

# Prise en compte de données séquentielles hétérogènes dans l'apprentissage profond : application aux données de soins intensifs

Mamadou Ben Hamidou CISSOKO\*, Nicolas LACHICHE\*, Vincent CASTELAIN\*\*

\* ICube, Université de Strasbourg, \*\* Hôpitaux Universitaires de Strasbourg

**Résumé.** Les données massives des dossiers médicaux électroniques (DME) constituent une opportunité pour l'analyse de la santé à grande échelle. La prédiction des résultats cliniques tels que la durée de séjour ou la mortalité dans les unités de soins intensifs (USI) joue un rôle important dans l'amélioration des performances des systèmes de santé. Une question difficile concerne les séries chronologiques d'événements multivariés contenues dans les DME. Cette question est de savoir comment modéliser correctement les dépendances temporelles entre de nombreux événements cliniques différents. Nous proposons une architecture d'apprentissage profond appelée Time-distributed LSTM (Long Short Term Memory) pour des données USI (TD-LSTM-USI), qui apprend les dépendances temporelles entre les différents événements cliniques afin d'en améliorer le pouvoir prédictif. Nous montrons l'efficacité de l'approche proposée sur deux tâches médicales de référence (mortalité et durée de séjour).

## 1 Introduction

L'analyse des séries chronologiques cherche à utiliser les données recueillies lors d'observations antérieures pour élaborer un modèle capable de représenter la structure de la série afin de prédire et de classer les événements futurs. En raison de la pertinence de cette compétence prédictive, l'analyse des séries chronologiques est devenue un aspect crucial de la modélisation des données dans divers secteurs, notamment les prévisions financières, les diagnostics médicaux et la surveillance de l'environnement ([Mudelsee, 2019](#)).

Ces dernières années, la disponibilité des données médicales a augmenté grâce à l'adoption généralisée des dossiers médicaux électroniques (DME) dans les systèmes d'information des hôpitaux. Dès lors, l'intérêt et la recherche se sont accrus pour l'application d'algorithmes d'apprentissage machine (ML), allant des modèles linéaires simples aux réseaux d'apprentissage profond aux données cliniques, notamment pour l'analyse des données des dossiers médicaux électroniques (DME) des unités de soins intensifs. Compte tenu de la nature des données physiologiques du DME, il existe souvent un grand nombre de valeurs manquantes par manque de collecte. Cet inconvénient limite l'utilisation des modèles ML traditionnels, qui nécessitent principalement des données de séries chronologiques périodiques. Un DME stocke les données relatives aux visites des patients à l'hôpital avec des caractéristiques hétérogènes, notamment les diagnostics, les résultats de laboratoire, les médicaments prescrits, les

procédures, les données d'image ainsi que certaines informations démographiques. Il contient également des notes cliniques rédigées par les praticiens médicaux pour rendre compte de l'état du patient et des événements médicaux survenus pendant son séjour ou sa visite. Les données résultantes comprennent des trajectoires de santé différentes avec des longueurs variables et des périodes différentes entre les observations consécutives pour chaque variable clinique, car toutes les variables cliniques ne sont pas mesurées à intervalles de temps réguliers. Ainsi, les mesures peuvent se produire sporadiquement dans le temps en fonction de l'état sous-jacent du patient. De ce fait les intervalles de temps entre les observations consécutives varient d'un patient à l'autre et même au sein du parcours de santé d'un même patient.

Afin de résoudre ce problème, la plupart des travaux existants ont directement modélisé les observations contenant des valeurs manquantes, par exemple en les transformant en une série temporelle de distributions sur les valeurs possibles. [Zheng et al. \(2017\)](#) ont proposé une méthode générale pour représenter ces séries temporelles d'événements multivariés en séries temporelles régulières non biaisées. Cette méthode est basée sur l'agrégation des mesures en intervalles de temps discrets à chaque épisode de soins pour créer des séries temporelles multivariées avec un intervalle de temps régulier pour des données DME échantillonnées irrégulièrement. A partir de là, un modèle d'inférence est utilisé pour apprendre la caractéristique pour chaque variable médicale à chaque point de temps.

Au fil des ans de nombreuses stratégies basées sur les DME ont été développées allant des procédures statistiques de base pour l'imputation, telles que le zéro, la moyenne, la moyenne mobile, la dernière observation ([Kantardzic, 2011](#); [Lipton et al., 2015](#)) à des approches basées sur les réseaux de neurones récurrents ont été utilisées sur les données DME en raison de leurs capacités de traitement des informations séquentielles ([Lipton et al., 2015](#); [Purushotham et al., 2018](#)). Cependant, les approches existantes basées sur les réseaux de neurones ne modélisent pas les données par heure mais donnent toutes les données du patient au LSTM en une seule fois. Cela peut facilement conduire à un ajustement excessif des données cliniques multivariées contenant des dépendances temporelles cruciales entre de nombreux événements cliniques. Car la nature séquentielle et la dépendance temporelle ne sont pas prises en compte. Dans les séries chronologiques d'événements multivariés produits à partir des DME, la question relative sur comment modéliser correctement les dépendances temporelles entre de nombreux événements cliniques différents reste d'actualité. Généralement, dans les réponses à cette question, les données sont fournies en une seule fois mais pas par heure. Par exemple, l'administration de norépinéphrine dépend de la présence d'une hypotension (pression artérielle basse) avec un niveau croissant de lactate qui peut être soumis à une intubation récente d'un patient présentant un choc septique. [Lee et Hauskrecht \(2019\)](#) montrent comment la modélisation de la dépendance à court terme peut améliorer la prévisibilité d'événements futurs multivariés mais il s'est avéré que les réseaux neuronaux récurrents peuvent facilement omettre ou atténuer ces informations (dépendances temporelles) lors de leurs calculs dans l'état caché ([Pascanu et al., 2013](#)). Par conséquent, dans ce travail, nous abordons le problème mentionné ci-dessus de la modélisation des dépendances à long terme dans les séries chronologiques d'événements cliniques multivariés en proposant une architecture appelée Time-distributed LSTM aux données USI (TD-LSTM-USI). Cette architecture est capable de traiter des séries chronologiques cliniques multivariées complexes avec une plus grande expressivité tout en apprenant des dépendances temporelles entre des nombreux événements cliniques. Pour évaluer notre modèle, nous utilisons les données cliniques réelles dérivées des DME des patients en soins intensifs

dans la base de données MIMIC-III (Johnson et al., 2016). Nous étudions également l'effet des différentes stratégies d'imputation sur la performance du modèle. Ainsi, les principales contributions de ce travail sont les suivantes :

- Nous proposons un réseau neuronal multimodal capable d'apprendre les dépendances temporelles sur les données cliniques pour améliorer la fiabilité de la prédiction de la mortalité et de la durée de séjour ;
- La structure **Time-distributed** est proposée pour apprendre les dépendances temporelles à court et à long terme entre différentes caractéristiques cliniques. Elle permet d'exploiter pleinement les informations à différentes échelles de temps ;
- Nous avons évalué notre modèle sur des ensembles de données de soins de santé du monde réel et obtenu de meilleurs résultats que l'état de l'art pour les différentes tâches de prédiction en utilisant moins de caractéristiques, validant ainsi l'efficacité de l'approche proposée dans un contexte clinique ;

Cet article est structuré comme suit : la section 2 est relative aux travaux connexes à notre étude. La section 3 décrit les jeux de données, les sujets d'études, le développement du modèle et les méthodes d'évaluation utilisés. Dans la section 4, nous décrivons la base de données MIMIC-III Johnson et al. (2016) et les étapes de prétraitement utilisées pour obtenir les différents ensembles de données de référence. Les différents résultats sont discutés dans la section 5. Enfin, la section 6 conclut cette étude.

## 2 Travaux connexes

Plusieurs méthodes de prévision et d'imputation ont été proposées au cours de la dernière décennie pour traiter les valeurs manquantes dans des séries chronologiques d'événements multivariés produites à partir de DME échantillonnés de manière irrégulière et/ou éparse. Ainsi de même pour les dépendances temporelles entre différentes caractéristiques cliniques. Ces méthodes peuvent être classées en deux types : (i) représentation temporelle et granularité temporelle, (ii) méthodes d'imputation conventionnelles.

### 2.1 Représentation temporelle et granularité temporelle

Les séries chronologiques initiales d'événements multivariés basées sur les DME sont constituées d'événements irréguliers enregistrés progressivement. Pour modéliser ces séries temporelles d'événements cliniques multivariés, on peut les transformer en une représentation en temps discret en utilisant une segmentation basée sur une fenêtre (Rajkomar et al., 2018; Lee et Hauskrecht, 2019) mettant en correspondance plusieurs événements qui se produisent dans une fenêtre temporelle donnée dans un vecteur binaire de taille fixe. Lee et Hauskrecht (2021) ont également proposé un cadre dans lequel ils incorporent trois modules de mécanismes différents : abstraction, contexte récent et mémoire de périodicité pour traiter les caractéristiques temporelles des séries chronologiques d'événements cliniques multivariés. Dans leur approche, la séquence d'entrée des événements cliniques est représentée par un vecteur binaire. Ces approches aboutissent à une représentation plus détaillée des états du patient, mais peuvent également conduire à des séquences plus longues et plus éparse, ce qui rend la modélisation de la dépendance des événements beaucoup plus difficile et coûteuse en termes de calcul.

## 2.2 Méthodes d'imputation conventionnelles

Ces méthodes d'imputation conventionnelles peuvent être classées en deux catégories. La première catégorie comprend les méthodes d'imputation statistique telles que les méthodes d'imputation par la moyenne simple, par la médiane (Acuna et Rodriguez, 2004) et par le ratio. D'autres méthodes statistiques classiques de séries chronologiques tels que la moyenne mobile intégrée autorégressive (ARIMA), qui élimine les parties non stationnaires d'une séquence et l'ajuste à un modèle stationnaire paramétré ont été utilisées. L'une des limites de leur approche de l'utilisation des données chronologiques du DME comme données non chronologiques est la sous-utilisation des données chronologiques du DME.

La deuxième catégorie comprend une série des modèles neuronaux, notamment les réseaux neuronaux récurrents (RNN). Baytas et al. (2017) ont proposé un LSTM sensible au temps (T-LSTM), en modifiant les cellules du LSTM qui ajuste l'état caché en fonction des intervalles de temps irréguliers. Chung et al. (2014) ont utilisé une approche simple pour traiter les valeurs manquantes et les intervalles de temps irréguliers dans les données de séries temporelles régulières en agrégeant les séries temporelles, et en proposant des méthodes d'imputation simples sur ces séries comprenant l'imputation moyenne et le remplissage avant avec les valeurs passées avant de les fournir à un classifieur. Lipton et al. (2016) montrent également que la performance de la prévision de la mortalité peut être améliorée en passant des vecteurs de masque (un vecteur binaire pour indiquer si la variable est renseignée ou pas à un instant  $t$ ) comme caractéristiques supplémentaires à un classifieur RNN. Mais ils perdent toujours des informations à grain fin en agrégeant chaque série temporelle en intervalles horaires. Harutyunyan et al. (2019) ont proposé un LSTM qui gère chaque variable séparément avec son vecteur de masque dans une couche LSTM bidirectionnelle indépendante, puis une concaténation de toutes les sorties est effectuée. Ainsi, les informations relatives à une variable unique peuvent être apprises. Cette approche est coûteuse car chaque variable doit être apprise dans un réseau indépendant, ce qui entraîne des calculs supplémentaires.

Notre méthode est conçue pour traiter des séries temporelles multivariées avec des valeurs manquantes massives. Elle peut traiter directement les séries temporelles et les indicateurs binaires pour les valeurs manquantes sans considérer une procédure en deux étapes. Les caractéristiques calculées manuellement des différentes variables ne sont pas nécessaires car elle peut explorer les caractéristiques des différentes variables en même temps automatiquement sans coût de calcul supplémentaire.

## 3 Matériels et méthodes

Dans cette section, nous présentons le cadre que nous proposons pour la prédiction des données hétérogènes. Nous commençons par présenter la notation et une description des séries temporelles physiologiques, puis nous présentons les modèles utilisés pour les tâches mentionnées ci-dessus.

### 3.1 Représentation des données & Architecture utilisée

Étant donné une série temporelle multivariée comportant  $D$  variables physiologiques, y compris les analyses de laboratoire et les signes vitaux, sur  $T$  points temporels pour chaque sé-

jour (icustay)  $n$  dans notre cohorte de  $N$  séjours, nous la désignons comme  $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_t, \dots, \mathbf{x}_T\}^T \in \mathbb{R}^{T \times D}$ , où  $\mathbf{x}_t \in \mathbb{R}^D$  représente la  $t$ -ième observation de toutes les variables observées aux instants  $t$ , et  $x_t^d$  est le  $d$ -ième élément ou variable dans  $\mathbf{x}_t$ . Dans ce contexte, étant donné que la série temporelle  $\mathbf{X}$  comprend des valeurs manquantes, nous introduisons le vecteur de masque dans la série temporelle,  $\mathbf{M} = \{\mathbf{m}_1, \dots, \mathbf{m}_t, \dots, \mathbf{m}_T\}^T \in \mathbb{R}^{T \times D}$ , de même taille que  $\mathbf{X}$ , pour indiquer quelles variables sont observées ou manquantes à l'instant  $t$ . En particulier,

$$m_t^d = \begin{cases} 1, & \text{si } x_t^d \text{ est renseignée} \\ 0, & \text{sinon} \end{cases} \quad (1)$$

Et un ensemble d'observations statiques noté  $\mathbf{S} = \{s_i, \dots, s_{i+1}\}$  où  $s_i \in \mathbf{N}^S$  est extrait pour chaque séjour. Étant donné un ensemble de données de séries chronologiques cliniques pour  $N$  séjours, nous définissons notre cohorte de  $N$  exemples de sujets comme  $\{(X^{(n)}; M^{(n)}; S^{(n)})\}_{n=1}^N$ .

La prédiction de la mortalité est un problème de classification binaire. La durée de séjour (LOS) est un problème de régression.

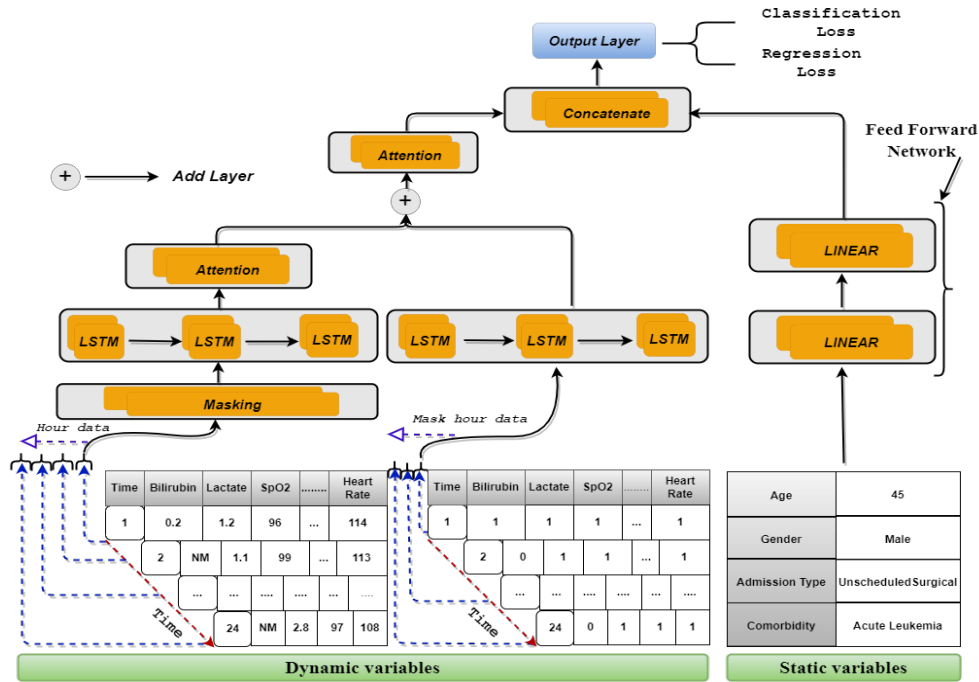


FIG. 1 – L'architecture globale de notre modèle. La séquence originale d'événements irréguliers est envoyée sous forme de données horaires dans une couche LSTM. Notre modèle apprend les dépendances temporelles entre les différentes caractéristiques cliniques afin d'améliorer la prédiction.

## 3.2 Développement du modèle & Mesures d'évaluation

Pour toutes les tâches de prédiction, nous évaluons notre architecture par une stratégie de validation croisée double en 5 plis. Le seuil de classification est réglé sur les données d'entraînement.

Toutes les mesures de performance sont évaluées sur les données de test. Pour les tâches de régression, nous mesurons l'erreur absolue moyenne (MAE) et la racine de l'erreur quadratique moyenne (RMSE), tandis que pour les tâches de classification, nous indiquons la moyenne et l'écart type 5 plis utilisant le F1-Score. Le F1-score est recommandé en présence de classes déséquilibrées. C'est notre cas car le nombre de décès est près de 10% voir tableau 1.

Pour toutes les tâches, nous utilisons la même cohorte et extrayons les données des 24 et 48 premières heures de chaque séjour en soins intensifs. Dans toutes les architectures, nous utilisons la fonction d'activation ReLU pour ajouter de la non-linéarité aux modèles, sauf dans la couche finale. Des lots d'exemples de ( $n=128$ ) pour les tâches de classification et ( $n=64$ ) pour les tâches de régression ont été fixés l'entraînement. Pour éviter le surapprentissage, une régularisation par décroissance de poids  $L_2$  est appliquée à toutes les unités de récurrence avec un facteur d'échelle de 0,001. Adam est utilisé comme optimiseur avec un taux d'apprentissage de  $10^{-3}$  et un taux de décroissance de  $10^{-4}$ . Le nombre d'époque est fixée à 100 pour tous les modèles et l'arrêt précoce est utilisé pour éviter le surapprentissage. Nous utilisons également un taux d'abandon de 0,3 tout en éliminant de manière aléatoire les neurones d'entrée du réseau pendant la phase d'apprentissage afin d'éviter un surapprentissage.

## 4 EXPÉRIMENTATIONS

### 4.1 La base des données

MIMIC-III (Medical Information Mart for Intensive Care III) (Johnson et al., 2016), est une base de données publique sur les soins intensifs gérée par le Laboratory for Computational Physiology du Massachusetts Institute of Technology (MIT). MIMIC-III comprend des données de santé anonymes associées à 46 520 patients distincts qui ont séjourné dans les différentes unités de soins intensifs du Beth Israel Deaconess Medical Center (BIDMC) entre 2001 et 2012. Les données couvrent 58 976 admissions à l'hôpital, dont 61 532 admissions distinctes dans les différentes unités de soins intensifs. Cette base de données comprend des informations sur la santé des patients telles que les données démographiques, les signes vitaux, les résultats des tests de laboratoire, les médicaments, les codes de diagnostic, ainsi que les notes cliniques.

### 4.2 Sélection de la cohorte

Pour créer la cohorte correspondante, nous avons d'abord exclu tous les patients âgés de moins de 15 ans et ceux qui ne disposaient d'aucun enregistrement des signaux temporels requis. L'admission d'un patient à l'hôpital peut correspondre à zéro ou plusieurs épisodes de soins intensifs c'est-à-dire transféré d'une unité de soins intensifs à une autre plusieurs fois au cours de son séjour. Dans cet article, nous considérons tous les séjours associés à un patient pendant son admission à l'hôpital. Après avoir appliqué ces critères d'exclusion, nous avons obtenu un total de 53 304 séjours distincts éligibles pour nos différentes tâches de prédiction

avec un âge médian de 65,8 ans (Q1-Q3 : 52,8-77,8). Le taux de mortalité en milieu hospitalier et en unités des soins intensifs étaient respectivement de 12,4% et 8,8%. La durée moyenne de séjour en unités de soins intensifs était de 4,15 jours (plus de détails sont fournis dans le tableau 1).

TABLE 1 Caractéristiques et mesures des résultats en matière de mortalité. Durée du séjour (LOS). Les variables continues sont présentées sous forme de médiane [écart interquartile Q1-Q3]; les variables binaires ou catégorielles sous forme de nombre (%).

		Overall	Alive at Hospital	Dead at Hospital
ICU ADMISSIONS		53304	46718	6586
AGE, median [Q1,Q3]		65.8 [52.8,77.8]	64.7 [52.0,76.9]	73.4 [60.1,82.8]
GENDER, n (%)	F	23276 (43.7)	20276 (43.4)	3000 (45.6)
	M	30028 (56.3)	26442 (56.6)	3586 (54.4)
ADMISSION TYPE, n (%)	Medical	32578 (61.1)	27729 (59.4)	4849 (73.6)
	ScheduledSurgical	6296 (11.8)	6104 (13.1)	192 (2.9)
	UnscheduledSurgical	14430 (27.1)	12885 (27.6)	1545 (23.5)
<b>Outcomes</b>				
HOSPITAL DEATH, n (%)	Alive	46718 (87.6)	-	-
	Dead	6586 (12.4)	-	-
ICU DEATH, n (%)	Alive	48591 (91.2)	-	-
	Dead	4713 (8.8)	-	-
ICU LOS* (days), median [Q1,Q3]		2.1 [1.2,4.2]	2.1 [1.2,4.0]	3.1 [1.3,7.2]

### 4.3 Sélection et Extraction des caractéristiques

Les ensembles de données de référence extraits sont traités pour obtenir les caractéristiques qui seront utilisées pour les tâches de prédiction. Nous collectons 4 caractéristiques démographiques statiques (âge, sexe, type d'admission, comorbidité) et 60 mesures de laboratoire et mesures vitales qui varient dans le temps pour chaque séjour du patient.

Nous avons choisi deux ensembles de caractéristiques, comme décrit ci-dessous, pour permettre une étude comparative exhaustive.

- Ensemble de caractéristiques A : Cet ensemble de caractéristiques comprend les 15 caractéristiques utilisées dans le calcul du score SAPS-II (Le Gall et al., 1993). Nous fusionnons les données en fonction des connaissances médicales. Ainsi, pour calculer le score de l'échelle de coma de Glasgow, nous additionnons les valeurs GCSVerbal, GCSMotor et GCSEyes. Pour calculer le débit urinaire, nous additionnons les caractéristiques représentant le débit urinaire. Pour la température corporelle, nous convertissons l'échelle Fahrenheit en Celsius et calculons également le rapport PaO2/FiO2 au lieu de le considérer comme des caractéristiques individuelles.
- Ensemble de caractéristiques B : Cet ensemble de caractéristiques se compose de 66 caractéristiques et comprend les 15 caractéristiques de l'ensemble de caractéristiques A.

#### Transformation en données de séries chronologiques horaires :

Nous avons transformé chaque caractéristique temporelle en données pour les premières 24 heures et les premières 48 heures de chaque admission en USI. Chaque caractéristique temporelle d'un séjour aux soins intensifs est agrégée en tranches horaires régulièrement espacées (0-1 heure, 1-2 heures, etc.). Certaines caractéristiques peuvent avoir plusieurs enregistrements au cours du processus d'échantillonnage. Comme nous avons besoin d'une valeur représentative pour chaque caractéristique à un pas de temps particulier, nous avons agrégé les enregistre-

## Données séquentielles hétérogènes dans l'apprentissage profond : cas des soins intensifs

ments multiples en fonction de la caractéristique en prenant soit le maximum, soit le minimum ou la somme à un pas de temps particulier. Par exemple, nous additionnons la caractéristique de sortie d'urine et prenons la valeur minimale la pression artérielle moyenne (PAM), GCSVerbal, GCSMotor, et GCSEyes et le maximum pour les autres caractéristiques.

Outre en échantillonnant les données toutes les heures, nous avons constaté que l'absence des valeurs pour les 60 variables de laboratoire et des signes vitaux est supérieur à plus de 93%, car chaque patient peut ne se voir prescrire que quelques tests en fonction de ses besoins médicaux, et ces tests sont peu fréquents dans le temps.

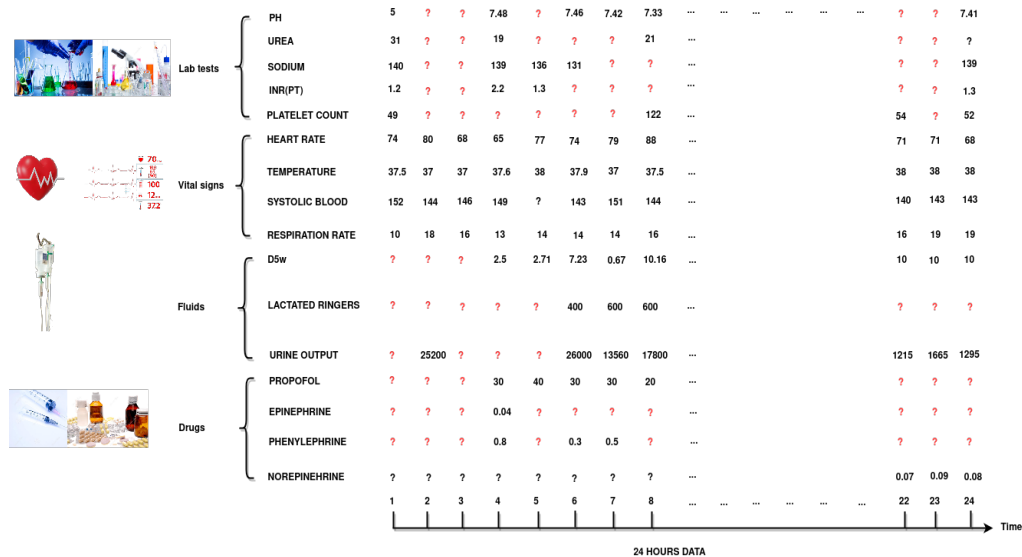


FIG. 2 – Exemple d'événements temporels hétérogènes

### 4.4 Tâches de prédiction

Dans cette partie, nous considérons deux tâches de prédiction de référence en soins intensifs à savoir la prédiction de la mortalité et de la durée du séjour. Plusieurs travaux s'intéressant à ces tâches ont été effectués allant de la prédiction de la mortalité des patients atteints de la septicémie ou des patients sous la ventilation mécanique (Purushotham et al., 2018; George et al., 2021; Kong et al., 2020; Rajkomar et al., 2018) pour évaluer les algorithmes d'apprentissage automatique sur les données des DME. Dans tous les travaux listés ci-dessus, les auteurs se sont intéressés à une tâche de prédiction spécifique. Cependant, dans cette étude, nous nous intéressons à une tâche de prédiction de la mortalité de façon générale.

1. **Prédiction de la mortalité** : La prédiction de la mortalité est reconnue comme l'un des principaux résultats d'intérêt. L'objectif principal de cette tâche est de prédire si un patient décède pendant son séjour à l'hôpital (milieu hospitalier ou USI).
2. **Prédiction de la durée de séjour** : La durée de séjour des patients est utilisée dans les hôpitaux pour mesurer la gravité de l'état d'un patient ainsi que pour la planifi-



cation et la gestion des ressources. Les patients dont la durée de séjour est prolongée consomment davantage de ressources hospitalières et présentent souvent des pathologies chroniques complexes qui ne mettent pas immédiatement leur vie en danger mais qui sont difficiles à traiter. Pour réduire les dépenses de santé, ces patients doivent être identifiés et traités le plus rapidement possible. La prédiction de la durée de séjour des patients pourrait donc jouer un rôle important dans la gestion des soins de santé.

## 5 Résultats

Dans cette section, nous évaluons l'architecture TD-LSTM-USI que nous proposons sur des tâches décrites précédemment 4.4 et le comparons au travail de (Song et al., 2018). Nous montrons que notre architecture surpasse le réseau d'interpolation-prédiction proposé par (Song et al., 2018) sur les différentes tâches. Nous nous posons également la question de l'impact de l'utilisation de différents ensembles de caractéristiques et de la durée d'observation sur la performance de prédiction. Nous discutons des différentes mesures d'évaluation utilisées en section 3.2.

### 5.1 Évaluation de la tâche de prédiction de la mortalité

La table 2 montre les différentes valeurs du F1-score de notre modèle, en utilisant chacune des méthodes d'imputations, sur la tâche de prédiction de la mortalité en milieu hospitalier. La table 3 évalue les mêmes méthodes d'imputation sur la tâche de prédiction de la mortalité en USI. De par ces résultats, plusieurs constats émergent. Le premier constat est que quelle que soit la méthode d'imputation utilisée notre architecture affiche un F1-score supérieur à 50% avec les données de 24 premières heures de séjour. Ce score reste le même avec l'ensemble des caractéristiques A ou B utilisées comme variables d'entrées. Le second constat est que lorsque plus de données sont utilisées par une durée d'observation de 48 heures, la performance du modèle augmente considérablement sur les différentes tâches de prédiction avec un F1-score moyen de plus de 60%. Par ailleurs, nous avons utilisé la méthode du Critical Difference Diagram (Demšar, 2006) afin de comparer par paires les six techniques d'imputations que nous avons utilisées. Les résultats obtenus à travers l'application de cette méthode nous permet d'affirmer qu'il y a une différence significative entre les techniques d'imputations sur les deux tâches de prédiction de la mortalité (figure 3) et que c'est la méthode Forward-Zero-Imputation qui est la meilleure parmi les six utilisées. En conclusion des différentes méthodes d'imputations des valeurs manquantes utilisées, nous pouvons affirmer que le choix de la méthode a un impact significatif sur la performance de notre modèle. Nous avons choisi d'utiliser la méthode Forward-Zero-Imputation par la suite.

Nous avons encore utilisé la méthode du CD pour mieux comparer la performance de notre modèle sur l'utilisation des différentes caractéristiques en l'occurrence l'ensemble A et l'ensemble B. Cette comparaison consiste à utiliser la seule méthode de Forward-Zero-Imputation sur les deux durées d'observation (24h & 48h) et qui sera ensuite utilisé afin de comparer notre modèle par rapport au modèle défini dans (Song et al., 2018). Les résultats (figure 4) montrent que quand plus de caractéristiques et plus de données sont collectées, la performance du modèle s'améliore. Cependant, d'après le diagramme du CD, les quatre types de données sont

Données séquentielles hétérogènes dans l'apprentissage profond : cas des soins intensifs

TABLE 2 Mortalité en milieu hospitalier

IMPUTATION METHODS	24 HOURS DATA		48 HOURS DATA	
	SAPS II FEATURES	ALL FEATURES	SAPS II FEATURES	ALL FEATURES
ZERO-IMPUTATION -WO-M	0.53	0.53	0.58	0.59
MEDIAN-IMPUTATION -WO-M	0.52	0.56	0.56	0.62
ZERO-IMPUTATION + MASKING VECTOR	0.53	0.58	0.58	0.62
ZERO-IMPUTATION	0.54	0.57	0.58	0.62
FORWARD-ZERO-IMPUTATION	0.55	0.58	0.61	0.63
FORWARD-MEDIAN-IMPUTATION	0.53	0.60	0.60	0.62

TABLE 3 Mortalité en USI

IMPUTATION METHODS	24 HOURS DATA		48 HOURS DATA	
	SAPS II FEATURES	ALL FEATURES	SAPS II FEATURES	ALL FEATURES
ZERO-IMPUTATION -WO-M	0.53	0.55	0.58	0.60
MEDIAN-IMPUTATION -WO-M	0.52	0.56	0.58	0.62
ZERO-IMPUTATION + MASKING VECTOR	0.53	0.58	0.61	0.65
ZERO-IMPUTATION	0.52	0.58	0.60	0.65
FORWARD-ZERO-IMPUTATION	0.55	0.59	0.62	0.66
FORWARD-MEDIAN-IMPUTATION	0.51	0.59	0.61	0.66

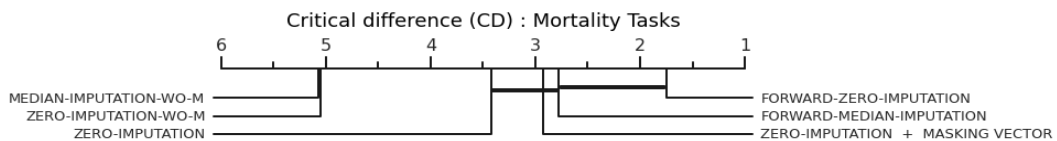


FIG. 3 – Diagramme de différence critique montrant la comparaison des différentes méthodes d'imputation par paire sur la prédiction de la mortalité sur les données de (24 hrs & 48 hrs) en utilisant les différents ensembles de caractéristiques

statistiquement différents les uns des autres. Nous avons retenu la représentation qui apporte le plus d'information : l'ensemble de caractéristiques B observées sur une durée de 48 heures.

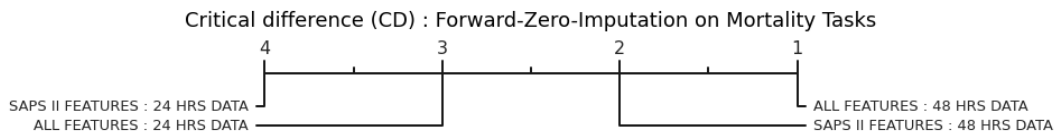


FIG. 4 – Diagramme de différence critique de la prédiction de la mortalité en utilisant les différents ensembles de caractéristiques (A & B) sur les données de (24 hrs & 48 hrs)

En dépit de l'évaluation de la capacité de notre modèle en utilisant la méthode du F1-score, nous l'avons aussi comparé avec l'approche proposée dans (Song et al., 2018). Cette

TABLE 4 TD-LSTM-ICU VS SAnD on MORTALITY TASKS : 48 HOURS DATA

TASKS	TD-LSTM-ICU		SAnD	
	SAPS II FEATURES	ALL FEATURES	SAPS II FEATURES	ALL FEATURES
HOSPITAL-MORTALITY	<b>0.61</b>	<b>0.63</b>	0.59	0.61
ICU-MORTALITY	<b>0.62</b>	<b>0.66</b>	0.60	0.64

comparaison est faite sur la base des tâches de prédiction de la mortalité en milieu hospitalier et dans les USI. La table 4 affiche les résultats obtenus. Ainsi, nous constatons que notre modèle dépasse celui proposé dans (Song et al., 2018) sur les deux tâches de prédictions de plus de 2%. La supériorité de la performance de notre approche à celle proposée par (Song et al., 2018) peut s’expliquer par le fait que le modèle issu de notre approche parvient à apprendre la dépendance temporelle entre les différents évènements puisque les données sont modélisées par heure, contrairement à beaucoup d’autres architectures (Purushotham et al., 2018) où les données du patient sont fournies en une seule fois.

## 5.2 Évaluation de la durée de séjour

La prédiction de la durée de séjour nécessite la prise en compte des dépendances temporelles entre les différents évènements cliniques. Notre architecture est capable de capturer cette dépendance de manière efficace et surpasse l’état de l’art en obtenant de meilleures performances en termes d’erreur quadratique moyenne (en jours). La table 5 compare l’erreur absolue moyenne et la racine de l’erreur quadratique des prédictions des durées de séjour prédites par notre approche et par celle de l’état de l’art (Song et al., 2018). Avec les données

TABLE 5 TD-LSTM-ICU VS SAnD on LOS TASK : 48 HRS DATA

TASK : LOS	TD-LSTM-ICU		SAnD	
	SAPS II FEATURES	ALL FEATURES	SAPS II FEATURES	ALL FEATURES
MAE	<b>1.951 ± 0.057</b>	<b>1.925 ± 0.076</b>	3.177 ± 0.155	3.172 ± 0.15
RMSE	<b>5.374 ± 0.06</b>	<b>5.323 ± 0.078</b>	6.947 ± 0.088	6.942 ± 0.086

des 48 premières heures et quelque soit les caractéristiques (ensemble A ou B) utilisées notre modèle affiche une performance meilleure que celui de l’état de l’art. Les prédictions de notre modèle se situent dans une fourchette de 5 à 5.5 jours par rapport aux données réelles.

À partir de tous les résultats des tâches de prédiction, nous faisons les observations suivantes : notre architecture est beaucoup plus performante lorsque davantage de caractéristiques sont utilisées pour la prédiction. Cela implique qu’elle peut apprendre de meilleures représentations des caractéristiques à partir de plusieurs modalités de données.

## 6 Conclusion

Dans cet article, nous avons identifié la nature des irrégularités présentes dans les données DME, tant au niveau temporel qu’au niveau des caractéristiques mesurées. Nous montrons que la capture des dépendances temporelles au niveau des caractéristiques cliniques peut améliorer

les performances de l'analyse des données DME. Pour répondre à cette question, nous avons proposé un modèle basé sur LSTM capable de gérer la dépendance temporelle au niveau de l'irrégularité des caractéristiques cliniques. Il apprend l'impact des informations précédentes sur l'état du patient pour améliorer les informations extraites à différents niveaux de fonctionnalités.

Nous démontrons également que l'approche proposée atteint une meilleure performance dans toutes les tâches que les modèles de l'état de l'art, avec un grand ou un petit nombre de caractéristiques cliniques utilisées en entrée pour les modèles de prédiction.

Nous notons que, comme les données de MIMIC-III sont générées dans un seul système de DME, elles peuvent contenir des biais systématiques. Une étude future intéressante consistera à explorer comment notre modèle formé sur cet ensemble de données se généralise à d'autres ensembles de données cliniques.

## Références

- Acuna, E. et C. Rodriguez (2004). The treatment of missing values and its effect on classifier accuracy. In *Classification, clustering, and data mining applications*, pp. 639–647. Springer.
- Baytas, I. M., C. Xiao, X. Zhang, F. Wang, A. K. Jain, et J. Zhou (2017). Patient subtyping via time-aware lstm networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 65–74.
- Chung, J., C. Gulcehre, K. Cho, et Y. Bengio (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv :1412.3555*.
- Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine learning research* 7, 1–30.
- George, N., E. Moseley, R. Eber, J. Siu, M. Samuel, J. Yam, K. Huang, L. A. Celi, et C. Lindvall (2021). Deep learning to predict long-term mortality in patients requiring 7 days of mechanical ventilation. *PLoS one* 16(6), e0253443.
- Harutyunyan, H., H. Khachatrian, D. C. Kale, G. Ver Steeg, et A. Galstyan (2019). Multitask learning and benchmarking with clinical time series data. *Scientific data* 6(1), 1–18.
- Johnson, A. E., T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, et R. G. Mark (2016). MIMIC-III, a freely accessible critical care database. *Scientific data* 3(1), 1–9.
- Kantardzic, M. (2011). *Data mining : concepts, models, methods, and algorithms*. John Wiley & Sons.
- Kong, G., K. Lin, et Y. Hu (2020). Using machine learning methods to predict in-hospital mortality of sepsis patients in the ICU. *BMC medical informatics and decision making* 20(1), 1–10.
- Le Gall, J.-R., S. Lemeshow, et F. Saulnier (1993). A new simplified acute physiology score (saps II) based on a European/north American multicenter study. *Jama* 270(24), 2957–2963.
- Lee, J. M. et M. Hauskrecht (2019). Recent context-aware lstm for clinical event time-series prediction. In *Conference on Artificial Intelligence in Medicine in Europe*, pp. 13–23. Springer.

- Lee, J. M. et M. Hauskrecht (2021). Modeling multivariate clinical event time-series with recurrent temporal mechanisms. *Artificial intelligence in medicine 112*, 102021.
- Lipton, Z. C., D. Kale, et R. Wetzel (2016). Directly modeling missing data in sequences with rnns : Improved classification of clinical time series. In *Machine learning for healthcare conference*, pp. 253–270. PMLR.
- Lipton, Z. C., D. C. Kale, C. Elkan, et R. Wetzel (2015). Learning to diagnose with lstm recurrent neural networks. *arXiv preprint arXiv :1511.03677*.
- Mudelsee, M. (2019). Trend analysis of climate time series : A review of methods. *Earth-science reviews 190*, 310–322.
- Pascanu, R., T. Mikolov, et Y. Bengio (2013). On the difficulty of training recurrent neural networks. In *International conference on machine learning*, pp. 1310–1318. PMLR.
- Purushotham, S., C. Meng, Z. Che, et Y. Liu (2018). Benchmarking deep learning models on large healthcare datasets. *Journal of biomedical informatics 83*, 112–134.
- Rajkomar, A., E. Oren, K. Chen, A. M. Dai, N. Hajaj, M. Hardt, P. J. Liu, X. Liu, J. Marcus, M. Sun, et al. (2018). Scalable and accurate deep learning with electronic health records. *NPJ digital medicine 1*(1), 1–10.
- Song, H., D. Rajan, J. Thiagarajan, et A. Spanias (2018). Attend and diagnose : Clinical time series analysis using attention models. In *Proceedings of the AAAI conference on artificial intelligence*, Volume 32.
- Zheng, K., J. Gao, K. Y. Ngiam, B. C. Ooi, et W. L. J. Yip (2017). Resolving the bias in electronic medical records. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2171–2180.

## Summary

The ever-growing mass of electronic health record (EHR) data provides an opportunity for large-scale data-driven health analysis and intelligent medical care. Predicting clinical outcomes such as length of stay or mortality in intensive care units (ICUs) plays an important role in improving the performance of healthcare systems. One difficult issue with predictive multivariate event time-series produced from EHRs that has not been adequately addressed is how to properly model the temporal dependencies between many different clinical events. We propose a deep learning model called Time-distributed LSTM to ICU data (TD-LSTM-ICU), which learns the temporal dependencies among different clinical events to improve the predictive power. To demonstrate the effectiveness of the proposed approach, we applied it to two benchmark medical tasks (mortality and length of stay).