



الجمهورية الجزائرية الديمقراطية الشعبية
The People's Democratic Republic of Algeria

وزارة التعليم العالي والبحث العلمي
The Ministry of Higher Education and Scientific Research

المدرسة العليا للإعلام الآلي - 8 ماي 1945 - سيدي بلعباس
The Higher School of Computer Science -08 May 1945- Sidi Bel Abbas

UNIVERSITÉ
TÉLUQ

GRADUATION THESIS

To obtain the diploma of **Engineering**

Field: **Computer Science**

Specialty: **Artificial Intelligence and Data Science**

Theme

**Design, Implementation and Deployment of a Self-Supervised
Pretraining Framework for Heterogeneous Physiological Signals**

Presented By
Yacine DAIT DEHANE

Dr. KHALDI Belkacem	President	Dr. NAIB Rabiaa	Examiner
Dr. Chaib Souleyman	Supervisor	Pr. BENABDESLEM Khalid	Co-supervisor
Pr. Henni Khadidja	Co-supervisor	Dr. Hakimi Yacine	Co-supervisor

Defended 23 September 2026 · Academic year 2025/2026

This research work was supported by



**Digital Research
Alliance of Canada**

Abstract

Continuous physiological monitoring generates heterogeneous time series characterized by irregular sampling, missing channels, and different acquisition frequencies. This PFE addresses these challenges through the design and implementation of GlucoAnchor, a multimodal foundation-model architecture capable of directly processing partially observed physiological signals without requiring explicit imputation.

The proposed architecture represents each recorded observation as a token encoding its value, timestamp, and source modality, while unmeasured channels contribute no tokens. This tokenization enables the model to accommodate variable channel combinations and sampling rates within a unified representation space. The architecture combines a permutation-equivariant set encoder with temporal phase representations and a self-supervised learning objective based on latent alignment and masked reconstruction.

The system is trained through a progressive pipeline, starting with CGM-based pre-training and extending to selected physiological channels and pretext tasks identified through ablation experiments. A common evaluation protocol is used to assess representation quality on downstream physiological tasks and to examine robustness across cohorts. The evaluation reported here comes from a checkpoint trained for 20 % of the full training schedule, 20 000 of 100 000 optimizer steps, owing to the time available before the defence; that checkpoint had completed 110 epochs over the pretraining corpus, so the limit is one of optimizer schedule rather than of data coverage, and it is a single seed evaluated under the frozen probe alone. The framework and the full experimental design stand independently of that budget. The resulting implementation provides a flexible foundation for multimodal physiological representation learning and is delivered together with a local offline demonstrator that serves the trained encoder on a CPU host, while compression for wearable-class hardware remains future work.

Keywords : Self-Supervised Learning, Foundation Models, Representation Learning, Irregularly Sampled Time Series, Partially Observed Data, Multi-Channel Physiological Signals, Continuous Glucose Monitoring

Résumé

La surveillance physiologique continue génère des séries temporelles hétérogènes caractérisées par un échantillonnage irrégulier, des canaux manquants et des fréquences d'acquisition différentes. Ce projet de fin d'études présente la conception et l'implémentation de GlucoAnchor, une architecture de modèle fondation multimodal capable de traiter directement des signaux physiologiques partiellement observés sans recourir à une imputation explicite.

L'architecture proposée représente chaque observation enregistrée sous la forme d'un jeton contenant sa valeur, son horodatage et sa modalité source, tandis qu'un canal non mesuré ne contribue à aucun jeton. Cette stratégie de tokenisation permet de gérer différentes combinaisons de canaux et fréquences d'échantillonnage au sein d'un espace de représentation commun. L'architecture combine un encodeur ensembliste équivariant par permutation, des représentations temporelles fondées sur la phase et un objectif d'apprentissage auto-supervisé associant alignement latent et reconstruction masquée.

Le système est entraîné selon une procédure progressive, en commençant par le pré-entraînement sur le CGM puis en intégrant les canaux physiologiques et objectifs de pré-entraînement sélectionnés par des expériences d'ablation. Un protocole d'évaluation commun permet d'étudier la qualité des représentations sur des tâches en aval ainsi que leur robustesse entre cohortes. L'évaluation rapportée ici provient d'un point de contrôle entraîné sur 20 % du calendrier d'entraînement complet, soit 20000 des 100000 étapes d'optimisation prévues, en raison du temps disponible avant la soutenance ; ce point de contrôle avait alors effectué 110 époques sur le corpus de pré-entraînement, de sorte que la limite porte sur le calendrier d'optimisation et non sur la couverture des données, et il s'agit d'une seule graine évaluée sous le seul protocole à encodeur gelé. Le cadre proposé et le plan expérimental complet ne dépendent pas de ce budget. L'implémentation obtenue constitue une base flexible pour l'apprentissage de représentations physiologiques multimodales et permet d'envisager différentes configurations de surveillance sans imposer une fréquence d'échantillonnage ou une disponibilité complète des canaux ; elle est accompagnée d'un démonstrateur local hors ligne qui exécute l'encodeur entraîné sur processeur, la compression pour matériel embarqué restant un travail futur.

Mots-clés : Apprentissage Auto-Supervisé, Modèles Fondamentaux, Apprentissage de Représentations, Séries Temporelles Irrégulières, Données Partiellement Observées, Signaux Physiologiques Multi-Canaux, Mesure Continue du Glucose

ملخص

تتسم بيانات المراقبة الفسيولوجية المستمرة بعدم انتظام أخذ العينات، واختلاف ترددات القياس بين القنوات، وغياب بعضها لدى بعض المرضى، مما يفرض تحديات في بناء تمثيل موحد. يقدم مشروع التخرج أجلوكونكوز نموذج أساس متعدد الوسائط لمعالجة الإشارات المرصودة جزئياً مباشرة، دون إعادة بناء صريحة للبيانات المفقودة.

يعتمد النظام على ترميز يمثل كل ملاحظة بقيمتها وطابعها الزمني ومصدرها، بينما لا تنتج القناة غير المقاسة أي رموز، مما يتيح التعامل مع اختلاف القنوات وترددات أخذ العينات ضمن فضاء موحد. وتتكوّن البنية من مُرمز لمجموعة متكافئ بالنسبة للتبديل، وتمثيل زمني قائم على شبكة طورية، وهدف للتعلّم الذاتي الإشراف يجمع بين المحاذاة في الفضاء الكامن وإعادة البناء المقّنع.

يبدأ التدريب المسبق باستخدام ثعب، ثم يتوسع ليشمل القنوات الفسيولوجية وأهداف التدريب التي تحددها تجارب الاستئصال، مع بروتوكول تقييم موحد لجودة التمثيلات ومئاتها بين المجموعات. وينتج عن ذلك إطار مرن للتعلّم متعدد الوسائط لا يشترط توفر جميع القنوات أو بنية زمنية موحدة، مما يسهل إعادة استخدامه في سيناريوهات مختلفة للمراقبة الفسيولوجية.

ويستند التقييم إلى نقطة تحقق عند ٢٠٠٠٠ خطوة من أصل ١٠٠٠٠٠، أي خمس جدول التدريب، نظراً لضيق الوقت قبل المناقشة، وقد أتمّ النموذج خلالها ١١٠ حقبات كاملة؛ لذا يقع القيد في جدول التحسين لا في تغطية البيانات، ولا يتوقف عليه الإطار أو التصميم التجريبي. كما يُرفق التنفيذ بتطبيق عرض محلي يعمل دون اتصال ويشغل المُرْمِز على وحدة المعالجة المركزية، بينما يبقى ضغط النموذج للعتاد القابل للارتداء عملاً مستقبلياً.

الكلمات المفتاحية: التعلّم الذاتي الإشراف؛ نماذج الأساس؛ تعلّم التمثيلات؛ السلاسل الزمنية غير المنتظمة؛ البيانات المرصودة جزئياً؛ الإشارات الفسيولوجية متعددة القنوات؛ المراقبة المستمرة للغلوكونز
