The unreasonable effectiveness of dissecting models: A case for trustworthy and efficient biomedical time-series models

Christodoulos Kechris[*]

EPFL, Lausanne, Switzerland
ORCID (Christodoulos Kechris): https://orcid.org/0009-0008-2653-447X

Abstract. Black-box machine learning in clinical time series poses trust and efficiency hurdles. We propose dissection as a key strategy to tackle this limitation. We apply our framework through all four phases of a model's lifecycle: data creation, model training, validation and deployment. Our diagnostic toolset exposes and eliminates spurious shortcuts. It also introduces inference optimisations inspired by signal-processing principles. These advances pave the way for trustworthy and efficient clinical machine learning.

1 Introduction

Recent trends in Machine Learning (ML) follow the big data, big models paradigm: training on massive datasets with ever larger architectures yields unprecedented performance and generalisability. Generative text and vision models illustrate this trend [3, 20]. On the wake of this success, large models, paired with continuous monitoring from wearable devices, promise the next technological breakthroughs in medicine [18, 15] with potential innovations such as biomarker detection [19], patient monitoring [9] and personalised medicine [22]. This paradigm stands at odds with the classical signal-processing approach based on mechanistic insight.

Central to our contribution is **dissection**, which we define as a phase-wise framework for systematically probing and decomposing an ML pipeline into its data, model, validation and deployment level mechanisms. This decomposition aims at exposing spurious correlations and shortcut risks and uncovering interpretability and efficiency gains. In my research, we ask the main research question:

Main Research Question. *How can model dissection be leveraged to build trustworthy and efficient biomedical time-series models?*

In what follows, we explore this question across four stages of a model's lifecycle: dataset creation (Section 2), model training (Section 3), validation (Section 4) and deployment (Section 5).

2 Data: Is knowledge-free data collection/processing trustworthy?

Many biomedical signals, such as electroencephalography [2] or electrocardiography [5], are well studied and their properties are extensively described in medical and engineering literature. *But if large scale data and models can automatically uncover latent medical information, why is this prior knowledge necessary?*

Research Question 2.1. *Can knowledge-free data collection and blind feature extraction reliably uncover biomedical signal biomarkers?*

To answer this question, we study a field in which black-box ML is the main tool for knowledge discovery: vibration arthrography [1]. The premise is that pathologic knees present structural abnormalities, creating acoustic emissions; machine learning models can identify these abnormal audio signatures, diagnosing the underlying pathology. From a medical perspective, little prior knowledge exists - or even if such an audio-pathological link even exists [6]. Modern works depend on automatically-extracted features and black-box models [21]. The main supporting argument: *The model's high diagnostic classification accuracy means unhealthy knees indeed produce "pathological" sounds.*

Instead of blind feature extraction and model training, we propose a reality-centric thorough investigation of the audio signal [12]. We argue that experimental results should be interpreted under a causal framework, linking the knee mechanism to a vibration generation process. To the best of our knowledge, this is the first ever result reproduction study in this field. We uncover model performance inflation due to external information introduced by the experimental setup: pathological knees do not present any detectable audio differences compared to healthy knees. Removing external information diagnostic (binary) accuracy dropped to 51.28% from reported 80.6% [21].

Remark 2.1. *Without prior signal analysis, models can learn experimental artefacts and shortcuts rather than pathology; robust biomarker discovery must begin with causal, domain-grounded data exploration.*

3 Model: Is knowledge-free model building trustworthy?

If prior knowledge is crucial for data collection and processing, is it also relevant when building models? After all, modern ML pipelines only require input-output pairs for training.

Research Question 3.1. *How does embedding domain knowledge into model architectures affect robustness and fidelity?*

[*] Corresponding Author. Email: christodoulos.kechris@epfl.ch

In this use case, we study an application with rich prior information: extracting heart rate (HR) from photoplethysmography (PPG) signals acquired from smartwatches. The heart rate component in PPG is comprised of two harmonics at the heart rate and its double frequency [4]. Additionally, hand motions introduce artefacts, interfering with the heart rate component [16]. In extreme cases, the heart rate information may be damaged beyond the ability of recovery. Deep learning models extracting heart rate are trained on a generic inference loss, e.g. Mean Absolute Error, matching the model's output to the ground truth values. No motion artefact separation task is explicitly given.

We propose Knowledge Informed Deep-learning (KID)-PPG [11], a deep model with guided training incorporating prior PPG knowledge. Although state-of-the-art models excel in extracting HR under minor interference, they cannot separate motion artefacts from heart rate. Importantly, they do not seem to learn the required motion-separation task. To the best of our knowledge, this is the first study investigating these limitations. Integrating an additional explicit motion artefact removal step reduces susceptibility to motion artefacts, reducing HR error by $\sim 35\%$. Furthermore, guiding the model's uncertainty via prior knowledge based data augmentation leads to more robust understanding of whether heart rate can be retrieved or not - $\sim 37\%$ decrease in negative log likelihood.

Remark 3.1. *KID-PPG demonstrates that explicit source separation and uncertainty guidance produce models that not only isolate the heart component but also know when they cannot do so.*

Future work. Our work demonstrates failure-modes of deep models, trained without any prior knowledge. But if deep networks are universal function approximations, why and when do these modes appear? What **dissection** tools do we need to answer this question. Such investigation could uncover failure modes beyond the specific application, providing insights on when models fail to generalise.

4 Validation: Can knowledge-informed evaluation enhance trustworthiness?

The use-cases of Sections 2 and 3 demonstrate that, out-of-context, small loss on a testing subset does not present a full picture of the data or the model behaviour in the real-world. Then,

Research Question 4.1. *How can we build pipelines to evaluate generalisability, augmented with meaningful quantitative metrics beyond accuracy?*

For this stage we have focused on EEG-based epileptic seizure detection. We introduce szCore [7], a seizure detection benchmark feature a private evaluation dataset, prohibiting researchers from overfitting their models on the specific characteristics of the dataset - simulating real-world deployment situations. Notably, szCore reveals that state of the art models failed to generalise (F1 43%) - lower than self-reported performance (F1 reaching $\sim 90\%$) found in the literature [17].

Furthermore, we explore saliency maps as an accuracy supplement. Such explainability methods have been proposed as a way of validating models on medical imaging applications, often revealing shortcut-learning phenomena [8]. Although these methods are useful for image inputs, their use in time-series models is limited: they can highlight individual important pixels, but they cannot decompose a signal into semantically meaningful frequency or morphological components. To tackle this limitation, we have proposed Cross-domain Integrated Gradients (IG) [13], attributing model outputs to semantically meaningful transformations of the original time-domain input. Prior information is integrated by choosing a suitable signal transformation, e.g., the Independent Component Analysis for EEG. The method then provides relevance scores for each component.

Remark 4.1. *Cross-domain IG bridges prior-knowledge-based signal transforms and attributions, providing practitioners with semantically meaningful insights and rich evaluation information.*

Future work. We proposed Cross-domain IG as a post-hoc analysis method - how can we integrate it in a large-scale validation benchmark like szCore? Importantly, can it be elevated to a real-time reporting tool for clinicians? With Cross-domain IG inevitably increasing the computational complexity during inference, are there optimisations to ensure these insights can be produced on the fly in deployed systems?

5 Deployment: Gaining computational efficiency

ML models, and especially deep ones, are often treated as *black-boxes*, without any control or understanding of their inner mechanisms. Apart from trustworthiness, this also limits available strategies for optimising computational efficiency. We exploit our study of the models' inner workings to propose strategies for computational efficiency gains. We ground our approach in a signal-processing based network analysis.

Research Question 5.1. *Can insights from model dissections unlock gains in model inference?*

We introduce StreamiNNC [14], a streaming inference optimisation approach for Convolutional neural networks (CNN). We take advantage of convolution's inherent translation invariance to reuse overlapping windows and reduce per-step complexity. Crucially, we perform a theoretical analysis to provide approximation error upper bounds for non-translation-invariant layers, e.g. pooling. We base these bounds on our analysis of the ReLU activation [10], which gives an exact Fourier-domain characterisation of the activation function: it passes all input frequencies, injects a controllable DC component and additional harmonics. Our StreamiNNC approach enables linear, w.r.t. overlap, reductions in operations without compromising accuracy - e.g. 2.03% NRMSE compared to full inference.

Remark 5.1. *By exploiting convolution's shift-invariance and ReLU's frequency properties, StreamiNNC reduces redundant work, critical for real-time streaming biomedical models.*

Future work. Translation invariance is an inherent convolution property. What other intermediate activation properties can we exploit for further computational optimisations? Can we guide the network to form properties which would reduce computations without damaging model accuracy or expressivity?

6 Conclusions

We have introduced **dissection** as an approach to study the inner mechanisms of ML models throughout four essential stages, demonstrating gains in robustness, explainability and computational efficiency. With ever-growing black-box models deployed in critical applications like healthcare, such analytical methodologies are crucial to maintain trust and capability of deployment. A natural overarching open question now surrounds our findings: *What is the limit to which dissection can open the ML black-box?*

References

[1] S. C. Abbott and M. D. Cole. Vibration arthrometry: a critical review. *Critical Reviews™ in Biomedical Engineering*, 41(3), 2013.

[2] A. Biasiucci, B. Franceschiello, and M. M. Murray. Electroencephalography. *Current Biology*, 29(3):R80–R85, 2019.

[3] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[4] J. Cho, Y. Sung, K. Shin, D. Jung, Y. Kim, and N. Kim. A preliminary study on photoplethysmogram (ppg) signal analysis for reduction of motion artifact in frequency domain. In *2012 IEEE-EMBS Conference on Biomedical Engineering and Sciences*, pages 28–33. IEEE, 2012.

[5] M. B. Conover. *Understanding electrocardiography*. Elsevier Health Sciences, 2002.

[6] J. L. Couch, M. G. King, D. D. O. Silva, J. L. Whittaker, A. M. Bruder, F. Serighelli, S. Kaplan, and A. G. Culvenor. Noisy knees-knee crepitus prevalence and association with structural pathology: a systematic review and meta-analysis. *British journal of sports medicine*, 59(2): 126–132, 2025.

[7] J. Dan, A. Shahbazinia, C. Kechris, and D. Atienza. Szcore as a benchmark: report from the seizure detection challenge at the 2025 ai in epilepsy and neurological disorders conference. *arXiv preprint arXiv:2505.18191*, 2025.

[8] A. J. DeGrave, J. D. Janizek, and S.-I. Lee. Ai for radiographic covid-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7):610–619, 2021.

[9] Z. Jeddi and A. Bohr. Remote patient monitoring using artificial intelligence. In *Artificial intelligence in healthcare*, pages 203–234. Elsevier, 2020.

[10] C. Kechris, J. Dan, J. Miranda, and D. Atienza. Dc is all you need: describing relu from a signal processing standpoint. *arXiv preprint arXiv:2407.16556*, 2024.

[11] C. Kechris, J. Dan, J. Miranda, and D. Atienza. Kid-ppg: Knowledge informed deep learning for extracting heart rate from a smartwatch. *IEEE Transactions on Biomedical Engineering*, 2024.

[12] C. Kechris, J. Thevenot, T. Teijeiro, V. A. Stadelmann, N. A. Maffiuletti, and D. Atienza. Acoustical features as knee health biomarkers: A critical analysis. *Artificial Intelligence in Medicine*, 158:103013, 2024.

[13] C. Kechris, J. Dan, and D. Atienza. Time series saliency maps: Explaining models across multiple domains. *arXiv preprint arXiv:2505.13100*, 2025.

[14] C. Kechris, J. Dan, J. Miranda, and D. Atienza. Don't think it twice: Exploit shift invariance for efficient online streaming inference of cnns. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 17805–17813, 2025.

[15] W. Khan, S. Leem, K. B. See, J. K. Wong, S. Zhang, and R. Fang. A comprehensive survey of foundation models in medicine. *IEEE Reviews in Biomedical Engineering*, 2025.

[16] H. Lee, H. Chung, H. Ko, A. Parisi, A. Busacca, L. Faes, R. Pernice, and J. Lee. Adaptive scheduling of acceleration and gyroscope for motion artifact cancelation in photoplethysmography. *Computer methods and programs in biomedicine*, 226:107126, 2022.

[17] A. Miltiadous, K. D. Tzimourta, N. Giannakeas, M. G. Tsipouras, E. Glavas, K. Kalafatakis, and A. T. Tzallas. Machine learning algorithms for epilepsy detection based on published eeg databases: A systematic review. *IEEE Access*, 11:564–594, 2022.

[18] M. Moor, O. Banerjee, Z. S. H. Abad, H. M. Krumholz, J. Leskovec, E. J. Topol, and P. Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023.

[19] S. Ng, S. Masarone, D. Watson, and M. R. Barnes. The benefits and pitfalls of machine learning for biomarker discovery. *Cell and tissue research*, 394(1):17–31, 2023.

[20] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[21] D. C. Whittingslow, J. Zia, S. Gharehbaghi, T. Gergely, L. A. Ponder, S. Prahalad, and O. T. Inan. Knee acoustic emissions as a digital biomarker of disease status in juvenile idiopathic arthritis. *Frontiers in Digital Health*, 2:571839, 2020.

[22] S. Zhang, S. M. H. Bamakan, Q. Qu, and S. Li. Learning for personalized medicine: a comprehensive review from a deep learning perspective. *IEEE reviews in biomedical engineering*, 12:194–208, 2018.