جامعة ابن طفيل
Ibn Tofaïl University
Faculté des Sciences

Ministère de l'Enseignement Supérieur, de la Recherche Scientifique et de l'Innovation
Université Ibn Tofail
Faculté des Sciences, Kénitra

Mémoire du projet de fin d'études
Présenté pour l'obtention du diplôme de
**Master Spécialisé Big Data et Cloud Computing**

---

# Apprentissage fédéré sous confidentialité différentielle pour la prédiction de la mortalité en réanimation : étude comparative et analyse de calibration sur MIMIC-IV

---

**Présenté par :** Achraf Birhrissen
**Soutenu le :** 29/12/2025, à Kénitra

**Jury :**

| | |
|---|---|
| Pr. Jaafar ABOUCHABAKA | FSK – Président |
| Pr. Idriss MOUMEN | Examinateur |
| Pr. Hatim DERROUZ | Examinateur |
| Pr. Najat RAFALIA | Encadrante |

Année universitaire : 2024/2025

# Dédicace

*À mes chers parents,*

*pour votre amour inconditionnel, vos sacrifices*
*et votre confiance qui m'ont porté jusqu'ici.*

*À mon frère,*

*pour ta présence, ton soutien*
*et notre complicité précieuse.*

*À mes deux sœurs,*

*pour votre tendresse, vos encouragements*
*et la force que vous m'inspirez.*

*À toute ma famille,*

*pour votre bienveillance et votre foi en moi.*

*À tous ceux qui m'ont soutenu et accompagné,*

*je vous dédie ce travail avec toute ma gratitude.*

# Remerciements

Je tiens tout d'abord à exprimer ma profonde gratitude à Dieu pour m'avoir accordé la force, la patience et la persévérance nécessaires pour mener à bien ce travail.

Je souhaite exprimer ma reconnaissance à ma famille pour son soutien constant, son encouragement et sa confiance, qui ont constitué un appui essentiel durant tout mon parcours.

J'adresse ensuite mes remerciements les plus sincères à mon encadrante, Pr. Najat Rafalia, pour son accompagnement, ses conseils scientifiques, sa disponibilité et la qualité de ses orientations tout au long de ce mémoire. Ses remarques pertinentes et son exigence méthodologique ont été déterminantes pour la structuration de ce travail et la consolidation de ses résultats.

Je remercie également les membres du jury qui ont accepté d'évaluer ce mémoire. Leurs observations et suggestions contribueront à enrichir ce travail et à améliorer sa portée.

Je remercie aussi mes amis, ainsi que mes collègues, pour leurs échanges, leur aide et l'esprit de collaboration qui ont facilité la réalisation de ce projet.

Enfin, je dédie ce travail à toutes les personnes qui m'ont soutenu de près ou de loin, avec une pensée particulière pour ma famille : leur présence a été ma plus grande motivation.

# Résumé

La prédiction précoce de la mortalité en réanimation à partir des DME peut soutenir l'aide à la décision clinique, mais se heurte à deux contraintes majeures : (i) la difficulté de mutualiser des données sensibles entre établissements et (ii) la nécessité de fournir des probabilités fiables, au-delà de la seule discrimination. Ce projet étudie un cadre d'DP-FL pour la prédiction de la mortalité en réanimation sur MIMIC-IV, en simulant un scénario multi-institutions via une partition par type d'ICU (MICU, SICU, CCU, CVICU, Neuro) à partir des 24 premières heures de séjour.

Nous comparons systématiquement quatre variantes DP-FL basées sur DP-SGD (DP-FedAvg, DP-FedProx, DP-FedBN, DP-Ditto), avec clipping des gradients et bruit gaussien. Les garanties de confidentialité sont quantifiées par une comptabilité de type RDP, pour $\delta = 10^{-5}$ et des budgets cibles $\varepsilon \in \{2, 4, 6, 8\}$. L'évaluation multi-graines (3 seeds) repose sur des métriques de discrimination (AUROC, AUPRC) et de calibration (ECE, score de Brier), complétées par une calibration post-hoc par *temperature scaling*, qui ne consomme pas de budget de confidentialité supplémentaire.

Les résultats mettent en évidence un compromis confidentialité–utilité favorable : l'AUROC moyen passe d'environ 0,826 ($\varepsilon = 2$) à 0,847 ($\varepsilon = 8$), et l'AUPRC de 0,359 à 0,417 lorsque la contrainte de confidentialité s'assouplit, sans différence statistiquement significative entre les quatre méthodes. La calibration post-hoc améliore nettement la fiabilité des probabilités (ECE réduite d'environ 0,10 à $\sim$0,05), renforçant la pertinence clinique des modèles DP-FL étudiés.

**Mots-clés :** FL ; DP ; DP-SGD ; RDP ; MIMIC-IV ; ICU ; Prédiction de mortalité ; Calibration.

# Abstract

Early prediction of Intensive Care Unit (ICU) mortality from Electronic Health Record (EHR) data can support clinical decision-making but faces two major constraints : (i) the difficulty of pooling sensitive patient data across institutions and (ii) the need to output reliable probabilities beyond discrimination alone. This thesis investigates a Differentially Private Federated Learning (DP-FL) framework for ICU mortality prediction on Medical Information Mart for Intensive Care IV (MIMIC-IV), simulating a multi-institution setting by partitioning data by ICU type (Medical Intensive Care Unit (MICU), Surgical Intensive Care Unit (SICU), Coronary Care Unit (CCU), Cardiovascular Intensive Care Unit (CVICU), Neurocritical Care Unit (Neuro)) using the first 24 hours of stay.

We compare four DP-FL methods based on Differentially Private Stochastic Gradient Descent (DP-SGD) : Differentially Private Federated Averaging (DP-FedAvg), Differentially Private Federated Proximal (DP-FedProx), Differentially Private Federated Batch Normalization (DP-FedBN), and Differentially Private Ditto (DP-Ditto). Privacy guarantees are quantified using Rényi Differential Privacy (RDP) accounting with $\delta = 10^{-5}$ and target budgets $\varepsilon \in \{2, 4, 6, 8\}$. Evaluation over three random seeds includes discrimination metrics (Area Under the Receiver Operating Characteristic Curve (AUROC), Area Under the Precision-Recall Curve (AUPRC)) and calibration metrics (Expected Calibration Error (ECE), Brier score), with post-hoc temperature scaling applied at no additional privacy cost.

Results show a favorable privacy–utility trade-off : AUROC increases from approximately 0.826 ($\varepsilon = 2$) to 0.847 ($\varepsilon = 8$), and AUPRC from 0.359 to 0.417, with no statistically significant differences between the four methods. Temperature scaling reduces ECE from roughly 0.10 to approximately 0.05, strengthening clinical interpretability of predicted risks.

**Keywords :** Federated Learning (FL) ; Differential Privacy (DP) ; DP-SGD ; RDP ; MIMIC-IV ; ICU ; Mortality Prediction ; Calibration.

# Table des matières

**Dédicace** **i**

**Remerciements** **ii**

**Résumé** **iii**

**Abstract** **iv**

**Liste des acronymes** **xi**

**1 Introduction générale** **1**

1.1 Contexte et motivation . . . . . 1
1.2 Problématique . . . . . 1
1.3 Cadre expérimental et accès aux données . . . . . 2
1.4 Objectifs et contributions . . . . . 3
1.5 Organisation du mémoire . . . . . 3

**2 État de l'art** **5**

2.1 Prédiction de la mortalité en réanimation à partir des DME . . . . . 5
2.1.1 Contexte clinique et scores classiques . . . . . 5
2.1.2 Tâches de prédiction précoce . . . . . 6
2.1.3 Particularités des données de réanimation . . . . . 7
2.1.4 Familles de modèles . . . . . 8
2.1.5 Protocoles d'évaluation en prédiction clinique . . . . . 10
2.2 Apprentissage fédéré en santé . . . . . 11
2.2.1 Motivation . . . . . 11
2.2.2 Principes de l'apprentissage fédéré . . . . . 12
2.2.3 Hétérogénéité non-iid . . . . . 13
2.2.4 Limites et besoin de personnalisation . . . . . 14
2.3 Risques de confidentialité dans l'apprentissage fédéré . . . . . 16
2.3.1 Fuite d'information via gradients et paramètres . . . . . 16
2.3.2 Attaques de reconstruction et d'inférence de présence . . . . . 17

2.3.3 Distinction entre absence de données brutes et garanties formelles . 19
2.4 Confidentialité différentielle pour l'apprentissage . . . . . 20
2.4.1 Définition de la confidentialité différentielle . . . . . 20
2.4.2 Descente de gradient stochastique différentiellement privée . . . . . 21
2.4.3 Spécificités du contexte médical . . . . . 22
2.5 Apprentissage fédéré sous confidentialité différentielle . . . . . 22
2.5.1 Principes généraux du DP-FL . . . . . 23
2.5.2 DP-FedAvg et DP-FedProx . . . . . 23
2.5.3 DP-FedBN . . . . . 24
2.5.4 DP-Ditto . . . . . 25
2.5.5 Discussion comparative . . . . . 25
2.6 Comptabilité de confidentialité par RDP . . . . . 26
2.6.1 Motivations de la RDP pour l'entraînement itératif . . . . . 26
2.6.2 Composition des itérations et calcul du budget . . . . . 27
2.6.3 Paramètres pratiques . . . . . 27
2.7 Calibration des probabilités en prédiction clinique . . . . . 28
2.7.1 Discrimination et calibration : définitions et enjeux . . . . . 28
2.7.2 Mesures de calibration . . . . . 29
2.7.3 Méthodes de calibration post-hoc . . . . . 29
2.7.4 Calibration et confidentialité différentielle . . . . . 30
2.8 Synthèse et formulation du problème scientifique . . . . . 31
2.8.1 Lacunes identifiées et verrou scientifique . . . . . 31
2.8.2 Positionnement du travail . . . . . 32
2.8.3 Lien avec les chapitres suivants . . . . . 32

**3 Préparation des données et établissement des baselines 33**
3.1 Base de données MIMIC-IV . . . . . 33
3.1.1 Description générale . . . . . 33
3.1.2 Définition de la cohorte et critères d'inclusion . . . . . 34
3.2 Pipeline de prétraitement . . . . . 36
3.2.1 Extraction et agrégation temporelle . . . . . 36
3.2.2 Gestion des manquants et normalisation . . . . . 37
3.2.3 Ingénierie des caractéristiques . . . . . 37
3.3 Modèles de base centralisés . . . . . 37
3.3.1 Panorama des modèles évalués . . . . . 37
3.3.2 Résultats et choix de l'architecture fédérée . . . . . 38
3.4 Simulation de l'environnement fédéré . . . . . 39
3.4.1 Analyse de l'hétérogénéité (non-IID) . . . . . 39
Synthèse . . . . . 39

**4 Implémentation et résultats expérimentaux** **41**
4.1 Environnement technique . . . 41
4.1.1 Infrastructure matérielle . . . 41
4.1.2 Stack logiciel . . . 41
4.1.3 Simulation de l'environnement fédéré . . . 42
4.2 Algorithmes d'apprentissage fédéré . . . 42
4.2.1 FedAvg : La référence . . . 42
4.2.2 FedProx : Gestion de la divergence . . . 42
4.2.3 FedBN : Normalisation locale . . . 43
4.2.4 Ditto : Personnalisation par régularisation . . . 43
4.2.5 Synthèse comparative . . . 43
4.3 Intégration de la confidentialité différentielle . . . 44
4.3.1 Mécanisme DP-SGD et Opacus . . . 44
4.3.2 Calibration du bruit et comptabilité RDP . . . 44
4.4 Protocole de calibration post-hoc . . . 45
4.4.1 Méthodologie . . . 45
4.4.2 Préservation de la confidentialité . . . 45
4.5 Configuration expérimentale . . . 45
4.6 Résultats et discussion . . . 46
4.6.1 Rappel des résultats centralisés . . . 46
4.6.2 Résultats de l'apprentissage fédéré sans DP . . . 46
4.6.3 Résultats de l'apprentissage fédéré sous DP . . . 48
4.6.4 Analyse de la calibration . . . 52
4.6.5 Tests statistiques . . . 53
4.6.6 Comparaison avec l'état de l'art . . . 54
4.6.7 Discussion et implications cliniques . . . 56
Synthèse . . . 59

**Conclusion générale** **60**
Synthèse des travaux . . . 60
Contributions principales . . . 60
Implications pratiques . . . 61
Limitations . . . 62
Perspectives . . . 62
Mot de fin . . . 63

# Table des figures

2.1 Schéma de la tâche de prédiction précoce de mortalité sur une fenêtre de 24 h. . . . . 7

3.1 Diagramme de flux de sélection de la cohorte à partir de MIMIC-IV. . . . . 35
3.2 Vue d'ensemble du pipeline de prétraitement : extraction, imputation, normalisation et ingénierie. . . . . 36
3.3 Caractérisation de l'hétérogénéité : (a) Volume de données par client, (b) Variabilité de la mortalité intra-ICU. . . . . 39

4.1 Compromis confidentialité-utilité : AUROC en fonction du budget $\varepsilon$ pour les quatre algorithmes FL. Les barres d'erreur représentent l'écart-type sur 3 seeds. . . . . 49
4.2 Compromis confidentialité-utilité : AUPRC en fonction du budget $\varepsilon$. . . . . 50
4.3 Heatmap des performances AUROC moyennes par algorithme et budget $\varepsilon$. 50

# Liste des tableaux


2.1 Caractéristiques des principaux scores de sévérité en réanimation. . . . . 6
2.2 Familles de modèles pour la prédiction de mortalité en ICU. . . . . 9
2.3 Comparaison des algorithmes DP-FL étudiés. . . . . 25

3.1 Caractéristiques de la cohorte finale (N = 68 322). . . . . 34
3.2 Impact de l'ingénierie des caractéristiques (Validation par ablation). . . . . 37
3.3 Performances comparatives sur l'ensemble de test. . . . . 38

4.1 Environnement logiciel pour les expériences DP-FL. . . . . 42
4.2 Comparaison des algorithmes FL implémentés. . . . . 44
4.3 Multiplicateurs de bruit $\sigma$ calculés pour les budgets cibles ($\delta = 10^{-5}$). . . . 44
4.4 Hyperparamètres globaux de l'apprentissage fédéré. . . . . 45
4.5 Récapitulatif des performances centralisées (ensemble de test, n = 13 665). 46
4.6 Performances des algorithmes FL sans DP (ensemble de test). . . . . 47
4.7 Performances de Ditto par client ICU (modèles personnalisés). . . . . 47
4.8 Comparaison de la calibration : centralisé vs fédéré (avant temperature scaling). . . . . 48
4.9 Performances DP-FL : moyenne ± écart-type sur 3 seeds (4 décimales). . . 49
4.10 Comparaison des algorithmes : AUROC moyen (4 décimales) par budget $\varepsilon$. 51
4.11 Coût de la confidentialité différentielle en termes d'AUROC. . . . . 51
4.12 ECE avant et après temperature scaling (extraits : $\varepsilon \in \{2, 8\}$, moyenne sur 3 seeds). . . . . 52
4.13 Score de Brier avant et après temperature scaling. . . . . 53
4.14 P-values des tests t de Welch pour les comparaisons pairwise d'AUROC (n=3 seeds par configuration). . . . . 54
4.15 Test t apparié pour l'efficacité du temperature scaling (ECE avant vs après) sur 48 configurations. . . . . 54
4.16 Comparaison indicative avec l'état de l'art : prédiction de mortalité en ICU. Les protocoles (définition du label, fenêtre d'observation, cohortes, variables, splits) diffèrent entre études ; les valeurs d'AUROC ne sont donc pas strictement comparables. . . . . 55

4.17 Comparaison avec l'état de l'art : apprentissage fédéré en santé. Les écarts sont rapportés à titre indicatif lorsque les baselines centralisées et les protocoles sont explicitement comparables dans l'étude. . . . . 55
4.18 Études DP-FL (sélection) et comparaison avec notre travail. . . . . 56
4.19 Résumé des contributions par rapport à l'état de l'art. . . . . 56

# Liste des acronymes

**APACHE** Acute Physiology and Chronic Health Evaluation
**AUPRC** Area Under the Precision-Recall Curve
**AUROC** Area Under the Receiver Operating Characteristic Curve
**BN** Batch Normalization
**CCU** Coronary Care Unit
**CVICU** Cardiovascular Intensive Care Unit
**DME** Dossier Médical Électronique
**DP** Differential Privacy
**DP-Ditto** Differentially Private Ditto
**DP-FedAvg** Differentially Private Federated Averaging
**DP-FedBN** Differentially Private Federated Batch Normalization
**DP-FedProx** Differentially Private Federated Proximal
**DP-FL** Differentially Private Federated Learning
**DP-SGD** Differentially Private Stochastic Gradient Descent
**ECE** Expected Calibration Error
**EHR** Electronic Health Record
**FedAvg** Federated Averaging
**FedBN** Federated Batch Normalization
**FedProx** Federated Proximal
**FL** Federated Learning
**FT-Transformer** Feature Tokenizer Transformer
**F1** F1-score
**GRU** Gated Recurrent Unit
**HIPAA** Health Insurance Portability and Accountability Act
**ICU** Intensive Care Unit
**LSTM** Long Short-Term Memory
**MICE** Multiple Imputation by Chained Equations
**MICU** Medical Intensive Care Unit

| | |
|---|---|
| **MIMIC-IV** | Medical Information Mart for Intensive Care IV |
| **MLP** | Multi-Layer Perceptron |
| **Neuro** | Neurocritical Care Unit |
| **non-IID** | non-Independent and Identically Distributed |
| **RDP** | Rényi Differential Privacy |
| **RF** | Random Forest |
| **RGPD** | Règlement Général sur la Protection des Données |
| **RNN** | Recurrent Neural Network |
| **SAINT** | Self-Attention and Intersample Attention Transformer |
| **SAPS** | Simplified Acute Physiology Score |
| **SGD** | Stochastic Gradient Descent |
| **SICU** | Surgical Intensive Care Unit |
| **SOFA** | Sequential Organ Failure Assessment |
| **TCN** | Temporal Convolutional Network |

# Chapitre 1

# Introduction générale

## 1.1 Contexte et motivation

Les unités de soins intensifs prennent en charge des patients présentant des défaillances aiguës, souvent multi-organes, avec des trajectoires cliniques qui peuvent basculer en quelques heures. La mortalité en réanimation reste un indicateur central de gravité et un enjeu à la fois médical et organisationnel. Estimer le risque de décès dès les premières heures d'un séjour permettrait d'améliorer la surveillance, de mieux cibler certaines interventions et d'accompagner la décision clinique. L'exploitation de données cliniques réelles soulève cependant des défis méthodologiques : hétérogénéité des mesures, valeurs manquantes, déséquilibre des classes, bruit. Au-delà de la capacité à discriminer les patients à risque, les systèmes d'aide à la décision doivent produire des probabilités interprétables, cohérentes avec les fréquences observées. Cette exigence de calibration est souvent négligée alors qu'elle conditionne l'usage pratique des modèles.

Les dossiers médicaux électroniques et les bases de données de soins intensifs offrent une opportunité pour développer des modèles d'apprentissage automatique capables d'extraire des signaux prédictifs à partir de variables hétérogènes : démographie, constantes vitales, analyses biologiques, interventions. La base MIMIC-IV regroupe des données hospitalières et de réanimation dé-identifiées, largement utilisées en recherche biomédicale. Ce mémoire se concentre sur une fenêtre d'observation limitée aux 24 premières heures de séjour, un choix motivé par la pertinence clinique d'une prédiction précoce et par la comparabilité avec les scores de sévérité classiques.

## 1.2 Problématique

Malgré leur intérêt scientifique, les données de santé demeurent hautement sensibles. Des contraintes réglementaires et éthiques limitent leur centralisation et leur partage entre établissements. Or, la robustesse des modèles prédictifs repose souvent sur la diversité des

populations, des pratiques de soins et des environnements hospitaliers. Une approche strictement centralisée, où les données seraient regroupées dans un même centre, reste fréquemment irréalisable.

L'apprentissage fédéré offre une alternative : il permet d'entraîner un modèle global en agrégeant des mises à jour locales calculées au sein de chaque établissement, sans déplacer les données brutes. L'absence de partage direct des données ne suffit toutefois pas à garantir la confidentialité. Des travaux ont montré que les gradients ou les paramètres échangés peuvent révéler des informations sensibles, notamment via des attaques d'inférence ou de reconstruction [1, 2]. Il devient alors nécessaire de compléter l'apprentissage fédéré par des mécanismes offrant des garanties formelles de protection, comme la confidentialité différentielle.

Toutefois, l'ajout de bruit pour la confidentialité tend à dégrader la performance (utilité) et la fiabilité des probabilités (calibration), particulièrement lorsque les données sont hétérogènes entre les hôpitaux.

**Formalisation du problème.** Afin de délimiter précisément le périmètre de ce mémoire, nous formalisons la problématique de recherche comme suit :

Soit un ensemble de $K$ hôpitaux disposant de jeux de données locaux hétérogènes $\{\mathcal{D}_k\}_{k=1}^{K}$. Étant donné un budget de confidentialité cible $(\varepsilon, \delta)$ et des exigences de performance en discrimination (AUROC, AUPRC) et en calibration (ECE), l'objectif est de trouver un algorithme d'entraînement qui :

1. Produit un modèle global $f_\theta$ (ou des modèles personnalisés $\{f_{\theta_k}\}_{k=1}^{K}$) maximisant l'exactitude prédictive ;
2. Satisfait la $(\varepsilon, \delta)$-confidentialité différentielle avec une comptabilité rigoureuse du budget consommé ;
3. Garantit des estimations de probabilité bien calibrées, adaptées à la décision clinique ;
4. Opère sans jamais centraliser les données brutes des patients.

Cette formulation impose un triple compromis entre confidentialité, utilité et hétérogénéité que nous explorerons tout au long de ce travail.

## 1.3 Cadre expérimental et accès aux données

Ce travail s'appuie sur la base MIMIC-IV, dont l'accès est contrôlé. L'utilisation de ces données nécessite une formation préalable et une validation par PhysioNet. Dans le cadre de ce mémoire, l'accès a été obtenu après la réussite du parcours de formation requis et l'obtention du certificat associé, attestant du respect des conditions d'usage et des bonnes pratiques en matière de manipulation de données cliniques dé-identifiées.

Pour étudier un scénario fédéré multi-institutions en l'absence de données multi-sites réelles, nous simulons un environnement distribué en partitionnant MIMIC-IV selon le type d'unité de réanimation : MICU, SICU, CCU, CVICU et Neuro. Cette partition induit une hétérogénéité des distributions (prévalence, profils de patients, pratiques), caractéristique des environnements fédérés non-IID.

## 1.4 Objectifs et contributions

L'objectif principal de ce mémoire est d'étudier et d'évaluer un cadre d'apprentissage fédéré sous confidentialité différentielle pour la prédiction de la mortalité en réanimation à partir des 24 premières heures de séjour. Ce travail vise à analyser l'impact de la confidentialité sur la performance prédictive, le compromis confidentialité–utilité lorsque le budget $\varepsilon$ varie, et l'effet de procédures de calibration post-hoc sur la fiabilité des probabilités produites.

Les contributions de ce mémoire sont les suivantes :

1. Mise en place d'un cadre expérimental de prédiction de mortalité en réanimation sur MIMIC-IV à partir d'une fenêtre d'observation de 24 heures, avec une simulation fédérée par type d'ICU (cinq clients).
2. Étude comparative de quatre méthodes DP-FL basées sur DP-SGD : DP-FedAvg, DP-FedProx, DP-FedBN et DP-Ditto.
3. Intégration d'une comptabilité de confidentialité de type RDP afin de quantifier la consommation de confidentialité $(\varepsilon, \delta)$ pour des budgets $\varepsilon \in \{2, 4, 6, 8\}$ et $\delta = 10^{-5}$.
4. Évaluation multi-graines (3 seeds) à l'aide de métriques de discrimination (AUROC, AUPRC) et de calibration (ECE, score de Brier), complétée par une calibration post-hoc par *temperature scaling*.
5. Analyse empirique du compromis confidentialité–utilité et discussion des implications pratiques.

## 1.5 Organisation du mémoire

Le reste de ce mémoire est organisé comme suit.

Le Chapitre 2 présente le contexte scientifique et l'état de l'art : prédiction de mortalité en réanimation, apprentissage fédéré, confidentialité différentielle, approches existantes en DP-FL et notions de calibration.

Le Chapitre 3 décrit le pipeline de données et la méthodologie de préparation sur MIMIC-IV, la définition de la cohorte, les variables utilisées, et les modèles de référence centralisés.

Le Chapitre 4 détaille l'implémentation des algorithmes fédérés et leur extension sous confidentialité différentielle, la configuration expérimentale et la comptabilité RDP, puis présente les résultats expérimentaux, l'analyse de calibration, la comparaison avec l'état de l'art, et une discussion critique.

Une conclusion générale synthétise les apports de ce travail et ouvre sur des perspectives.

# Chapitre 2

# État de l'art

## 2.1 Prédiction de la mortalité en réanimation à partir des DME

### 2.1.1 Contexte clinique et scores classiques

Les unités de soins intensifs accueillent des patients dont l'état clinique peut évoluer de manière rapide et imprévisible. L'estimation du pronostic, et notamment du risque de mortalité hospitalière, répond à plusieurs besoins : stratifier la gravité à l'admission, orienter l'allocation des ressources, comparer les performances entre établissements et ajuster le risque dans les études cliniques. Ces exigences ont conduit, dès les années 1980, au développement de scores de sévérité reposant sur un nombre limité de variables physiologiques, biologiques et cliniques.

**APACHE II.** Le score APACHE II (*Acute Physiology and Chronic Health Evaluation II*) figure parmi les plus anciens et les plus utilisés [3]. Il combine des mesures physiologiques de routine, l'âge du patient et des indicateurs de comorbidités chroniques pour produire une valeur comprise entre 0 et 71. Plus le score est élevé, plus le risque de décès hospitalier augmente. APACHE II sert avant tout à la stratification pronostique et à l'ajustement du risque lors de comparaisons inter-unités ; son usage comme outil décisionnel individuel reste limité.

**SAPS II et SAPS 3.** Les scores Simplified Acute Physiology Score (SAPS) (*Simplified Acute Physiology Score*) visent une estimation du risque de mortalité hospitalière à partir d'un nombre restreint de variables. SAPS II a été construit sur une large cohorte multicentrique internationale et permet de convertir directement le score en probabilité de décès [4]. SAPS 3 étend cette approche en intégrant des informations disponibles dans l'heure précédant et suivant l'admission en réanimation, ce qui le rend plus adapté à une évaluation très précoce [5].

**SOFA.** Le score SOFA (*Sequential Organ Failure Assessment*) poursuit un objectif différent : quantifier l'atteinte de six systèmes d'organes (respiratoire, cardiovasculaire, hépatique, neurologique, coagulation, rénal) au moyen de sous-scores gradués [6]. Sa dimension séquentielle permet de suivre l'évolution de la défaillance d'organes au cours du séjour, ce qui en fait un outil pronostique dynamique, particulièrement utilisé dans les populations septiques.

Le Tableau 2.1 résume les caractéristiques de ces trois scores.

TABLE 2.1 – Caractéristiques des principaux scores de sévérité en réanimation.

| **Score** | **Objectif** | **Fenêtre** | **Variables** |
|---|---|---|---|
| APACHE II | Pronostic global | 24 h | 12 phys. + âge + comorbidités |
| SAPS II/3 | Mortalité hospitalière | Adm. / 24 h | 12–17 variables |
| SOFA | Défaillance d'organes | Séquentiel | 6 systèmes d'organes |

**Limites des scores classiques.** Ces outils présentent plusieurs atouts : standardisation, interprétabilité et faible coût de déploiement. Leur capacité de représentation reste toutefois restreinte. Les pondérations, fixées lors de la construction du score, ne capturent pas les interactions complexes entre variables ni les trajectoires temporelles fines. Les performances peuvent aussi se dégrader lorsque les pratiques de soins, les protocoles ou la composition de la population s'écartent des cohortes de développement. Ces limitations justifient l'intérêt croissant pour des modèles d'apprentissage automatique capables d'exploiter la richesse des DMEs.

### 2.1.2 Tâches de prédiction précoce

La prédiction de la mortalité en réanimation se formule généralement comme un problème de classification binaire. L'objectif est d'estimer, à partir d'informations collectées dans les premières heures suivant l'admission, une probabilité de décès durant le séjour en ICU ou avant la sortie de l'hôpital. Dans ce mémoire, nous retenons la **mortalité en ICU** comme variable cible.

La fenêtre d'observation joue un rôle central. Limiter cette fenêtre aux 24 premières heures présente plusieurs avantages [7]. D'abord, cette période correspond à une phase de surveillance rapprochée où les mesures sont fréquentes et les bilans biologiques répétés. Ensuite, une prédiction à 24 h est compatible avec un usage opérationnel : elle intervient suffisamment tôt pour influencer la prise en charge sans attendre plusieurs jours. Les scores classiques (APACHE II, SAPS II) utilisent cette même fenêtre, ce qui facilite les comparaisons. La prédiction est effectuée à l'issue de la fenêtre et non de manière continue au fil du séjour.

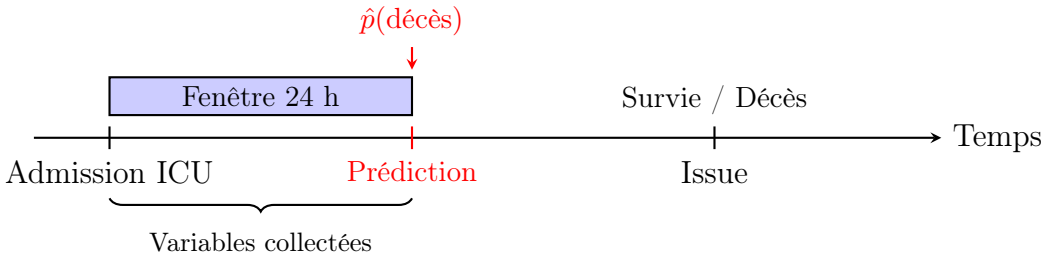


FIGURE 2.1 – Schéma de la tâche de prédiction précoce de mortalité sur une fenêtre de 24 h.

La Figure 2.1 illustre ce schéma de prédiction.

Les variables d'entrée sont hétérogènes : informations statiques (âge, sexe, comorbidités), séries temporelles de mesures (signes vitaux, bilans biologiques), et données sur les interventions (ventilation, perfusions) [8, 9]. Deux grandes familles d'approches existent pour exploiter ces données. La première consiste à résumer les séries temporelles en statistiques agrégées (dernière valeur, moyenne, minimum, maximum, variance), produisant un vecteur de caractéristiques fixe par patient. Cette stratégie convient aux modèles tabulaires et simplifie l'implémentation, mais elle peut perdre une partie de la dynamique intra-fenêtre [7]. La seconde approche traite la fenêtre comme une séquence multivariée, généralement après discrétisation horaire et imputation, permettant aux modèles récurrents ou attentionnels de capturer des dépendances temporelles.

Un défi transversal concerne la gestion des valeurs manquantes. En réanimation, l'absence d'une mesure n'est pas aléatoire : elle peut refléter un jugement clinique (patient stable), une contrainte de ressources ou une évolution rapide de l'état [10]. Certains modèles intègrent des masques d'observation et des intervalles depuis la dernière mesure afin de modéliser cette *missingness* informative.

Les tâches de mortalité sont souvent déséquilibrées, la prévalence du décès étant relativement faible. L'AUROC reste la métrique la plus courante, mais l'AUPRC apporte une information complémentaire lorsque la classe positive est rare [11]. Les protocoles d'évaluation imposent un découpage au niveau patient pour éviter les fuites d'information entre apprentissage et test [7].

### 2.1.3 Particularités des données de réanimation

Les données issues des DMEs en réanimation présentent des caractéristiques qui compliquent la modélisation et expliquent, en partie, les écarts entre performances obtenues sur une base de recherche et celles observées en déploiement réel [9].

**Irrégularité des mesures.** Les variables physiologiques et biologiques sont enregistrées à des fréquences très variables. Les signes vitaux peuvent être quasi continus, tandis que les bilans biologiques ou les gaz du sang sont prélevés à intervalles irréguliers. Cette hété-

rogénéité impose soit une étape d'alignement temporel, soit l'usage de modèles capables de traiter des séquences irrégulières [7].

**Valeurs manquantes.** Les données manquantes sont omniprésentes et leur structure n'est généralement pas aléatoire. L'absence d'un examen peut traduire un jugement clinique, une contrainte de ressources ou une évolution rapide de l'état du patient. Ignorer cette dimension risque d'introduire des biais. Des architectures comme GRU-D intègrent explicitement des masques et des intervalles depuis la dernière observation afin de modéliser ces motifs [10].

**Non-stationnarité.** On distingue trois formes de non-stationnarité : (i) intra-séjour, liée à l'évolution rapide de l'état du patient ; (ii) inter-patients, due à la variété des profils cliniques ; (iii) temporelle, reflétant les changements de pratiques et protocoles au fil du temps. Un modèle peut ainsi apprendre des associations dépendantes du contexte organisationnel plutôt que des relations physiopathologiques robustes [9].

**Censure et événements concurrents.** Certaines variables n'existent qu'après un acte (les paramètres de ventilation n'apparaissent qu'après intubation). Des issues intermédiaires comme un transfert ou une limitation thérapeutique modifient la trajectoire d'observation. Il en résulte un risque de confusion entre signal pronostique et marqueurs de prise en charge.

**Déséquilibre des classes.** La mortalité en ICU présente une prévalence relativement faible. L'AUROC peut rester élevée même lorsque les performances sur la classe positive sont insuffisantes. L'AUPRC et les métriques de calibration (ECE, score de Brier) apportent une lecture complémentaire [11, 12].

Ces particularités imposent des choix méthodologiques récurrents : définition stricte de la fenêtre et de la cible, traitement explicite des manquants, gestion du déséquilibre et protocoles d'évaluation rigoureux. Ces éléments seront mis en œuvre dans le cadre expérimental décrit au Chapitre 3.

### 2.1.4 Familles de modèles

La littérature sur la prédiction de mortalité en ICU fait appel à un large éventail de modèles, des méthodes statistiques classiques aux architectures profondes. Le choix dépend de la représentation des données, des contraintes de déploiement et des objectifs d'évaluation.

**Modèles linéaires.** La régression logistique, appliquée à un vecteur de caractéristiques agrégées, constitue une baseline interprétable et facile à calibrer. Sa capacité de représen-

tation reste limitée lorsque les relations entre variables sont non linéaires ou fortement interactives [9].

**Ensembles d'arbres.** Les forêts aléatoires combinent de nombreux arbres entraînés sur des sous-échantillons et des sous-ensembles de variables, réduisant la variance et améliorant la robustesse [13]. Les méthodes de boosting, notamment XGBoost [14] et LightGBM [15], offrent des performances compétitives sur données tabulaires. Ces approches exploitent toutefois moins naturellement la dynamique temporelle si l'information est fortement résumée [7].

**Réseaux profonds tabulaires.** Les réseaux feed-forward (MLP) permettent un apprentissage de représentations non linéaires avec un contrôle fin de la régularisation (dropout, weight decay) [9]. Des architectures comme TabNet introduisent des mécanismes d'attention pour sélectionner des sous-ensembles de variables à chaque étape de décision [16]. Leur complexité accrue peut cependant poser problème dans un cadre fédéré sous contraintes de confidentialité.

**Modèles temporels.** Lorsque les données sont représentées comme des séquences, les architectures récurrentes (LSTM, GRU) capturent les dépendances temporelles [17]. GRU-D intègre masques et intervalles depuis la dernière mesure [10]. Les convolutions temporelles (TCN) offrent une alternative non récurrente [18], tandis que les Transformers ont popularisé l'attention multi-têtes pour le traitement séquentiel [19]. Dans les DMEs, des modèles attentionnels comme RETAIN permettent de visualiser les contributions temporelles [20]. Des pipelines à grande échelle ont montré la faisabilité de modèles profonds pour la prédiction d'issues cliniques [21].

Le Tableau 2.2 récapitule ces familles.

TABLE 2.2 – Familles de modèles pour la prédiction de mortalité en ICU.

| **Famille** | **Entrée** | **Caractéristiques** |
|---|---|---|
| Régression logistique | Tabulaire | Interprétable, calibrable |
| Ensembles d'arbres (RF, XGBoost) | Tabulaire | Robuste, performant |
| MLP | Tabulaire | Flexible, compatible DP-SGD |
| RNN (LSTM, GRU-D) | Séquentielle | Capture la dynamique |
| Transformers | Séquentielle | Attention, parallélisable |

**Choix retenu.** Dans ce mémoire, l'objectif est une prédiction précoce sur 24 h dans un cadre fédéré sous confidentialité différentielle. Des représentations tabulaires agrégées, couplées à un MLP relativement compact, offrent un équilibre raisonnable entre expressivité, stabilité d'entraînement, coût de communication et intégration de mécanismes DP.

Les sections suivantes précisent les protocoles d'évaluation et, plus loin, la justification de ce choix au regard des contraintes DP-FL.

### 2.1.5 Protocoles d'évaluation en prédiction clinique

L'évaluation d'un modèle de prédiction clinique doit garantir l'absence de fuite d'information, une estimation fiable de l'incertitude et une lecture conjointe de la discrimination et de la calibration. Des recommandations méthodologiques soulignent l'importance d'une description précise des cohortes, des critères d'inclusion, des transformations des variables et des procédures de validation [22, 23].

**Découpage des données.** Dans les bases de réanimation, un même patient peut présenter plusieurs séjours. Un découpage au niveau patient (ou au niveau séjour, en empêchant qu'un même patient apparaisse dans plusieurs ensembles) est essentiel pour éviter que le modèle ne bénéficie d'informations corrélées [23]. Le schéma usuel est un découpage train/validation/test disjoint, la validation servant au réglage des hyperparamètres et le test étant réservé à l'estimation finale [22]. Un découpage temporel peut être envisagé lorsque la dérive représente une préoccupation.

**Reproductibilité.** La variabilité liée à l'initialisation et à l'optimisation peut être non négligeable. Répéter les expériences sur plusieurs graines aléatoires et rapporter des statistiques résumées (moyenne, écart-type, intervalles de confiance) évite des conclusions fondées sur une unique initialisation [23].

**Discrimination.** L'AUROC mesure la capacité à ordonner correctement les patients selon leur risque. Lorsque l'événement est rare, l'AUPRC apporte une information complémentaire sur le compromis précision–rappel [11]. Il est pertinent de rapporter les deux métriques et, si un seuil opérationnel est envisagé, d'y adjoindre sensibilité, spécificité et valeurs prédictives.

**Calibration.** La calibration évalue la concordance entre probabilités prédites et fréquences observées. Le score de Brier fournit une mesure quadratique globale [24], tandis que l'ECE approxime l'écart moyen entre confiance et exactitude après discrétisation des probabilités [12]. Les courbes de fiabilité offrent une visualisation sur l'intervalle $[0, 1]$. En santé, la calibration est cruciale car la probabilité prédite est interprétée comme un risque et peut influencer la décision clinique [22].

**Comparaison de modèles.** Des intervalles de confiance, généralement estimés par bootstrap, accompagnent les métriques. Pour comparer des AUROC sur les mêmes patients, le test de DeLong est fréquemment utilisé [25]. Lorsque plusieurs méthodes ou

budgets de confidentialité sont comparés, il convient de préciser le plan statistique : comparaison appariée, agrégation sur plusieurs graines et contrôle des comparaisons multiples.

**Synthèse.** Un protocole robuste en prédiction de mortalité sur DMEs combine un découpage disjoint au niveau patient, des répétitions multi-graines, des métriques de discrimination (AUROC, AUPRC) adaptées au déséquilibre, des métriques de calibration (Brier, ECE) et une quantification de l'incertitude. Ces principes guident l'évaluation expérimentale présentée au Chapitre 4.

## 2.2 Apprentissage fédéré en santé

### 2.2.1 Motivation

Les données de santé issues des DMEs sont particulièrement sensibles : elles décrivent des trajectoires cliniques individuelles, des diagnostics, des traitements et des événements à fort enjeu éthique. Dans la plupart des contextes, la centralisation de ces données à grande échelle est contrainte par des exigences réglementaires, des procédures de gouvernance et des risques opérationnels (fuites, mésusages, ré-identification). Ces contraintes limitent la constitution de jeux de données réellement multi-institutionnels, alors même que la robustesse des modèles prédictifs dépend fortement de la diversité des populations et des pratiques de soins [26].

Un second obstacle est la fragmentation inter-sites. Les hôpitaux et les services de soins intensifs diffèrent par la composition des patients (âge, comorbidités, sévérité), les pratiques cliniques (protocoles, seuils de prescription) et les systèmes d'information (codage, granularité, fréquence de mesure). Un modèle appris sur un seul site ou une seule unité peut donc mal se transférer vers d'autres environnements, en particulier lorsque des décalages de distribution sont présents [27]. En réanimation, ces écarts sont marqués : les unités spécialisées (MICU, SICU, CCU) correspondent à des profils cliniques et des prises en charge distincts, ce qui rend la généralisation non triviale.

Dans ce cadre, l'apprentissage fédéré (FL) est motivé par l'idée de collaboration sans mutualisation des données brutes. Plutôt que de transférer les DMEs vers un centre, chaque site entraîne localement un modèle (ou une mise à jour) et ne partage que des informations agrégées avec un serveur de coordination, ce qui réduit la circulation des données sensibles et facilite certaines formes de collaboration multi-institutionnelle [26, 28]. Des applications en santé ont montré l'intérêt pratique de ce paradigme dans des contextes où la centralisation est difficile, notamment pour la prédiction d'issues cliniques à partir de données hospitalières [29].

La motivation du FL ne se limite pas à la conformité : elle est aussi méthodologique. En agrégeant des signaux provenant de plusieurs distributions locales, le FL vise à améliorer

la robustesse et la capacité de généralisation par rapport à un modèle entraîné sur un site unique, tout en conservant la possibilité d'adapter le modèle aux contraintes locales. Cette perspective est particulièrement pertinente en soins intensifs, où l'hétérogénéité est structurelle et où l'objectif est d'obtenir des prédictions fiables dans des environnements non identiques.

### 2.2.2 Principes de l'apprentissage fédéré

L'apprentissage fédéré (FL) désigne une famille de méthodes d'apprentissage distribué dans lesquelles plusieurs entités (*clients*) entraînent conjointement un modèle, tout en conservant localement leurs données. Le schéma le plus courant est une architecture *client–serveur* : un serveur de coordination orchestre les rounds d'entraînement, tandis que chaque client effectue des calculs locaux sur son jeu de données et renvoie une mise à jour au serveur. L'algorithme FedAvg (*Federated Averaging*), proposé par McMahan et al. [30], constitue la référence fondatrice de ce paradigme [28, 26].

**Objectif d'optimisation fédéré.** Dans une formulation standard, chaque client $k \in \{1, \ldots, K\}$ possède un jeu de données local $D_k$ de taille $n_k$. L'objectif global consiste à minimiser une somme pondérée de risques empiriques locaux :

$$\min_{w \in \mathbb{R}^d} \; F(w) \;=\; \sum_{k=1}^{K} p_k \, F_k(w), \qquad p_k = \frac{n_k}{\sum_{j=1}^{K} n_j}, \tag{2.1}$$

où $w$ désigne les paramètres du modèle et $F_k(w)$ la perte moyenne sur $D_k$ [30, 28]. Cette pondération reflète une agrégation proportionnelle à la quantité de données de chaque client, souvent utilisée pour approximer l'optimisation sur l'union des données sans les centraliser.

**Boucle d'entraînement par rounds.** Un round d'apprentissage fédéré suit généralement les étapes suivantes [30, 28] :

1. **Diffusion :** le serveur envoie le modèle global courant $w^t$ (ou une partie des paramètres) à un ensemble de clients sélectionnés $S_t \subseteq \{1, \ldots, K\}$.
2. **Entraînement local :** chaque client $k \in S_t$ entraîne le modèle sur ses données (par exemple quelques époques de descente de gradient stochastique) et produit une mise à jour locale $\Delta w_k^t$ ou des paramètres mis à jour $w_k^{t+1}$.
3. **Agrégation :** le serveur agrège les mises à jour reçues afin d'obtenir le nouveau modèle global $w^{t+1}$.

Dans un cadre synchronisé (le plus fréquent), l'agrégation se fait après réception des mises à jour d'un sous-ensemble (ou de tous) les clients du round. Les choix de nombre d'époques

locales, de pas d'apprentissage, de taille de mini-lots et de fraction de clients participants contrôlent un compromis entre coût de communication, vitesse de convergence et stabilité.

**Principe d'agrégation.** L'agrégation la plus courante est une moyenne pondérée des paramètres (ou des incréments), pondérée par la taille des données locales. Si les clients renvoient des paramètres $w_k^{t+1}$, une forme standard est :

$$w^{t+1} \;=\; \sum_{k \in S_t} \tilde{p}_k \, w_k^{t+1}, \qquad \tilde{p}_k \;=\; \frac{n_k}{\sum_{j \in S_t} n_j}, \tag{2.2}$$

ce qui correspond à une approximation pratique de l'optimisation globale lorsque les clients effectuent des étapes locales de SGD [30]. Cette agrégation suppose implicitement un certain alignement entre les objectifs locaux et l'objectif global, hypothèse qui devient fragile en présence d'hétérogénéité non-IID (cf. Section 2.2.3).

**Participation partielle et contraintes systèmes.** Dans de nombreux scénarios, tous les clients ne participent pas à chaque round (participation partielle). Ce mécanisme est motivé par des contraintes de disponibilité et par la réduction du coût de communication, mais il introduit une variabilité supplémentaire dans l'optimisation [28]. Le FL se décline souvent en deux catégories : *cross-device* (très grand nombre de clients, participation partielle forte) et *cross-silo* (petit nombre de clients stables, typiquement des institutions), ce dernier étant le plus proche des scénarios hospitaliers [28, 26].

**Confidentialité et sécurité.** Le FL réduit la circulation de données brutes, mais ne garantit pas à lui seul une confidentialité formelle. D'un point de vue systèmes, des mécanismes cryptographiques tels que l'agrégation sécurisée peuvent empêcher le serveur d'observer les mises à jour individuelles en clair [31]. D'un point de vue statistique, des garanties formelles (par exemple la confidentialité différentielle) sont nécessaires pour borner l'information pouvant être inférée à partir des échanges. Ces aspects seront détaillés dans les Sections 2.3 et 2.4.

### 2.2.3 Hétérogénéité non-iid

Dans un cadre fédéré, l'hypothèse classique d'échantillons *i.i.d.* (indépendants et identiquement distribués) est rarement vérifiée. Les données locales peuvent différer en quantité, en distribution de variables et en distribution des labels. Cette hétérogénéité statistique est l'un des facteurs majeurs qui distingue l'apprentissage fédéré de l'optimisation distribuée traditionnelle [30, 28].

**Formes d'hétérogénéité.** On observe typiquement plusieurs sources de non-iid : (i) le décalage de covariables (*covariate shift*) lorsque les distributions des variables d'entrée

diffèrent entre clients (profils de patients, pratiques de mesure) ; (ii) le décalage de prévalence (*label shift*) lorsque le taux d'événement varie (mortalité plus élevée dans certaines unités) ; (iii) le décalage de concept lorsque la relation entrée–issue dépend du contexte clinique (protocoles, seuils d'intervention) ; (iv) l'hétérogénéité de quantité (tailles $n_k$ très différentes), qui influence le poids effectif des clients dans l'agrégation [28]. En santé, et en particulier en réanimation, ces phénomènes sont structurels : la spécialisation des unités et la diversité des prises en charge se traduisent par des distributions locales distinctes.

**Conséquences sur la convergence et la stabilité.** L'algorithme FedAvg, fondé sur des mises à jour locales suivies d'une moyenne au serveur, peut subir une dérive des clients (*client drift*) lorsque les objectifs locaux $F_k$ sont éloignés les uns des autres. Intuitivement, plusieurs étapes locales optimisées sur des distributions différentes peuvent produire des directions de mise à jour incompatibles, ce qui ralentit la convergence et peut dégrader la performance globale [30, 28]. Des observations empiriques sur des scénarios fortement non-iid montrent que la performance peut chuter lorsque chaque client voit une distribution de classes très différente de la distribution globale [32].

**Variations de pratique clinique et biais de transportabilité.** Au-delà des aspects purement statistiques, les différences de protocoles (fréquence des bilans, stratégies thérapeutiques, seuils de mise sous ventilation) peuvent induire des corrélations contextuelles : certaines variables deviennent des marqueurs de décision plutôt que des marqueurs physiologiques. Dans un environnement fédéré, ces biais peuvent être amplifiés si le modèle apprend principalement des signaux spécifiques à certains clients. Ceci motive des analyses de robustesse inter-clients et des méthodes visant à réduire la sensibilité à l'hétérogénéité.

**Pistes méthodologiques face au non-iid.** La littérature distingue plusieurs stratégies : (i) ajuster l'optimisation globale pour mieux tolérer l'hétérogénéité (par exemple via des régularisations proximales), (ii) personnaliser les modèles (composante globale + adaptation locale), et (iii) contrôler la participation et le nombre d'époques locales pour limiter la dérive [28]. Dans cette direction, FedProx introduit un terme proximal pénalisant l'écart au modèle global durant l'entraînement local, ce qui vise à stabiliser la convergence en contexte hétérogène [33]. Ces considérations sont directement pertinentes en santé, où la variabilité inter-sites (ou inter-unités) est attendue et où l'objectif est d'obtenir un modèle fiable sur l'ensemble des clients, tout en évitant de dégrader les performances pour certains sous-groupes.

### 2.2.4 Limites et besoin de personnalisation

L'apprentissage fédéré standard, tel qu'il est implémenté par FedAvg, repose sur l'hypothèse que les clients partagent un objectif commun et que l'agrégation simple de leurs

mises à jour produit un modèle globalement performant. Cette hypothèse devient fragile lorsque l'hétérogénéité inter-clients est importante, ce qui est fréquent dans les applications médicales.

**Biais d'agrégation.** Lorsque les distributions locales diffèrent fortement, la moyenne pondérée des paramètres peut converger vers un modèle qui ne correspond à l'optimum d'aucun client. Un client avec une faible représentation (données peu nombreuses ou distribution atypique) voit sa contribution diluée, et les performances locales sur ce client peuvent être nettement inférieures à celles obtenues par un modèle entraîné uniquement sur ses propres données [28]. Ce phénomène pose un problème d'équité : les sous-populations minoritaires risquent d'être défavorisées par le modèle global.

**Dérive de distribution et généralisation.** Même si le modèle fédéré atteint de bonnes performances moyennes, il peut mal se transférer vers de nouveaux clients dont la distribution s'écarte de celles observées durant l'entraînement. En santé, cette limitation est critique : un modèle développé sur un consortium d'hôpitaux peut échouer lors du déploiement dans un établissement aux pratiques ou à la patientèle différentes [26]. La robustesse à la dérive de distribution reste un défi ouvert.

**Besoin de personnalisation.** Pour pallier ces limites, plusieurs travaux proposent des approches de personnalisation : plutôt que d'imposer un modèle unique à tous les clients, on autorise une adaptation locale. Parmi les stratégies existantes, on trouve (i) le fine-tuning local après convergence du modèle global, (ii) l'ajout de termes de régularisation pénalisant l'écart au modèle global (FedProx), (iii) le maintien de paramètres locaux pour certaines couches (FedBN), et (iv) l'entraînement simultané de modèles globaux et personnalisés (Ditto) [33, 34, 35]. Ces méthodes seront détaillées dans le contexte de la confidentialité différentielle à la Section 2.5.

**Tension entre collaboration et spécificité.** La personnalisation introduit une tension : plus le modèle s'adapte aux données locales, moins il bénéficie de la collaboration fédérée. Un client disposant de peu de données peut préférer s'appuyer sur le modèle global, tandis qu'un client avec une distribution très particulière peut tirer profit d'une forte personnalisation. Trouver le bon équilibre dépend du contexte applicatif et du degré d'hétérogénéité [28].

L'apprentissage fédéré standard présente donc des limites face à l'hétérogénéité non-iid, et la personnalisation apparaît comme une réponse méthodologique pertinente. La section suivante examine un autre aspect critique : les risques de confidentialité qui persistent malgré l'absence de partage de données brutes.

## 2.3 Risques de confidentialité dans l'apprentissage fédéré

L'apprentissage fédéré repose sur l'idée que les données brutes ne quittent pas les sites locaux. Cette propriété réduit certains risques liés au transfert et au stockage centralisé, mais elle ne constitue pas, à elle seule, une garantie de confidentialité. Les mises à jour échangées entre clients et serveur (gradients, paramètres) encodent une information dérivée des données d'entraînement, et des travaux ont montré que cette information peut être exploitée pour inférer des propriétés sensibles, voire reconstruire des exemples individuels. Cette section examine les principaux vecteurs de fuite et distingue l'absence de données brutes d'une protection formelle de la vie privée.

### 2.3.1 Fuite d'information via gradients et paramètres

Dans un schéma fédéré classique, chaque client calcule des gradients (ou des mises à jour de paramètres) sur ses données locales et les transmet au serveur. Ces gradients agrègent les contributions de tous les exemples du mini-lot, ce qui pourrait sembler protéger les informations individuelles par effet de moyennage. En pratique, cette intuition est trompeuse : les gradients conservent une signature des données sous-jacentes, et plusieurs travaux ont démontré qu'un adversaire ayant accès aux gradients peut en extraire des informations sensibles.

**Attaques par inversion de gradient.** Zhu et al. [1] ont proposé une attaque dite *Deep Leakage from Gradients* (DLG). L'idée consiste à reconstruire les données d'entrée $(x, y)$ en résolvant un problème d'optimisation : partant d'une initialisation aléatoire, l'attaquant cherche des données fictives dont le gradient, calculé sur le même modèle, soit aussi proche que possible du gradient observé. Plus formellement, l'attaquant résout :

$$\min_{x',y'} \left\| \nabla_w \ell(f_w(x'), y') - g \right\|^2, \tag{2.3}$$

où $g$ désigne le gradient intercepté et $\ell$ la fonction de perte. Des raffinements ultérieurs ont amélioré la qualité de la reconstruction, notamment en exploitant des régularisations sur les images ou en utilisant des techniques de second ordre [36]. Ces attaques sont particulièrement efficaces lorsque le mini-lot est de petite taille ; avec un seul exemple, la reconstruction peut être quasi parfaite pour des images de taille modérée.

**Conditions favorisant la fuite.** Plusieurs facteurs accroissent le risque de reconstruction : (i) un mini-lot de petite taille, qui limite l'effet de moyennage ; (ii) un modèle surparamétré, dont les gradients encodent davantage d'information ; (iii) un nombre limité de classes, facilitant la recherche du label ; (iv) des données de faible dimension ou fortement

structurées. En contexte médical, les données tabulaires présentent une dimensionnalité souvent plus faible que les images, mais contiennent des attributs discrets (diagnostics, codes) dont l'inférence peut être facilitée par la structure des gradients.

**Portée en apprentissage fédéré.** Dans un scénario *cross-silo* avec peu de clients, le serveur observe directement les mises à jour de chaque participant. Un serveur malveillant, ou un adversaire ayant compromis le serveur, peut tenter d'inverser les gradients reçus. Même dans un scénario honnête mais curieux, le simple stockage des mises à jour crée un risque de fuite différée si ces données sont ultérieurement compromises. L'agrégation sécurisée (*secure aggregation*) [31] permet de masquer les contributions individuelles en ne révélant au serveur que la somme des mises à jour, mais elle n'empêche pas les attaques au niveau du modèle agrégé ni les fuites côté client.

**Fuite via les paramètres du modèle.** Au-delà des gradients, le modèle lui-même, une fois entraîné, peut révéler des informations sur les données d'entraînement. Des attaques d'inférence de propriété (*property inference*) permettent de déduire des caractéristiques statistiques du jeu de données (par exemple, la proportion d'un groupe démographique), tandis que des attaques d'inférence de présence (*membership inference*) déterminent si un individu précis faisait partie de l'entraînement [2]. Ces attaques exploitent la différence de comportement du modèle entre exemples vus et non vus, et s'appliquent indépendamment du caractère fédéré ou centralisé de l'entraînement.

Les échanges de gradients ou de paramètres ne sont donc pas anodins : ils transportent une information exploitable par un adversaire disposant de ressources suffisantes. L'apprentissage fédéré réduit l'exposition directe aux données brutes, mais ne protège pas contre ces attaques d'inférence. Une protection renforcée nécessite des mécanismes complémentaires, tels que l'agrégation sécurisée ou la confidentialité différentielle.

### 2.3.2 Attaques de reconstruction et d'inférence de présence

Les risques de fuite évoqués précédemment se concrétisent à travers plusieurs familles d'attaques, dont les objectifs et les hypothèses diffèrent. On distingue principalement les attaques de reconstruction, qui visent à retrouver les données d'entraînement, et les attaques d'inférence de présence (*membership inference*), qui cherchent à déterminer si un individu a contribué à l'apprentissage.

**Attaques de reconstruction.** L'objectif est de reconstruire, partiellement ou totalement, des exemples du jeu d'entraînement à partir d'informations accessibles à l'adversaire. L'attaque DLG [1] et ses variantes [36] relèvent de cette catégorie lorsqu'elles ciblent les gradients. D'autres approches exploitent le modèle final : un attaquant peut entraîner

un réseau génératif conditionné à produire des échantillons cohérents avec les représentations internes du modèle cible [37]. En imagerie médicale, des travaux ont montré la possibilité de reconstruire des images de patients à partir de modèles de classification entraînés sur des données privées, soulevant des préoccupations éthiques majeures.

**Attaques d'inférence de présence (*membership inference*).** Shokri et al. [38] ont introduit un cadre d'attaque où l'adversaire entraîne un classificateur binaire pour distinguer les exemples utilisés lors de l'entraînement de ceux qui ne l'ont pas été. L'intuition repose sur le fait qu'un modèle se comporte différemment sur ses données d'entraînement (généralement avec une confiance plus élevée ou une perte plus faible) que sur des données inédites. Cette attaque ne nécessite qu'un accès en boîte noire au modèle (prédictions) et peut atteindre des taux de réussite significatifs sur des modèles surappris.

Nasr et al. [2] ont étendu cette analyse au contexte fédéré et ont montré que les attaques d'inférence de présence restent efficaces, voire plus faciles, lorsque l'adversaire observe les mises à jour intermédiaires. Dans un scénario où le serveur est honnête mais curieux, l'accès aux gradients de chaque round fournit un signal plus riche que le modèle final seul.

**Facteurs de vulnérabilité.** Plusieurs caractéristiques augmentent la vulnérabilité aux attaques d'inférence : (i) le surapprentissage, qui accentue la différence de comportement entre données vues et non vues ; (ii) la complexité du modèle, qui offre davantage de signaux exploitables ; (iii) le déséquilibre des classes, qui peut faciliter l'identification de membres de la classe minoritaire ; (iv) la présence d'attributs rares ou identifiants dans les données. En santé, les dossiers médicaux contiennent souvent des combinaisons d'attributs quasi uniques (âge, diagnostics, dates), ce qui accroît le risque d'identification même sans reconstruction complète.

**Implications pour l'apprentissage fédéré.** L'architecture fédérée n'élimine pas ces menaces. Un serveur malveillant peut mener des attaques de reconstruction sur les gradients reçus ; un adversaire externe disposant du modèle final peut tenter des inférences de présence. Les mécanismes de défense se répartissent en deux catégories : (i) techniques cryptographiques (agrégation sécurisée, chiffrement homomorphe), qui protègent les communications mais pas le modèle final ; (ii) techniques statistiques (confidentialité différentielle), qui limitent l'information que le modèle peut révéler sur tout individu. La section suivante précise la distinction entre l'absence de données brutes et les garanties formelles de confidentialité.

### 2.3.3 Distinction entre absence de données brutes et garanties formelles

L'apprentissage fédéré est parfois présenté comme une solution de protection de la vie privée au motif que les données brutes ne quittent pas les sites locaux. Cette affirmation, bien que techniquement exacte, peut induire en erreur quant au niveau de protection effectivement offert. Il convient de distinguer plusieurs notions de confidentialité et de situer l'apprentissage fédéré dans ce cadre.

**Confidentialité opérationnelle vs formelle.** L'absence de transfert de données brutes réduit certains risques opérationnels : vol lors du transit, stockage dans un centre non sécurisé, accès non autorisé à une base centralisée. On parle parfois de *data minimization*, un principe recommandé par plusieurs cadres réglementaires. Cette forme de protection dépend cependant de l'implémentation et ne fournit pas de borne quantifiable sur l'information pouvant être inférée par un adversaire. À l'opposé, une garantie formelle, telle que la confidentialité différentielle, offre une définition mathématique et un paramètre ($\varepsilon$) mesurant le niveau de protection indépendamment des capacités de l'attaquant [39].

**Limites de l'approche fédérée seule.** Comme l'ont montré les travaux sur l'inversion de gradients et l'inférence de présence, les mises à jour échangées en FL transportent une information exploitable. Un serveur honnête mais curieux, un participant malveillant ou un adversaire externe disposant du modèle final peuvent mener des attaques d'inférence. L'agrégation sécurisée [31] empêche le serveur de voir les contributions individuelles en clair, mais elle n'offre pas de protection contre les attaques ciblant le modèle agrégé ni contre les inférences de présence. De plus, elle ne couvre pas les scénarios où un client lui-même est compromis ou coopère avec l'adversaire.

**Nécessité d'une protection formelle.** Pour obtenir des garanties quantifiables et indépendantes du modèle d'attaque, il est nécessaire d'introduire un mécanisme de bruit ou de perturbation dont les propriétés mathématiques bornent la fuite d'information. La confidentialité différentielle remplit ce rôle : en ajoutant du bruit calibré lors de l'entraînement, elle assure que la contribution d'un individu ne modifie la distribution des sorties que d'un facteur contrôlé par $\varepsilon$. Cette garantie vaut pour tout adversaire, quelle que soit sa puissance de calcul ou sa connaissance auxiliaire [39].

**Complémentarité des approches.** En pratique, une stratégie de protection peut combiner plusieurs couches : (i) l'apprentissage fédéré pour éviter le transfert de données brutes ; (ii) l'agrégation sécurisée pour masquer les contributions individuelles au serveur ; (iii) la confidentialité différentielle pour borner l'information pouvant être inférée du modèle ou des mises à jour. Ces mécanismes ne sont pas exclusifs et peuvent être

déployés conjointement. La section suivante introduit les fondements de la confidentialité différentielle et son application à l'entraînement de modèles.

## 2.4 Confidentialité différentielle pour l'apprentissage

La confidentialité différentielle (DP) fournit un cadre mathématique pour quantifier et limiter la fuite d'information lors de l'analyse de données sensibles. Introduite par Dwork et al. [40], cette notion est devenue une référence pour la protection de la vie privée dans les systèmes d'apprentissage automatique. Cette section présente les définitions fondamentales, le mécanisme DP-SGD pour l'entraînement de réseaux de neurones, et les spécificités du contexte médical.

### 2.4.1 Définition de la confidentialité différentielle

La confidentialité différentielle vise à garantir que la sortie d'un algorithme ne révèle pas trop d'information sur un individu particulier du jeu de données. L'idée centrale est de borner l'influence qu'un seul enregistrement peut avoir sur le résultat.

**Définition $(\varepsilon, \delta)$-DP.** Un mécanisme aléatoire $\mathcal{M}$ satisfait la $(\varepsilon, \delta)$-confidentialité différentielle si, pour tous jeux de données adjacents $D$ et $D'$ (différant d'au plus un enregistrement) et pour tout ensemble mesurable $S$ de sorties possibles :

$$\Pr[\mathcal{M}(D) \in S] \leq e^{\varepsilon} \cdot \Pr[\mathcal{M}(D') \in S] + \delta. \tag{2.4}$$

Le paramètre $\varepsilon$ (budget de confidentialité) contrôle le niveau de protection : plus $\varepsilon$ est petit, plus la garantie est forte. Le paramètre $\delta$ représente une probabilité d'échec de la garantie ; dans les applications pratiques, on choisit généralement $\delta \ll 1/n$, où $n$ est la taille du jeu de données [39].

**Interprétation.** La définition assure qu'un adversaire, même disposant d'une connaissance auxiliaire illimitée, ne peut distinguer de manière fiable si un individu donné était présent ou absent dans le jeu de données. Cette propriété est indépendante des capacités de calcul de l'attaquant et reste valide face à des attaques futures non encore connues.

**Propriétés clés.** La confidentialité différentielle possède plusieurs propriétés utiles : (i) la composition, qui permet de combiner plusieurs mécanismes DP en contrôlant le budget total ; (ii) l'immunité au post-traitement, qui garantit que toute transformation d'une sortie DP reste DP sans coût additionnel ; (iii) la robustesse aux connaissances auxiliaires, qui distingue la DP des approches basées sur l'anonymisation [39].

### 2.4.2 Descente de gradient stochastique différentiellement privée

L'application de la confidentialité différentielle à l'entraînement de réseaux de neurones passe par le mécanisme DP-SGD, introduit par Abadi et al. [41]. Ce mécanisme modifie la descente de gradient stochastique standard pour garantir une borne sur la fuite d'information à chaque itération.

**Clipping des gradients.** La première étape consiste à borner la contribution de chaque exemple au gradient. Pour un exemple $i$ du mini-lot, le gradient individuel $g_i = \nabla_w \ell(w; x_i, y_i)$ est clippé :

$$\bar{g}_i = g_i \cdot \min\left(1, \frac{C}{\|g_i\|_2}\right), \tag{2.5}$$

où $C$ est la norme maximale autorisée. Cette opération garantit que $\|\bar{g}_i\|_2 \leq C$, bornant ainsi la sensibilité du gradient agrégé à la présence ou l'absence d'un exemple.

**Ajout de bruit gaussien.** Après clipping, du bruit gaussien calibré est ajouté au gradient moyen :

$$\tilde{g} = \frac{1}{B}\left(\sum_{i=1}^{B} \bar{g}_i + \mathcal{N}(0, \sigma^2 C^2 I)\right), \tag{2.6}$$

où $B$ est la taille du mini-lot et $\sigma$ est le multiplicateur de bruit. Le paramètre $\sigma$ détermine le niveau de confidentialité : plus $\sigma$ est grand, plus le bruit est important et plus la garantie est forte.

**Impact sur l'optimisation.** Le clipping et le bruit introduisent un biais et une variance supplémentaires dans l'estimation du gradient. Le clipping peut tronquer les gradients informatifs, particulièrement en début d'entraînement ou pour des exemples atypiques. Le bruit dégrade le rapport signal/bruit, ralentissant la convergence et pouvant limiter la performance finale. Ces effets sont d'autant plus marqués que le budget $\varepsilon$ est faible (confidentialité forte) [41].

**Choix des hyperparamètres.** La configuration de DP-SGD implique plusieurs compromis : (i) la norme de clipping $C$ doit être suffisamment grande pour préserver le signal mais suffisamment petite pour limiter la sensibilité ; (ii) le multiplicateur de bruit $\sigma$ est déterminé par le budget $\varepsilon$ cible, le nombre d'itérations et le taux d'échantillonnage ; (iii) la taille du mini-lot $B$ influence le rapport signal/bruit (un $B$ plus grand améliore ce rapport). Des travaux récents proposent des stratégies de clipping adaptatif pour simplifier le réglage [42].

### 2.4.3 Spécificités du contexte médical

L'application de la confidentialité différentielle aux données de santé soulève des considérations particulières, tant sur le plan du compromis utilité–confidentialité que sur le choix des paramètres.

**Compromis utilité–confidentialité.** En prédiction clinique, la performance du modèle a des implications directes sur la qualité des soins. Une dégradation excessive de l'AUROC ou de la calibration peut rendre le modèle inutilisable en pratique. Le choix du budget $\varepsilon$ doit donc tenir compte des exigences de performance minimales pour l'application visée. Des études empiriques suggèrent que des budgets $\varepsilon \in [1, 10]$ permettent souvent de conserver une utilité raisonnable sur des tâches de classification [41].

**Ordres de grandeur de $\varepsilon$.** Il n'existe pas de consensus sur la valeur de $\varepsilon$ garantissant une protection « suffisante ». Des valeurs $\varepsilon < 1$ offrent une garantie forte mais peuvent dégrader significativement les performances. Des valeurs $\varepsilon \in [2, 8]$ représentent un compromis courant dans la littérature appliquée. Le paramètre $\delta$ est généralement fixé à une valeur inférieure à $1/n$ (par exemple $10^{-5}$ ou $10^{-6}$) pour garantir une probabilité d'échec négligeable [39].

**Sensibilité des attributs médicaux.** Les données de réanimation contiennent des attributs particulièrement sensibles : diagnostics, traitements, issues. Certains attributs peuvent être quasi-identifiants (combinaisons rares de caractéristiques). La confidentialité différentielle protège contre l'inférence sur tout attribut, y compris ceux non explicitement considérés, ce qui la rend particulièrement adaptée aux données médicales où les risques de ré-identification sont élevés.

**Exigences réglementaires.** Les cadres réglementaires (RGPD en Europe, HIPAA aux États-Unis) encouragent la minimisation des données et la protection par conception. Bien que la confidentialité différentielle ne soit pas explicitement mentionnée dans ces textes, elle constitue une approche technique compatible avec leurs principes et offre une base formelle pour démontrer la protection des données personnelles.

## 2.5 Apprentissage fédéré sous confidentialité différentielle

La combinaison de l'apprentissage fédéré et de la confidentialité différentielle (DP-FL) vise à obtenir les avantages des deux approches : collaboration sans centralisation des

données et garanties formelles de protection. Cette section présente les principes généraux du DP-FL et détaille les quatre algorithmes étudiés dans ce mémoire.

### 2.5.1 Principes généraux du DP-FL

Dans un cadre DP-FL, le bruit nécessaire à la confidentialité différentielle peut être introduit à différents niveaux de l'architecture fédérée. On distingue plusieurs configurations selon le point d'injection du bruit et le niveau de granularité de la protection.

**DP locale vs centrale.** En DP locale, chaque client ajoute du bruit à ses mises à jour avant de les transmettre au serveur. Le serveur ne reçoit que des données bruitées et ne peut inférer d'information précise sur les données locales. En DP centrale (ou globale), le bruit est ajouté après agrégation au niveau du serveur, ce qui suppose que le serveur est de confiance. La DP locale offre une protection plus forte mais au prix d'un bruit total plus élevé, dégradant davantage l'utilité [43].

**Protection au niveau enregistrement vs client.** On distingue généralement (i) la protection au niveau *enregistrement* (*record-level*), qui vise à protéger la contribution d'un individu (ici, un patient), et (ii) la protection au niveau *client* (*client-level*), qui vise à protéger la contribution d'un client entier (ici, une unité/hôpital). Dans ce mémoire, l'objectif est une protection *record-level*, obtenue en appliquant DP-SGD lors de l'entraînement local (clipping par exemple et bruit gaussien), ce qui borne l'influence d'un patient sur les mises à jour locales [41, 39]. À titre de contraste, des approches *client-level* ont également été étudiées dans la littérature fédérée [44].

**Intégration avec DP-SGD.** L'approche la plus courante consiste à remplacer la descente de gradient standard par DP-SGD lors de l'entraînement local. Chaque client effectue ses époques locales avec clipping et bruit, puis transmet les paramètres mis à jour au serveur. L'agrégation procède ensuite comme dans le FL standard. Cette configuration assure que chaque mise à jour locale satisfait une garantie DP, et la composition sur les rounds permet de calculer le budget total consommé.

### 2.5.2 DP-FedAvg et DP-FedProx

**DP-FedAvg.** L'algorithme DP-FedAvg applique le protocole FedAvg standard avec DP-SGD au niveau local. À chaque round, les clients sélectionnés reçoivent le modèle global, effectuent $E$ époques d'entraînement avec DP-SGD, et renvoient leurs paramètres au serveur pour agrégation par moyenne pondérée. La garantie DP résulte de la composition des itérations locales sur l'ensemble des rounds [30, 43].

**DP-FedProx.** L'algorithme DP-FedProx étend DP-FedAvg en ajoutant un terme de régularisation proximale à l'objectif local :

$$\min_w F_k(w) + \frac{\mu}{2}\|w - w^t\|^2, \tag{2.7}$$

où $w^t$ est le modèle global au début du round et $\mu$ contrôle la force de la régularisation [33]. Ce terme pénalise les écarts importants au modèle global, limitant la dérive des clients en contexte non-iid. Le terme proximal est déterministe et ne modifie pas la sensibilité des gradients par exemple, de sorte que DP-SGD s'applique directement à l'objectif modifié.

**Gestion de l'hétérogénéité.** FedProx a été conçu pour améliorer la convergence en présence d'hétérogénéité statistique et systémique (clients avec des capacités de calcul différentes). Sous contraintes DP, la régularisation proximale peut aider à stabiliser l'optimisation face au bruit, en empêchant les mises à jour locales de diverger excessivement du modèle global.

### 2.5.3 DP-FedBN

L'algorithme FedBN (*Federated Batch Normalization*) traite l'hétérogénéité de caractéristiques (*feature shift*) en maintenant les paramètres de normalisation locaux à chaque client [34].

**Principe.** Dans un réseau avec couches de normalisation (batch normalization ou layer normalization), les statistiques (moyenne, variance) et les paramètres appris (scale, shift) capturent des informations sur la distribution des activations. Lorsque les distributions d'entrée diffèrent entre clients (par exemple, des pratiques de mesure ou des populations différentes), partager ces paramètres peut dégrader les performances locales. FedBN propose de ne pas agréger les couches de normalisation : chaque client conserve ses propres paramètres de normalisation, tandis que les autres couches sont partagées et agrégées normalement.

**Intégration avec DP.** Dans DP-FedBN, l'entraînement local utilise DP-SGD pour toutes les couches, y compris les couches de normalisation. Lors de l'agrégation, seuls les paramètres non liés à la normalisation sont moyennés. Les paramètres de normalisation, bien qu'entraînés avec DP, restent locaux et ne sont pas transmis au serveur, ce qui peut offrir une forme de protection supplémentaire pour les statistiques locales.

**Avantages et limites.** FedBN peut améliorer les performances en présence de *feature shift* marqué, ce qui est plausible en santé où les pratiques de mesure varient entre établissements. La limite est que l'approche n'adresse pas directement le *label shift* (variation

de prévalence) et introduit une forme de personnalisation implicite qui peut réduire la généralisation à de nouveaux clients.

### 2.5.4 DP-Ditto

Ditto est une méthode de personnalisation fédérée qui maintient simultanément un modèle global et des modèles personnalisés pour chaque client [35].

**Principe.** Chaque client $k$ optimise un modèle personnalisé $v_k$ en minimisant :

$$\min_{v_k} F_k(v_k) + \frac{\lambda}{2}\|v_k - w\|^2, \tag{2.8}$$

où $w$ est le modèle global et $\lambda$ contrôle la régularisation vers ce modèle. Le modèle global $w$ est mis à jour par FedAvg standard (ou DP-FedAvg sous DP). La régularisation permet au modèle personnalisé de bénéficier de la collaboration tout en s'adaptant aux spécificités locales.

**Intégration avec DP.** Dans DP-Ditto, les deux phases (mise à jour du modèle global et mise à jour du modèle personnalisé) utilisent DP-SGD. Le budget de confidentialité doit comptabiliser l'ensemble des itérations des deux modèles. Le multiplicateur de bruit est calculé pour satisfaire le budget $\varepsilon$ cible sur la totalité de l'entraînement.

**Avantages et limites.** Ditto permet une personnalisation explicite et contrôlée, ce qui peut améliorer les performances sur des clients avec des distributions atypiques. La méthode a également montré des propriétés de robustesse et d'équité [35]. Sous contraintes DP, le coût en budget de confidentialité est potentiellement plus élevé (deux modèles à entraîner), et les avantages de la personnalisation peuvent être atténués par le bruit.

### 2.5.5 Discussion comparative

Le Tableau 2.3 résume les caractéristiques des quatre algorithmes DP-FL étudiés.

TABLE 2.3 – Comparaison des algorithmes DP-FL étudiés.

| **Algorithme** | **Mécanisme** | **Cible d'hétérogénéité** |
|---|---|---|
| DP-FedAvg | Agrégation standard + DP-SGD local | Baseline, pas de traitement spécifique |
| DP-FedProx | Régularisation proximale + DP-SGD | Hétérogénéité statistique (convergence) |
| DP-FedBN | Normalisation locale + DP-SGD | Décalage de caractéristiques |
| DP-Ditto | Modèles personnalisés + DP-SGD | Personnalisation explicite |

**Hypothèses et objectifs.** Les quatre méthodes partagent le même mécanisme de base (DP-SGD local) mais diffèrent par leur traitement de l'hétérogénéité. DP-FedAvg constitue la baseline ; DP-FedProx vise à stabiliser la convergence ; DP-FedBN cible les variations de distribution d'entrée ; DP-Ditto propose une personnalisation explicite. Le choix de l'algorithme dépend du type d'hétérogénéité dominant dans l'application.

**Impact du budget de confidentialité.** Sous contraintes DP strictes (faible $\varepsilon$), le bruit injecté peut dominer les différences entre algorithmes. Des études empiriques suggèrent que lorsque le bruit est élevé, les avantages spécifiques de chaque méthode de personnalisation peuvent être atténués, et le facteur déterminant de la performance devient le budget $\varepsilon$ lui-même plutôt que le choix de l'algorithme FL.

**Complexité et coût.** DP-FedAvg est le plus simple à implémenter. DP-FedProx ajoute un terme à l'objectif mais ne modifie pas significativement la complexité. DP-FedBN nécessite une gestion séparée des couches de normalisation lors de l'agrégation. DP-Ditto double potentiellement le nombre de paramètres à maintenir et le coût de calcul local.

## 2.6 Comptabilité de confidentialité par RDP

La comptabilité de confidentialité (*privacy accounting*) permet de suivre la consommation du budget $\varepsilon$ au fil des itérations d'entraînement. La confidentialité différentielle de Rényi (RDP) offre des bornes de composition plus serrées que la DP standard, ce qui la rend particulièrement adaptée à l'entraînement itératif.

### 2.6.1 Motivations de la RDP pour l'entraînement itératif

L'entraînement d'un réseau de neurones implique un grand nombre d'itérations de gradient. Chaque itération de DP-SGD consomme une partie du budget de confidentialité. La composition naïve (additive pour $\varepsilon$) conduit rapidement à des budgets prohibitifs. Des théorèmes de composition avancés permettent d'obtenir des bornes sous-linéaires, mais la RDP offre une approche plus naturelle et plus serrée pour ce type de composition [45].

**Définition de la RDP.** Un mécanisme $\mathcal{M}$ satisfait la $(\alpha, \varepsilon_\alpha)$-RDP si, pour tous jeux de données adjacents $D$ et $D'$ :

$$D_\alpha \left(\mathcal{M}(D) \| \mathcal{M}(D')\right) \leq \varepsilon_\alpha, \tag{2.9}$$

où $D_\alpha$ est la divergence de Rényi d'ordre $\alpha > 1$ [45]. La RDP offre une famille de garanties paramétrées par $\alpha$, permettant une analyse plus fine que la DP standard.

**Avantages pour la composition.** La composition de $T$ mécanismes $(\alpha, \varepsilon_\alpha)$-RDP indépendants donne un mécanisme $(\alpha, T \cdot \varepsilon_\alpha)$-RDP. Cette propriété de composition additive en $\varepsilon_\alpha$ (pour un $\alpha$ fixé) est plus favorable que les bornes de composition en $(\varepsilon, \delta)$-DP. De plus, le mécanisme gaussien (utilisé dans DP-SGD) admet une caractérisation RDP exacte, ce qui permet un calcul précis du budget consommé.

### 2.6.2 Composition des itérations et calcul du budget

Dans un entraînement DP-SGD, chaque étape de gradient correspond à un mécanisme gaussien avec sous-échantillonnage. La comptabilité RDP procède comme suit.

**Étape élémentaire.** Pour une itération avec taux d'échantillonnage $q = B/n$ (ratio taille de mini-lot sur taille du jeu de données) et multiplicateur de bruit $\sigma$, la garantie RDP est caractérisée pour chaque ordre $\alpha$. Des formules analytiques permettent de calculer $\varepsilon_\alpha$ en fonction de $q$, $\sigma$ et $\alpha$ [45].

**Composition sur $T$ itérations.** Après $T$ itérations, le budget RDP à l'ordre $\alpha$ est $T \cdot \varepsilon_\alpha$. Le nombre total d'itérations dans un cadre fédéré est $T = R \times E \times \lceil n_k/B \rceil$, où $R$ est le nombre de rounds, $E$ le nombre d'époques locales, $n_k$ la taille des données du client $k$ et $B$ la taille du mini-lot.

**Conversion RDP vers $(\varepsilon, \delta)$-DP.** Une garantie $(\alpha, \varepsilon_\alpha)$-RDP peut être convertie en garantie $(\varepsilon, \delta)$-DP par :

$$\varepsilon = \min_{\alpha > 1} \left( \varepsilon_\alpha + \frac{\log(1/\delta)}{\alpha - 1} \right). \tag{2.10}$$

En pratique, on évalue cette expression sur une grille d'ordres $\alpha$ et on retient le minimum. Cette conversion permet de rapporter un budget $\varepsilon$ interprétable pour un $\delta$ fixé [45].

### 2.6.3 Paramètres pratiques

La configuration de la comptabilité RDP implique plusieurs choix pratiques.

**Choix de $\delta$.** Le paramètre $\delta$ est généralement fixé à une valeur inférieure à $1/n$. Des valeurs courantes sont $\delta = 10^{-5}$ ou $\delta = 10^{-6}$, garantissant que la probabilité d'échec de la garantie DP est négligeable par rapport à la taille du jeu de données.

**Calcul du multiplicateur de bruit.** Étant donné un budget cible $\varepsilon$, un $\delta$ fixé, un nombre d'itérations $T$ et un taux d'échantillonnage $q$, le multiplicateur de bruit $\sigma$ est déterminé par recherche binaire : on cherche le plus petit $\sigma$ tel que le budget consommé (calculé par RDP) reste inférieur ou égal à $\varepsilon$. Des bibliothèques comme Opacus [46] fournissent des fonctions pour ce calcul.

**Ordres $\alpha$ considérés.** Le calcul RDP évalue la garantie sur une grille d'ordres $\alpha$, typiquement $\alpha \in [2, 64]$ ou $[2, 128]$. La conversion vers $(\varepsilon, \delta)$-DP sélectionne l'ordre optimal. En pratique, des ordres modérés ($\alpha \in [10, 30]$) sont souvent optimaux pour les configurations courantes de DP-SGD.

**Budget dépensé vs budget cible.** En raison de la discrétisation des itérations et du caractère conservatif de certaines bornes, le budget effectivement dépensé $\varepsilon_{\text{spent}}$ peut être légèrement inférieur au budget cible. Il est recommandé de rapporter $\varepsilon_{\text{spent}}$ plutôt que le budget cible pour une description précise des garanties obtenues.

## 2.7 Calibration des probabilités en prédiction clinique

En prédiction clinique, la capacité d'un modèle à produire des probabilités fiables est aussi importante que sa capacité à discriminer les classes. Cette section présente les concepts de calibration, les métriques associées, les méthodes de calibration post-hoc et leur interaction avec la confidentialité différentielle.

### 2.7.1 Discrimination et calibration : définitions et enjeux

**Discrimination.** La discrimination mesure la capacité d'un modèle à ordonner correctement les individus selon leur risque. Un modèle parfaitement discriminant attribue des scores plus élevés à tous les cas positifs qu'à tous les cas négatifs. L'AUROC quantifie cette propriété : elle correspond à la probabilité qu'un cas positif choisi au hasard reçoive un score supérieur à un cas négatif choisi au hasard.

**Calibration.** La calibration mesure l'adéquation entre les probabilités prédites et les fréquences observées. Un modèle parfaitement calibré vérifie :

$$\Pr(Y = 1 \mid \hat{p}(X) = p) = p, \quad \forall p \in [0, 1]. \tag{2.11}$$

Autrement dit, parmi les patients auxquels le modèle attribue une probabilité de décès de 30%, environ 30% décèdent effectivement.

**Indépendance des deux notions.** Discrimination et calibration sont des propriétés distinctes. Un modèle peut être excellent en discrimination mais mal calibré (par exemple, en surestimant systématiquement les probabilités). Inversement, un modèle peut être bien calibré mais peu discriminant. Les réseaux de neurones profonds sont souvent bien discriminants mais tendent à être surconfiants (mal calibrés) [12].

**Enjeux cliniques.** En pratique clinique, les probabilités prédites sont souvent interprétées comme des risques et peuvent influencer des décisions thérapeutiques. Une surestimation du risque peut conduire à des traitements excessifs ; une sous-estimation peut retarder des interventions nécessaires. La calibration est donc essentielle pour l'usage opérationnel des modèles prédictifs en santé [22].

### 2.7.2 Mesures de calibration

Plusieurs métriques permettent d'évaluer la calibration d'un modèle.

**Expected Calibration Error (ECE).** L'ECE approxime l'écart moyen entre confiance et exactitude. Les prédictions sont regroupées en $B$ intervalles (bins) de probabilité, et pour chaque bin $b$, on calcule l'écart entre la confiance moyenne $\text{conf}(b)$ et l'exactitude observée $\text{acc}(b)$ :

$$\text{ECE} = \sum_{b=1}^{B} \frac{|B_b|}{n} \left|\text{acc}(B_b) - \text{conf}(B_b)\right|, \tag{2.12}$$

où $|B_b|$ est le nombre d'exemples dans le bin $b$ et $n$ le nombre total d'exemples [12]. Un ECE de zéro indique une calibration parfaite.

**Score de Brier.** Le score de Brier mesure l'erreur quadratique moyenne entre les probabilités prédites et les issues observées :

$$\text{Brier} = \frac{1}{n} \sum_{i=1}^{n} (\hat{p}_i - y_i)^2, \tag{2.13}$$

où $\hat{p}_i$ est la probabilité prédite et $y_i \in \{0, 1\}$ l'issue observée [24]. Le score de Brier combine discrimination et calibration ; il peut être décomposé en composantes de fiabilité (calibration), résolution (discrimination) et incertitude.

**Courbes de fiabilité.** Les courbes de fiabilité (*reliability diagrams*) offrent une visualisation de la calibration. Pour chaque bin de probabilité, on trace l'exactitude observée en fonction de la confiance moyenne. Une calibration parfaite correspond à la diagonale. Les écarts à la diagonale indiquent une surconfiance (en dessous) ou une sous-confiance (au-dessus).

### 2.7.3 Méthodes de calibration post-hoc

Lorsqu'un modèle est mal calibré, des méthodes de calibration post-hoc permettent d'ajuster les probabilités sans réentraîner le modèle.

**Temperature scaling.** Le *temperature scaling* est une méthode simple et efficace introduite par Guo et al. [12]. Elle consiste à diviser les logits (sorties avant la fonction sigmoïde ou softmax) par un paramètre scalaire $T > 0$ :

$$\hat{p}_{\text{calibré}} = \sigma(z/T), \tag{2.14}$$

où $z$ est le logit et $\sigma$ la fonction sigmoïde. Le paramètre $T$ est optimisé sur un ensemble de validation en minimisant la perte de cross-entropie. Une température $T > 1$ adoucit les probabilités (réduit la surconfiance) ; $T < 1$ les accentue.

**Avantages du temperature scaling.** Le temperature scaling ne modifie pas l'ordonnancement des prédictions (il préserve la discrimination) et n'introduit qu'un seul paramètre, ce qui réduit le risque de surapprentissage sur l'ensemble de validation. Malgré sa simplicité, il surpasse souvent des méthodes plus complexes (Platt scaling, isotonic regression) sur de nombreuses architectures [12].

**Autres méthodes.** D'autres approches existent : le *Platt scaling* ajuste une régression logistique sur les logits ; la *régression isotonique* apprend une fonction monotone de recalibration ; les méthodes basées sur les histogrammes ajustent les probabilités par bin. Ces méthodes peuvent offrir plus de flexibilité mais au prix d'un risque accru de surapprentissage.

### 2.7.4 Calibration et confidentialité différentielle

L'interaction entre calibration et confidentialité différentielle mérite une attention particulière.

**Effet potentiel du bruit DP sur la calibration.** Le bruit injecté par DP-SGD, ainsi que le clipping, modifient la dynamique d'optimisation et la distribution des sorties (logits/probabilités). En conséquence, la calibration d'un modèle entraîné sous DP peut être affectée, ce qui rend nécessaire une évaluation explicite de la calibration (p. ex. ECE, score de Brier) et, si besoin, l'application de méthodes de calibration post-hoc.

**Post-traitement sans coût de confidentialité.** Une propriété fondamentale de la confidentialité différentielle est l'immunité au post-traitement : toute fonction appliquée à la sortie d'un mécanisme DP reste DP sans consommation supplémentaire de budget [39]. Le temperature scaling, qui ne dépend que des logits produits par le modèle et des labels de l'ensemble de validation (distinct de l'ensemble d'entraînement), constitue un post-traitement. Il peut donc être appliqué sans affecter les garanties de confidentialité du modèle.

**Utilisation d'un ensemble de validation séparé.** L'optimisation du paramètre de température s'effectue sur un ensemble de validation qui n'a pas été utilisé pour l'entraînement du modèle. Cet ensemble peut être constitué à partir des données locales (validation locale chez chaque client) ou d'un ensemble de validation global (agrégation des validations locales). Dans ce mémoire, nous utilisons un ensemble de validation global pour ajuster une température unique appliquée à toutes les prédictions.

**Implications pratiques.** Le temperature scaling représente une amélioration « gratuite » en termes de confidentialité : il améliore la calibration sans consommer de budget $\varepsilon$ supplémentaire. Cette propriété est particulièrement précieuse dans les applications médicales où la calibration est essentielle et où le budget de confidentialité est contraint.

## 2.8 Synthèse et formulation du problème scientifique

Ce chapitre a permis de dresser le panorama des méthodes de prédiction de mortalité en réanimation, des architectures d'apprentissage fédéré et des mécanismes de confidentialité différentielle. Cette section synthétise les limites de l'existant pour formaliser la problématique scientifique traitée dans ce mémoire.

### 2.8.1 Lacunes identifiées et verrou scientifique

L'analyse de la littérature révèle que si les briques technologiques (FL, DP, modèles cliniques) sont matures individuellement, leur intégration conjointe dans un cadre rigoureux soulève des questions non résolues :

- **Manque de comparabilité :** Les études existantes comparent rarement les algorithmes fédérés (FedAvg, FedProx, Ditto) dans un cadre strictement homogène (mêmes données, même tâche, même budget de confidentialité), rendant difficile l'identification de la méthode la plus adaptée aux données de réanimation.
- **Impact méconnu de la DP sur la personnalisation :** Les méthodes comme FedProx ou Ditto sont conçues pour gérer l'hétérogénéité (non-IID). Cependant, on ignore si leurs bénéfices théoriques persistent lorsqu'un bruit de confidentialité significatif (faible $\varepsilon$) est injecté dans les gradients.
- **Absence d'analyse de la calibration :** La majorité des travaux se focalisent sur la discrimination (AUROC). Or, l'ajout de bruit DP risque de dégrader la fiabilité des probabilités (calibration), un aspect critique pour la décision clinique qui reste largement inexploré.
- **Rigueur de la comptabilité :** De nombreux travaux rapportent des budgets théoriques sans détailler la consommation réelle via RDP ou sans inclure l'impact des hyperparamètres d'échantillonnage.

**Formulation du problème.** Le verrou scientifique de ce mémoire peut donc se formuler ainsi : *Comment garantir un apprentissage fédéré performant et bien calibré pour la prédiction de mortalité en réanimation, en présence d'une forte hétérogénéité des données (non-IID), tout en assurant des garanties formelles de confidentialité différentielle ?*

Plus spécifiquement, il s'agit de déterminer si les mécanismes de gestion de l'hétérogénéité (régularisation proximale, personnalisation) conservent leur utilité lorsque l'entraînement est soumis à un bruit de confidentialité significatif (faible $\varepsilon$).

### 2.8.2 Positionnement du travail

Pour répondre à cette problématique, ce mémoire propose une étude empirique systématique caractérisée par :

**Comparaison multi-algorithmes sous contraintes identiques.** Nous évaluons quatre approches (DP-FedAvg, DP-FedProx, DP-FedBN, DP-Ditto) sur une tâche commune de prédiction à 24 h, en utilisant la même comptabilité RDP pour assurer une équité stricte des comparaisons.

**Analyse dimensionnelle de la performance.** Au-delà de l'AUROC, nous intégrons systématiquement l'AUPRC (pour le déséquilibre de classe) et des métriques de calibration (ECE, Brier), en évaluant l'efficacité du *temperature scaling* comme remède à la dégradation induite par la DP.

**Robustesse expérimentale.** Contrairement à de nombreuses études préliminaires, nos résultats sont consolidés par des répétitions multi-graines (3 seeds) et des tests de significativité statistique, sur une partition réaliste de la base MIMIC-IV simulant cinq types d'unités de soins distincts.

### 2.8.3 Lien avec les chapitres suivants

La résolution de cette problématique nécessite la mise en place d'un pipeline expérimental rigoureux.

Le Chapitre 3 décrira la construction de la cohorte à partir de MIMIC-IV, le prétraitement des données, et l'établissement des modèles de référence centralisés.

Le Chapitre 4 détaillera l'implémentation technique des algorithmes fédérés et du mécanisme DP-SGD, puis présentera les résultats expérimentaux, l'analyse de calibration, et la comparaison avec l'état de l'art, apportant ainsi les réponses empiriques aux questions soulevées ici.

# Chapitre 3

# Préparation des données et établissement des baselines

Ce chapitre définit le cadre expérimental de notre étude. Nous présentons en premier lieu la base de données MIMIC-IV et la méthodologie rigoureuse de sélection de la cohorte. Nous détaillons ensuite le pipeline de traitement des données, incluant les stratégies d'imputation et d'ingénierie des caractéristiques. Une étude comparative approfondie des modèles centralisés (baselines) est menée pour établir le plafond de performance théorique. Enfin, nous décrivons la simulation de l'environnement fédéré, spécifiquement conçue pour reproduire une hétérogénéité réaliste entre unités de soins intensifs.

## 3.1 Base de données MIMIC-IV

### 3.1.1 Description générale

La base de données MIMIC-IV (*Medical Information Mart for Intensive Care IV*) constitue une référence majeure en recherche biomédicale. Elle regroupe les dossiers médicaux électroniques dé-identifiés de patients admis au *Beth Israel Deaconess Medical Center* (Boston, USA) entre 2008 et 2019 [8]. Cette ressource se distingue par sa richesse, documentant plus de 94 458 séjours en soins intensifs avec une granularité temporelle fine.

L'architecture de la base s'articule autour de modules thématiques. Notre étude exploite principalement les modules `hosp` (données intra-hospitalières : démographie, biologie) et `icu` (données de réanimation : signes vitaux, fluides, interventions). La structure relationnelle repose sur trois clés primaires : `subject_id` (le patient), `hadm_id` (l'admission à l'hôpital) et `stay_id` (le séjour spécifique en ICU). Conformément à notre problématique de prédiction intra-séjour, l'unité d'analyse retenue est le `stay_id`.

L'accès aux données a été obtenu après validation de la certification PhysioNet (formation CITI), garantissant le respect des normes éthiques et de confidentialité inhérentes à la manipulation de données de santé sensibles.

### 3.1.2 Définition de la cohorte et critères d'inclusion

La constitution de la cohorte suit un diagramme de flux strict (Figure 3.1) visant à isoler une population adulte pour laquelle la prédiction précoce de mortalité présente un intérêt clinique.

Les critères d'inclusion retiennent tout patient adulte ($\geq$ 18 ans) dont le séjour en ICU excède 24 heures, afin de garantir une fenêtre d'observation suffisante pour la collecte des variables prédictives. Les critères d'exclusion s'appliquent en cascade :

1. **Durée de séjour :** 19 615 séjours de moins de 24 heures ont été écartés. Une durée aussi courte correspond souvent à des surveillances post-opératoires simples ou à des décès très précoces, hors du périmètre de la prédiction à 24h.
2. **Qualité des données :** 6 507 séjours présentant un taux de valeurs manquantes supérieur à 40% sur les variables critiques ont été exclus pour garantir la fiabilité de l'apprentissage.

Le processus aboutit à une cohorte finale de **68 322 séjours**. Le Tableau 3.1 en synthétise les propriétés. La variable cible est la mortalité intra-ICU, observée avec une prévalence de 10.84%. Ce déséquilibre de classe modéré motive l'usage de métriques complémentaires à l'exactitude, telles que l'AUPRC.

TABLE 3.1 – Caractéristiques de la cohorte finale (N = 68 322).

| **Caractéristique** | **Valeur / Proportion** |
|---|---|
| Nombre total de séjours | 68 322 |
| Taux de mortalité en ICU | 10.84% |
| Répartition (Train / Val / Test) | 70% / 10% / 20% |

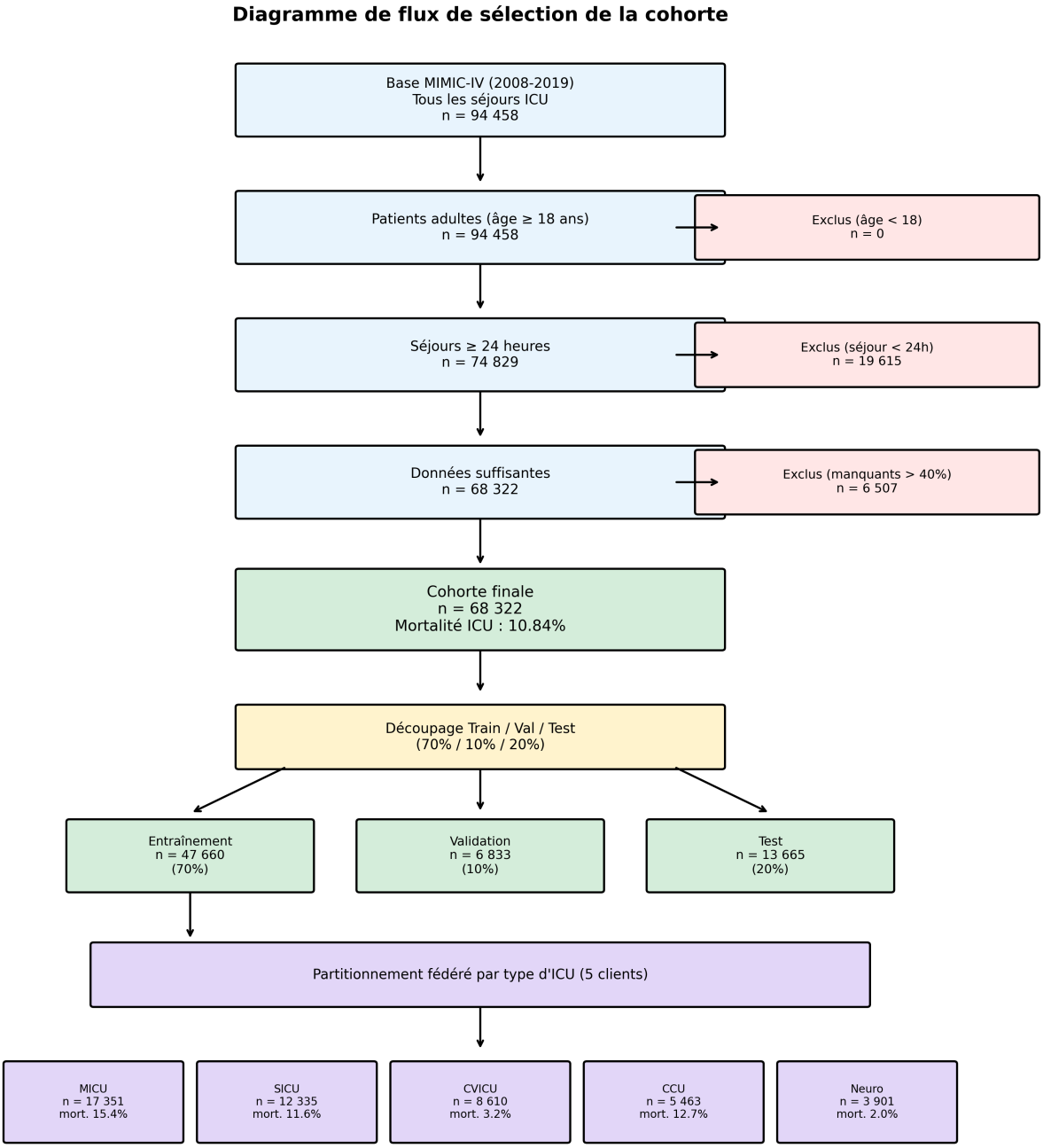


FIGURE 3.1 – Diagramme de flux de sélection de la cohorte à partir de MIMIC-IV.

## 3.2 Pipeline de prétraitement

La qualité des données étant déterminante pour la performance et la stabilité des modèles, notamment dans un contexte différentiellement privé bruité, un pipeline de prétraitement robuste a été mis en place (Figure 3.2).

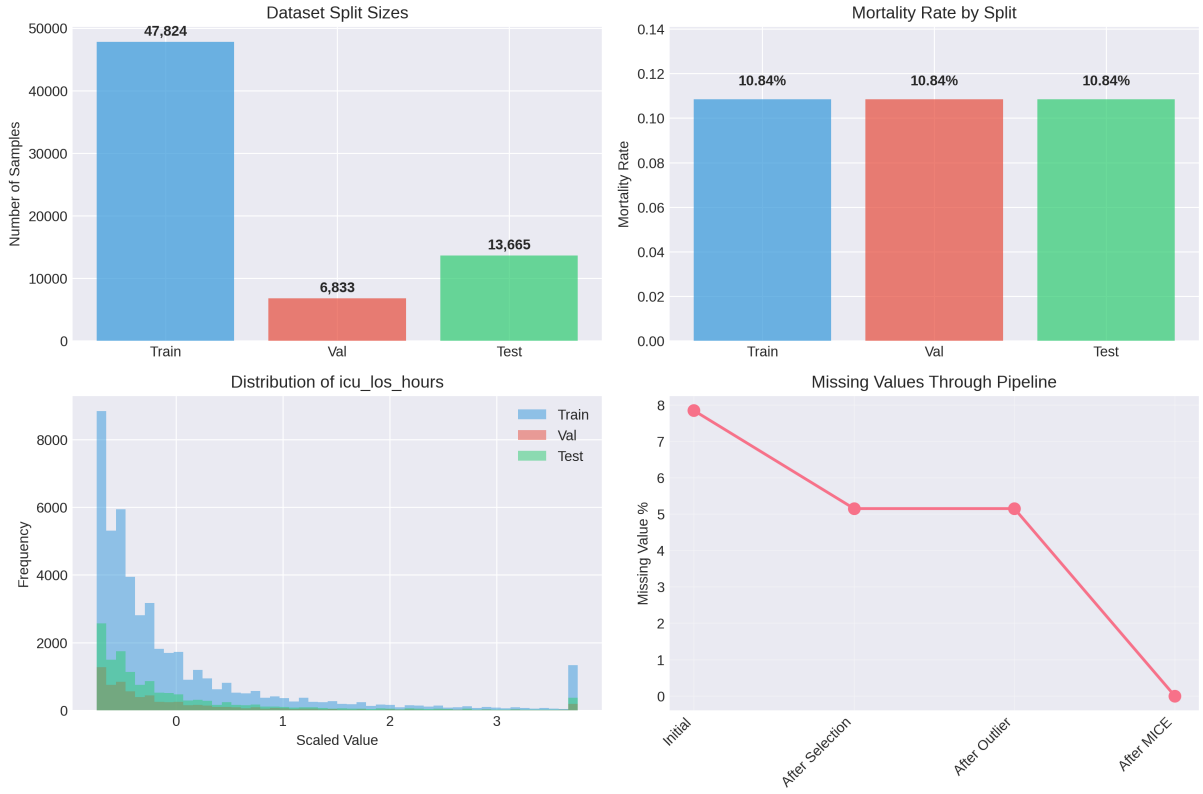


FIGURE 3.2 – Vue d'ensemble du pipeline de prétraitement : extraction, imputation, normalisation et ingénierie.

### 3.2.1 Extraction et agrégation temporelle

Pour chaque séjour, les données sont extraites sur la fenêtre [0h, 24h] suivant l'admission en réanimation. Les séries temporelles (signes vitaux, biologie) sont résumées par des statistiques agrégées (min, max, moyenne, écart-type, dernière valeur) afin de produire un vecteur de caractéristiques de taille fixe.

Les variables couvrent quatre dimensions cliniques :

- **Démographie :** Âge, sexe, ethnie.
- **Signes vitaux :** Fréquence cardiaque, pression artérielle, fréquence respiratoire, $SpO_2$, température.
- **Biologie :** Marqueurs rénaux (créatinine, urée), hépatiques (bilirubine), inflammatoires (leucocytes, CRP), gaz du sang (pH, lactates).
- **Support clinique :** Score de Glasgow, ventilation mécanique, vasopresseurs.

### 3.2.2 Gestion des manquants et normalisation

L'absence de données en réanimation est souvent informative (absence de mesure = patient stable). Cependant, pour les besoins de la modélisation, nous appliquons une imputation multiple par équations chaînées (*MICE*), qui préserve les corrélations inter-variables mieux qu'une imputation par la moyenne [10]. Les valeurs aberrantes sont traitées par *winsorization* (écrêtage à 5 écarts-types).

La normalisation (Z-score) est effectuée en utilisant strictement les statistiques (moyenne, écart-type) calculées sur l'**ensemble d'entraînement**. Ces paramètres sont ensuite appliqués aux ensembles de validation et de test, prévenant ainsi toute fuite d'information (*data leakage*).

### 3.2.3 Ingénierie des caractéristiques

Pour enrichir la représentation, 23 variables dérivées ont été créées, portant le total à 97 caractéristiques. Ces ajouts incluent des ratios physiopathologiques pertinents (ex : BUN/Créatinine pour l'hydratation, Index de choc) et des scores composites de défaillance d'organes (SOFA partiel). Comme le montre le Tableau 3.2, cet enrichissement améliore significativement la discrimination (+0.97% d'AUROC), confirmant la valeur de l'expertise métier intégrée au modèle.

TABLE 3.2 – Impact de l'ingénierie des caractéristiques (Validation par ablation).

| **Configuration** | **AUROC** | **AUPRC** |
|---|---|---|
| Variables brutes (74) | 0.8305 | 0.3960 |
| Avec Feature Engineering (97) | **0.8385** | **0.4059** |

## 3.3 Modèles de base centralisés

Avant de distribuer l'apprentissage, il est essentiel d'établir des performances de référence (*baselines*) en environnement centralisé. Cette étape valide la qualité des données et justifie le choix de l'architecture finale.

### 3.3.1 Panorama des modèles évalués

Nous avons comparé un large spectre d'algorithmes, allant des méthodes classiques au Deep Learning moderne :

— **Gradient Boosting (XGBoost, LightGBM) :** État de l'art actuel sur données tabulaires, reconnus pour leur gestion native des manquants et leur robustesse.

— **Architectures Deep Tabulaires :** TabNet (mécanisme d'attention séquentielle), FT-Transformer (Feature Tokenizer) et SAINT, qui adaptent les principes des Transformers aux données structurées.

— **MLP amélioré :** Un perceptron multicouche optimisé avec connexions résiduelles, normalisation et dropout.

L'optimisation des hyperparamètres a été réalisée via le framework Optuna (algorithme TPE, 150 essais), en maximisant l'AUROC sur l'ensemble de validation.

### 3.3.2 Résultats et choix de l'architecture fédérée

Les résultats sur l'ensemble de test (Tableau 3.3) confirment la suprématie des méthodes d'ensemble. Le *Stacked Ensemble* domine avec un AUROC de 0.8715. Parmi les modèles simples, LightGBM (0.8694) devance légèrement les approches Deep Learning.

Cependant, pour la phase d'apprentissage fédéré sous confidentialité différentielle (DP-FL), nous sélectionnons l'**Enhanced MLP** (AUROC 0.8626). Ce choix stratégique se justifie par trois facteurs :

1. **Compatibilité DP-SGD :** Contrairement aux arbres de décision, le MLP est entièrement différentiable, permettant le calcul direct des gradients par échantillon (*per-sample gradients*) nécessaire au mécanisme de privatisation.
2. **Coût communicationnel :** L'architecture MLP (env. 38 000 paramètres) est compacte, réduisant la charge réseau lors de l'envoi des mises à jour au serveur.
3. **Performance compétitive :** L'écart de performance avec le meilleur modèle d'arbre est minime ($< 0.7\%$), un compromis acceptable pour l'intégration des garanties de confidentialité.

TABLE 3.3 – Performances comparatives sur l'ensemble de test.

| **Modèle** | **AUROC** | **AUPRC** | **Brier Score** |
|---|---|---|---|
| Régression Logistique | 0.8341 | 0.3719 | 0.1665 |
| XGBoost | 0.8686 | 0.4841 | 0.0839 |
| LightGBM | **0.8694** | **0.4858** | 0.0841 |
| **Enhanced MLP (Retenu)** | 0.8626 | 0.4820 | 0.0738 |
| TabNet | 0.8567 | 0.4574 | 0.0760 |
| *Stacked Ensemble* | *0.8715* | *0.4967* | *0.0724* |

## 3.4 Simulation de l'environnement fédéré

En l'absence d'infrastructure multi-hôpitaux réelle, nous simulons un environnement fédéré en exploitant la structure interne de MIMIC-IV. Le partitionnement des données est effectué selon le type d'unité de soins (*Care Unit*), créant cinq clients virtuels distincts : MICU (Médicale), SICU (Chirurgicale), CCU (Coronarienne), CVICU (Cardiovasculaire) et Neuro.

### 3.4.1 Analyse de l'hétérogénéité (non-IID)

Ce découpage induit une hétérogénéité statistique forte, représentative des défis réels (Figure 3.3). Nous observons deux types de décalages distributionnels :

— **Déséquilibre de quantité :** Le client MICU possède 4,4 fois plus de données que le client Neuro (17 351 vs 3 901 séjours).

— **Décalage de prévalence (Label Shift) :** Le taux de mortalité varie drastiquement, allant de 2.00% en Neuro-réanimation à 15.37% en réanimation médicale.

Un tel écart (ratio de 7.7) suggère que l'optimisation d'un modèle global unique sera complexe, les clients ayant des objectifs locaux divergents. Cette configuration justifie pleinement l'évaluation de méthodes fédérées avancées comme FedProx ou Ditto, conçues pour gérer ce type de données non identiquement distribuées (non-IID).

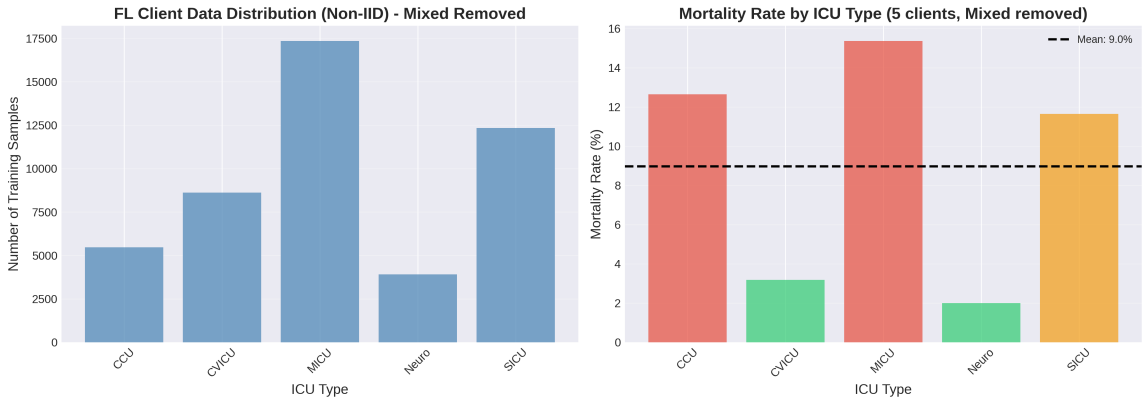


FIGURE 3.3 – Caractérisation de l'hétérogénéité : (a) Volume de données par client, (b) Variabilité de la mortalité intra-ICU.

## Synthèse

Ce chapitre a posé les bases méthodologiques de notre expérimentation. À partir de la base MIMIC-IV, nous avons construit une cohorte robuste de 68 322 séjours et un pipeline de traitement validé par des baselines centralisées performantes (AUROC $> 0.86$). Le choix de l'architecture MLP a été motivé par la nécessité technique d'appliquer le DP-SGD, malgré une performance brute légèrement inférieure aux méthodes de boosting. Enfin, la

simulation fédérée par type d'unité révèle une forte hétérogénéité de prévalence, préparant le terrain pour l'analyse comparative des algorithmes DP-FL au chapitre suivant.

# Chapitre 4

# Implémentation et résultats expérimentaux

Ce chapitre détaille la concrétisation technique de notre système d'apprentissage fédéré sous contraintes de confidentialité différentielle. Nous y décrivons l'environnement matériel et logiciel, l'architecture des algorithmes implémentés (FedAvg, FedProx, FedBN, Ditto), ainsi que l'intégration critique de la bibliothèque Opacus pour la gestion du mécanisme DP-SGD. Enfin, nous présentons le protocole de calibration post-hoc et la configuration expérimentale rigoureuse qui sous-tend les résultats présentés au chapitre suivant.

## 4.1 Environnement technique

### 4.1.1 Infrastructure matérielle

Les expériences ont été conduites sur la plateforme de calcul Kaggle, utilisant des accélérateurs graphiques NVIDIA Tesla P100 (16 Go VRAM). L'environnement dispose de 13 Go de RAM et d'un temps d'exécution limité à 12 heures par session. Ces contraintes de ressources ont guidé nos choix architecturaux, notamment l'utilisation d'un modèle MLP compact plutôt que de grands modèles de langage ou de vision, afin de permettre l'exécution complète des simulations fédérées (50 rounds) dans le temps imparti.

### 4.1.2 Stack logiciel

Le développement repose sur l'écosystème Python et PyTorch. Le Tableau 4.1 recense les bibliothèques clés. L'élément central est la bibliothèque **Opacus**, développée par Meta Research, qui permet une vectorisation efficace du calcul des gradients par échantillon (*per-sample gradients*), indispensable à DP-SGD.

TABLE 4.1 – Environnement logiciel pour les expériences DP-FL.

| **Composant** | **Version** | **Rôle principal** |
|---|---|---|
| Python | 3.11 | Langage de programmation |
| PyTorch | 2.1+ | Framework d'apprentissage profond |
| Opacus | 1.4 | Implémentation de DP-SGD et comptabilité RDP |
| NumPy / Pandas | 1.24+ / 2.0+ | Manipulation matricielle et gestion des données |
| Optuna | 3.0+ | Optimisation bayésienne des hyperparamètres |
| Scikit-learn | 1.3+ | Métriques et prétraitement |

### 4.1.3 Simulation de l'environnement fédéré

Conformément aux standards de la littérature [30], nous simulons l'environnement distribué de manière centralisée. Les données partitionnées par type d'ICU sont chargées en mémoire, et l'orchestrateur exécute séquentiellement l'entraînement local de chaque client sur le GPU disponible. Bien que l'échange réseau soit simulé, la logique algorithmique (isolation des données, agrégation des poids) et le mécanisme de confidentialité (bruitage local) sont strictement identiques à un déploiement réel.

## 4.2 Algorithmes d'apprentissage fédéré

Nous avons implémenté une architecture modulaire où chaque algorithme hérite d'une classe `BaseFLTrainer`. Cette conception assure que tous les modèles partagent le même pipeline d'évaluation et de gestion des graines aléatoires (*seeds*), garantissant une comparaison équitable.

### 4.2.1 FedAvg : La référence

L'algorithme FedAvg sert de baseline. À chaque round $t$, le serveur diffuse le modèle $w^t$. Chaque client $k$ effectue $E$ époques de descente de gradient stochastique (SGD) sur ses données locales pour produire $w_k^{t+1}$. L'agrégation est une moyenne pondérée par la taille des jeux de données $n_k$.

### 4.2.2 FedProx : Gestion de la divergence

Pour contrer l'hétérogénéité des données qui éloigne les modèles locaux de l'optimum global, FedProx introduit un terme proximal dans la fonction de perte locale [33] :

$$\min_w \mathcal{L}_k(w) + \frac{\mu}{2}\|w - w^t\|^2 \tag{4.1}$$

Ce terme pénalise les changements drastiques des paramètres. Nous avons implémenté

ce mécanisme en calculant explicitement la norme L2 de la différence des poids lors de la passe avant. Le paramètre $\mu$ a été optimisé par validation croisée ($\mu = 0.01$).

### 4.2.3 FedBN : Normalisation locale

FedBN part du principe que le décalage de distribution (*feature shift*) se reflète principalement dans les statistiques de normalisation. L'algorithme exclut donc les couches de normalisation de l'agrégation : chaque client conserve ses propres paramètres (moyenne, variance, poids affines) d'un round à l'autre [34].

**Note architecturale et compatibilité DP.** Une spécificité importante de notre implémentation concerne le choix de la couche de normalisation. La *Batch Normalization* standard pose problème avec la confidentialité différentielle (DP-SGD) car elle introduit une dépendance entre les échantillons d'un même lot. Pour garantir la validité du calcul des gradients par échantillon, nous utilisons la **Layer Normalization** (*LayerNorm*). Dans notre cas, le modèle retenu est un MLP tabulaire. Nous appliquons le *principe* de FedBN en conservant localement les paramètres de normalisation (ici *LayerNorm* pour compatibilité DP-SGD) et en n'agrégeant que les autres paramètres du réseau.

### 4.2.4 Ditto : Personnalisation par régularisation

Ditto apprend simultanément un modèle global et des modèles personnalisés pour chaque client [35]. Chaque client $k$ maintient deux jeux de paramètres : le modèle global $w$ et le modèle personnalisé $v_k$. Le modèle personnalisé est entraîné en minimisant :

$$\mathcal{L}_k^{\text{ditto}}(v_k) = \mathcal{L}_k(v_k) + \frac{\lambda}{2}\|v_k - w^t\|^2 \tag{4.2}$$

Notre implémentation réutilise la fonction `computed_proximal_loss` définie pour FedProx, mais l'applique dans une phase d'optimisation distincte propre à chaque client.

### 4.2.5 Synthèse comparative

Le Tableau 4.2 résume les propriétés structurelles des quatre algorithmes déployés.

TABLE 4.2 – Comparaison des algorithmes FL implémentés.

| **Caractéristique** | **FedAvg** | **FedProx** | **FedBN** | **Ditto** |
|---|---|---|---|---|
| Agrégation pondérée | ✓ | ✓ | ✓ | ✓ |
| Régularisation proximale | — | ✓ | — | ✓ |
| Paramètres locaux | — | — | LayerNorm | Modèle complet |
| Hyperparamètre spécifique | — | $\mu$ | — | $\lambda$ |
| Cible d'hétérogénéité | — | Label shift | Feature shift | Les deux |
| Complexité mémoire | 1× | 1× | 1× | 2× |

## 4.3 Intégration de la confidentialité différentielle

L'ajout de garanties formelles repose sur le remplacement de l'optimiseur standard par DP-SGD. Nous utilisons `Opacus` pour gérer la complexité mathématique de cette opération.

### 4.3.1 Mécanisme DP-SGD et Opacus

Le processus suit trois étapes critiques gérées par le `PrivacyEngine` d'Opacus :

1. **Clipping :** Les gradients sont calculés pour chaque échantillon individuel. Si la norme L2 d'un gradient $g_i$ dépasse un seuil $C$, il est mis à l'échelle. Nous fixons $C = 1.0$.
2. **Agrégation et Bruitage :** Les gradients clippés sont sommés, et un bruit gaussien $\mathcal{N}(0, \sigma^2 C^2 \mathbf{I})$ est ajouté.
3. **Comptabilité :** Le budget $\varepsilon$ consommé est suivi en temps réel.

### 4.3.2 Calibration du bruit et comptabilité RDP

Pour garantir un budget exact, nous utilisons la méthode inverse : fixer le budget cible $\varepsilon$ et calculer le $\sigma$ requis via l'analyseur RDP. Pour 50 rounds de 5 époques locales, le Tableau 4.3 illustre les niveaux de bruit calculés. On note que pour $\varepsilon = 2.0$, le bruit est très élevé ($\sigma \approx 9.9$), ce qui représente un défi majeur pour la convergence.

TABLE 4.3 – Multiplicateurs de bruit $\sigma$ calculés pour les budgets cibles ($\delta = 10^{-5}$).

| **Budget Cible ($\varepsilon$)** | **Multiplicateur ($\sigma$)** | **Budget Réel Dépensé** |
|---|---|---|
| 2.0 | 9.92 | 1.83 |
| 4.0 | 4.96 | 3.66 |
| 6.0 | 3.31 | 5.50 |
| 8.0 | 2.48 | 7.34 |

## 4.4 Protocole de calibration post-hoc

Les modèles différentiellement privés souffrent souvent de mauvaise calibration. Nous appliquons le *Temperature Scaling* pour corriger les probabilités.

### 4.4.1 Méthodologie

Une température scalaire $T$ est apprise pour redimensionner les logits $z$ avant la fonction sigmoïde : $\hat{p} = \sigma(z/T)$. L'optimisation de $T$ se fait par minimisation de la *Negative Log Likelihood* (NLL) sur l'ensemble de validation.

### 4.4.2 Préservation de la confidentialité

Un point théorique crucial est que cette étape ne consomme pas de budget de confidentialité supplémentaire. En vertu du **théorème de post-traitement** de la confidentialité différentielle [39], toute fonction appliquée à la sortie d'un mécanisme $(\varepsilon, \delta)$-DP reste $(\varepsilon, \delta)$-DP, tant qu'elle n'accède pas aux données privées d'entraînement. Ici, $T$ est calibré sur l'ensemble de validation (disjoint de l'entraînement), préservant ainsi l'intégrité des garanties formelles.

## 4.5 Configuration expérimentale

Pour assurer la reproductibilité et la robustesse statistique, nous avons défini un protocole strict résumé dans le Tableau 4.4.

TABLE 4.4 – Hyperparamètres globaux de l'apprentissage fédéré.

| **Paramètre** | **Valeur** |
|---|---|
| Nombre de clients ($K$) | 5 (Partitionnement par ICU) |
| Rounds globaux | 50 |
| Époques locales | 5 |
| Batch size | 64 |
| Taux d'apprentissage | $10^{-4}$ (Adam) |
| Norme de clipping ($C$) | 1.0 |
| Graines aléatoires (Seeds) | 0, 1, 2 (pour chaque expérience) |

La grille expérimentale complète comporte 48 configurations principales (4 algorithmes $\times$ 4 niveaux de budget $\times$ 3 graines), comparées aux baselines non-privées.

## 4.6 Résultats et discussion

Cette section présente les résultats expérimentaux de notre étude. Nous analysons successivement les performances des modèles centralisés (rappel), l'impact de la fédération sans contrainte de confidentialité, puis les résultats complets de l'apprentissage fédéré sous confidentialité différentielle. L'analyse de la calibration, les tests statistiques et la comparaison avec l'état de l'art complètent cette évaluation. Nous concluons par une discussion des implications cliniques et des limitations.

### 4.6.1 Rappel des résultats centralisés

Les modèles centralisés, présentés en détail à la Section 3, établissent le plafond de performance atteignable sans contraintes de distribution des données ni de confidentialité. Le Tableau 4.5 rappelle les performances clés.

TABLE 4.5 – Récapitulatif des performances centralisées (ensemble de test, n = 13 665).

| **Modèle** | **AUROC** | **AUPRC** | **Brier** |
|---|---|---|---|
| Stacked Ensemble | **0.8715** | **0.4967** | **0.0724** |
| LightGBM | 0.8694 | 0.4858 | 0.0841 |
| XGBoost | 0.8686 | 0.4841 | 0.0839 |
| Enhanced MLP | 0.8626 | 0.4820 | 0.0738 |

L'ensemble stacking (LightGBM + XGBoost + MLP) atteint un AUROC de 0.8715, représentant la borne supérieure de performance. Le MLP amélioré, retenu pour les expériences fédérées en raison de sa compatibilité avec DP-SGD, atteint 0.8626, soit une différence de 1.0% par rapport à l'ensemble optimal.

### 4.6.2 Résultats de l'apprentissage fédéré sans DP

Avant d'introduire les contraintes de confidentialité différentielle, nous évaluons l'impact de la fédération seule sur les performances. Cette baseline permet d'isoler le coût de la distribution des données (hétérogénéité non-IID) du coût de la confidentialité.

#### Performances globales

Le Tableau 4.6 présente les résultats des quatre algorithmes FL sans confidentialité différentielle.

**Analyse.** Ditto surpasse le MLP centralisé (+0.34% en AUROC), démontrant que la personnalisation peut compenser, voire dépasser, les effets négatifs de la fédération. Ce

TABLE 4.6 – Performances des algorithmes FL sans DP (ensemble de test).

| **Algorithme** | **AUROC** | **AUPRC** | **F1** | **Écart vs Centralisé** |
|---|---|---|---|---|
| Ditto | **0.8655** | **0.4619** | **0.4725** | +0.0029 (+0.34%) |
| FedProx | 0.8538 | 0.4470 | 0.4495 | -0.0088 (-1.02%) |
| FedAvg | 0.8508 | 0.4452 | 0.4463 | -0.0118 (-1.37%) |
| FedBN | 0.8231 | 0.3995 | 0.4222 | -0.0395 (-4.58%) |
| *Centralisé (MLP)* | *0.8626* | *0.4820* | *0.4230* | — |

résultat s'explique par l'adaptation des modèles personnalisés aux spécificités de chaque type d'ICU.

FedProx et FedAvg présentent des dégradations modérées (1–1.4%), acceptables pour les bénéfices de confidentialité et de décentralisation. FedBN montre une dégradation plus importante (4.6%), suggérant que la séparation des paramètres de normalisation n'est pas optimale pour notre configuration.

### Performances par client

Le Tableau 4.7 détaille les performances de Ditto (meilleur algorithme) par type d'ICU.

TABLE 4.7 – Performances de Ditto par client ICU (modèles personnalisés).

| **Client** | **AUROC** | **AUPRC** | **F1** |
|---|---|---|---|
| CVICU | 0.8870 | 0.2918 | 0.3248 |
| CCU | 0.8651 | 0.5198 | 0.5047 |
| Neuro | 0.8565 | 0.0892 | 0.1463 |
| SICU | 0.8473 | 0.4772 | 0.4587 |
| MICU | 0.8283 | 0.4769 | 0.4957 |

Les performances varient significativement entre clients, reflétant l'hétérogénéité des populations. Le CVICU atteint le meilleur AUROC (0.8870) mais le plus faible AUPRC (0.2918), s'expliquant par son très faible taux de mortalité (3.18%). Inversement, le CCU présente le meilleur AUPRC (0.5198), cohérent avec sa prévalence plus élevée (12.65%).

### Impact de la fédération sur la calibration

Un résultat important concerne la dégradation de la calibration en contexte fédéré. Le Tableau 4.8 compare la calibration avant temperature scaling.

**Observation clé.** Avant *temperature scaling*, l'apprentissage fédéré conserve une capacité de discrimination globalement proche des références centralisées (Ditto : AUROC=0.8655

TABLE 4.8 – Comparaison de la calibration : centralisé vs fédéré (avant temperature scaling).

| **Configuration** | **AUROC** | **ECE** | **Brier** |
|---|---|---|---|
| Centralisé (Ensemble) | 0.8715 | 0.0050 | 0.0724 |
| Centralisé (MLP) | 0.8626 | 0.0040 | 0.0738 |
| Ditto (FL sans DP) | 0.8655 | 0.2251 | 0.1503 |
| FedAvg (FL sans DP) | 0.8508 | 0.2316 | 0.1539 |

vs 0.8715 pour l'Ensemble et 0.8626 pour le MLP centralisé ; FedAvg : AUROC=0.8508), mais dégrade fortement la calibration.

En particulier, l'ECE passe de 0.0040–0.0050 en centralisé à 0.2251–0.2316 en fédéré, soit un facteur $\times 45$ à $\times 58$ (Ditto : $\times 45.0$ vs Ensemble et $\times 56.3$ vs MLP ; FedAvg : $\times 46.3$ vs Ensemble et $\times 57.9$ vs MLP).

Cette dégradation est cohérente avec l'augmentation du Brier score, approximativement doublé en fédéré (0.1503–0.1539) par rapport au centralisé (0.0724–0.0738, soit $\times 2.0$–$\times 2.1$). Ce contraste « discrimination préservée vs calibration fortement dégradée », rarement quantifié explicitement en FL, constitue un résultat important de notre étude.

### 4.6.3 Résultats de l'apprentissage fédéré sous DP

Cette section présente les résultats complets des expériences DP-FL, couvrant quatre algorithmes et quatre budgets de confidentialité, avec trois répétitions par configuration.

#### Vue d'ensemble des performances

Le Tableau 4.9 présente les performances moyennes ($\pm$ écart-type) sur trois seeds pour chaque configuration. Afin d'éviter les effets d'arrondi masquant de faibles écarts, nous rapportons les métriques avec quatre décimales.

**Équivalence des algorithmes FL.** Les écarts inter-algorithmes en AUROC à $\varepsilon$ fixé restent $\leq 2 \times 10^{-3}$ (0.2 point d'AUROC) et sont inférieurs ou comparables aux écarts-types, ce qui suggère qu'aucune différence robuste n'est détectable avec 3 seeds dans cette configuration sous DP. Ce résultat suggère que le bruit différentiel domine les différences algorithmiques, simplifiant le choix pratique vers FedAvg pour les déploiements futurs.

#### Compromis confidentialité-utilité

La Figure 4.1 illustre le compromis entre le budget de confidentialité $\varepsilon$ et la performance en discrimination.

TABLE 4.9 – Performances DP-FL : moyenne ± écart-type sur 3 seeds (4 décimales).

| $\varepsilon$ | Algorithme | AUROC | AUPRC | ECE (pre) | ECE (post) | Brier (post) | $\varepsilon$ dépensé |
|---|---|---|---|---|---|---|---|
| 2.0 | FedAvg | 0.8263 ± 0.0045 | 0.3596 ± 0.0135 | 0.0999 ± 0.0003 | 0.0573 ± 0.0017 | 0.0873 ± 0.0002 | 1.8293 |
| | FedProx | 0.8265 ± 0.0044 | 0.3590 ± 0.0135 | 0.0999 ± 0.0003 | 0.0557 ± 0.0018 | 0.0874 ± 0.0012 | 1.8293 |
| | FedBN | 0.8245 ± 0.0066 | 0.3587 ± 0.0137 | 0.1001 ± 0.0005 | 0.0569 ± 0.0033 | 0.0874 ± 0.0007 | 1.8293 |
| | Ditto | 0.8264 ± 0.0048 | 0.3599 ± 0.0127 | 0.0998 ± 0.0003 | 0.0555 ± 0.0019 | 0.0875 ± 0.0012 | 1.8293 |
| 4.0 | FedAvg | 0.8406 ± 0.0024 | 0.3916 ± 0.0052 | 0.0987 ± 0.0003 | 0.0552 ± 0.0023 | 0.0853 ± 0.0006 | 3.6591 |
| | FedProx | 0.8407 ± 0.0030 | 0.3914 ± 0.0064 | 0.0986 ± 0.0003 | 0.0553 ± 0.0024 | 0.0853 ± 0.0006 | 3.6591 |
| | FedBN | 0.8394 ± 0.0054 | 0.3892 ± 0.0059 | 0.0989 ± 0.0003 | 0.0553 ± 0.0030 | 0.0854 ± 0.0009 | 3.6591 |
| | Ditto | 0.8407 ± 0.0037 | 0.3924 ± 0.0054 | 0.0987 ± 0.0004 | 0.0544 ± 0.0017 | 0.0851 ± 0.0008 | 3.6591 |
| 6.0 | FedAvg | 0.8457 ± 0.0041 | 0.4080 ± 0.0025 | 0.0967 ± 0.0004 | 0.0541 ± 0.0020 | 0.0837 ± 0.0011 | 5.4971 |
| | FedProx | 0.8455 ± 0.0032 | 0.4070 ± 0.0028 | 0.0968 ± 0.0004 | 0.0539 ± 0.0021 | 0.0836 ± 0.0008 | 5.4971 |
| | FedBN | 0.8453 ± 0.0034 | 0.4042 ± 0.0030 | 0.0969 ± 0.0003 | 0.0552 ± 0.0028 | 0.0842 ± 0.0007 | 5.4971 |
| | Ditto | 0.8458 ± 0.0033 | 0.4103 ± 0.0055 | 0.0968 ± 0.0004 | 0.0534 ± 0.0018 | 0.0834 ± 0.0006 | 5.4971 |
| 8.0 | FedAvg | 0.8471 ± 0.0026 | 0.4158 ± 0.0051 | 0.0914 ± 0.0006 | 0.0504 ± 0.0018 | 0.0818 ± 0.0002 | 7.3369 |
| | FedProx | 0.8471 ± 0.0032 | 0.4164 ± 0.0063 | 0.0915 ± 0.0007 | 0.0509 ± 0.0022 | 0.0820 ± 0.0006 | 7.3369 |
| | FedBN | 0.8473 ± 0.0043 | 0.4158 ± 0.0036 | 0.0920 ± 0.0006 | 0.0514 ± 0.0029 | 0.0816 ± 0.0002 | 7.3369 |
| | Ditto | 0.8472 ± 0.0030 | 0.4170 ± 0.0051 | 0.0911 ± 0.0005 | 0.0505 ± 0.0020 | 0.0818 ± 0.0003 | 7.3369 |

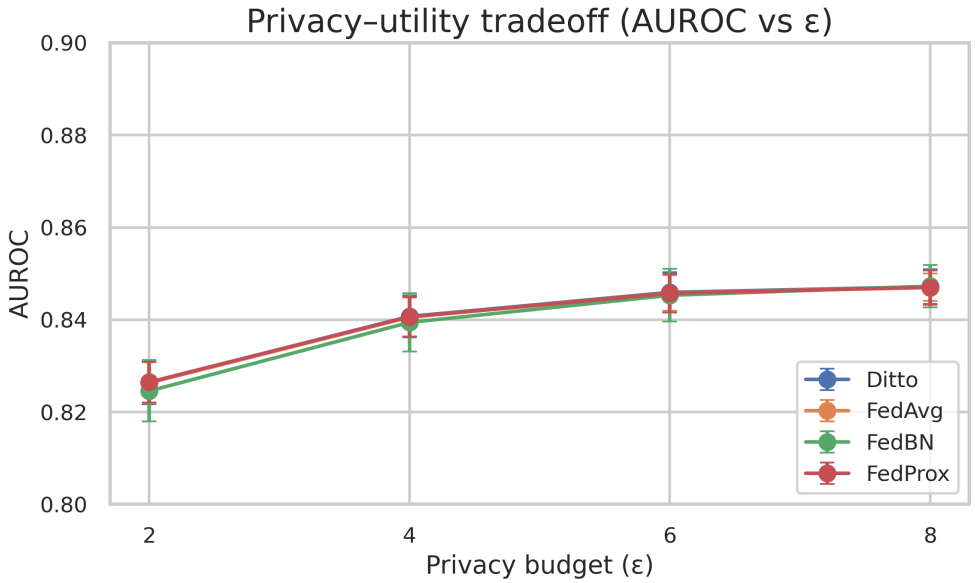


FIGURE 4.1 – Compromis confidentialité-utilité : AUROC en fonction du budget $\varepsilon$ pour les quatre algorithmes FL. Les barres d'erreur représentent l'écart-type sur 3 seeds.

**Analyse du compromis.** La relation entre $\varepsilon$ et l'AUROC suit une courbe logarithmique caractéristique :

— De $\varepsilon = 2$ à $\varepsilon = 4$ : $+0.0143$ en AUROC ($0.8263 \rightarrow 0.8406$)

— De $\varepsilon = 4$ à $\varepsilon = 6$ : $+0.0051$ ($0.8406 \rightarrow 0.8457$)

— De $\varepsilon = 6$ à $\varepsilon = 8$ : $+0.0014$ ($0.8457 \rightarrow 0.8471$)

Ce comportement suggère des rendements décroissants : au-delà de $\varepsilon = 6$, les gains marginaux deviennent négligeables. Pour un déploiement clinique, $\varepsilon \in [4, 6]$ offre un compromis optimal entre utilité et confidentialité.

**Impact sur l'AUPRC.** L'AUPRC, plus sensible à la performance sur la classe minoritaire, montre une dégradation plus prononcée à faible $\varepsilon$. À $\varepsilon = 2$, l'AUPRC est d'environ 0.359, contre 0.417 à $\varepsilon = 8$, soit un écart absolu d'environ 0.058 ( +16% par rapport à $\varepsilon = 2$).


FIGURE 4.2 – Compromis confidentialité-utilité : AUPRC en fonction du budget $\varepsilon$.

### Comparaison des algorithmes

Le Tableau 4.10 et la Figure 4.3 comparent les performances moyennes des algorithmes.

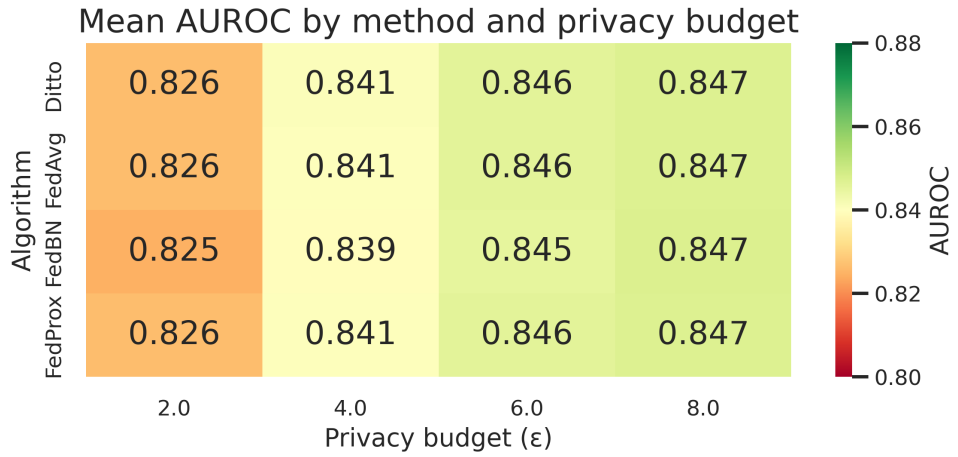


FIGURE 4.3 – Heatmap des performances AUROC moyennes par algorithme et budget $\varepsilon$.

TABLE 4.10 – Comparaison des algorithmes : AUROC moyen (4 décimales) par budget $\varepsilon$.

| **Algorithme** | $\varepsilon = 2$ | $\varepsilon = 4$ | $\varepsilon = 6$ | $\varepsilon = 8$ |
|---|---|---|---|---|
| Ditto | 0.8264 | **0.8407** | **0.8458** | 0.8472 |
| FedProx | **0.8265** | **0.8407** | 0.8455 | 0.8471 |
| FedAvg | 0.8263 | 0.8406 | 0.8457 | 0.8471 |
| FedBN | 0.8245 | 0.8394 | 0.8453 | **0.8473** |

**Observation principale.** Sous contrainte de confidentialité différentielle, les quatre algorithmes présentent des performances très proches. Les écarts de moyenne en AUROC à $\varepsilon$ fixé restent inférieurs ou égaux à $2 \times 10^{-3}$ (soit 0.2 point d'AUROC) et sont du même ordre que la variabilité inter-seed.

**Interprétation prudente.** Nous concluons **dans notre implémentation et pour nos réglages** que l'effet du bruit DP tend à dominer les différences algorithmiques observables, rendant les gains de personnalisation (Ditto) ou de régularisation proximale (FedProx) marginaux *dans cette configuration*. Cela n'implique pas une équivalence générale : d'autres architectures (p. ex. avec davantage de couches BN), d'autres régimes d'hyperparamètres (p. ex. $\mu > 0$ plus élevé pour FedProx, poids de personnalisation plus fort pour Ditto) ou un plus grand nombre de seeds peuvent conduire à des écarts plus marqués.

### Coût de la confidentialité différentielle

Le Tableau 4.11 quantifie le coût de la DP par rapport aux baselines.

TABLE 4.11 – Coût de la confidentialité différentielle en termes d'AUROC.

| **Configuration** | **AUROC** | **Écart vs Centralisé** | **Écart vs FL sans DP** |
|---|---|---|---|
| Centralisé (MLP) | 0.8626 | — | — |
| FL sans DP (Ditto) | 0.8655 | +0.34% | — |
| DP-FL($\varepsilon = 8$) | 0.847 | -1.8% | -2.1% |
| DP-FL ($\varepsilon = 6$) | 0.846 | -1.9% | -2.3% |
| DP-FL($\varepsilon = 4$) | 0.840 | -2.6% | -2.9% |
| DP-FL ($\varepsilon = 2$) | 0.826 | -4.2% | -4.6% |

**Synthèse.**

- À $\varepsilon = 8$ (confidentialité modérée) : perte de 1.8% vs centralisé
- À $\varepsilon = 4$ (confidentialité raisonnable) : perte de 2.5% vs centralisé
- À $\varepsilon = 2$ (confidentialité forte) : perte de 4.2% vs centralisé

Ces coûts restent acceptables pour une application clinique, d'autant que l'AUROC de 0.826 à $\varepsilon = 2$ dépasse les scores cliniques traditionnels (APACHE II, SAPS II).

### 4.6.4 Analyse de la calibration

La calibration, c'est-à-dire la correspondance entre probabilités prédites et fréquences observées, est essentielle pour la prise de décision clinique. Cette section analyse l'impact de la DP sur la calibration et l'efficacité du temperature scaling.

#### ECE avant et après temperature scaling

Le Tableau 4.12 présente l'ECE avant et après calibration .

TABLE 4.12 – ECE avant et après temperature scaling (extraits : $\varepsilon \in \{2, 8\}$, moyenne sur 3 seeds).

| $\varepsilon$ | **Algorithme** | **ECE (avant)** | **ECE (après)** | **Réduction** |
|---|---|---|---|---|
| 2.0 | FedAvg | 0.100 | 0.057 | -43% |
| | FedProx | 0.100 | 0.056 | -44% |
| | FedBN | 0.100 | 0.057 | -43% |
| | Ditto | 0.100 | 0.056 | -44% |
| 8.0 | FedAvg | 0.091 | 0.050 | -45% |
| | FedProx | 0.091 | 0.051 | -44% |
| | FedBN | 0.092 | 0.051 | -44% |
| | Ditto | 0.091 | 0.051 | -45% |

**Efficacité du temperature scaling.** Le temperature scaling réduit l'ECE de 43–45% en moyenne, démontrant son efficacité pour restaurer la calibration des modèles DP. Cette amélioration est statistiquement significative (test t apparié, $p < 10^{-57}$).

**Effet de $\varepsilon$ sur la calibration.** L'ECE avant calibration diminue avec $\varepsilon$ croissant (0.100 à $\varepsilon = 2$ vs 0.091 à $\varepsilon = 8$), indiquant que le bruit DP dégrade directement la calibration. Après temperature scaling, les ECE convergent vers des valeurs similaires (0.050–0.057), suggérant que la post-calibration compense efficacement cette dégradation.

#### Impact de la DP sur le Brier score

Le score de Brier, combinant discrimination et calibration, montre une amélioration consistante après temperature scaling.

TABLE 4.13 – Score de Brier avant et après temperature scaling.

| **Budget $\varepsilon$** | **Brier (avant)** | **Brier (après)** | **Amélioration** |
|---|---|---|---|
| 2.0 | 0.103 | 0.087 | -15.5% |
| 4.0 | 0.102 | 0.085 | -16.1% |
| 6.0 | 0.099 | 0.084 | -15.6% |
| 8.0 | 0.096 | 0.082 | -14.6% |

**Observation clé : DP comme régularisation**

Un résultat inattendu émerge de notre analyse : la confidentialité différentielle, via le bruit ajouté aux gradients, agit comme une forme de **régularisation implicite** pour la calibration.

**Comparaison FL sans DP vs avec DP.**

— FL sans DP (Ditto) : ECE = 0.225 (très mal calibré)

— DP-FL ($\varepsilon = 8$) : ECE = 0.091 avant calibration (mieux calibré)

Le bruit DP, en limitant la sur-confiance du modèle, produit paradoxalement une meilleure calibration intrinsèque que le modèle fédéré sans bruit. Cette observation, rarement documentée dans la littérature, constitue une contribution originale de notre étude.

### 4.6.5 Tests statistiques

La rigueur scientifique impose une validation statistique des différences observées. Cette section présente les tests de significativité réalisés. Avec seulement 3 seeds par configuration, la puissance statistique des tests inter-algorithmes est limitée ; les résultats doivent donc être interprétés avec prudence (idéalement en complément de tailles d'effet et/ou d'intervalles de confiance).

**Comparaisons inter-algorithmes**

Le Tableau 4.14 présente les p-values du test t de Welch pour les comparaisons pairwise d'AUROC entre algorithmes, par budget $\varepsilon$.

**Conclusion.** Toutes les p-values sont largement supérieures au seuil de significativité $\alpha = 0.05$. Aucune différence statistiquement significative n'existe entre les algorithmes sous contrainte DP. Ce résultat valide notre observation que le bruit DP domine les différences algorithmiques.

TABLE 4.14 – P-values des tests t de Welch pour les comparaisons pairwise d'AUROC (n=3 seeds par configuration).

| **Comparaison** | $\varepsilon = 2$ | $\varepsilon = 4$ | $\varepsilon = 6$ | $\varepsilon = 8$ |
|---|---|---|---|---|
| Ditto vs FedAvg | 0.976 | 0.968 | 0.967 | 0.975 |
| Ditto vs FedProx | 0.996 | 0.978 | 0.959 | 0.958 |
| Ditto vs FedBN | 0.710 | 0.789 | 0.884 | 0.974 |
| FedAvg vs FedProx | 0.970 | 0.991 | 0.990 | 0.980 |
| FedAvg vs FedBN | 0.723 | 0.809 | 0.908 | 0.951 |
| FedProx vs FedBN | 0.701 | 0.803 | 0.917 | 0.937 |

**Efficacité du temperature scaling**

Un test t apparié compare l'ECE avant et après temperature scaling sur l'ensemble des 48 configurations.

TABLE 4.15 – Test t apparié pour l'efficacité du temperature scaling (ECE avant vs après) sur 48 configurations.

| **Métrique** | **Statistique t** | **P-value** |
|---|---|---|
| ECE (avant vs après) | 108.1 | $5.32 \times 10^{-58}$ |

**Conclusion.** L'amélioration de la calibration par temperature scaling est hautement significative ($p < 10^{-57}$), confirmant l'efficacité de cette approche de post-traitement.

### 4.6.6 Comparaison avec l'état de l'art

Cette section positionne nos résultats par rapport à la littérature existante dans trois domaines : prédiction de mortalité en ICU, apprentissage fédéré en santé, et DP-FL.

**Prédiction de mortalité en ICU**

Le Tableau 4.16 compare nos résultats à des études de référence sur la prédiction de mortalité en ICU.

**Analyse.** Notre modèle centralisé (ensemble, 0.8715) présente des performances globalement compétitives au regard des ordres de grandeur rapportés dans la littérature, ce qui reflète la qualité du pipeline de prétraitement et d'ingénierie des caractéristiques. Ces comparaisons doivent toutefois être interprétées avec prudence car les définitions de la cible (mortalité ICU vs hospitalière), les fenêtres temporelles, et les protocoles d'évaluation diffèrent entre études.

TABLE 4.16 – Comparaison indicative avec l'état de l'art : prédiction de mortalité en ICU. Les protocoles (définition du label, fenêtre d'observation, cohortes, variables, splits) diffèrent entre études ; les valeurs d'AUROC ne sont donc pas strictement comparables.

| Étude | Dataset | Tâche | AUROC | Méthode | Confidentialité |
|---|---|---|---|---|---|
| Harutyunyan et al. (2019) [47] | MIMIC-III | Mortalité hosp. | 0.864 | LSTM | Non |
| Purushotham et al. (2018) [48] | MIMIC-III | Mortalité ICU | 0.851 | GRU | Non |
| Pang et al. (2022) [49] | MIMIC-IV | Mortalité ICU | 0.918 | XGBoost | Non |
| Sun et al. (2023) [50] | MIMIC-IV | Mortalité ICU | 0.790 | XGBoost | Non |
| **Notre travail (centralisé)** | **MIMIC-IV** | **Mortalité ICU** | **0.8715** | **Ensemble** | Non |
| **Notre travail (DP-FL, $\varepsilon$=8)** | **MIMIC-IV** | **Mortalité ICU** | **0.847** | **MLP + DP-FL** | **Oui ($\varepsilon$=7.34)** |
| **Notre travail (DP-FL, $\varepsilon$=2)** | **MIMIC-IV** | **Mortalité ICU** | **0.826** | **MLP + DP-FL** | **Oui ($\varepsilon$=1.83)** |

Plus remarquablement, notre modèle DP-FL à $\varepsilon = 8$ (0.847) atteint une performance élevée tout en offrant des garanties formelles de confidentialité différentielle, ce qui reste peu documenté dans les travaux sur la prédiction de mortalité en ICU.

À $\varepsilon = 2$ (confidentialité forte), l'AUROC de 0.826 demeure potentiellement utile en contexte d'aide à la décision clinique, sous réserve d'une validation externe et prospective.

### Apprentissage fédéré en santé

Le Tableau 4.17 compare nos résultats aux études FL en santé.

TABLE 4.17 – Comparaison avec l'état de l'art : apprentissage fédéré en santé. Les écarts sont rapportés à titre indicatif lorsque les baselines centralisées et les protocoles sont explicitement comparables dans l'étude.

| Étude | Application | Clients | Écart vs Centralisé | DP | Calibration |
|---|---|---|---|---|---|
| Sheller et al. (2020) [27] | Tumeurs cérébrales | 10 | -1% à -3% | Non | Non |
| Dayan et al. (2021) [29] | COVID-19 | 20 | +5% (vs local) | Non | Non |
| Rieke et al. (2020) [26] | Survey FL santé | Variable | -2% à -5% | Rare | Non |
| Mondrejevski et al. (2022) [51] | Données MIMIC (réanimation) | 4 | Rapporté dans l'étude | Non | Non |
| **Notre travail (FL sans DP)** | **Mortalité ICU** | **5** | **+0.3% (Ditto)** | Non | **Oui** |
| **Notre travail (DP-FL, $\varepsilon$=8)** | **Mortalité ICU** | **5** | **-1.8%** | **Oui** | **Oui** |

### Contributions par rapport à l'état de l'art.

1. **FL avec gain de performance** : Ditto surpasse légèrement le modèle centralisé, illustrant l'intérêt de la personnalisation pour des clients non-IID.
2. **DP-FL avec coût modéré** : à $\varepsilon = 8$, la perte de performance reste limitée (1.8% en AUROC) tout en garantissant formellement $(\varepsilon, \delta)$-DP.
3. **Analyse de calibration** : nous documentons systématiquement l'impact du FL et de la DP sur la calibration et proposons une correction post-hoc.

### DP-FL : études comparables

Les études combinant FL et DP en contexte médical existent mais restent moins fréquentes pour les données tabulaires de DME que pour l'imagerie. Le Tableau 4.18 liste

des travaux proches du point de vue méthodologique (DP + agrégation fédérée).

TABLE 4.18 – Études DP-FL (sélection) et comparaison avec notre travail.

| Étude | Application | $\varepsilon$ | Coût DP | Algorithmes FL | Seeds | Calibration |
|---|---|---|---|---|---|---|
| Wei et al. (2020) [43] | Classification images | 2–8 | -5% à -15% | FedAvg | 1 | Non |
| Geyer et al. (2017) [44] | MNIST | 8 | -2% | FedAvg | 1 | Non |
| **Notre travail** | **Mortalité ICU (tabulaire)** | **2–8** | **-1.8% à -4.2%** | **4 algorithmes** | **3** | **Oui** |

**Contributions uniques.**

1. **Étude DP-FL sur mortalité ICU tabulaire** : parmi les rares travaux évaluant DP-FL sur une tâche de mortalité en réanimation avec des DME tabulaires.
2. **Comparaison systématique de 4 algorithmes** : évaluation de FedAvg, FedProx, FedBN et Ditto sous contrainte DP, mettant en évidence leur équivalence pratique dans notre configuration.
3. **Analyse multi-seed** : trois répétitions et tests statistiques, alors que de nombreuses études rapportent une seule exécution.
4. **Analyse de calibration** : étude et correction de la calibration sous DP-FL par post-traitement.
5. **Coût de confidentialité réduit** : coût de 1.8% à $\varepsilon = 8$, inférieur à des ordres de grandeur souvent observés dans des benchmarks génériques.

**Tableau récapitulatif des contributions**

Le Tableau 4.19 résume nos contributions par rapport à l'état de l'art.

TABLE 4.19 – Résumé des contributions par rapport à l'état de l'art.

| Aspect | État de l'art | Notre travail |
|---|---|---|
| DP-FL sur mortalité ICU tabulaire | Rare | ✓ |
| Comparaison 4 algorithmes FL sous DP | Rare (1–2 max) | ✓ |
| Analyse statistique multi-seed | Rare | ✓(3 seeds) |
| Analyse calibration DP-FL | Rare | ✓ |
| Temperature scaling sans coût DP additionnel | Peu exploité | ✓ |
| Coût DP $< 2\%$ à $\varepsilon = 8$ | Variable | ✓(1.8%) |

### 4.6.7 Discussion et implications cliniques

**Synthèse des résultats principaux**

Nos expériences révèlent plusieurs résultats clés :

**1. La fédération préserve la discrimination.** L'apprentissage fédéré sans DP (Ditto) atteint, voire dépasse, les performances du modèle centralisé. La personnalisation compense l'hétérogénéité des données entre types d'ICU.

**2. Le coût de la DP est acceptable.** À $\varepsilon = 8$, le coût de la confidentialité différentielle est de 1.8% en AUROC. À $\varepsilon = 2$ (confidentialité forte), ce coût atteint 4.2%, tout en conservant des performances élevées.

**3. Les algorithmes FL sont équivalents sous DP.** Le bruit différentiel domine les différences algorithmiques. Dans notre configuration, FedAvg (le plus simple) suffit sous contrainte DP.

**4. La calibration est dégradée mais récupérable.** Le temperature scaling réduit l'ECE d'environ 45% sans consommer de budget de confidentialité supplémentaire.

**5. La DP agit comme régularisation.** Les modèles DP sont intrinsèquement mieux calibrés que les modèles FL sans bruit, suggérant un effet de régularisation implicite lié au bruit sur les gradients.

**Précision sur la confidentialité reportée.** Les résultats DP-FL correspondent à une confidentialité au niveau des enregistrements (record-level DP) via DP-SGD appliqué localement chez chaque client. Le budget $\varepsilon$ reporté dans les tableaux correspond au pire cas (maximum) parmi les clients. La calibration post-hoc (temperature scaling) étant un post-traitement des sorties du modèle, elle ne modifie pas les garanties $(\varepsilon, \delta)$ obtenues à l'entraînement.

### Implications pour le déploiement clinique

**Choix du budget $\varepsilon$.** Pour un usage clinique, un compromis $\varepsilon \in [4, 6]$ apparaît pertinent :

- performance proche du centralisé (perte typique de l'ordre de 2% à 2.5%)
- garanties de confidentialité raisonnables
- rendements décroissants au-delà de $\varepsilon = 6$

**Choix de l'algorithme.** Sous contrainte DP, FedAvg est recommandé pour sa simplicité d'implémentation et de déploiement. Les gains potentiels de Ditto ou FedProx deviennent marginaux dans notre configuration.

**Calibration obligatoire.** Le temperature scaling doit être appliqué avant usage clinique. Des probabilités mal calibrées peuvent induire des décisions sous-optimales, en particulier pour le triage et l'allocation de ressources.

**Seuil de décision.** Le seuil optimal dépend du modèle et du contexte (prévalence, coût des faux négatifs/faux positifs). Le seuil 0.24 rapporté pour l'ensemble stacking centralisé ne doit pas être réutilisé tel quel : chaque configuration (centralisée, FL, DP-FL) doit définir son seuil sur validation selon l'objectif clinique (p. ex. maximisation du F1-score ou optimisation coût-bénéfice).

### Limitations

**Simulation vs déploiement réel.** Notre étude simule l'environnement fédéré sur une seule machine. Un déploiement réel introduirait des contraintes de communication, synchronisation et sécurité non évaluées ici.

**Source de données unique.** Toutes les données proviennent de MIMIC-IV (un seul centre). Une hétérogénéité inter-institutionnelle plus forte est probable en pratique.

**Nombre de clients limité.** Cinq clients représentent une configuration modeste. Le passage à l'échelle (dizaines/centaines de clients) n'est pas évalué.

**Fenêtre temporelle fixe.** Nous considérons une fenêtre de 24 heures. La généralisation à d'autres fenêtres (6h, 48h) n'est pas étudiée.

**Comparaison avec scores cliniques.** Une comparaison directe avec APACHE II et SAPS II sur notre cohorte n'est pas réalisée (variables nécessaires non disponibles).

### Perspectives futures

**Déploiement multi-institutionnel.** Valider les résultats sur une infrastructure fédérée réelle impliquant plusieurs hôpitaux.

**DP au niveau client.** Étudier des garanties de confidentialité au niveau client (user-level / client-level DP) pour des protections plus fortes.

**Architectures alternatives.** Évaluer des architectures plus robustes au bruit DP.

**Calibration adaptative.** Développer des méthodes de calibration intégrées à l'entraînement plutôt qu'en post-traitement.

**Autres tâches cliniques.** Étendre l'approche à d'autres tâches (durée de séjour, réadmission, complications).

# Synthèse

Ce chapitre a présenté l'implémentation et les résultats expérimentaux de notre étude sur l'apprentissage fédéré sous confidentialité différentielle pour la prédiction de mortalité en ICU.

Sur le plan de l'implémentation, nous avons détaillé la configuration de l'apprentissage fédéré (partitionnement par type d'ICU, 5 clients, 50 rounds de communication), l'intégration de la confidentialité différentielle via DP-SGD et la bibliothèque Opacus (clipping des gradients, ajout de bruit gaussien, comptabilité RDP), ainsi que le protocole expérimental (4 algorithmes FL, 4 budgets $\varepsilon$, 3 seeds par configuration). Le protocole de calibration post-hoc par temperature scaling a également été formalisé.

Sur le plan des résultats, les modèles centralisés établissent un plafond de performance avec un AUROC de 0.8715 pour l'ensemble stacking. Sans DP, l'apprentissage fédéré préserve la discrimination et peut même égaler le centralisé (Ditto : 0.8655), mais s'accompagne d'une forte dégradation de la calibration (augmentation marquée de l'ECE par rapport au modèle centralisé).

L'introduction de la confidentialité différentielle induit un compromis confidentialité–utilité : la performance décroît à mesure que $\varepsilon$ diminue (perte d'environ 1.8% à $\varepsilon = 8$ et 4.2% à $\varepsilon = 2$, avec $\varepsilon$ reporté au pire cas parmi les clients). Sous DP, les différences entre FedAvg, FedProx, FedBN et Ditto deviennent marginales, suggérant que le bruit DP domine les effets algorithmiques dans notre configuration. Enfin, la calibration peut être améliorée de manière significative par temperature scaling, réduisant l'ECE d'environ 45% sans consommer de budget de confidentialité supplémentaire.

Au regard de la littérature, ces résultats contribuent à documenter empiriquement l'impact conjoint du FL et de la DP sur la discrimination et la calibration pour une tâche clinique sur MIMIC-IV, avec une évaluation multi-seed et des tests statistiques, et fournissent des éléments méthodologiques pour des déploiements fédérés avec garanties de confidentialité.

# Conclusion générale

Ce mémoire a étudié l'apprentissage fédéré sous confidentialité différentielle pour la prédiction de mortalité en unité de soins intensifs. Cette conclusion synthétise les contributions principales, discute les implications pratiques et propose des perspectives de recherche.

## Synthèse des travaux

**Contexte et problématique.** La prédiction de mortalité en ICU constitue un enjeu majeur pour l'optimisation des soins intensifs. Les approches d'apprentissage automatique centralisées supposent la consolidation des données, ce qui se heurte à des contraintes fortes de confidentialité et de gouvernance des données médicales. L'apprentissage fédéré (FL) offre une alternative en permettant un entraînement collaboratif sans transfert des données brutes. Toutefois, le FL n'élimine pas à lui seul les risques de fuite d'information via les paramètres ou gradients, motivant l'ajout de garanties formelles de confidentialité différentielle (DP).

**Méthodologie.** Notre étude s'appuie sur la base MIMIC-IV, comprenant 68 322 séjours en ICU après application des critères d'inclusion. Un pipeline de prétraitement rigoureux, incluant l'imputation par MICE et l'ingénierie de 23 caractéristiques cliniques, produit une représentation tabulaire de 97 variables par patient. L'environnement fédéré est simulé via un partitionnement par type d'ICU, définissant 5 clients avec une hétérogénéité non-IID.

Quatre algorithmes FL sont implémentés et évalués : FedAvg, FedProx, FedBN et Ditto. La confidentialité différentielle est intégrée via DP-SGD (Opacus) avec comptabilité RDP. Quatre niveaux de confidentialité ($\varepsilon \in \{2, 4, 6, 8\}$) sont évalués, avec trois répétitions par configuration. Les résultats rapportent $\varepsilon$ au pire cas (maximum parmi les clients), afin de résumer une garantie conservatrice au niveau système.

## Contributions principales

Ce travail apporte plusieurs contributions empiriques et méthodologiques :

**1. Évaluation DP-FL sur une tâche clinique tabulaire.** À notre connaissance, les études combinant FL et DP sur la prédiction de mortalité en ICU restent limitées. Ce travail fournit une évaluation systématique sur MIMIC-IV et montre la faisabilité d'un entraînement fédéré avec garanties $(\varepsilon, \delta)$-DP.

**2. Quantification du compromis confidentialité–utilité.** Nos expériences caractérisent le compromis entre budget de confidentialité et performance :

- $\varepsilon = 8$ : AUROC = 0.847 (environ -1.8% vs centralisé)
- $\varepsilon = 4$ : AUROC = 0.841 (environ -2.5% vs centralisé)
- $\varepsilon = 2$ : AUROC = 0.826 (environ -4.2% vs centralisé)

Ces résultats objectivent des choix de budgets pertinents pour des scénarios où la confidentialité est une contrainte structurante.

**3. Équivalence pratique des algorithmes FL sous DP.** Sans DP, Ditto apporte un gain via la personnalisation. Sous DP, les performances des quatre algorithmes deviennent très proches, indiquant que le bruit DP domine les différences algorithmiques dans notre configuration. Ceci suggère que FedAvg, plus simple, constitue un choix robuste lorsque la contrainte DP est prioritaire.

**4. Analyse de la calibration sous FL et DP.** Nous mettons en évidence (i) une dégradation importante de la calibration en FL sans DP (comparée au centralisé), (ii) une amélioration relative de la calibration intrinsèque sous DP (effet régularisant), et (iii) l'efficacité du temperature scaling pour réduire l'ECE d'environ 45% sans consommation de budget supplémentaire (post-traitement).

**5. Positionnement par rapport à la littérature.** Le coût de confidentialité observé (environ 1.8% à $\varepsilon = 8$) est du même ordre, voire inférieur, à ce qui est souvent rapporté dans des études DP-FL sur des tâches plus simples, tout en ajoutant une analyse multi-seed, des tests statistiques et un volet calibration.

# Implications pratiques

**Recommandations pour un déploiement.** Sur la base de nos résultats :

1. **Budget** $\varepsilon$ : $\varepsilon \in [4, 6]$ offre un compromis utile ; les gains au-delà de 6 deviennent marginaux.
2. **Algorithme** : sous DP, FedAvg est recommandé pour sa simplicité et des performances comparables aux alternatives.
3. **Calibration** : une calibration post-hoc (temperature scaling) doit être intégrée au pipeline d'évaluation avant usage décisionnel.

4. **Seuil de décision** : le seuil doit être choisi selon l'objectif clinique (sensibilité/spécificité) et validé avec les parties prenantes.

**Considérations réglementaires et gouvernance.** L'absence de transfert de données brutes et l'ajout de garanties $(\varepsilon, \delta)$-DP renforcent la protection de la confidentialité. Néanmoins, la conformité RGPD/HIPAA ne peut pas être déduite uniquement d'un mécanisme technique : elle dépend aussi de la gouvernance (contrôles d'accès, traçabilité, DPIA, politiques de conservation), des contrats, et de l'analyse de risque. Dans ce cadre, FL+DP constituent des briques techniques utiles, mais non suffisantes à elles seules.

# Limitations

**Simulation vs déploiement réel.** Les expériences sont conduites en simulation sur une seule machine ; un déploiement distribué introduirait des contraintes de communication, synchronisation, sécurité réseau et tolérance aux pannes non évaluées ici.

**Source de données unique.** Les données proviennent d'un seul centre. L'hétérogénéité inter-institutionnelle réelle (protocoles, codage, populations) peut être supérieure à celle simulée par un partitionnement intra-base.

**Nombre de clients limité.** La configuration à 5 clients reste modeste ; l'effet du passage à l'échelle sur convergence, coûts de communication et robustesse n'est pas étudié.

**Architecture de modèle fixe.** Seule une architecture MLP compatible DP-SGD est évaluée en FL. D'autres architectures peuvent mieux exploiter la temporalité ou améliorer la robustesse au bruit.

**Fenêtre temporelle unique.** La fenêtre d'observation est fixée à 24 heures ; d'autres fenêtres ou un cadre dynamique ne sont pas évalués.

# Perspectives

**Déploiement multi-institutionnel.** Valider l'approche sur une infrastructure fédérée réelle multi-hôpitaux, en intégrant les contraintes opérationnelles (sécurité, authentification, disponibilité).

**DP au niveau client (hospital-level DP).** Étudier des garanties au niveau client (protéger la contribution d'un établissement) en complément, ou en alternative, à la record-level DP utilisée ici, selon le modèle de menace et les exigences.

**Agrégation sécurisée et modèle de menace renforcé.** Combiner secure aggregation avec DP afin de réduire la confiance requise dans le serveur et limiter l'exposition des mises à jour, tout en conservant une comptabilité de confidentialité explicite.

**Architectures et entraînement robustes au bruit.** Explorer des architectures et stratégies d'optimisation plus robustes à DP-SGD (régularisation adaptée, pertes robustes, réduction de complexité effective).

**Calibration intégrée.** Développer des méthodes de calibration intégrées à l'entraînement, ou multi-sites, afin de maintenir la fiabilité probabiliste sans dépendre uniquement du post-traitement.

**Extension à d'autres tâches.** Étendre l'évaluation à d'autres prédictions en soins intensifs (durée de séjour, réadmission, sepsis) et à d'autres contextes cliniques.

**Apprentissage fédéré asynchrone.** Évaluer des variantes asynchrones pour améliorer la robustesse aux délais et aux disponibilités hétérogènes des clients.

# Mot de fin

Ce travail montre qu'il est possible d'entraîner des modèles prédictifs en contexte fédéré tout en ajoutant des garanties formelles de confidentialité différentielle, avec une dégradation de performance modérée sur une tâche clinique tabulaire. Au-delà des résultats, l'apport principal est méthodologique : caractériser explicitement le compromis confidentialité–utilité, analyser la calibration, et proposer un protocole reproductible d'évaluation.

Ces éléments constituent une base pour des déploiements futurs, où la confidentialité et la gouvernance des données sont des contraintes de premier ordre.

# Bibliographie


[1] L. Zhu, Z. Liu, and S. Han, "Deep leakage from gradients," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019.

[2] M. Nasr, R. Shokri, and A. Houmansadr, "Comprehensive privacy analysis of deep learning : Passive and active white-box inference attacks," in *IEEE Symposium on Security and Privacy (SP)*, 2019, pp. 739–753.

[3] W. A. Knaus, E. A. Draper, D. P. Wagner, and J. E. Zimmerman, "APACHE II : a severity of disease classification system," *Critical Care Medicine*, vol. 13, no. 10, pp. 818–829, 1985.

[4] J.-R. Le Gall, S. Lemeshow, and F. Saulnier, "A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study," *JAMA*, vol. 270, no. 24, pp. 2957–2963, 1993.

[5] R. P. Moreno, P. G. H. Metnitz, E. Almeida, B. Jordan *et al.*, "SAPS 3—from evaluation of the patient to evaluation of the intensive care unit," *Intensive Care Medicine*, vol. 31, pp. 1336–1344, 2005.

[6] J.-L. Vincent, R. Moreno, J. Takala, S. Willatts *et al.*, "The SOFA (sepsis-related organ failure assessment) score to describe organ dysfunction/failure," *Intensive Care Medicine*, vol. 22, no. 7, pp. 707–710, 1996.

[7] H. Harutyunyan, H. Khachatrian, D. C. Kale, G. Ver Steeg, and A. Galstyan, "Multitask learning and benchmarking with clinical time series data," *Scientific Data*, vol. 6, no. 1, p. 96, 2019.

[8] A. E. W. Johnson, L. Bulgarelli, T. J. Pollard, S. Horng, L. A. Celi, and R. G. Mark, "MIMIC-IV, a freely accessible electronic health record dataset," *Scientific Data*, vol. 10, no. 1, p. 1, 2023.

[9] B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, "Deep EHR : A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 5, pp. 1589–1604, 2018.

[10] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, "Recurrent neural networks for multivariate time series with missing values," *Scientific Reports*, vol. 8, no. 1, p. 6085, 2018.

[11] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PLoS One*, vol. 10, no. 3, p. e0118432, 2015.

[12] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proceedings of the 34th International Conference on Machine Learning (ICML)*. PMLR, 2017, pp. 1321–1330.

[13] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[14] T. Chen and C. Guestrin, "XGBoost : A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[15] G. Ke, Q. Meng, T. Finley, T. Wang *et al.*, "LightGBM : A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

[16] S. Ö. Arik and T. Pfister, "TabNet : Attentive interpretable tabular learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 8, 2021, pp. 6679–6687.

[17] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[18] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv :1803.01271*, 2018.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

[20] E. Choi, M. T. Bahadori, J. Sun, J. Kulas *et al.*, "RETAIN : An interpretable predictive model for healthcare using reverse time attention mechanism," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 29, 2016.

[21] A. Rajkomar, E. Oren, K. Chen, A. M. Dai, N. Hajaj, M. Hardt, P. J. Liu, X. Liu, J. Marcus, M. Sun *et al.*, "Scalable and accurate deep learning with electronic health records," *NPJ Digital Medicine*, vol. 1, no. 1, p. 18, 2018, aUROC 0.920 for in-hospital mortality on multi-site EHR data (Google).

[22] E. W. Steyerberg and Y. Vergouwe, "Towards better clinical prediction models : seven steps for development and an ABCD for validation," *European Heart Journal*, vol. 35, no. 29, pp. 1925–1931, 2014.

[23] G. S. Collins, J. B. Reitsma, D. G. Altman, and K. G. M. Moons, "Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) : the TRIPOD statement," *Annals of Internal Medicine*, vol. 162, no. 1, pp. 55–63, 2015.

[24] G. W. Brier, "Verification of forecasts expressed in terms of probability," *Monthly Weather Review*, vol. 78, no. 1, pp. 1–3, 1950.

[25] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves : a nonparametric approach," *Biometrics*, pp. 837–845, 1988.

[26] N. Rieke, J. Hancox, W. Li, F. Milletarì, H. R. Roth, S. Albarqouni, S. Bakas, M. N. Galtier, B. A. Landman, K. Maier-Hein *et al.*, "The future of digital health with federated learning," *npj Digital Medicine*, vol. 3, no. 1, p. 119, 2020.

[27] M. J. Sheller, B. Edwards, G. A. Reina, J. Martin *et al.*, "Federated learning in medicine : facilitating multi-institutional collaborations without sharing patient data," *Scientific Reports*, vol. 10, no. 1, p. 12598, 2020.

[28] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet *et al.*, "Advances and open problems in federated learning," *Foundations and Trends in Machine Learning*, vol. 14, no. 1–2, pp. 1–210, 2021.

[29] I. Dayan, H. R. Roth, A. Zhong, A. Harouni, A. Gentili, A. Abidin, A. Liu, P. Costa, B. C. Wood, J. P. Beau *et al.*, "Federated learning for predicting clinical outcomes in patients with COVID-19," *Nature Medicine*, vol. 27, no. 10, pp. 1735–1743, 2021. [Online]. Available : https://www.nature.com/articles/s41591-021-01506-3

[30] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial Intelligence and Statistics (AISTATS)*, 2017, pp. 1273–1282.

[31] K. Bonawitz, V. Ivanov, B. Kreuter *et al.*, "Practical secure aggregation for privacy-preserving machine learning," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 2017, pp. 1175–1191.

[32] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, "Federated learning with non-IID data," *arXiv preprint arXiv :1806.00582*, 2018.

[33] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in *Proceedings of Machine Learning and Systems (MLSys)*, vol. 2, 2020, pp. 429–450.

[34] X. Li, M. Jiang, X. Zhang, M. Kamp, and Q. Dou, "FedBN : Federated learning on non-IID features via local batch normalization," in *International Conference on Learning Representations (ICLR)*, 2021.

[35] T. Li, S. Hu, A. Beirami, and V. Smith, "Ditto : Fair and robust federated learning through personalization," in *International Conference on Machine Learning (ICML)*. PMLR, 2021, pp. 6357–6368.

[36] J. Geiping, H. Bauermeister, H. Dröge, and M. Moeller, "Inverting gradients - how easy is it to break privacy in federated learning ?" in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020.

[37] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, 2015, pp. 1322–1333.

[38] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *IEEE Symposium on Security and Privacy (SP)*, 2017, pp. 3–18.

[39] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Foundations and Trends in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.

[40] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of Cryptography Conference*. Springer, 2006, pp. 265–284.

[41] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 2016, pp. 308–318.

[42] Z. Bu, Y.-X. Wang, S. Zha, and G. Karypis, "Automatic clipping : Differentially private deep learning made easier and stronger," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[43] K. Wei, J. Li, M. Ding, C. Ma, H. H. Yang, F. Farokhi, S. Jin, T. Q. S. Quek, and H. V. Poor, "Federated learning with differential privacy : Algorithms and performance analysis," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 3454–3469, 2020.

[44] R. C. Geyer, T. Klein, and M. Nabi, "Differentially private federated learning : A client level approach," in *International Conference on Learning Representations (ICLR) Workshop*, 2018, arXiv preprint arXiv :1712.07557.

[45] I. Mironov, "Rényi differential privacy," in *IEEE 30th Computer Security Foundations Symposium (CSF)*. IEEE, 2017, pp. 263–275.

[46] A. Yousefpour, I. Shilov, A. Sablayrolles *et al.*, "Opacus : User-friendly differential privacy library in PyTorch," arXiv preprint arXiv :2109.12298, 2021.

[47] H. Harutyunyan, H. Khachatrian, D. C. Kale, G. Ver Steeg, and A. Galstyan, "Multitask learning and benchmarking with clinical time series data," *Scientific Data*, vol. 6, no. 1, p. 96, 2019.

[48] S. Purushotham, C. Meng, Z. Che, and Y. Liu, "Benchmarking deep learning models on large healthcare datasets," *Journal of Biomedical Informatics*, vol. 83, pp. 112–134, 2018.

[49] K. Pang, L. Li, W. Ouyang, X. Liu, and Y. Tang, "Establishment of ICU mortality risk prediction models with machine learning algorithm using MIMIC-IV database," *Diagnostics*, vol. 12, no. 5, p. 1068, 2022.

[50] Y. Sun, Z. He, J. Ren, and Y. Wu, "Prediction model of in-hospital mortality in intensive care unit patients with cardiac arrest : A retrospective analysis of MIMIC-IV database based on machine learning," *BMC Anesthesiology*, vol. 23, no. 1, p. 178, 2023.

[51] L. Mondrejevski, I. Miliou, A. Montanino, D. Pitts, J. Hollmén, and P. Papapetrou, "FLICU : A federated learning workflow for intensive care unit mortality prediction," in *2022 IEEE 35th International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE, 2022, pp. 32–37.