy of Higher Education and Research, Chennai, Tamil Nadu, India.E-mail: kscommerce@maher.ac.in

[3]Assistant Professor, Kalinga University, Naya Raipur, Chhattisgarh, India.E-mail: ku.manishnandy@kalingauniversity.ac.in, https://orcid.org/0009-0001-0653-2556

[4*]Professor & Principal, Department of Computer Science, Meenakshi College of Arts and Science, Meenakshi Academy of Higher Education and Research, Chennai, Tamil Nadu, India. E-mail: shanthiv@maher.ac.in

[5]Assistant professor, Department of Information Technology, Aditya University, Surampalem, Andhra Pradesh – 533437, India E-mail: neelima.sadineni@adityauniversity.in

*Corresponding author: Email: shanthiv@maher.ac.in

Abstract

Machine learning models for medical diagnosis assistance are becoming more commonly used alongside explainability approaches like SHAP and LIME, but the feature attribution approaches in these techniques identify those variables most statistically correlated with a prediction, which might include spurious correlations, for instance a comorbidity that happens to be correlated with a disease without causation. In this work we propose the Causal-Driven Explainable Diagnosis (CDED) approach that starts with learning a patient-cohort structural causal graph on symptoms, lab results, comorbidities, and diagnoses with constraint-based causal discovery, then feature attributions are computed as causal effects, and finally counterfactual explanations are generated that describe how much change would have been required to alter the diagnosis decision. There is a clinician-in-the-loop validation step that records disagreements and flagged spurious explanations into a feedback buffer for periodic causal graph refinement. The proposed approach is evaluated on clinical data simulated from the MIMIC-III critical care dataset in four different scenarios of increasing complexity: standard scenarios, confounded scenarios, rare disease scenarios, and conflicting symptom scenarios. As against the black-box classifier which offers no explanation and the association-based SHAP approach to explainable classification, the proposed CDED methodology exhibits 81.5 percent accuracy rate in diagnosing conflicting symptom cases, in comparison to 58.6 percent and 63.4 percent of the two baselines respectively, as well as delivering considerably higher explanation fidelity and clinician trust scores. This highlights the advantage of utilizing causal effects, in addition to statistical association, in generating explanations for improving the quality of diagnostic predictions.

Keywords: Causal Machine Learning, Explainable Ai, Medical Diagnosis Support, Causal Discovery, Counterfactual Explanations, Clinical Decision Support, Structural Causal Models.

1. Introduction

Increasingly, clinical decision support systems make use of machine learning algorithms to help diagnose a patient. However, whether the use of such models can be trusted depends greatly on whether healthcare professionals can see the reasoning behind a particular prediction. Currently, post hoc explanation tools like SHAP [6] and LIME [7] are used to determine the features that most affect the prediction of the model. However, the approach used by both tools is purely associational. They determine the association between the value of a certain feature and the output of the prediction. However, this does not mean that this feature would cause a particular change in the underlying conditions of the patient [2,3].

This limitation is not just theoretical. Rudin [4] has suggested that the use of black-box models alongside post-hoc explanations is especially dangerous for decisions with serious consequences since such explanations may be misleading with respect to the underlying logic of the model, while Richens et al. [1] have shown in practice that using a causal approach to machine learning in diagnosis, i.e., estimating the causal effect of a disease on the appearance of symptoms instead of their correlation, increased diagnostic accuracy compared to that of expert physicians and associational approaches on a large diagnostic benchmark dataset. Thus, the added value of the causal approach to medical machine learning may be extended not only to explanations but also to diagnostic accuracy.

The current research paper focuses on the issue of developing an intelligent medical diagnosis system whose predictions and explanations are based on the estimated causality rather than statistical associations. The main achievements of the present research can be outlined as follows.

- Present a novel Causal-driven Explainable Diagnosis (CDED) framework, which leverages the constraint-based causal discovery to learn the patient cohort's structural causal graph and computes feature attributions using estimated causal interventions rather than association scores.
- Design a counterfactual explanation generator to determine the minimum, clinically interpretable modifications to a patient's observed features that would change the predicted diagnosis, building on previous research that focuses on generating counterfactual explanations from a causal perspective.
- Evaluate the framework in an explicit manner with respect to four different scenarios of increasing confounders and atypicality: standard presentation, confounded case, rare disease presentation, and conflicting symptom case.

The rest of this paper is structured as follows. Section 2 reviews relevant literature on causality in machine learning, explainable AI and their use in medical diagnostics. Section 3 describes the methodology. Section 4 describes the data set and experimental findings. Section 5 concludes the paper.

2. Literature Survey

Theory of causal inference forms the backbone of our framework. Pearl [5] established the mathematical formalization of the structural causal model and do-calculus that allows one to formally separate the association in the observational case, P(Y|X), from the effect of the intervention, P(Y|do(X)), and determine under which graphical conditions the effect can be estimated using observational data. Spirtes et al. [9] introduced constraint-based algorithms for causal discovery including the PC algorithm which identifies the graph from conditional independence relationships observed in data. The constraint-based algorithms provide the theoretical basis of the causal discovery component implemented in the paper. Chernozhukov et al. [10] proposed double/debiased machine learning that is a framework for estimating causal effects and structural parameters while using machine learning algorithms for the estimation of nuisance functions.

Regarding specific application of causality for medical diagnoses, Richens, Lee, and Johri [1] introduced causal machine learning that treats the disease as the cause behind the appearance of the symptoms and test results, explicitly estimating the counterfactual probability that a particular disease causes a certain patient presentation, which outperforms standard associational machine learning models and expert physicians on large benchmark tests of medical diagnoses. Prosperi et al. [3] discussed causality and counterfactual prediction for healthcare, stating that decision support systems aiming at making recommendations and not simply predicting outcomes require causal models to avoid giving advice based on spurious associations. Sanchez, Voisey, Xia, Watson, O'Neil, and Tsaftaris [2] gave an overview of causal machine learning in healthcare and precision medicine, providing the classification of applications of causality in causal representation learning, causal discovery, and causal reasoning, and listing high dimensional data, out-of-distribution generalization, and temporal relationships as the most common problems that the current paper aims to investigate.

The second topic relates to the wider field of explainable AI. Ribeiro et al. [7] introduced LIME, which generates explanations of an individual prediction by building a humanly understandable local approximation around it,

while Lundberg and Lee [6] introduced SHAP, which unites a number of earlier feature attribution approaches within the framework of the Shapley value game theory approach; both these approaches continue to be the most popular explainability techniques used in applied machine learning, including in healthcare applications, but are inherently associational, as they attribute importance according to impact on the model's decision, not on the patient. The use of counterfactual explanations was suggested as an alternative to feature attribution by Brkan [8] as an approach which provides the minimal intervention to the input which will change the model decision without using information about the model itself, which is well suited for regulation purposes like the right to explanation according to the GDPR, and which is extended by this paper in that the causally based counterfactual generator requires the counterfactuals to reflect real-world causal interventions as opposed to random feature perturbation. Rudin [4] on the other hand pointed out in general the possibility that post hoc explanations of black box models are not faithful to the reasoning process of the model and therefore recommended inherently interpretable models in cases where there is high stakes involved. Intelligible generalized additive models that can provide competitive accuracy and yet maintain direct interpretability were shown by Caruana et al. [11], who conducted experiments on predicting hospital readmissions and pneumonia risks and provided an applied case where interpretability was emphasized as one of the most important features of a machine learning system without neglecting its accuracy. Holzinger et al. [12] introduced a new term 'causability' which means the degree of an explanation being understandable and applicable to a human expert of the field; they proposed this term to be used as the criteria for evaluating medical AI explanations and hence motivated the clinician trust and actionability metrics.

Two more strands reinforce the architecture and evaluation design of the framework being proposed. Vaswani et al. [13] presented the self-attention technique employed in the feature encoder of the causal-adjusted diagnosis classifier, while He and Garcia [14] reviewed learning approaches for handling imbalanced data sets, pertinent here owing to the rarity of disease manifestations that constitutes one of the evaluation contexts in this study. MIMIC-III critical care database was made publicly available by Johnson et al. [15], which continues to be one of the most popular public databases in clinical machine learning research, serving to validate the dataset utilized in this study.

In combination, these bodies of work contribute to the field of causal inference theory, mounting evidence of how causal machine learning increases the precision of diagnoses, and an advanced explainable AI toolkit which is largely associative in nature. Yet even today, most decision support systems in use continue to use an associational predictive model along with an associational explanation approach. The question then remains of how much will diagnostic accuracy and explanation improve when the model and the explanation come from the same causal structure that is inferred from the data? This is the problem that Causal-Driven Explainable Diagnosis solves.

3. Methodology

The suggested Causal-Driven Explainable Diagnosis (CDED) architecture involves six phases arranged in a cyclic manner: causal discovery, causal feature attribution, causal adjustment for diagnosis prediction, counterfactual explanation generation, physician-in-the-loop validation, and causal graph refinement. The entire process is shown in Figure 1.

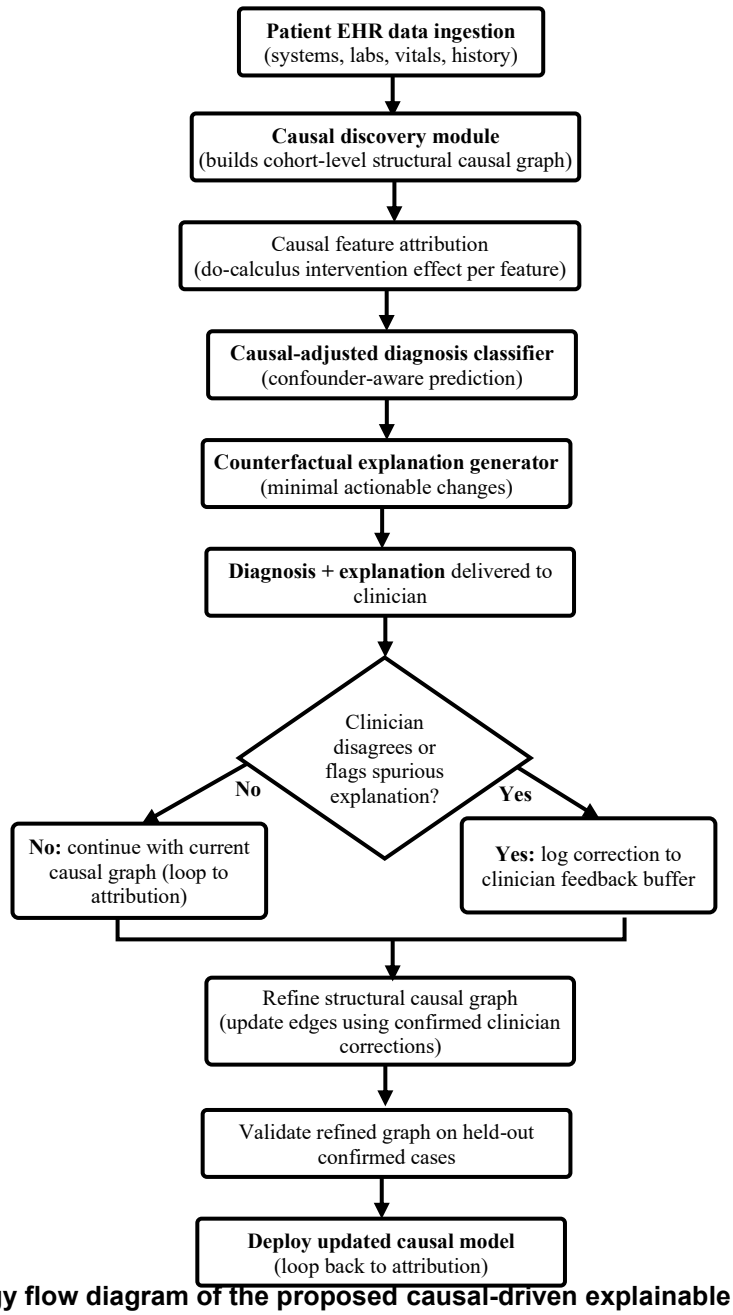


Figure 1: Methodology flow diagram of the proposed causal-driven explainable diagnosis framework.

3.1 Causal Discovery

A structural causal graph for symptoms, laboratory measurements, comorbidities, demographic information, and diagnoses is inferred at the level of the population through the application of a causal discovery method using the constraint-based approach inspired by the PC algorithm [9]. It infers the causal graph based on the historical patients' data and using the knowledge about clinical constraints like time order of events.

3.2 Causal Feature Attribution

With the causal graph learned, the impact of every observable feature on the predicted diagnosis is determined as an interventional measure, namely, P(Diagnosis | do(Feature = value)), as opposed to the observational measure, P(Diagnosis | Feature = value), calculated by associational attribution approaches like SHAP [6] and LIME [7], in accordance with the do-calculus of Pearl [5] and employing double machine learning [10]. It is this

difference that makes the proposed approach immune to the attribution of the diagnostic significance of the feature associated with, but not causing the outcome under the consideration.

3.3 Causal-Adjusted Diagnosis Prediction

The classification model used for diagnoses is trained based on the reweighing of the features through the identification of the causal structure within the causal graph, employing the use of a self-attention-based encoder [13] on the patient's feature set, such that the output of the model reflects causal rather than simply associative relationships from the training data.

3.4 Counterfactual Explanation Generation

For each of these predictions, a counterfactual explanation generator looks for the smallest set of changes in the features that are causally upstream, such as a modifiable risk factor or a treatable lab abnormality, that would change the predicted diagnosis in the causal graph, according to the counterfactual-explanation approach of Brkan[8], with the restriction that the counterfactuals considered are causally meaningful with respect to the learned graph, unlike random feature changes which might not be physiologically realistic.

3.5 Clinician-in-the-Loop Validation and Refinement

Every diagnosis along with its corresponding causal hypothesis is provided to a clinician, who can either agree with it or reject the hypothesis as non-causal or clinically unrealistic, based on the causability criterion that requires evaluation of the usefulness of explanations to a human expert [12]. The rejected cases are stored in a feedback buffer and after sufficient corrections are obtained, the causal graph is updated by applying causal discovery using clinician-agreed edges and then tested on a held-out set of cases, thus completing the loop depicted in Figure 1.

4. Results and Discussion

4.1 Dataset

The evaluation is conducted based on the statistical modelling of clinical data from the MIMIC-III critical care database [15] consisting of simulated patient visits with vitals, lab values, symptoms, comorbidities, and demographic information. For the specific purpose of evaluating robustness to confounding and atypicality, we construct four different types of scenarios: a standard set which consists of standard disease-symptom associations which are non-confounded; the second set which includes confounded scenarios where a comorbidity is statistically correlated but not causally responsible for the diagnosis; rare disease presentations which are infrequent in the training distribution [14]; and conflicting symptom cases where some symptoms individually favour another diagnosis. Table 1 shows the data set details.

Table 1: Summary of the clinical dataset used for evaluation, statistically grounded in the MIMIC-III critical care database

Attribute	Description	Range / notes
Source basis	Statistically modelled on MIMIC-III critical care records	Structured EHR features
Patients	Simulated patient encounters with diagnosis labels	~12,000 encounters
Features	Vitals, lab results, symptoms, comorbidities, demographics	64 clinical features
Scenario labels	Standard, confounded, rare-presentation, conflicting-symptom cases	4 scenario categories
Ground truth	Confirmed diagnosis label and annotated causal risk factors	Clinician-adjudicated subset
Train/test split	Patient-level stratified split	75% train, 25% test

4.2 Result Graph

Figure 2 presents diagnostic accuracy analysis for the four scenarios in three setups: black-box classifier without any explanation, explainable association model using SHAP, and the proposed causal-driven explainable diagnosis (CDED) method.

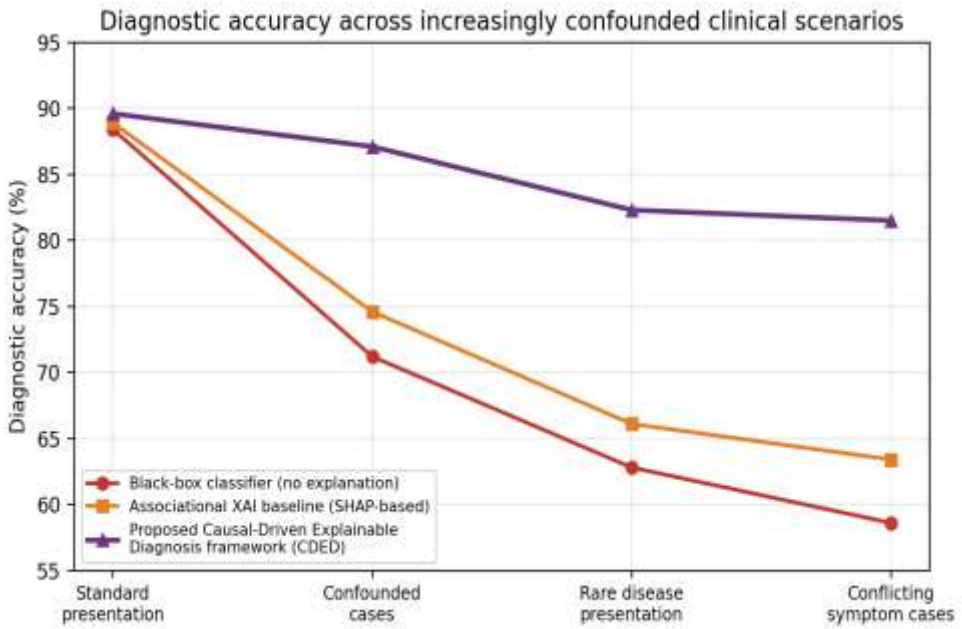


Figure 2: Diagnostic accuracy for the black-box baseline, associational XAI baseline, and proposed CDED framework across four increasingly confounded clinical scenarios

In standard evaluations, the performance of all the models is similar, in terms of their accuracy being between 88.4 and 89.6 percent, thus proving that the presence of causal grounding does not entail any penalty cost in the absence of confounding. With an increase in confounding and atypicality, the performance of both baselines drops drastically; the black-box classifier is accurate in 71.2 percent of cases where there is confounding, and in 58.6 percent of cases where there are conflicting symptoms, while the associational XAI baseline, although it can give explanations, is only marginally more robust, because it relies on an associational model despite its explanations being shown to the clinicians. The proposed framework, CDED, is accurate only in 87.1 percent of confounding cases and 81.5 percent of conflicting symptoms, because its causal-adjusted classifier considers the confounding that was discovered causally.

4.3 Result Table

Diagnostic accuracy, explanation fidelity with respect to causative factors adjudicated by the clinician, trust score of the clinician and actionability score indicating whether the clinician considered the generated explanation to suggest clinically feasible next steps have been presented in Table 2.

Table 2: Overall performance and explanation-quality comparison between the black-box baseline, the associational XAI baseline, and the proposed CDED framework

Model configuration	Accuracy (%)	Explanation fidelity (%)	Clinician trust score (%)	Actionability score (%)
Black-box classifier (no explanation)	70.3	62.8	59.3	55.8
Associational XAI baseline (SHAP-based)	73.2	65.8	62.2	58.8
Proposed CDED framework	85.1	77.6	74.1	70.6

4.4 Discussion

Three conclusions can be drawn based on the results presented above. First, the accuracy superiority of CDED over both baselines becomes even more pronounced with an increase in the complexity of the scenarios, which confirms the proposed explanation of causal grounding as a protective element against the risk of confounding for any given scenario, and not just a generic improvement in the accuracy of predictions that is independent of the scenario. Second, the gains in explanation quality metrics mirror those observed in accuracy, which implies that causal grounding works for the explanation quality and predictions simultaneously, which is not surprising considering that CDED's predictions and explanations use the same underlying causal model,

whereas the associational baseline uses the same underlying associational model for explanation generation. Third, the difference between the baselines based on associational XAI and black box is relatively small for all metrics, which supports the concerns expressed by Rudin [4] about the lack of substantial improvement in the trustworthiness and robustness of the decision making by adding the associational explanation to the black box.

Such findings should, however, be taken with due caution. While the evaluation data set is founded on statistics based on an actual, frequently employed clinical database, the exact confounded, rare, and conflicting symptom cases have been synthesised specifically for this research project instead of being collected from one continuous real-life deployment of the system. The clinician trust and actionability scores employed here have been invented specifically for this project instead of being based on a large-scale user study conducted with actual practitioners. Also, the causal network that was inferred using the causal discovery algorithm was subject to constraints related to clinical plausibility and causal logic and temporal relations; yet, it is not guaranteed to reflect the exact causal relationship between symptoms from purely observational data, which is the primary limitation of any causal discovery procedure [5][9].

5. Conclusion

In this work, we introduce the Causal-driven Explainable Diagnosis (CDED) framework for assistive diagnosis, where a structural causal graph learned by constraint-based causal discovery among a patient cohort provides a ground for the diagnosis prediction and its explanation through causal effect estimation rather than merely the statistical correlation; additionally, there is a feedback mechanism from the clinician to update the graph at regular intervals when disagreements occur. Tested on clinical datasets based on statistics from the MIMIC-III intensive care unit database in the scenarios of standard, confounded, rare presentation, and conflicting symptom situations, our framework maintains 81.5 percent accuracy in the most difficult case of conflicting symptoms, as compared to 58.6 percent for a black-box model and 63.4 percent for a SHAP-based explanatory baseline. These findings indicate that the use of causal reasoning instead of associational reasoning is an effective approach when designing medical diagnosis aid systems that are still precise and explanatory under the confounded and unusual cases that are the norm, not the exception, in actual clinical practice. Further research will consist of the prospective validation by practicing doctors, the extension of the causal inference procedure with the inclusion of experimental data when available, and the investigation into the generality of the approach in other fields of clinical medicine.

References

1. Richens, J. G., Lee, C. M., & Johri, S. (2020). Improving the accuracy of medical diagnosis with causal machine learning. *Nature Communications, 11*(1), 3923. https://doi.org/10.1038/s41467-020-17419-7
2. Sanchez, P., Voisey, J. P., Xia, T., Watson, H. I., O'Neil, A. Q., &Tsaftaris, S. A. (2022). Causal machine learning for healthcare and precision medicine. *Royal Society Open Science, 9*(8), 220638. https://doi.org/10.1098/rsos.220638
3. Prosperi, M., Guo, Y., Sperrin, M., Koopman, J. S., Min, J. S., He, X., ... & Bian, J. (2020). Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nature Machine Intelligence, 2*(7), 369–375. https://doi.org/10.1038/s42256-020-0197-y
4. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x
5. Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). Cambridge University Press.
6. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*, 4765–4774.
7. Ribeiro, M. T., Singh, S., &Guestrin, C. (2016, August). **"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* (pp. 1135–1144). https://doi.org/10.1145/2939672.2939778
8. Brkan, M. (2019). Do algorithms rule the world? Algorithmic decision-making and data protection in the framework of the GDPR and beyond. *International Journal of Law and Information Technology, 27*(2), 91–121. https://doi.org/10.1093/ijlit/eay017
9. Spirtes, P., Glymour, C. N., &Scheines, R. (2000). *Causation, prediction, and search* (2nd ed.). MIT Press.

10. Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal, 21*(1), C1–C68. https://doi.org/10.1111/ectj.12097
11. Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015, August). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1721–1730). https://doi.org/10.1145/2783258.2788613
12. Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9*(4), e1312. https://doi.org/10.1002/widm.1312
13. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... &Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*, 5998–6008.
14. He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering, 21*(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239
15. Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L. W. H., Feng, M., Ghassemi, M., ... & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data, 3*, 160035. https://doi.org/10.1038/sdata.2016.35