Université Paris Cité

École doctorale Hématologie, Oncogenèse et Biothérapies (HOB) 561

UMR-S 942 - MASCOT - Marqueurs Cardiovasculaires En Situations de Stress

Perioperative monitoring enhanced by artificial intelligence techniques based on deep neural network learning in a big data environment in anaesthesia

Par JADE PERDEREAU

Thèse de doctorat en Biothérapies et Biotechnologies

Dirigée par FABRICE VALLÉE

et co-encadrée par THOMAS MOREAU

Présentée et soutenue publiquement le 15/12/2025

Devant un jury composé de :

FABRICE VALLEE, PH HDR	Université Paris Cité	Directeur de thèse
ROMAIN PIRRACCHIO, FULL PROFESSOR	UCSF, USA	Rapporteur
HUGUES TALBOT, PU	CentraleSupélec	Rapporteur
MATHILDE MOUGEOT, PU	Université Paris Saclay	Examinatrice
JANE-LISE SAMUEL, DR	Université Paris Cité	Examinatrice
THOMAS MOREAU, CR HDR	Université Paris Saclay	Membre invité et co-encadrant
ETIENNE GAYAT, PU-PH	Université Paris Cité	Membre invité

Titre: Monitorage péri-opératoire augmenté par techniques d'intelligence artificielle basées sur l'apprentissage par réseaux de neurones profonds dans un environnement de données massives en anesthésie.

Mots clefs: Photopléthysmographie ; Apprentissage profond ; Temps réel ; Apprentissage multimodal ; Apprentissage self-supervisé ; Anesthésie générale

Résumé: Les données physiologiques du bloc opératoire sont générées passivement par les capteurs et, lorsqu'elles sont enregistrées, constituent un socle commun de connaissance du patient. Elles peuvent permettre de développer de nouveaux outils pour les anesthésistes, pour de l'aide à la décision pendant l'opération ou pour le soin postopératoire. Cette thèse a pour but de proposer des estimateurs en temps réel aux anesthésistes en utilisant les signaux non-invasifs monitorés en routine.

Parmi les métriques surveillées en anesthésie, la pression artérielle revêt une place centrale. On s'appliquera à l'estimer à partir des signaux monitorés en routine. Pour ce faire, on s'appuie sur des méthodes d'apprentissage profond, que l'on tente d'adapter aux connaissances biophysiques régissant le système cardiovasculaire. Dans un second temps, on s'intéresse à la profondeur d'anesthésie, qu'il est essentiel de maîtriser afin d'éviter les réveils accidentels pendant l'opération.

Une approche graduelle, au regard de la complexité des données, des paramètres des modèles et des méthodologies utilisées, sera employée. Après avoir introduit les concepts généraux, une première partie montrera l'intérêt de l'apprentissage profond et d'indicateurs cliniques dans la prédiction de la pression artérielle. Ensuite, une étude multicentrique permettra de comparer différentes méthodes d'apprentissage et de montrer l'intérêt d'une calibration basée sur l'architecture des réseaux récurrents, ainsi que la nécessité de l'usage de convolutions causales. Enfin, une approche self-supervisée sur l'électroencéphalogramme permettra de montrer notre capacité à prédire l'indice bispectral, qui est un indicateur clé de la profondeur d'anesthésie.

Title: Perioperative monitoring enhanced by artificial intelligence techniques based on deep neural network learning in a big data environment in anaesthesia.

Keywords: Photoplethysmography; Deep learning; Real time; Multimodal learning; Self-supervised learning; General anaesthesia

Abstract: Physiological data in the operating room is generated passively by sensors and, when recorded, forms a common base of knowledge about the patient. It can be used to develop new tools for anaesthetists, to aid decision-making during surgery or for post-operative care. The aim of this thesis is to propose real-time estimators for anaesthetists using routinely monitored non-invasive signals.

Among the metrics monitored in anaesthesia, blood pressure plays a central role. We will attempt to estimate it from routinely monitored signals. To do this, we will rely on deep learning methods, which we will try to adapt, where appropriate, to the biophysical knowledge governing the cardiovascular system. Secondly, we will also look at the depth of anaesthesia, which must be carefully controlled in order to prevent accidental awakening during surgery.

Given the complexity of the data, model parameters and methodologies used, a gradual approach will be employed. After introducing the general concepts, the first part will demonstrate the value of deep learning and clinical indicators in predicting blood pressure. Next, a multicentric study will compare different learning methods and demonstrate the value of calibration based on recurrent network architecture, as well as the need to use causal convolutions to guarantee causality. Finally, a self-supervised approach to electroencephalograms will demonstrate our ability to predict the bispectral index, which is a key indicator of anaesthesia depth.

Résumé substanciel en français

Cette thèse s'intéresse aux données périopératoires issues des moniteurs et à leur utilisation pour proposer des outils d'aide à la décision clinique. En utilisant les signaux non-invasifs monitorés en routine, on souhaite proposer des estimateurs en temps réel afin d'améliorer le soin pendant l'opération, ainsi que pour le soin post-opératoire en proposant des indicateurs de risque pour rediriger les patients vers des services de soins spécialisés si nécessaire. Parmi les métriques surveillées en anesthésie, la pression artérielle revêt une place centrale. Elle caractérise la perfusion sanguine des organes. A ce jour, la pression artérielle est monitorée principalement de deux façons : soit via un brassard à tension, qui est le monitoring usuel de bloc opératoire, et qui génère un signal discontinu, généralement toutes les cinq minutes, soit via un cathéter artériel invasif placé au niveau du bras. Celui-ci permet une mesure continue de l'onde de pression sanguine et est indiqué pour les patients à risque. En revanche, du fait de sa nature invasive, il nécessite un personnel soignant mieux formé et plus d'attention au cours de l'opération pour éviter les risques d'infection locale. Au cours de l'opération, de subites chutes de pression peuvent avoir des effets délétères sur les organes vitaux, même pour des durées courtes d'hypotension de moins de cinq minutes. Elles augmentent le risque de morbi-mortalité ainsi que le risque d'insuffisance rénale aigüe et d'infarctus du myocarde. Il importe donc d'être capable de détecter des chutes de pression entre deux prises de brassard. On s'appliquera à l'estimer à partir des signaux non invasifs monitorés en routine au bloc opératoire. Parmi ceux-ci, on distingue notamment la photopléthysmographie qui correspond à l'oxygénation du sang mesurée au niveau du doigt, l'électrocardiogramme placé au niveau du coeur, et le brassard à tension. En plus d'être non invasifs, ces signaux comportent également des caractéristiques physiologiques qui les associent à la pression artérielle. L'onde de photoplethysmographie peut

être perçue comme le reflet de la pression artérielle, et le deuxième mode de cette onde caractérise la résistance des vaisseaux. Couplé à la fréquence cardiaque, ce marqueur peut permettre de fournir une première approximation des variations de pression artérielle. L'ajout de l'électrocardiogramme peut permettre de déduire le temps de parcours de l'onde entre le coeur et les organes périphériques - la pléthysmographie étant mesurée au doigt - et fournir un proxy pour le temps de transit de l'onde. Enfin, le brassard à tension étant de fait toujours disponible au bloc opératoire, il est utilisé afin de calibrer le modèle. Afin d'apprendre des relations non linéaires entre ces signaux, nous nous appuyons sur des méthodes d'apprentissage profond, que l'on adapte, lorsque tel est le cas, aux connaissances biophysiques régissant le système cardiovasculaire. On observera une approche graduelle concernant la complexité des modélisations, en commençant par des approches supervisées, pour finir sur des approches non-supervisées dont le but est de tirer profit des données non étiquetées. Une approche séquentielle au regard de la complexité des données, des paramètres des modèles et des méthodologies utilisées sera utilisée. Concernant les données d'étude, une première étude se basera sur une cohorte préliminaire constituée au sein de l'hôpital de 220 patients issus pour moitié de chirurgie digestive. Pour ce type de chirurgie, les complications et variations hémodynamiques sont fréquentes et accrues. Une seconde étude addresse également le sujet par une approche supervisée, et élargit l'échelle à un niveau multicentrique. Après les étapes de processing des données, cette étude compte plus de 5060 patients de l'hôpital Lariboisière (APHP) et 3000 patients de VitalDB, une base de données d'anesthésie haute fréquence issue de l'hôpital de Séoul en Corée du Sud. Enfin, dans la troisième étude, on s'attachera à répondre au sujet du risque de réveil au cours de l'anesthésie, qui tout comme l'hypotension perpopératoire, a des conséquences néfastes pour le patient qui vont jusqu'à des conséquences long terme comme la dépression ou le syndrôme du choc post traumatique. Pour monitorer la profondeur d'anesthésie, on proposera une approche non supervisée appliquée à l'électroencéphalogramme, qui nous permettra de prédire un biomarqueur appelé l'indice bispectral, proposé par certains capteurs. Cette dernière étude nécessitant une cohorte comportant l'électroencéphalogramme et l'indice bispectral, caractéristique de la profondeur d'anesthésie, et se basera sur une cohorte externe issue de VitalDB de 797 patients. L'organisation de la thèse est la suivante :

1. Le chapitre 1 replace la thèse dans son contexte général. On présente d'abord les données monitorées en routine en anesthésie générale, ainsi que les opportunités qu'elles représentent pour améliorer le soin, que ce soit par la génération de nouveaux biomarqueurs peropératoires pour faire de l'aide à la décision au bloc opératoire, ou pour déduire des indicateurs de fragilité cérébrale, cardiaques ou globaux liés aux patients. La structure des données de bloc opératoire est également détaillée, ainsi que les étapes nécessaires à la génération de datasets exploitables. Enfin, on parcourt les principales catégories d'architectures d'apprentissage profond utilisées, des réseaux convolutifs, récurrents aux transformers. On introduira notamment les méthodes employées en dernière partie concernant l'apprentissage self-supervisé, qui consiste à diviser l'entraînement en deux étapes. Une première étape consiste à

Les chapitres 2 à 4 présentent les contributions de la thèse.

2. Le chapitre 2 met en avant la première contribution de la thèse. On y décrit plus précisément les indicateurs physiologiques du signal de photopléthysmographie qui sont utiles pour l'apprentissage. Un modèle mécanistique d'estimation de la pression artérielle est construit selon la loi d'Ohm-Poiseuille appliquée aux vaisseaux. Il définit la pression artérielle comme le produit de la résistance vasculaire, du débit cardiaque et de la fréquence cardiaque. Des proxys pour ces deux premières grandeurs peuvent être proposés à partir d'indicateurs du signal de photopléthysmographie, à savoir respectivement le deuxième mode d'onde et l'indice de perfusion défini comme le ratio de l'onde pulsatile sur l'onde non pulsatile. Ce modèle est comparé à un modèle d'apprentissage profond dont le but est d'apprendre les relations non linéaires entre les variables, et la reconnaissance de motifs sur le signal. Cette première version d'algorithme d'apprentissage profond est basée sur des réseaux convolutifs et récurrents. Les différentes versions de modèles sont évalués selon les standards du domaines, qui doivent vérifier un seuil d'erreur et de variabilité d'erreur. L'erreur moyenne absolue est également utilisée pour comparer les modèles, et les prédictions sont représentées sur des graphiques de Bland-Altman pour identifier d'éventuels biais et les intervalles de confiance des prédictions par rapport aux mesures. L'étude montre que le modèle le plus performant est le modèle de deep learning avec couches convolutives et récurrentes. L'interprétabilité des résultats sera également explorée grâce à l'utilisation de cartes de chaleurs (Grad-CAM)

illustrant les régions d'intérêt pour le réseau. Celles-ci mettent en avant le fait que la région de l'onde dichrote est informative pour le réseau, ce qui est cohérent avec le rationnel physiologique.

3. Le chapitre 3 approfondit le sujet abordé en chapitre II de prédiction de la pression artérielle. Dans cette étude, on ajoute une dimension multicentrique en comparant une version évoluée du modèles aux modèles de l'état de l'art. Contrairement à une partie substancielle des études existantes, on compare les modèles sur les mêmes jeux de données. On distingue deux catégories de modèles, les modèles non calibrées et les modèles calibrés. Cette distinction s'effectue notamment au regard de la multiplicité d'études pronant une approche sans calibration, convoitée pour le monitorage à domicile. Parmi les modèles utilisés, on distingue notamment une approche en U appelée Wave-U-Net, une approche basée sur des réseaux résiduels (Resnet), une approche mécanistique telle que vue au chapitre précédent ainsi qu'une approche convolutive avec calibration. Concernant notre approche, les améliorations apportées par rapport au chapitre précédent se situe à deux niveaux. Le premier est celui qui consiste à augmenter la taille du champ récepteur des réseaux convolutifs en utilisant des convolutions dilatées, qui consistent à augmenter le facteur de dilatation au fur et à mesure des couches. Cette approche est un moyen de conserver une architecture convolutive tout en bénéficiant des avantages des réseaux récurrents, à savoir une forme de "mémoire" long terme. Le second axe consiste en l'utilisation de la calibration non sous forme de variable mais comme état initial des couches de mémoire à court et long terme (LSTM). Cette calibration en utilisant l'architecture du réseau permet une calibration dynamique sans nécessiter de réentraînement. Les variables de cette étude incluent celles décrites au chapitre précédent complétées de l'électrocardiogramme (ECG), qui, couplé à la photopléthysmographie, permet comme énoncé plus haut une compréhension plus précise des variations cardiovasculaires du patient. Cette étude multicentrique se base sur deux cohortes, une cohorte de l'hôpital Lariboisière et une cohorte de VitalDB.

4. Le chapitre 4 aborde le sujet de la prévalence de données non étiquettées au bloc opératoire, et tente de répondre à cette problématique via une modélisation frugale basée sur l'apprentissage semi-supervisé. Dans cette partie, on s'intéresse plus particulièrement à l'évaluation de la profondeur d'anesthésie, principalement évaluée via des indicateurs sur l'électroencéphalogramme

(EEG). On appliquera des méthodes ayant porté leurs fruits dans le domaine de la polysomnographie, notamment le "relative positioning", qui consiste à apprendre au modèle à classifier si des paires de fenêtres de signal sont proches ou non. Cette méthode se révèle particulièrement utile sur l'EEG, possiblement en ce qu'elle capte les ondes lentes qui caractérisent l'endormissement, ou la profondeur d'anesthésie. Cette méthode se révélant infructueuse pour des signaux pseudo-périodiques comme la photoplethysmographie ou l'électrocardiogramme, nous testerons d'autres approches non supervisées sur ces signaux dans un second temps. Les autoencodeurs masqués permettront notamment de reconstruire fidèlement le signal. Afin de caractériser le lien entre la structure des embeddings apprise au cours du préentrainement, ceux-ci seront représentés via la méthode d'approximation et projection uniformes d'une manifold (UMAP). Cette représentation permettra de mettre en exergue le gradient par rapport à l'indice bispectral. Lors du tracé de trajectoire par catégorie de risque - caractérisé par l'âge ou le score ASA, on illustrera le fait que selon ces groupes de risque on observera un second gradient sur la manifold sur le deuxième axe pour une représentation 3D. Ce résultat semble indiquer que les représentations obtenues caractérisent la réponse du cerveau aux anesthésiants, et peut constituer un pas supplémentaire vers la formation continue pour mieux appréhender la profondeur d'anesthésie chez les cliniciens.

La thèse se termine par une section concernant le déploiement des modèles au bloc opératoire, qui a été testé sur l'infrastructure de l'hôpital. Les modèles nécessitant exclusivement des capteurs déjà utilisés en routine, le déploiement repose essentielle sur l'utilisation d'outils de déploiement (Docker, CI/CD). Cette courte section est suivie d'une conclusion qui retrace les principales contributions ainsi que les perspectives futures. On mentionne notamment les perspectives pour la prédiction de la pression artérielle, notamment en questionnant les approches sans calibration, l'apprentissage multimodal qui fait partie intégrantes des travaux présentés, et de la prédiction de l'incertitude, qui est un volet de recherche qui permettrait d'approfondir ces travaux.

Remerciements

Je souhaite tout d'abord adresser mes premiers mots de remerciements à mes directeurs de thèse, Fabrice et Thomas.

Fabrice, ta pédagogie et ta clairvoyance ont orienté le cheminement de mes travaux tout au long de la thèse. Ta sincérité et ta bienveillance rayonnent au sein de l'équipe et apportent une teinte humaine à une aventure qui peut paraître technique de l'extérieur. Si ton humour ne cache pas tout à fait ta sagesse, il a apporté beaucoup de légèreté à mes années dans le service et imprimé de durables souvenirs de camaraderie. Je ne pense pas exagérer en formulant le fait que tu es, plus généralement, un pilier de l'hôpital.

Thomas, ton excellence et ton exigence seront pour longtemps source d'inspiration pour moi, et m'ont permis de viser toujours plus loin. Je te remercie d'avoir été si présent au quotidien. Ta confiance et ton enthousiasme ont su maintenir ma motivation tout au long de ce chemin. Il serait difficile d'omettre tes qualités fédératrices dans l'équipe qui ont apporté une couleur chaleureuse et joyeuse à ces années à Turing.

Je remercie aussi chaleureusement Jona pour m'avoir guidée avec intuition et esprit, dans les environnements clinique et technique de l'hôpital, présentant chacun leur complexité. Je souhaite également remercier tout particulièrement Jérôme, avec qui j'ai eu le plaisir de travailler au cours de ma thèse. Je te remercie pour ta confiance, et pour tes conseils à la rigueur affutée. Ma gratitude va également aux professeurs Etienne Gayat et Alexandre Mebazaa, et à Philippe, Bertrand et Gaël, grâce à qui j'ai pu avoir la chance de conduire cette thèse au sein des équipes et de participer aux conférences et séminaires. Je tiens également à exprimer ma gratitude aux membres du jury pour leur participation à ma soutenance et leurs précieuses remarques. Je remercie tout particulièrement les rapporteurs pour le temps qu'ils ont consacré à la lecture de mon manuscrit, et pour leurs retours détaillés et constructifs.

Ensuite, je tiens à remercier mes collègues qui ont apporté gaité et constance dans mon quotidien de thèse. J'ai toujours eu plaisir à vous retrouver

au bureau, fusse au prix de volées de marches ou d'occasionnelles attentes nocturnes de train. Merci tout particulièrement à Virginie, pour ton soutien sans faille, ta bienveillance, le plaisir de travailler ensemble et la douceur des pauses tournesols. Merci à Julie pour tous ces fous rires, Oriane pour la légèreté des après-midi d'été, François, Alexis, Joachim, Cyril, Kenzha, Emilia, Ibrahim et tant d'autres pour la sympathie au temple et -1, et d'avoir tant de fois pris mes créneaux de modeste cuisine. Merci à Pierre-Louis, Marie, Ambroise, Félix, Célestin, Nicolas, Sébastien, Jad-Jad, Pierre-Antoine, Mansour, Samuel, Lionel, Iovan et tant d'autres, pour la bonne humeur à Turing, vos comptes ideal garden, et tant d'autres choses. Je remercie aussi les anciens de MIND, Louis et Raphaël, Théo, Alexis, Julia, les deux Alex, Cédric. Une pensée chaleureuse également pour les membres des équipes de l'EDS comme Thibaut et Alexandre avec qui j'ai eu plaisir à travailler.

A mes amis, la thèse est aussi un bout de vie et vous m'avez accompagnée tout au long de ce chemin. Grâce à vous la vie est plus douce, plus intense. Merci à Lucie, Raphaëlle, Hélène, pour votre amitié sans faille depuis tant d'années. Merci à Linayef, Simon, Robin, Nicolas, Clémentine, et tant d'autres pour ces belles années à vos côtés.

Je tiens également à remercier mes parents d'avoir guidé ma vie vers une expérience si riche, de leur écoute et soutien en toutes circonstances, à toute heure. Merci enfin à Médéric, d'être là pour moi, de me faire rire chaque jour et de rendre la vie si légère, comme un nuage.

Contents

Résumé ii

v

Remerciements x

List of figures xx

List of tables xxii

1 General Background 1

1.1 Background on general anesthesia and monitoring 1

1.1.1 Available vital signs 3

1.1.2 The process of anaesthesia 7

1.1.3 Detrimental events during GA: The case of intraoperative hypotension 8

1.1.4 Challenges in the definition and management of IOH . 11

1.1.5 Synthesis 12

1.2 Background on hospital data availability 13

1.2.1 Lariboisière Paris Hospital operating room database . . 13

1.2.2 Challenges in cohorts generation 16

1.2.3 The challenge of data quality 17

1.2.4 Clinical data warehouse 18

1.2.5 External open-source databases 19

1.3 Opportunities for machine learning in the operating room . . 22

1.3.1 Leverage underutilised GA data for diagnostics 22

1.3.2 Building alternatives to expensive sensors 24

1.3.3 Towards developing integrative biomarkers for predictive anesthesia 25

1.3.4 Towards closed loop drug delivery? 26

1.4 Background on sequence-to-sequence learning 27
1.4.1 Multimodal autoregressive modelling 27
1.4.2 The learning process of neural networks 28
1.4.3 Architectures 30
1.4.4 Self-supervised paradigm 32
1.4.5 Explainability in machine learning 36
1.5 Objective of the thesis and contributions 38
1.5.1 Objectives 38
1.5.2 Contributions 39
1.5.3 Publications 41

2 Physiologically inspired modelling based on the photoplethysmography 42
2.1 Introduction 42
2.1.1 Physiological rationale 42
2.1.2 Mechanistic Model 44
2.1.3 Concurrent studies 45
2.2 Methods 46
2.2.1 Patients 46
2.2.2 Anesthesia Protocol 47
2.2.3 Data 48
2.2.4 Deep Learning Model 49
2.2.5 Naïve Model 51
2.2.6 Statistical Analysis 51
2.2.7 Visualization of results using Bland-Altman representations 52
2.2.8 Mixed-effects modelling to handle correlated data . . . 53
2.3 Results 54
2.3.1 Patients 54
2.3.2 Models performance 54
2.3.3 Repeated measurements correlation and Bland-Altman analysis 56
2.3.4 CNN + LSTM model explainability 58
2.4 Discussion 58
2.4.1 Main findings 58
2.4.2 Limits of the mechanistic approach 60
2.4.3 Perspectives 61

2.5 Chapter conclusion 61
2.6 Appendix 61

3 Benchmarking models on common datasets 65
3.1 Introduction 65
3.1.1 Why use ECG and PPG together? 65
3.1.2 Arterial pressure 66
3.1.3 Existing methods 67
3.1.4 Operating room constraints 68
3.2 Methods 69
3.2.1 Approval for Data Acquisition 69
3.2.2 Data selection/Pre-processing 70
3.2.3 Study design: Comparison to SOTA architectures . . . 71
3.2.4 AnesthNet model characteristics 73
3.2.5 Statistical Analysis 76
3.3 Results 77
3.3.1 Datasets characteristics 77
3.3.2 Considered models 79
3.3.3 Models performance assessment 80
3.3.4 The impact of distance to cuff calibration 84
3.4 Discussion 85
3.4.1 Calibration as a field limit 85
3.4.2 Implementability at the operating room 86
3.4.3 Limitations on field characteristics 86
3.5 Chapter conclusion 87
3.6 Appendix 88

4 Self-supervised learning approaches to unveil brain and vessels representations 91
4.1 Introduction 92
4.1.1 On the depth of anesthesia 92
4.1.2 SSL methods applied to EEG 94
4.1.3 Defining similarity in anesthesia 94
4.2 Methods 95
4.2.1 Patients 95
4.2.2 Training pipeline 96
4.2.3 Quantitative evaluation on downstream tasks of interest 97

4.2.4 Qualitative assessment using visual representations . . 99
4.2.5 Statistical analysis 99
4.3 Results 100
4.3.1 Patients 100
4.3.2 SSL on the EEG reveals the depth of anesthesia 101
4.3.3 Patient trajectories 103
4.4 Experiments on other modalities 105
4.4.1 Leveraging multimodal input signals 105
4.4.2 Visual representations of our results on the PPG 108
4.5 Discussion 110
4.5.1 BIS prediction 110
4.5.2 Masked autoencoders performance 111
4.5.3 Limitations 112
4.5.4 Perspectives 112
4.6 Conclusions 113
4.7 Appendix 113
4.7.1 Supplementary result: Relative positioning is ineffective in capturing periodic signals 113
4.7.2 Supplementary result: Fine-tuning the masked autoencoder on predicting the arterial pulse pressure 114
4.7.3 Supplementary table: curvature statistics per age group 115

5 Real-time deployment perspective 117
5.1 Model deployment at the operating room 117
5.1.1 Model serving application using FastAPI 117
5.1.2 A consumer application to process messages in real-time 119
5.1.3 Deployment perspectives 120

6 Conclusion and perspectives 121
6.1 Summary of contributions 121
6.2 Open questions 123
6.2.1 Cuffless blood pressure monitoring perspective 123
6.2.2 Prediction of uncertainty 124
6.2.3 Multimodal learning 125

List of abreviations 126

7 Appendix 128

Bibliography **166**

List of Figures

1.1 Top ten global causes of death in 2016 2

1.2 General anesthesia vital signs of interest to this study 6

1.3 Illustration of the biosignals acquired in the operating room. From top to bottom, the ECG, PPG, arterial pressure, CO2, EEG, Cuff pressure. Perfusion index will be developed in chapter 2. . . 9

1.4 Data interoperability at the operating room 14

1.5 Physiological data platform at Lariboisiere 16

1.6 Volume of AI-related general anesthesia publications (Ou et al., 2025) and volume of high-frequency data storage in Lariboisiere 23

2.1 Physiologically informative features on the ppg signal 43

2.2 Patients inclusion diagram 47

2.3 CNN + LSTM model architecture 50

2.4 General study design and features for cuff, mechanistic, and deep-learning models. 52

2.5 Correlation plots and Bland-Altman representations. The first row corresponds to the whole test set. Then, each row displays a comorbidity. From top to bottom, the plots correspond to: the whole test set, hypertensive vs non-hypertensive patients, patients aged above or below 55 years old, and men versus women. 57

2.6 Saliency map. Deep learning model first convolutional layer saliency map for 1 batch of 5 s of raw PPG waveform measured at a 125 Hz frequency for the test cohort. PPG indicates photoplethysmograhy. 58

3.1 Pulse wave velocity explanatory diagram 67

3.2 Proposed deep learning model architecture. **a**. Multi-modal features are shaped as sequences of inputs between two cuff measurements. They include the cuff oscillometer recorded every five minutes, the high-frequency ECG and PPG (respectively acquired at 500Hz and 125Hz and downsampled to 125Hz), and PPG-derived features which carry physiological information about MAP variations: the Notch Relative Amplitude (NRA), Perfusion Index (PI), and Heart Rate (HR). **b**. High-frequency signals are processed using a convolutional network with increased dilation to maintain causality throughout the network while increasing the receptive field within the sequence without increasing the number of parameters. The output of this network is then concatenated on the channels axis with the low-frequency derived features (NRA, PI, HR) to feed a Long Short Term Memory network. **c**. The LSTM network is initialized using the cuff value as the initialization parameter. The output sequence corresponds to the predicted MAP sequence between the cuff measures. 74

3.3 Study design. a. Exploitable datasets sizes: LaribDB (5060 patients) and VitalDB (2833 patients). b. Multimodal features categories which include high-frequency data (ECG, PPG), PPG-derived features, and the cuff oscillometer. c. Preprocessing steps, filtering, and generation of ready-to-train datasets. d. State-of-the-art models compared in this study, including physiology-based, calibration-free and calibration-based approaches. e. Each model takes as input a combination of non invasive features and produces a prediction of MAP per window of data. Results are then aggregated per patient to generate the main results metrics. 79

4.1 Depth of anesthesia and bispectral index values correspondance diagram 93

4.2 Self-supervised learning training pipeline on EEG 98

4.3 SSL model predictions colored by bispectral index values 102

4.4 Self-supervised and supervised learning comparison. Lower is better. Standard deviation of the absolute error is represented as the shaded area. 103

4.5 Trajectories on the test dataset by age group: young (age<35), adult (35<age<60), senior (60<age<75) and old (75<age), in three dimensions on the left and on the X/Y plane on the right. The trajectories of each group are represented in a different color, with a gradient from green, light green, yellow to orange. 105
4.6 Masked autoencoder study design on semi-periodic signals (PPG) 107
4.7 A. Masked autoencoder training and validation curves on pretext task. B. Example of masked autoencoder reconstructed PPG signal. C. Example masked regions on the PPG signal. 108
4.8 Masked autoencoder UMAP representation on the PPG colored by heart rate values. 109
4.9 UMAP representations of the catheter arterial pressure colored by a color code based on vasopressor concentration. 114
4.10 UMAP representation of the fine-tuned masked autoencoder, colored by the amplitude of the PPG. The manifold exhibits two distinctive clusters of extreme values of photoplethysmography pulse pressure. This result suggests that to learn arterial pulse pressure, the model has learned to discriminate between high and low PPG amplitudes. 115

5.1 Deployment architecture at the operating room 120

List of Tables

1.1 Comparison of Larib, VitalDB, and MIMIC Databases 21
1.2 Comparison of architecture types 33
2.1 Summary of patient characteristics and features 55
2.2 Model performance metrics 56
3.1 Model features definition and eventual calibration 76
3.2 Datasets characteristics 78
3.3 Models performance comparison among various learning- and PPG-based cuffless BP estimation architectures. Estimation was based on AAMI standard: The number of test subjects needs to be higher than 85, the ME should be $\leq \pm 5$ mmHg, and the SD of error should be ≤ 8 mmHg. The highlights in bold represent the model reaching the best performance. 81
3.4 Models performance comparison among various learning- and PPG-based cuffless BP estimation architectures. Estimation was based on the BHS standard: The grades based on the BHS standard are given based on the error percentage as follows: if 60%, 50%, and 40% of errors ≤ 5 mmHg, then grades A, B, and C are given, respectively; if 85%, 75%, and 65% of errors ≤ 10 mmHg, then grades A, B, and C, are given, respectively; if 95%, 90%, and 85% of errors ≤ 15 mmHg, then grades A, B, and C, are given, respectively. The highlights in bold represent the model reaching the best performance. 81
3.5 Models performance regarding mean inference time. Mean inference time was defined as the time taken by each model to perform inference on a batch of features. It was averaged over 15 batches and latencies were reported in milliseconds. 83

3.6 Calibration-based models performance regarding distance to cuff calibration. Segments of respectively 3, 5 and 10 minutes were generated to model the variation of cuff measurement frequency. Inference is performed during these intervals for calibrated models. 84

4.1 Descriptive statistics for continuous variables by subset 101
4.2 BIS prediction performance comparison. N/A indicates that the authors did not report the corresponding metric. 103
4.3 Masked-Auto Encoder hyperparameter search space and best configuration for PPG 109
4.4 Downstream tasks performance of the masked autoencoder using the PPG on the PPG-BP dataset - external validation cohort. Classification tasks such as hypertension are compared using area under the curve AUC - unitless, between 0 and 1. Regression metrics are presented as absolute error mean and 95% confidence intervals for consistency with previous studies. 110

1 General Background

This thesis lies at the intersection between the scientific fields of computational science and medicine. The core of this project was to evaluate the benefits of applying machine learning and deep learning methods in the environment of high-frequency operating room data, particularly using real-world hospital data. Before delving into the details of this thesis, it is necessary to introduce the general concepts that will be essential for understanding the study. In this first introductory section, we will explain the general background of this work: perioperative care and general anesthesia. Next, we will explain the context of anesthesia databases. Naturally, at the intersection of the two fields of research, we will examine the opportunities for deep learning on anaesthesia data. Finally, the learning methods used in the scientific community and in this study will be introduced.

1.1 Background on general anesthesia and monitoring

Anesthetists often compare general anesthesia (GA) to flying an airplane. Both guarantee overall safety, are theoretically as safe, and are composed of successive characteristic phases. First, the plane must take off, which corresponds to the induction process. This phase is the injection of drugs to transition the patient to a state of unconsciousness, and is generally the most challenging phase and the one that requires the most attention from the pilot or clinician. After that, once the patient is asleep and stable, risks become less likely and originate most probably from surgical operations. This could correspond to a stationary phase. Finally, a plane landing would correspond to the phase during which a patient wakes up. It requires vigilance but is primarily a natural process. To date, postoperative complications are

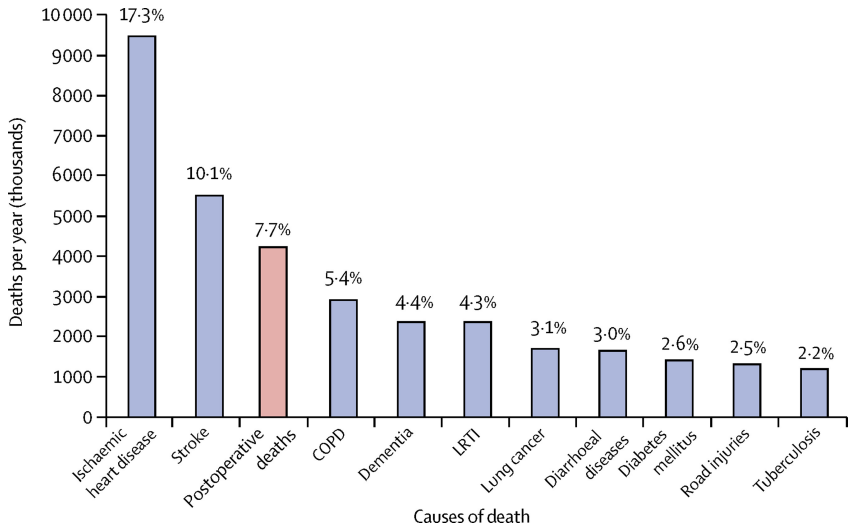


Figure 1.1: Top ten global causes of death in 2016

the third leading cause of mortality around the world, as shown in Figure 1.1. Modern anesthesia-related mortality has decreased dramatically from historical rates of 6.40 per 10,000 anesthetics in the 1950s to current rates of 0.4 per 100,000 anesthetics, as highlighted by (Gottschalk et al., 2011). Better anesthetic management, such as the establishment of international standards, has evidently played an important role in improving these surgical outcomes. However, the overall 30-day postoperative mortality remains concerning at approximately 2.1% (Fecho et al., 2008). These range from cardiac events - including myocardial infarction - to acute kidney injuries, stroke, sepsis, and other organ dysfunction. The most recent comprehensive data from the 7th national audit project of the Royal College of Anesthetists in London (United Kingdom), reveals that serious complications occur in 1 in 18 cases (6%) during anesthesia, with circulatory events accounting for the majority (36%) of all complications. This finding underscores the critical importance of maintaining adequate hemodynamic stability throughout the perioperative period. These complications are mostly cardiovascular in origin. Another study conducted in France by (Lienhart et al., 2006) found that two-thirds of deaths occurring during and after surgery were directly or indirectly related to cardiovascular causes.

1.1.1 Available vital signs

To guarantee patient safety during general anesthesia, clinicians around the world follow international standards and guidelines regarding the process of anesthesia and monitoring equipment. Renowned associations such as the American Society of Anesthesiologists recommend the following non-invasive physiological signals as the minimum standard for monitoring:

The electrocardiogram

The heart's activity is of crucial importance during general anesthesia, as it perfuses the other organs with blood. If we place two recording electrodes on the body and connect them to an electrocardiograph, we observe a signal called the electrocardiogram (ECG). It represents the electric potential variations from the heart during a systole. More precisely, on the ECG signal, we distinguish several deflections, such as the P wave, which characterises the depolarization of the atria, the QRS complex, which represents the depolarization of the ventricles, and the T wave, which characterises the repolarization of the ventricles. Q is always negative, R is the peak, and S is a second peak that can exist after the R wave. These are illustrated in Figure 1.2.

The frequency and morphology of the ECG, as well as the amplitude of the different parts of the patterns, help physicians characterize normal or pathological patients. From primary care to intensive care units, the ECG's accessibility, speed, and cost-effectiveness make it one of the most widely used tools for diagnosing heart conditions—ranging from atrial fibrillation to acute myocardial infarction—worldwide. In clinical practice, the number of electrodes can vary but most devices propose 12-lead or 5-lead.

The photoplethysmography

In parallel with electric activity, periodically, with each beat, the heart ejects blood through the aorta into the peripheral vessels, and this wave, although slightly attenuated by distance and travel time, is sufficient to relax the arteries and arterioles in the subcutaneous tissues. By placing a pulse oxymeter,

which is a sensor that illuminates the skin via an LED and measures variations in light absorption by the skin using a photodiode, it is possible to measure blood perfusion in microvascular tissue using this wave, known as photoplethysmography (PPG). Traditionally, in the operating room, this wave is measured on the finger and provides a continuous measure of pulse oximetry (SpO_2), which is the universal quantifier of adequate tissue oxygenation during general anesthesia (GA). Its shape is similar to that of the blood pressure wave. It should be noted that, as with other biomedical devices, due to the variety of sensors marketed for clinical or wellness applications, the exact function of calculating the waveform may depend on filters or operations performed by the manufacturer and is not necessarily known. Monitoring the PPG is a non-invasive, lightweight, and cheap method to address multiple operating room uses, from heart rate, cardiac cycle, respiration, hyper and hypovolemia - an increased or decreased volume of blood circulating in the body, and blood pressure monitoring. Some cyclic features on this signal will be described in more detail in chapter 2.

The cuff oscillometer vs arterial catheter

To ensure adequate circulatory function in all patients undergoing anesthesia, arterial blood pressure must be measured and assessed at minimum intervals of five minutes, in addition to continuous electrocardiographic (ECG) monitoring. This protocol is essential for maintaining patient safety and optimizing perioperative outcomes. To acquire this discontinuous metric, mercury sphygmomanometers (which originate from the Greek term "pulse" and the scientific term manometer for "pressure meter") are considered as the gold standard measurement. This device is composed of an inflatable cuff to collapse and release the artery under the cuff in a controlled manner, and a mercury manometer to measure the pressure. It is a routine non-invasive monitor that provides an intermittent non-invasive blood pressure measurement, typically every 3–5 minutes. The cuff provides a measure of systolic blood pressure - maximum per cycle, diastolic blood pressure - minimum per cycle, and mean arterial pressure per cycle. The mean arterial pressure (MAP) is defined as

$$MAP = \frac{1}{3}(SBP - DBP) + DBP \tag{1.1}$$

where SBP corresponds to the systolic blood pressure and DBP to the diastolic blood pressure. Later on, the cuff series of measurements will be referred to as the calibration on which a naive measurement can be built.

Alternatively, invasive arterial pressure monitoring through catheters, or arterial lines (A-line), can be placed in high-risk cases. Additionally, these systems deliver continuous arterial pressure waveforms, supplemented by cyclic measurements including mean, systolic, and diastolic values. This device remains optional due to its increased demand on nursing attention and the elevated risk of local infection.

Capnography

Continuous monitoring of ventilation is achieved through capnography, which corresponds to the concentration of partial pressure of carbon dioxide (CO_2) in the respiratory gases. Capnography is considered essential for detecting hypoventilation and confirming airway device placement. The waveform also confirms the endotracheal tube placement and can prevent injuries due to ventilatory failure or defective breathing circuits during the intervention.

The electroencephalogram

Even though it is not subject to mandatory monitoring during general anaesthesia, measuring the electroencephalogram (EEG), or electrical activity of the brain, is gaining popularity in both clinical practice and intraoperative research. This measure of the electrical activity from neurons' communication weak electromagnetic fields reflects the brain's instantaneous activity and can be measured with conductive electrodes. The voltage difference between two electrodes indicates the brain activity occurring in the volume between their locations. The availability of polysomnography databases has propelled research in this field, particularly through artificial intelligence algorithms, and an increasing number of clinicians are using this non-invasive measurement to monitor the depth of anaesthesia (DoA).

Synthesis

These routinely measured signals are endorsed by major organizations such as the World Health Organization, World Federation of Societies of Anaesthesiologists, the French National Society of anesthetists and the American Society of Anesthesiologists (Bi et al., 2023). Facilities unable to meet these standards are considered to be providing care below the acceptable safety threshold.

In the following work, we attempt to maintain a frugal approach, in the sense that the features we will use in our models will come from sensors already in routine use. The implementation of our results will therefore not require any additional equipment. The signals of interest to our studies, which is a subsample of the aforementioned signals, can be represented in Figure 1.2.

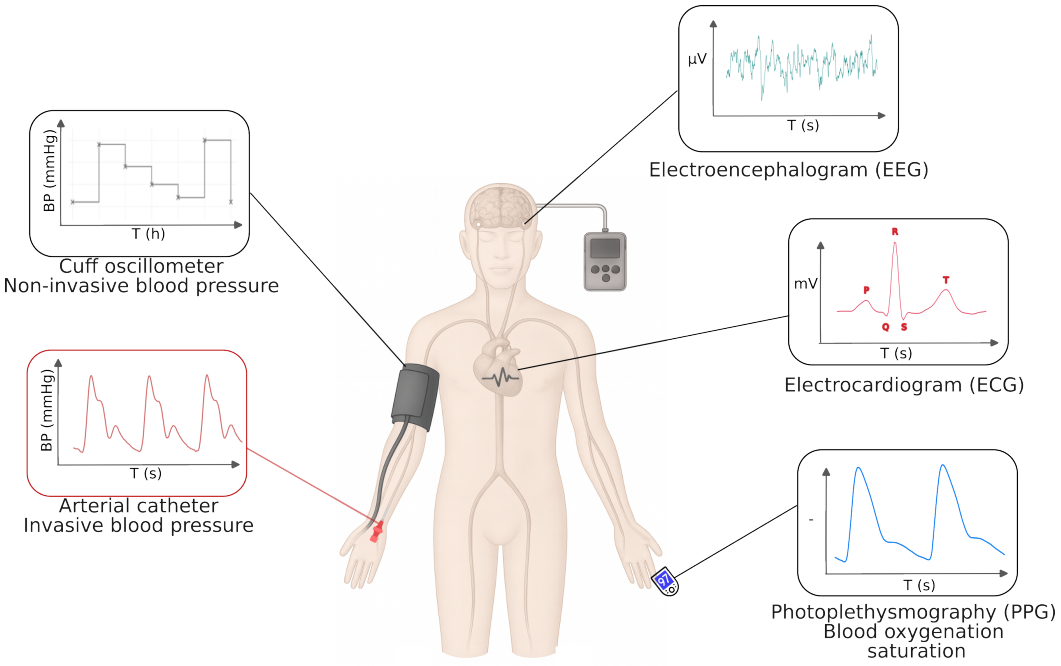


Figure 1.2: General anesthesia vital signs of interest to this study

1.1.2 The process of anaesthesia

The following outlines the sequential steps and key objectives involved in general anesthesia for adult patients. While this framework provides a standardized approach, it is important to note that individual patient characteristics and surgical requirements may necessitate adaptations.

The induction phase

First, venous access - a drip - must be established in the patient to administer the anaesthetic drugs and rehydration solutions. Once the drip is in place and before the injection of the anaesthetic drugs, the patient is given pure oxygen to breathe, ensuring their oxygen reserves are adequate. Then, three anaesthetic drugs are usually injected in doses calculated empirically based on the patient's weight, height, and age. The following are injected: a powerful morphine derivative such as remifentanil or sufentanil to ensure sufficient analgesia, a sedative such as propofol to put the patient into a deep sleep, and a curare-type muscle relaxant - such as atracurium or rocuronium - to paralyse the patient's muscles. Next comes the airway control phase. As the patient can no longer breathe on their own, a tube is usually placed in the trachea -intubation process- to allow the patient to breathe with the aid of a ventilator. The correct position of the tube is checked using a capnograph.

The maintenance of a stable hypnosis phase

Next comes the phase of adjusting the drugs and maintaining anaesthesia. The level of sedation and analgesia is set by modifying the administration of the sedative flow rate - intravenous propofol or halogenated gases such as sevoflurane administered via the respiratory tract - and morphine - remifentanil administered intravenously or sufentanil administered in repeated rapid, single-dose administration called boluses - while monitoring the depth of anaesthesia (DoA) using the EEG. More specifically, clinicians control aggregated values on the EEG signal, such as the Bispectral Index (BIS) available on the sensor commercialized by Medtronic - a unitless continuous scalar which varies between 0 and 100 and which has to be maintained within the range 40-60 during surgery. Another widely clinically validated indicator is

the Patient State Inde, which is monitored using the SedLine © EEG monitor, which characterizes sedation and drug effect. Arguably, one could note that these objectives are based on low-frequency mean values aggregated from high-frequency signals, and wonder if the whole signals hold greater potential.
The level of paralysis is adjusted using a curarimeter monitor according to the needs of the surgery. Throughout this phase, hemodynamic stability is ensured, targeting a mean arterial pressure above 65 mmHg and a heart rate maintained between 60 and 80 beats per minute. To achieve this, vasopressors - such as baby-noradrenaline or ephedrine and vascular filling are administered if necessary. All these drugs are maintained and adjusted during the maintenance phase of anaesthesia, which allows the surgical procedure to be performed. Continuous monitoring of EEG, MAP, and HR enables real-time adjustment of drug administration, ensuring an optimal balance between anesthetic depth, hemodynamic stability, and the progression of the surgical procedure.

The awakening

At the end of the surgery, sedatives, powerful morphine derivatives, and muscle relaxants are discontinued. Medications are administered to ensure post-operative analgesia, namely using paracetamol, anti-inflammatories or long-acting morphine. When the patient is awake and breathing independently, the endotracheal tube is removed - extubation - and the patient is transferred to the recovery room.

Figure 1.3 represents the successfully acquired signals, which are usually displayed together on a scope during general anesthesia.

1.1.3 Detrimental events during GA: The case of intraoperative hypotension

Modern anaesthesia is safer than ever, with remarkable success rates. The risk of mortality directly related to anaesthesia for a healthy person undergoing routine surgery is estimated at around 1 in 100,000 or even 200,000 cases.

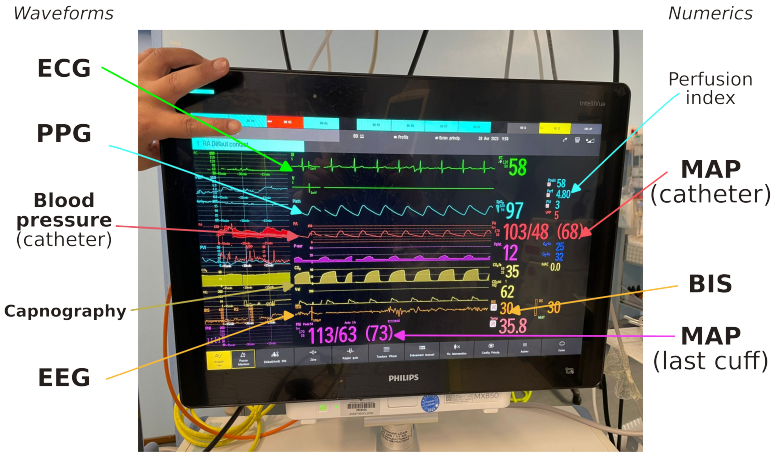


Figure 1.3: Illustration of the biosignals acquired in the operating room. From top to bottom, the ECG, PPG, arterial pressure, CO2, EEG, Cuff pressure. Perfusion index will be developed in chapter 2.

This success is the result of significant advances in anaesthetist training, monitoring equipment, and medication.

However, despite these reassuring figures, it is essential to recognise that adverse events can occur during general anaesthesia, which can affect patients' long-term quality of life. Accidental awareness during general anesthesia (AAGA), along with burst suppressions - a pathological EEG pattern characteristic of profound hypnosis state, can induce lifelong complications on daily life, from post traumatic stress disorder to immunosuppression. Besides, intraoperative hypotension (IOH) – a drop in blood pressure during surgery – is an independent risk factor for complications like myocardial injury and acute kidney injury (Bijker, Klei, et al., 2007; Walsh et al., 2013). Its management is the subject of intensive research. It has even been associated with extended hospital stays and increased risks of postoperative mortality (Sessler et al., 2012; Cartailler et al., 2024; Wesselink et al., 2018), with a pooled odds ratio for mortality of approximately 1.94 (95% CI: 1.32–2.84) in patients experiencing IOH compared to those who do not. IOH is a frequent intraoperative event. Its occurrence happens in a significant proportion of surgeries—up to 65% when defined as mean arterial pressure (MAP) below 65 mmHg. Its occurrence results from a compromise from the anesthetists regarding blood pressure monitoring.

To date, there are two principal methods of monitoring blood pressure: continuous monitoring using an arterial catheter and discrete measurements obtained using the cuff oscillometer. Arterial catheterization is the gold standard for monitoring blood pressure in high-risk patients in intensive care units despite its invasiveness (Lehman et al., 2013; Roach and Thiele, 2019). It requires technical expertise and particular attention during surgery due to the risk of local infections and intra-operative complications (Murray, 1981; Sladen, 1988). Consequently, standard cuff oscillometry is often preferred for its less invasive aspect (Meidert et al., 2019), despite its intermittent nature, leading to inaccuracies in detecting intraoperative hypotension (Roach and Thiele, 2019).

Recent evidence suggests that both the magnitude and duration of hypotensive episodes are critical determinants of patient outcome. Even brief periods of severe hypotension (MAP $<$ 55 mmHg) can be more harmful than more extended periods of moderate hypotension. This finding emphasizes the need for rapid recognition of hypotensive episodes, immediate intervention to restore adequate perfusion pressure as well as individualized blood pressure targets based on patient comorbidities. In a population of high-risk patients undergoing major abdominal surgery, a randomized controlled study of (Futier et al., 2017) showed that individualized blood pressure management strategies, aimed at minimizing IOH exposure, had the potential in reducing perioperative organ damage. Therefore, screening and treating IOH can be crucial to guarantee patient safety.

As the clinical monitoring effort focuses on the mean value of arterial pressure, a valuable feature is to provide an interpolation of mean arterial pressure between brachial cuff measures. In fact, other signals from the operating room hold valuable information on pressure variations. This will be developed further in chapters 2 and 3. In addition to applications in the operating room, it should be noted that monitoring mean arterial pressure goes beyond the hospital setting, as it is also used for remote monitoring, particularly for hypertensive patients. One notable difference is that blood pressure is already monitored in the operating room, but at too low a frequency to prevent cases of hypotension. Part of the work, therefore, involves attempting to use the calibration obtained from the cuff intelligently and naturally.

1.1.4 Challenges in the definition and management of IOH

IOH complex definition

To avoid IOH, which mean arterial pressure (MAP) value to target, then? While no consensus threshold exists for detecting perioperative arterial hypotension, both a MAP decrease exceeding 20% from baseline and an absolute MAP below 60 mmHg—even for brief durations—have been linked to a significant increase in postoperative complications. (Walsh et al., 2013; Bijker, Persoon, et al., 2012). In a retrospective cohort of 57,315 patients, Salmasi et al., 2017 recently demonstrated that hypotension defined by a relative change in MAP (drop $>$ 20%) compared to an absolute MAP value ($<$ 65 mmHg) did not offer any advantage in predicting the occurrence of renal or myocardial damage. This study suggests that anaesthetic management could be based on an absolute MAP value. Nonetheless, in everyday clinical practice, the haemodynamic strategy in most centres is based on absolute MAP values, with a target range of 65 to 75 mmHg. Latest studies suggest that at-risk patients should not meet MAP values lower than 60mmHg. In fact, the literature does not converge towards a standard definition, and some advocate that such a definition is not likely to be found (Brady and Hogue, 2013).

IOH management at the operating room

The etiology of IOH is multifactorial, often involving overlapping mechanisms. One potential cause is excessively low vascular resistance, which can be addressed through the administration of vasopressors to enhance arterial tone and afterload, thereby ensuring adequate aortic valve opening pressure. Secondly, it is possible that the heart is not filling sufficiently. In this case, venous flow must be increased, which can be done by adding saline solution or blood through auto- or heterotransfusion. Finally, if the source of IOH is a lack of myocardial strength, the clinician administers inotropes, such as dobutamine. In practice, these mechanisms frequently coexist, as vasopressors often exhibit both vasoconstrictive and inotropic effects. Regarding clinical practice, the most natural way to avoid IOH is to

administer vasopressors, which are drugs that constrict blood vessels. These agents induce peripheral vasoconstriction in veins and terminal arterioles, redistributing blood flow toward vital organs like the coronary and carotid systems, albeit at the expense of less critical regions such as the renal and muscular networks. It is initially accompanied by an increase in preload due to venoconstriction, but also by the rise in left ventricular afterload (Thiele et al., 2011; Goertz et al., 1993). As a consequence, cardiac output has been shown to decrease after boluses of vasopressors (Vallée et al., 2019). Reaching adequate organ perfusion remains a complex objective. Future IOH strategies will likely include thresholds for patient-specific hypotension, organ-specific targets, and postoperative hypotension, which has been less monitored and studied so far.

1.1.5 Synthesis

Intraoperative hypotension (IOH) remains a critical challenge, especially for patients with cardiovascular risks, as it demands precise hemodynamic management and personalized treatment. Non-invasiveness is a crucial characteristic for upcoming clinical monitoring tools. To achieve this, representative databases and learning methodologies could enhance real-time monitoring and customized interventions, helping clinicians anticipate and mitigate IOH while optimizing patient outcomes.

1.2 Background on hospital data availability

Analyzing events that occur during anaesthesia requires access to data recorded during treatment. Therefore, the generation and availability of time series databases within the scientific community is a catalyst for accelerating research. As our work is based in a hospital that has developed its own data collection system—rare for high-frequency data in France—we introduce here this technical environment and the challenges it presents for scientists wishing to contribute to this field.

1.2.1 Lariboisière Paris Hospital operating room database

Database overview

In recent years, technological developments and falling storage costs have enabled the recording of high-frequency signals, including those issued in the operating room. In France, this opportunity was seized in 2018 by the Anesthesiology department at the Lariboisiere Hospital inside the Assistance Publique Hôpitaux de Paris (APHP), in partnership with the APHP clinical data warehouse central services and the industrial biomedical vendor Philips. I was part of the project and contributed to make it more scalable. The APHP is a grouping of hospitals and aggregates data from 38 hospitals—including Lariboisière Hospital—from which the French data in this study originate. The Lariboisière Hospital is the first one of these centres to collect physiological data. The recording was enabled with the use of the Data Warehouse Connect (DWC), the solution commercialized by Philips to record vital signs for clinical use. Since then, fifty-two operating room monitors from various departments such as maternity, neurosurgery, orthopaedics, neuroradiology etc, have been connected to a central station, which aggregates the interoperable flow of HL7-formatted data in real-time and sends it to a Kafka platform. Figure 1.4 represents the main data sources at the operating room.

The high-frequency data is then stored in an SQL Server database - a technical choice made by the vendor. From 2018 to 2024, more than 60,000 raw physiological recordings have been collected within the hospital, including more than 10,000 patients with invasive arterial lines (A-line). Since

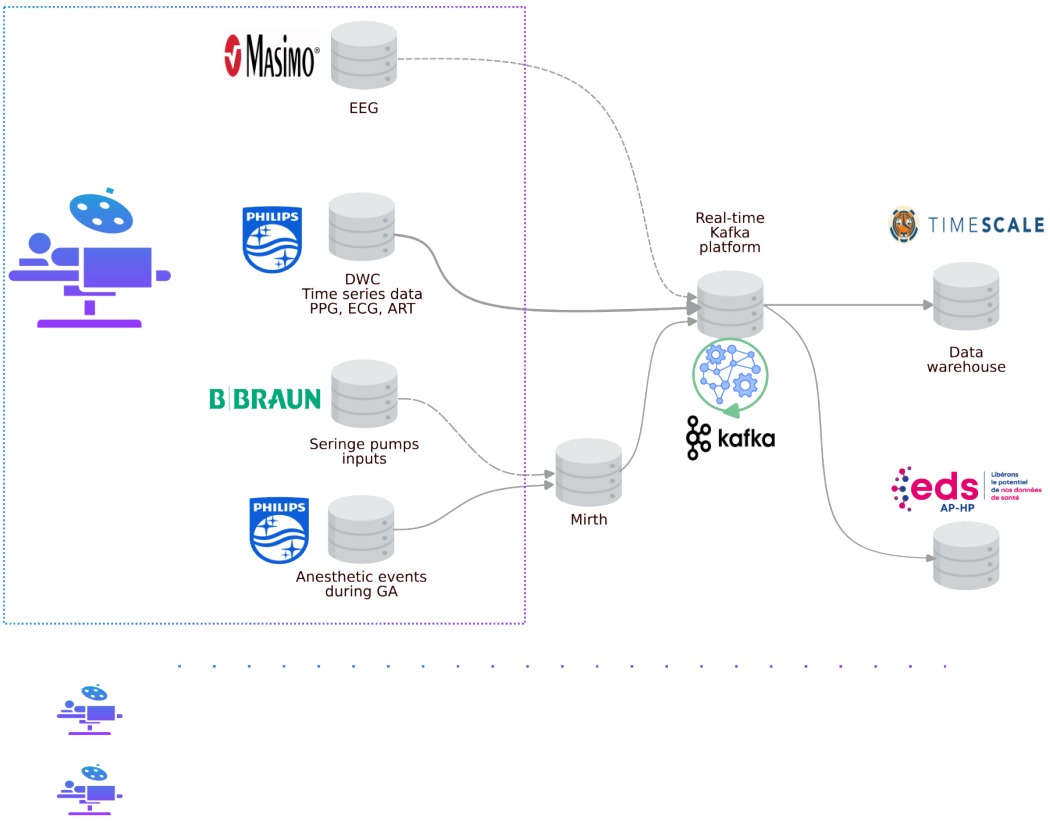


Figure 1.4: Data interoperability at the operating room

2019, the records generated each month have reached one terabyte. This volume of data makes it one of the largest anesthesia databases around the world. Considering physiological data, biosignals are divided into two main categories: high-frequency data, commonly referred to as 'waveform', whose acquisition frequency generally varies between 100 and 500 Hz, and cyclical data computed from these signals, which typically have a frequency close to one beat per second, i.e. 1 Hz. Different sources of data are being added as things progress, with the inclusion of the events occurring during GA along with the seringe pumps in 2023. As the volume and sources of data increase over time, the complexity of the datalake also increases, necessitating custom tools to interact with and work on the data.

Accessing operating room data

Creating a high-frequency biosignals database requires the establishment of a solid data environment. A first challenge resides in accessing high-frequency

data. SQL Server's ACID-compliant transactional architecture, while essential for maintaining data consistency, introduces significant overhead that becomes prohibitive under high-frequency data ingestion scenarios. The database's page-based storage model and sequential transaction log writing create inherent bottlenecks, as every write operation must traverse multiple layers of indexing, locking, and durability mechanisms. Consequently, the system exhibits degraded throughput and increased latency as insertion rates scale, making it fundamentally unsuitable for applications requiring sustained high-frequency data streams. To facilitate the work on the data, a custom database named DWCmeta has been developed on top of the vendor's database architecture to enable almost real-time access by loading the original database metadata only. It addresses both research and clinical needs, from the exploration of the volume of data with given signals of interest, to the actual recording of signals from a particular patient who has just been through surgery. This offers insights into the raw data. Then, to extract the raw signals, the records need to be fetched using another tool, an open-source Python library named dwclib. Historically, not all devices were recorded on the same infrastructure, notably the EEG - although its profound interest to the research community and under the spotlight of the latest developments in machine learning, which sensor does not fit the DWC acquisition format without an expensive module. This is why its acquisition channel is shown separately in the following diagram. The relational PostgreSQL database is illustrated by the timescaleDB, which intends to aggregate different sources of data to enable scientists to build cohorts efficiently. As we see, working with synchronized and matched physiological and clinical data (such as age, pathologies, and ID) is a non-trivial matter. The data collection is a complex process that is described in Figure 1.5, which illustrates the numerous technical steps and interconnections.

A strength of this acquisition infrastructure lies in the Kafka time series platform, which enables the deployment of models as APIs for real-time applications. This perspective will notably be discussed in chapter 6.

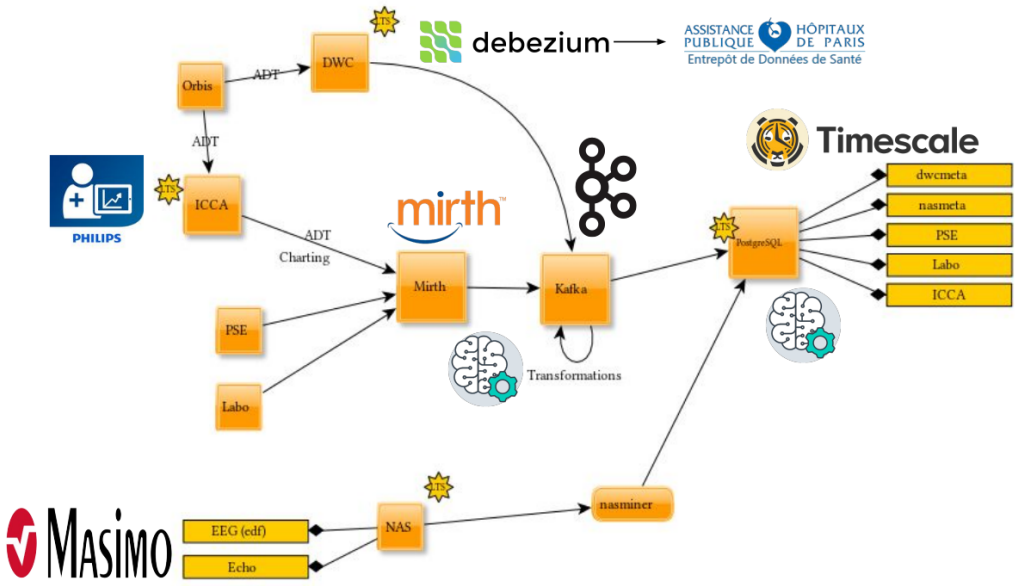


Figure 1.5: Physiological data platform at Lariboisiere

1.2.2 Challenges in cohorts generation

A fundamental distinction must be drawn between raw clinical data and curated research datasets. Using real-world data presents significant logistical and technical hurdles not typically encountered with standardized benchmarks, primarily due to issues of data integrity, infrastructure constraints, and computational complexity. Unlike preprocessed datasets, raw clinical data is characterized by extreme heterogeneity, requiring a granular, multi-stage integration and extraction protocol. In the absence of prior analysis, these datasets frequently include missing values and extraction failures that require extensive cleaning procedures. Furthermore, recording frequencies remain dependent on the specific sensors connected and their availability, requiring a robust strategy to standardize formats and address missing data constraints before the signals can be used for research.

These technical difficulties are compounded by strict medical data privacy protocols. Access is generally restricted to the hospital's internal network, which precludes remote connectivity and complicates collaboration with external partners, such as the National Institute for Research in Digital Science and Technology (INRIA). Additionally, restricted internet bandwidth within these secure environments poses a practical challenge when deploying large-scale software libraries, such as PyTorch. Beyond connectivity, the

volume of high-frequency waveforms presents a substantial computational bottleneck. The duration of GA records is highly variable—ranging from 2 to 155 hours—which often exceeds the memory capacity of high-performance local workstations, leading to systemic execution failures. Reliability is further challenged by discrepancies in the DWCmeta database, where the presence of an entry does not guarantee the existence of the corresponding signal or the accuracy of its labels.

To address these constraints, optimizing resource utilization through multi-core parallelization was critical for preserving the integrity and depth of the analysis. We leveraged Dask, a parallel computing library in Python, to distribute transformation operations across available CPU cores. This approach mitigates the risk of memory exhaustion by processing data in manageable chunks, thereby maximizing the allocation of computational resources. Furthermore, this transformation stage ensured dataset homogeneity and facilitated the selection of an optimal storage format, such as GPU-compatible tensors, to enhance efficiency during model training. By accelerating iteration speed and enabling complex analyses within fixed computational constraints, this parallelized framework played a pivotal role in maintaining both the performance and scalability of the research pipeline.

1.2.3 The challenge of data quality

Following cohort generation, a critical phase of data qualification was conducted to assess the viability of the extracted information. A significant discrepancy was observed, with a loss rate of up to 30% between the initial selection criteria and the actual existence of corresponding files. Identifying the root causes of this attrition required extensive reverse-engineering, revealing a range of issues from physical sensor disconnection to systemic failures within the data acquisition workflow. Furthermore, synchronizing data from disparate sources presented a substantial administrative and technical hurdle. Mapping patients to their clinical information was restricted by the availability of identifiers; specifically, such information could only be retrieved for the approximately one-third of patients who were formally identified during their time in the operating room. This identifier is essential for retrospective access to auxiliary medical databases—including electronic

health records and syringe pump injection logs—which are necessary to reconstruct a comprehensive overview of the clinical care provided.

Beyond administrative synchronization, the inherent nature of biosignals in an operative environment introduces significant physiological noise, which can compromise the accuracy of both real-time monitoring and subsequent analysis. This interference stems from biological variability, including respiratory sinus arrhythmia, motion artifacts from surgical manipulation, muscle contractions, and hemodynamic fluctuations. To mitigate these effects, a multi-stage filtering strategy was implemented. This included the application of Savitzky-Golay filters (Savitzky and Golay, 1964) to preserve signal morphology and key features while reducing high-frequency noise, as well as Butterworth filters (Butterworth, 1930), used for their maximally flat frequency response in the passband to isolate specific frequency components without signal distortion.

To ensure the reliability of the processed data, signal quality indicators (SQI), such as the signal-to-noise ratio (SNR), were developed to quantify signal integrity and flag unusable segments. These custom methodologies were integrated into a specialized Python library, physiocurve (Joachim, 2022), providing a standardized framework for generating metrics and ensuring the quality of the physiological signals used in this research.

1.2.4 Clinical data warehouse

While the on-site collection infrastructure leverages deployment tools that also facilitate application deployment, the integration processes were also inspired by the industrial procedures employed at the APHP clinical data warehouse (CDW). These are essential for making data available outside the hospital in an industrial, secure, and anonymised manner. Within the APHP health data warehouse, data integration, transformation, and enrichment rely on standardized frameworks to ensure interoperability and consistency. These follow worldwide recommendations regarding the standardization of clinical data warehouses (Khalaf Hamoud et al., 2018; Doutreligne et al., 2023). Among existing approaches, the FHIR standard is the most widely adopted, providing a common nomenclature and storage strategy that is

essential for interfacing heterogeneous tools developed by different medical technology providers and for enabling the deployment of new solutions. Because the integration of identifying health data raises significant privacy concerns, pseudonymization methods are systematically applied to safeguard patient information. Mature frameworks to store and transform voluminous data such as Spark or PyArrow were used to load data in packets according to indices, thus avoiding loading all the data into memory. Also, the integration pipeline is continuously monitored through platforms such as KafkaUI, which allow real-time supervision of acquisition processes. To streamline scalability and maintainability, data enrichment and structuring are further supported by automated workflows for metadata generation. These workflows rely on periodic computations that detect and register updates to newly written data, implemented through scheduled tasks (cron jobs) embedded in a continuous integration and deployment (CI/CD) environment. As part of our work in the warehouse, these CI/CD pipelines—using notably GitLab CI, Helm charts, and Grafana—have been deployed to reinforce automation and to ensure a robust and sustainable data infrastructure.

1.2.5 External open-source databases

The landscape of physiological data research benefits from publicly shared data from the operating room. These have also emerged over the last years thanks to the growth of storage capacities and the modernization of medical equipment. Several high-quality, open-access databases provide time series physiological data collected during anesthesia and surgical procedures. These databases are crucial for research in perioperative medicine, machine learning, and physiological signal analysis, notably in that they enable faster proofs of concept and multicentric studies.

Medical Information Mart for Intensive Care MIMIC

Created in 2016, MIMIC (Johnson et al., 2016), now renamed PhysioNet, is the oldest large-scale medical time series database, developed in North America from intensive care units (ICU) high-frequency records. It has been consolidated over time, leading to multiple developments that concentrate

and share medical time series. It includes biosignals from over 16,000 ICU stays across more than 9,000 adult patients, including some surgical and anesthesia cases. The dataset captures key physiological measurements such as heart rate, respiratory rate, SpO2, and arterial blood pressure at minute-level resolution. Recently, patient clinical information and laboratory results have been added too. While this dataset is publicly available, access requires proper credentialing to maintain patient privacy and data security standards. Recent individual datasets have been added to the renamed PhysioNet datamart, which increases the available data but also introduces data heterogeneity, and the entry cost to use this database is relatively high.

VitalDB

The Seoul National Hospital opened access to its 6,388 surgical cases under general anesthesia. To date, VitalDB (H.-C. Lee, Park, et al., 2022) is the largest open-source dataset of intraoperative and perioperative high-frequency biosignals. It includes over 486,451 waveform and numeric data tracks from 196 intraoperative monitoring parameters. Each case includes up to 12 waveforms and 184 numeric data streams. Data is acquired from multiple anesthesia devices: patient monitors, anesthesia machines, brain monitors, cardiac monitors, and infusion pumps. It regroups 73 perioperative clinical parameters and 34 time-series laboratory result parameters. It includes both low-frequency data referred to as numeric data (1–7 seconds period) and high-frequency waveform data (62–500 Hz). The access is public, anonymized, and available via a cloud and Python API. It has fostered anesthesia research, notably by enabling multicentric studies, and its easy access using common programming languages makes it a database of choice to test new methods in deep learning for arterial pressure, pharmacokinetic-pharmacodynamic studies, or intraoperative event prediction.

The different databases' information can be synthesized in the following table:

Table 1.1: Comparison of Larib, VitalDB, and MIMIC Databases

Database	Advantages	Drawbacks
Lariboisière (2018-today, France) 60,000 patients	• Increases clinical adoption and involvment • Possibility to access more specific patient context	• Unprocessed database • Restricted access for researchers • Dataset not public
VitalDB (2023, South Korea) 6,388 patients	• Anesthesia • Multi-modal data integration • Open access for academic use	• Some missing data • Limited long-term follow-up data
MIMIC / PhysioNet (2016 - today, USA) 50,000 patients	• Highly diverse patient population	• Requires credentialed access • Fragmentation of datasets • Mainly composed of ICU patients with fewer hemodynamic variations

1.3 Opportunities for machine learning in the operating room

The mandatory monitoring of GA has made it a safer procedure. But there is an extra opportunity; it also created large databases of monitored patients, which could be leveraged to further improve the GA safety through machine learning. This led to a very active field with various opportunities, which we describe in this section.

1.3.1 Leverage underutilised GA data for diagnostics

Latest studies analyzing the trends in anesthesia research exhibit the increasing prevalence of AI-related work, with a significant growth that began after 2019, shows an annual growth rate of around 75%. It has a notable impact on several areas of research that are gaining attention, including depth of anesthesia monitoring or hypnosis state classification (Aubin et al., 2023), risk assessment and fragility prediction, clinical efficiency, and education. This global research trend is associated with the improvement of local capacities for storing and analyzing data, as the volume of high-frequency data acquisition and storage has soared over the past years. As shown in Figure 1.6, the gap between labelled and unlabelled data is widening, which encourages the use of semi- or unsupervised learning methods.

A second major observation is that public health and postoperative care could benefit from this amount of perioperative data. The induction can be perceived as a stress test, characterized by the monitoring targets that anesthetists tend to reach. The nature and amount of the drug, along with the characteristics of the physiological signals resulting from it, form a piece of information on a patient's health. Outside the hospital, most of these vitals are rarely monitored, especially altogether, and are collected only in consultation with a specialist. Insofar as treatment exists and allows the patient to be referred to appropriate care services, it seems worthwhile to make the most of this readily available data recorded during GA to generate information on the patient's further care. In a nutshell, the rationale of the

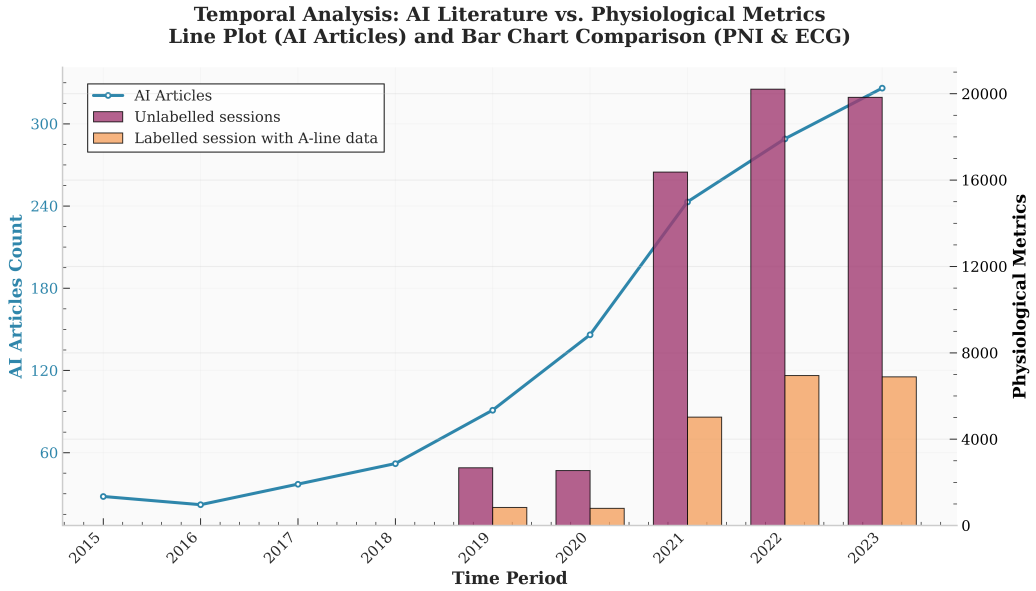


Figure 1.6: Volume of AI-related general anesthesia publications (Ou et al., 2025) and volume of high-frequency data storage in Lariboisiere

"Asleep medical appointment" is to position general anesthesia as a diagnostic service of preventive medicine.

As discussed above, general anaesthesia offers a unique opportunity to observe the human body's resistance to drug injection (Ayeb et al., 2025). However, once a resistance index – vascular, cerebral, etc. – has been quantified, what impact does this have on public health? It seems necessary to question the real benefit to the patient of such a diagnosis, and for long-term detection, to focus on pathologies that benefit from treatment or at least a therapeutic approach. For example, when it comes to heart problems, an ECG test is key to diagnosing atrial fibrillation. If left untreated—for instance, with anticoagulants—this condition can lead to severe complications such as heart failure. Furthermore, the shape of the pulse wave visible on the PPG waveform is characteristic of vascular stiffness, which is a cardiovascular risk factor that varies independently of other cardiovascular risks (Vlachopoulos et al., 2010; Pecha et al., 2023; Duan et al., 2022). Its prevalence among young people and adolescents is increasing year on year. It is now a new risk factor for hypertension and insulin resistance, and therefore diabetes, in adulthood, later in life. Diagnosis is to benefit from the latest advances in deep learning, given that randomized controlled trials prove their positive effect (Aasvang and Meyhoff, 2023). In this thesis, some general prediction

metrics will be used to characterize improved patient understanding for postoperative care. These will include the length of stay at the hospital, hypertension diagnosis, and will be discussed further in chapter 4 of the manuscript.

1.3.2 Building alternatives to expensive sensors

By marketing medical devices to hospitals, manufacturers in the biomedical sector are also attempting to address the issues faced by clinicians, notably regarding IOH prevention. However, their response, which prioritises commercial considerations, often seems insufficient. Biomedical companies also attempt to reduce the occurrence of hypotension episodes during GA. When an arterial line is deemed necessary and used, the American firm Edwards Lifesciences©markets the monitoring of a risk index measured on arterial lines named the Hypotensive prediction index (HPI). This index, based on invasive arterial waveform features, predicts hypotension defined as mean arterial pressure less than 65 mmHg for at least one minute (Maheshwari et al., 2020). Although randomized controlled trials have been conducted, Shirmohamadi et al., 2025 showed that despite its sensitivity and specificity in predicting hypotensive events, it may not improve cardiovascular or renal outcomes. When invasive lines are not placed, but the interval between cuff measurements is considered too long, other commercialized approaches, such as the Clearsight©(Y. Wang et al., 2022) or CNAP, propose a non-invasive estimation of arterial pressure based on the surveillance of the PPG. Still, the cost of such devices is a hurdle to their generalized usage across hospitals. The Clearsight©is intended for single use, has a maximal usage duration of three days, and costs around $90 - $120 for multiple monitoring metrics, including the cardiac output. Their single-use nature also raises questions about the impact of their widespread use from a sustainable development perspective. Some fairly new acquisition techniques based on resonance sonomanometry hold great potential for noninvasive and continuous blood pressure monitoring with minimal patient discomfort. However, they still require further clinical validation and regulatory approval before any adoption stage in anesthesia practice can be considered.

Although it seems highly unlikely that arterial catheters will be entirely replaced by non-invasive sensors as described above due to the risks involved, the latter can nevertheless be used as a complementary measure. In developing countries, more limited budgets mean that priority is given to essential expenditure, and invasive catheters, which require better-trained medical staff, and represent a significant cost of approximately 45$/intervention, are often not used. The ability to predict blood pressure may also have applications outside the operating room – as do commercially available products – for outpatient care or home monitoring. These multiple uses make non-invasive, continuous blood pressure monitoring a key area of development in the biomedical field.

1.3.3 Towards developing integrative biomarkers for predictive anesthesia

Data collected from the operating room is massive and originates from multiple sensors, which follow different acquisition frequencies and standards. One of the complexities of general anesthesia (GA) data stems from its inherent multimodality. On a daily basis, anesthetists follow numerous high-frequency values that vary simultaneously. Interpreting and synthesizing such a large amount of information is not an evident task and can benefit from machine learning and informatics. How can we learn from various sources? Theoretically, combining data from numerous modalities should be able to provide a more accurate, complete, and reliable understanding than could be achieved from any single source alone (S. Li and Tang, 2024). Multimodal data fusion aims to combine information from different sources to compensate for individual weaknesses and leverage complementary strengths. The process is typically categorized by the stage at which integration occurs (Castanedo, 2013).

Early fusion combines raw multimodal features at the input, capturing fine-grained interactions but requiring strict synchronization and suffering from high dimensionality. Intermediate fusion extracts modality-specific features first, balancing flexibility and correlation capture, though its performance depends on feature quality. Late fusion merges independent modality outputs, ensuring robustness and modularity but missing cross-modal interactions.

Modern data fusion relies heavily on deep learning to automatically learn both feature representations and the fusion process itself (Gao et al., 2020). This is intended to go beyond the fusion level and enable representation learning(Baltrusaitis et al., 2019). Notably, cross-modal attention is compelling. It enables the model to learn direct alignments between modalities. The Transformer architecture, built entirely on self-attention and cross-attention, is the current state-of-the-art for many complex fusion tasks. In this study, late fusion will be performed in chapters 2 and 3 while 4 will propose a comparative analysis when aggregating data from multiple sensors to enable both brain and vessels representations.

1.3.4 Towards closed loop drug delivery?

Considering that induction of anesthesia represents a reproducible phase to transition from a state of wakefulness to a state of hypnosis during surgery. This allows the deduction of haemodynamic or cerebral variables that implies that our models are capable of characterising the body's response to drug injections. It should be noted that this modelling can also be formulated in the opposite direction. Maintaining patients in a stable depth of anesthesia (DoA) during general anesthesia relies on the understanding of both the effects of the drugs on the body and the effect of the body characteristics on the drug assimilation. In other words, the process involves dovetailing pharmacokinetics, which refers to the effect of the body on the concentration of the administered drug, and pharmacodynamics, which is the study of how a drug affects an organism. These systems are intended to provide even more personalized care and control laws (Paolino et al., 2024). Studies have shown that closed-loop systems outperformed open systems (Casas-Arroyave et al., 2019; Y. Liu et al., 2015). An ultimate extension of the scope of the Asleep Medical Consultation described above would consist of a modelization of the patient using his vitals to predict the optimal drug dose to administer to ensure a safe anesthesia, along with a family of alerting algorithms to predict detrimental events during surgery. Closed-loop controllers are encountered in all aspects of modern life in applications ranging from air-conditioning to spaceflight. Although these systems are virtually ubiquitous, they are infrequently used in anesthesiology because of the complexity of physiologic

systems and the difficulty in obtaining reliable and valid feedback data from the patient. Despite these challenges, closed-loop systems are being increasingly studied and improved for medical use.

1.4 Background on sequence-to-sequence learning

The previous sections have given an overview of the operating room environment, key clinical questions and the data at stake along with its acquisition and exploitation hurdles. The next step is to use the tools at our disposal to analyze and understand this data, making the results of our analyses actionable and intelligible to the medical community. Generic learning challenges include, but are not limited to, forecasting and prediction, anomaly detection, classification and clustering, and representation learning. This section outlines methods for understanding time series and the families of learning paradigms that seem relevant. These will be developed and adapted to the specificities of the thesis question in the following section.

1.4.1 Multimodal autoregressive modelling

What will the weather be like tomorrow? Meteorology is often used as a characteristic natural phenomenon to illustrate autoregressive models. Numerous physical systems evolution depend on their state at the previous moment. Forecasting tasks have guided the development of time series modelling and its initial building blocks, which include autoregressive models. Considering a system X, several steps p back in time can influence such systems. To predict the next state of system X, the equation for an autoregressive model of order p, is given by:

$$X_t = \sum_{i=1}^{p} \alpha_i X_{t-i} + \varepsilon_t \tag{1.2}$$

where X_t is the value of the time series at time t, $a_1, \ldots, \alpha_p$ are the parameters of the autoregressive model, X_{t-i} are the past values of the series and ε_t is the white noise error term at time t. This modelling can be useful to

reconstruct time series that are known at previous timesteps.

As we have seen in previous sections, the signal from one sensor captures only a fraction of the patient's condition. Similar to neighbouring words in natural language processing, the various sources of signals and events recorded during anaesthesia make up the context of the models. Context refers to all the additional or environmental information surrounding a piece of data, task, or prediction that influences how a model learns, generates representations, or makes decisions. Taking context into account promotes robustness and personalisation of predictions, and is known to reduce errors. In our case, if we consider the prediction of our variable of interest X and make the prediction based on m other biosignals Y in the operating room, and a cuff oscillometer Z that provides a proxy for X, we can redefine the autoregressive problem of order p within the framework of multimodal autoregressive models using the following equation:

$$\hat{X}_t = \sum_{i=1}^{p} \alpha_i \hat{X}_{t-i} + \sum_{i=1}^{p} \sum_{j=1}^{m} \beta_i^j Y_{t-i}^j + \lambda Z_{t-\tau} + \varepsilon_t \tag{1.3}$$

To capture more complex non-linear relationships between modalities and improve the prediction of X_t (MAP), concurrent methods such as neural networks are a natural extension of this linear model. Convolutional architectures can exploit local temporal patterns, recurrent networks (RNN/LSTM) effectively model long-term temporal dependencies, and Transformers capture attention relationships between all modalities for more accurate prediction of the latent variable X.

1.4.2 The learning process of neural networks

Before describing core architectures, a few concepts regarding the learning process of neural network learning are necessary. The most common paradigm is supervised learning, where a neural network learns from a labeled dataset of size n, where each example includes both the input and the known target value to be predicted. Each feature is paired with its corresponding target y. The model learns from most of the dataset, called the

training set, and is evaluated on a previously unseen test set, which should characterise the model's ability to generalise. The quality and representativeness of the training set are crucial, as they directly influence the model's performance in capturing underlying patterns.

Learning in neural networks relies on stochastic gradient descent, a fundamental optimization algorithm that iteratively adjusts the model's parameters in the direction that reduces error. This process is guided by the loss function, which quantifies how well the model's predictions match the actual data. The goal of training is to minimize this loss, thereby improving the model's accuracy and generalization. Gradient descent is typically combined with backpropagation, a method that efficiently computes the gradients of the loss function with respect to each parameter. These gradients are then used to update the parameters, steering the model toward optimal performance.
In regression tasks, where the output is a continuous value, the choice of loss function is crucial. A straightforward approach is to minimize the mean squared error (MSE), which measures the average squared difference between the predicted values $\hat{y}$ and actual values y:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{1.4}$$

It is sensitive to outliers due to the squaring of errors. Despite the use of methods to reduce noise in the data, outliers are frequently encountered in the use of medical time series. A more robust approach to outliers than MSE is the mean absolute error (MAE), which computes the average absolute difference between predictions and true values.

$$\mathcal{L}_{\text{MAE}} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{1.5}$$

This metric will be used in chapters 2 and 3 as a central metric to discriminate models regarding arterial pressure estimation. To make the most of available resources, which often represent a bottleneck, gradient descent is usually performed over multiple samples simultaneously, which is called minibatch gradient descent. This approach leverages GPU parallelization for faster computation while reducing gradient noise. Modern optimizers also

incorporate techniques that make the training process less sensitive to the choice of hyperparameters — the fixed algorithm settings that are not learnt during training.

1.4.3 Architectures

Convolutional networks

Convolutional neural networks (CNNs) are a widely used architecture designed to learn hierarchical, nonlinear spatial patterns in grid-structured data such as images, time series, or signals. This is the primary method that has given neural networks their current popularity, and it remains the main component of deep learning literature, particularly in computer vision. Many of the pre-trained models used today are based on these operations. CNNs excell at tasks on 2D images such as handwriting recognition and categorization (Lecun et al., 1998). Time series such as GA signal data are the 1D equivalent to images, and have been the subject of numerous developments using CNN architectures, which captured non-inherent nonlinear patterns. Applying CNN to contiguous sequences of data integrates the immediate context of the target data into the calculation of the intermediate representation. The equation for a 1D convolution operation is given by:

$$S(i) = (I * K)(i) = \sum_{m} I(i - m)K(m) \tag{1.6}$$

where: - $S(i)$ is the value in the output feature map at position i.
- I represents the input sequence.
- K is the convolution kernel (or filter).
- i is the index of the output sequence.
- m is the index for the kernel K.
- The $*$ symbol denotes the convolution operation.

Considering contiguous segments, CNNs capture the context of fixed length limited by the length of the sequences. Specific implementations of CNN aim to increase the context size considered. These methods will be developed in chapter 3.

Recurrent neural networks

While convolutional filters enable learning from patterns present in the signal , they learn using a fixed context length. For this purpose, recurrent neural network (RNN) architectures are commonly used, as they enable learning from previous steps (Hopfield, 1982; Elman, 1990). Recurrent layers impose a structure whereby a layer receives both the preceding layer's activations and its own activations from the previous time step as input. RNNs integrate context sequentially and can memorize information that is more distant in time. Models composed of recurrent layers are thus encouraged to make use of the sequential structure of data, allowing them to achieve high performance in tasks such as natural language processing (NLP). The operations carried out by a vanilla Recurrent Neural Network (RNN) layer are governed by the following equations for the hidden state $\mathbf{h}(t)$ and the output $\mathbf{o}(t)$ at time step t:

$$\mathbf{h}(t) = g_h(\mathbf{W}\mathbf{h}(t-1) + \mathbf{U}\mathbf{x}(t) + \mathbf{b}_h) \tag{1.7}$$

$$\mathbf{o}(t) = g_o(\mathbf{V}\mathbf{h}(t) + \mathbf{c}) \tag{1.8}$$

where $\mathbf{x}(t) \in \mathbb{R}^{D_x}$ is the input vector, $\mathbf{h}(t) \in \mathbb{R}^{D_h}$ is the hidden state vector, and $\mathbf{o}(t) \in \mathbb{R}^{D_o}$ is the output vector. The weight matrices are $\mathbf{W} \in \mathbb{R}^{D_h \times D_h}$ (recurrent), $\mathbf{U} \in \mathbb{R}^{D_h \times D_x}$ (input), and $\mathbf{V} \in \mathbb{R}^{D_o \times D_h}$ (output), with bias vectors $\mathbf{b}_h \in \mathbb{R}^{D_h}$ and $\mathbf{c} \in \mathbb{R}^{D_o}$. The functions g_h and g_o are element-wise non-linear activation functions (e.g., $\tanh$ or ReLU). This structure enforces a critical form of **parameter sharing**, where the same parameters are reused across all time steps. A significant drawback of such architectures is that the input must be "unfolded" in time, leading to a **bottleneck** when the model attempts to compress a long sequence of data points into a single, fixed-size hidden state vector. This made it difficult for models to capture long-range dependencies and keep track of critical past events, leading to inaccurate forecasts for long sequences. Approaches based of gradient truncation to force constant error flows and drastically reduce training time were proposed in long-short term memory (LSTM) units (Hochreiter and Schmidhuber, 1997). Some approaches, such as using LSTM or hybrid modelling, will be explored in chapters 2 and 3.

Transformer-based architectures

Instead of relying on a single final hidden state, an attention mechanism allows the model to "look back" at all previous time steps and dynamically weigh their relevance when making a prediction. This is particularly useful for long-term forecasting. A key reference in this area is the work on temporal attention mechanisms, such as those used in models like the Informer (H. Zhou et al., 2021) and the Autoformer (Wu et al., 2022). These models use a form of attention to efficiently handle extremely long time series by identifying dominant time steps and dependencies, which has significantly improved performance in various long-term forecasting tasks (Bahdanau et al., 2016; Jianpeng Cheng et al., 2016).

Just as in NLP, self-attention has also proven highly effective for time series. This mechanism allows the model to learn relationships and dependencies between different steps within the same series, without needing a separate encoder-decoder structure. It can isolate complex, non-linear patterns, such as periodicities or seasonalities, by comparing each time step to all other time steps. This approach has led to state-of-the-art results in tasks like anomaly detection and imputation, where understanding the internal structure of the time series is critical. The Transformer architecture (Vaswani et al., n.d.), with its self-attention mechanism, has been successfully adapted for time series analysis, providing a powerful tool for learning complex dependencies and increasing the context size. Later developments in chapter 4 will propose some implementations.

The characteristics and usability of these architectures can be summarized in a table as follows:

1.4.4 Self-supervised paradigm

What if labelled data is scarce and insufficient to reach satisfactory performance? This is the case in many fields of application in the physical world. As we saw in section 1.3.1, arterial lines are also limited in the operating rooms. In these cases, other paradigms can be leveraged to address the scarcity of real-life labels. Self-taught learning, or self-supervised learning (SSL), is a paradigm in machine learning introduced to the community by

Table 1.2: Comparison of architecture types

Architecture type	Advantages	Drawbacks
CNNs	• Good at extracting local patterns and physical shapes • Parallelizable • Low computational cost	• Limited global context • Causality has to be enforced • Fixed-length kernels
RNNs	• Handle sequences of varying length • Retain information from previous time steps and capture long-term context with hidden states	• Sequential processing makes them slower to train • Limited parallelization
Transformers	• Self-attention mechanisms capture context across the entire sequence • Handle variable-length input and model complex interactions • Highly scalable	• May overfit small datasets • Computational cost

Raina et al (Raina et al., 2007), which has enjoyed growing success in recent years, outperforming supervised approaches in many computer vision tasks where the generation of labels is both time-consuming and cost-prohibitive, such as in medical imagery (S.-C. Huang et al., 2023). In this framework, the model learns in two steps. First, a *pretext task* is solved based on an auxiliary or pretext classification task using pseudo-labels, which help to initialize the model parameters. This enables the training of an embedder that captures information from the original signal and compacts it in an embedding or vector in the latent space, whose dimension is usually smaller than the original feature space. Next, the actual task called the *downstream task* is performed with supervised or unsupervised learning.

The substance often lies in the ability to define a pretext task that captures an underlying structure of the data that is closely linked to the downstream task. The aim is to reduce the problem to a known one and to cling to classic

and solved optimization problems. Two major approaches consist in data reconstruction and contrastive learning.

Reconstructing the data

In language and vision, auto-associative learning, or learning to reconstruct the data itself, achieved interesting results. Autoencoders are a category of neural networks that compress data into a latent space - using an encoder - and then reconstruct it - via a decoder. They have proven their efficiency in detecting anomalies, removing noise, or generating synthetic data. Among the reconstruction methods, one notable example is masking, or randomly hiding part of the signal while teaching the network to reconstruct it. Although this type of approach already existed for denoising autoencoders, its popularity exploded following its success in the field of natural language processing, with the performance of transformers architectures such as BERT (Devlin et al., 2019). Previous models were outperformed from 86% to 93,2% in F1-score on large text comprehension datasets such as the Stanford Question Answering Dataset. Masked autoencoders also proved their scalability in vision tasks and outperformed supervised methods on ImageNet (He, X. Chen, et al., 2021). When reconstructing the data - and for our case time series, the predictions are continuous values, and the model is trained using MSE loss applied to the masked patches. The masked autoencoder is trained in two stages. First, the encoder f_θ processes only the visible patches:

$$\mathbf{z} = f_\theta(\{x_j\}_{j \in \mathcal{V}}) \tag{1.9}$$

where $\mathcal{V} = \{1, \ldots, N\} \setminus \mathcal{M}$ denotes the set of visible (non-masked) patch indices. The decoder g_ϕ then reconstructs the masked patches from the encoded representation:

$$\hat{x}_i = g_\phi(\mathbf{z}, i) \quad \forall i \in \mathcal{M} \tag{1.10}$$

The reconstruction loss is computed only on masked patches:

$$\mathcal{L}_{\text{MAE}} = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \|x_i - \hat{x}_i\|_2^2 \tag{1.11}$$

This formulation forces the encoder to learn meaningful representations from visible patches that enable accurate reconstruction of masked content.

Numerous approaches adapted the masked autoencoding framework to time series using approaches that range from neural ordinal equations to point-level reconstruction (Z. Li et al., 2023; Q. Liu et al., 2025).

Contrastive learning

When learning a meaningful representation of data using an encoder, a robust approach is to adopt a geometric perspective in the embedding space. The goal is to pull similar data points closer together while pushing dissimilar points farther apart, effectively structuring the space for improved discrimination. This principle is particularly effective in multi-class settings, where the model learns to distinguish between classes by optimizing the relative positions of embeddings. Central to this approach is the definition of similarity: How do we determine which samples are "close" (positive pairs) and which are "distant" (negative pairs)? Contrastive learning addresses this challenge by explicitly contrasting positive and negative pairs, enabling the model to learn robust and discriminative representations (Oord, Y. Li, et al., 2019). By leveraging this paradigm, models can capture the underlying structure of the data, even in the absence of explicit labels.

The most widely used loss function in contrastive learning is the InfoNCE loss, which takes the form:

$$\mathcal{L}_{\text{contrastive}} = -\sum_{i \in I} \log \frac{\exp(z_i \cdot z_{j(i)}/\tau)}{\sum_{a \in A(i)} \exp(z_i \cdot z_a/\tau)} \tag{1.12}$$

where z_i represents the embedding of sample i, $j(i)$ denotes the positive sample for anchor i, $A(i)$ is the set of all other samples in the batch, and τ characterizes the sharpness of the probability distribution which is why it is called a "temperature" parameter, in an analogy to statistical mechanics.

For binary classification problems, training using InfoNCE loss may be considered excessive; self-supervised learning can also be performed using classical metrics for this well-documented category of subjects in machine learning (Ericsson et al., 2022; Assran et al., 2023; Banville et al., 2021; Guetschel et al., 2025).

These modelling methods will be explored in more detail in chapter 4 of the manuscript.

1.4.5 Explainability in machine learning

After learning is achieved and training has converged, it is necessary to get clinicians on board with the new technological solutions. Thus, it is necessary to explain the predictions of the various models. This section outlines several key features of our development that informed our work and were crucial for interdisciplinary communication. Visual representations held a central place.

Is it possible to reconstruct the learning process of a deep learning model? Explainable artificial intelligence (XAI) is a field of research that explores methods that provide humans with the ability of intellectual oversight over AI algorithms. It offers a technical and reproducible solution to it. Broadly speaking, this involves using reverse engineering methods to trace the learning path and ultimately identify features of interest or regions of interest in the initial features.
A unified approach used within the community is inspired by game theory. Indeed, in cooperative game theory, the Shapley value is a method for fairly distributing the total gains or costs among a group of players who have collaborated (Lundberg and S.-I. Lee, 2017). For example, in a team project where each member contributed differently, the Shapley value provides a way to determine how much credit or blame each member deserves. In the context of machine learning, the "players" are the individual predictors or variables in the model, and the "gain" is the total explained variance or predictive power of the model. By interpreting a model trained on a set of features as a value function on a coalition of players, Shapley values provide a natural way to compute which features contribute to a prediction or con-

tribute to the uncertainty of a prediction. When applied to deep learning models such as convolutional networks on time series data, Shapley values facilitate interpretability by identifying which specific regions in the temporal sequence most influenced the network's decision. For an input instance x, the Shapley value ϕ_i for feature i is calculated as

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(|N| - |S| - 1)!}{|N|!} \left[f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S) \right] \tag{1.13}$$

where N is the set of all features, S is a subset of features not containing i, and $f_S(x_S)$ represents the model prediction when only features in S are present. Visual explanatory maps can then be generated using gradient-weighted class activation mapping (Grad-CAM) (Selvaraju et al., 2017). In the time series context, this approach reveals how different temporal segments contributed to the prediction, enabling both local and global interpretability. By mapping the Shapley values onto the time axis, the intervals the temporal convolutional network relied upon most for its decision can be visually and quantitatively localized. The application of this method will be explored in chapter 2.

1.5 Objective of the thesis and contributions

1.5.1 Objectives

Today, anaesthetists lack tools for analysing and predicting perioperative risk. While doctors have an individualised understanding of their patients based on their experience, patient records and the anaesthesia process, the standards of care used by the community are still historically based on average values. These standards have changed little over the past few decades, largely due to the opacity maintained by biomedical manufacturers. This has long hindered the development and refinement of methods embedded in biomedical decision support tools. The current period marks a turning point in this sector, as the opening up of open-source databases to a large community of researchers is accelerating scientific advances in this field. Recent advances in data science and engineering are making it possible to take advantage of data generated passively during the perioperative period, in particular by saving vital signs collected during surgery. The objective of this thesis falls within the scope of precision medicine. It aims to use real-world clinical data to provide data-driven decision support to anesthetists, with a view to reduce postoperative risks. This thesis initially explores arterial pressure monitoring using non-invasive signals, specifically PPG and ECG. The work progresses step-by-step, culminating in the development of indicators to characterize frailty. The final section shifts focus to augmented depth of anesthesia monitoring, employing a self-supervised approach based on EEG signals and their features. The learning methods prioritize explainability and clarity to encourage discussion and adoption within the medical community.

Prior to the thesis, an initial mechanistic model had been developed by the clinical team to estimate blood pressure non-invasively and continuously from cyclical indicators present in the PPG signal. These developments led to the creation of several libraries for working on physiological signals in the operating room, such as physiocurve, graphysio and dwclib. With clinical intuition guided by observations in the operating room, these tools make it possible to immediately test an idea on a patient whose progress is known, thereby accelerating iterations. Despite developments concerning the health

data warehouse, the preliminary work preceding this study predated the implementation of the DWC collection structure. It was therefore based on a cohort collected outside the described collection circuits. The cohort for the first part of the study was constituted explicitly by the clinical team, including digestive surgery patients. These are known to experience significant haemodynamic variations and illustrate the problem of detecting pressure drops. Following the publication of external databases, we sought to incorporate other state-of-the-art models to establish a benchmark grounded in common data on this subject. We felt it was necessary to propose this approach because most other studies do not compare the same data, which obscures the comparison of models and advances in the field. Next, intending to model the body as a whole, as well as the interactions between vital organs such as the heart and brain, we considered self-supervised learning. Methods that have proven themselves in other fields, such as polysomnography, seemed interesting and allowed us to visualise the brain's response to anaesthetics. We then attempted to extend these results to other signals in the study, such as PPG and ECG, which necessitated adapting the methods and developing strategies for merging the different modalities.

1.5.2 Contributions

The thesis is organised as follows.
First, Chapter 1 - this chapter - places the thesis in its general context. It first presents and defines the ins and outs of general anesthesia, from usual sensors to the challenges met in modern medicine. Data routinely monitored in general anaesthesia, as well as the opportunities they represent for improving care, whether through the generation of new intraoperative biomarkers to aid decision-making in the operating room, or to deduce indicators of cerebral, cardiac, or overall patient frailty. The structure of operating room data is also detailed, as well as the steps required to generate usable datasets.
Chapters 2 to 4 follow the chronological order of the thesis and present the associated contributions.
Chapter 2 highlights the first contribution of the thesis. It develops physiological indicators of the photoplethysmography signal that are useful for learning, as well as a first version of a deep learning algorithm based on

convolutional and recurrent networks. The interpretability of the results will also be explored using heat maps illustrating areas of interest for the network.

Chapter 3 expands on the subject addressed in chapter 2, adding a multicentric dimension and comparing it to state-of-the-art models. It also includes the electrocardiogram, which, when combined with photoplethysmography, provides a more accurate understanding of the patient's cardiovascular variations.

Chapter 4 addresses the prevalence of unlabelled data in the operating room and attempts to respond to this issue through Frugal modelling based on self-supervised learning. Promising results have been obtained for the assessment of anaesthesia depth using electroencephalograms so that we will focus on the fusion of multimodal data and the adaptation of pre-training tasks to semi-periodic data such as the plethysmography.

Finally, chapter 5 succinctly describes the technical developments regarding the models real-time deployment.

1.5.3 Publications

The work presented in this thesis led to the following publications:

- **Jade Perdereau*, Thibaut Chamoux***, Etienne Gayat, Arthur Le Gall, Fabrice Vallée, Jérôme Cartailler, Jona Joaquim. *Blood Pressure Estimation Using Explainable Deep-Learning Models Based on Photoplethysmography*. Anesthesia Analgesia, 140 (1), 119-128
- **Jade Perdereau**, Jona Joachim, Fabrice Vallée, Jérôme Cartailler and Thomas Moreau. *Harnessing operating room signals to estimate Mean Arterial Pressure with AnesthNet.* Scientific Reports, July 2025
- **Jade Perdereau**, Virginie Loison, Fabrice Vallée, Thomas Moreau, Jérôme Cartailler. *Self-supervised learning to unveil brain representations during general anesthesia*. Manuscript in preparation, 2025

And the following work:

- **Oriane Saint Aubin*, Inès Khemir***, Jade Perdereau, Cyril Touchard, Fabrice Vallée, Jérôme Cartailler. *Repurposing the electroencephalographic signal for automatic segmentation of intra-operative periods under general anesthesia*. IEEE EUROCON 2023-20th International Conference on Smart Technologies, 286-290
- **Théophile Harlé**, Jona Joachim, Pierre Boutouyrie, Joaquim Mateo, Jade Perdereau, Alexandre Mebazaa, Jérôme Cartailler, Fabrice Vallée. *Continuous measurement of carotid-femoral pulse wave velocity (PWVcf.) during general anaesthesia using Doppler: a preliminary study*. Journal of Clinical Monitoring and Computing, 2025
- **Jona Joaquim**, Thibaut Chamoux, Jade Perdereau, Valentin Iovène, Jérôme Cartailler, Fabrice Vallée, *Integrating Heterogeneous Perioperative Data Sources: Platform Design and Implementation*. Submission in progress.

2

Physiologically inspired modelling based on the photoplethysmography

In the previous part, we have gone through the general framing of the technical environment of general anesthesia, and the associated opportunities. As previous work has shown, routine signals from the operating room inform on blood pressure variations. Indeed, and notably for the PPG, the signal generated by the optical sensor can be valuable for MAP estimation, as it characterizes the oxygenation of the blood in the tissues. In this chapter, we will explain the rationale for using such features, and combine them using a mechanistic approach based on physiology. Then, we will challenge its performance using deep convolutional networks. This chapter led to the publication of the article entitled: "Blood Pressure Estimation Using Explainable Deep-Learning Models Based on Photoplethysmography".

2.1 Introduction

2.1.1 Physiological rationale

An initial observation reveals that synchronously monitored signals, since they describe the same system, can serve as predictive indicators for one another. In this chapter, we focus on the PPG and notably on some hand-crafted features derived from the PPG raw signal, and we show that they are informative on blood pressure variations.

Dicrotic Notch

The first PPG characteristic that comes to mind when we think of hemodynamics is the dicrotic notch. It is a slight, downward deflection on the descending limb of the photoplethysmography (PPG) waveform, occurring just after the systolic peak. It corresponds physiologically to the closure of the aortic valve, marking the end of systole and the beginning of diastole in the cardiac cycle. The notch results from the brief backflow of blood as the aortic valve shuts, producing a pressure wave that is reflected in both arterial blood pressure and PPG signals. The position and prominence of the dicrotic notch provide information about arterial compliance, vascular resistance, and cardiac function. Since the notch marks the transition from systole to diastole, its timing and amplitude are influenced by changes in arterial pressure and vascular tone. Variations in the dicrotic notch's location and depth can reflect changes in mean arterial pressure (MAP), as these are affected by stroke volume, peripheral resistance, and arterial stiffness—key determinants of MAP. Various approaches intend to detect or use it for analysis purposes. On the one hand, it corresponds to the abscissa of the second derivative of the signal, when it exists. On the other hand, a more precise method resides in the intersection of the tangent to the PPG signal at the systolic peak and the baseline of the PPG signal.

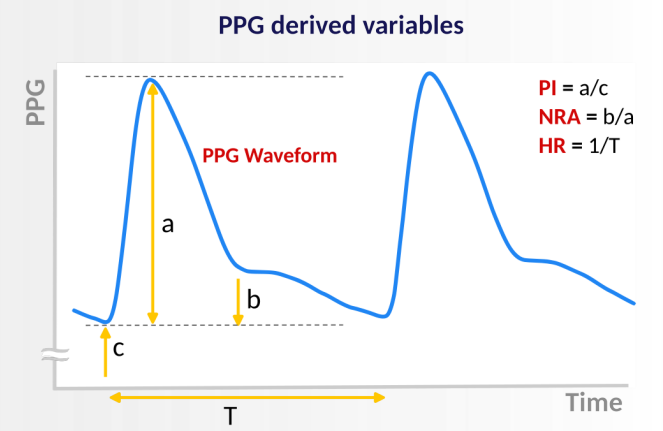


Figure 2.1: Physiologically informative features on the ppg signal

Perfusion index

Another key metric from the PPG signal is the perfusion index (PI), which is a numerical value derived from the PPG signal, defined as the ratio of the pulsatile (AC) component to the non-pulsatile (DC) component of the PPG waveform. It reflects the strength of the arterial pulse relative to the overall blood flow and tissue perfusion at the measurement site. PI provides an indirect measure of peripheral perfusion and vascular tone, both of which are influenced by mean arterial pressure. Changes in MAP can lead to vasoconstriction or vasodilation, altering the amplitude of the pulsatile component and thus the PI. A decrease in MAP typically reduces peripheral perfusion, lowering the PI, while an increase in MAP can enhance perfusion and raise the PI. PI is sensitive to acute hemodynamic changes and is used in clinical settings to monitor circulatory status, making it a valuable feature for estimating MAP variations.

2.1.2 Mechanistic Model

From correlation to estimation, how to bridge the gap ? In fact, the association of the previously defined metrics can be used complementarily to develop a mechanistic model as a white-box counterpart. The rationale for such modelling lyies on the electric analogy of blood vessels. In a simplistic approximation based on The Ohm-Poiseuille law applied to blood vessels suggests that MAP is linearly associated with both heart rate (HR) and stroke volume. Specifically,

$$MAP = SV * HR * R \tag{2.1}$$

Where R is the systemic vascular resistance, and the product represents the cardiac output.

Theoretically, notch relative amplitude (NRA) is linearly related to R, while PI is physiologically linked to SV. However, PI often demonstrates abrupt, non-linear variations in response to MAP fluctuations, which can introduce noise or disproportionate influence in predictive models. To mitigate this

effect and better capture the nuanced relationship between PI and MAP, we introduce the transformed feature

$$mPI = ln(1 + \frac{1}{PI}) \tag{2.2}$$

This logarithmic scaling attenuates extreme values, stabilizes variability, and aligns more closely with the gradual physiological changes observed in MAP, thereby improving the robustness and interpretability of our predictions. Using these considerations, we can use PI and NRA to build an MAP estimator:

$$\widehat{MAP} = \sigma\left(mPI + \beta\right)\left(NRA + \beta'\right)HR + \delta \tag{2.3}$$

Or in its polynomial form:

$$\widehat{MAP} = a1 + a2 \circ HR + a3 \circ HR \circ NRA + a4 \circ HR \circ mPI + a5 \circ HR \circ mPI \circ NRA \tag{2.4}$$

The polynomial coefficients can then be estimated using regression methods using machine learning libraries such as scikit-learn.

2.1.3 Concurrent studies

The accurate and continuous estimation of MAP from non-invasive measurements remains a significant challenge. The photoplethysmography signal has shown to carry information on MAP variations (Allen, 2007; Nafisi and Shahabi, 2018). Our prior research indicates that variables derived from the PPG signal routinely collected, such as dicrotic notch relative amplitude or perfusion index, are valuable for reconstructing MAP (Coutrot et al., 2019; Joachim et al., 2020). This suggests that a mechanistic or a deep-learning model approach could be deployed to rebuild a continuous MAP from these variables. Deep-learning models are particularly well-suited as universal approximators, which can be leveraged to transform PPG inputs into continuous MAP data. These models provide a robust framework to handle the complexity associated with multimodal data (S. Lee et al., 2021; Y. Zhou et al., 2022). Specifically, they can integrate time-series- specific layers that can capture patterns within a complex multi-dimensional signal (Fritz et al., 2019). Additionally, architectures such as convolutional layers (CNN) and

Long Short Term Memory (LSTM) units (Dahl et al., 2012), have shown to outperform single-structure models (Belur Nagaraj et al., 2020). These models are becoming increasingly appealing in the field of clinical science as methods to address the critical issue of explainability emerge. In this study, we develop both mechanistic and deep-learning models to reconstruct continuous MAP for comparison with data from invasive arterial lines. We propose a deep learning architecture that incorporates CNN and LSTM layers, utilizing the PPG signal and derivatives as inputs. Explanability will be explored using Grad-CAM, a visualization technique that locates the input section used in the decision- making process (Selvaraju et al., 2017). We shall adhere to international standards of the field to assess models' performance (Advancing Safety in Medical Technology (AAMI), 2013; Williams et al., 2004). Our method aims to estimate MAP using routine monitoring signals, offering an affordable, eco-friendly, and universally applicable solution for continuous, non-invasive blood pressure monitoring in real-time.

2.2 Methods

2.2.1 Patients

This study was approved by the Société de Réanimation de Langue Française (SRLF) research ethics committee under the reference number CE SRLF 14-34. All analyses were conducted on a biometric research database containing patient monitor data from the Lariboisière Hospital operating rooms, in accordance with the MR-004 methodology and French law 3rd May 2018. Patients were provided with an information letter, and verbal consent was obtained from each patient before anaesthesia. Patient were recruited from the Lariboisière Hospital for two sets of procedures: neuroradiology interventions from January 2015 to September 2017, digestive surgery procedures in the Department of Surgery from January 2018 to September 2018. Patients were eligible for inclusion if they were over 18 years old and required invasive blood pressure monitoring. The analysis was performed on anonymized patient data. Patients were excluded if they had less than 30 minutes of data, a median arterial pressure below 40 mmHg, no 10 consecutive minutes window of arterial pressure data without missing values, or more than 50for

PI. These criteria resulted in a cohort of 85 neurosurgery-related patients for the training and validation set, and 32 patients from digestive surgery for the test cohort. Figure 2.2 presents the flowchart of patient selection.

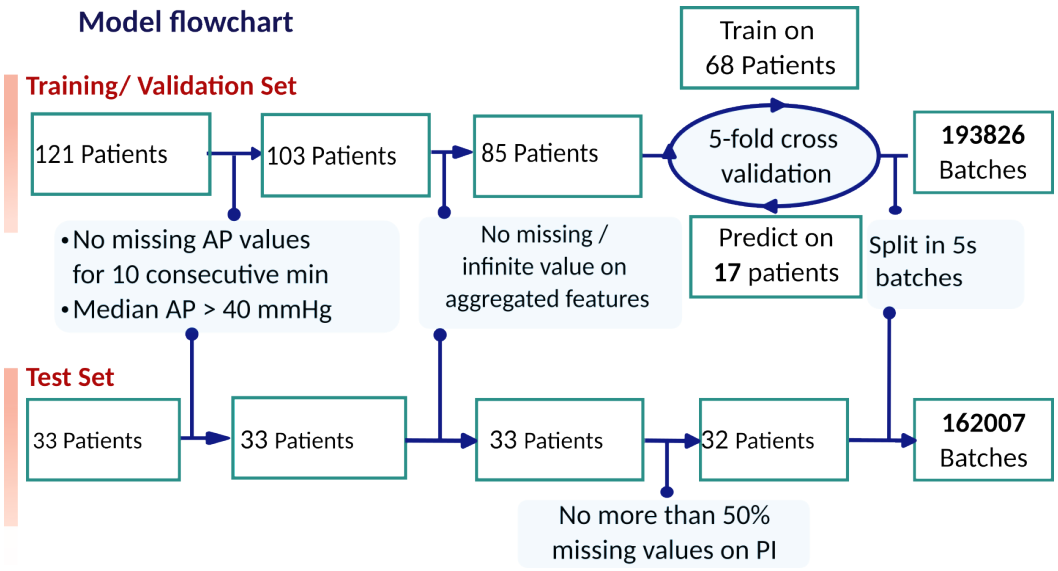


Figure 2.2: Patients inclusion diagram

2.2.2 Anesthesia Protocol

General anesthesia was induced and maintained using target-controlled infusion of Propofol and Remifentanil, adjusted to maintain a bispectral index between 40 and 60. All patients were intubated after induction of neuromuscular blockade using Atracurium besilate (0.5 mg/kg), then mechanically ventilated with a tidal volume of 6 to 8 ml/kg, and a respiratory rate adjusted to maintain end-tidal CO2 between 35 and 38 mmHg. Ventilation parameters were adjusted based on arterial blood gas analysis. A radial or femoral artery catheter was inserted after anesthesia induction and before skin incision. Dilute Norepinephrine tartrate at a concentration of 5 g/ml was used as a standard vasopressor, administered by infusion or bolus using a syringe pump.

2.2.3 Data

Data acquisition

Patient biometric data was monitored using a Philips IntelliVue MX850 bedside monitor (Philips, Eindhoven, The Netherlands), which was connected to the Data Warehouse Connect storage solution from the same company. This setup allowed for real-time, automated database storage of data from the patient monitor for subsequent offline analysis. Data was recorded at a sampling rate of 1 Hz for low-frequency data and 125 Hz for high-frequency waveform data.

Data features

This study used only non-invasive monitoring methods to collect features. The features included oscillometrically measured cuff arterial pressure, recorded every five minutes. High frequency Photoplethysmography (PPG, 125Hz) signal was measured using a finger pulse oximeter optical sensor. From the PPG waveform, three additional features were derived using a custom open-source library (Joachim, 2022) : dicrotic notch relative amplitude (NRA), which reflect the vascular resistance; perfusion index (PI), which indicate tissue thickness and is defined as the ratio between the PPG pulsatile wave and the non-varying wave; and heart rate (HR) characterizing the duration of one cardiac cycle (see Figure 1B). One value was obtained per cardiac cycle. Each model developed in this study used either some or all of these features. The MAP, measured every second via a catheter, served as the gold standard for validation (ground truth). Diastolic and systolic blood pressure values were not consistently available and PPG-derived features offer limited insight into these parameters: their estimation was not conducted in this investigation.

Data preprocessing

Intervals containing infinite or negative values in any waveform features were removed. Missing heart rate values were generated from the high-frequency PPG waveform, and we excluded data points that fell outside the

range of three standard deviations from the mean. To reduce acquisition noise and smooth the signal, we built a custom filter for arterial pressure, using a successive application of a Savitsky-Golay filter (order 4; 81 points). Additionally, acquisition noise in numeric features was also filtered using a combination of third order Butterworth and Savitsky-Golay filters. All derived features underwent standardization by subtracting the mean and dividing by the standard deviation. For the Deep Learning models, predictions were based on the preceding 5 seconds of the signal. Specifically, the PPG waveform signal, acquired at a 125Hz, was segmented into batches of 625 (5x125) points each, corresponding to the 5 seconds preceding the MAP value to be predicted. Waveform features were similarly splited into batches of five consecutive values to predict the subsequent MAP value.

2.2.4 Deep Learning Model

Model architecture

We developed a deep learning model by increasing architecture complexity iteratively, evaluating various combinations of input features. Intermediate models incorporated part of the model layers tailored to the associated inputs: LSTM for PPG-derived features and CNN for the PPG waveform. The optimal architecture using a multi-threaded CNN + LSTM model - architecture combines convolutional blocks (O'Shea and Nash, 2015) for PPG waveform, LSTM layers (Dahl et al., 2012), and Linear layers for output (see Figure 2.3). The convolutional blocks are designed to extract valuable information from the PPG waveform shape, while the LSTM layers preserve temporal patterns from previous data points. Training hyperparameters are detailed in Supp. Table ST1. The main results are presented on the test cohort using the model trained on the whole training cohort of 85 patients. Predictions on the validation set were also generated using a k-fold (k=5) cross-validation method on the train/validation cohort. These results are gathered in Supplementary Table ST2. We used Tensorboard (Wongsuphasawat et al., 2018) as a visualization tool to monitor the training process and gradient descent. The model was implemented using the Pytorch framework and trained on a NVIDIA Quadro RTX4000 at Lariboisière Hospital.

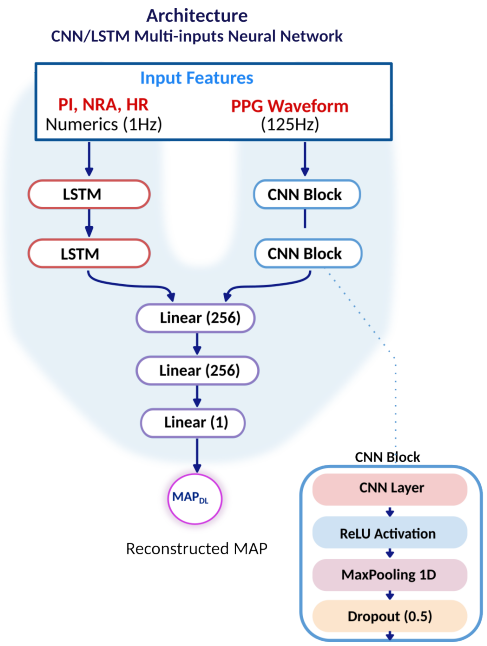


Figure 2.3: CNN + LSTM model architecture

Calibration

The cuff measurements are an opportunity to calibrate predictions according to the ground truth variations at runtime. The model calibration is performed by subtracting the difference from the error calculated at the last cuff measure timestep. The model predicted values are therefore calibrated as follows:

$$\Phi\left[X\right](t, calibrated) = \Phi\left[X\right](t) - \left(Y\left[t_0\right] - \Phi\left[X\right](t_0)\right) \tag{2.5}$$

where X represents the features which depend on the estimator, $Y\left(t_0\right)$ corresponds to the cuff oscillometer value at time t_0, $\Phi\left[X\right](t_0)$ is the prediction of the model at time t_0 . For each patient, when the cuff oscillometer value was not available immediately, the first initial calibration is performed using the initial value of the arterial line to which a Gaussian noise of zero-mean and a standard deviation of $\sigma = 2$ mmHg was added, which corresponds to the observed bias in cuff values.

2.2.5 Naïve Model

To assess the performance of deep-learning and mechanistic models against a reference, we constructed a cuff model that consists solely of cuff pressure measurements taken every five minutes. This model provides continuous values at 1 Hz, which are defined by the most recent cuff pressure measurement. This step interpolator requires no training. Missing cuff values were imputed from downsampled invasive pressure, to which a Gaussian noise of zero-mean and a standard deviation of $\sigma = 2$ mmHg was added, corresponding to the observed bias in cuff values. Study designs, the different models, and features are shown in Figure 2. More specifically,

$$\widehat{MAP}(t) = Y(t_0) \tag{2.6}$$

at time t and t_0 corresponding to the last cuff value.

2.2.6 Statistical Analysis

To assess agreement between models' predictions and MAP from the arterial line, we used the mean absolute error (MAE), mean error (ME), and the root mean squared error (RMSE). The distribution around the mean values was quantified using the standard deviation of the error. The standard deviation was used as a deviation metric for MAE and ME. Correlation between the model's output and the gold standard was assessed using the repeated measures correlation library rmcorr. Bland-Altman analysis for repeated measures was performed using a mixed-effects model, with the random effect accounting for patient-specific variability in the mean difference. Then, subgroup analysis was performed regarding patients' gender, age over 65, and pre-existing hypotension using the Bland-Altman representations. Bias was defined as the mean difference between measurements, while the upper and lower limits of agreement were set at ± 1.96 standard deviations of the bias. To enhance readability and accommodate the substantial volume of data, we downsampled the plots to represent one point for every five data points and applied a transparency coefficient of 0.1 to the scatter plots. Furthermore, to evaluate the explainability of the Deep Learning model, we generated a

saliency map using Grad-CAM. This heatmap highlights the intensity of the neural network attention on the PPG waveform signal to indicate regions with the most influence.

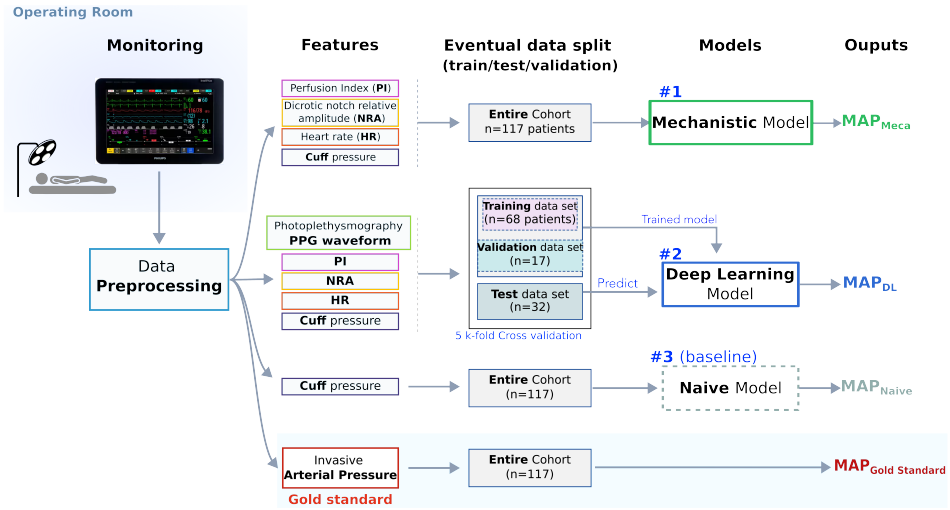


Figure 2.4: General study design and features for cuff, mechanistic, and deep-learning models.

2.2.7 Visualization of results using Bland-Altman representations

The Bland-Altman plot is a graphical representation widely used in scientific research to assess the agreement between two different quantitative measurement methods. Unlike a simple correlation plot, which only shows the relationship or association between two variables, the Bland-Altman plot focuses on the differences between the measurements. It is constructed by plotting the mean of the two measurements on the x-axis and the difference between the two measurements on the y-axis. This visualization is handy for identifying systematic bias (a consistent positive or negative difference) and for determining the range of agreement, typically represented by 95% limits of agreement. The plot thereby provides a more insightful and clinically relevant assessment of method agreement than a correlation coefficient, as it clearly reveals whether one method consistently overestimates or underestimates the other and whether the differences are clinically significant.

2.2.8 Mixed-effects modelling to handle correlated data

From a statistical point of view, time-series data points correspond to frequently repeated measures from a set of patients. The conventional representation of medical measurement methods often involve the use of Bland-Altman plots, which put individual measurements and the reference measurement into perspective. This method had to be adapted to take into account the fact that the data were initially correlated. To do this, we used a mixed model, which incorporates a fixed effect,t but also a random effect to take into account the fact that the measurements have a hierarchical structure as opposed to a random variable. Fixed effects encapsulate the tendencies that are consistent at the levels of primary interest. These effects are considered fixed because they are non-random and assumed to be constant for the population being studied.

Random effects introduce statistical variability at different levels of the data hierarchy. These account for the unmeasured sources of variance that affect specific groups in the data. In matrix notation a linear mixed model can be represented as

$$y = X\beta + Z\alpha + \epsilon \tag{2.7}$$

Where: y is a known vector of observations, with mean E (y) = X β ;

- β is an unknown vector of fixed effects;
- α is an unknown vector of random effects, with mean $\mathbb{E}(\alpha) = 0$ and variance–covariance matrix var$(\alpha) = G$;
- ϵ is an unknown vector of random errors, with mean $\mathbb{E}(\epsilon) = 0$ and variance var$(\epsilon) = R$;
- **X** is the known design matrix for the fixed effects, relating the observations **y** to β;
- **Z** is the known design matrix for the random effects, relating the observations **y** to α;
- The random effect α accounts for patient-specific variability in the mean difference.

2.3 Results

2.3.1 Patients

The train/validation cohort included 85 patients (mean age (std) 54 (± 6) y.o., 61% female) and the test cohort included 32 patients (mean age (std) 56 (± 5) y.o., 60% female), with a total record duration of 494 hours. After segmenting the signal, we obtained a median of 760 samples (IQR: 512 to 4108, total: 193825) of 5-second duration in the training set and a median of 4778 samples (IQR: 2709 to 6463, total: 162007) in the test cohort. Procedures ranged from intracranial surgery to bowel surgery. A large proportion of patient presented comorbidities such as arterial hypertension (37%, respectively 38% in the test cohort), tobacco use (39%, respectively 58% in the test cohort), dyslipidemia (17%, respectively 25% in the test cohort), and diabetes (10%, respectively 22% in the test cohort). MAP values in the train/validation and test sets were 75.1 (± 8.1) mmHg and 72.2 (± 11.1) mmHg, respectively. Cuff pressure distribution sensibly followed that of MAP. Table 2.1 provides a description of the population demographics, PPG derived features (PI, NRA, HR), and MAP values for the training/validation sets before normalization.

2.3.2 Models performance

We compared model performance relative to the arterial line gold standard. First, the Cuff model, serving as our baseline reference, achieved an MAE of 6.1 (± 14.5) mmHg. The Mechanistic model reached an MAE of 9.1 (± 12.2) mmHg. For the deep-learning models, the architecture based on CNN alone achieved an MAE of 5.3 (± 4.7) mmHg, whereas the model using solely LSTM on derived features obtained an MAE of 8.7 (± 8.4) mmHg. The deep-learning model combining both CNN and LSTM achieved the best accuracy with an MAE of 3.5 (± 4.4) mmHg and the lowest RMSE of 5.5 mmHg. Simpler deep-learning models using only one type of neural network showed poorer performance than the CNN + LSTM model. Hereafter, the term deep-learning model will refer to the CNN + LSTM architecture. Table 2.2 gathers the results for all models obtained on the test cohort. We extracted

Characteristics	Training/validation set	Test set
5-s samples (n)	193,826	162,007
Patients (n)	85	32
Demographic characteristic		
Age, mean (std), y	54 (± 6)	56 (± 5)
Gender, male n (%)	34 (43)	12 (39)
Weight, mean (std), kg	72 (± 6)	69 (± 4)
Height, mean (std), cm	165 (± 2)	161 (± 4)
BMI, mean (std), $kg{\cdot}m^{-2}$	24.5 (± 3.1)	26.1 (± 3.4)
ASA score classification n (%)		
I	14 (17)	6 (19)
II	58 (69)	21 (65)
III	11 (14)	4 (12)
IV	0 (0)	0 (0)
Comorbidities n (%)		
Age >55	34 (41)	11 (36)
Arterial hypertension	31 (37)	12 (38)
Tobacco use	33 (39)	18 (58)
Dyslipidemia	15 (17)	8 (25)
Diabetes mellitus (type 1 or 2)	9 (10)	7 (22)
Cardiovascular event	5 (6)	0 (0)
Procedures n (%)		
Intracranial surgery	9 (11)	8 (26)
Neurovascular embolization	62 (73)	6 (18)
Bowel major surgery	13 (16)	22 (68)
Features distribution before standardization		
Perfusion index	3.4 (± 3.1)	3.8 (± 3.3)
Dicrotic notch relative amplitude	0.18 (± 0.4)	0.26 (± 0.3)
Heart rate	71.3 (± 10.3)	78.3 (± 16.6)
Arterial pressure, mm Hg	75.1 (± 8.1)	72.2 (± 11.1)

Table 2.1: Summary of patient characteristics and features

the cumulative percentages of errors of the different models to determine quality grades based on standards from the British Hypertension Society (BHS). We used 5, 10, and 15 mmHg thresholds to define grade A, B, and C, respectively. Only the CNN + LSTM reached a grade A, while Cuff and Mechanistic models reached grades B and C, respectively. The Cuff model results were significantly better in the test cohort. Results are summarized in Supplementary Table ST2. Supplementary Table ST3 gathers the results on the train/validation cohort of 85 patients for each model. cohort.

Table 2.2: Model performance metrics

Models	MAE (±STD)	ME (±STD)	RMSE
Cuff reference model (cuff only)	6.1 (±14.5)	−1.8 (±16.1)	15.9
Mechanistic model (derived features)	9.1 (±12.2)	−0.2 (±12.4)	13.9
LSTM (derived features)	8.7 (± 8.4)	4.8 (± 9.4)	10.6
CNN (high-frequency features)	5.3 (± 4.7)	2.7 (± 6.9)	6.7
CNN + LSTM (all features)	3.5 (± 4.4)	−0.1 (± 6.2)	5.5

2.3.3 Repeated measurements correlation and Bland-Altman analysis

We compared the agreement between models and gold standard measures using repeated measures correlation and Bland-Altman plots. The deep-learning model achieved a correlation coefficient of 0.86 ($p<0.001$) value with gold standard, compared to 0.58 and 0.48 for the cuff and mechanistic models, respectively (both $p<0.001$). Bland-Altman bias analysis revealed that biases and 95% limits of agreement between predictions and gold standard were -0.05 mmHg (-12.7 to + 12.6 mmHg) for the deep-learning model (see Fig. 2.5 A,B). Of note: the biases were larger for Cuff and Mechanistic models: 1.61 mmHg (-38.1 to + 34.9 mmHg) and 0.41 mmHg (-23.2 to + 24.0 mmHg), respectively. The corresponding results on the test cohort are available in Supplementary Figure SF1. Considering demographic subgroup analysis (see Fig. 3 C, D, E, F, G, H), results showed that the Deep Learning Model achieved biases and 95% limits of agreement between predictions and gold standard of 0.02 mmHg (-13.4 to 13.4 mmHg) for hypertensive patients as opposed to -0.1 mmHg (-12.1 to 11.9 mmHg) for non hypertensive patients. Regarding ageing, the biases and 95% limits of agreement were - 0.12 mmHg (-13.6 to 13.4 mmHg) for patients over 55 years old and 0.1 mmHg (-11.0 to 11.2 mmHg) for patients under the age of 55. Finally, they were of -0.04 mmHg (-13.1 to 13.0 mmHg) for female patients as opposed to -0.08 mmHg (-11.9 to 11.8 mmHg) for male patients.

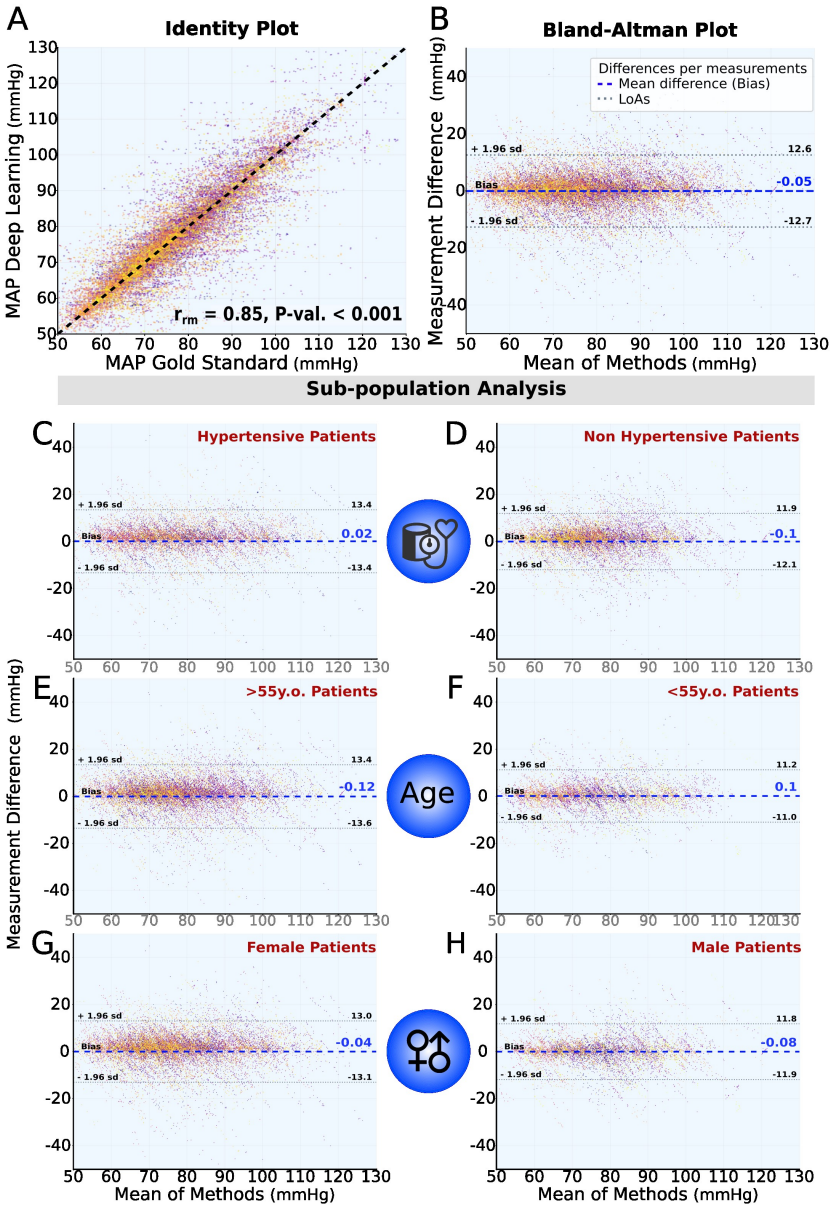


Figure 2.5: Correlation plots and Bland-Altman representations. The first row corresponds to the whole test set. Then, each row displays a comorbidity. From top to bottom, the plots correspond to: the whole test set, hypertensive vs non-hypertensive patients, patients aged above or below 55 years old, and men versus women.

2.3.4 CNN + LSTM model explainability

Within the CNN+LSTM deep-learning model, the PPG waveform is fully processed by the CNN layers, which automatically extract, in a black-box fashion, informative patterns to predict MAP. To explain this decision-making process within CNN, specifically to determine and inspect the region of the PPG signal most impactful on MAP prediction, we generated a saliency map (see Figure 2.6 and Material and Methods). The figure reveals that the most informative region is concentrated around the dicrotic notch, notably in the first and last cycles.

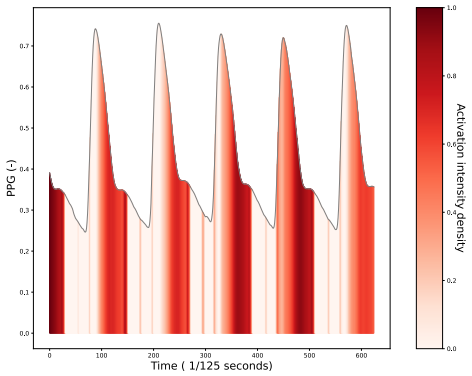


Figure 2.6: Saliency map. Deep learning model first convolutional layer saliency map for 1 batch of 5 s of raw PPG waveform measured at a 125 Hz frequency for the test cohort. PPG indicates photoplethysmograhy.

2.4 Discussion

2.4.1 Main findings

This study demonstrates that continuous, non-invasive mean arterial pressure monitoring in the operating room is feasible. The present work uses only data that are routinely and mandatorily collected during surgical interventions. Among several model tested, the best results were achieved using a CNN + LSTM deep-learning model. This approach capitalized not only on specific features computed from the PPG waveform but also on the raw

waveform itself. Such model reached a grade A equivalent according to the BHS standard and proved robust to external validation. Using saliency map method allowed for assessment of the deep-learning decision-making process. In the context of arterial pressure monitoring, these results follow and confirm previous work which demonstrated the performance of NRA and PI in estimating arterial pressure (Joachim et al., 2020). The CNN + LSTM model prediction mean error of 0.1 (± 6.2) mmHg on the test cohort is comparable to other PPG-based studies performance (Joung et al., 2023; Yilmaz et al., 2023) using fewer training samples, even though these studies were focused on systolic and diastolic blood pressure estimation. Subgroup analysis on patients demographics showed that the model performance appears to be robust even for different subgroup populations. Supplementary Figure SF3 also shows that the Mechanistic and Deep Learning Models tend to get better pressure estimates between cuff measurements, notably shorter delays detecting pressure drops. Furthermore, it should be noted that the present approach differs from existing deep-learning- based IOH detection approaches that, contrary to our approach, rely on invasive monitoring features (S. Lee et al., 2021). The model explainability approach revealed that the deep-learning model appears to extract most of its information from the PPG waveform around the dicrotic notch area – the only derived feature which can be spatially localized using the methodology. This approach in interesting for several reasons: first, it confirms that the neural network is selecting regions clinically recognized as relevant. Additionally, it shows how the model recreates feature-like inputs from the waveform in a data-driven fashion, making them less arbitrary than those designed by humans. The possibility to visually represent, explain, and interpret the model's predictions is a key feature to clinical acceptability for decision support, and mirrors explainable models developed in other clinical areas (Molenaar et al., 2024). Surprisingly, the mechanistic model achieved a poorer performance on the digestive cohort. This could be explained by the complexity of digestive surgical interventions, notably resulting in temperature variations. The mechanistic model being highly sensitive to the perfusion index values, and the perfusion index being dependent on temperature, this could explain the results difference for this model. On the contrary, the Cuff model reached a lower MAE on the digestive cohort. This is notably due to the use of more recent measuring devices in this test cohort, resulting in fewer cases of missing data interpolation.

2.4.2 Limits of the mechanistic approach

Considering that we proposed both a mechanistic and a deep-learning modelizations, one could argue that the drawback of the mechanistic model is that it is too simplistic. A more refined approach than the simple Ohm-Poiseuille law is the Windkessel model. Although the Ohm-Poiseuille law provides a useful basis by assimilating pressure, flow and resistance, it does not take into account the elasticity of blood vessels, a crucial factor in pulsatile blood circulation. The Windkessel model, introduces this elasticity by assimilating the cardiovascular system to a more complex electrical circuit. It uses resistance to represent peripheral vascular resistance and capacitance to model the compliance (elasticity) of large arterial vessels, particularly the aorta. This approach provides a better approximation of pressure and flow dynamics throughout the cardiac cycle, notably differentiating between systolic pressure and diastolic pressure.
Other limitations to the study include the fact that it was conducted in a single center, and its results should be validated in other university hospital centers through follow-up studies. Moreover, our study was retrospective and did not involve real-time display. Unfortunately, a key feature of our model, the perfusion index, though routinely used in the operating room and available on most PPG devices, is not recorded in open source databases such as MIMIC III and VitalDB (H.-C. Lee, Park, et al., 2022; Johnson et al., 2016). For broader applicability, our results would benefit from expending the cohort to include different medical procedure categories. Additionally, building our models, neither the ECG raw signal nor drugs administration time series were available. Integrating the ECG waveform with the PPG waveform could have improved our results (B. Huang et al., 2022), especially if associated variables like pulse transit time were available (Kim et al., 2013). Furthermore, from a clinical standpoint, relying solely on MAP as a hypotension marker may not provide sufficient information regarding its underlying causes. Therefore, when feasible, hemodynamic monitoring (e.g., cardiac output, stroke volume) should be used when possible to guide treatment and reduce the risk of vasopressor overtreatment.

2.4.3 Perspectives

For future investigations, we plan to develop a hybrid model that will include deep learning layers enhanced with mechanistic intelligence, aiming to leverage the strengths of both approaches, for instance using neural ordinary differential equations or physics informed neural networks (R. T. Q. Chen et al., 2019). Combining this approach with Bayesian neural networks could also be beneficial in the assessment of model uncertainty (Shen, L. Liu, et al., 2022). In line with the frugal design of these models, and the existence of input signals during GA, we plan to test the display of their predictions in real-time at the ICU. Near real-time decision support systems based on notifications in the operating room have shown to reduce gaps in blood pressure monitoring and provide insights for adapting anesthetic procedures (Nair et al., 2014). To conclude, we have built and compared two fundamentally different approaches to estimate MAP non-invasively at the ICU using real-world data. The CNN + LSTM model outperformed the cuff-based MAP estimator. This model was trained on the PPG signal which is always available in the operating room and can be used by any physician around the world without the need for additional sensors, while its explainability fosters clinical adoption.

2.5 Chapter conclusion

In this chapter, we have seen how physiologically insightful features and deep learning could be combined to propose an efficient estimation of blood pressure at the operating, based exclusively on routine monitoring. A natural step further is to extend this work to external datasets, and assess its generalizability. This is the subject of the next chapter, where we will use the VitalDB database to evaluate the model performance on a large cohort of patients whilst specifying the model architecture.

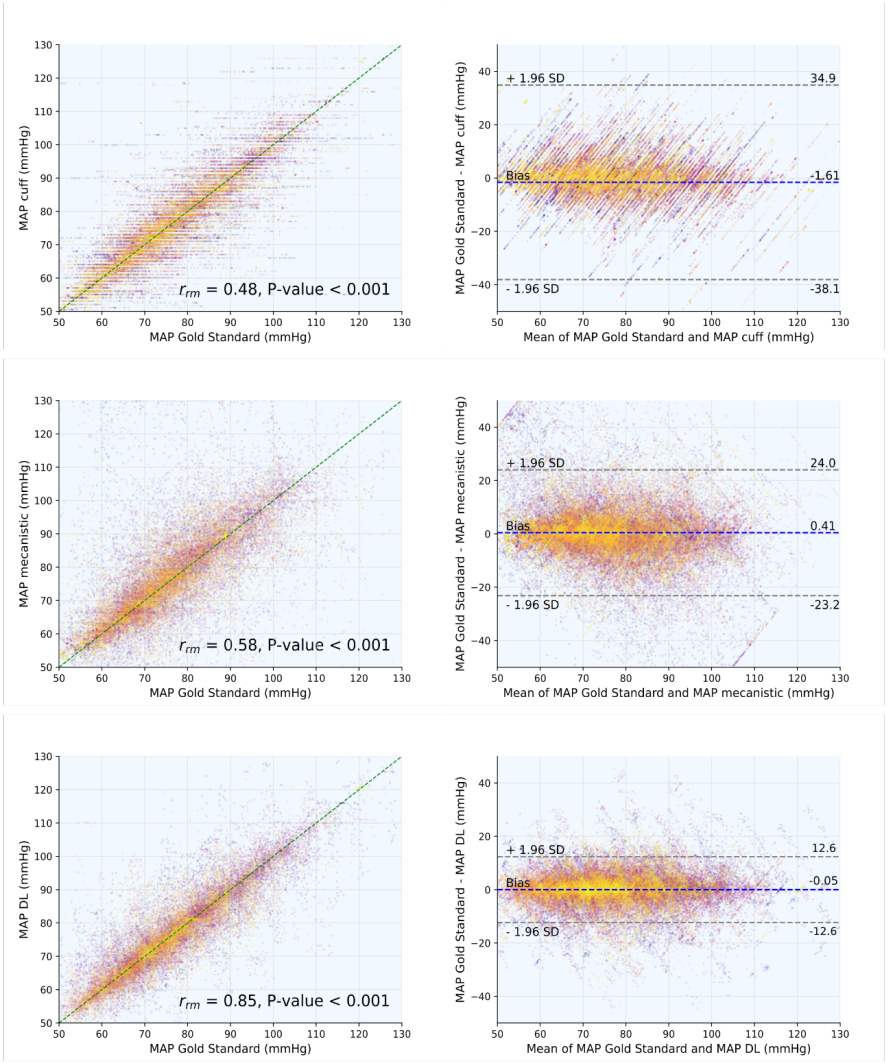
2.6 Appendix

Parameter	Default Value	Search Space
N_patients	117	-
N_epochs	10	[1, 2, 5, 10, 20]
Kfold	5	[3, 5, 10]
Shuffle	False	[True, False]
Beta	0.3	[0.2, 0.3, 0.4, 0.5]
Early Stopping Patience	3	[1, 3, 5]
Learning_rate	0.001	[0.01, 0.001, 0.0001, 0.00001]
Batch_size	8	[4, 8, 16, 32, 64, 128]
Optimizer	RMSProp	[Adam, RMSProp, SGD]
LSTM_size	16	[8, 16, 32, 64, 128]
Dense, Conv Layer initializer	He	[None, zeros, He]
LSTM Layer initializer	GlorotUniform	[GlorotUniform, He, Uniform]

Supplementary Table ST1: Deep Learning Model hyperparameters. This table details the values tested and then chosen for each hyperparameter of the Deep Learning Model. The values were tested using grid-search.

Models	$<$ 5 mmHg	$<$ 10 mmHg	$<$ 15 mmHg	Grade reached
Cuff Model	71.2%	77.5%	88.1%	B
Mechanistic Model	53.1%	72.6%	91.5%	C
CNN + LSTM	77.2%	91.5%	96.7%	A
BHS standard definition				
Grade A	60%	85%	95%	-
Grade B	50%	75%	90%	-
Grade C	40%	65%	85%	-

Supplementary Table ST2 : Models cumulative error performance on the test cohort regarding the British Hypertension Standard. The cumulative percentage errors below 5 mmHg (resp. 10, resp. 15 mmHg) correspond to the standards of the British Hypertension Society (BHS) for BP measuring devices. The grades associated with the proportion of values respecting this standard are also displayed below the models results.



Supplementary Figure SF1: Cuff, Mechanistic and Deep Learning models patient-level performance with repeated measures correlation and Bland-Altman plot for repeated measures on the test cohort. For lisibility purposes and because of the high number of individual measurements, a transparency coefficient of 0.1 was applied to the scatter plots. The CNN + LSTM reached the highest correlation results along with the lowest confidence interval

Model Description	**MAE (± STD)**	**ME (± STD) (mm)**	**RMSE**
Cuff Reference Model (Cuff only)	8.9 (± 11.7)	-0.8 (± 16.8)	11.7
Mechanistic Model (Derived features)	6.2 (± 8.7)	0.1 (± 14.1)	11.3
CNN + LSTM (All features)	2.5 (± 3.9)	-0.1 (± 4.9)	3.7

Supplementary Table ST3: Models performance on the validation set averaged by fold. For the CNN + LSTM model, the results were obtained by 5-fold cross-validation, ie by training the model on 80% of the patients and predicting on the 20% remaining, for each of the five splits. Therefore, predictions are always performed on unseen data. The results are obtained by concate- nating the predictions on all five validation folds. Values are presented as mean values and standard deviation of the error. The best results are obtained for the CNN + LSTM model.

3

Benchmarking models on common datasets

In the previous chapter, we proposed an approach based on physiology and a mechanistic model to estimate mean arterial pressure (MAP) from non-invasive operating room data. It showed satisfying results regarding MAE estimation performance, using the international standards of the field. However, one could argue that it relies on a single monocentric cohort and that performances were not compared to preexisting state-of-the-art models. Most clinical studies in the field compare their findings to those of other datasets using the reported metrics. To account for these aspects, the next chapter will develop a benchmark of some best-performing and easily reproducible state-of-the-art architectures for MAP estimation. This chapter led to the publication entitled: "Harnessing operating room signals to estimate Mean Arterial Pressure with AnesthNet".

3.1 Introduction

3.1.1 Why use ECG and PPG together?

As we have seen over previous chapters, maintaining appropriate blood pressure is essential to ensure adequate oxygen delivery to tissues (Wesselink et al., 2018; Bijker, Klei, et al., 2007). Previous studies have demonstrated the potential of non-invasive vital signs measurements to estimate continuous blood pressure variations, primarily aiming to enable remote monitoring aid for hypertensive event detection. Notably, blood oxygenation measured at the fingertip through photoplethysmography (PPG) has been associated with blood pressure (Allen, 2007; El-Hajj and Kyriacou, 2021) and used to estimate it continuously (Perdereau et al., 2025). PPG-derived markers were

shown to offer valuable insights for estimating continuous MAP (Joachim et al., 2020). However, the generalizability of such models on external and larger cohorts remains challenging. Other studies considered predicting the BP from the central measure of the electrocardiogram (ECG). However, ECG alone was not deemed a viable standalone option for blood pressure estimation (Finnegan et al., 2023) and should be used in conjunction with other modalities. Indeed, the simultaneous monitoring of ECG and PPG provides a measure of the central waveform (ECG) and the radial waveform at the fingertip. Theoretically, this process enables the deduction of the time taken by the wave to reach the periphery of the body (here, the fingertip), also called the transit time (TT), as illustrated in Figure 3.1. Using this time, and as the distance travelled is constant, the speed of the wave can then be inferred. This methodology provides an equivalent of the measure of the pulse wave velocity (PWV), which is a well-known marker of arterial stiffness and has been shown to be associated with blood pressure. Its reference measurement method is based on peripheral arterial tonometry, a technology that measures changes in blood flow and arterial pulsatile volume in the fingertips. The time taken by the wave to reach the periphery is defined as the Pulse Arrival Time (PAT), is the time taken by the wave to travel from the heart to the fingertip. It is linearly associated to the TT by the relationship : $PAT = TT + PEP$ where PEP is the pre-ejection period, defined as the time between the onset of the QRS complex and the opening of the aortic valve. Studies have shown varying degrees of correlation between PAT and BP, with some finding strong correlations, particularly for systolic BP changes(Ben-Shlomo et al., 2014), while others have noted that the relationship can be complex and affected by factors like the pre-ejection period of the heart (Harlé et al., 2025).

3.1.2 Arterial pressure

While systolic and diastolic blood pressure are routinely monitored in outpatient settings, anesthesiologists rely on mean arterial pressure (MAP) as the standard metric for intraoperative blood pressure management. International consensus guidelines recommend maintaining a MAP above 60 mmHg, particularly in high-risk patients (Saugel et al., 2024). Moreover,

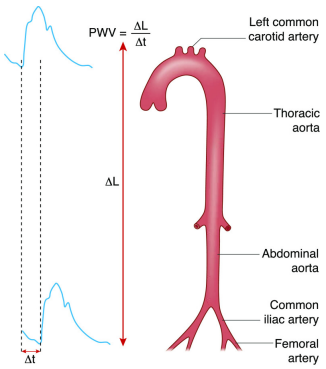


Figure 3.1: Pulse wave velocity explanatory diagram

owing to MAP linear definition from systolic and diastolic pressures, and recent findings showing they share redundant information content under intraoperative conditions (Cartailler et al., 2024), the scope of this study will be focused on MAP estimation. As for calibration, brachial cuff measurements are part of the operating room's routinely acquired vitals. Considering the drifts observed for cuff-less blood pressure predictions (Schutte, 2022), operating-room applications strongly benefit from leveraging all available data, notably calibration using cuff oscillometry.

3.1.3 Existing methods

Reconstructing blood pressure from other vital signs requires capturing their nonlinear relationships. Many methodologies have been explored, including features-to-signal and signal-to-signal study designs. Features-to-signal methods are based on hand-crafted time-domain and frequency-domain features, measured or calculated on the waveform signal. They reduce the problem complexity by extracting key features from the vitals time series to predict blood pressure. They have mostly been developed with small-scale datasets ($n < 100$) using classical signal processing and machine learning techniques, such as spectral analysis (X.-R. Ding et al., 2016), wavelet transforms (Attivissimo et al., 2023), or random forest regression (Yilmaz et al., 2023). With the emergence of larger datasets, with thousands of records, deep learning-based methods have become increasingly popular, with various architectures proposed to tackle the signal-to-signal classification, based on class of neural networks such as multi-layer perceptron (Bousefsaf et al., 2022; B. Huang

et al., 2022), convolutional neural networks (CNN) (Juan Cheng et al., 2021; Athaya and Choi, 2021), recurrent networks such as long short term memory (LSTM) and GRU (K et al., 2021), up to U-shaped architectures such as Wave-U-Net (Miyaguchi et al., 2021). These methods leverage the increased sample sizes to learn models that better generalize.

3.1.4 Operating room constraints

While previous studies have explored arterial pressure estimation using physiological data, many have overlooked critical operating room constraints. Two fundamental challenges persist: maintaining causality and incorporating calibration effectively. First, causality must be preserved—a requirement violated by architectures like bidirectional LSTMs (El-Hajj and Kyriacou, 2021; H. Xiao et al., 2022). Although causal convolutional neural networks (e.g., TCNs) (Dai et al., 2024; Zabihi et al., 2022) have been employed, they often lack long-term memory and fail to integrate physiological priors such as calibration. Second, while cuffless architectures have gained traction, their performance degrades significantly on large datasets or during hemodynamic fluctuations. Unlike wearable applications, where calibration is impractical, anesthesiologists require precise estimates leveraging all available intraoperative data. However, prior attempts to incorporate calibration often compromised causality. For instance, PPG2BP-Net (Joung et al., 2023) uses random segment selection and future data access, rendering it unsuitable for real-world clinical use. Moreover, at inference time, the arterial line might be unavailable, questioning the deployability of such a solution. Additionally, methodological inconsistencies across studies, such as comparisons on disparate datasets, hinder objective performance evaluation. To address these limitations, this study introduces AnesthNet, a deep learning architecture designed for real-time continuous MAP estimation in operating room settings. The key contributions of this work include: 1) the use of dilated causal CNNs to expand the receptive field and capture long-range dependencies while mitigating short-term memory limitations, 2) the integration of cuff-derived MAP values as initialization parameters to incorporate clinical knowledge into the model, and 3) rigorous benchmarking on standardized datasets including the open-source VitalDB dataset and our in-house

LaribDB dataset, comprising over 10,000 intraoperative recordings. This custom dataset notably contains valuable physiological features such as the brachial cuff oscillometry or the perfusion index, which are missing from open source datasets. Performance was assessed using established international standards from the British Hypertension Society and the Association for the Advancement of Medical Instrumentation (AAMI).

3.2 Methods

3.2.1 Approval for Data Acquisition

The VitalDB dataset was downloaded using the provided Python library (H.-C. Lee, Park, et al., 2022). Data derived from the MIMIC-III matched subset are released under the Open Database License (ODbL), while data derived from the VitalDB dataset are released under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) license.

As for LaribDB, it is a retrospective cohort study, conducted using the Assistance Publique-Hopitaux de Paris (AP-HP) Health Data Warehouse ("Entrepôt de Données de Santé (EDS)," https://eds.aphp.fr/). This data warehouse contains electronic health records (EHRs) of all inpatients from the 39 greater Paris University hospitals. The Institutional Review Board of the AP-HP Health Data Warehouse approved this study on April 2nd, 2024, under the number CSE-24-05_PLETHOMAP. This cohort included 10117 patient records in the Department of Anesthesiology and Critical Care at Lariboisière University Hospital from January 2018 to September 2023, stored at the Assistance Publique-Hopitaux de Paris (AP-HP) Health Data Warehouse deployed at the Lariboisière Hospital. Patient biometric data was monitored using a Philips IntelliVue MX850 bedside monitor (Philips, Eindhoven, The Netherlands), which was connected to the Data Warehouse Connect storage solution (Philips, Eindhoven, The Netherlands). This allowed for real-time, automated database storage of the data from the patient monitor for offline analysis. Data was recorded at a sampling rate of 1 Hz for low-frequency data and 125 Hz for high-frequency waveform data.

3.2.2 Data selection/Pre-processing

Datasets generation

The patients relevant to this study from the VitalDB database were selected based on the simultaneous presence of PPG, ECG, HR and ABP signals. The records were generated using the Python vitaldb library and relevant features such as the notch relative amplitude and cycle MAP were calculated from ABP using the open-source physiocurve library. After the preprocessing, data was saved as patient-level `PyTorch` tensors to maximize GPU usage and enable parallel processing using the `joblib` library.
The Lariboisière dataset was retrieved from the Data Warehouse Connect using the `dwclib` Python library, and its initial format was Parquet. Considering the high variability of patients' records length, notably for intensive care data, some patient records reached over 100 gigabytes of data and would not fit in memory. To address such memory issues, `Dask` was used for chunk data reading, processing, and rewriting. The use of chunk data read resulted in a reduction of peak memory of 68% and an increase of memory efficiency - defined as the ratio of the actual data size to the peak memory usage - of 170%.Abnormal records were defined as records for which either of the features was unavailable, and the minimal patient record length was set to one hour.

Derived features generation and missing values

Derived features were generated from raw waveform data. The dicrotic notch calculated on the PPG waveform was generated using the open-source Python library `physiocurve`[1]. The VitalDB dataset lacked certain features, specifically the perfusion index and brachial cuff measurements. To retrospectively generate the missing cuff values, the Lariboisière Dataset was utilized to determine the distribution of the error between cuff measurements and the corresponding arterial line values. This discrepancy was then applied as a Gaussian noise (mu=0, sigma=3.1) to the associated arterial line value, resulting in the generation of the missing cuff values.

[1] `https://framagit.org/jaj/physiocurve`

Unified quality metrics generation

Quality metrics on patients' records were defined and calculated to assess data quality at a segment level between two brachial cuff measurements. These metrics included the mean, variance, and extremum values for each of the derived features. The corresponding table was generated and saved locally. Abnormal segment removal was then performed using simultaneous selections for each feature, and abnormal segments were eliminated considering the following criteria:

T1 : 50 mmHg $<$ segment MAP mean $<$ 150 mmHg
T2 : MAP dispersion within segment $<$ 40
T3 : MAP gradient within segment $<$ 2 mmHg/s
T4 : MAP segment standard deviation $>$ 1
T5 : MAP segment missing proportion $<$ 20 %
T6 : PI missing proportion within segment $<$ 20 %
T7 : PI unique values within segment $>$ 5
T8 : HR missing proportion within segment $<$ 20 %
T9 : PPG missing proportion within segment $<$ 1 %

The models requiring the perfusion index feature were adapted to bypass its absence using model-specific datasets written in the Pytorch tensor format. Model-specific data loaders were defined to generate segments of input sliding windows fit for each model architecture.

3.2.3 Study design: Comparison to SOTA architectures

State-of-the-art architectures from the literature were compared to estimate mean arterial pressure. Depending on the models' architecture, model features included high-frequency data but also derived features, depending on the associated models' definition. The work presented here differentiates between calibration and calibration-free approaches. Whenever available, the code provided by the associated authors was directly used. Table 3.1 details the features and input window length for each model.

Physiologic approach based on derived features exclusively

- **Mechanistic model**: This model (Perdereau et al., 2025) is a parametric model based on derived features alone. It is based on the Ohm-Poiseuille law applied to blood vessels, which states that MAP is linearly associated to HR and stroke volume (SV), such that $\widehat{MAP} = \alpha \left(\ln\left(1 + \frac{1}{PI}\right) + \beta\right)(NRA + \beta')\, HR + \delta$. It proved to achieve good results in a preliminary cohort. As one of its main features - the perfusion index - was unavailable in the VitalDB cohort, its results are reported on LaribDB exclusively. It is a parametric model that requires no training.

Calibration-free approaches

- **PulseNet**: This calibration-free model was excerpted from Wang et al (W. Wang et al., 2023) and required no adaptation as its inputs are solely composed of raw windows of PPG and ECG sampled at 125 Hz.
- **Wave-U-Net**: Instead of predicting derived mean arterial pressure directly, the performance of the best-performing model was assessed on signal generation. The mean value per segment was generated a posteriori. This architecture was chosen for its results in signal-to-signal analysis. It is a calibration-free architecture.

Calibration-based approaches

- **PPG2BP-Net**: As it was presented in the introduction section, some of these models were not designed for operating room purposes and did not fit the associated criteria. To overcome this, slight adaptations were introduced to the network's definition, notably the patient-specific SBP and DBP calibration segments for the PPG2BP-Net model (Joung et al., 2023) were replaced. Two calibration configurations were tested: the replacement with the patient's initial calibration measure and the last cuff measurement during the intervention. The associated results are available in the supplementary material. The model architecture was reproduced and kept unchanged.

- **AnesthNet**, the proposed model, whose characteristics are detailed in the following section.

The study design is detailed in Figure 3.3. Model-specific dataloaders were developed to address the varying length, feature shape, and availability. For all models, a parameter search was performed using Optuna. Models were implemented using the Pytorch framework and trained on the APHP Clinical DataWarehouse infrastructure using a Nvidia GPU.

3.2.4 AnesthNet model characteristics

To address the specificity and challenges in operating rooms, a novel deep-learning architecture was developed to estimate MAP. A major operating room characteristic is the discontinuous measure of arterial pressure using the cuff oscillometer. Thus, the objective was to interpolate the values of arterial pressure between cuff measurements. To this end, input windows were designed as segments of data between two cuff measurements.

The proposed multimodal network takes three types of input features: high-frequency waveforms from the PPG and ECG, derived cyclic features, and cuff oscillometry. Figure 3.2 illustrates the integration strategy for these three inputs, detailed below. As for layer choice, CNN and LSTM building blocks were used to benefit from their complementarity: the facility to optimize CNNs and the unbounded context of recurrent networks. Cuff oscillometry, which measures mean arterial pressure, was used as calibration. A CNN architecture was used to process the high-frequency waveform data; more details are available in the following paragraph regarding causal convolutions. Cyclic derived features were resampled at a 1Hz frequency and concatenated with the CNN output of the same length on the channels dimension after interpolating to preserve the time window length. The resulting tensor was processed using LSTM layers, with the inner states initialized using the segment corresponding cuff value. More details on this part of the architecture are available in the corresponding section.

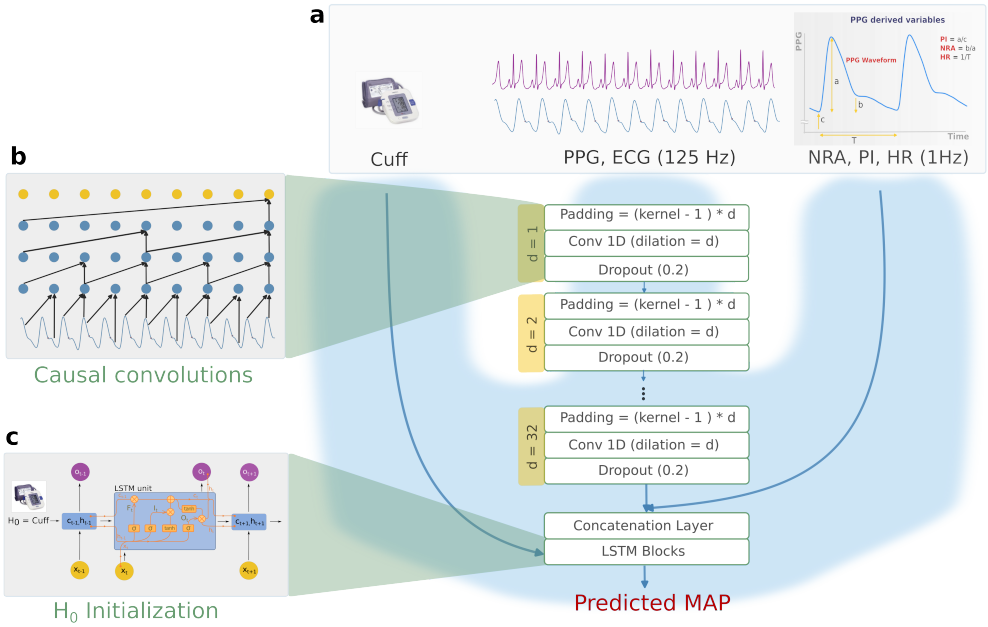


Figure 3.2: Proposed deep learning model architecture. **a**. Multi-modal features are shaped as sequences of inputs between two cuff measurements. They include the cuff oscillometer recorded every five minutes, the high-frequency ECG and PPG (respectively acquired at 500Hz and 125Hz and downsampled to 125Hz), and PPG-derived features which carry physiological information about MAP variations: the Notch Relative Amplitude (NRA), Perfusion Index (PI), and Heart Rate (HR). **b**. High-frequency signals are processed using a convolutional network with increased dilation to maintain causality throughout the network while increasing the receptive field within the sequence without increasing the number of parameters. The output of this network is then concatenated on the channels axis with the low-frequency derived features (NRA, PI, HR) to feed a Long Short Term Memory network. **c**. The LSTM network is initialized using the cuff value as the initialization parameter. The output sequence corresponds to the predicted MAP sequence between the cuff measures.

Causal Convolutions

To guarantee the feasibility of inference at runtime, and for physiological consistency, the convolutional network was built using causal convolutions, *i.e.*, using only past information to make predictions. This concept was first introduced for speech and audio generation to capture long-term dependencies (Oord, Dieleman, et al., 2016). Inspired by deep generative models,

causal filters with dilated convolutions allowed the receptive fields to grow exponentially with depth, which is crucial to modeling the long-range temporal dependencies observed in physiological records. Because sequences were relatively long (>30,000 points), dilated convolution aimed at maximizing the memory effect on PPG waveform sequences. Dilation was increased by a factor of two for each layer. A left-side padding equal to (k - 1) * d was used where i corresponds to the layer depth within the convolutional network. Following a hyperparameter grid search, the convolutional layers' parameters were a kernel size of 15 and a stride of 1, followed by a dropout of 0.2. Linear layers were not used as they do not preserve causality.

Calibration through LSTM initial inner state h_0

Recurrent layers, such as LSTM, account for the observed features' continuity and enable us to learn across segments. Input channels from the CNN and derived features were stacked before being given as input to an LSTM network. As the h_0 parameter in LSTM corresponds to the initial hidden state of the network, the inner state used during initialization determines its initial values. Instead of using default random or zero initialization methods, conventionally used when training LSTM networks, these terms were replaced with the segment cuff value.

Optimizing training to avoid I/O Bottleneck

Datasets were stored as batched arrays of windows between two cuff measurements. Models' training time was proportional to the GPU usage proportion. Indeed, smaller data files resulted in almost immediate calculations on the GPU. To make the most of available resources, the optimal file size was assessed to improve training efficiency. The two original datasets were transformed into a standard PyTorch tensor format. The file length that maximized GPU usage was one hour. This tensor format was found to minimize I/O bottlenecks while keeping storage size to a minimum. This format allowed for the design a posteriori of model-level dataloaders, adapting model processing on the fly.

Model	High Frequency Features	Derived Features	Window length	Calibration	Operating-room adapted calibration
PulseNet	[ECG, PPG]	-	10 seconds	None	
PPG2BP-Net	[PPG]	-	10 seconds	Random SBP/DBP and PPG segment	Initial cuff and PPG segment
Wave-U-Net	[PPG, PPG', PPG"]	-	10 seconds	None	
Mechanistic	-	[HR, DIC, PI]	5 seconds (agnostic)	Last five cuff measurements	
Proposed	[ECG, PPG]	[HR, DIC, PI]	5 minutes	Last cuff measurement	

Table 3.1: Model features definition and eventual calibration

3.2.5 Statistical Analysis

For each dataset, 10% of the patients were kept as a hold-out test set, while the rest of the data was split between training and validation. Results are presented on the hold-out test set of each dataset. Mean absolute error (MAE), mean error (ME), and cumulative error percentages were reported to assess model performance and the associated standard deviation to quantify error dispersion. Model predictions were all compared to the arterial line. In accordance with the AAMI standard, the Mean Error (respectively Standard Deviation of the error) should be inferior to 5 mmHg (respectively to 8 mmHg). The British Hypertension Society (BHS) defines grades of performance regarding the cumulative error percentages. The cumulative frequency of errors above the thresholds of 5, 10, and 15 mmHg was assessed for each model to compare to the standard threshold for each grade. Inference time was calculated per batch on the same machine for each dataset and then divided by the segments' sample size. Line plots were used to represent cumulative error percentages, with colors representing each grade of the standard. Computational performance was evaluated across three processing environments: GPU, multi-core CPU, and single-core CPU configurations to simulate deployment scenarios ranging from high-performance systems to resource-constrained edge devices and microcontroller units. Single-core processing was conducted using an Intel(R) Xeon(R) CPU E5-2660 v2 @ 2.20GHz with processing restricted to a single core to emulate embedded system constraints.

3.3 Results

3.3.1 Datasets characteristics

The study's eligible data included 2,833 patients in VitalDB and 5,060 patients in LaribDB, for a total of 7,893 patient records. Table 3.2 gathers the descriptive characteristics of both datasets. For LaribDB, the dataset includes data from post-anesthetic recovery room (29%), neurosurgery (24%), surgical intensive care (19%), emergency department (7%), otolaryngology (2%), maternity (<1%) and more. In VitalDB, the main procedures reported ranged from lung and thoracic surgery (29%), digestive surgery (25%), patients from the emergency department(16 %), gynecology (5%) and urology (4%). Systolic, diastolic, and mean arterial pressure followed similar distributions with mean pressure 81.5 (± 14.3) mmHg for the training set of MAP in VitalDB (respectively 80.1 (± 20.8) mmHg in LaribDB) and 82.0 (± 14.1) mmHg in the VitalDB test set (respectively 82.4 (± 20.4) mmHg in the LaribDB test set). LaribDB patients were monitored using either a Philips (19%) or a Massimo device (81%) whereas the patients within the VitalDB database were monitored using the Tram-Rac 4A from GE Healthcare. Patients were included based on the availability of all records in the dataset, thresholds on duration, and quality metrics validation. Details about patients inclusion procedures are available in the quality metrics section and more globally in supplementary figure SF1. The study included 2,549 patients in the training set in VitalDB (respectively 4,554 in LaribDB) and 284 patients in the test set of VitalDB (respectively 506 in LaribDB). The intervals of interest were considered, namely inter-brachial periods. After segmenting the signal, a total of 41,285 non-overlapping 5-minute long windows were obtained in VitalDB (respectively 37,179 in the training set and 4,106 in the test set) and 91,663 windows in LaribDB (respectively 82,497 in the training set and 9,166 in the test set).

Characteristics \ Dataset	VitalDB		LaribDB	
	Train	Test	Train	Test
Number of patients, n	2549	284	4554	506
Number of 5-minute-segments, n	37179	4106	82497	9166
Age, n mean (± std)	59.1 (± 14.6)	58.1 (± 15.6)	54.1 (± 19.2)	55.3 (± 17.6)
Gender, female proportion (%)	44	45	37	38
Pressure variability, mmHg mean (± std)				
SBP	121.5 (± 22.1)	121.7 (± 21.8)	115.7 (± 28.0)	115.1 (± 28.1)
DBP	61.6 (± 13.0)	62.2 (± 12.5)	64.3 (± 20.2)	64.4 (± 18.6)
MAP	81.5 (± 14.3)	82.0 (± 14.1)	80.1 (± 20.8)	82.4 (± 20.4)
Procedures, n (%)				
Post-anesthetic recovery room	-	-	1334	156
Neurosurgery	-	-	1116	120
Surgical intensive care	-	-	885	65
Lung/Thoracic surgery	749	83	-	-
Digestive Surgery	652	81	13	2
Emergency department	-	-	357	38
Gynecology	122	11	-	-
Otolaryngology	-	-	110	18
Urology	124	16	-	-
Maternity	-	-	65	6
Orthopedic Surgery	-	-	39	4
Other general surgery	902	93	1	2
PPG Monitoring Device, n (%)				
M1191BL SpO2 (Philips)	-	-	889	97
MightySat Rx (Massimo)	-	-	3665	409
Tram-Rac 4A (SNUADC, GE HealthCare)	2549	284	-	-

Table 3.2: Datasets characteristics

3.3.2 Considered models

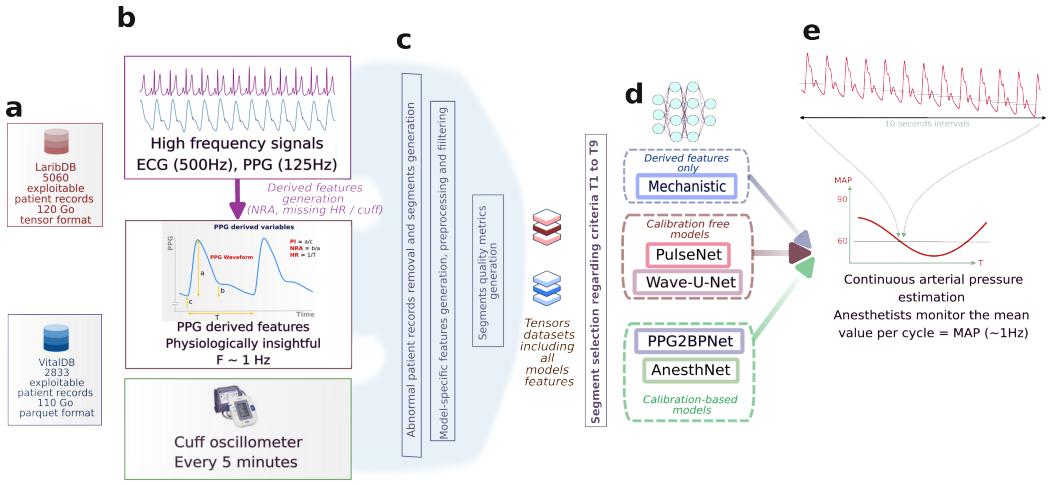


Figure 3.3: Study design. a. Exploitable datasets sizes: LaribDB (5060 patients) and VitalDB (2833 patients). b. Multimodal features categories which include high-frequency data (ECG, PPG), PPG-derived features, and the cuff oscillometer. c. Preprocessing steps, filtering, and generation of ready-to-train datasets. d. State-of-the-art models compared in this study, including physiology-based, calibration-free and calibration-based approaches. e. Each model takes as input a combination of non invasive features and produces a prediction of MAP per window of data. Results are then aggregated per patient to generate the main results metrics.

This comparative study assessed mechanistic, calibration-based, and calibration-free state-of-the-art models. When relevant, architectures were adapted to operating-room settings, notably regarding realistic calibration. Random systolic and diastolic segments were systematically replaced by the last cuff measurement. The most straightforward calibration-free approach was composed of a ResNet architecture using only the raw ECG and PPG (W. Wang et al., 2023). A slightly more complex calibration-free architecture was based on the U-Net architecture originally designed for audio signals, with successive downsampling and upsampling blocks to reconstruct the desired time series (Oord, Dieleman, et al., 2016). This model was also applied to raw ECG and PPG signals. Then, calibration-based models included notably PPG2BP-Net, a CNN-MLP network calibrating using a pair of random PPG and arterial pressure segments for each entry. A multilinear model inspired by physiology was trained on low-frequency features derived from the raw signal of the PPG and recalibrated every five minutes using the cuff. In the

following section, this model will be referred to as the mechanistic model. The model proposed, AnesthNet, estimates arterial pressure during the interval between cuff measurements, which guided the design of the study towards its input characteristics as segments of five consecutive minutes of data. Input features belong to three categories: cuff calibration, high-frequency PPG and ECG signals, and physiologically informative low-frequency features. High-frequency features were fed to a causal convolutional network to account for long-term dependencies within segments, and later on concatenated with low-frequency inputs into LSTM recurrent layers. Calibration is performed using the LSTM inner state affected with the cuff value whenever available.

3.3.3 Models performance assessment

Performance metrics were chosen based on international standards criteria. According to the AAMI standard, the Mean Error (respectively Standard Deviation of the error) should be inferior to 5 mmHg (respectively to 8 mmHg). Cumulative error percentages for 5, 10, and 15 mmHg were assessed for all models to compare their performance to the British Hypertension Society (BHS) standard criteria. As a result, mean absolute error (MAE) and mean error (ME) were reported along with their associated standard deviation, as much as cumulative error percentages to compare model performance.

Tables 3.3 and 3.4 gather the results of each model on VitalDB and LaribDB. The AnesthNet model reached an MAE of 4.6 (± 4.6) mmHg on VitalDB and 3.8 (± 5.7) mmHg on LaribDB. The PulseNet reached an MAE of 11.6 (± 9.3) mmHg on VitalDB and 11.7 (± 11.1) mmHg on LaribDB and PPG2BP-Net reached an MAE of 5.3 (± 4.7) mmHg on VitalDB and 8.1 (± 7.2) mmHg on LaribDB. The Wave-U-Net model reached a performance of 10.2 (± 8.7) mmHg on VitalDB and 11.5 (± 9.2) mmHg on LaribDB. As for cumulative error percentages, the AnesthNet model achieved a grade A performance on both databases. PPG2BP-Net achieved a grade B performance on VitalDB and grade C on LaribDB. The mechanistic model reached a grade C performance on LaribDB. Other models did not meet the BHS standard criteria for the lowest grade, grade C.

Database	Models	Error Metrics	
		ME (mmHg)	MAE (mmHg)
VitalDB	Mechanistic	-	-
	PulseNet	-1.1 (± 14.8)	11.6 (± 9.3)
	Wave-U-Net	-2.8 (± 12.8)	10.2 (± 8.7)
	PPG2BP-Net	0.67 (± 6.7)	5.5 (± 5.1)
	AnesthNet (Proposed)	**0.3 (± 6.3)**	**4.6 (± 4.7)**
LaribDB	Mechanistic	-0.8 (± 14.2)	7.9 (± 11.9)
	PulseNet	3.7 (± 15.4)	11.7 (± 11.1)
	Wave-U-Net	-3.6 (± 14.3)	11.5 (± 9.2)
	PPG2BP-Net	-2.42 (± 10.4)	8.1 (± 7.2)
	AnesthNet (Proposed)	**0.2 (± 6.8)**	**3.8 (± 5.7)**

Table 3.3: Models performance comparison among various learning- and PPG-based cuffless BP estimation architectures. Estimation was based on AAMI standard: The number of test subjects needs to be higher than 85, the ME should be $\leq \pm 5$ mmHg, and the SD of error should be ≤ 8 mmHg. The highlights in bold represent the model reaching the best performance.

Database	Models	BHS Error Metrics			
		e $\leq$ 5 mmHg (%)	e $\leq$ 10 mmHg (%)	e $\leq$ 15 mmHg (%)	BHS Grade
VitalDB	Mechanistic	-	-	-	-
	PulseNet	33.1	60.1	78.7	Below C
	Wave-U-Net	30.2	57.0	76.5	Below C
	PPG2BP-Net	57.7	83.5	93.4	Grade B
	AnesthNet (Proposed)	**65.7**	**88.3**	**96.1**	Grade A
LaribDB	Mechanistic	60.1	77.2	84.1	Grade C
	PulseNet	31.8	59.2	79.4	Below C
	Wave-U-Net	29.4	55.1	73.9	Below C
	PPG2BP-Net	40.8	70.1	86.3	Grade C
	AnesthNet (Proposed)	**72.1**	**91.2**	**95.3**	Grade A

Table 3.4: Models performance comparison among various learning- and PPG-based cuffless BP estimation architectures. Estimation was based on the BHS standard: The grades based on the BHS standard are given based on the error percentage as follows: if 60%, 50%, and 40% of errors ≤ 5 mmHg, then grades A, B, and C are given, respectively; if 85%, 75%, and 65% of errors ≤ 10 mmHg, then grades A, B, and C, are given, respectively; if 95%, 90%, and 85% of errors ≤ 15 mmHg, then grades A, B, and C, are given, respectively. The highlights in bold represent the model reaching the best performance.

Performance regarding models latency

Considering the mean latency induced by the models, they remained within the same order of magnitude for GPU and 24 CPU cores, ie, below 30 milliseconds (ms) per batch. The lowest mean latencies were observed for PPG2BP-Net and PulseNet, with respective mean latencies of 1.8 ms and 1.6 ms per batch on VitalDB and LaribDB. The use of a GPU showed an overall improvement by a factor of 2 to 10 compared to the usage of 24 CPU cores. For the proposed AnesthNet model, the mean latency introduced by the inference process was respectively 3.1 and 3.3 milliseconds for VitalDB and LaribDB. Supplementary Table ST2 shows that even under the deliberately constrained conditions of single-core CPU execution, the latency of all models remained below 500 milliseconds. These latency values are to be compared to the intrinsic latency of five seconds introduced by the operating room acquisition unit at data collection time.

Database	Models	Inference Time	
		GPU (ms)	24-core CPU (ms)
VitalDB	Mechanistic	-	-
	PulseNet	2.3	9.1
	Wave-U-Net	2.6	11.6
	PPG2BP-Net	**1.8**	**3.7**
	AnesthNet (Proposed)	3.1	21
LaribDB	Mechanistic	-	1.8
	PulseNet	**1.5**	6.4
	Wave-U-Net	2.6	27.4
	PPG2BP-Net	1.6	**3.6**
	AnesthNet (Proposed)	3.3	24

Table 3.5: Models performance regarding mean inference time. Mean inference time was defined as the time taken by each model to perform inference on a batch of features. It was averaged over 15 batches and latencies were reported in milliseconds.

Variable	Metric (Unit)	VitalDB		LaribDB	
		PPG2BP-Net	AnesthNet	PPG2BP-Net	AnesthNet
Cuff - 3 minutes	ME (mmHg)	-0.76 (± 6.1)	0.1 (± 4.85)	0.7 (± 10.4)	0.8 (± 6.1)
	MAE (mmHg)	4.8 (± 4.6)	**3.9 (± 4.1)**	7.8 (± 6.9)	**3.3 (± 5.1)**
	e ≤ 5 mmHg (%)	60.9	71.1	41.9	81.9
	e ≤ 10 mmHg (%)	86.0	90.5	71.2	93.9
	e ≤ 15 mmHg (%)	94.9	96.5	87.1	97.1
Cuff - 5 minutes	ME (mmHg)	0.67 (± 6.7)	0.3 (± 6.3)	-2.42 (± 10.9)	0.2 (± 6.8)
	MAE (mmHg)	5.5 (± 5.1)	**4.6 (± 4.7)**	8.1 (± 7.2)	**3.8 (± 5.7)**
	e ≤ 5 mmHg (%)	57.7	65.7	40.8	72.1
	e ≤ 10 mmHg (%)	83.5	88.3	70.1	91.2
	e ≤ 15 mmHg (%)	93.4	96.1	86.3	95.3
Cuff - 10 minutes	ME (mmHg)	-5.0 (± 8.1)	0.9 (± 7.8)	-0.4 (± 11.8)	0.5 (± 8.7)
	MAE (mmHg)	7.3 (± 5.9)	**6.1 (± 5.8)**	8.9 (± 7.8)	**5.7 (± 6.6)**
	e ≤ 5 mmHg (%)	42.2	55.2	37.8	62.1
	e ≤ 10 mmHg (%)	73.8	79.8	66.1	83.7
	e ≤ 15 mmHg (%)	89.1	91.9	82.9	92.1

Table 3.6: Calibration-based models performance regarding distance to cuff calibration. Segments of respectively 3, 5 and 10 minutes were generated to model the variation of cuff measurement frequency. Inference is performed during these intervals for calibrated models.

3.3.4 The impact of distance to cuff calibration

To address the variability of practices or conventions observed between medical sites, the models' robustness to the variation of brachial cuff frequency was assessed. The associated datasets were generated for cuff frequencies of 3, 5, and 10 minutes and the associated windows of data. Results were compared for calibrated models only. On VitalDB, PPG2BP-Net reached a MAE of 4.8 (± 4.6) mmHg for three-minute-intervals (respectively 5.5 (± 5.1) mmHg for five-minute-intervals and 7.3 (± 5.9) mmHg for ten-minute-intervals). On the same database, the proposed AnesthNet model reached a MAE of 3.9 (± 4.1) mmHg for three-minute-intervals (respectively 4.6 (± 4.7) mmHg for five-minute-intervals and (6.1 ± 5.8) mmHg for ten-minute-intervals). On LaribDB, PPG2BP-Net reached a MAE of 7.8 (± 6.9) mmHg for three-minute-intervals (respectively 8.1 (± 7.2) mmHg for five-minute-intervals and 8.9 (± 7.8) mmHg for ten-minute-intervals). On the same database, the AnesthNet model reached a MAE of 3.3 (± 5.1) mmHg for three-minute-intervals (respectively 3.8 (± 5.7) mmHg for five-minute-intervals and 5.7 (± 6.6) mmHg for ten-minute-intervals). AnesthNet proved more robust to an increased duration between two brachial cuff measurements and obtained significantly better results, notably on the more challenging dataset LaribDB.

3.4 Discussion

This multicentric study offers valuable insights into the performance of state-of-the-art models for non-invasive blood pressure estimation in the operating room, using the largest datasets available to date. Both recent-and-competitive and traditional-yet-effective approaches were included to reflect the range of existing methods. Comparing and evaluating these models from an anesthetist's perspective emphasizes deployable architectures suitable for practical use and underscores the performance disparity between cuff-based and cuffless methods. The proposed architecture, AnesthNet, is a multi-modal network that takes both raw waveform and cyclic features, along with brachial cuff measurements, and proposes estimates between cuff measurements using built-in deep learning model parameters. Generally speaking, calibration-based architectures performed better than calibration-free and mechanistic approaches. More specifically, AnesthNet outperformed the other models independently of the interval between cuff measurements. Notably, LaribDB appears to be more discriminative across models, revealing larger performance differences, especially with PPG2BP-Net, which did not meet the BHS grade C. This may be due to the greater variability and complexity in signal collected in the LaribDB dataset compared to benchmark datasets like VitalDB. The integration of a clinical prior via LSTM initialization in the AnesthNet could have contributed to its robustness and generalizability across possibly more challenging, real-world conditions.

3.4.1 Calibration as a field limit

In practice, calibration is always available, but the frequency with which it is taken may vary depending on the stage of the operation or the usual practices in a hospital. Resilience to varying calibration is crucial to ensure the robustness of an estimator within real-life conditions. Although providing the best results, the calibration-based models were sensitive to the distance to the last cuff. However, performance error remained acceptable for the usual frequency ranges of cuff measurement. The use of the proposed MAP estimation is compatible with the conventional use of cuff calibration. None

of the models achieved an A quality level beyond 10 minutes from calibration, which appears to be a field limit, beyond which it becomes more interesting to carry out a recalibration, *i.e.* to take an additional oscillometry measurement. This also demonstrates the complementarity of this approach with clinical practice, serving as an assistant to clinicians by balancing real-time estimation with recalibration as necessary.

3.4.2 Implementability at the operating room

All models provided acceptable results regarding latency. Using a GPU systematically and significantly reduced inference time. The AnesthNet approach, along with the Wave-U-Net architectures, showed the highest benefit of GPU usage, which is consistent with their larger number of parameters. Nonetheless, all models' prediction times respected the order of magnitude of current measurement methods ($\Delta t < 1$ second per sample) and are compatible with operating room inference infrastructure. While the concept of a trade-off between performance and inference time could be discussed, the notable differences in model performance suggest that the accuracy of predictions prevails here.

3.4.3 Limitations on field characteristics

Traditional research in this domain evaluates performance outcomes using internationally recognized benchmarks such as the AAMI and BHS standards, predicated on estimating near real-time beat-to-beat arterial pressure values. These could be over-optimistic in the case of an over-representation of patients with a low BP variation range. Apart from this, modelling the occurrence of hypotension is also a subject in itself. Some approaches to framing the problem also allow for a more specific visualization of performance concerning hypotension, such as classifying hypotensive event detection; however, this relies on a rather unsettled clinical definition of IOH. Also, such an approach does not address the issue of prediction uncertainty, though it could provide more representative conclusions regarding model accuracy.

Moreover, considering the growing complexity of the models used, it could foster clinical acceptability. Future work will notably attempt to assess the model uncertainty using state-of-the-art methods, such as the conformal uncertainty prediction framework (Angelopoulos and Bates, 2022). Such an approach would make it possible to add to the arterial pressure estimator a quantification of the risk associated with the uncertainty of the predictions. The deployment of the model in the operation room will also be performed to test the system in real-life conditions. In future investigations, examining systolic blood pressure (SBP) and diastolic blood pressure (DBP) as distinct variables could also provide valuable insights beyond those captured by mean arterial pressure alone. While MAP is central to anesthetists, analyzing SBP and DBP separately may reveal differential associations with clinical outcomes, as these parameters reflect distinct aspects of cardiac function and vascular dynamics. Such an approach could enhance our understanding of the independent contributions of systolic and diastolic pressures to advanced methods in monitoring or cardiology.

3.5 Chapter conclusion

In this study, various architectures were compared for the non-invasive estimation of Mean Arterial Pressure (MAP) using operating room data. Our proposed model combines causal convolutions on physiological signals to capture long-term dependencies with an initialization strategy for deep learning parameters. Specifically, the inner state of Long Short-Term Memory (LSTM) networks was initialized using field calibration data from cuff oscillometry, leveraging domain knowledge and enabling real-time capabilities. Our model demonstrated superior performance compared to existing architectures on both databases, while maintaining inference times compatible with near real-time deployment.

Regarding the interactions between features, simultaneous and synchronous monitoring of ECG and PPG provided better results. Physiologically, this could be explained by the fact that it enables the estimation of the pulse arrival time (PAT), defined as the time interval between the R-wave of the ECG and the arrival of the corresponding pulse wave at a peripheral site, typically measured using a PPG sensor. As a surrogate of the pulse transit

time (PTT), and often used as an indicator of cardiovascular function, the PAT is related to arterial stiffness and blood pressure, notably in high frequency ranges (Hoshide et al., 2022). However, since the PAT varies by the same orders of magnitude as the shifts that can be observed in signals due to measurement artefacts, its systematic calculation remains untrustworthy.
These first two sections have enabled us to visualise the state of the art in blood pressure estimation. One observation that can be made about these two approaches is that they are both based on supervised learning. How can we go beyond the threshold of 'labelled' data which necessitates an arterial catheter? How can we leverage the volume of data available during 'conventional' general anaesthesia to maximise the performance of our models? We will explore this in the next section.

3.6 Appendix

Database	**Models**	**Error Metrics**		**BHS Error Metrics**		
		ME (mmHg)	MAE (mmHg)	e $\leq$ 5 mmHg	e $\leq$ 10 mmHg	e $\leq$ 15 mmHg
VitalDB	PPG2BP-Net: Initial cuff	1.3 (± 9.1)	6.9 (± 6.2)	48.2	77.3	90.6
	PPG2BP-Net: Last cuff	0.67 (± 6.7)	5.5 (± 5.1)	57.7	83.5	93.4
LaribDB	PPG2BP-Net: Initial cuff	6.9 (± 12.8)	10.7 (± 10.1)	35.6	59.7	74.3
	PPG2BP-Net: Last cuff	-2.42 (± 10.4)	8.1 (± 7.2)	40.8	70.1	86.3

Supplementary table ST1: PPG2BP-Net performance comparison depending on the replaced calibration definition. A conservative approach close to the calibration performed by the authors is to calibrate using a unique calibration point corresponding to the beginning of the intervention. A more precise approach leveraging the latest measure from the operating room is to use the previous cuff measure. In the main results, this model's results are presented with the configuration that maximized performance, that is to say, the latest.

Dataset	Architecture	ME (mmHg)	MAE (mmHg)
VitalDB	Dilated CNN	-1.1 (± 6.7)	4.8 (± 5.1)
	Dilated CNN + Hand-crafted features	0.3 (± 6.3)**	4.6 (± 4.7)**
LaribDB	Dilated CNN	-0.4 (± 7.2)	3.9 (± 6.0)
	Dilated CNN + Hand-crafted features	0.2 (± 6.8)**	3.8 (± 5.7)**

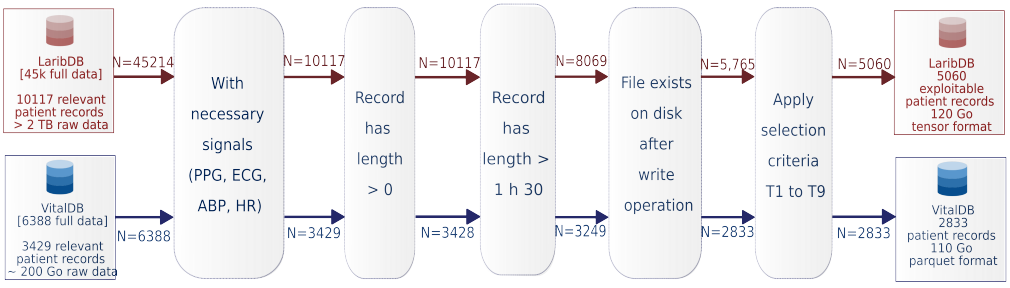
Supplementary Table ST2: Ablation study regarding hand-crafted features. The ablation study was conducted on VitalDB. For each dataset, a paired t-test revealed that the error distributions of our model and the other model were significantly different ($p<0.001$), as indicated by the double asterisk in the table.

Reference	Model name	Dataset (dataset category)	Number of training subjects	Number of test subjects	SBP error (mmHg)	DBP error (mmHg)	Obtained MBP error (mmHg)
[18]	Adaboost	MIMIC II (ICU)	At least 441	At least 41	-0.05 (± 8.9)	0.18 (±4.17)	-
[18]	Random Forest Regressor	MIMIC II (ICU)	At least 441	At least 41	0.15 (± 10.6)	0.19 (± 4.73)	-
[18]	SVR	MIMIC II (ICU)	At least 441	At least 41	-0.9 (± 16.71)	-0.6 (±7.5)	-
[21]	ABP-Net	ICU DB (ICU)	1620	162	1.64 (±7.42)	-0.28 (±5.81)	-2.8 (± 12.8)
[25]	U-Net	UCI Dataset (ICU)	unknown	unknown	0.06 (± 20.9)	-2.46 (± 10.8)	-
[32]	PPG2BPNet	VitalDB (Anesthesia)	2987	410/797 (whole: hold-out)	-0.231 (±10.263)	0.062 (±6.252)	0.67 (± 6.7)
[33]	PulseNet	VitalDB / MIMIC IIII (Anesthesia/ ICU)	2506	279	0.8 (± 18.6)	0.5 (± 12.3)	-1.1 (± 14.8)

Supplementary Table ST3: Performance Comparison of Blood Pressure Estimation Models. This table shows the performance obtained by other models on previously published work. The 'dataset category' column stands for anesthesia versus intensive care unit, to account for the patient admission category. The obtained MBP error corresponds to the model performance reached in our study for the VitalDB dataset.

Model name	Number of parameters	Latency on single-core CPU (ms)
Mechanistic	5	86
PulseNet	3,844,865	266
Wave-U-Net	5,484,695	415
PPG2BP-Net	1,136,763	154
AnesthNet	2,978,873	197

Supplementary Table ST4: Resource-constrained evaluation. We tested all models using single-core CPU execution. This represents a severely limited processing environment relative to the multi-core systems standard in modern operating rooms. Under these restrictive conditions, all models achieved latencies below 500 milliseconds. We note that these performance benchmarks were deliberately conservative to establish computational feasibility even in resource-constrained settings. These processing times are substantially faster than the 5-second acquisition delay inherent to the operating room data collection system, confirming that model inference does not introduce computational bottlenecks into the clinical pipeline.



Supplementary Figure SF1: Detailed patient selection process and criteria for both databases.

4

Self-supervised learning approaches to unveil brain and vessels representations

In the preceding chapters, we demonstrated how supervised learning techniques can be leveraged to model and estimate mean arterial pressure from non-invasive and mandatory operating-room data. Another specificity of operating-room data lies in the abundance of unlabelled signals. In the frame of blood pressure monitoring, label data refer to arterial lines which are not always placed on the patient.

In this chapter, we focus on SSL, which leverages unlabelled data, to propose novel and richer representations of patients' physiological state during general anesthesia. We will demonstrate that different physiological signals require distinct modeling strategies and pose unique challenges. Among these signals, the electroencephalogram, though exhibiting periodic components, is generally less strictly rhythmic than highly periodic biosignals such as ECG or PPG. More precisely, applied to electroencephalography, such methods provide a comprehensive representation of a patient's depth of anesthesia. Then, we will leverage multimodal learning strategies to address the synchronous acquisition of operating room vitals, in an attempt to provide a more complete view of a patient's state. This is preliminary work for a manuscript in preparation.

4.1 Introduction

Despite the abundance of approaches in the fields of computer vision and NLP, self-supervised learning approaches are less commonly applied to physiological signals. One of the most studied biosignals is the electroencephalogram (EEG), which has been the subject of numerous studies. For this reason, this work begins by studying EEG under general anesthesia, before attempting to extend this work to other intraoperative signals.

4.1.1 On the depth of anesthesia

How can we be sure that a patient is adequately sedated? Accidental awareness during general anesthesia (AAGA) is a distressing complication, as it can induce patient psychological trauma and pain. In worst cases, long-term sequelae include notably post-traumatic stress disorder, depression, anxiety, and sleep disturbance (Ghoneim, 2007; Sandin et al., 2000). Its psychological impact can be lifelong. AAGA incidence is estimated at 1:19,000 during general anesthesia. Still, it is subject to high variations across anesthesia procedures, with a rate of up to 1:670 in caesarean sections (Cook et al., 2014), and some studies suggest that its incidence could even be higher (Goddard and Smith, 2013).

Furthermore, AAGA has been shown to occur especially during the induction phase - at the time of drug injection (Pandit, Andrade, et al., 2014). To avoid such risks, EEG analysis provides crucial information for characterising the depth of anaesthesia (DoA) and is endorsed by clinical recommendations (Checketts et al., 2016). EEG-based Depth of Anesthesia (DoA) monitoring is becoming a crucial component of modern anesthesiology, aiming to prevent both intraoperative awareness and anesthetic overdose (Hajat et al., 2017). This field encompasses several established commercial systems and emerging research approaches that process EEG signals to provide quantitative measures of consciousness level during general anesthesia (Ledowski, 2019; Shander et al., 2018). However, the validation processes of all these devices are not equivalent (Bruhn et al., 2006). More specifically, it has been shown that given several commercialized products with augmented EEG-based met-

rics, in up to two-thirds of the cases, the recommendations proposed by the devices diverged, highlighting the need for personalized EEG understanding and interpretation (Hight et al., 2023).

Among these indicators, the Bispectral Index (BIS) is the first and remains one of the most widely adopted EEG-based DoA monitors in clinical practice. Introduced in 1994 by Aspect Medical Systems, BIS uses a proprietary algorithm that combines time-domain, frequency-domain, and high-order spectral EEG parameters to generate a single dimensionless index ranging from 0 (isoelectric EEG) to 100 (fully awake). Figure 4.1 illustrates the association between BIS values and the depth of anesthesia of the corresponding patient. BIS values of 40-60 are considered optimal for general anesthesia, providing adequate unconsciousness with minimal awareness risk. The B-Aware randomized controlled trial showed that the use of BIS could reduce the risk of AAGA by 82% (Myles et al., 2004). To date, however, monitoring the BIS requires a commercial licence and not all operating rooms using EEG monitoring possess the ability to benefit from the BIS index during GA. Therefore, there is a basis for proposing novel BIS predictors based on the raw EEG signals. Such approaches also hold valuable potential for efficient decision-making at the OR.

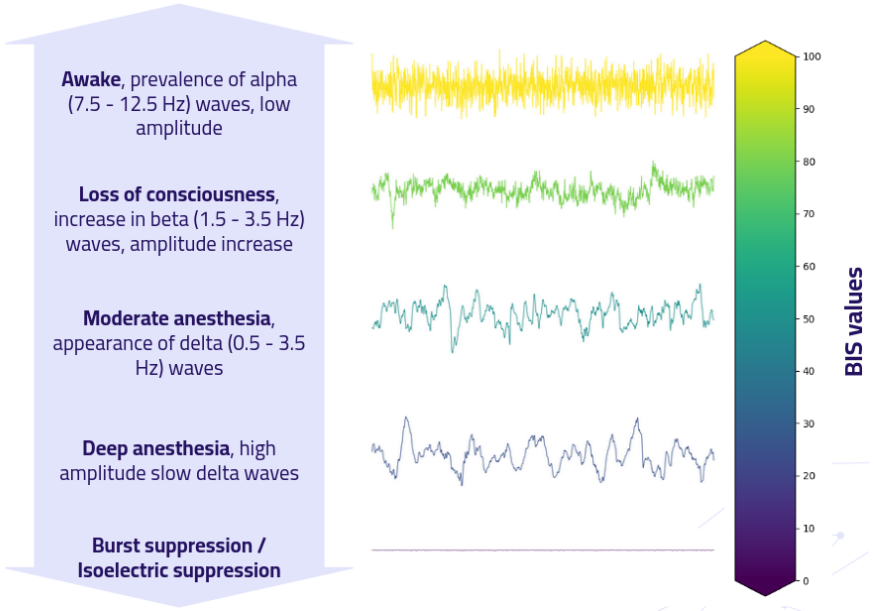


Figure 4.1: Depth of anesthesia and bispectral index values correspondance diagram

4.1.2 SSL methods applied to EEG

Electroencephalography used during general anesthesia relies on similar principles to those applied in sleep research and epilepsy monitoring, although it typically employs a reduced number of electrodes. In epilepsy, attention-based transformers have been found effective in detecting seizures (T. Xiao et al., 2024). In the context of SSL approaches, models have been developed to recognize emotions from raw EEG signals. This is achieved through a pretext task, wherein various transformations are applied to the EEG data, and the model is trained to identify the specific transformation that was applied. (X. Wang et al., 2023) . However, considering the specificities of the EEG signal, knowledge-driven SSL frameworks and custom signal-oriented pretext tasks are valuable avenues worth exploring (Weng et al., 2024). In fact, for EEG recorded during sleep, convolutional neural network architectures have been designed to effectively process the temporal and spectral characteristics of raw EEG signals, enabling for instance a robust automatic sleep stage classification (Chambon, Galtier, et al., 2017). Building on this paradigm, self-supervised learning approaches have been developed to exploit large collections of unlabeled EEG data. In particular, Banville et al., 2021 proposed a contrastive learning framework based on relative positioning tasks, in which the model learns to determine whether two signal segments are temporally close. This method effectively captures semantic representation of EEG signals. It holds great potential to be adapted to neighboring fields such as general anesthesia monitoring.

4.1.3 Defining similarity in anesthesia

Monomodal approaches consider only one signal source. They assume that pairs of windows are positive under certain conditions, for instance if they belong to the same subject (Bhattachrjee et al., 2023). This modelization requires numerous patients so the model learns variability. Frequency domain spectrograms (Zhang et al., 2024) or morphology-based foundation models (Pillai et al., 2025) have also been shown to improve results compared to patient-level similarity definition. Then, once we have learned from different signals separately, how do we aggregate and learn from various modalities

altogether? One specificity of GA is the multimodality of the available biosignals. In this context, various pretraining approaches exist. Notably, different definitions of positive pairs can be tried. In the CLIP paradigm (Radford et al., 2021), images and their descriptive text are paired and share the same meaning. For operating room data, this matching is less evident. A straightforward approach consists of pairing simultaneous signal windows, assuming they represent the same state (Thapa et al., 2024). However, this approach can be misleading, notably considering the low level of information in the signals considered, with few actual pairs of signals being representative of valuable concepts to learn.

4.2 Methods

4.2.1 Patients

The VitalDB dataset was downloaded using the provided Python library (H.-C. Lee, Park, et al., 2022). The patients relevant to this study were selected based on the simultaneous presence of waveform 2-lead EEG, BIS, PPG, ECG, HR, arterial blood pressure and propofol injection signals. ECG, PPG and ABP were monitored using the General Electric Tram-Rac 4A and EEG was monitored with the Medtronic Vista sensor. Data was recorded at a sampling rate of 1 Hz for low-frequency data and 128 Hz for high-frequency waveform data. Patients were maintained unconscious using propofol and rémifentanyl. EEG waveforms were filtered using a low-pass filter of 30 Hertz frequency.

The dataset was split between train, validation and test sets using a 60/20/20 ratio at the patient level. For each patient, the data was split into segments of 30 seconds of continuous data. EEG windows with mean BIS values outside of the [10, 80] range were excluded. When attempting to predict maximal values for ECG and PPG, the systolic peak was computed using the physiocurve and neurokit2 libraries. Downstream tasks of interest were defined as the evaluation of the mean BIS value over the window for EEG and the heart rate for cyclic signals such as the PPG. The records were generated using the Python vitaldb library.

4.2.2 Training pipeline

The SSL training pipeline is two-fold. The first phase consists in the training of the full model on the pretext task, with the aim of extracting the encoder. The rationale behind the pretext tasks is to create meaningful supervision from unlabeled data by designing auxiliary objectives that encourage the model to learn generalizable features without explicit human annotations. The main categories of SSL pretext tasks include generative approaches - autoencoders, masked autoencoding, contrastive methods (He, Fan, et al., 2020; T. Chen et al., 2020), and discriminative tasks. Each of these tasks category is designed to capture different aspects of the data's underlying structure. Here, we focus on constrastive learning on the EEG. Reconstruction methods will be developed further down the manuscript. During the second phase, the downstream phase, we plug the pretrained encoder to a shallower network, which is trained on the downstream task of interest. The aim is to improve downstream tasks' performances, notably in the low-label regime. In this section, we focus on one specific pretext task: relative positioning. This method was found most conclusive on the EEG in our experiments. Other SSL methods exist and can be applied, such as reconstruction tasks, which we detail one example in section 4.4.1.

Relative positioning

The process of self-supervised learning using relative positioning is detailed in Figure 4.2. A pretaining phase is performed on the pretext task of interest. This task was defined by Banville et al., 2021 as the relative positioning for two EEG windows. A contrastive model learns whether two windows of EEG are close using the following definition of proximity:
Given $\tau_{\text{pos}} \in \mathbb{N}$, which controls the duration of the positive context, and $\tau_{\text{neg}} \in \mathbb{N}$, which corresponds to the negative context around each window X_i, we sample N labelled pairs:

$$\mathrm{Z}_N = \left\{((X_{t_i}, X_{t'_i}), y_i) \mid i \in \{1, \ldots, N\}, (t_i, t'_i) \in T, y_i \in Y\right\} \quad (4.1)$$

We consider that the set of labels defining proximity is defined as $Y = \{-1, 1\}$. The set of time index pairs T is defined as:

$$\mathrm{T} = \{(t, t') \in \{M - L + 1\}^2 \mid |t - t'| \leq \tau_{\text{pos}} \quad \text{or} \quad |t - t'| > \tau_{\text{neg}}\} \tag{4.2}$$

Intuitively, T is the set of all pairs of time indices (t, t') which can be constructed from windows of size L in a time series of size M, given the duration constraints imposed by the particular choices of τ_{pos} and τ_{neg}.

Here $y_i \in Y$ is specified by the positive or negative contexts parameters:

$$y_i = \begin{cases} 1, & \text{if } |t_i - t'_i| \leq \tau_{\text{pos}} \\ -1, & \text{if } |t_i - t'_i| > \tau_{\text{neg}} \end{cases} \tag{4.3}$$

τ_{neg} and τ_{pos} are training hyperparameters. Windows sampled outside of the positive and negative contexts are ignored.
In this experiment, the EEG embedder was built on the architecture from Chambon, Thorey, et al., 2018, which has shown excellent results at sleep staging classification on polysomnography data. The Python Braindecode library was used for biosignal datasets generation.

Downstream phase

Once successful training is observed, the associated embedder is saved for downstream tasks prediction. Subsequently, another shallow model such as a Random Forest, Support Vector Regressor, or Multi Layer Perceptron can be used to predict the target of interest - the BIS in this case.

4.2.3 Quantitative evaluation on downstream tasks of interest

We defined two categories of downstream tasks based on their nature. Characterizing the patient state using cerebral and cardiac biosignals opens the door for global downstream tasks of interest. For the EEG, the quantification

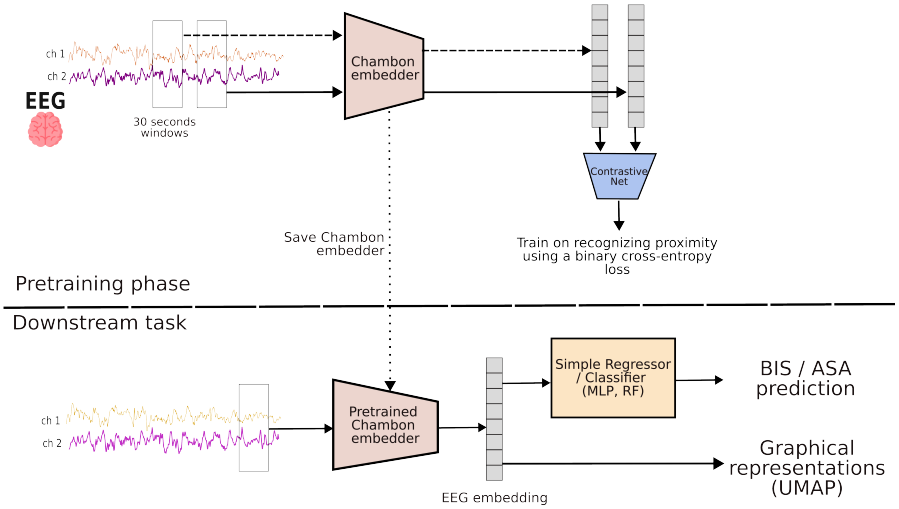


Figure 4.2: Self-supervised learning training pipeline on EEG

of the depth of anesthesia using the bispectral index appears natural. For other modalities, the definition of downstream tasks is more focused on long-term comorbidities that could be diagnosed using operating room data.

Real-time downstream tasks

A first natural tasks category is composed of near-real-time biomarkers which vary along the operation. Tasks performances are evaluated based on the robustness of their predictions regarding harmful intraoperative events or improvements in patient monitoring. In the former case, the prevalence of such events within the dataset, the length of the records, and the synchronicity of signals in the case of multimodal learning is most valuable to improve the variability of the training set. In general anesthesia, we distinguish between downstream tasks of interest related to the brain and some related to the cardiovascular system. Regarding the depth of anesthesia, a straightforward metric to predict at the window level is the bispectral index (BIS). Then, for the cardiovascular system, an a priori relatively simple downstream task could be inferring the heart rate, which is usually computed from the PPG waveform. Then, an ultimate biomarker of interest lies in arterial pressure or associated metrics. Considering blood oxygenation, arterial pressure would

provide insights into actionable metrics such as IOH prevention as we have seen in chapters 2 and 3.

Long-term downstream tasks

Patient-dependent tasks are descriptive tasks relative to the patient, or long-term comorbidities. They hold value in improving postoperative care. Considering patient descriptive characteristics, general downstream tasks can include predicting the age, length of stay at the hospital. Mortality was rejected as too few patients were concerned. Considering long-term cardiac comorbidities, hypertension and diabetes appear as helpful metrics to redirect the patient to relevant care departments.

4.2.4 Qualitative assessment using visual representations

After considering SSL-learned features performance on downstream tasks, one can wonder what structure was captured by SSL. Hence, we examined the relationship between the embeddings and patient characteristics available in the dataset. We projected the 100-dimensional embeddings obtained on VitalDB to three dimensions using Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2020). This manifold learning technique for nonlinear dimensionality reduction that preserves the topological structure of high-dimensional data. UMAP projects embeddings into a lower-dimensional space suitable for visualization while maintaining both local neighborhoods and global data structure. Then, to assess the results regarding comorbidities, we focused on age and split patients into four age groups considering the age distribution in the dataset: young (age35), adult (35<age60), senior (60<age75) and old (75<age). Each age group was represented in a specific color.

4.2.5 Statistical analysis

The pretrained encoder was extracted and used to generate embeddings. These embeddings were passed to a shallow classifier or regressor depending

on the downstream task of interest. Performances are displayed for the unseen test set. As for the metrics used, models were evaluated based on classification and regression performance on downstream tasks of interest. Error and absolute errors distribution and R2 score were used for regression tasks, while area under the curve (AUC) was used for classification.

4.3 Results

4.3.1 Patients

Table 4.1 summarizes the cohort characteristics. The cohort of interest included 797 patients from the VitalDB Dataset. The patients were split between train, validation and test sets using a 60/20/20 ratio. The average record duration was of 3.11 $\pm$ 1.68 hours in the train and validation sets and 3.13 $\pm$ 1.82 in the test set. The mean age of patients was of 57.36 $\pm$ 15.1 in the train/validation set and 55.61 $\pm$ 14.58 in the test set. They were respectively composed of 50.7 % and 50.3 % of women. Most patients were of ASA score 1 and 2. Procedures ranged from general surgery to thoracic surgery. Considering comorbidities, preoperative diabetes and hypertension are clinically known to be strongly correlated, as both conditions often coexist due to shared risk factors such as obesity, aging, and metabolic syndrome. However, in this dataset, the reported occurence of hypertension (mean 26–35%) and diabetes (mean 6–9%) exhibits an unusually large disparity. Given the established epidemiological link between these comorbidities, such a pronounced difference in prevalence rates is unexpected and suggests potential inaccuracies in data collection, coding, or reporting.

Characteristics	Train+Val (n=625)	Test (n=155)
Duration (mean ± SD)	3.11 ± 1.68 hours	3.13 ± 1.82 hours
Age (mean ± SD)	57.36 ± 15.12	55.61 ± 14.58
Gender (female, %)	50.7	50.3
Weight (mean ± SD)	61.50 ± 11.41	61.71 ± 11.30
Height (mean ± SD)	162.10 ± 8.91	162.09 ± 8.51
Preoperative hypertension (%)	28.5	24.5
Preoperative diabetes (%)	8.9	6.5
ASA Score		
1	184	46
2	368	87
3	58	20
4	4	1
ICU Days	**Train+Val**	**Test**
0 days	502	131
1 day	87	17
2–8 days	30	4
Department		
General surgery	493	123
Thoracic surgery	103	21
Gynecology	20	6
Urology	9	5

Table 4.1: Descriptive statistics for continuous variables by subset

4.3.2 SSL on the EEG reveals the depth of anesthesia

We considered the bispectral index as the downstream task of interest. To visualize the semantic embedding structure, we generated the UMAP representations of the SSL approach. The manifold exhibited a clear gradient regarding DoA, with two distinct regions regarding BIS extreme values. These regions are coherent with clinical interpretation as they correspond to awake and hypnosis states. This representation is available on Figure 4.3. In fact, the BIS can be interpreted as a continuous equivalent of sleep stages that are classified in Banville et al., 2021. A distinctive feature of anaesthesia is that patients transition from a state of wakefulness to a state of hypnosis and then back to a state of wakefulness. Therefore, it may be interesting to examine the trajectory of patients on the manifold during anaesthesia.

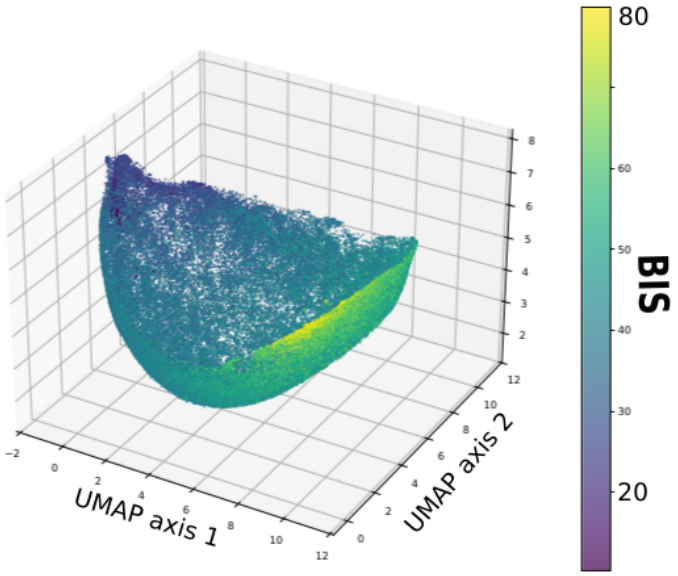


Figure 4.3: SSL model predictions colored by bispectral index values

Considering downstream task prediction performance, the self-supervised EEG model predicted the bispectral index accurately. As a comparison, its performance was compared to a supervised model trained on the raw EEG signal. For this purpose, we implemented a 1D CNN network based on the CNN architecture from (Chambon, Thorey, et al., 2018) on the 2-lead raw EEG channels. These results are available in Table 4.2. On the test set, the SSL model predicted the BIS with a mean error of -0.1 (± 4.0), a mean absolute error of 3.1 (± 2.6), and a reconstruction score R2 of 0.83. The supervised model reached a mean absolute error of 4.0 (± 4.7) and a R2 score of 0.57. Results from existing studies using supervised learning on other datasets are also provided for comparison. They emphasize the fact that our model shows the lowest error variability and is compatible with clinical use.

We compared supervised versus unsupervised learning performance with a varying training set size, for a number of patients in training set in the range [10, 50, 100, 300, 480]. The self-supervised approach largely reduced the prediction error for all training sizes, with a decrease in error variability from 41% to 29%. Predictions from the SSL model are more precise and more stable. These results are available on Figure 4.4.

Publication	Dataset (n patients)	Model	Mean error	Mean absolute error	R2 score
This work	VitalDB (797)	SSL (Relative Positioning)	-0.1 (± 4.0)	3.1 (± 2.6)	0.83
Chambon, Thorey, et al., 2018	VitalDB (797)	Supervised (1D CNN)	-	4.1 (± 3.8)	0.67
H.-C. Lee, Ryu, et al., 2019	VitalDB (5427)	Decision tree / linear regression	N/A	4.1 (± N/A)	N/A
H. Li et al., 2024	N/A	Generative adversarial networks	N/A	3.61 (± N/A)	N/A

Table 4.2: BIS prediction performance comparison. N/A indicates that the authors did not report the corresponding metric.

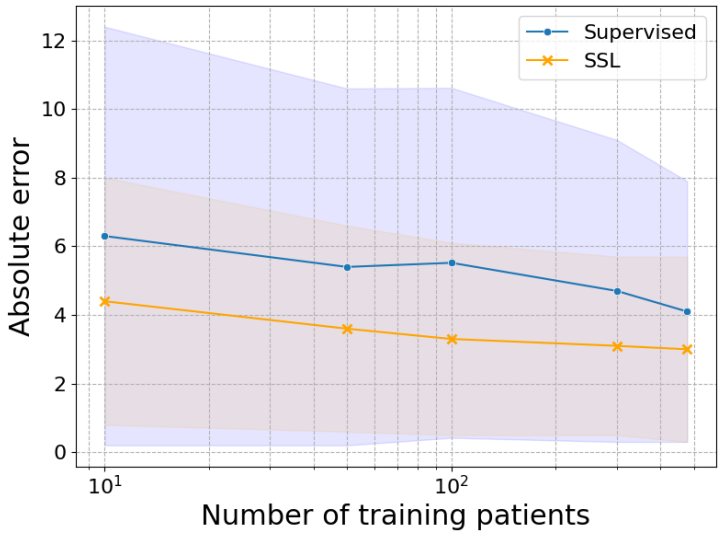


Figure 4.4: Self-supervised and supervised learning comparison. Lower is better. Standard deviation of the absolute error is represented as the shaded area.

4.3.3 Patient trajectories

To illustrate the differences in trajectories according to age, we defined four age groups based on the age distribution. The patients were split between young (age<35), adult (35<age<60), senior (60<age<75) and old (75<age). We averaged the duration of the trajectory for all patients and

computed the median trajectory for each age group. To quantify the noise in the trajectories, we defined the geometric curvature κ such that:

$$\kappa = \frac{\|\mathbf{r}'(t) \times \mathbf{r}''(t)\|}{\|\mathbf{r}'(t)\|^3} \tag{4.4}$$

Where:

- $r'(t) = v(t)$ is the first derivative, representing the tangent vector.
- $r''(t) = a(t)$ is the acceleration.
- $\times$ denotes the cross product.
- $||.||$ denotes the magnitude (Euclidean norm) of the vector.

Supplementary Table ST1 displays the curvature statistics for each age group. Considering the high variability among these curvatures, a Savitzky-Golay filter of order 3 and window length 51 was applied to the median trajectory of each group to smooth the trajectories. We represented these median trajectories in bold in four distinctive colors for each age group with each color representing an age group. This representation is available in Figure 4.5. This result underscores the capacity of the learned manifold to discriminate between patients based on age, serving as a proxy for fragility. It supports the broader hypothesis that general anesthesia can function as a diagnostic indicator, revealing differences in patient fragility through brain responses to anesthetic agents. It corroborates the "asleep medical appointment" assumption defined in section 1.3.1.

Considering the statistical significance, a Kolmogorov-Smirnov test on each axis of coordinates of the manifold showed that the distribution of the coordinates was significantly different between groups. Statistics confirmed the visual insight that young and older patients observed different paths on the manifold, notably on the second axis.

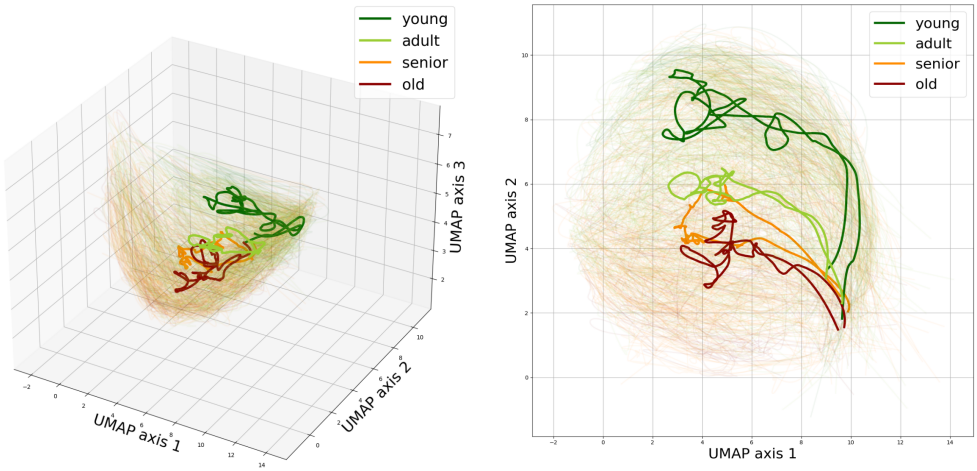


Figure 4.5: Trajectories on the test dataset by age group: young (age<35), adult (35<age<60), senior (60<age<75) and old (75<age), in three dimensions on the left and on the X/Y plane on the right. The trajectories of each group are represented in a different color, with a gradient from green, light green, yellow to orange.

4.4 Experiments on other modalities

4.4.1 Leveraging multimodal input signals

If the EEG provides information on the depth of anaesthesia, how can the patient's cardiovascular status be inferred from the other signals? The anaesthetist's role is to interpret a multiplicity of signals simultaneously. Some studies describe the limitations of methods based on EEG alone (Pandit and Cook, 2013). In the field of epilepsy, multimodal patient representation combining EEG and ECG features was found helpful (Vandecasteele et al., 2021). One could be tempted to apply the same principle to other monitored signals at the OR, to build a larger set of characteristics of the patient. Unfortunately, for the ECG, PPG and arterial pressure (ART), relative positioning was found unsatisfactory in capturing insightful features. The associated results are available in the chapter appendix 4.7. In fact, the semi-periodic nature of these signals makes them a weaker target for temporal proximity classification using contrastive learning. Conversely, the low-frequency changes in the EEG which can be captured by relative positioning also characterize the awakening status of the brain. Other SSL methods exist and have shown promising

results on semi-periodic signals. For instance, representation learning has been addressed using masked autoencoders (He, X. Chen, et al., 2021), primarily designed for computer vision applications. The concept has also been adapted to time series and has achieved convincing performances in time series analysis using neural ordinal equations (Q. Liu et al., 2025) or masking (Z. Li et al., 2023). Among the cyclic signals acquired at the operating room, SSL methods applied to the PPG have exhibited high inter-subject variability (Ghorbani et al., 2022). A method based on contrastive learning on the PPG associating pairs belonging to the same patient yielded satisfactory results (Bhattachrjee et al., 2023). Latent forecasting and instance discrimination reached similar performance to supervised learning on 12-lead ECG (Mehari and Strodthoff, 2022). Among the self-supervised methods, reconstruction methods hold the potential to provide valuable insights from these signals. In this section, we describe preliminary results we obtained in this direction.

Semi-periodic modalities from the operating room

A simple explanation could reside in the semi-periodic natures of these biosignals, for which the proximity lies less in the temporal proximity to other signal windows than in the shape of the pattern. The information they carry can also be condensed and interpreted using cyclic features, which guide clinicians to interpret patients' decisions. To try to overcome this issue, masked autoencoders were used to learn helpful features from signal reconstruction. Based on the work from He, X. Chen, et al., 2021, we defined a training pipeline for a masked auto-encoder working on PPG, ECG and ART biosignals. As previously developed in the 1, the learning process can be defined as follows:

Given an input signal $\mathbf{x} \in \mathbb{R}^{T \times d}$, we randomly masked a fraction ρ of the temporal tokens to create a corrupted input $\mathbf{x}_{\mathcal{V}}$, where $\mathcal{V}$ denotes the set of visible (non-masked) token indices. The encoder f_θ processes only the visible tokens to produce latent representations:

$$\mathbf{z} = f_\theta(\mathbf{x}_{\mathcal{V}}) \tag{4.5}$$

The decoder g_ϕ then reconstructs the masked tokens from the encoded representation:

$$\hat{\mathbf{x}}_{\mathcal{M}} = g_\phi(\mathbf{z}, \mathcal{M}) \tag{4.6}$$

where $\mathcal{M} = \{1, \ldots, T\} \setminus \mathcal{V}$ denotes the masked token indices. The reconstruction loss is computed only on masked tokens:

$$\mathcal{L}_{\text{MAE}} = \frac{1}{|\mathcal{M}|} \sum_{t \in \mathcal{M}} \|\mathbf{x}_t - \hat{\mathbf{x}}_t\|_2^2 \tag{4.7}$$

where $|\mathcal{M}|$ is the number of masked tokens.

The study design of this experiment is detailed in Figure 4.6. The pretext reconstruction performance of the masked auto-encoder was evaluated using mean squared error applied to the decoder on the masked regions. The best pretraining results were obtained for the PPG and are the ones presented below. ECG and ART were not conclusive at the pretraining stage, perhaps due to the superior artefacts and noise observed for these signals.

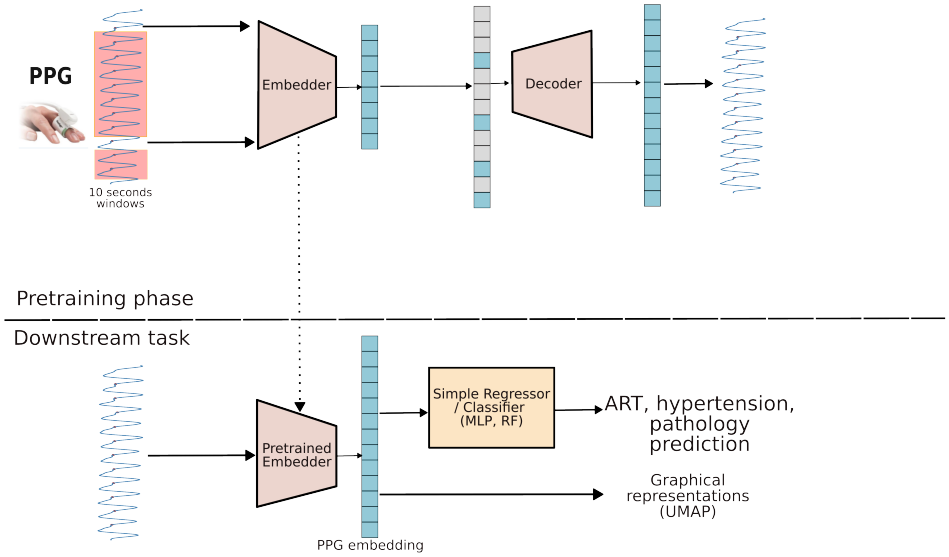


Figure 4.6: Masked autoencoder study design on semi-periodic signals (PPG)

The training hyperparameters were defined by a hyperparameter search performed using the library Optuna. Patches lengths were thought to include up to one cycle. We defined two patching methods, one based on linear

layers and another based on convolutional layers. The number of training epochs was controlled using early stopping. We used cosine annealing as a learning rate scheduler and Adam with warm restart as an optimizer. The associated search results are available in Table 4.3. Using these parameters, the training curves and a pretext task performance example is available in Figure 4.7.

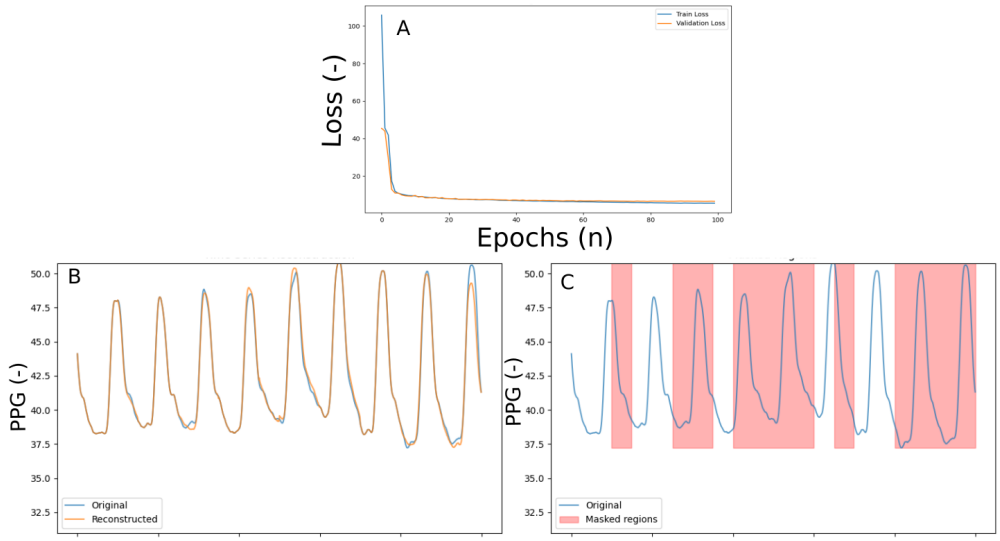


Figure 4.7: A. Masked autoencoder training and validation curves on pretext task. B. Example of masked autoencoder reconstructed PPG signal. C. Example masked regions on the PPG signal.

4.4.2 Visual representations of our results on the PPG

To assess downstream task performance, we generated the UMAP representations on the test set according to heart rate values. This variable is usually calculated from the PPG, and characterizes the capacity of the model to learn the frequency of the signal, although it has no direct link with the shape of the wave that is being learned. This representation shows that the embeddings has some embryos of clusters but do not yet show a clear gradient. The associated UMAP representations are available in Figure 4.8.

To evaluate the model's performance on long-term downstream tasks of interest, we aggregated a second dataset for external validation purposes. This dataset contains 657 PPG records of 219 patients along with comorbidities

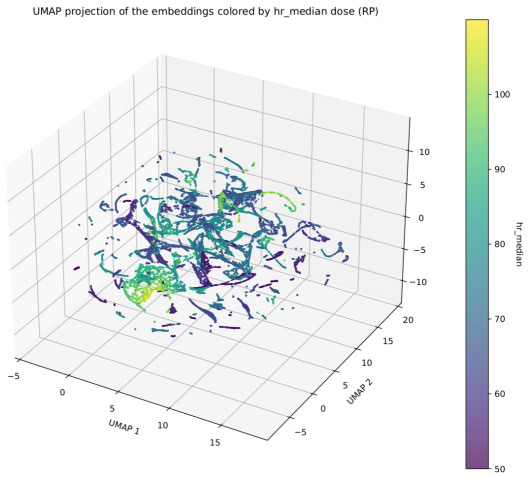


Figure 4.8: Masked autoencoder UMAP representation on the PPG colored by heart rate values.

Parameter	Search Space	Best Value
patch_len	[20, 50, 100]	50
embed_dim	[128, 256, 512]	512
encoder_depth	[6, 8, 12]	8
decoder_depth	[6, 8, 12]	6
num_heads	[4, 8]	8
mask_ratio	Uniform(0.5, 0.75)	0.576
batch_size	[64, 128]	128
learning_rate	LogUniform(1e-5, 1e-3)	2.62e-4
patch_method	["linear", "conv"]	"conv"

Table 4.3: Masked-Auto Encoder hyperparameter search space and best configuration for PPG

such as hypertension and diabetes, along with systolic and diastolic arterial pressures. This dataset is composed of two to three windows of 2.1 seconds of PPG per patients, sampled at 1000Hz. The dataset includes the according heart rate, systolic and diastolic blood pressure mean value. The patients age range from 20 to 89 years old and the majority of patients have hypertension. To adapt the PPG sequences to the trained MAE expected length, PPG sequences were resampled at 100 Hz, concatenated per patient, and repeated, tile and trimed to reach the equivalent of 1000-point-sequences. Finally, PPG sequences were standardized.

Task	This work	Chronos (FM)	Papagei-S (FM)
Hypertension ↑	0.64	0.57	0.77
Systolic BP (mmHg) ↓	16.9 [13.54–20.37]	16.91 [13.31–19.34]	14.65 [12.50–16.78]
Diastolic BP (mmHg) ↓	8.29 [6.61–10.22]	10.26 [8.13–12.57]	8.71 [7.18-10.01]
Heart rate (bpm) ↓	7.9 [5.89–9.84]	8.51 [7.05-10.07]	4.00 [3.34-4.67]

Table 4.4: Downstream tasks performance of the masked autoencoder using the PPG on the PPG-BP dataset - external validation cohort. Classification tasks such as hypertension are compared using area under the curve AUC - unitless, between 0 and 1. Regression metrics are presented as absolute error mean and 95% confidence intervals for consistency with previous studies.

At this stage, our model reached predictions better than chance, as shown in table 4.4. The masked autoencoder model reached an AUC of 0.64, with mean absolute errors similar to those observed by foundation models regarding systolic pressure, diastolic pressure and heart rate. However, hypertension prediction performance is still lower than the PPG-specific foundation model approach defined in Pillai et al., 2025. The plurality and volume of the pretraining dataset and rich morphology-aware metrics definition could explain this gap. Morphology-adapted approaches leveraging the learning of physiological characteristics on the PPG curve could be helpful in bridging this gap.

4.5 Discussion

4.5.1 BIS prediction

Unlike proprietary BIS algorithms, our approach provides a transparent prediction of the depth of anesthesia along with interpretable state representations. The SSL model performance is highly competitive with current research standards. Our approach demonstrated superior performance compared to supervised methods, particularly in low-label regimes, where the error rate decreased significantly from approximately 30% to 12% as the number of available labels increased. The SSL approach also significantly reduce the error variability. UMAP-based visualization offers innovative insights into anesthetic state structure. It also enables the distinction between

patient age groups and presumably characterizes the brain's response to anesthetics. As discussed in the introduction, patients are expected to reach a steady state, which corresponds to the anaesthetised state, during which the surgeon can operate. The trajectory to reach this state is a proxy of the brain's response to drug injection. By focusing on patients administered with remyfentanyl and propofol only, as opposed to desflurane and sevoflurane - gaseous anaesthetics that induce a different response from the brain, the cohort is homogenized towards drugs and enables the observation of trends. For all groups, we observed a loop on the trajectory, which characterises the route, and a different path during the induction and awakening of the patient. Focus on BIS prediction maintains the compatibility with established clinical workflows and provides insights into patient-level response to anesthetic drugs. These representations can enable us to take a further step forward in assisting future generations of clinicians regarding anaesthetic depth.

4.5.2 Masked autoencoders performance

Regarding masked autoencoders, this study showed that we were able to learn to reconstruct cyclic signals such as the PPG. The failures observed for other signals might indicate a more prevalent noise or artefacts for such signals. The embeddings showed, to some extent, a capacity to form clusters regarding continuous metrics such as the heart rate. In order to compare the performance to existing foundation models, the predictions on arterial pressure are kept calibration-free and do not take into account previous developments from chapters 2 and 3. The orders of magnitude are therefore a lot higher than in previous chapters. The masked autoencoder representation in figure 4.8 is however not as smooth as those of the EEG. Efforts are still needed to extract clinically relevant features from such signals. To date, following the path of large language models, approaches based on foundation models have soared. They address the multiplicity of datasets and potential tasks, and the explosion of computing power and cluster servers training efficiency. Indeed, morphology-aware modelizations outperformed their patient-proximity counterparts, notably on the PPG (Pillai et al., 2025). In this sense, one can wonder how these methods can be combined to learn

from the multiple modalities and propose a global understanding of the patient.

4.5.3 Limitations

First, this is a monocentric study. The findings would require to be validated on an external cohort. If the representations by age groups were conclusive, the classification between ASA groups was not as clear, perhaps as there were too few patients classified with a score of 4 and a form of subjectivity from the interviewer to classify between scores 2 and 3. To pursue our developments on other modalities, the cohort should reflect the entirety of the patients with BIS monitoring, and future studies could also extend these results to the full cohort of the VitalDB dataset - independently of other features - to strengthen the conclusions. Considering learning on semi-periodic signals, despite the visual ability to reconstruct the signals and converge on the training task, the current embeddings were not deemed sufficiently performant on downstream tasks of interest. These are very ambitious considering the signal itself. To reconstruct the curve, some beginnings of clusters could be observed on the UMAP representations, but they were not sufficiently clear to be interpreted. Predicting long-term comorbidities such as diabetes or hypertension from these signals is a challenging task, as these conditions may not manifest prominently in the short-term waveform data.

4.5.4 Perspectives

We highlighted the fact that trajectories on the manifold depend on patient age, observing in particular a gradation of curves along axis 2 in relation to age, which can be seen as a proxy for the brain's response to anaesthetics. One avenue for further research could be to parameterize these trajectories using a geometric approach. Mathematical formalism offers tools such as topological persistence that could measure the topological features of the curves obtained (Edelsbrunner and Harer, 2008; Bubenik, n.d.). If found successful, these methods could push the results one step further to an objective, quantitative, and topological characterization of the data manifold (Chazal et al., 2013).

The manifold also highlights the value of the "asleep medical appointment" concept. Then, considering semi-periodic signals, the use of more specific masking strategies or modelization methods to learn from semi-periodic signals could be useful. Because they show superior performance in cross-dataset downstream tasks of interest on the PPG, morphologically-aware modelling, though tested, still holds good promises for future experiments. Considering the performance across datasets, domain adaptation methods have shown promising results on EEG datasets and could be a key element to improve results (Gnassounou et al., n.d.).

4.6 Conclusions

In this chapter, we have leveraged a self-supervised learning approach on EEG high-frequency data to estimate the depth of anesthesia during GA. We used relative positioning, a pretraining task based on the learning of the concept of proximity of pairs of EEG windows in the time dimension. We achieved good results in the prediction of the bispectral index, which is a key biomarker for DoA quantification. UMAP graphical representations highlighted diverging trajectories for patients regrouped by age groups and opens up new perspectives on the use of these manifolds to understand the brain's response to drugs. Integrating cyclic features is still a challenging task, and preliminary results showed that representation learning using masked autoencoders has slightly improved its performance.

4.7 Appendix

4.7.1 Supplementary result: Relative positioning is ineffective in capturing periodic signals

Unlike the EEG, ECG, and PPG, which mostly carry a semi-periodic form of information as a repetition of a pattern, proximity is not naturally predictable from a neural network. Hyperparameter tuning did only enable us to reach suboptimal performances on the pretext task, and the learned visualizations

did not exhibit a clear gradient. An example on the arterial pressure is available on Figure 4.9. Clinically, one may assume that arterial pressure waveform should carry information regarding the dose of remyfentanyl administered. This is however an ambitious task regarding all the exogenous events which happen at the operating room. Downstream tasks' performance of interest was not exceptional: despite clinical relevance, pretraining on arterial pressure values was unable of discriminating regarding the administered drug dose. This could be because, unlike EEG, ECG, or PPG signals, which contain little long-range information, a well-informed convolutional network can capture this information from the EEG signal. Instead, other generic methods, such as signal reconstruction or informative feature retrieval, are more prone to exhibiting valuable information from the signals.

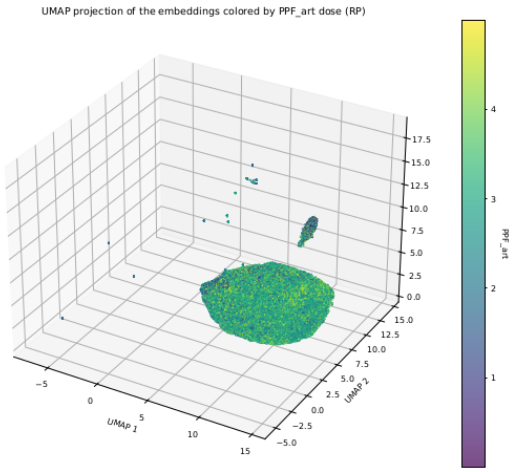


Figure 4.9: UMAP representations of the catheter arterial pressure colored by a color code based on vasopressor concentration.

4.7.2 Supplementary result: Fine-tuning the masked autoencoder on predicting the arterial pulse pressure

To improve the downstream tasks results of masked autoencoders, we tested the addition of a fine tuning step. More specifically, during this step, we performed fine-tuning by adding a regression head at the end of the masked autoencoder architecture. This MLP head is retrained for 20 epochs on an

auxiliary task, namely the prediction of the pulse pressure of the arterial pressure. The pulse pressure is defined as the difference between systolic and diastolic pressures. After having performed this step, the embeddings colored by the amplitude of the PPG clearly showed two distinct clusters. This could indicate that the amplitude of the PPG and the amplitude of arterial pressure are associated and that the model used one to predict the other. These results should, however, be comforted by other experiments. This result is available in Figure 4.10.

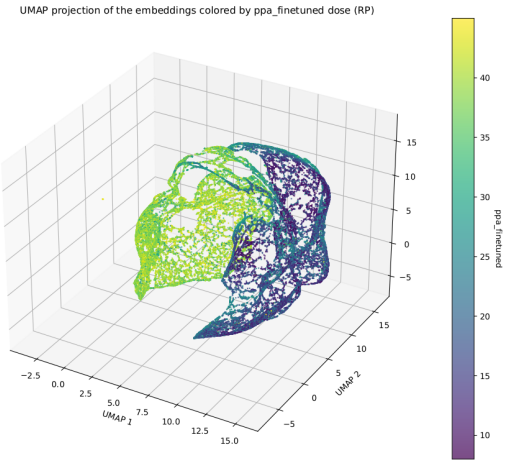


Figure 4.10: UMAP representation of the fine-tuned masked autoencoder, colored by the amplitude of the PPG. The manifold exhibits two distinctive clusters of extreme values of photoplethysmography pulse pressure. This result suggests that to learn arterial pulse pressure, the model has learned to discriminate between high and low PPG amplitudes.

4.7.3 Supplementary table: curvature statistics per age group

Supplementary Table ST1: Curvature statistics by age group across different resampling lengths

Subset (Median Length)	Age Group	Max	Std Dev
Test (503)	Senior	595.9	52.0
	Young	333.3	24.7
	Old	660.4	32.8
	Adult	919.6	56.5
Train (483)	Adult	541.1	48.6
	Senior	725.0	48.6
	Young	544.9	33.6
	Old	61666.9	2799.9
Validation (449)	Adult	231.3	28.6
	Senior	409.3	34.9
	Old	333.9	34.9
	Young	667.2	44.0

5

Real-time deployment perspective

We have now been through the main contributions of this manuscript. Before we conclude, we take the time to introduce here some the techniques at work to deploy our models in real-time. Indeed, throughout this thesis, we have constrained ourselves to use time series which are routinely monitored. This gives the opportunity to deploy them live at the operating room, and opens the door to engineering challenges. In this short chapter, we trace the developments relating to models deployment.

5.1 Model deployment at the operating room

As explained in the previous chapters, we have developed several models that validates international standards for blood pressure monitoring. After having assessed the model's performance, a crucial last step consisted in deploying it as a real-time microservice at the operating room. In a nutshell, this deployment can and is carried out through the interaction between two main applications.

5.1.1 Model serving application using FastAPI

The first application consists of serving the model via a software program that packages the trained deep learning model and exposes it as a service, typically through an Application Programming Interface (API), so other applications can use it to make predictions or perform tasks. It's the crucial step of taking a model from a development environment to a production setting where it can be used by real-world users or systems. This AI model serving application

has several key components. First, the machine learning model itself is saved in a - .pt - file that can be loaded into the application. This model contains the "learned intelligence" and is the core of the service. Then, a web framework like FastAPI is used to create the web server and define the API endpoints. It handles the incoming requests (like a POST request for most and for our case) and sends back the responses. A POST request is a method supported by the HTTP protocol used by the World Wide Web. By design, the POST request method requests that a web server accepts the data enclosed in the body of the request message, most likely for storing it. It is often used when uploading a file or when submitting a completed web form. The API endpoint, a specific URL within the application - in our case, and for most cases /predict - is the entry point for other applications to send data to the model. An endpoint is defined to handle a specific type of request, like a POST request for sending data to get a prediction. The Data format is validated using tools like Pydantic, which is integrated with FastAPI. It is used to define the expected format of the input data - usually a JSON object with specific features and data types. This ensures that the data sent to the API is in the correct format before it's passed to the model, preventing errors. Finally, an inference function: This is the main logic of the application. It receives the input features from the API request, loads the pre-trained model, pre-processes the input data into the format the model expects. Then, it calls the model's predict() function to generate the prediction. It is then formatted into a response - a JSON object - and sent back to the client. Essentially, such an application acts as a wrapper around the deep learning model. It provides a simple, standard interface using the API for a complex, non-standard component - the model. To package the application with all the necessary scripts and libraries, a widely used strategy is to containerize the application in a container, for which Docker is a natural tool. To avoid starting from scratch, i.e. reinstalling basic languages or processes, such as python, a common practice is to build containers on open-source prepackaged entities, called images, that are shared on online platforms, such as Docker Hub. In our case, and as the model uses graphics processing units (GPU), a relevant family of images are the images proposed by Nvidia that interact with the GPU and are intended to facilitate the deployment of applications that use the GPU.

5.1.2 A consumer application to process messages in real-time

Then, a second application is a Faust application which works by using a stream processing agent to continuously process messages from the Kafka topic of interest. The topics are arranged in three main categories: the patient identification, the numerics data - low frequency inputs - and the waveform data - high-frequency signals. As the model inputs are a combination of numerics and waveforms, the application listens to both topics. As each message arrives, the agent processes it and select the messages of interest based on their labels - which are the names of the signal. The selected messages are stored in an in-memory dictionary; a distributed key-value store, which acts as a buffer. In addition to the continuous stream processing, a separate, timed function is activated every few seconds. This timer is responsible for checking the buffer dictionary to see if there is any data to send. If the dictionary is long enough, i.e. if it's length is superior to the minimal length to generate a prediction, the function iterates through the buffered entries, converts each one into a JSON-friendly format, and asynchronously sends a POST request to the /predict endpoint of the FastAPI model serving application that is already defined. After a successful batch of requests, the buffer is cleared, and the cycle repeats, ensuring that the model is served with data efficiently in batches without having to be queried for every single incoming Kafka message. The generated predictions can then be converted to kafka messages and sent to the topic as a new signal identified by its label. This process is illustrated in Figure 5.1. For offline analysis purposes, they can be automatically saved in the datalake.

This basic deployment applications connect the real-time data platform to the model serving API. Future developments could include the development of error handling, retries for failed API calls, and metrics to monitor the rate of messages being processed and sent. Notably, Grafana is often used for deployed applications monitoring.

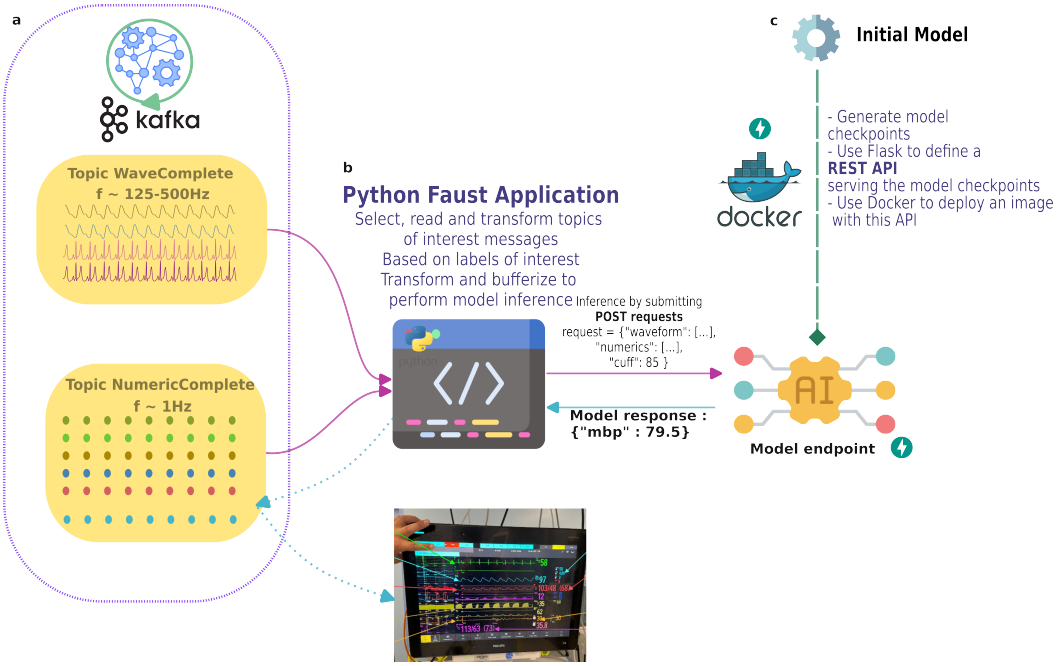


Figure 5.1: Deployment architecture at the operating room

5.1.3 Deployment perspectives

Current developments were focused on deploying a model that is trained offline, i.e. on a static dataset. However, the operating room is a dynamic environment where data is continuously generated. After real-time deployment is performed, the next natural step is to leverage this dynamic and continue the learning process while discovering new patients characteristics. This leverages both theoretical paradigms and engineering building blocks. To adapt to this changing environment, the model could be updated in real-time using online learning techniques. More specifically, online learning is known to avoid prediction drifts, a common pitfall observed in real-time deployment processes.

6 Conclusion and perspectives

It is now soon time for us to conclude. Here, we summarize our work and give a few of the perspectives for future research.

6.1 Summary of contributions

In this thesis, I addressed the challenge of generating actionable biomarkers for anesthetists using deep learning methods on large-scale hospital databases. We have taken particular care to restrict ourselves to the use of non-invasive features, so as not to necessitate additional sensors. This provides a real-time deployment opportunity that has been tested on the hospital data infrastructure. Considering clinical decision support during the operation, we focused on the prediction of near-real-time metrics of the operating room, namely the arterial pressure and the bispectral index. This thesis explored the methodologies to predict biomarkers from simultaneous high frequency time series.

When we started the work on our thesis, a feature-based approach had just been tested on the preliminary cohort. The technology and the deep learning environment has been evolving rapidly over the past few years, and it oriented some of the technical choices of this thesis. A first natural line of research was to bridge between clinical insights and machine learning research. This led to the inclusion of PPG-derived features and the comparison of pure deep learning methods to a mechanistic approach in chapter 2. In this chapter, we proposed a deep learning model based on both clinically insightful metrics such as the dicrotic notch and the perfusion index, and the raw PPG signal. We demonstrated that the hybrid CNN + LSTM reduced the error variability by a factor of three compared to a mechanistic approach.

Also, both the raw signal and aggregated features were useful in that their synchronized use reduce error variability by a factor of two compared to derived features sole use. Grad-CAM representations showed that the CNN model attention was localized around the dicrotic notch. Despite the recent surge of approaches based on large language models or foundation models, which suggest that a learning capacity is primarily a function of model scale and data volume, clinical acceptance is still highly dependent on physiological consistency. This also guided our thoughts in the representation of model explainability, or the UMAP representations in chapter 4.

Considering the prevalence of benchmarks in tabular data or medical imagery (Moreau et al., 2022), this practice was an inspiration to compare arterial pressure monitoring methods in a fair way. I was interested in addressing the subject using a benchmark angle and proposing a fair comparison of deep learning methods. In chapter 3, we conducted a multicentric study and evaluated both calibration-based and calibration-free approaches on two large scale databases of 5060 and 2833 patients each. We demonstrated the superiority of our approach using causal convolutions and LSTM inner state initialization. Calibration remains a central topic for arterial pressure estimation and more recent architectures could be leveraged to address it too.

Exploring self-supervised learning in chapter 4 also widened the scope of our research. We extended our studies to the electroencephalogram to address the monitoring of the depth of anesthesia. Our SSL model used relative positioning and demonstrated superior performance in prediction the bispectral index compared to existing supervised approaches. Compared to a 1D CNN from Chambon, Thorey, et al., 2018, the error variability was reduced by 44%. Our UMAP manifolds on the EEG highlighted the fact that patients' trajectories showed a clear gradient regarding age, describing a loop which is consistent with clinical intuition. To some extent, these results answer the question of the relevance of using modern deep learning methods at the operating room. Indeed, the learned embeddings structure provide new visualizations of the response of the body to the drugs, and pushes further the understanding of complex concepts while maintaining clinical acceptance. Though masked autoencoding has not provided similar results

yet, we are confident that a mix of morphology-aware and reconstruction pretext tasks will ultimately be found successful.

Then, we delved into the real-time deployment specificity of our project. Indeed, our frugal approaches reduce the deployment cost to real-time technical issues and require no additional sensor implementation. Beyond the contributions and answers presented in this thesis, our research has uncovered several questions that could guide future research efforts.

6.2 Open questions

6.2.1 Cuffless blood pressure monitoring perspective

Over the course of this thesis, we have seen the thrive of studies proning the possibility to perform cuffless blood pressure (BP) monitoring (El-Hajj and Kyriacou, 2021). If measurement practices are not about to be changed in the operating room, a growing part of the hypertension research community is interested in remote monitoring devices (Stergiou et al., 2022). The latest advances in electronics, such as the development of portable mechanical and optical sensors incorporating signal processing or machine learning modules, offer a technical alternative to blood pressure cuff oscillometers, which patients may find cumbersome to use at home (Hua et al., 2024). Despite the promises of studies proposing deep learning models to address this issue, the results must be examined carefully. A comparative study conducted comparing measurement conditions in the laboratory and in the operating room showed differences in blood pressure variability depending on age group, measurement environment, and the sensor used for measurements (Mieloszyk et al., 2022; Mukkamala et al., 2023). Indeed, in the conclusion section of chapter 3, we mentioned the fact that despite their widespread use by clinicians, international standards requirements for arterial pressure are based on error bounds instead of more representative metrics, such as those based on the explained variance. Studies often present a fairly new method that outperforms previous ones, but are these metrics sufficient to prove the superiority of a newly developed method? In medical imaging research, machine learning studies have been shown to incur methodological pitfalls

regarding the strength of drawn conclusions (Varoquaux and Cheplygina, 2022; Radford et al., 2021). In this realm, studies have proposed adjusted metrics and methodologies for thorough results reporting (Maier-Hein et al., 2024). As pointed out by the European society of Hypertension (Stergiou et al., 2022), common pitfalls for blood pressure prediction performance include the representation of predicted BP versus measured BP without taking intra and inter patient variability into account, which can result in misleadingly high performance regarding intra-patient performance prediction. Particular attention should be put on the stability of post-calibration measurements.

6.2.2 Prediction of uncertainty

While learning-based models offer numerical predictions, these values should be treated with caution. Uncertainty remains, which may be linked to variables that have not been taken into account, such as the environment, or the nature of the surgery. Methods exist to reduce the impact of this uncertainty and to clarify the model's confidence in its predictions. For classification problems, the prediction probability of each class is usually used to quantify uncertainty. This is commonly referred to as the confidence index. Developing a confidence index for regression models requires specialized paradigms for quantifying predictive uncertainty, especially in high-stakes domains such as clinical decision support. Traditional point estimates offer no measure of reliability, making uncertainty quantification a crucial step toward actionable model outputs. Quantile regression and its extension through conformal prediction provide distribution-free statistical tools to construct valid predictive intervals with guaranteed coverage (Angelopoulos and Bates, 2022; Romano et al., 2019). These methods are attractive for clinical use because they do not require strong distributional assumptions and remain robust under model misspecification.

In contrast, the Bayesian framework integrates prior knowledge with observed data to derive posterior distributions over model parameters or predictions (Shen, L. Liu, et al., 2022; Kendall and Gal, 2017). This yields probabilistic predictive intervals that can incorporate domain expertise and hierarchical structures common in biomedical data. Approximate Bayesian deep learning techniques, such as Monte Carlo dropout (Gal and Ghahramani,

2016) , deep ensembles (Lakshminarayanan et al., 2017), and variational inference, offer scalable implementations for complex regression architectures, enabling uncertainty estimation without exact posterior computation.

Comparing these approaches reveals a trade-off between statistical guarantees, computational efficiency, and interpretability. Quantile or conformal methods ensure finite-sample validity but may ignore uncertainty in model parameters, while Bayesian-inspired approaches capture both epistemic and aleatoric uncertainty at the cost of heavier computation. Promising results on conformal prediction applied to wearable sensors data pave the way for future investigations (Shen, Chakraborti, et al., 2025).

6.2.3 Multimodal learning

As we observed throughout the manuscript, learning from operating room data implies to extract information from multiple sources. On top of technical challenges incurred by the variety of sensors, frequencies, and technical characteristics, it also holds challenges in the modality fusion method (Gao et al., 2020; Baltrusaitis et al., 2019; S. Li and Tang, 2024). Many challenges remain, particularly with regard to self-supervised learning applied to semi-periodic data. Beyond our work, limited to time series data, future developments may include additional sources categories-clinical data, events occurring during GA or syringe pump injected drugs values are immediate candidates to strengthen our analyses. Alignment, noise resistance or disparities in features representations are keys challenges of multimodal learning. This was found particularly true for self-supervised learning. Applying on-the-shelf methods was not deemed satisfactory for all signals, and a particular attention had to be put on the design of modality specific pretext tasks. Indeed, despite the plurality of data fusion methods, reviews showed that methods performance were highly dependent on the data and task at hand (Krones et al., 2025). Emerging research highlights the necessity of tailoring modality-specific pretext tasks, such as contrastive objectives tailored to physiological time series, to ensure robust feature learning and downstream transferability R. Ding et al., 2023.

List of abreviations

AAGA accidental awareness during general anesthesia. 9, 92

A-line arterial line. 5, 13

API Application Programming Interface. 117, 118

AUC area under the curve. xxii, 100, 110

BIS Bispectral index. xvi, 7, 93, 97, 98, 101, 110

BP Blood pressure. 66, 123, 124

CDW clinical data warehouse. 18

CNN Convolutional Neural Network. 30, 49, 68, 121, 122

DoA Depth of anesthesia. 5, 7, 26, 92, 93, 101, 113

DWC Data warehouse connect. 13

ECG Electrocardiogram. xviii, 3, 9, 38, 106

EEG Electroencephalogram. xviii, 5, 7, 9, 15, 38, 92, 94, 96, 105, 122

GA general anesthesia. xiii, 1, 4, 14, 17, 22, 24, 25, 30, 95, 113, 125

GPU Graphical Processing Units. 29, 118

Grad-CAM Gradient-weighted Class Activation Mapping. 37, 46, 122

HPI Hypotensive prediction index. 24

HR Heart rate. 48

ICU Intensive care unit. 19

IOH Intraoperative hypotension. 9–12, 24, 99

LSTM Long Short Term Memory Unit. 31, 49, 68, 121

MAE Mean Absolute Error. 29, 51

MAP Mean Arterial Pressure. 4, 11, 28, 44, 67

ME Mean error. 51

MIMIC Medical Information Mart for Intensive Care. 19

MSE Mean Squared Error. 29, 34

NRA Notch relative amplitude. 44, 48

PAT Pulse arrival time. 66

PI Perfusion index. 44, 48

PPG Photoplethysmography. xviii, xxii, 4, 9, 24, 38, 42, 98, 107–111, 115, 121

PWV pulse wave velocity. 66

RMSE Root Mean squared error. 51

RNN Recurrent Neural Network. 31

SSL Self-supervised learning. 32, 91, 94, 96, 102, 106, 110, 122

TT transit time. 66

UMAP Uniform Manifold Approximation and Projection. 99, 101, 108, 113, 122

XAI Explainable artificial intelligence. 36

7

Appendix: related articles not presented during the thesis

Oriane Saint Aubin, Ines Khemir, Jade Perdereau, Cyril Touchard, Fabrice Vallée, Jérôme Cartailler. *Repurposing electroencephalographic signal for automatic segmentation of intra-operative periods under general anesthesia*. IEEE EUROCON 2023-20th International Conference on Smart Technologies, 286-290

Repurposing electroencephalographic signal for automatic segmentation of intra-operative periods under general anesthesia

O. Saint Aubin*, 1, 2, I. Khemir*,3, J. Perdereau1, 4C,. Touchard1, F. Vallée1, 5, 6, J. Cartailler1, 5, 4, †, *Member, IEEE*

***Abstract*— During general anaesthesia, brain response during loss (LoC) and recovery (RoC) are informative about cerebral fragility. These periods are yet difficult to assess. We propose an automatic method to identify Loc and Roc from intraoperative electro-encephalogram EEG. Dataset includes 289 intraoperative electroencephalograms (EEG) of patients under general anaesthesia (propofol or sevoflurane) for which LoC and RoC have been labelled by qualified clinician. EEG was decomposed into 30 seconds long epochs, where power spectral density was used as features. We used a gradient-boosting (GB) algorithm with a 3-state label to classify states of unconsciousness. LoC and RoC times were identified by fitting a logistic function on the sequence of probability from classifier. GB test accuracy was 95.58% for a micro-average AUC of 0.994. LoC and RoC were predicted with a median[IQR] error of 25.4[-82.1; 77.2] sec and -15.7[-65; 26.5] sec respectively. We showed that LoC and RoC in the perioperative period can be automatically identified using only the intraoperative EEG. Our work opens the possibility of studying LOC/ROC-based biomarkers for large volumes of data.**

***Clinical Relevance*— Our method provides an automatic mean to identify brain critical period in response to general anesthesia that are loss and recovery of consciousness.**

***Index Terms*— Electroencephalography (EEG), loss of consciousness, recovery of consciousness, general anesthesia, gradient boosting.**

I. Introduction

Electroencephalographic (EEG) monitoring during general anaesthesia (GA) can help to assess patient brain fragility, especially during loss (LoC) and recovery (RoC) of consciousness [1]–[3]. However, identifying and labelling these transition periods on EEG requires qualified practitioners and time. Previously, Subramanian and colleagues developed an automatic method based on 9 patients that successfully identifies loss and regain of consciousness during propofol anaesthesia. They used heart rate, blood pressure and electrodermal activity as features [4]. This method relies on multimodal inputs that could be difficult to get in daily practice or may not be available for retrospective investigation. We propose a segmentation method using only EEG signals from the Fp2 frontal electrode to detect both LoC and RoC. The EEG signal comes from a depth-of-anesthesia monitor used routinely in the operative room. We first used a gradient-boosting method to classify state of unconsciousness based on 289 intraoperative EEGs, then applied logistic regressions to identify LoC and RoC times.

II. Materiel and Methods

A.Data and features

Intra-operative raw EEG traces were collected during neuroradiology or orthopaedic interventions using the Sedline brain function monitor (Masimo Corporation©, Irvine, California, USA). Settings parameters stayed fixed during intervention for all patients [5]. GA was induced with propofol and maintained with propofol or sevoflurane (details on GA in Supp. Mat.). Data were collected between November 2018 and December 2020 at Lariboisiere hospital (France) under the ClinicalTrials.gov Identifier: NCT04481347.

A clinician qualified in EEG reading identified times of LoC and RoC. Unconscious state was defined from EEG signal according to changes in the frontal alpha or beta bands concomitant to propofol injection. We use three labels to classify patient state: 0 before LoC, 1 during unconscious state and 2 after RoC (Fig. 1.A).

Affiliations:1: Department of Anesthesia and Critical Care Medicine, AP-HP, Hôpital Lariboisière, F-75010, Paris, France; 2: Université de Paris, UFR Mathématiques et Informatique F-75006, Paris, France; 3: EPITA Majeure Santé, 94270, Le Kremlin-Bicêtre, France; 4: Fondation de l'Assistance Publiques - Hopitaux de Paris, 75004 Paris. 5: Université de Paris, INSERM, U942 MASCOT, F-75006, Paris, France; 6: Equipe M3DISIM, Inria-Saclay, Université Paris Saclay, 91190 Saint-Aubin, France.

*Equally contributing author; †Corresponding author: jerome.cartailler@inserm.fr

EEG signal was not pre-processed (no filtering or artefacts removal) before computing predictors for our machine-learning algorithm. Building classifier features, we first down sampled all traces to 63Hz to match the minimal sampling frequency in our dataset. We then decomposed the EEG signal into 30 seconds long epochs (Fig. 1.B). For each epoch we computed the power spectral density (psd) using Welch's method (Fig. 1.C). The resulting psd, in dB, was linearly interpolated every 0.5Hz for discretization. The 50 values ranging from 0.5 Hz to 25 Hz were kept as input features. Consequently, for each epoch we had 50 features computed from a single EEG trace (Fig. 1.D).

B.Machine learning model

We first classified state of arousal for each epoch. We used a 3-states gradient boosting method (GB) to handle unbalanced sets and possible nonlinear features interactions. Then we identify LoC and RoC times by fitting two logistic functions $\left(1+e^{-a_1(t-t_{LoC})}\right)^{-1}$ for the LoC and $1-\left(1+e^{-a_2(t-t_{RoC})}\right)^{-1}$ for the RoC) over the computed sequence of probability obtained from the GB classifier. Identification of parameters $(a_1,\ t_{LoC},\ a_2,\ t_{RoC})$ was done using a soft L1-regularisation in the loss function (see Supp. Mat.). A schematic representation of the process is shown (Fig. 1.E-F).

For testing, 20% of the patients were randomly selected and set aside (TEST_Set) and 80% were kept for training (TRAIN_Set). We corrected label unbalance in TRAIN_Set – patients spend more time being unconscious during general anaesthesia (label ratio 1/10) – by keeping within each patient, all the n_0 epochs labelled 0 or 2 (awake), then randomly selecting $n_1=2n_0$ epochs labelled 1. We trained our algorithm with a stratified k-fold (k=3) cross validation method. Within each fold, we estimated GB hyper-parameters using halving grid search cross-validation (see Supp. Mat.). Classifiers over each fold were pooled into a voting ensemble (soft voting) that we challenged on the TEST_Set. We reported balanced accuracy and micro-averaging AUC obtained by applying our classifier on TEST_Set and TRAIN_Set. At this stage we had output as sequences of probability (one versus the rest) for being unconscious (state 1), computed for each epoch. This time series was smoothed by a 5-epochs rolling-average, then we applied a tanh function to push small probabilities towards 0 and larger ones towards 1. To predict LoC/RoC times, we fitted logistic functions and collected parameters $t_{_LoC}$ and $t_{_RoC}$. Error in prediction time between estimated and gold standard LoC (resp. RoC) was denoted $\Delta_{_LoC}$ (resp. $\Delta_{_RoC}$) and reported in seconds. Prediction residue computed with a 42 random seed was reported as IQR. The standard deviation was then obtained running 50 times the full training process, including the 20%/80% TEST_Set/TRAIN_Set random split.Variable importance was computed using Shapley additive explanations (SHAP) method [6]. Model training and performance evaluation were carried out using the Python-based scikit-learn librairies [7].

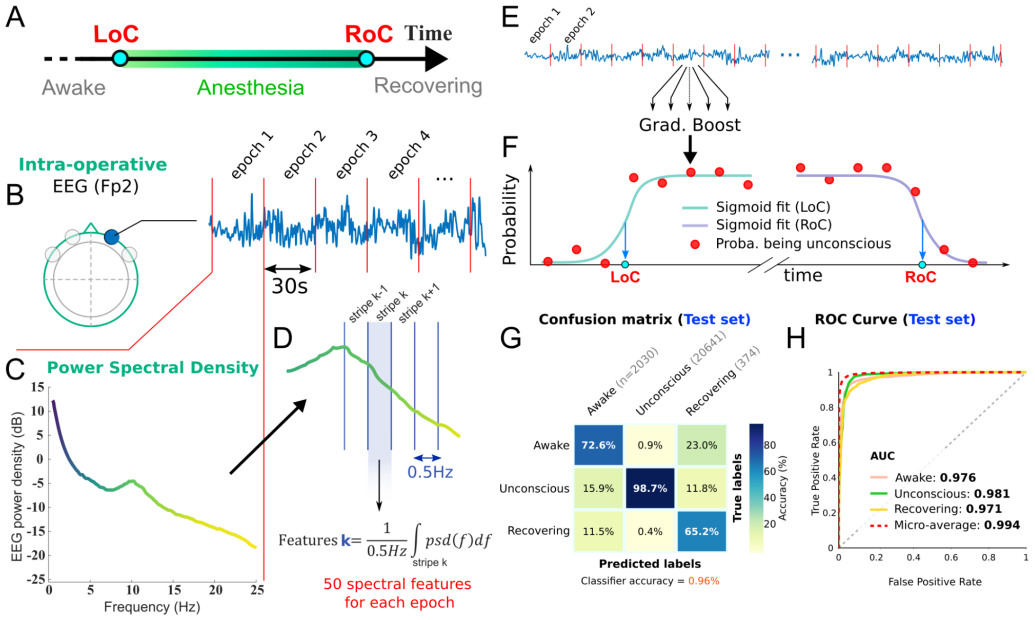


Figure 1: Schematic of EEG features design, unconsciousness classification and Loc/RoC estimation.
A: Loss (LoC) and recovery (RoC) of consciousness within the perioperative time course.
B: Schematic representation of EEG electrode positioning and trace 30s long epoching.
C: Example of intra-operative power spectral density (psd, colours linked to figure 2A).
D: Classifier features computation from psd discretisation over 0.5Hz windows.
E-F: From each EEG epochs, 50 spectral features are computed and used to estimate the probability of being unconscious (red dots). A logistic function is fitted (blue and purple) to determine LoC and RoC times (cyan dots).
G: Confusion matrix obtained from GP classifier evaluated on the test set.
H: ROCAUC over test set for each GP labels and unconsciousness "one versus the rest" (micro-averaged).

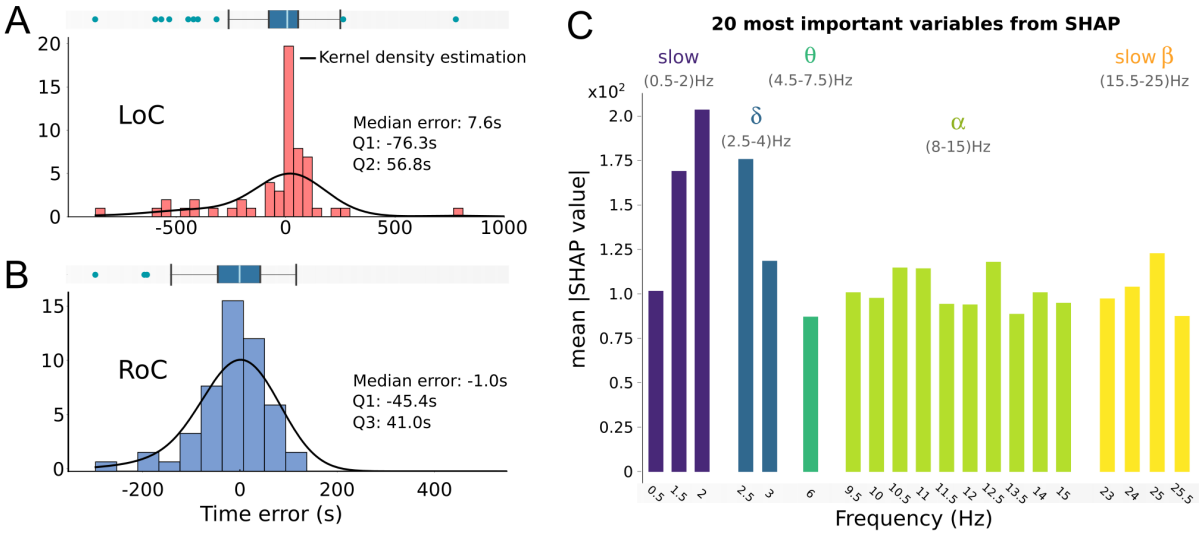


Figure 2: Impact of spectral features and Loc/Roc error distribution.

A-B: Δ_{LoC} and Δ_{RoC} distribution computed from all 293 patients.

C: 20 most important features from GB classifier based on SHAP values. Colors are used to distinguish common EEG oscillations.

III. Results

GP classifier accuracy on Test_Set was 95.58% with a micro-averaged AUC of 0.994. Comparatively we obtained on Train_Set a 97.90 % and a micro-averaged AUC of 0.998 on TRAIN_Set (see. Supp Material). From all patient pooled, we found an error Δ_{LoC} values of Median[Q1-Q3]=25.4[-82.1, 77.2]s (see Fig. 2A) and 15.7[-64.9, 26.5]s for Δ_{RoC} (see Fig. 2B). Importance variable analysis revealed that the most important components on the spectrum concern the low frequency waves delta (≤ 4 Hz) or alpha-beta frequency range (9-25 Hz, fig 2. C, Supp. Fig S1).

IV. Discussion

We showed that perioperative LoC and RoC can be automatically identified using intraoperative EEG with an error of 25.4s and -15.7s respectively. We developed a supervised machine learning algorithm - combining a classification and a regression step - based on 289 unprocessed EEG. Our method relies on a single sub-hairline electrode and a depth-of-anaesthesia device routinely used for anaesthesia monitoring.

From a clinical perspective, previous work found that brain response to GA collected the first ten minutes right after LoC could reveal cognitive abnormality [1]. During this same period, it is also possible to anticipate occurrence of burst-suppression, an EEG pattern marker of too deep anaesthesia causing postoperative delirium among a fraction of fragile patients [2], [8]. The possible reason why the period right after LoC is informative may be because it reflects brain response to the stressor that represents GA. Similarly, tagging RoC may help identifying factors of delayed emergence that represent an inadequate recovery possibly followed by complications [3], [9].

Currently, EEG inspection is a tedious process where trained neurophysiologists review the signal 'page by page'. Therefore, awake/unconscious state segmentation represents a possible bottleneck for anaesthesia research. Automated LoC and RoC detection could promote intra-operative EEG research on large volumes of data. In addition, the present method is parsimonious as it relies only on EEG traces that contrast with existing multimodal approaches [4].

The RMSE, 14.79s and 20.56s, we found is pertinent as sleep scoring performed by trained clinicians is done over 30s long windows [10]. Interestingly, we had better results considering three different labels (awake, unconscious, recovering) compared to a binary classification (awake/unconscious). This agrees with physiological observations describing non-symmetrical behaviors whether a patient loses or recovers consciousness [11], [12].

We can report several limitations to be considered in future work. First, we define loss and recovery of consciousness as a significant change in the patient's EEG determined by a qualified clinician and that is consistent with propofol pharmacodynamics. However, various definitions co-exist and therefore the present algorithm may not fully agree with all of them. In addition, our method works for GA induced with propofol and maintained with either propofol or halogenated gases, both enhancing inhibitory GABA-ergic currents (gamma-aminobutyric acid). This method is not validated for drugs such as (S-)ketamine or dexmedetomidine that lead to very different EEG signatures [13]. Finally, our method works offline and thus supplementary developments need to be done in order to provide an online solution, for instance using filter theory.

V. Acknowledgment and Fundings

J.C. research was funded by Fondation APHP. O.S.A. is financed by Année recherche (Ministère de la Santé), I. K. is financed by Epita and APHP.

VI. Conflict of Interest

The authors declare having no conflict of interests

References

[1] J. Cartailler, C. Touchard, P. Parutto, E. Gayat, C. Paquet, and F. Vallée, "Brain fragility among middle-aged and elderly patients from electroencephalogram during induction of anaesthesia," *Eur. J. Anaesthesiol. EJA*, vol. 38, no. 12, pp. 1304–1306, Dec. 2021, doi: 10.1097/EJA.0000000000001524.

[2] J. Cartailler, P. Parutto, C. Touchard, F. Vallée, and D. Holcman, "Alpha rhythm collapse predicts iso-electric suppressions during anaesthesia," *Commun. Biol.*, vol. 2, no. 1, pp. 1–10, 2019.

[3] M. Cascella, S. Bimonte, and R. Di Napoli, "Delayed Emergence from Anaesthesia: What We Know and How We Act," *Local Reg. Anesth.*, vol. 13, pp. 195–206, Nov. 2020, doi: 10.2147/LRA.S230728.

[4] S. Subramanian, R. Barbieri, P. L. Purdon, and E. N. Brown, "Detecting Loss and Regain of Consciousness during Propofol Anaesthesia using Multimodal Indices of Autonomic State," *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. IEEE Eng. Med. Biol. Soc. Annu. Int. Conf.*, vol. 2020, pp. 824–827, Jul. 2020, doi: 10.1109/EMBC44109.2020.9175366.

[5] F. von Dincklage, C. Jurth, G. Schneider, P. S García, and M. Kreuzer, "Technical considerations when using the EEG export of the SEDLine Root device," *J. Clin. Monit. Comput.*, vol. 35, no. 5, pp. 1047–1054, Oct. 2021, doi: 10.1007/s10877-020-00578-9.

[6] Y. Nohara, K. Matsumoto, H. Soejima, and N. Nakashima, "Explanation of Machine Learning Models Using Improved Shapley Additive Explanation," in *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, New York, NY, USA, Sep. 2019, p. 546. doi: 10.1145/3307339.3343255.

[7] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Mach. Learn. PYTHON*, p. 6

[8] B. A. Fritz *et al.*, "Preoperative Cognitive Abnormality, Intraoperative Electroencephalogram Suppression, and Postoperative Delirium: A Mediation Analysis," *Anesthesiology*, vol. 132, no. 6, pp. 1458–1468, 2020, doi: 10.1097/ALN.0000000000003181.

[9] G. A. Mashour *et al.*, "Recovery of consciousness and cognition after general anaesthesia in humans," *eLife*, vol. 10, p. e59525, May 2021, doi: 10.7554/eLife.59525.

[10] Raman K Malhotra and Alon Y Avidan. Sleep stages and scoring technique. Atlas of Sleep Medicine, pages 77–99, 2013.

[11] M. Cascella, S. Bimonte, and M. R. Muzio, "Towards a better understanding of anaesthesia emergence mechanisms: Research and clinical implications," *World J. Methodol.*, vol. 8, no. 2, pp. 9–16, Oct. 2018, doi: 10.5662/wjm.v8.i2.9.

[12] V. Tarnal, P. E. Vlisides, and G. A. Mashour, "The Neurobiology of Anaesthetic Emergence," *J. Neurosurg. Anesthesiol.*, vol. 28, no. 3, p. 250, Jul. 2016, doi: 10.1097/ANA.0000000000000212.

[13] P. L. Purdon, A. Sampson, K. J. Pavone, and E. N. Brown, "Clinical Electroencephalography for Anesthesiologists: Part I: Background and Basic Signatures," *Anesthesiology*, vol. 123, no. 4, pp. 937–960, Oct. 2015, doi: 10.1097/ALN.0000000000000841.

[14] Lundberg S, Lee S-I, "A unified approach to interpreting model predictions", *Adv. Neural Inf. Process. Syst*, 1–10 (2017).

[15] Lundberg S, Erion G, Lee S-I, "Consistent individualized feature attribution for tree ensembles", *arXiv:1802.03888 (2018)*

Supplementary Materials

Patient selection and intraoperative data collection: Between November 2018 and December 2020, were selected patients eligible for endovascular embolization of a neurovascular disease or an orthopaedic surgery performed under general anaesthesia. Nevertheless, patients with a history of neurosurgery, aneurysmal subarachnoid haemorrhage, or intracerebral lesions were excluded. This study is part of the SOMEDYSPO perspective, monocentric and observational study was a prospective, observational, single-center study (ClinicalTrials.gov Identifier: NCT04481347).

Frontal EEG data were recorded using the Sedline brain function monitor (Masimo Corporation©, Irvine, California, USA) with recording subhairline electrodes located at approximately at Fp1, Fp2, F7, F8, a ground electrode at Fpz and a common reference electrode 1 cm above Fpz. Raw EEG traces were exported as European Data Format files (.edf) and analysed using Python 3.8.8. Channels were rejected if the electrode impedance was greater than 5 kΩ.

Anaesthetic protocol: General anaesthesia was done in a standard fashion. It was always induced with propofol at a dose of 2.5mg/kg. Analgesia was done administering opioids: sufentanil 0,2-0,5 μg/kg for orthopaedic procedures and remifentanil with an effect site concentration of 3-5ng/ml which pharmacokinetic controlled using syringe drivers based on Minto's model for neuroradiology procedures. Finally, a non-depolarizing neuromuscular blocker (curare) was administered during induction (atracurium besylate). Notice that some patients received Ketamine during the induction at the dose of 0.1-0.3 mg/kg. Departmental protocol indicates to maintain intraoperative blood pressure above 90% of its preoperative baseline, a task achieved by administering diluted norepinephrine (5 μg/ml). During maintenance of anaesthesia (roughly 15-20min after beginning of GA), patients were administered either intravenous propofol or volatile sevoflurane. The propofol target concentration was chosen by the anaesthetist in charge and maintained automatically by syringe drivers using Schnider's model and ranged between 3 and 5ug/ml. For sevoflurane, end-tidal gas concentration was initiated at 1.5 to 2%. For both propofol and sevoflurane, departmental protocol dictates maintaining the spectral edge frequency (SEF_{95}) from depth-of-anaesthesia monitor (Sedline) between 8 and 15 Hz and avoiding intra-operative burst-suppression as much as possible.

Table 1 : Cohort description

Age (year)	
Median (Q1, Q3)	56 (37, 69)
Min - Max	18 - 99
Gender	
Female	213 (64.5%)
Male	117 (35.5%)
ASA (Physical status score)	
ASA1	69 (21.3%)

ASA2	185 (57.1%
ASA3	70 (21.6%)
Maintenance drug (number of patient)	
Propofol	196 (66.7%)
Sevoflurane	98 (33.3%)

Signal processing: Use of Fp2 electrode alone has been motivated by the output data from the depth-of-anaesthesia monitor. Indeed, collect .edf file report signals visualised on screen which results in clipped signals on the other electrodes which affect in particular Fp1 and F8 electrodes. The remaining files were downsampled to a common sampling frequency of 63Hz before concatenation. From 293 EEG, we obtained 34309 30s epochs for the Train-SET and 24773 30s epochs for the Test-SET ((**Fig. 1.B,C,D**)

Machine Learning model

Classifier: Classifier hyperparameters were tuned on the training set (Train_set), which was further split using a 3-fold cross-validation split based "HalvingGridSearchCV" class. The kept hyperparameters are the following:

Table 2 : Hyperparameters settings

loss	*log_loss*
learning rate	0.25
min_samples_leaf	1
max_depth	10
min_samples_split	2
max_features	6
n_estimators	125
subsample	1.0
criterion	*friedman_mse*
min_weight_fraction_leaf	*0.0*
min_impurity_decrease	*0.0*

SHAP values: We decided to exploit SHapley Additive exPlanation method as a variable importance metric for our GB based classifier (SHAP library). We briefly explain the idea behind such an approach to ease the interpretation of the results presented in the paper. The algorithm was initially proposed by Lundberg and al. in 2017 [14] to improve consistency of existing feature attribution methods for tree ensemble algorithms. It is based on shapley values that come from cooperative game theory. Shapley value is the average of the marginal gain for players in order to fairly distribute the gain among cooperating players. Each player *{i}* will be assigned a value that represents the effect of him playing on the total gain (f_x). To compute this effect, a model $f_x(S\cup\{i\})$ is trained with that feature *{i}* present, and another model $f_x(S)$ is trained with the feature withheld. Then, predictions from the two models are compared on the current input $[f_x(S\cup\{i\})-f_x(S)]$. Since the effect of withholding a feature depends on other features in the model, the preceding differences are computed for all possible subsets $S\subseteq M\backslash\{i\}$:

$$\phi_i = \sum_{S\subseteq N\{i} \frac{|S|!(M-|S|-1)!}{M!}[f_x(S\cup i)-f_x(S)]$$

where, S a subset of non-zero features used in the model, N a set of all input features and M the total number of features. In other words, the shapley value of a feature is its contribution to the payout, weighted and summed over all possible feature value combinations.

In the SHAP method, i are the Shapley values of a conditional expectation function of the original model where $f_x(S) = E[f_x|x_s]$ is the prediction function conditioned on the feature values in the subset S. The SHAP method is additive and can be written as a sum of the effects ϕ_i of each feature X_i (figure 1).

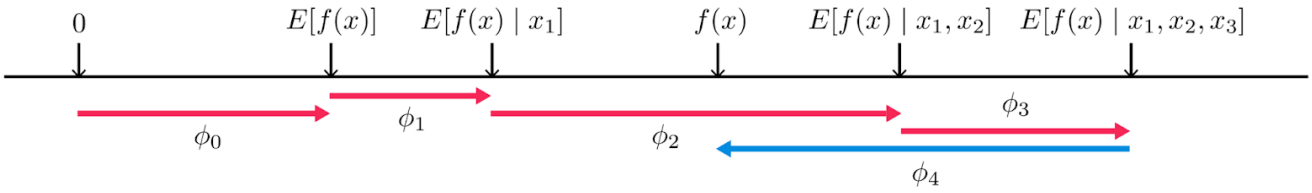


Figure 1 : SHAP values concept developed by Lundberg and al. [14].

Because of the exponential complexity of the shapley equation, Lundberg et al proposed in 2018 an algorithm that runs in $O(TLD)^2$ [15].

In our model, *f* is a gradient boosting model and *x* is the set of all the frequencies (50 features) from a certain epoch. Interestingly, the important features integrates by our algorithm to detect the loss of consciousness are coherent to neurophysiological states:

- Beta oscillation, seen during propofol induced sedation (paradoxical excitation) (features between 26 and 49).
- Alpha oscillations shift to the front of the head with loss of consciousness (features between 18 and 25).
- Slow delta oscillation anaesthetic state corresponding to the decreasing excitatory inputs from the thalamus and the brainstem. The appearance of this slow oscillation means that the patient sleeps (features between 0 and 7).

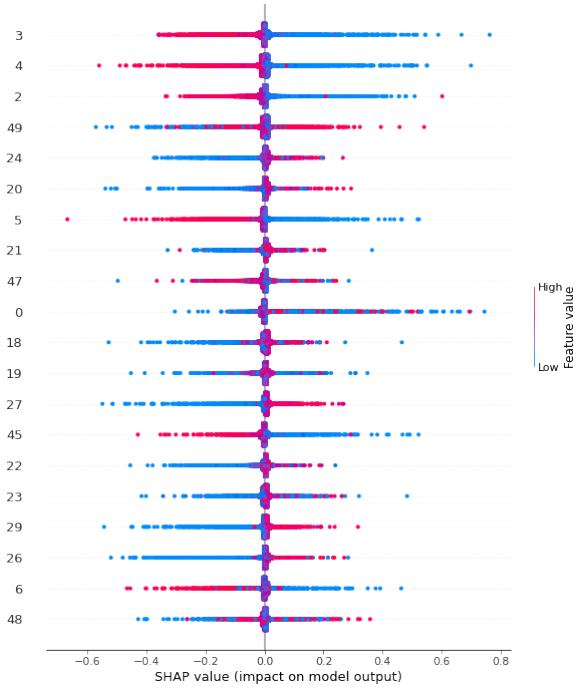


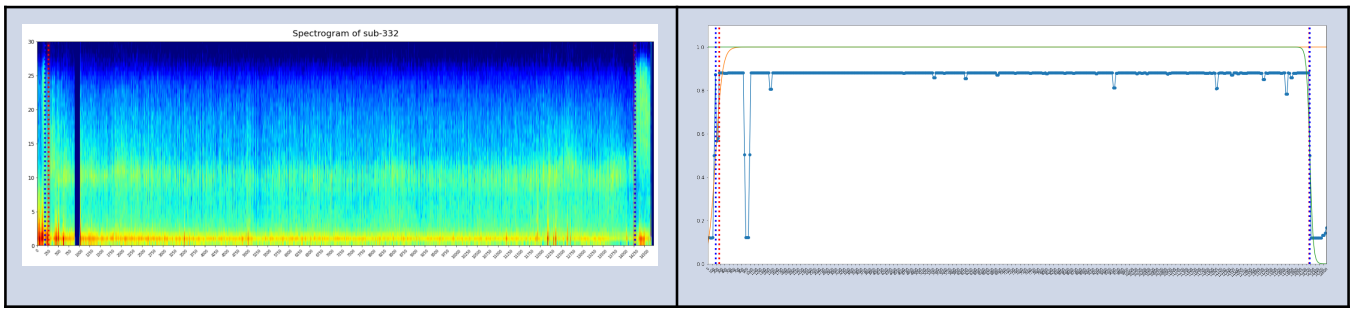
Figure 2: SHAP value summary plot.

The Shap values are represented for each variable in their order of importance : each point represents a SHAP value for a specific data point. It delivers the following information:

- *Feature importance: Variables are ranked in descending order.*
- *Impact: The horizontal location shows whether the effect of that value is associated with a higher or lower prediction.*
- *Original value: Colour shows whether that variable is high (in red) or low (in blue) for that observation.*

Identification of LoC and RoC:

LoC and RoC were found at the mid-point of the corresponding logistic function (**Figure 2).** Identification of our parameters was done using least square method considering a L1-norm regularisation.



A: example of one spectrogram from a patient including signal before, during, and after propofol-based general anaesthesia.

B: time series representing change in unconsciousness probability based as a result of gradient boosting classification over each 30s epochs (blue dots). Logistic regression associated with LoC (red) and RoC (orange) estimation are shown in plain lines. **C:** reports performances obtained in estimating unconsciousness among patients anaesthetised with propofol.

Additional results

Table 3Gradient boosting (BG) classification performance and Loc/RoC estimation error for two classes (unconsciousness / awakeness).

	TRAIN_SET		TEST_SET	
	LoC	**RoC**	**LoC**	**RoC**
Accuracy		0.996705		0.9684
AUC		0.999		0.981
Mean	-11.806931	6740.201564	-32.626957	15.198250
Median	15.516872	5904.691095	5.500000	-0.999999
sdt	452.526477	4839.029623	342.788007	130.876631
CI 95% (low)	-70.090462	6116.954649	-121.957998	-18.908386
CI 95% (high)	46.476599	7363.448479	56.704084	49.304886
Q1	0.578691	3242.459812	-82.111297	-30.499976
Q3	30.517756	9169.858837	62.974248	48.174023
MAE	56.489030	6845.831082	192.764396	63.392516
RMSE	6192.815498	8444.384278	19.373290	22.644341

Table 4 : Gradient boosting (BG) classification performance and Loc/RoC estimation error for three classes (before LoC / unconsciousness / after RoC).

	TRAIN_SET		TEST_SET	
	LoC	**RoC**	**LoC**	**RoC**
Accuracy	0.9790		0.9558	
AUC	0.998		0.994 (0=0.976, 1=0.981, 2=0.971)	
Mean	-9.530243	6738.883154	-17.550213	0.040694
Median	17.188660	5901.842544	7.603086	-0.999998
sdt	452.668527	4838.111799	319.811811	131.650473
CI 95% (low)	-67.832069	6115.754451	-100.893625	-34.267606
CI 95% (high)	48.771582	7362.011857	65.793199	34.348994
Q1	2.698445	3247.382245	-76.315351	-45.373429
Q3	31.806562	9159.014738	56.756768	41.000007
MAE	57.515041	6844.353040	166.278805	67.114260
RMSE	6192.878430	8433.315520	14.787375	20.557836

Table 5 : Mean Gradient boosting (GB) classification performance and Loc/RoC estimation error for three classes (before LoC / unconsciousness / after RoC) performed on 50 iterations.

	TEST_SET	
	LoC	**RoC**
Mean	-32.20442401	-54.9488382
Median	25.43321265	-15.70830923
sdt	307.2024122	372.8839523
CI 95% (low)	-112.2618086	-152.1229254
CI 95% (high)	47.85296058	42.22524896
Q1	-82.14757184	-64.92407395

Q3	77.19399899	26.47038324
RMSE	112.7965112	103.9012878

Jona Joaquim, Thibaut Chamoux, Jade Perdereau, Valentin Iovène, Jérôme Cartailler, Fabrice Vallée, *Integrating Heterogeneous Perioperative Data Sources: Platform Design and Implementation*. Submission in progress.

Integrating Heterogeneous Perioperative Data Sources: Platform Design and Implementation

Jona Joachim[1,3,2*], Thibaut Chamoux[4,3], Jade Perdereau[1,3,2], Valentin Iovène[5,3], Jérôme Cartailler[1,2], Fabrice Vallée[1,3,2]

[1], Université Paris Cité, 85 boulevard Saint-Germain, Paris, 75006.

[2], INSERM, U942 MASCOT, Paris, France, 2 rue Ambroise Paré, Paris, 75010.

[3], Department of Anaesthesia and Critical Care, APHP, Hôpital Lariboisière, 2 rue Ambroise Paré, Paris, 75010.

[4], Entrepôt de données de santé, Assistance Publique Hôpitaux de Paris, 55, boulevard Diderot, Paris, 75012.

[5], Physight, 5 rue Marie Pape-Carpantier, Paris, 75006.

*Corresponding author(s). E-mail(s): jona.joachim@aphp.fr;

Keywords

anesthesia, intensive care, perioperative data, digital twins, real-time processing, machine learning, clinical decision support, interoperability

Introduction

Healthcare providers universally strive to enhance the quality of patient care. Recent advances in information technology have created unprecedented opportunities for healthcare improvement that would have been unimaginable a decades ago. Modern intensive care units (ICUs) and operating rooms (ORs) now rely on sophisticated patient monitoring systems that generate vast quantities of valuable data. Until relatively recently, this wealth of information was routinely discarded, until researchers

recognized its immense potential, as evidenced by pioneering efforts like the MIMIC database established in 2000[1].

In the years since, numerous databases capturing patient monitoring data from ICU and OR environments have been developed. Some, like MIMIC[2] and VitalDB[3], have made their data publicly available, allowing research teams worldwide to apply computational statistics to develop generalizable models of human physiology.

Despite these advances, the healthcare data landscape remains insufficient to fully address the complex challenges of physiological modeling. Significant hurdles persist in effectively leveraging machine learning techniques to develop truly reliable artificial intelligence systems for clinical medicine, especially in order to leverage these models in real time.

A comprehensive perioperative data platform must capture sufficiently diverse patient populations to ensure that derived insights generalize across clinical settings. This presents significant challenges as perioperative care practices, patient demographics, and disease prevalence vary substantially across institutions, regions, and healthcare systems. Existing databases have inherent geographic and specialty biases—MIMIC data comes primarily from Massachusetts (USA) with a cardiovascular procedure predominance, while VitalDB originates from South Korea with concentration in general surgery, thoracic, urological, and gynecological procedures. These limitations raise concerns about model applicability across global populations, as algorithms can only recognize patterns present in their training data. Therefore, demographic and clinical diversity in data collection becomes as critical as raw data volume for developing robust perioperative models. A purpose of this article is to provide a blueprint on how to build interoperable databases in order to permit multinational collaboration on wider data sets.

Goals

In order to support clinical decision making we believe that a modern data platform needs to address short term and long term needs and support real time algorithms as well as artificial intelligence (AI) model training on big data workloads. We identified the three following workloads which the platform must address.

Immediate Clinical Support

Establishing a comprehensive perioperative database with real-time acquisition capabilities enables immediate improvements in patient care through what is called digital twin modeling[4]. By leveraging this infrastructure, clinicians can access personalized virtual representations of patients that continuously update with incoming vital sign measurements during procedures. These digital twin models simulate patient-specific physiological responses to interventions, enabling predictive capabilities beyond conventional monitoring. The platform processes raw physiological signals to calibrate these virtual models in real-time, calculating novel biomarkers and predicting potential

physiological trajectories before they manifest clinically. This approach presents medical teams with both current status and anticipated responses to planned interventions, enriching clinical decision-making with more detailed, contextual, and forward-looking information tailored to each patient's unique physiological profile.

Predictive Risk Mitigation

In the medium term, such a database supports predictive analytics for adverse events like sepsis during hospitalization. Machine learning models trained on the rich dataset—including vital signs, laboratory values, and patient demographics—can identify early warning patterns before clinical manifestations become evident. This predictive capability allows medical teams to implement proactive interventions, such as targeted antibiotic therapy or enhanced monitoring[5], potentially reducing morbidity and mortality through early intervention rather than reactive treatment.

Population Health Advancement

The long-term value of comprehensive perioperative data extends beyond individual patient care to broader public health improvements. Systematic analysis of previously underutilized physiological signals can reveal disease patterns and risk indicators—such as subtle cardiac arrhythmias on ECG recordings—that might otherwise remain undetected. These discoveries facilitate appropriate referrals to specialized services and earlier implementation of preventive measures. Consequently, patients benefit from reduced diagnostic delays, decreased symptom progression, and lower rates of post-operative complications, which remain a significant cause of preventable mortality worldwide. This approach transforms perioperative data collection from a transient monitoring tool into a valuable component of longitudinal health surveillance.

We designed a data platform with the goal to respond to these research and care demands. Additionally, we strived to build a system which is reproducable, independent and future proof. To this end, we tried to adhere as much as possible to medical and industry standards and use only open source software.

Methods

Data Governance

All data was stored on premise and the infrastructure was isolated from the Internet according to hospital standards. The healthcare data warehouse (Entrepôt de Données de Santé-EDS) of the Greater Paris University Hospitals (AP-HP) obtained approval from the Commission nationale de l'informatique et des libertés (CNIL), the national data protection authority for France (autorisation n°1980120) to store AP-HP healthcare data for research purposes in compliance with the European General Data Protection Regulation (GDPR). EDS obtained the ISO/IEC 27001 information security standard label.

Usage of the data for research projects and collaboration requires approval from the EDS Institutional Review Board (IRB), upon which data access can be shared after anonymisation.

Data sources

Data in the operating room as well as the ICU is heterogeneous and originates from many different sources. In order to dispose of a comprehensive database, it is important to consider all the various data sources which contribute to the bigger picture of patient physiology and response to treatment and insult. The following data sources were considered for our platform:

Patient monitor and vitals

Patient monitors constitute the major bulk of high frequency data in the operating room. They provide continuous monitoring of vital signs such as heart rate, blood pressure, respiratory rate, oxygen saturation, and electrocardiogram (ECG) waveforms. This data is of high importance for biomechanical modeling and digital twin models since it describes the physiological state of the patient in real time and in great detail. These devices are typically connected to a central station or a computer system that allows for real-time visualization and storage of data. Several types of data are usually available depending on vendor support:

- Aggregated numerics like heart rate, SpO2, etc. typically sampled at 1-5 Hz.
- Raw waveforms like ECG, PPG, etc. typically sampled at 100-200 Hz and up to 500 Hz for ECG data.
- Event based data like alarms
- Metadata like current device settings
- Demographic data like patient name, age, gender, etc.

Usually the data can be exported directly from the monitor using a proprietary interface or preferably from the central station using a standard interface. Patient monitors can usually interconnect with other devices like ventilators and cardiac output monitors and will export the data from these devices as well. Future standards like IEEE 11073 service-oriented device connectivity (SDC) might further improve interoperability between devices. At our institution, Philips patient monitors are in use as well as Mindray which were deployed at later stages. Mindray (Shenzhen, Guangdong , China) central stations provide HL7 export of numerics (1 Hz) and waveforms data. Philips (Eindhoven, The Netherlands) provide two high fidelity data platforms: Datawarehouse Connect (DWC) and Clinical Insights Manager (CIM). This Philips platform exports comprehensive patient monitor data with numerics, waveform, alarms, metadata and demographic data to a central database.

Since March 2019, we have collected data from Philips patient monitors using Philips DataWarehouse Connect (DWC), a Microsoft (Redmond, WA, US) SQL Server-based storage solution. This system captures monitor data in near real-time and has

proven reliable for long-term storage over six years across ICU and OR environments, accumulating more than 42 TB of data.

Despite its reliability, the DWC architecture presented several limitations. Most significantly, the proprietary Philips architecture stored data directly to the database without providing access to real-time data streams. To address this constraint, we implemented a Change Data Capture (CDC) platform[6] using Debezium that leverages Microsoft SQL Server's change data capture (CDC) functionality. This implementation, though not officially supported by Philips, was accomplished in collaboration with their R&D team. Detailed instructions for CDC activation are available in our supplementary materials.

Additional challenges emerged from SQL Server's limited time-series data support (prior to the 2022 version), making time-window queries inefficient. As a result, data often required bulk extraction and post-processing. The DWC table design itself introduced performance bottlenecks, necessitating complex joins across multiple tables that demanded substantial server memory.

A further complication arose from waveform data storage. Since SQL Server lacks native ARRAY type support, waveform data was stored in binary format, requiring resource-intensive conversion operations for queries.

To simplify access for researchers unfamiliar with the underlying database structure, we developed the dwclib Python library (https://github.com/larib-data/dwclib). This tool provides an abstraction layer enabling efficient queries to the SQL database while hiding the complexity of the DWC architecture.

Philips has developed Clinical Insights Manager (CIM) and High Fidelity Data Export (HFDE) as next-generation replacements for DWC. CIM operates as a cloud-based platform where Philips manages all data storage and processing. Meanwhile, HFDE serves as an on-premise alternative that provides an Apache Kafka[7,8] data stream for real-time processing, giving users complete flexibility over how data is stored and processed. Our team collaborated with Philips R&D on HFDE development to ensure it could adapt to the rapidly evolving healthcare data landscape. By eliminating the Microsoft SQL Server dependency, HFDE removes previous technical limitations and allows for seamless migration between storage technologies as more efficient solutions emerge.

At our institution we were early testers of DWC and helped develop CIM and HFDE.

Syringe pumps

While historically, syringe pumps provided an RS232 serial port, current syringe pumps have adopted the same infrastructure as patient monitors. They are typically connected to the hospital computer network and communicate with a vendor server over a secure channel. This server exports data in HL7 using the Integrating the Healthcare Enterprise® (IHE) Patient Care Device (PCD) profile. The existance of a dedicated HL7 profile has permitted standardized and interoperable data export

across vendors. At our institution we deployed the B. Braun (Melsungen, Germany) Space+ syringe pumps together with the OnlineSuite+ gateway server.

Ventilators and Anesthesia machines

Ventilators provide deep insight into pulmonary mechanics, gas exchanges and heart-lung interations. Modern ventilators provide a vast amount of data which has been used in the intensive care setting to get deeper understanding of acute lung diseases like Acute Respiratory Distress Syndrome (ARDS). Anesthesia machines provide additional data concerning anesthetic gases. While these devices can be connected to patient monitors, the shear amount of parameter they provide overwhelms the patient monitors and only a subset of parameters are typically available on the monitor. In order to obtain high fidelity data from the ventilators, the data needs to be extracted directly from the device. At our institution, we use Mindray (Shenzhen, Guangdong , China) A9 ventilators in the OR and Maquet Servo-i (Getinge, Gothenburg, Sweden) ventilators in the ICU. A9 ventilators export numerics and waveform data in HL7 through the vendor gateway, while Servo-i ventilators are connected to the patient monitor and transmite numerics and waveform data though the patient monitor using IntelliBridge EC10 connectors (Philips, Eindhoven, The Netherlands).

EEG monitor

Over the last years, depth of anesthesia monitoring using EEG has become a standard in many operating rooms. However, obtaining high quality EEG data for research from these devices has remained challenging due to limitations of EEG channels or export facilities. Bispectral index (BIS) modules are often integrated in the patient monitor but export only one EEG waveform channel. In our institution, we use Masimo (Irvine, California, US) Root monitors with Sedline EEG which derives four frontal EEG channels. These devices are connected to the Masimo PatientSafetyNet server, consisting of numerics and waveform data.

Patient data management system (PDMS)

The PDMS software used in the OR and ICU and its interaction with the electronic medical record (EMR) is crucial to get an understanding of events occurring during a perioperative procedure. The PDMS provides a centralized platform for managing patient information, including demographics, medical history, allergies, medications, and procedures. In the operating room, it give information concerning surgical periods, drugs administered, fluids administered, patient equipment and many other crucial information. In terms of machine learning, this data is often important to label the data for supervised learning tasks. Indeed, a change in blood pressure or heart rate can be an indicator of pain or bleeding during surgery but could also be the result of drugs administered. It is crucial to cross data with PDMS annotations to ensure accurate

interpretation and decision-making. In our institution, we use Philips IntelliSpace Critical Care and Anaesthesia (ICCA) in the operating room and Orbis (Dedalus, Milano, Italy) in the ICU. These systems usually have their own database. Current efforts in standardizing these databases is ongoing and many PDMS databases have been converted to the open Observational Medical Outcomes Partnership OMOP Common Data Model (CDM)[9].

Laboratory Values and Arterial Blood Gas Analysis

Laboratory data represents an essential component of comprehensive perioperative monitoring. Arterial blood gas (ABG) analysis provides critical insights into a patient's oxygenation status, acid-base equilibrium, and electrolyte composition. These measurements enable clinicians to make evidence-based decisions regarding ventilation parameters, fluid therapy strategies, and medication administration. Furthermore, laboratory values serve as objective biomarkers that reflect the patient's underlying physiological state and can be leveraged for outcome prediction and risk stratification. Laboratory information systems use standardized HL7 protocols to ensure interoperability. For our implementation, we integrated ABL90 (Radiometer, Copenhagen, Denmark) analyzers directly into our data ecosystem, allowing for real-time capture and analysis alongside other physiological data streams.

General trends in data export

Traditionally, medical devices were not interconnected and data could be extracted using proprietary software or serial interface. However, there has been a significant shift towards networked devices that provide a gateway instead of a serial interface for data extraction. This allows for easier integration with other systems and facilitates the exchange of data between different devices and platforms. These gateways usually provide standardized data export using HL7 or Fast Healthcare Interoperability Resources (FHIR) standards and lately Kafka, which enable interoperability and seamless data sharing between various healthcare systems and applications. An illustration of the Lariboisière data platform structure can be seen in Figure 1. One additional challenge when intergrating data from different vendors is the data synchronisation. To address this challenge, it is crucial that devices and servers are synchronized using network time servers (NTP)[10]. This ensures that the timestamps of the data are consistent across all sources, enabling accurate analysis and correlation of data from multiple devices or systems. Unfortunately, signal processing will invariably introduce a delay between the actual measurement and the availability of the data for analysis. This delay can vary depending on various factors such as the complexity of the signal processing algorithms. The amount of delay is often not documented and can introduced challenges, especially when jointly analyzing waveform data from different devices. Strict documentation of these delays by vendors will be necessary in the future to address these shortcomings.

Real time processing

Patient monitoring data in the perioperative and ICU setting is generated as a real-time stream of time series. The data will often be exported as an HL7 / FHIR stream by a vendor gateway but could also be exported directly from the device as a stream over a serial port for instance. This data needs to be processed and transformed in real time to join with other data streams and to normalize the data into a format that can be used for analysis. This process needs dedicated software components that are able to handle the high volume of data, perform the necessary transformations, and ensure low latency. We built a real-time data platform centered around Apache Kafka, which acts as a high-throughput real-time messaging system. Kafka has established itself as an industry standard used by major corporations worldwide to handle low latency high volume real-time data. Kafka provides a publish-subscribe architecture, facilitating the deployment of microservices, dedicated software bricks providing algorithms and transformations. While Kafka is a well established industry standard, it is not yet widely adopted by the healthcare industry. Healthcare data pipelines are dominated by Health Level 7 (HL7) streams. We deployed the open source Mirth Connect software (NextGen Healthcare, Atlanta, GA, US), which is a solution that can interact with HL7 streams and write the data to Apache Kafka.

Stream processing: ksqlDB and Flink

Data streams ingested into Kafka require real-time processing through specialized frameworks that enable immediate analysis, transformation, and enrichment. These stream processors join multiple data sources, perform calculations, and either store results in databases or forward them to other systems. For our platform, we implemented ksqlDB[11], a stream processing engine built on Apache Kafka that offers a SQL-like interface for querying data streams. This familiar SQL syntax makes it accessible for performing complex transformations and aggregations on real-time data without requiring specialized programming expertise. While ksqlDB has served our immediate needs effectively, we are currently transitionning our pipelines to Apache Flink[12] which has established itself as the new industry standard for stream processing. Flink represents a more modern alternative with enhanced capabilities for stateful processing and improved handling of event-time operations, which can better support our expanding use cases.

Long term storage solutions

Big data storage platforms have evolved significantly over the past decade, going from data warehouses to data lakes and finally the data lakehouse infrastructure.

Data warehouses are built using relational databases, which store well-structured data in predefined schemas and require upfront transformations before ingestion. Their structure is rigid and predefined at the creation of the database and while efficient, as needs change it can be difficult to adapt to new requirements or data sources.

A data lake, on the other hand, is an unstructured storage system that holds a vast amount of raw data in its native format. Data lakes are often used for big data analytics, machine learning (ML), and AI because they can handle structured and unstructured data types.

The evolution from data warehouses to data lakes occurred due to the rise of big data and the need for real-time analysis and insights. Data lakes offer scalability, flexibility, and cost efficiency over traditional data warehouses by eliminating the need for extensive data modeling and upfront transformation processes. However, while data lakes provide flexibility and scalability, they can also lead to data duplication, inconsistency, and increased query latency due to their unstructured nature.

A more recent trend is the emergence of a hybrid solution known as a data lakehouse. A data lakehouse combines the benefits of both data warehouses and data lakes. It provides scalability, flexibility (for unstructured and semi-structured data), and the ability to run queries efficiently on large datasets (similar to a data warehouse). At the same time, it allows for real-time analytics, ML, and AI applications, just like a data lake. This is achieved by maintaining a single copy of data in the data lake but creating an additional layer that can process and query data using SQL interfaces, similar to how a data warehouse works. This helps organizations leverage both traditional tools as well as modern analytics capabilities.

TimescaleDB

For our low-frequency data streams (1-2 Hz) consolidated in Kafka, we implemented TimescaleDB (TigerData, NY, US) as our primary time-series database. TimescaleDB offers an ideal solution because it efficiently manages time-series data while maintaining SQL compatibility. As an open-source PostgreSQL extension, TimescaleDB automatically partitions data chronologically, which dramatically enhances query performance by limiting scan operations to relevant time segments only.

We particularly valued TimescaleDB's advanced features for time-series optimization, including continuous aggregates and downsampling capabilities. These features allow us to precalculate aggregate values across various time intervals, significantly improving query response times for analytical operations. To transfer data from our Kafka streams into TimescaleDB, we deployed the Apache Camel JDBC Kafka connector, which provided the necessary flexibility and reliability for our data pipeline.

Our selection of TimescaleDB was strategically aligned with our existing infrastructure, which already utilized PostgreSQL. This choice leveraged PostgreSQL's well-established reputation for robustness and scalability, while also taking advantage of its extensive ecosystem of tools and community support.

For the database storage backend, we chose ZFS RAID-Z[13] on Linux. This open-source file system, originally developed by Sun Microsystems, provides critical features for medical data storage including robust data integrity protection, fault tolerance mechanisms, and efficient storage resource utilization.

For high frequency data and waveforms, we deployed the Apache Iceberg datalake infrastructure. Iceberg currently serves as the main storage backend for new data and is considered for migration of the existing DWC database. A key advantage of Apache Iceberg is its vendor-neutral architecture and wide support while delivering state of the art performance.

The data lakehouse architecture enables us to store information in efficient column-oriented formats like Apache Parquet while maintaining SQL query capabilities. This hybrid approach combines the structured querying power of traditional databases with the scalability advantages of distributed file systems.

To support our storage requirements, we deployed Purestorage hardware solutions (Pure Storage Inc., Mountain View, CA, US): FlashBlade systems providing AWS S3-compatible object storage for our Iceberg data lakehouse architecture, and FlashArray instances delivering high-performance block storage for our Microsoft SQL Server databases running Philips DWC.

The Duality of Real-Time and Long-Term Data

Real-time data streams and long-term storage databases are complementary and closely related[14]. A comprehensive data platform needs to provide both forms of data presentation. Indeed, long term big data storage is key to train modern AI models through machine learning, while real-time data can make the power of these algorithms available to the clinicians caring for the patient. There exists a duality between these two paradigms. Real-time data can be stored in a database as the data comes in to consolidate it in long term storage for big data applications. On the other hand, long-term storage databases can be monitored for changes and trigger events through CDC.

Data Records

Patient Monitor

Our DWC database currently contains patient monitoring data collected from March 2019, encompassing various clinical specialties. The number of cases across specialties is presented in the table below:

Clinical Specialty	Number of Cases
Ophthalmology	3,255
Ambulatory Bowel Surgery	2,987
Emergency Cases (Total)	14,059
- Emergency Orthopedics	6,321
ENT (Ear, Nose & Throat)	8,170
Neurosurgery	5,879

Clinical Specialty	Number of Cases
Gynecology OR	3,434
Maternity Birth Ward	108
Interventional Neuroradiology	5,500
Elective Orthopedics	4,099
PACU (Post-Anesthesia Care)	63,965
ICU (Intensive Care Unit)	7,459

Syringe pump data

Clinical Specialty	Number of Cases
Ambulatory Bowel Surgery	114
Emergency Cases (Total)	1,494
- Emergency Orthopedics	455
ENT (Ear, Nose & Throat)	1,701
Neurosurgery	2,259
Interventional Neuroradiology	1,528

EEG data

Clinical Specialty	Number of Cases
Ambulatory Bowel Surgery	763
Emergency Cases (Total)	2,218
- Emergency Orthopedics	865
ENT (Ear, Nose & Throat)	1,757
Neurosurgery	1,531
Gynecology OR	964
Interventional Neuroradiology	1,467
Elective Orthopedics	1,162

Discussion

Data Volume Requirements

Establishing effective perioperative big data platforms faces fundamental uncertainty regarding minimal dataset size requirements. The data volume needed varies dramatically based on the intended analytical applications, from basic statistical analyses to complex deep learning models with millions of parameters. Unfortunately, determining sufficiency is often only possible retrospectively through model validation. Advanced machine learning applications typically require data from thousands or tens of thousands of procedures to achieve convergence and generalizability. Current leading perioperative databases—such as VitalDB with over 6,000 operating room

cases-while substantial, represent only a minute fraction of the estimated 300 million anesthesia procedures performed annually worldwide[15]. This relative scarcity underscores why high-volume data acquisition remains a cornerstone challenge in medical AI development.

The development of these platforms is only possible through the close collaboration between medical doctors, biomedical engineers and IT engineers.

Signal Quality and Preprocessing

Building effective perioperative data platforms requires addressing pervasive data quality challenges that demand sophisticated engineering solutions. Raw physiological data from operating rooms and intensive care units contain numerous artifacts stemming from patient movement, surgical manipulations, electromagnetic interference, sensor disconnections, and calibration issues. Without proper cleaning and preprocessing pipelines, these quality issues propagate through analytical models, potentially leading to erroneous conclusions or unsafe recommendations. Consequently, robust data validation, artifact detection, and signal reconstruction algorithms must be integral components of any comprehensive perioperative data platform. While certain algorithms could be implemented in real time for data cleaning, we chose to store raw data and filter the data at extraction as a data engineering step. This reduces the real time pipeline complexity and ensures that no data is lost du to filter misspecification.

Reliance on standards

While certains pipelines rely heavily on interfacing with individual devices, often using RS232 serial ports[3], we chose to develope an infrastructure based on standardised communication protocols. The reliance on formal standards such as HL7 and industry standards such as Kafka mean that we can abstract over the underlying devices and are not bound to individual vendors. This means that we can typically adopt new hardware early without having to develop costly device drivers and adapt as public markets change.

The Critical Value of Real-Time Processing

The data generated in operating rooms and intensive care units is inherently real-time in nature, flowing continuously as patients undergo procedures and receive ongoing monitoring. This real-time characteristic offers significant clinical value, enabling the deployment of algorithms that can generate biomarkers and insights immediately available to clinicians for informed decision-making. Beyond immediate analytics, real-time data streams provide opportunities for adaptive patient-specific modeling—a cornerstone of precision medicine. For example, in pharmacometric applications, population-based pharmacokinetic/pharmacodynamic (PKPD) models initially calibrated across broad patient groups can be dynamically refined through Bayesian inference using concurrent syringe pump data and physiological measurements. This real-time model adaptation creates a feedback loop that continuously improves predictive accuracy for the specific patient, potentially enhancing therapeutic outcomes through increasingly personalized interventions.

Clinical Applications

The data platform which we described supports several research topics in the domains outlined in the introduction, notably immediate clinical support[16,17], predictive risk mitigation[18] and population health advancement[19].

Conclusion

This paper has described the design and implementation of a comprehensive data platform supporting medical digital twin models in perioperative and intensive care settings. Our work addresses a critical gap in healthcare informatics by creating an integrated system that combines real-time data processing capabilities with robust long-term storage solutions, enabling both immediate clinical applications and advanced research opportunities. The platform we've developed successfully integrates diverse data sources—including patient monitors, ventilators, syringe pumps, EEG systems, laboratory values, and patient data management systems—creating a holistic view of patient physiology that was previously unattainable. We hope that this work will serve as a foundation for future research and development in the field of digital twins and perioperative inform and enable wider deployment of data platforms in clinical settings.

Bibliography

1. Johnson, A., Pollard, T. & Mark, R. MIMIC-III Clinical Database. (2015) doi:10.13026/C2XW26.
2. Johnson, A. E. W. *et al.* MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data* **10**, 1 (2023).
3. Lee, H.-C. *et al.* VitalDB, a high-fidelity multi-parameter vital signs database in surgical patients. *Scientific Data* **9**, 279 (2022).
4. Sel, K. *et al.* Building Digital Twins for Cardiovascular Health: From Principles to Clinical Impact. *Journal of the American Heart Association: Cardiovascular and Cerebrovascular Disease* **13**, e031981 (2024).
5. Lal, A. *et al.* Development and Verification of a Digital Twin Patient Model to Predict Specific Treatment Response During the First 24 Hours of Sepsis. *Critical Care Explorations* **2**, e0249 (2020).
6. Ankorion, I. Change data capture efficient ETL for real-time bi. *Information Management* **15**, 36 (2005).
7. Goodhope, K. *et al.* Building LinkedIn's real-time activity data pipeline. *IEEE Data Eng. Bull.* **35**, 33–45 (2012).
8. Sax, M. J. Apache Kafka. in *Encyclopedia of Big Data Technologies* 1–8 (Springer, Cham, 2018). doi:10.1007/978-3-319-63962-8_196-1.
9. Voss, E. A. *et al.* Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. *Journal of the American Medical Informatics Association: JAMIA* **22**, 553–564 (2015).
10. Martin, J., Burbank, J., Kasch, W. & Mills, P. D. L. *Network Time Protocol Version 4: Protocol and Algorithms Specification*. https://datatracker.ietf.org/doc/rfc5905 (2010) doi:10.17487/RFC5905.

11. Jafarpour, H. & Desai, R. KSQL: Streaming SQL Engine for Apache Kafka. (2019) doi:10.5441/002/EDBT.2019.48.
12. Katsifodimos, A. & Schelter, S. Apache Flink: Stream Analytics at Scale. in *2016 IEEE International Conference on Cloud Engineering Workshop (IC2EW)* 193–193 (2016). doi:10.1109/IC2EW.2016.56.
13. Pugh, C., Henderson, P., Silber, K., Carroll, T. & Ying, K. Utilizing zfs for the storage of acquired data. in *2009 23rd IEEE/NPSS Symposium on Fusion Engineering* 1–4 (2009). doi:10.1109/FUSION.2009.5226462.
14. Sax, M. J., Wang, G., Weidlich, M. & Freytag, J.-C. Streams and Tables: Two Sides of the Same Coin. in *Proceedings of the International Workshop on Real-Time Business Intelligence and Analytics* 1–10 (Association for Computing Machinery, New York, NY, USA, 2018). doi:10.1145/3242153.3242155.
15. Gao, M. *et al.* Global trends in anesthetic research over the past decade: a bibliometric analysis. *Annals of Translational Medicine* **10**, 607 (2022).
16. Perdereau, J. *et al.* Blood Pressure Estimation Using Explainable Deep-Learning Models Based on Photoplethysmography. *Anesthesia and Analgesia* **140**, 119–128 (2025).
17. Perdereau, J., Joachim, J., Vallée, F., Cartailler, J. & Moreau, T. Harnessing operating room signals to estimate mean arterial pressure with AnesthNet. *Scientific Reports* **15**, 33988 (2025).
18. Vallée, F. *et al.* The ICEBERG: A score and visual representation to track the severity of traumatic brain injury: Design principles and preliminary results. *The Journal of Trauma and Acute Care Surgery* **93**, 229–237 (2022).
19. El Ayeb, K. *et al.* Using intraoperative physiological data to predict postoperative cognitive decline and cardiovascular events: results from a 30-month prospective study. *British Journal of Anaesthesia* https://doi.org/10.1016/j.bja.2025.03.036 (2025) doi:10.1016/j.bja.2025.03.036.

Abstract

Background

Modern intensive care units and operating rooms generate vast quantities of valuable physiological data that remain largely underutilized for clinical decision support and research. Existing perioperative databases exhibit geographic and specialty biases, limiting the generalizability of derived models. Additionally, healthcare lacks standardized approaches for integrating heterogeneous data sources while supporting both real-time analytics and long-term machine learning applications.

Objectives

We designed and implemented a comprehensive perioperative data platform to support three clinical workloads: immediate clinical support through digital twin modeling, predictive risk mitigation for adverse events, and population health advancement through discovery of previously undetected disease patterns.

Methods

We developed an integrated data infrastructure at the Greater Paris University Hospitals combining real-time stream processing with robust long-term storage. The platform ingests data from patient monitors, syringe pumps, ventilators, EEG systems, laboratory analyzers, and patient data management systems using standardized protocols. Real-time processing utilizes Apache Kafka with ksqlDB/Flink for stream analytics, while long-term storage employs TimescaleDB for low-frequency data (1-2 Hz) and Apache Iceberg data lakehouse architecture for high-frequency waveforms. All infrastructure adheres to open-source standards and GDPR compliance.

Results

Since March 2019, we have collected over 42 TB of data encompassing 122,815 cases across multiple specialties including ICU (7,459 cases), emergency surgery (14,059 cases), and post-anesthesia care (63,965 cases). The platform successfully integrates diverse data streams with synchronized timestamps and supports both retrospective analysis and real-time clinical applications.

Conclusions

This work provides a reproducible blueprint for building interoperable perioperative data platforms that enable multinational collaboration while supporting both immediate clinical decision-making and advanced research applications. The vendor-neutral, standards-based architecture ensures adaptability as technologies evolve and facilitates wider deployment across healthcare institutions.

Figures

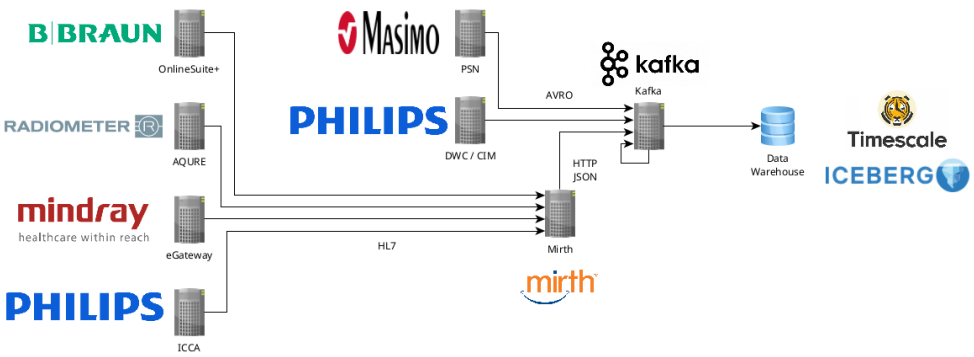


Figure 1: The federated network infrastructure of the real-time data pipeline

Théophile Harlé, Jona Joachim, Pierre Boutouyrie, Joaquim Mateo, Jade Perdereau, Alexandre Mebazaa, Jérôme Cartailler, Fabrice Vallée. *Continuous measurement of carotid-femoral pulse wave velocity (PWVcf.) during general anaesthesia using Doppler: a preliminary study*. Journal of Clinical Monitoring and Computing, 2025

Journal of Clinical Monitoring and Computing (2025) 39:495–503
https://doi.org/10.1007/s10877-024-01256-w

ORIGINAL RESEARCH

Continuous measurement of carotid-femoral pulse wave velocity ($PWV_{cf.}$) during general anaesthesia using Doppler: a preliminary study

Théophile Harlé[1,2] · Jona Joachim[1,2] · Pierre Boutouyrie[3] · Joaquim Mateo[1,2] · Jade Perdereau[1,2] · Alexandre Mebazaa[1,2] · Jérome Cartailler[1,2] · Fabrice Vallée[1,2]

Received: 23 September 2024 / Accepted: 12 December 2024 / Published online: 13 March 2025
© The Author(s), under exclusive licence to Springer Nature B.V. 2024

Abstract
This study explores the feasibility of continuous pulse wave velocity (PWV) monitoring during general anaesthesia (GA), particularly in response to blood pressure fluctuations. Our aim is to evaluate whether dynamic PWV can provide new insight to detect cardiovascular risks. From December 2022 to February 2023, continuous carotid and femoral Doppler monitoring was performed on patients scheduled for surgery with GA, to collect PWV data at awakening (PWV_{AW}) and during GA (PWV_{GA}). The study investigated PWV's response to MAP fluctuations using the α-angle, a dynamic stiffness parameter. We evaluated PWV and α-angle efficacy in discriminating between low (CVR-) and high (CVR+) cardiovascular risk patients. Among 43 patients, 41 (95%) had successful PWV measurements. PWV_{AW} was significantly higher than PWV_{GA} (8.1 vs. 7.4 $m.s^{-1}$, $p<0.0001$). This difference vanished after matching MAP levels. A strong correlation was found between PWV_{AW} and PWV_{GA} ($r=0.88$, and $r=0.97$ at the same MAP levels). PWV_{GA}, α-angle and their product (α x PWV_{GA}) were significantly higher in CVR+patients (8.1 vs. 6.9 $m.s^{-1}$, $p<0.01$; 2.6 vs. 1.3 degrees, $p<0.001$; 21.8 vs. 8.1 $degrees.m.s^{-1}$, $p<0.001$, respectively), with AUC values indicating good predictive capabilities for cardiovascular risk (PWV_{GA}: AUC [95%CI]=0.80 [0.65–0.95]; α-angle: 0.83 [0.69–0.96]; product: 0.86 [0.74–0.97]). Measurement of PWV under GA using carotid and femoral Doppler is a feasible method to continuously assess arterial stiffness under general anaesthesia. Further studies are required to validate the α-angle parameter in different physiological conditions.

Keywords anaesthesia · arterial stiffness · pulse wave velocity · Doppler

1 Introduction

Cardiovascular disease, a leading cause of hospitalization and mortality in France, present an important healthcare challenge, with over 1 million hospitalizations and more than 140 000 deaths annually [1]. While traditional cardiovascular risk factors, such as age, sex, hypertension (HT), are well described in the literature, recent studies highlight arterial stiffness, a loss of compliance of large arteries, as an early and independent predictor of cardiovascular events [2–5]. As such, arterial stiffness promotes later in life acute life-threatening events like myocardial infarctions or strokes [6, 7].

The assessment of arterial stiffness relies on measuring carotid-to-femoral pulse wave velocity (PWV) [8]. However, this approach has limitations in routine practice. PWV measurement requires significant operator skill and time.

Fabrice Vallée was equally involved in the realization of this work and should be considered as senior author.

✉ Théophile Harlé
theophile.harle@gmail.com

1 Université Paris Cité, Inserm UMRS 942 Mascot, Paris F-75006, France

2 Department of Anesthesia and Critical Care, Lariboisière Hospital, APHP, Paris F-75010, France

3 Pharmacology, Inserm PARCC U970, Georges-Pompidou European Hospital, Paris-Cité University, Assistance publique-Hôpitaux de Paris, Paris, France

Most importantly, its value is influenced acutely by blood pressure levels, as demonstrated by Bramwell-Hill [9, 10], but also by chronic remodelling of the vessel wall after exposition to CV risk factors. The added value of PWV over blood pressure being related to arterial wall remodelling, there is a call for a better understanding of PWV changes in response to acute changes in blood pressure levels. In fact, some authors have attempted to approach this goal by measuring PWV in the standing and supine positions, after administration of vasopressors, or using 24-hour recordings, but the variations in arterial pressure observed were often small [11–13].

In this context, general anaesthesia (GA), numbering around 10 million patients annually, offers an untapped resource for monitoring cardiovascular health. Since 2019, the anaesthesiology department at CHU Lariboisière has been exploring the use of GA data to predict cardiovascular damage, transforming routine anaesthesia into a "sleeping medical consultation". Through this approach, we could identify early signs of cardiovascular fragility by providing a real-time assessment of arterial stiffness/blood pressure relationship, taking advantage of the procedure-induced blood pressure changes, together with easy instrumentation of the patients and dedicated time to surveillance.

This study explores the feasibility of continuous PWV monitoring during GA using carotid and femoral Doppler, particularly in response to blood pressure fluctuations commonly observed during GA. Our aim is to evaluate whether dynamic PWV can provide new insight to detect cardiovascular risks.

2 Materials and methods

2.1 Patients

Between December 2022 and February 2023, patients scheduled for surgery under GA at Lariboisière hospital (Paris, France), were considered for inclusion in this study unless they expressed any objection to participating. Exclusion criteria included being under 18 years old, having cardiac arrhythmia, advanced obliterative arteriopathy of the lower limbs (Leriche and Fontaine grade III and IV), or being pregnant. Oral consent was obtained the day before the surgery, and patients were informed of their right to object to the use of their computerized data for medical research. This research was approved by the French patient's protection committee Nord-Ouest IV (NIVOP: Noninvasive Measurement of Arterial Stiffness by Pulse Wave Velocity) as a standard of care research.

2.2 Study protocol

Starting before induction of general anaesthesia (Supp. Mat. 1), patients were monitored using standard procedures, complemented by the placement of two non-invasive pulsed Doppler probes, one probe positioned over the right common carotid artery, and the other over the common femoral artery. This was achieved by optimizing the ultrasound beam depth, gain, and power settings. A laminar flow pattern in the flow envelope indicated a high-quality signal. These probes were then fixed to the patient in the position to obtain the optimum signal using a self-adhesive elastic band and were connected to a WAKIe R3 2TC® monitor (Athys Medical, France) for continuous recording of both carotid and femoral arterial flows. The WAKIe R3 2TC® was integrated with the IntelliVue MX850® monitor, which centralized the signals and enabled export of the data for post-processing (Supp. Mat. 2).

All intraoperative recordings were manually extracted and analysed using Graphysio software, which provided a graphical, synchronized representation of all recorded intraoperative signals. PWV measurements, in $mm.ms^{-1}$, were computed using Graphysio software, based on the formula:

$$PWV = \frac{0.8\ x\ D_{cf}}{TT_{cf}},$$

where $D_{cf.}$ is the carotid-femoral distance in millimetres, measured manually between the two Doppler probes using a tape measure before the procedure. As the full distance overestimates the distance travelled by the pulse wave, we have corrected this by 0.8, as recommended in the 2012 consensus document [14]. $TT_{cf.}$, the carotid-femoral transit time in milliseconds, is defined as the time between the feet of the carotid and femoral flow waves, using the "foot-to-foot" method. This measurement uses the maximum acceleration from the signal second order derivative, automatically calculated by the software. We visually verified the accuracy of point placement throughout the entire period studied. $TT_{cf.}$ values were averaged over 30 consecutive cardiac cycles (Supp. Mat. 3).

2.3 Data analysis

2.3.1 Periods of interest (Fig. 1A)

Continuous PWV recording during the procedure allowed the targeting of key periods of interest, with PWV measured during both awakening (PWV_{AW}) and under general anaesthesia (PWV_{GA}). Due to possible Doppler signal disturbances during anaesthetic induction, for instance caused by face mask ventilation and oro-tracheal intubation, PWV_{GA}

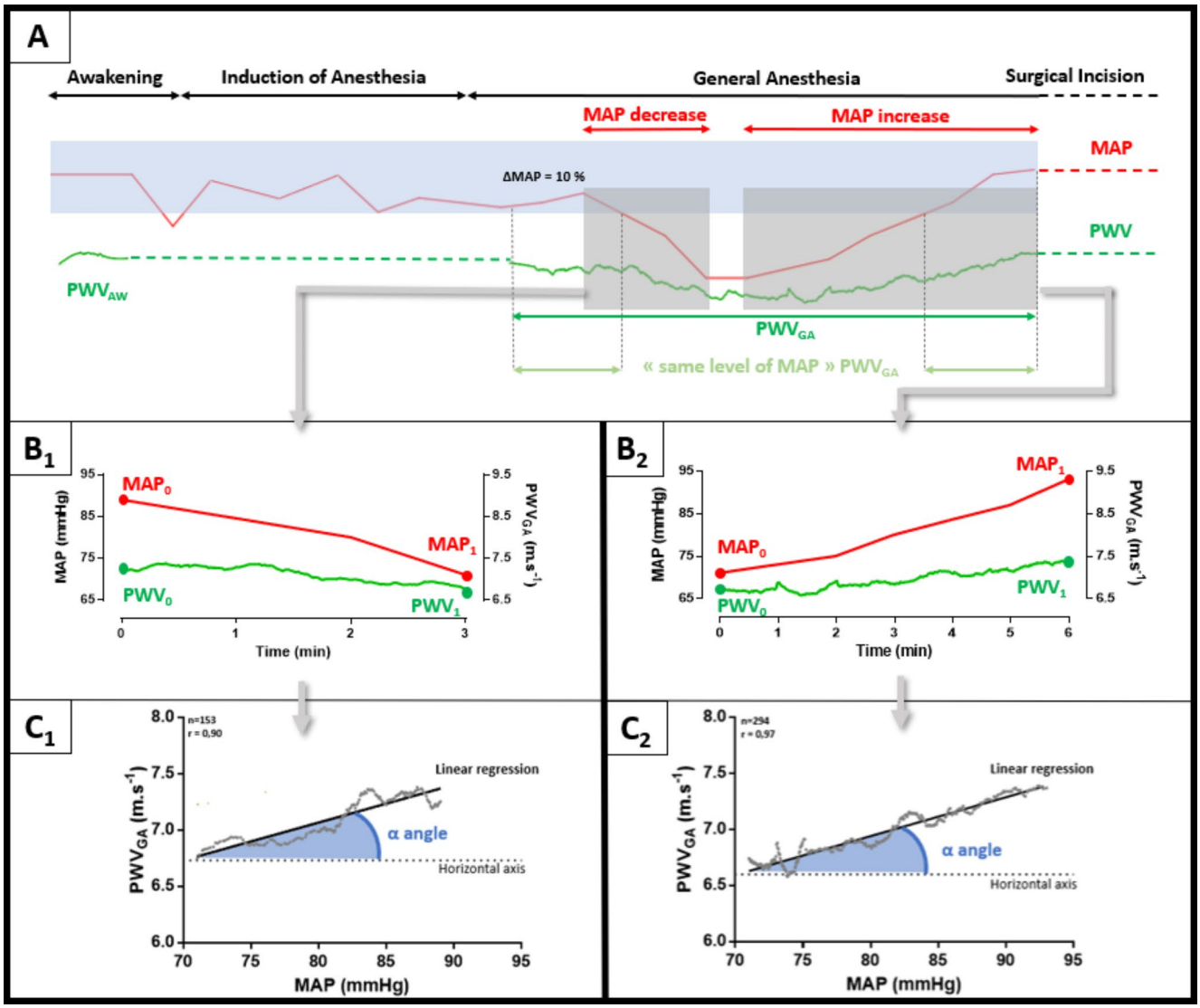


Fig. 1 Description of the method and α-angle measurement. **A**: MAP and PWV changes during the various stages of anaesthesia and during MAP variations (PWV_{AW}: PWV when awake, PWV_{GA}: PWV under GA). **B**: changes of MAP and PWV over time during an increase [1] and decrease [2] in MAP ($_0$: start of MAP variation, $_1$: end of MAP variation). **C**: PWV plotted against MAP during an increase [1] and decrease [2] in MAP, with α-angle depicted as the angle formed at the intersection of the linear regression line of PWV versus MAP and the horizontal axis

measurements were limited to the period between the end of induction and the start of surgery. In addition, during variations in MAP we studied variations in PWV values. Our PWV measurements included, PWV_{AW} averaged over one minute while the patient was awake before anaesthesia induction and the average PWV_{GA} calculated from all PWV measurements during GA from the end of induction to the start of surgery. PWV_{GA} "at the same MAP levels" was calculated as the average PWV_{GA} considering only data values measured when MAP under GA (MAP_{GA}) varied less than 10% compared to awake MAP (MAP_{AW}). We also included variations in PWV under GA due to MAP changes, considering only those over 15%. After induction, norepinephrine was used at the anaesthetist discretion to counteract anaesthetic-induced vasoplegia (Supp. Mat. 1). This study focused on MAP variations resulting solely from vasopressor (norepinephrine) dosage changes, maintaining constant sedation to obtain a bispectral index (BIS) between 40 and 60 without surgical stimuli. For each patient, we analysed a maximum of one episode of MAP increase and one decrease. All measurement times are summarized in Fig. 1A.

2.3.2 Coupled analysis of PWVGA and MAP by constructing the α-angle

In this study, our objective was to continuously analyse PWV_{GA} in relation to MAP. For this analysis, continuous recordings of PWV_{GA} and MAP were needed. Given that most patients are monitored with non-invasive blood pressure methods, we assumed that blood pressure varied linearly between 2 oscillometric cuff measurements, taken at one-minute intervals.

During each MAP variation episode, we plotted MAP versus PWV_{GA} and computed the associated linear regression. To quantify the impact of MAP on PWV_{GA}, we introduced the α-angle, an index of dynamic arterial stiffness, defined by the angle at which the regression line intersects the horizontal axis. The methodology for constructing the α-angle is illustrated in Fig. 1B and C.

2.3.3 Comparing PWVA and PWVGA

We compare PWV_{AW} with PWV_{GA} and investigate their correlation. Given that PWV is influenced by MAP, we compared PWV_{AW} to both the average of all measured PWV_{GA} values and the average PWV_{GA} values at a MAP level similar (+/-10%) to the patient's awake state (Fig. 1A).

2.3.4 Groups for cardiovascular risk stratification

To assess the discriminative ability of PWV_{GA} and the α-angle in categorizing patients based on cardiovascular risk, we defined two groups: a high-risk group (CVR+) and a low-risk group (CVR-). The classification criteria included a major criterion (age > 50 years) and minor criteria (active smoking with cessation < 1 year, diabetes, dyslipidemia, arterial hypertension, or a previous cardiovascular event such as acute coronary syndrome, peripheral arterial disease, or stroke. Patients with one major criterion or at least two minor criteria were categorized as high cardiovascular risk (CVR+), while those without were considered low cardiovascular risk (CVR-).

2.4 Statistical analysis

For all tests, a 5% risk was used as a significance level. Categorical variables were expressed as percentages, and quantitative variables as medians with interquartile ranges. For paired data comparison, Wilcoxon tests were used. Spearman correlation analysis was used to compare PWV_{AW} and PWV_{GA}. A Mann-Whitney test was used to compare PWV_{GA} and α-angle values between high-risk (CVR+) and low-risk (CVR-) patients. For patients with two dynamic assessments (an increase and a decrease in MAP), we used the mean of the two α-angle values. The discriminating ability of these parameters for cardiovascular risk was assessed Receiver Operating Characteristic (ROC) curve construction based on a logistic regression model. The area under each ROC curve was measured for each parameter, and the optimal threshold for the best sensitivity and specificity was determined using the Youden index. All statistical analyses were conducted using Prism® software (GraphPad software, California, USA).

3 Results

3.1 Patients

In this study, 43 patients were initially included. However, 2 were excluded due to poor Doppler signal quality. Of the 41 patients, 22 were CVR- and 19 were CVR+. The demographic and clinical characteristics of these patients are detailed in Table 1. Of the 41 patients, 32 underwent both PWV_{AW} and PWV_{GA} evaluations, while 38 patients received at least one continuous dynamic PWV_{GA} evaluation, accounting for a total of 54 episodes of MAP changes, comprising 24 events of MAP increases and 30 MAP of decreases (Supp. Mat. 4). The median duration of these episodes was 6 [4–10] min, and the median number of PWV_{GA} and MAP values collected per episode were 350 [244–509], accounting for 93 [82–98] % of the total cardiac beats.

3.2 Comparison between PWVAW and PWVGA

Among 32 patients who underwent both awake and under GA PWV assessments, in a paired analysis, PWV_{AW} was significantly higher compared to PWV_{GA} (8.1 [6.9–10.3] vs. 7.4 [6.6–8.4] $m.s^{-1}$, $p < 0.0001$). A strong correlation between PWV_{AW} and PWV_{GA} ($r = 0.88$, $p < 0.001$) was found. Concurrently, MAP_{AW} values were also significantly higher than the mean of MAP_{GA} (97 [91–106] vs. 81 [74–90] mmHg, $p < 0.0001$). When considering PWV_{GA} values at comparable MAP levels as during awakening (variation < 10%), no significant difference was found between PWV_{AW} and PWV_{GA} (8.1 [6.9–10.3] vs. 7.9 [7.1–10.2] $m.s^{-1}$, $p = 0,13$), yet a strong correlation between them persisted ($r = 0.97$, $p < 0.0001$). Of note, the median duration for PWV_{GA} recordings was 16 [14–24] min, while the median duration of PWV_{GA} recordings analysed "at the same MAP level" was 4 [3–6] min. Results are presented in Fig. 2.

3.3 PWVAW and PWVGA according to CVR groups

We found that PWV_{AW} was significantly higher in the CVR + group compared to the CVR- group (10.7 [9.1–11.6]

Table 1 Characteristics of the population. Results are presented as median [interquartile range] or as counts (percentage). Patients were classified as being at high cardiovascular risk (CVR+) if they were over 50 years or had two of the following criteria: hypertension, active smoking, dyslipidemia, diabetes, a history of cardiovascular events (coronary heart disease, ischemic stroke, or lower extremity artery disease). Patients were classified as having a low cardiovascular risk (CVR-) if they did not meet these criteria. * ASA for American Society Anaesthesiologists score

	All	CVR -	CVR +	*P* value
	(*n* = 41)	(*n* = 22)	(*n* = 19)	
Demographic				
Age (years)	47 [35–66]	36 [31–44]	67 [59–76]	**< 0.0001**
Female gender	22 (54%)	13 (59%)	9 (47%)	0.45
Height (cm)	168 [165–179]	168 [165–181]	169 [163–178]	0.67
Weight (kg)	74 [64–81]	74 [66–83]	74 [59–82]	0.51
BMI (kg/m2)	24.5 [22.1–27.9]	25.1 [23.3–27.6]	22.9 [21.1–29.0]	0.58
Carotid-femoral distance (cm)	65 [61–68]	65 [62–67]	64 [59–68]	0.51
Hypertension	13 (32%)	3 (14%)	10 (53%)	**< 0.001**
Smokers	11 (27%)	5 (23%)	6 (32%)	0.52
Dyslipidemia	6 (15%)	1 (5%)	5 (26%)	**< 0.05**
Diabetes Mellitus	4 (10%)	1 (5%)	3 (16%)	0.22
Cardiovascular history	2 (5%)	0	2 (11%)	0.12
ASA*				
ASA 1	9 (22%)	8 (36%)	1 (5%)	**< 0.05**
ASA 2	29 (71%)	14 (64%)	15 (79%)	0.28
ASA 3	2 (5%)	0	2 (11%)	0.12
ASA 4	1 (2%)	0	1 (5%)	0.28
Surgical procedures				
Interventional neuroradiology	32 (78%)	19 (86%)	13 (68%)	0.17
Orthopedic surgery	4 (10%)	1 (5%)	3 (16%)	0.23
Visceral surgery	5 (12%)	2 (9%)	3 (16%)	0.51

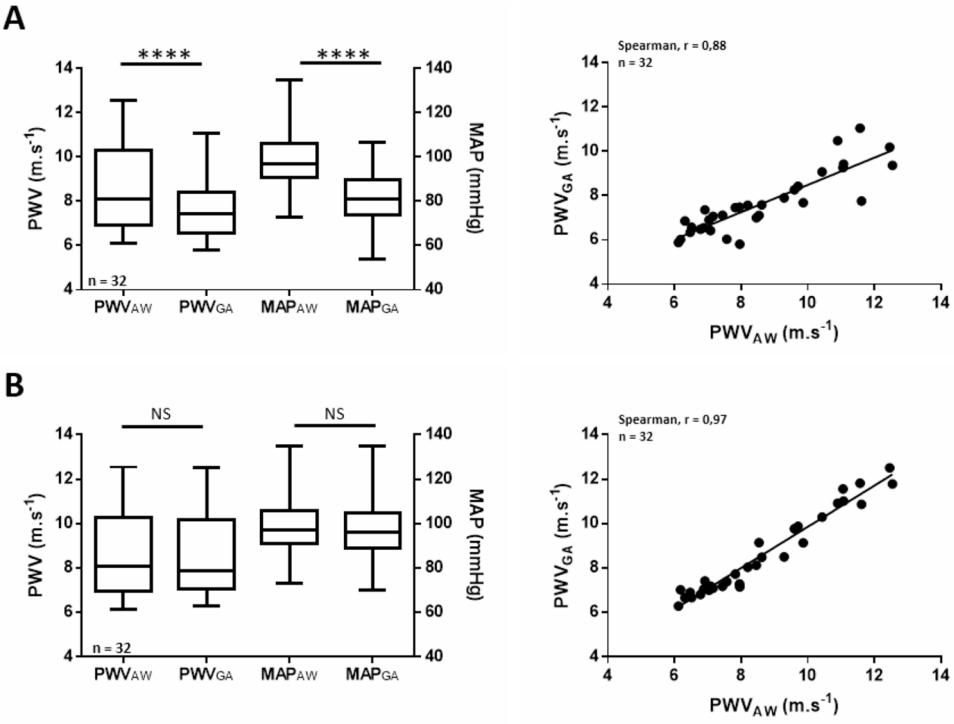


Fig. 2 Comparison and correlation between PWV_{AW} and PWV_{GA}. **A**: comparison between PWV_{AW} and the mean of PWV_{GA} measurements. **B**: comparison between PWV_{AW} and PWV_{GA} readings at identical MAP level (variation in MAP between awakening and GA < 5%). ****$p < 0.0001$ and NS = not significant

$m.s^{-1}$ vs. 7.1 [6.7-8.0], $p<0.001$). Similarly, PWV_{GA} was also lower in the CVR- group than in the CVR + group (6.7 [6.1–7.2] vs. 8.1 [7.1–9.4] $m.s^{-1}$, $p<0.001$, see Supp. Mat. 5).

3.4 PWV_{GA} during MAP variation episodes and α-angle analysis

For each of the 54 MAP changes episodes, we constructed a linear regression using PWV_{GA} and MAP signals. The median value of the α-angle computed from these linear regression was 1.9 [1.0-2.7] degrees. Importantly, we found no significant differences in α-angle values between episodes of MAP increase and decrease (1.9 [1.2–2.3] vs. 1.9 [0.9–3.1] degrees, $p=0.88$). Furthermore, the median correlation coefficient between MAP and PWV_{GA} across all MAP episodes was 0.85 [0.72–0.93]. PWV_{GA} varied in response to MAP changes: it increased from 6.6 [5.9–7.4] to 7.0 [6.6–8.4] m.s-^{1} when MAP rose from 69 [63–73] to 95 [90–101] mmHg ($p<0.0001$), and decreased from 7.4 [6.9–9.2] to 7.0 [6.2–7.8] $m.s^{-1}$ when MAP fell from 97 [90–108] to 71 [65–77] mmHg ($p<0.0001$). No significant variations in heart rate were observed during these dynamic assessments of PWV_{GA} (Supp. Mat. 6).

3.5 PWV and α-angle analysis according to the cardiovascular risk status

We were then interested in comparing the α-angle based on cardiovascular risk factors. We found that the α-angle was significantly lower in the CVR- group compared to the CVR + group (1.3 [0.7-2.0] vs. 2.6 [2.0-3.5] degrees, $p<0.001$). Observing that CVR- patients had lower PWV_{GA} and α-angles compared to CVR + patients, we investigated the product of these two parameters (α-angle x PWV_{GA}), calculated using mean values of PWV_{GA} and α-angle measured during a MAP episode. This product was significantly lower in the CVR- group compared to the CVR + group (8.1 [4.8–13.8] vs. 21.8 [14.7–34.7] $degrees.m.s^{-1}$, $p<0.0001$). The mean MAP under GA did not differ significantly between the CVR- and CVR + groups (81 [76–86] mmHg vs. 87 [74–90] mmHg, $p=0,56$). Results are shown in Fig. 3.

3.6 CVR Prediction using PWV_{GA} and α-angle

To assess the ability to predict cardiovascular risk, we constructed ROC curves for each of three parameters studied: PWV_{GA}, α-angle, and the product (α x PWV_{GA}). The AUC were 0.80 [0.65–0.95] for PWV_{GA}, 0.83 [0.69–0.96] for α-angle, and 0.86 [0.74–0.97] for the product (α x PWV_{GA}). Using the Youden index, we determined that the optimal threshold for PWV_{GA} was at 7.6 $m.s^{-1}$, leading to a sensitivity of 65 [38–86] % and a specificity of 95 [76–100] %. For the α-angle, the threshold was at 2.0 degrees, yielding to a sensitivity of 76 [50–93] % and a specificity of 81 [58–95] %. Finally, for the product (α x PWV_{GA}), a value of 15.2 $degrees.m.s^{-1}$ resulted in a sensitivity of 76 [50–93] % and a specificity of 81 [58–95] %. The ROC curves are shown in Fig. 4.

4 Discussion

The present study has revealed several significant insights. First and foremost, we found that assessing arterial stiffness using carotid and femoral Doppler under GA was feasible in a high proportion of our patients (41 out of 43, or 95%). Furthermore, our results showed that PWV while awake

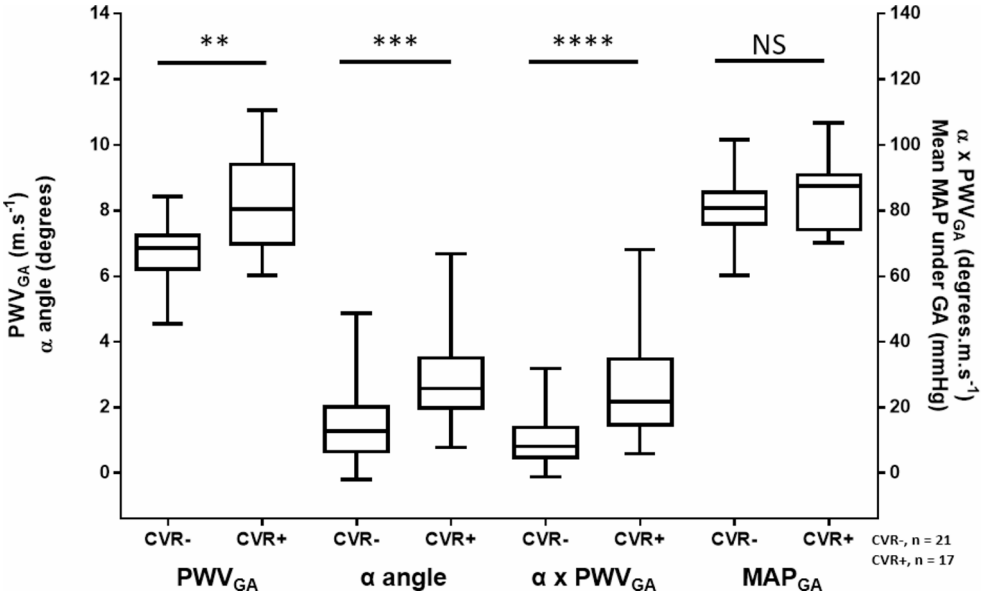


Fig. 3 PWV_{AG}, α-angle, product (α x PWV_{GA}) and mean MAP under GA (MAP_{GA}) as a function of cardiovascular risk status. $^{**}p<0.01$. $^{***}p<0.001$, $^{****}p<0.0001$, NS = not significant

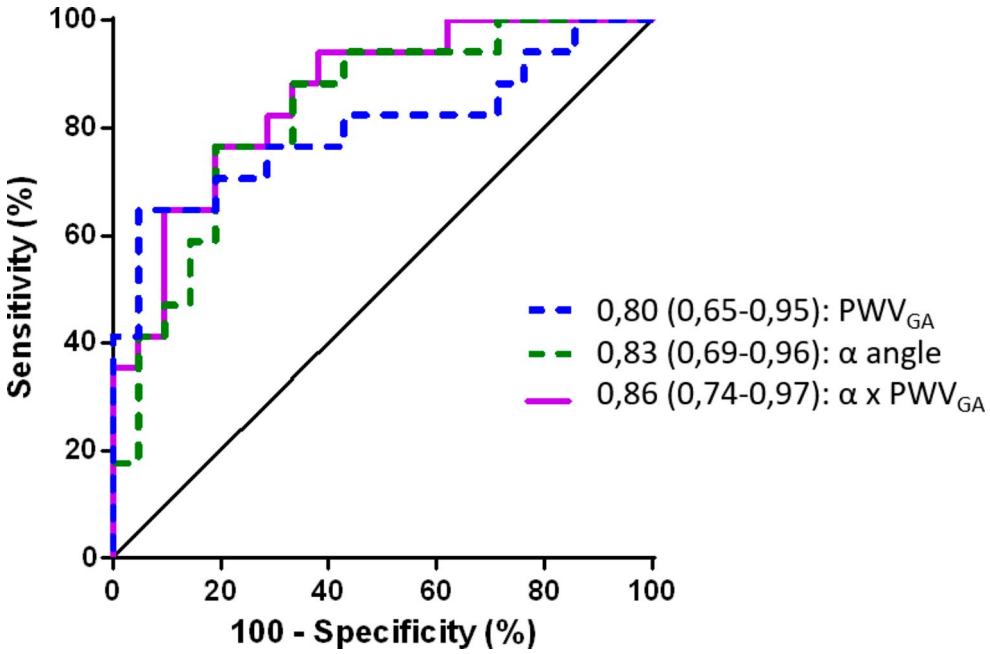


Fig. 4 ROC AUC curves to predict cardiovascular risk with different parameters. PWV_{GA} (dotted blue line), α-angle (dotted green line) and product (PWV_{GA} x α) (purple line)

(PWV_{AW}) were significantly higher than the mean PWV under GA (PWV_{GA}), yet were strongly correlated. Moreover, this correlation improved when analysis was conducted at comparable MAP levels. We also observed in our cohort a significant correlation between PWV and age, both in awake and anaesthetized states. These findings illustrate the technical feasibility and reliability of PWV measurements under GA.

Our findings demonstrate that PWV_{GA} is influenced by the MAP level. This dependency aligns with outcomes from various experimental and clinical studies that have investigated the changes in pulse wave transit time during GA [15]. For instance, Yurie Obata and colleagues [16] found a positive correlation between MAP and PWV measured under GA in 25 cardiac surgery patients using transesophageal echocardiography and invasive blood pressure monitoring. Our study echoes these findings, demonstrating a similar correlation through a non-invasive evaluation of arterial stiffness, thereby highlighting the relationship between PWV and MAP.

Examination of the α-angle as a marker of dynamic arterial stiffness allowed us to identify that the relationship between PWV and MAP varied according to cardiovascular risk status. Indeed, the α-angle may indirectly reflect the sensitivity of arterial stiffness to change in MAP. Our observation revealed that patients with low cardiovascular risk (CVR-) had a lower α-angle, suggesting a reduced influence of MAP on their PWV, indicative of lower arterial stiffness. In contrast, high-risk patients (CVR+) had a higher α-angle, implying in a PWV considerably affected by MAP levels and indicative of a higher arterial stiffness. Findings also in line with the recent study of Salvi and colleagues, which showed a decline in arterial distensibility with age among 75 healthy volunteers [17].

Using PWV and α-angle measurements provide a complementary approach to differentiate patients according to their cardiovascular risk. In particular, PWV_{GA} measurement demonstrated high specificity at a threshold of 7.6 m.s^{-1}, indicating that patients above this threshold are at an increased cardiovascular risk. Conversely, the α-angle showed good sensitivity at a 1.4 degrees threshold, indicating a lower cardiovascular risk for patients below this level. Finally, using the (α x PWV_{GA}) product as a composite index enhanced the effectiveness in discriminating patients according to cardiovascular risk.

4.1 Limitations

Our study has several limitations. Firstly, the patient cohort was highly selected, comprising mainly individuals with few comorbidities and a low to moderate anaesthetic risk. This means that the groups are essentially discriminated by age and not by other cardiovascular risks such as hypertension or diabetes. Additionally, these patients had little impairment of their cardiovascular function, with only 8 having a PWV exceeding the 10 m.s^{-1} threshold proposed by European guidelines for defining arterial stiffness (constituting 20% of our study population).

The methodology we used to define cardiovascular risk, according to age and cardiovascular risk factors, isn't a widely recognized classification system. Traditional anaesthesia risk assessment tools such as American Society of Anaesthesiologists score or the Lee score focus on cardiovascular risk factors but omit age, which is a crucial factor

in assessing arterial stiffness [18]. Moreover, the lack of systematic collection of certain biological data during the anaesthetic consultation precluded the use of established cardiology tools like the SCORE diagram [19] or the Framingham score [20] for cardiovascular risk stratification.

Another limitation concerns the lack of standardization in the episode of MAP increase and decrease during dynamic PWV_{GA} measurements. We were unable to predict the duration, linearity, or values of low and high MAP in the dynamic tests because of the variability in individual responses to drugs. Indeed, the relative doses of hypnotics and vasopressors (norepinephrine) used to achieve a certain MAP level were not accounted for yet could directly impact arterial stiffness.

Intermittently blood pressure monitoring using oscillometric cuffs, while common in anaesthesia, has its accuracy debated, especially in patients receiving vasopressors [21]. Furthermore, the assumption that MAP changes linearly between 2 consecutive measurements taken at 1-minute intervals is also questionable. It would have been interesting to obtain a continuous measurement of MAP to improve the reliability of PWV and MAP relationship. Finally, exploring correlation between PWV and other hemodynamic parameters such as systolic and diastolic blood pressure, pulse pressure, or stroke volume, as done in previous studies [15, 16], would have added value to our findings.

4.2 Perspectives

Our study successfully demonstrates the feasibility of non-invasively assessing arterial stiffness under GA, a significant consideration given the 9 million GA procedures performed annually in France. However, using carotid-femoral method remains difficult for large-scale use, and could also be hindered considering the access limitations to carotid and femoral sites during certain surgeries. It would therefore be interesting to find a simple alternative method using the usual monitoring tools in the operating room.

The measurement of PWV is now well integrated in European guidelines for hypertension management [18, 22]. The assessment of damage to target organs (in organs like the heart, brain, and kidneys), with a threshold value of 10 $m.s^{-1}$, has enhanced the prediction of cardiovascular risk [23], and led to patient reclassification for better therapeutic management [24]. Nevertheless, the effect of reducing arterial stiffness on reducing cardiovascular events remains uncertain. For instance, a study on patient with end-stage chronic renal failure showed improvement survival with reduced PWV [25], yet the multicenter SPARTE study did not find clear benefit in normalizing arterial stiffness in a hypertensive cohort [26]. It would be interesting to evaluate if arterial stiffness dynamically measured could improve the evaluation of cardiovascular risk from a single PWV measurement. Such an approach could potentially target at an earlier stage, patients who could benefit from normalising arterial stiffness.

Finally, arterial stiffness may serve as a valuable monitoring parameter in anaesthesia. In France, two-thirds of perioperative deaths are attributed to cardiovascular causes. Optimising haemodynamic is crucial to prevent these complications, and number of recent studies suggest that MAP targets should be individualized for each patient, with higher MAP targets in hypertensive patients [27, 28]. Monitoring arterial stiffness, a key determinant of cardiac afterload, can help to adjust fluid, vasopressor or inotrope administration to improve ventriculo-arterial coupling and intraoperative haemodynamic. The α-angle could significantly enhance the traditional application of PWV, particularly for real-time cardiovascular monitoring [29].

5 Conclusion

Measurement of PWV under GA using carotid and femoral Doppler is a feasible method to continuously assess arterial stiffness under general anaesthesia. Further studies are required to validate the α-angle parameter in different physiological conditions.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10877-024-01256-w.

Author contributions TH: conceptualization, methodology, formal analysis, investigation and writing; JJ: methodology, resources and formal analysis; PB: resources, review and editing; JM: resources, review and editing; JP: methodology and formal analysis; AM: review and editing; JC: formal analysis, review and editing; FV: conceptualization, methodology, formal analysis, investigation, supervision and writing.

Data availability Data is provided within the manuscript or supplementary information files.

Declarations

Competing interests The authors declare no competing interests.

References

1. O'Donnell MJ, Xavier D, Liu L, Zhang H, Chin SL, Rao-Melacini P, et al. Risk factors for ischaemic and intracerebral haemorrhagic stroke in 22 countries (the INTERSTROKE study): a case-control study. Lancet 10 juill. 2010;376(9735):112–23.
2. Ben-Shlomo Y, Spears M, Boustred C, May M, Anderson SG, Benjamin EJ et al. Aortic pulse wave velocity improves cardiovascular event prediction: an individual participant meta-analysis

of prospective observational data from 17,635 subjects. J Am Coll Cardiol. 25 févr. 2014;63(7):636–46.
3. Vlachopoulos C, Aznaouridis K, Stefanadis C. Prediction of cardiovascular events and all-cause mortality with arterial stiffness: a systematic review and meta-analysis. J Am Coll Cardiol. 30 mars. 2010;55(13):1318–27.
4. Laurent S, Boutouyrie P, Asmar R, Gautier I, Laloux B, Guize L, et al. Aortic stiffness is an independent predictor of all-cause and cardiovascular mortality in hypertensive patients. Hypertens mai. 2001;37(5):1236–41.
5. Mitchell GF, Hwang SJ, Vasan RS, Larson MG, Pencina MJ, Hamburg NM, et al. Arterial stiffness and cardiovascular events: the Framingham Heart Study. Circulation 2 févr. 2010;121(4):505–11.
6. Boutouyrie P, Tropeano AI, Asmar R, Gautier I, Benetos A, Lacolley P, et al. Aortic stiffness is an independent predictor of primary coronary events in hypertensive patients: a longitudinal study. Hypertens janv. 2002;39(1):10–5.
7. Laurent S, Katsahian S, Fassot C, Tropeano AI, Gautier I, Laloux B, et al. Aortic stiffness is an independent predictor of fatal stroke in essential hypertension. Stroke mai. 2003;34(5):1203–6.
8. Laurent S, Cockcroft J, Van Bortel L, Boutouyrie P, Giannattasio C, Hayoz D, et al. Expert consensus document on arterial stiffness: methodological issues and clinical applications. Eur Heart J nov. 2006;27(21):2588–605.
9. Reference Values for Arterial Stiffness' Collaboration. Determinants of pulse wave velocity in healthy people and in the presence of cardiovascular risk factors: « establishing normal and reference values ». Eur Heart J oct. 2010;31(19):2338–50.
10. Bramwell JC, Hill AV, VELOCITY OF TRANSMISSION OF THE PULSE-WAVE. Lancet mai. 1922;199(5149):891–2.
11. Mattace-Raso FUS, van der Cammen TJM, Knetsch AM, van den Meiracker AH, Schalekamp MADH, Hofman A, et al. Arterial stiffness as the candidate underlying mechanism for postural blood pressure changes and orthostatic hypotension in older adults: the Rotterdam Study. J Hypertens févr. 2006;24(2):339–44.
12. Stewart AD, Millasseau SC, Kearney MT, Ritter JM, Chowienczyk PJ. Effects of Inhibition of Basal Nitric Oxide Synthesis on Carotid-Femoral Pulse Wave Velocity and Augmentation Index in Humans. Hypertens nov. 2003;42(5):915–8.
13. Omboni S, Posokhov IN, Rogoza AN. Relationships between 24-h blood pressure variability and 24-h central arterial pressure, pulse wave velocity and augmentation index in hypertensive patients. Hypertens Res avr. 2017;40(4):385–91.
14. Van Bortel LM, Laurent S, Boutouyrie P, Chowienczyk P, Cruickshank JK, De Backer T, et al. Expert consensus document on the measurement of aortic stiffness in daily practice using carotid-femoral pulse wave velocity. J Hypertens mars. 2012;30(3):445–8.
15. Chen YT, Chiayg CY, Wang MC, Tsai WC, Wu HT, Liu CC. Serial changes of pulse wave velocity and correlations with hemodynamic parameters during general anesthesia. Acta Anaesthesiol Taiwan déc. 2006;44(4):193–8.
16. Obata Y, Mizogami M, Singh S, Nyhan D, Berkowitz DE, Steppan J, et al. The Effects of Hemodynamic Changes on Pulse Wave Velocity in Cardiothoracic Surgical Patients. Biomed Res Int. 2016;2016:9640457.
17. Salvi P, Valbusa F, Kearney-Schwartz A, Labat C, Grillo A, Parati G, et al. Non-Invasive Assessment of Arterial Stiffness: Pulse Wave Velocity, Pulse Wave Analysis and Carotid Cross-Sectional Distensibility: Comparison between Methods. J Clin Med janv. 2022;11(8):2225.
18. Williams B, Mancia G, Spiering W, Agabiti Rosei E, Azizi M, Burnier M, et al. 2018 ESC/ESH Guidelines for the management of arterial hypertension: The Task Force for the management of arterial hypertension of the European Society of Cardiology and the European Society of Hypertension: The Task Force for the management of arterial hypertension of the European Society of Cardiology and the European Society of Hypertension. J Hypertens oct. 2018;36(10):1953–2041.
19. Conroy RM, Pyörälä K, Fitzgerald AP, Sans S, Menotti A, De Backer G, et al. Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. Eur Heart J juin. 2003;24(11):987–1003.
20. Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. Circulation. 12 mai. 1998;97(18):1837–47.
21. Wax DB, Lin HM, Leibowitz AB. Invasive and concomitant noninvasive intraoperative blood pressure monitoring: observed differences in measurements and associated therapeutic interventions. Anesthesiology nov. 2011;115(5):973–8.
22. Wilkinson I, Mceniery C, Schillaci G, Boutouyrie P, Segers P, Donald A, et al. ARTERY Society guidelines for validation of non-invasive haemodynamic measurement devices: Part 1, arterial pulse wave velocity. Artery Res 1 juin. 2010;4:34–40.
23. Safar ME, Levy BI, Struijker-Boudier H. Current Perspectives on Arterial Stiffness and Pulse Pressure in Hypertension and Cardiovascular Diseases. Circulation. 10 juin. 2003;107(22):2864–9.
24. Sehestedt T, Jeppesen J, Hansen TW, Wachtell K, Ibsen H, Torp-Pedersen C, et al. Risk prediction is improved by adding markers of subclinical organ damage to SCORE. Eur Heart J avr. 2010;31(7):883–91.
25. Guerin AP, Blacher J, Pannier B, Marchais SJ, Safar ME, London GM. Impact of aortic stiffness attenuation on survival of patients in end-stage renal failure. Circulation. 20 févr. 2001;103(7):987–92.
26. Laurent S, Chatellier G, Azizi M, Calvet D, Choukroun G, Danchin N, et al. SPARTE Study: Normalization of Arterial Stiffness and Cardiovascular Events in Patients With Hypertension at Medium to Very High Risk. Hypertens oct. 2021;78(4):983–95.
27. Lienhart A, Auroy Y, Péquignot F, Benhamou D, Warszawski J, Bovet M, et al. Survey of anesthesia-related mortality in France. Anesthesiology déc. 2006;105(6):1087–97.
28. Futier E, Lefrant JY, Guinot PG, Godet T, Lorne E, Cuvillon P, et al. Effect of individualized vs standard blood pressure management strategies on postoperative organ dysfunction among high-risk patients undergoing major surgery: a randomized clinical trial. JAMA. 2017;10(14):1346–57.
29. Wu HT, Hsu PC, Liu AB, Chen ZL, Huang RM, Chen CP, et al. Six-channel ECG-based pulse wave velocity for assessing whole-body arterial stiffness. Blood Press juin. 2012;21(3):167–76.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Bibliography

Aasvang, Eske K. and Christian S. Meyhoff (Dec. 2023). “The future of postoperative vital sign monitoring in general wards: improving patient safety through continuous artificial intelligence-enabled alert formation and reduction”. en-US. In: *Current Opinion in Anesthesiology* 36.6, p. 683 (cit. on p. 23).

Advancing Safety in Medical Technology (AAMI) (2013). *ANSI/AAMI/ISO 81060-2:2013 Non-invasive sphygmomanometers - part 2: clinical investigation of automated measurement type.* (Cit. on p. 46).

Allen, John (Feb. 2007). “Photoplethysmography and its application in clinical physiological measurement”. en. In: *Physiological Measurement* 28.3, R1 (cit. on pp. 45, 65).

Angelopoulos, Anastasios N. and Stephen Bates (Dec. 2022). *A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification*. en (cit. on pp. 87, 124).

Assran, Mahmoud, Quentin Duval, Ishan Misra, et al. (Apr. 2023). *Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture*. en (cit. on p. 36).

Athaya, Tasbiraha and Sunwoong Choi (Mar. 2021). “An Estimation Method of Continuous Non-Invasive Arterial Blood Pressure Waveform Using Photoplethysmography: A U-Net Architecture-Based Approach”. In: *Sensors (Basel, Switzerland)* 21.5, p. 1867 (cit. on p. 68).

Attivissimo, Filippo, Vito Ivano D’Alessandro, Luisa De Palma, Anna Maria Lucia Lanzolla, and Attilio Di Nisio (Jan. 2023). “Non-Invasive Blood Pressure Sensing via Machine Learning”. en. In: *Sensors* 23.19, p. 8342 (cit. on p. 67).

Aubin, O. Saint, I. Khemir, J. Perdereau, et al. (July 2023). “Repurposing electroencephalographic signal for automatic segmentation of intra-operative periods under general anesthesia”. In: *IEEE EUROCON 2023 - 20th International Conference on Smart Technologies*, pp. 286–290 (cit. on p. 22).

Ayeb, Kanssaa El, Cyril Touchard, Valentin Iovene, et al. (July 2025). “Using intraoperative physiological data to predict postoperative cognitive decline and cardiovascular events: results from a 30-month prospective study”.

English. In: *British Journal of Anaesthesia* 135.1, pp. 254–256 (cit. on p. 23).

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio (May 2016). *Neural Machine Translation by Jointly Learning to Align and Translate*. en (cit. on p. 32).

Baltrusaitis, Tadas, Chaitanya Ahuja, and Louis-Philippe Morency (Feb. 2019). "Multimodal Machine Learning: A Survey and Taxonomy". en. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41.2, pp. 423–443 (cit. on pp. 26, 125).

Banville, Hubert, Omar Chehab, Aapo Hyvärinen, Denis-Alexander Engemann, and Alexandre Gramfort (Aug. 2021). "Uncovering the structure of clinical EEG signals with self-supervised learning". en. In: *Journal of Neural Engineering* 18.4, p. 046020 (cit. on pp. 36, 94, 96, 101).

Belur Nagaraj, Sunil, Sowmya M. Ramaswamy, Maud A. S. Weerink, and Michel M. R. F. Struys (May 2020). "Predicting Deep Hypnotic State From Sleep Brain Rhythms Using Deep Learning: A Data-Repurposing Approach". en-US. In: *Anesthesia & Analgesia* 130.5, p. 1211 (cit. on p. 46).

Ben-Shlomo, Yoav, Melissa Spears, Chris Boustred, et al. (Feb. 2014). "Aortic pulse wave velocity improves cardiovascular event prediction: an individual participant meta-analysis of prospective observational data from 17,635 subjects". eng. In: *Journal of the American College of Cardiology* 63.7, pp. 636–646 (cit. on p. 66).

Bhattachrjee, Soumyadeep, Huining Li, Jun Xia, and Wenyao Xu (June 2023). "*SimPPG*: Self-supervised photoplethysmography-based heart-rate estimation via similarity-enhanced instance discrimination". In: *Smart Health* 28, p. 100396 (cit. on pp. 94, 106).

Bi, Yuhua, Junping Huang, Mei Li, Siying Li, and Heshou Lei (Nov. 2023). "Monitoring and evaluation of anesthesia depth status data based on neuroscience". In: *Open Life Sciences* 18.1, p. 20220719 (cit. on p. 6).

Bijker, Jilles B., Wilton A. van Klei, Teus H. Kappen, et al. (Aug. 2007). "Incidence of Intraoperative Hypotension as a Function of the Chosen Definition: Literature Definitions Applied to a Retrospective Cohort Using Automated Data Collection". en. In: *Anesthesiology* 107.2, pp. 213–220 (cit. on pp. 9, 65).

Bijker, Jilles B., Suzanne Persoon, Linda M. Peelen, et al. (Mar. 2012). "Intraoperative Hypotension and Perioperative Ischemic Stroke after General Surgery". en. In: *Anesthesiology* 116.3, pp. 658–664 (cit. on p. 11).

Bousefsaf, Frédéric, Théo Desquins, Djamaleddine Djeldjli, et al. (Sept. 2022). "Estimation of blood pressure waveform from facial video using a deep U-shaped network and the wavelet representation of imaging photo-

plethysmographic signals". In: *Biomedical Signal Processing and Control* 78, p. 103895 (cit. on p. 67).
Brady, Kenneth and Charles W. Hogue (Sept. 2013). "Intraoperative Hypotension and Patient Outcome". In: *Anesthesiology* 119.3, pp. 495–497 (cit. on p. 11).
Bruhn, J., P. S. Myles, R. Sneyd, and M. M. R. F. Struys (July 2006). "Depth of anaesthesia monitoring: what's available, what's validated and what's next?" English. In: *British Journal of Anaesthesia* 97.1, pp. 85–94 (cit. on p. 92).
Bubenik, Peter (n.d.). "Statistical Topological Data Analysis using Persistence Landscapes". en. In: () (cit. on p. 112).
Cartailler, Jérôme, Victor Beaucote, Bernard Trillat, et al. (Sept. 2024). "Deciphering the explanatory potential of blood pressure variables on postoperative length of stay through hierarchical clustering: A retrospective monocentric study". en. In: *PLOS ONE* 19.9, e0308910 (cit. on pp. 9, 67).
Casas-Arroyave, Fabián David, Juan Manuel Fernández, and John Jairo Zuleta-Tobón (Apr. 2019). "Evaluation of a closed-loop intravenous total anesthesia delivery system with BIS monitoring compared to an open-loop target-controlled infusion (TCI) system: randomized controlled clinical trial". en. In: *Colombian Journal of Anesthesiology* 47.2, pp. 84–91 (cit. on p. 26).
Castanedo, Federico (2013). "A Review of Data Fusion Techniques". en. In: *The Scientific World Journal* 2013.1, p. 704504 (cit. on p. 25).
Chambon, Stanislas, Mathieu Galtier, Pierrick Arnal, Gilles Wainrib, and Alexandre Gramfort (Nov. 2017). *A deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series*. en (cit. on p. 94).
Chambon, Stanislas, Valentin Thorey, Pierrick J. Arnal, Emmanuel Mignot, and Alexandre Gramfort (July 2018). *A deep learning architecture to detect events in EEG signals during sleep*. en (cit. on pp. 97, 102, 103, 122).
Chazal, Frederic, Vin de Silva, Marc Glisse, and Steve Oudot (Mar. 2013). *The structure and stability of persistence modules*. en (cit. on p. 112).
Checketts, M. R., R. Alladi, K. Ferguson, et al. (2016). "Recommendations for standards of monitoring during anaesthesia and recovery 2015 : Association of Anaesthetists of Great Britain and Ireland". en. In: *Anaesthesia* 71.1, pp. 85–93 (cit. on p. 92).
Chen, Ricky T. Q., Yulia Rubanova, Jesse Bettencourt, and David Duvenaud (Dec. 2019). *Neural Ordinary Differential Equations*. en (cit. on p. 61).

Chen, Ting, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton (July 2020). *A Simple Framework for Contrastive Learning of Visual Representations*. en (cit. on p. 96).

Cheng, Jianpeng, Li Dong, and Mirella Lapata (Sept. 2016). *Long Short-Term Memory-Networks for Machine Reading*. en (cit. on p. 32).

Cheng, Juan, Yufei Xu, Rencheng Song, et al. (Nov. 2021). "Prediction of arterial blood pressure waveforms from photoplethysmogram signals via fully convolutional neural networks". In: *Computers in Biology and Medicine* 138, p. 104877 (cit. on p. 68).

Cook, T. M., J. Andrade, D. G. Bogod, et al. (Oct. 2014). "5th National Audit Project (NAP5) on accidental awareness during general anaesthesia: patient experiences, human factors, sedation, consent, and medicolegal issues†‡". In: *British Journal of Anaesthesia* 113.4, pp. 560–574 (cit. on p. 92).

Coutrot, Maxime, Jona Joachim, François Dépret, et al. (Mar. 2019). "Non-invasive continuous detection of arterial hypotension during induction of anaesthesia using a photoplethysmographic signal: proof of concept". en. In: *British Journal of Anaesthesia*, S0007091219300960 (cit. on p. 45).

Dahl, G. E., Dong Yu, Li Deng, and A. Acero (Jan. 2012). "Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition". en. In: *IEEE Transactions on Audio, Speech, and Language Processing* 20.1, pp. 30–42 (cit. on pp. 46, 49).

Dai, Dong, Zhaohui Ji, and Haiyan Wang (Jan. 2024). "Non-Invasive Continuous Blood Pressure Estimation from Single-Channel PPG Based on a Temporal Convolutional Network Integrated with an Attention Mechanism". en. In: *Applied Sciences* 14.14, p. 6061 (cit. on p. 68).

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (May 2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. en (cit. on p. 34).

Ding, Ruiwen, Anil Yadav, Erika Rodriguez, Ana Cristina Araujo Lemos da Silva, and William Hsu (Nov. 2023). "Tailoring pretext tasks to improve self-supervised learning in histopathologic subtype classification of lung adenocarcinomas". In: *Computers in Biology and Medicine* 166, p. 107484 (cit. on p. 125).

Ding, Xiao-Rong, Yuan-Ting Zhang, Jing Liu, Wen-Xuan Dai, and Hon Ki Tsang (May 2016). "Continuous Cuffless Blood Pressure Estimation Using Pulse Transit Time and Photoplethysmogram Intensity Ratio". eng. In: *IEEE transactions on bio-medical engineering* 63.5, pp. 964–972 (cit. on p. 67).

Doutreligne, Matthieu, Adeline Degremont, Pierre-Alain Jachiet, Antoine Lamer, and Xavier Tannier (July 2023). "Good practices for clinical data

warehouse implementation: A case study in France". en. In: *PLOS Digital Health* 2.7, e0000298 (cit. on p. 18).
Duan, Junbo, Qing Wang, Bo Zhang, et al. (Aug. 2022). "Accurate detection of atrial fibrillation events with R-R intervals from ECG signals". en. In: *PLOS ONE* 17.8, e0271596 (cit. on p. 23).
Edelsbrunner, Herbert and John Harer (2008). "Persistent homology—a survey". en. In: *Contemporary Mathematics*. Ed. by Jacob E. Goodman, János Pach, and Richard Pollack. Vol. 453. Providence, Rhode Island: American Mathematical Society, pp. 257–282 (cit. on p. 112).
Elman, Jeffrey L. (Apr. 1990). "Finding structure in time". In: *Cognitive Science* 14.2, pp. 179–211 (cit. on p. 31).
Ericsson, Linus, Henry Gouk, Chen Change Loy, and Timothy M. Hospedales (May 2022). "Self-Supervised Representation Learning: Introduction, Advances and Challenges". en. In: *IEEE Signal Processing Magazine* 39.3, pp. 42–62 (cit. on p. 36).
Fecho, Karamarie, Anne T Lunney, Philip G Boysen, Peter Rock, and Edward A Norfleet (Aug. 2008). "Postoperative mortality after inpatient surgery: Incidence and risk factors". In: *Therapeutics and Clinical Risk Management* 4.4, pp. 681–688 (cit. on p. 2).
Finnegan, Eoin, Shaun Davidson, Mirae Harford, et al. (Jan. 2023). "Features from the photoplethysmogram and the electrocardiogram for estimating changes in blood pressure". en. In: *Scientific Reports* 13.1, p. 986 (cit. on p. 66).
Fritz, Bradley A., Zhicheng Cui, Muhan Zhang, et al. (Nov. 2019). "Deep-learning model for predicting 30-day postoperative mortality". en. In: *British Journal of Anaesthesia* 123.5, pp. 688–695 (cit. on p. 45).
Futier, Emmanuel, Jean-Yves Lefrant, Pierre-Gregoire Guinot, et al. (2017). "Effect of Individualized vs Standard Blood Pressure Management Strategies on Postoperative Organ Dysfunction Among High-Risk Patients Undergoing Major Surgery: A Randomized Clinical Trial". eng. In: *JAMA* 318.14, pp. 1346–1357 (cit. on p. 10).
Gal, Yarin and Zoubin Ghahramani (June 2016). "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning". en. In: *Proceedings of The 33rd International Conference on Machine Learning*. PMLR, pp. 1050–1059 (cit. on p. 124).
Gao, Jing, Peng Li, Zhikui Chen, and Jianing Zhang (May 2020). "A Survey on Deep Learning for Multimodal Data Fusion". en. In: *Neural Computation* 32.5, pp. 829–864 (cit. on pp. 26, 125).

Ghoneim, Mohamed M. (Sept. 2007). “Incidence of and risk factors for awareness during anaesthesia”. In: *Best Practice & Research Clinical Anaesthesiology*. Awareness during Anaesthesia 21.3, pp. 327–343 (cit. on p. 92).

Ghorbani, Ramin, Marcel J. T. Reinders, and David M. J. Tax (Dec. 2022). *Self-Supervised PPG Representation Learning Shows High Inter-Subject Variability*. en (cit. on p. 106).

Gnassounou, Theo, Rémi Flamary, and Alexandre Gramfort (n.d.). “Convolutional Monge Mapping Normalization for learning on sleep data”. en. In: () (cit. on p. 113).

Goddard, Nick and David Smith (Dec. 2013). “Unintended awareness and monitoring of depth of anaesthesia”. English. In: *Continuing Education in Anaesthesia, Critical Care and Pain* 13.6, pp. 213–217 (cit. on p. 92).

Goertz, Axel W., Michael Schmidt, Christian Seefelder, Karl H. Lindner, and Michael Georgieff (Aug. 1993). “The Effect of Phenylephrine Bolus Administration on Left Ventricular Function During Isoflurane-Induced Hypotension”. en-US. In: *Anesthesia & Analgesia* 77.2, p. 227 (cit. on p. 12).

Gottschalk, André, Hugo Van Aken, Michael Zenz, and Thomas Standl (July 2011). “Is Anesthesia Dangerous?” In: *Deutsches Ärzteblatt International* 108.27, pp. 469–474 (cit. on p. 2).

Guetschel, Pierre, Thomas Moreau, and Michael Tangermann (2025). “S-JEPA: towards seamless cross-dataset transfer through dynamic spatial attention”. In: () (cit. on p. 36).

Hajat, Z., N. Ahmad, and J. Andrzejowski (2017). “The role and limitations of EEG-based depth of anaesthesia monitoring in theatres and intensive care”. en. In: *Anaesthesia* 72.S1, pp. 38–47 (cit. on p. 92).

El-Hajj, C and P. A Kyriacou (Sept. 2021). “Cuffless blood pressure estimation from PPG signals and its derivatives using deep learning models”. In: *Biomedical Signal Processing and Control* 70, p. 102984 (cit. on pp. 65, 68, 123).

Harlé, Théophile, Jona Joachim, Pierre Boutouyrie, et al. (June 2025). “Continuous measurement of carotid-femoral pulse wave velocity (PWVcf.) during general anaesthesia using Doppler: a preliminary study”. en. In: *Journal of Clinical Monitoring and Computing* 39.3, pp. 495–503 (cit. on p. 66).

He, Kaiming, Xinlei Chen, Saining Xie, et al. (Dec. 2021). *Masked Autoencoders Are Scalable Vision Learners*. en (cit. on pp. 34, 106).

He, Kaiming, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick (Mar. 2020). *Momentum Contrast for Unsupervised Visual Representation Learning*. en (cit. on p. 96).

Hight, Darren, Matthias Kreuzer, Gesar Ugen, et al. (May 2023). “Five commercial ‘depth of anaesthesia’ monitors provide discordant clinical recommendations in response to identical emergence-like EEG signals”. English. In: *British Journal of Anaesthesia* 130.5, pp. 536–545 (cit. on p. 93).

Hochreiter, Sepp and Jürgen Schmidhuber (Nov. 1997). “Long Short-Term Memory”. In: *Neural Computation* 9.8, pp. 1735–1780 (cit. on p. 31).

Hopfield, J J (Apr. 1982). “Neural networks and physical systems with emergent collective computational abilities.” In: *Proceedings of the National Academy of Sciences of the United States of America* 79.8, pp. 2554–2558 (cit. on p. 31).

Hoshide, Satoshi, Akiomi Yoshihisa, Fumihiro Tsuchida, et al. (June 2022). “Pulse transit time-estimated blood pressure: a comparison of beat-to-beat and intermittent measurement”. en. In: *Hypertension Research* 45.6, pp. 1001–1007 (cit. on p. 88).

Hua, Jiangbo, Mengrui Su, Jing Wu, et al. (Dec. 2024). “Wearable cuffless blood pressure monitoring: From flexible electronics to machine learning”. In: *Wearable Electronics* 1, pp. 78–90 (cit. on p. 123).

Huang, Bin, Weihai Chen, Chun-Liang Lin, Chia-Feng Juang, and Jianhua Wang (Mar. 2022). “MLP-BP: A novel framework for cuffless blood pressure measurement with PPG and ECG signals based on MLP-Mixer neural networks”. In: *Biomedical Signal Processing and Control* 73, p. 103404 (cit. on pp. 60, 67).

Huang, Shih-Cheng, Anuj Pareek, Malte Jensen, et al. (Apr. 2023). “Self-supervised learning for medical image classification: a systematic review and implementation guidelines”. en. In: *npj Digital Medicine* 6.1, p. 74 (cit. on p. 33).

Joachim, Jona (July 2022). *physiocurve 2022.7.14 library on Zenodo. Link https://zenodo.org/record/7544687* (cit. on pp. 18, 48).

Joachim, Jona, Maxime Coutrot, Sandrine Millasseau, et al. (Feb. 2020). “Real-time estimation of mean arterial blood pressure based on photoplethysmography dicrotic notch and perfusion index. A pilot study”. eng. In: *Journal of Clinical Monitoring and Computing* (cit. on pp. 45, 59, 66).

Johnson, Alistair E. W., Tom J. Pollard, Lu Shen, et al. (May 2016). “MIMIC-III, a freely accessible critical care database”. en. In: *Scientific Data* 3.1, p. 160035 (cit. on pp. 19, 60).

Joung, Jingon, Chul-Woo Jung, Hyung-Chul Lee, et al. (May 2023). “Continuous cuffless blood pressure monitoring using photoplethysmography-based PPG2BP-net for high intrasubject blood pressure variations”. en. In: *Scientific Reports* 13.1, p. 8605 (cit. on pp. 59, 68, 72).

K, Rishi Vardhan, Vedanth S, Poojah G, et al. (Nov. 2021). *BP-Net: Efficient Deep Learning for Continuous Arterial Blood Pressure Estimation using Photoplethysmogram*. en (cit. on p. 68).

Kendall, Alex and Yarin Gal (Oct. 2017). *What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?* en (cit. on p. 124).

Khalaf Hamoud, Alaa, Ali Salah Hashim, and Wid Akeel Awadh (Dec. 2018). "CLINICAL DATA WAREHOUSE A REVIEW". en. In: *Iraqi Journal for Computers and Informatics* 44.2 (cit. on p. 18).

Kim, Sung-Hoon, Jun-Gol Song, Ji-Hyun Park, et al. (Jan. 2013). "Beat-to-Beat Tracking of Systolic Blood Pressure Using Noninvasive Pulse Transit Time During Anesthesia Induction in Hypertensive Patients". en-US. In: *Anesthesia & Analgesia* 116.1, p. 94 (cit. on p. 60).

Krones, Felix, Umar Marikkar, Guy Parsons, Adam Szmul, and Adam Mahdi (Feb. 2025). "Review of multimodal machine learning approaches in healthcare". In: *Information Fusion* 114, p. 102690 (cit. on p. 125).

Lakshminarayanan, Balaji, Alexander Pritzel, and Charles Blundell (Nov. 2017). *Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles*. en (cit. on p. 125).

Lecun, Y., L. Bottou, Y. Bengio, and P. Haffner (Nov. 1998). "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11, pp. 2278–2324 (cit. on p. 30).

Ledowski, Thomas (Aug. 2019). "Objective monitoring of nociception: a review of current commercial solutions". In: *British Journal of Anaesthesia* 123.2, e312–e321 (cit. on p. 92).

Lee, Hyung-Chul, Yoonsang Park, Soo Bin Yoon, et al. (June 2022). "VitalDB, a high-fidelity multi-parameter vital signs database in surgical patients". In: *Scientific Data* 9, p. 279 (cit. on pp. 20, 60, 69, 95).

Lee, Hyung-Chul, Ho-Geol Ryu, Yoonsang Park, et al. (Sept. 2019). "Data Driven Investigation of Bispectral Index Algorithm". en. In: *Scientific Reports* 9.1, p. 13769 (cit. on p. 103).

Lee, Solam, Hyung-Chul Lee, Yu Seong Chu, et al. (Apr. 2021). "Deep learning models for the prediction of intraoperative hypotension". en. In: *British Journal of Anaesthesia* 126.4, pp. 808–817 (cit. on pp. 45, 59).

Lehman, Li-wei H., Mohammed Saeed, Daniel Talmor, Roger Mark, and Atul Malhotra (Jan. 2013). "Methods of Blood Pressure Measurement in the ICU". In: *Critical care medicine* 41.1, pp. 34–40 (cit. on p. 10).

Li, Hongna, Yuhan Li, and Jingjie Zhou (Apr. 2024). "An Improved Neural Network Bispectral Index Monitoring Algorithm in Time and Frequency DomainAn Improved Bispectral Index Monitoring Algorithm". In: *Proceedings of the 2023 7th International Conference on Electronic Information Technol-*

ogy and Computer Engineering. EITCE '23. New York, NY, USA: Association for Computing Machinery, pp. 645–650 (cit. on p. 103).
Li, Songtao and Hao Tang (Nov. 2024). *Multimodal Alignment and Fusion: A Survey*. en (cit. on pp. 25, 125).
Li, Zhe, Zhongwen Rao, Lujia Pan, Pengyun Wang, and Zenglin Xu (Jan. 2023). *Ti-MAE: Self-Supervised Masked Time Series Autoencoders*. en (cit. on pp. 35, 106).
Lienhart, André, Yves Auroy, Françoise Péquignot, et al. (Dec. 2006). "Survey of anesthesia-related mortality in France". eng. In: *Anesthesiology* 105.6, pp. 1087–1097 (cit. on p. 2).
Liu, Qian, Junchen Ye, Haohan Liang, Leilei Sun, and Bowen Du (Feb. 2025). "TS-MAE: A masked autoencoder for time series representation learning". In: *Information Sciences* 690, p. 121576 (cit. on pp. 35, 106).
Liu, Yu, Min Li, Dong Yang, et al. (Apr. 2015). "Closed-Loop Control Better than Open-Loop Control of Profofol TCI Guided by BIS: A Randomized, Controlled, Multicenter Clinical Trial to Evaluate the CONCERT-CL Closed-Loop System". In: *PLoS ONE* 10.4, e0123862 (cit. on p. 26).
Lundberg, Scott and Su-In Lee (Nov. 2017). *A Unified Approach to Interpreting Model Predictions*. en (cit. on p. 36).
Maheshwari, Kamal, Tetsuya Shimada, Dongsheng Yang, et al. (Dec. 2020). "Hypotension Prediction Index for Prevention of Hypotension during Moderate- to High-risk Noncardiac Surgery: A Pilot Randomized Trial". en-US. In: *Anesthesiology* 133.6, p. 1214 (cit. on p. 24).
Maier-Hein, Lena, Annika Reinke, Patrick Godau, et al. (Feb. 2024). "Metrics reloaded: recommendations for image analysis validation". en. In: *Nature Methods* 21.2, pp. 195–212 (cit. on p. 124).
McInnes, Leland, John Healy, and James Melville (Sept. 2020). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. en (cit. on p. 99).
Mehari, Temesgen and Nils Strodthoff (Feb. 2022). "Self-supervised representation learning from 12-lead ECG data". en. In: *Computers in Biology and Medicine* 141, p. 105114 (cit. on p. 106).
Meidert, A. S., J. Briegel, and B. Saugel (Sept. 2019). "[Principles and pitfalls of arterial blood pressure measurement]". ger. In: *Der Anaesthesist* 68.9, pp. 637–650 (cit. on p. 10).
Mieloszyk, Rebecca, Hope Twede, Jonathan Lester, et al. (July 2022). "A Comparison of Wearable Tonometry, Photoplethysmography, and Electrocardiography for Cuffless Measurement of Blood Pressure in an Ambulatory Setting". en. In: *IEEE Journal of Biomedical and Health Informatics* 26.7, pp. 2864–2875 (cit. on p. 123).

Miyaguchi, Naoki, Koh Takeuchi, Hisashi Kashima, Mizuki Morita, and Hiroshi Morimatsu (Dec. 2021). “Predicting anesthetic infusion events using machine learning”. en. In: *Scientific Reports* 11.1, p. 23648 (cit. on p. 68).

Molenaar, Mitchel A, Berto J Bouma, Folkert W Asselbergs, et al. (Mar. 2024). “Explainable machine learning using echocardiography to improve risk prediction in patients with chronic coronary syndrome”. In: *European Heart Journal - Digital Health* 5.2, pp. 170–182 (cit. on p. 59).

Moreau, Thomas, Mathurin Massias, Alexandre Gramfort, et al. (Oct. 2022). *Benchopt: Reproducible, efficient and collaborative optimization benchmarks*. en (cit. on p. 122).

Mukkamala, Ramakrishna, Sanjeev G. Shroff, Cederick Landry, et al. (Mar. 2023). “The Microsoft Research Aurora Project: Important Findings on Cuffless Blood Pressure Measurement”. In: *Hypertension* 80.3, pp. 534–540 (cit. on p. 123).

Murray, I. P. (Apr. 1981). “Complications of invasive monitoring”. eng. In: *Medical Instrumentation* 15.2, pp. 85–89 (cit. on p. 10).

Myles, P. S., K. Leslie, J. McNeil, A. Forbes, and M. T. V. Chan (May 2004). “Bispectral index monitoring to prevent awareness during anaesthesia: the B-Aware randomised controlled trial”. English. In: *The Lancet* 363.9423, pp. 1757–1763 (cit. on p. 93).

Nafisi, Vahid Reza and Mina Shahabi (Apr. 2018). “Intradialytic hypotension related episodes identification based on the most effective features of photoplethysmography signal”. en. In: *Computer Methods and Programs in Biomedicine* 157, pp. 1–9 (cit. on p. 45).

Nair, Bala G., Mayumi Horibe, Shu-Fang Newman, et al. (Jan. 2014). “Anesthesia information management system-based near real-time decision support to manage intraoperative hypotension and hypertension”. eng. In: *Anesthesia and Analgesia* 118.1, pp. 206–214 (cit. on p. 61).

O’Shea, Keiron and Ryan Nash (Dec. 2015). *An Introduction to Convolutional Neural Networks* (cit. on p. 49).

Oord, Aaron van den, Sander Dieleman, Heiga Zen, et al. (Sept. 2016). *WaveNet: A Generative Model for Raw Audio*. en (cit. on pp. 74, 79).

Oord, Aaron van den, Yazhe Li, and Oriol Vinyals (Jan. 2019). *Representation Learning with Contrastive Predictive Coding*. en (cit. on p. 35).

Ou, Yi, Xiaoyi Hu, Cong Luo, and Yajun Li (Apr. 2025). “Global trends in artificial intelligence research in anesthesia from 2000 to 2023: a bibliometric analysis”. In: *Perioperative Medicine* 14.1, p. 47 (cit. on pp. xviii, 23).

Pandit, J. J., J. Andrade, D. G. Bogod, et al. (2014). “The 5th National Audit Project (NAP5) on accidental awareness during general anaesthesia:

summary of main findings and risk factors". en. In: *Anaesthesia* 69.10, pp. 1089–1101 (cit. on p. 92).
Pandit, J. J. and T. M. Cook (Mar. 2013). "I. National Institute for Clinical Excellence guidance on measuring depth of anaesthesia: limitations of EEG-based technology". In: *British Journal of Anaesthesia* 110.3, pp. 325–328 (cit. on p. 105).
Paolino, Nicola, Michele Schiavo, Nicola Latronico, Massimiliano Paltenghi, and Antonio Visioli (Mar. 2024). "PK/PD model based design of PID control for closed-loop anesthesia". In: *IFAC Journal of Systems and Control* 27, p. 100247 (cit. on p. 26).
Pecha, Simon, Paulus Kirchhof, and Bruno Reissmann (Aug. 2023). "Perioperative Arrhythmias". In: *Deutsches Ärzteblatt International* 120.33-34, pp. 564–574 (cit. on p. 23).
Perdereau, Jade, Thibaut Chamoux, Etienne Gayat, et al. (Jan. 2025). "Blood Pressure Estimation Using Explainable Deep-Learning Models Based on Photoplethysmography". en-US. In: *Anesthesia & Analgesia* 140.1, p. 119 (cit. on pp. 65, 72).
Pillai, Arvind, Dimitris Spathis, Fahim Kawsar, and Mohammad Malekzadeh (Feb. 2025). *PaPaGei: Open Foundation Models for Optical Physiological Signals*. en (cit. on pp. 94, 110, 111).
Radford, Alec, Jong Wook Kim, Chris Hallacy, et al. (Feb. 2021). *Learning Transferable Visual Models From Natural Language Supervision*. en (cit. on pp. 95, 124).
Raina, Rajat, Alexis Battle, Honglak Lee, Benjamin Packer, and Andrew Y. Ng (June 2007). "Self-taught learning: transfer learning from unlabeled data". en. In: *Proceedings of the 24th international conference on Machine learning*. Corvalis Oregon USA: ACM, pp. 759–766 (cit. on p. 33).
Roach, Joshua K. and Robert H. Thiele (June 2019). "Perioperative blood pressure monitoring". en. In: *Best Practice & Research Clinical Anaesthesiology*. Advances in hemodynamic monitoring in perioperative medicine 33.2, pp. 127–138 (cit. on p. 10).
Romano, Yaniv, Evan Patterson, and Emmanuel J. Candès (May 2019). *Conformalized Quantile Regression*. en (cit. on p. 124).
Salmasi, Vafi, Kamal Maheshwari, Dongsheng Yang, et al. (Jan. 2017). "Relationship between Intraoperative Hypotension, Defined by Either Reduction from Baseline or Absolute Thresholds, and Acute Kidney and Myocardial Injury after Noncardiac Surgery: A Retrospective Cohort Analysis". eng. In: *Anesthesiology* 126.1, pp. 47–65 (cit. on p. 11).

Sandin, Rolf H., Gunnar Enlund, Peter Samuelsson, and Claes Lennmarken (Feb. 2000). “Awareness during anaesthesia: a prospective case study”. English. In: *The Lancet* 355.9205, pp. 707–711 (cit. on p. 92).

Saugel, Bernd, Nick Fletcher, Tong J. Gan, et al. (Aug. 2024). “PeriOperative Quality Initiative (POQI) international consensus statement on perioperative arterial pressure management”. English. In: *British Journal of Anaesthesia* 133.2, pp. 264–276 (cit. on p. 66).

Savitzky, Abraham. and M. J. E. Golay (July 1964). “Smoothing and Differentiation of Data by Simplified Least Squares Procedures.” In: *Analytical Chemistry* 36.8, pp. 1627–1639 (cit. on p. 18).

Schutte, Aletta E (Nov. 2022). “The promise and pitfalls of novel cuffless blood pressure devices”. en. In: *European Heart Journal* 43.41, pp. 4222–4223 (cit. on p. 67).

Selvaraju, Ramprasaath R., Michael Cogswell, Abhishek Das, et al. (Oct. 2017). “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization”. In: *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626 (cit. on pp. 37, 46).

Sessler, Daniel I., Jeffrey C. Sigl, Scott D. Kelley, et al. (June 2012). “Hospital stay and mortality are increased in patients having a "triple low" of low blood pressure, low bispectral index, and low minimum alveolar concentration of volatile anesthesia”. eng. In: *Anesthesiology* 116.6, pp. 1195–1203 (cit. on p. 9).

Shander, Aryeh, Gregg P. Lobel, and Donald M. Mathews (Feb. 2018). “Brain Monitoring and the Depth of Anesthesia: Another Goldilocks Dilemma”. en-US. In: *Anesthesia & Analgesia* 126.2, p. 705 (cit. on p. 92).

Shen, Zhan, Tapabrata Chakraborti, Christopher R. S. Banerji, and Xiaorong Ding (July 2025). “Conformal prediction quantifies wearable cuffless blood pressure with certainty”. en. In: *Scientific Reports* 15.1, p. 26697 (cit. on p. 125).

Shen, Zhan, Lei Liu, and Xiaorong Ding (July 2022). “Bayesian Model Averaging for Improving the Accuracy of Cuffless Blood Pressure Estimation”. In: *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 3981–3984 (cit. on pp. 61, 124).

Shirmohamadi, Erfan, Reza Hosseini Dolama, Narjes Mohammadzadeh, Navid Ebrahimi, and Negar Ghasemloo (July 2025). “Intraoperative hypotension prediction in cardiac and noncardiac procedures: is HPI truly worthwhile? A systematic review and meta-analysis”. In: *BMC Anesthesiology* 25, p. 388 (cit. on p. 24).

Sladen, Arnold (Feb. 1988). “Complications of invasive hemodynamic monitoring in the intensive care unit”. en. In: *Current Problems in Surgery* 25.2, pp. 75–145 (cit. on p. 10).

Stergiou, George S., Ramakrishna Mukkamala, Alberto Avolio, et al. (Aug. 2022). “Cuffless blood pressure measuring devices: review and statement by the European Society of Hypertension Working Group on Blood Pressure Monitoring and Cardiovascular Variability”. en-US. In: *Journal of Hypertension* 40.8, p. 1449 (cit. on pp. 123, 124).

Thapa, Rahul, Bryan He, Magnus Ruud Kjaer, et al. (May 2024). *SleepFM: Multi-modal Representation Learning for Sleep Across Brain Activity, ECG and Respiratory Signals*. en (cit. on p. 95).

Thiele, Robert H., Edward C. Nemergut, and Carl III Lynch (Aug. 2011). “The Clinical Implications of Isolated Alpha1 Adrenergic Stimulation”. en-US. In: *Anesthesia & Analgesia* 113.2, p. 297 (cit. on p. 12).

Vallée, Fabrice, Jona Joachim, Maxime Coutrot, et al. (June 2019). “System for Measuring the Mean Arterial Pressure”. WO2019122406A1 (cit. on p. 12).

Vandecasteele, Kaat, Thomas De Cooman, Christos Chatzichristos, et al. (2021). “The power of ECG in multimodal patient-specific seizure monitoring: Added value to an EEG-based detector using limited channels”. en. In: *Epilepsia* 62.10, pp. 2333–2343 (cit. on p. 105).

Varoquaux, Gaël and Veronika Cheplygina (Apr. 2022). “Machine learning for medical imaging: methodological failures and recommendations for the future”. en. In: *npj Digital Medicine* 5.1, p. 48 (cit. on p. 124).

Vaswani, Ashish, Noam Shazeer, Niki Parmar, et al. (n.d.). “Attention is All you Need”. en. In: () (cit. on p. 32).

Vlachopoulos, Charalambos, Konstantinos Aznaouridis, and Christodoulos Stefanadis (Mar. 2010). “Prediction of cardiovascular events and all-cause mortality with arterial stiffness: a systematic review and meta-analysis”. eng. In: *Journal of the American College of Cardiology* 55.13, pp. 1318–1327 (cit. on p. 23).

Walsh, Michael, Philip J. Devereaux, Amit X. Garg, et al. (Sept. 2013). “Relationship between intraoperative mean arterial pressure and clinical outcomes after noncardiac surgery: toward an empirical definition of hypotension”. eng. In: *Anesthesiology* 119.3, pp. 507–515 (cit. on pp. 9, 11).

Wang, Weinan, Pedram Mohseni, Kevin L. Kilgore, and Laleh Najafizadeh (Feb. 2023). “PulseDB: A large, cleaned dataset based on MIMIC-III and VitalDB for benchmarking cuff-less blood pressure estimation methods”. English. In: *Frontiers in Digital Health* 4 (cit. on pp. 72, 79).

Wang, Xingyi, Yuliang Ma, Jared Cammon, et al. (2023). “Self-Supervised EEG Emotion Recognition Models Based on CNN”. eng. In: *IEEE transactions on neural systems and rehabilitation engineering: a publication of the IEEE Engineering in Medicine and Biology Society* 31, pp. 1952–1962 (cit. on p. 94).

Wang, Yuefu, Weiqin Huang, Jiange Han, et al. (June 2022). “A comparison of ClearSight noninvasive cardiac output and pulmonary artery bolus thermodilution cardiac output in cardiac surgery patients”. In: *Perioperative Medicine* 11, p. 24 (cit. on p. 24).

Weng, Weining, Yang Gu, Shuai Guo, et al. (Jan. 2024). *Self-supervised Learning for Electroencephalogram: A Systematic Survey*. en (cit. on p. 94).

Wesselink, E.M., T.H. Kappen, H.M. Torn, A.J.C. Slooter, and W.A. van Klei (Oct. 2018). “Intraoperative hypotension and the risk of postoperative adverse outcomes: a systematic review”. en. In: *British Journal of Anaesthesia* 121.4, pp. 706–721 (cit. on pp. 9, 65).

Williams, Bryan, Neil R Poulter, Morris J Brown, et al. (Mar. 2004). “British Hypertension Society guidelines for hypertension management 2004 (BHS-IV): summary”. In: *BMJ : British Medical Journal* 328.7440, pp. 634–640 (cit. on p. 46).

Wongsuphasawat, Kanit, Daniel Smilkov, James Wexler, et al. (Jan. 2018). “Visualizing Dataflow Graphs of Deep Learning Models in TensorFlow”. en. In: *IEEE Transactions on Visualization and Computer Graphics* 24.1, pp. 1–12 (cit. on p. 49).

Wu, Haixu, Jiehui Xu, Jianmin Wang, and Mingsheng Long (Jan. 2022). *Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting*. en (cit. on p. 32).

Xiao, Hanguang, Chang Liu, and Banglin Zhang (Apr. 2022). “Reconstruction of central arterial pressure waveform based on CNN-BILSTM”. In: *Biomedical Signal Processing and Control* 74, p. 103513 (cit. on p. 68).

Xiao, Tiantian, Ziwei Wang, Yongfeng Zhang, et al. (Jan. 2024). “Self-supervised Learning with Attention Mechanism for EEG-based seizure detection”. In: *Biomedical Signal Processing and Control* 87, p. 105464 (cit. on p. 94).

Yilmaz, Gizem, Xingyu Lyu, Ju Lynn Ong, et al. (Jan. 2023). “Nocturnal Blood Pressure Estimation from Sleep Plethysmography Using Machine Learning”. en. In: *Sensors* 23.18, p. 7931 (cit. on pp. 59, 67).

Zabihi, Soheil, Elahe Rahimian, Fatemeh Marefat, et al. (Sept. 2022). “BP-Net: Cuff-less and non-invasive blood pressure estimation via a generic deep convolutional architecture”. In: *Biomedical Signal Processing and Control* 78, p. 103850 (cit. on p. 68).

Zhang, Zexing, Huimin Lu, Songzhe Ma, et al. (Dec. 2024). “A general framework for generative self-supervised learning in non-invasive estimation of physiological parameters using photoplethysmography”. In: *Biomedical Signal Processing and Control* 98, p. 106788 (cit. on p. 94).

Zhou, Haoyi, Shanghang Zhang, Jieqi Peng, et al. (Mar. 2021). *Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting*. en (cit. on p. 32).

Zhou, Yang, Yuan He, Jianwei Wu, et al. (Jan. 2022). “A method of parameter estimation for cardiovascular hemodynamics based on deep learning and its application to personalize a reduced-order model”. eng. In: *International Journal for Numerical Methods in Biomedical Engineering* 38.1, e3533 (cit. on p. 45).