Mini-Project 2.4: Physiological time-series analysis using approximate entropy and sample entropy

ML for Time Series - MVA 2024/2025

Raphaël Faure raphael.faure@student-cs.fr
Victor Jéséquel victor.jesequel@gmail.com

December 17, 2024

1 Contexte

Etant donné une série temporelle de longueur N, une problématique récurrente est d'extraire D features caractérisant cette série temporelle, afin de réaliser des tâches de Machine Learning classiques comme la classification ou le clustering. Parmi le panel de catégories de features possible, une particulière provient de la théorie de l'information. Pincus [1] a en effet défini en 1991 l'*Approximate Entropy* qui caractérise la structure chaotique ou au contraire régulière d'une série temporelle. Cependant cette méthode souffre de plusieurs inconvénients, notamment qu'elle introduit un biais dans le calcul en prenant en compte les auto-correspondances des sous-séquences. L'étude en question de ce mini-projet [2] étudié propose une méthode améliorée de $ApEn$, appelé $SampEn$ pour *Sample Entropy*, pour l'estimation plus précise de l'entropie de séries temporelles issues de données physiologiques courtes et bruitées. Dans ce rapport, nous nous attarderons d'abord sur la description concise de la théorie derrière ces méthodes. Nous analyserons ensuite les données réelles que nous utiliserons dans les expériences, et particulièrement nous montrerons en quoi les donénes sélectionnées vérifient bien les hypothèses fixées dans cette étude. Enfin, la dernière section sera consacrée aux résultats des principales expériences, établissant les propriétés importantes de $SampEn$. Concernant notre mode de travail :

- La répartition du travail s'est fait comme suit : Raphaël a rédigé la partie Méthode, pendant ce temps Victor a cherché des données physiologiques sur lesquelles conduire les expériences pour rédiger la partie Data. Nous avons tous les deux contribué à la partie résultats, où quand l'un avançait l'autre relisait, et chacun a généré des graphes sur les expériences pour les données simulées et réelles
- Par rapport au code, certaines fonctions utilitaires vues au TP2 ont été reprises pour conduire le diagnostique des séries temporelles réelles. Sur la partie Résultats, nous avons implémenté les fonctions *ApEn* et *SampEn* puis coder les expériences afin de créer les graphes nécessaires à la mise en avant des principaux résultalts. Nous avons utilisé la bibliothèque AntroPy, EntropieHub et une version optimisée avec la librairie Numba afin de comparer avec les résultats de notre implémentation et d'accélérer les calculs.

2 Méthode

L'entropie mesure le taux de génération d'information dans un système. Les méthodes classiques pour estimer l'entropie (par exemple, l'entropie de Kolmogorov) nécessitent des séries temporelles longues et peu bruitées, rarement disponibles dans les études biologiques. Le but de l'article est d'étudier les limites de l'*Approximate Entropy (ApEn)* et de présenter leur méthode *Sample Entropy* qui est plus performante aussi bien en terme de complexité algorithmique que d'approximations sur des séries courtes bruitées.

L'*Approximate Entropy* ($ApEn$) mesure la complexité d'une série temporelle en estimant la probabilité conditionnelle qu'une paire de vecteurs similaires pour m points reste similaire pour $m+1$ points, sous une tolérance r. Soit $\mathbf{x}_m(i)$ un vecteur de longueur m issu de la série, la distance entre deux vecteurs est définie par $d(\mathbf{x}_m(i), \mathbf{x}_m(j)) = \max_k |\mathbf{x}(i+k) - \mathbf{x}(j+k)|$, avec $k \in [0, m-1]$. On compte ensuite les vecteurs $\mathbf{x}_m(j)$ tels que $d(\mathbf{x}_m(i), \mathbf{x}_m(j)) \leq r$, on dénote par $B_i = Card(\{j;\ d(\mathbf{x}_m(i), \mathbf{x}_m(j)) \leq r\})$, $A_i = Card(\{j;\ d(\mathbf{x}_{m+1}(i), \mathbf{x}_{m+1}(j)) \leq r\})$ et $C_i^m(r) = \frac{B_i}{N-m+1}$ ce qui permet de calculer la moyenne logarithmique normalisée des correspondances, donnée par

$\Phi^m(r) = \frac{1}{N-m+1} \sum_{i=1}^{N-m+1} \ln(C_i^m(r))$. L'entropie $ApEn(m, r, N)$ est alors définie comme :

$$ApEn(m, r, N) = \Phi^m(r) - \Phi^{m+1}(r).$$

Nous remarquons directement la présence d'un biais, asymptotiquement nul, dû à la présence de l'auto-correspondance. Cette auto-correspondance permet que les ensembles B_i, A_i ne soient jamais vide pour ne pas avoir de problème de calcul logarithmique. Cependant, bien que le biais disparaît lorsque le nombre de points grandit, il est très impactant sur des séries courtes car il suggère plus de similarité qu'il n'y en a, ce qui cause une sous-évaluation de la complexité.

La *Sample Entropy* (*SampEn*) affine cette méthode sur les séries courtes en supprimant les auto-correspondances (*self-matches*), réduisant ainsi le biais. Au lieu d'estimer la probabilité moyenne logarithmique, *SampEn* calcule directement la probabilité conditionnelle globale en excluant les modèles individuels. Elle est donnée par :

$$SampEn(m, r, N) = -\ln\left(\frac{A(r)}{B(r)}\right),$$

où $A(r)$ est le nombre total de correspondances pour $m+1$ points et $B(r)$ celui pour m. Cet indicateur nous donne une régularité plus "globale" sur les séries courtes que *ApEn* car on ne regarde plus selon chaque vecteur. En enlevant cette focalisation et l'auto-correspondance, on s'expose à nouveau à des problèmes de calculs logarithmique mais avec des occurences plus faible étant donné qu'ils ne peuvent se produire qu'en cas d'abscence de correspondance entre des motifs de taille m ou de taille m+1 ce qui traduit davantage d'une absence de régularité que précédemment. L'un des avantages majeurs de SampEn est la possibilité de calculer des intervalles de confiance grâce à la loi de Student, en modélisant la probabilité conditionnelle comme une moyenne d'échantillon. Cela permet de quantifier la variabilité des estimations et d'évaluer la fiabilité des résultats obtenus. Dans des applications biologiques, où les données sont souvent bruitées et les échantillons de taille limitée, ces intervalles fournissent une validation statistique essentielle pour comparer différentes séries temporelles ou évaluer l'effet d'interventions. Ils renforcent ainsi la crédibilité des analyses dans des environnements incertains. Ce point représente une vraie valeur ajoutée comparé à *ApEn* où il n'est pas simple d'obtenir de tels intervalles.

Les deux indicateurs précédent sont conçues pour comparer une série temporelle avec elle-même. Dans le cadre d'étude de signaux biologiques, il est souvent intéressant de comparer deux signaux différents. En ce sens, les auteurs abordent les indicateurs *Cross-ApEn* et *Cross-SampEn* des outils développés pour analyser la similarité ou l'asynchronie entre deux séries temporelles distinctes. Ces indicateurs reposent en partie sur les méthodes présentées dans les paragraphes précédents. Tandis que *ApEn* et *SampEn* évaluent la régularité d'une seule série, leurs versions croisées comparent des motifs de longueur m issus d'une série dite *template* (u) à une autre série appelée *target* (v). La distance entre deux vecteurs $\mathbf{u}_m(i)$ (extrait de u) et $\mathbf{v}_m(j)$ (extrait de v) est définie comme $d[\mathbf{u}_m(i), \mathbf{v}_m(j)] = \max_{k \leq m} |u(i+k) - v(j+k)|$. On note de même $B_i(r) = Card(\{\mathbf{v}_m(j), d[\mathbf{u}_m(i), \mathbf{v}_m(j)] \leq r\})$, $A_i(r) = Card(\{\mathbf{v}_{m+1}(i), d[\mathbf{u}_{m+1}(i), \mathbf{v}_{m+1}(j)] \leq r\})$. On pose cette fois $C_i^m(r) = \frac{B_i(r)}{N-m+1}$, $C_i^{m+1}(r) = \frac{A_i(r)}{N-m}$ et $F^m(r)(v \mid u) = \frac{1}{N-m+1} \sum_{i=1}^{N-m+1} \ln\left(C_i^m(r)\right)$. L'indicateur *Cross-ApEn* est alors donné par :

$$Cross\text{-}ApEn(m, r, N)(v \mid u) = F^m(r)(v \mid u) - F^{m+1}(r)(v \mid u).$$

Remarquons que le biais d'auto-correspondance n'est plus présent ici. Néanmoins, des problèmes calculatoires liés au *log* persistent et l'indicateur peut être indéfini lorsque les correspondances

($C_i^m(r)$) sont nulles, ce qui se produit fréquemment pour de petites séries ou des tolérances r faibles. Des méthodes introduisant un biais le biais-max ou le biais 0 sont souvent utilisées pour corriger ce problème De plus, de par sa définition, *Cross-ApEn* est directionnelle (ie. $\textit{Cross-ApEn}(v \mid u) \neq \textit{Cross-ApEn}(u \mid v)$). Cela pose des problèmes lors d'analyse bidirectionnelle.

En posant de même, $B^m(r)$ (resp. $A^m(r)$) est le nombre total de correspondances entre u et v pour m points (resp. $m+1$ points), on définit *Cross-SampEn* par :

$$\textit{Cross-SampEn}(m, r, N)(v \mid u) = -\ln\left(\frac{A^m(r)}{B^m(r)}\right).$$

Avec cette défintion, *Cross Sample Entropy* est un indicateur robuste et symétrique, adapté pour évaluer la similarité ou l'asynchronie entre deux séries temporelles. Contrairement à Cross-ApEn, il est directionnellement indépendant, ce qui le rend idéal pour analyser des interactions bidirectionnelles ou des couplages dynamiques. De plus, il est défini même pour des petites séries ou des correspondances rares, garantissant une analyse fiable dans des environnements bruités ou avec des données limitées.

L'article ne donne pas d'implémentations et n'aborde que brièvement les questions de complexité algorithmique. La complexité des indicateurs *Approximate Entropy* (ApEn), *Sample Entropy* (SampEn), *Cross-ApEn* et *Cross-SampEn* dépend principalement de la longueur des séries temporelles (N), de la taille des motifs (m), et des comparaisons nécessaires pour identifier les correspondances sous une tolérance donnée (r). Pour ApEn et SampEn, la complexité asymptotique est $O((N-m+1)^2 \cdot m)$. Cependant, SampEn est souvent plus rapide dans la pratique, car il élimine les auto-correspondances, évitant des biais coûteux à corriger, bien qu'il nécessite des vérifications supplémentaires pour assurer des correspondances valides.

Pour les versions croisées, Cross-ApEn présente une complexité équivalente à celle d'ApEn ($O((N-m+1)^2 \cdot m)$) mais devient directionnel, nécessitant une gestion explicite des cas où les correspondances sont nulles. Cross-SampEn, quant à lui, double presque la charge computationnelle ($O(2 \cdot (N-m+1)^2 \cdot m)$) en raison du calcul séparé des proportions globales $A^m(r)$ et $B^m(r)$ pour garantir la symétrie directionnelle et une robustesse accrue face aux données bruitées ou aux séries courtes. Ainsi, si SampEn est plus performant que ApEn en termes de complexité, cet avantage s'inverse dans les versions croisées, où Cross-ApEn est moins coûteux que Cross-SampEn.

3 Analyse des données

Cette section est destinée à analyser les données mises en jeu dans ce papier, de quel domaine elles sont issues, les hypothèses posées pour ces dernières. Beaucoup de données simulées comme nous l'avons précédemment mentionné sont utilisées dans le papier, ainsi nous avons choisi le jeu de données ECGFiveDays du site www.timeseriesclassification.com pour tester les expériences. Il s'agit des données d'ECG d'un homme de 67 ans prises à deux dates différentes.

La stationnarité des signaux est évaluée à l'aide du test d'Augmented Dickey-Fuller (ADF). On obtient une p-value de 0.02 Notons cependant une légère tendance qu'on peut visualiser sur la figure 1.1. La documentation précisait que le signal original a été segmenté, chaque segment correspondant à un battement de cœur, comme il est courant de faire pour obtenir un signal stationnaire et ergodique. Les niveaux de bruit sont estimés en calculant le rapport signal/bruit (SNR). Le signal étudié ici, avec un SNR de 4.33 dB, présente un niveau de bruit significatif par rapport au

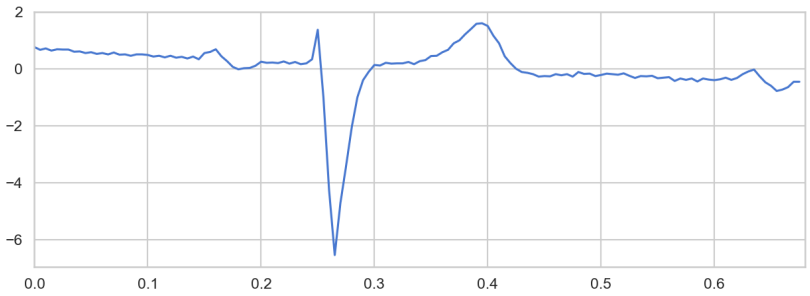

Figure 1: Un signal extrait du dataset, correspondant à un battement cardiaque

signal. Nous utilisons plusieurs outils pour analyser les données plus en profondeur disponible en annexe : 5.3

Ces observations sont en accord avec les caractéristiques décrites des signaux dans le papier, notamment la complexité et l'irrégularité locale des signaux cardiaques, qui peuvent être étudiées avec des outils comme l'entropie d'approximation (ApEn) et l'entropie d'échantillon (SampEn) pour différencier les classes et détecter les variations régulières ou bruitées.

4 Résultats

L'article ne fournissait pas d'implémentation. Nous avons donc implémenté $ApEn$, $SampEn$ ainsi que $CrossSampEn$ et $CrossApEn$. Ces algorithmes sont fonctionnels mais peu efficace, nous les avons donc optmisés à l'aide du package Numba. Nous avons ensuite comparé nos résultats avec les fonctions des packages AntroPy et EntropieHub. Enfin nous avons comparé les temps d'executions des différentes fonctions avec celles des packages. Par souci de concision, certaines figures sont annexe (cf. Figure 7, 8 ,9) le reste est sur le notebook. Globalement Numba à largement accéleré nos premières fonctions et les complexités respectent l'odre établie dans l'article. Les fonctions d'AntroPy sont globalement plus perfomantes. Notons cependant une différence de valeur entre la fonction $CrossApEn$ du package Entropie Hub et la nôtre. La documentation officielle ne détaille pas leur méthode de calcul mais d'après le code source dans le fichier local, la différence réside dans le traitement du $log(0)$.

Une première propriété attendue et plutôt intuitive est que l'entropie croît lorsque r décroît : en effet cela revient à diminuer la tolérance et donc trouver moins de séquences proches à r près. $SampEn$ est plus en accord avec la théorie que $ApEn$ au niveau des variations par rapport à r. En annexe 5.2 sont montrés les résultats sur des données réelles ECG: le signal 1 provient du dataset ECGFiveDays, le signal 2 de ECG200 et le signal 3 de ECG5000.

Une autre propriété critique attendue pour l'*Approximate Entropy* est sa **relative consistance**: Si $ApEn(m_1, r_1)(S) \leq SampEn(m_1, r_1)(T)$ alors $ApEn(m_2, r_2)(S) \leq SampEn(m_2, r_2)(T)$, S et T étant deux séries temporelles. Or plusieurs tests mentionnés dans le papier [2] montrent que $ApEn$ ne possède pas cette propriété, tandis que $SampEn$ a plutôt tendance à la respecter. Nous avons testé ce résultat sur 3 autres signaux issus des datasets précédemment mentionnés. On fixe $m = 3$ et on étudie les variations par rapport à r.

Un des inconvénients majeurs de $ApEn$ est le biais introduit par la prise en compte des auto-

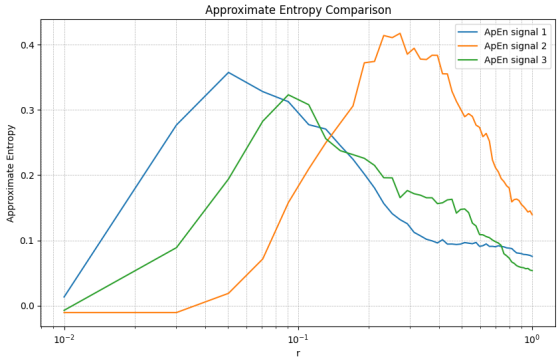

Figure 2: Comparaison des $ApEn$ pour les 3 signaux. Les courbes se croisent.

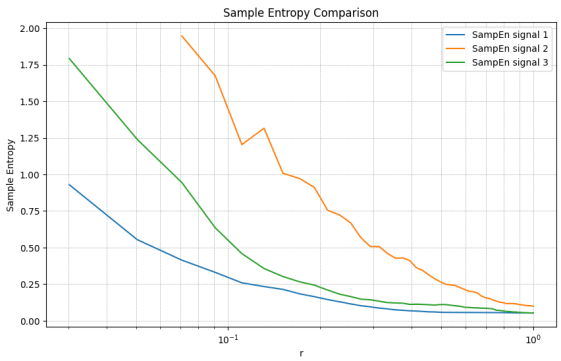

Figure 3: Comparaison des $SampEn$ pour les 3 signaux. Les courbes restent parallèles.

correspondances. Nous avons mise en avant ce défaut en utilisant les processus $MIX(P)$ introduit par les auteurs. Les processus stochastiques $MIX(P)$ qui consistent à générer une sinusoïde dont $N \times P$ points sont remplacés par un bruit aléatoire, P désignant une probabilité entre 0 et 1. Nous avons tracé les valeurs de $AmpEn$ et $SampEn$ en fonction de la proportion de bruit pour des différentes valeurs de r. $ApEn$ sous-estime l'entropie lorsque la proportion de bruit est faible, particulièrement visible sur les graphes où r est faible : elle reste constante alors que $SampEn$ détecte plus précisément l'irrégularité introduite par le bruit. Lorsque r augmente l'écart se réduit car la tolérance devient plus large ce qui masque partiellement l'impact des corrélations.

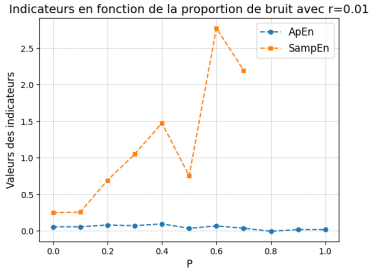

Figure 4: $ApEn$ et $SampEn$ en fonction du bruit, r = 0.01

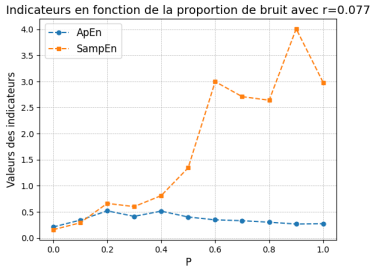

Figure 5: $ApEn$ et $SampEn$ en fonction du bruit, r = 0.077

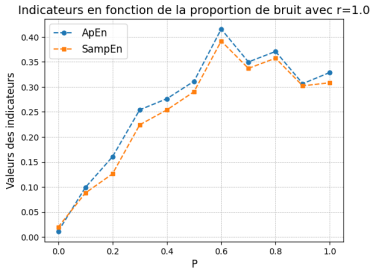

Figure 6: $ApEn$ et $SampEn$ en fonction du bruit, r = 1

Pour finir, nous voulions également tester la *Cross Sample Entropy* ($CrossSampEn$) comme distance dans un problème de classification avec l'algorithme *k-NN*. En effet, cette dernière présente l'avantage d'être bidirectionnelle comparée à la $CrossApEn$, qui par ailleurs échoue à juger de l'ordre de deux séries en les comparant à une 3ème. Il se trouve que lors de l'implémentation d'un k-NN avec la DTW comme distance pour le dataset **"human-locomotion-dataset"**, représentant des données de pas de patients recueillies avec des unités de mesure inertielle (accéléromètre+gyroscope), nouos obtenions un $F1 - score$ de 0.49 après sélection des hyperparamètres optimaux par validation croisée. En remplaçant la DTW par la $CrossSampEn(x, y, m, r)$ avec m=3 et r=0.2 fixés, nous obtenons un F1-score de 0.82, pour k=5 voisins. Cela peut s'expliquer par la capacité de cette dernière à mieux capturer la complexité et les irrégularités des signaux, contrairement à la DTW, qui se concentre sur l'alignement temporel. CrossSampEn est plus sensible aux motifs dynamiques et aux variations locales. CrossSampEn peut être plus adaptée pour le type de données où des variations subtiles et des irrégularités jouent un rôle clé dans la classifica-

tion.

5 Annexe

5.1 Temps d'execution

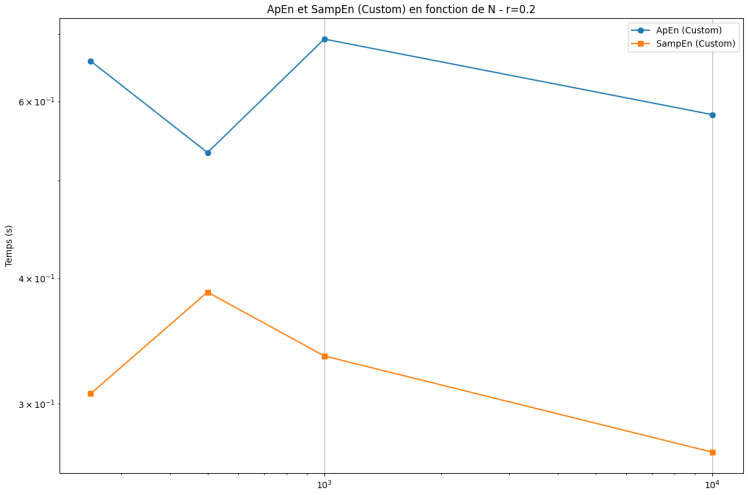

Figure 7: *ApEn* vs *SampEn* signal 1

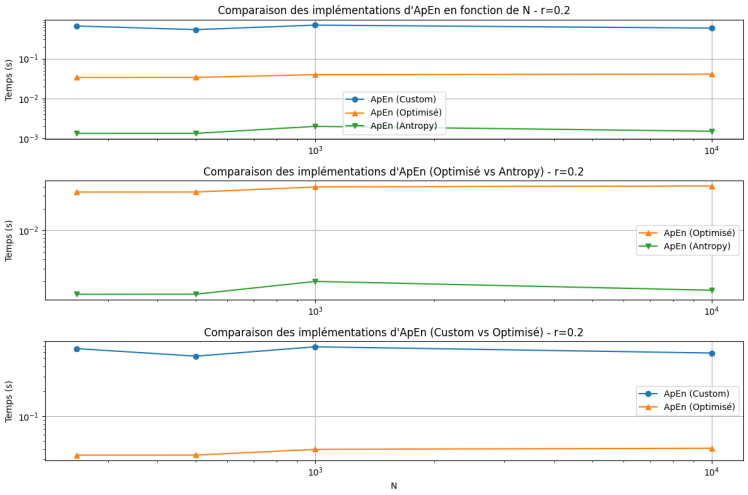

Figure 8: Comparaison *ApEn*

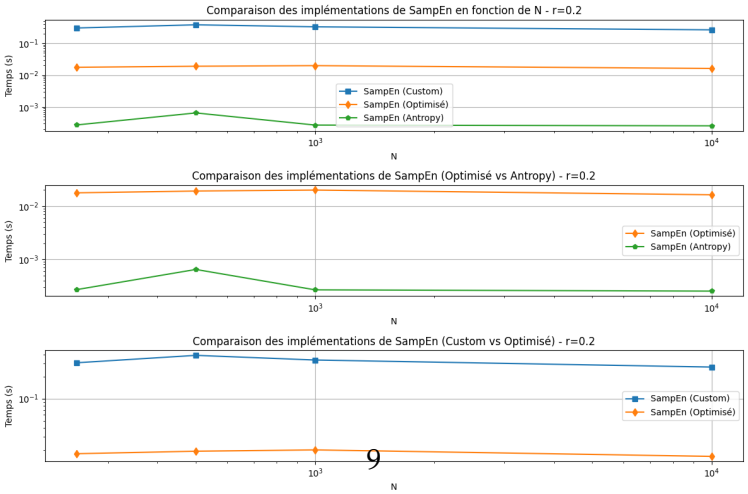

Figure 9: Comparaison *SampEn* signal 3

5.2 Variation de l'entropie avec r

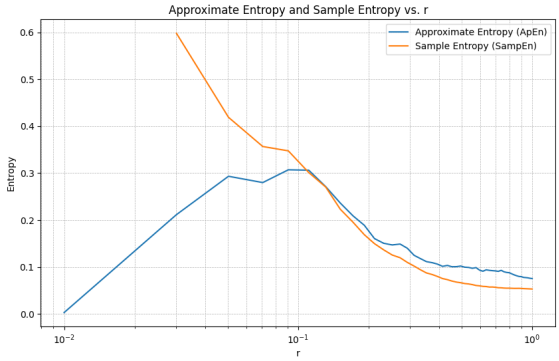

Figure 10: *ApEn* vs *SampEn* signal 1

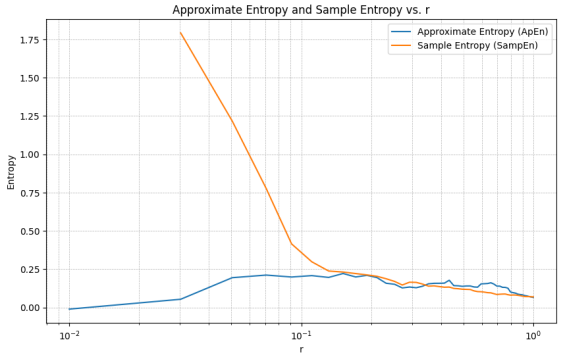

Figure 11: *ApEn* vs *SampEn* signal 2

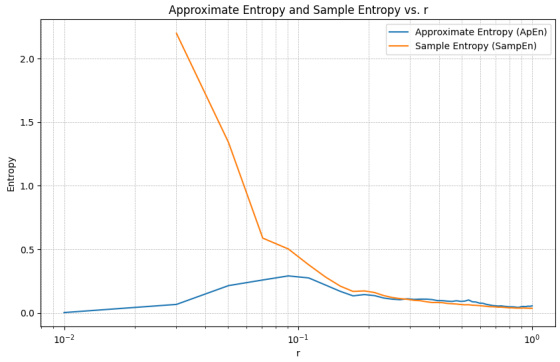

Figure 12: *ApEn* vs *SampEn* signal 3

5.3 Analyse approfondie du signal

- Fonction d'autocorrélation (ACF) : Le déclin progressif de l'ACF montre une diminution de la dépendance temporelle à mesure que le décalage augmente, indiquant que les valeurs du signal deviennent progressivement moins corrélées avec leurs valeurs passées. Il n'y a pas d'oscillation marquée dans le graphe, ce qui suggère l'absence de périodicité claire
- Analyse spectrale : Le spectrogramme révèle une concentration d'énergie dans les basses fréquences, ce qui est attendu pour des signaux ECG

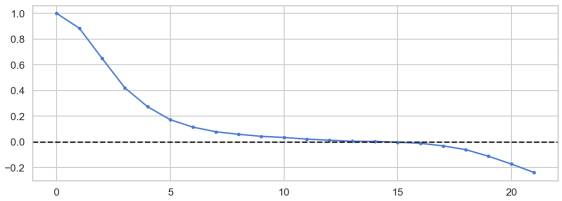

Figure 13: Tracé de la fonction d'auto-corrélation.

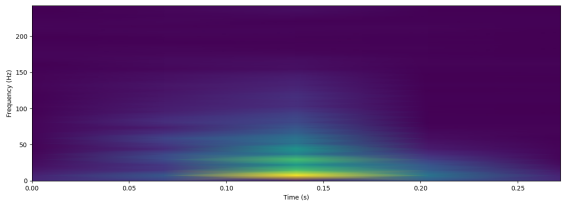

Figure 14: Spectrogramme

5.4 Matrices de confusion pour le problème de classification de pas

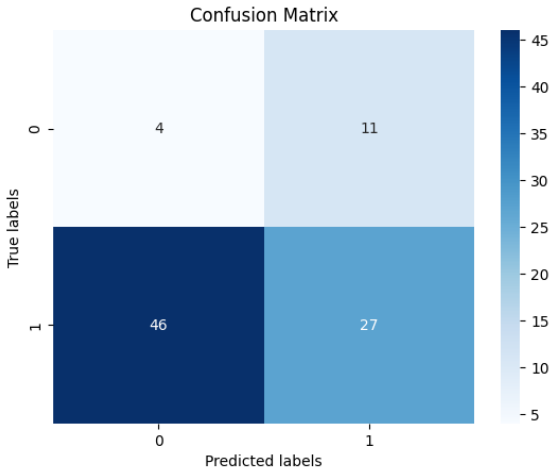

Figure 15: Matrice de confusion avec la DTW comme distance

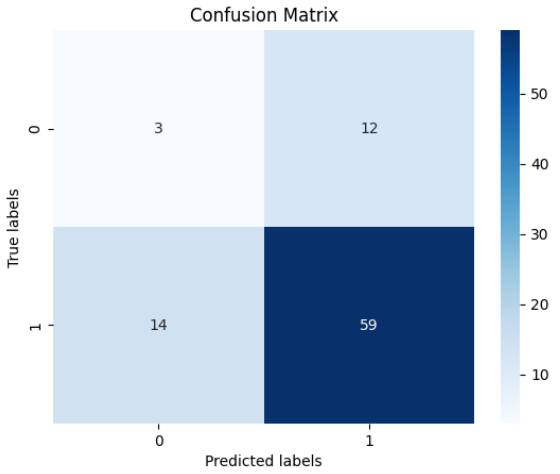

Figure 16: Matrice de confusion avec la *CrossSampEn* comme distance

References

[1] Pincus SM. "Approximate entropy as a measure of system". In: *Proc Natl Acad Sci USA* (1991).

[2] Joshua S. Richman and J. Randall Moorma. "Physiological time-series analysis using approximate entropy and sample entropy". In: *Am J Physiol Heart Circ Physiol 278: H2039–H2049* (2000).