# ConceptECGxAI : une approche post-hoc à base de concepts médicaux pour expliquer un modèle d'apprentissage profond d'aide au diagnostic cardiaque

Victoria Bourgeais, Fall Ahmad, Lence Alex, Salem Joe-Elie, Zucker Jean-Daniel, Prifti Edi, Hanczar Blaise

# ConceptECGxAI : une approche post-hoc à base de concepts médicaux pour expliquer un modèle d'apprentissage profond d'aide au diagnostic cardiaque

Victoria Bourgeais*, Ahmad Fall**, Alex Lence**, Joe-Elie Salem****,♯
Jean-Daniel Zucker**,***, Edi Prifti**,***, Blaise Hanczar♯♯♯

* LaBRI (CNRS/UMR 5800), Université de Bordeaux, Talence, France
Auteur correspondant : victoria.bourgeais@u-bordeaux.fr
** IRD, Sorbonne Université, UMMISCO, F-93143, Bondy, France
*** Sorbonne Université, INSERM, NutriOmique, AP-HP, Hôpital Pitié-Salpêtrière, Paris, France
**** Vanderbilt University Medical Center, Nashville, USA
♯ Centre d'Investigation Clinique Paris-Est, INSERM, AP-HP, Hôpital Pitié-Salpêtrière, Paris, France
♯♯ Laboratoire IBISC (EA 4526), Université Paris-Saclay (Univ Evry), Evry, France

**Résumé.** Le développement de l'intelligence artificielle (IA) participe à l'émergence d'une nouvelle forme de médecine dite personnalisée, qui vise à mieux prendre en compte les caractéristiques des patients. Dans ce contexte, nous nous intéressons à l'application de l'apprentissage profond pour détecter des maladies cardiovasculaires telles que le syndrome du QT-long à partir d'électrocardiogrammes. Cependant, une préoccupation majeure réside dans la nature souvent opaque de ces modèles d'IA dits "boîte noire". Cet article vise à rendre ces derniers plus interprétables en intégrant des concepts médicaux dans une nouvelle approche post-hoc, ConceptECGxAI, fournissant des explications plus intelligibles aux médecins.

## 1 Introduction

L'apprentissage profond est une avancée majeure dans le domaine de l'intelligence artificielle (IA) de ces dernières années. Il s'est rapidement imposé comme un nouveau standard dans plusieurs domaines en surpassant les performances des méthodes antérieures considérées comme l'état de l'art. Ses domaines de prédilection sont principalement l'analyse d'images et le traitement du langage naturel. Un des futurs enjeux majeurs de cette approche est lié aux applications dans le domaine de la santé. Au sein du projet **ANR DeepECG4U** [1], nous nous intéressons à l'application des approches d'apprentissage profond pour la détection de maladies cardio-vasculaires comme le syndrome du QT-long qui peut déclencher des arythmies mortelles, telles que la Torsades-de-Pointes (TdP), à partir des données d'électrocardiogrammes (ECG) des patients. La prise de certains médicaments peut induire l'allongement de l'intervalle QT et en être la cause de la TdP, comme peuvent également être certaines mutations

1. https://anr.fr/fr/projets-finances-et-impact/projets-finances/projet/funded/project/anr-20-ce17-0022/

spécifiques (*i.e.*, le QT-long congénital). Nous souhaitons ainsi proposer un outil d'aide au diagnostic à destination des médecins afin de prévenir ces risques. Dans ce sens, un premier outil à base d'apprentissage profond a été développé (Prifti et al., 2021).

Cependant, les modèles d'apprentissage profond, ainsi que d'autres méthodes d'apprentissage automatique comme les machines à vecteurs de support, sont considérés comme des « boîtes noires », dans lesquelles les données de patients sont injectées en entrée, puis une prédiction est retournée en sortie sans aucune explication. Ceci est un gros problème et un point de réflexion actif chez les législateurs. L'Union Européenne a récemment adopté un texte imposant aux utilisateurs d'algorithmes d'apprentissage automatique d'être capables d'expliquer les décisions d'un modèle prédictif (Goodman et Flaxman, 2017). Premièrement, il est important de s'assurer que les modèles d'apprentissage automatique basent leurs prédictions sur des représentations fiables des patients et ne se concentrent pas sur des artefacts non pertinents présents dans les données d'apprentissage, autrement dit qu'ils ne soient pas sensibles aux biais. Deuxièmement, un modèle performant pour la prédiction d'une certaine maladie ou condition, peut avoir identifié une signature dans les données qui pourrait être une piste de recherche pour les médecins, pouvant renseigner sur la physiopathologie de la maladie en question.

Dans l'état de l'art actuel, il existe deux approches principales pour interpréter les réseaux de neurones : en créant des modèles qui sont par essence interprétables (approche dite *ante-hoc* ou *auto-explicative*), ou en ayant recours à une méthode tierce dédiée à l'interprétation du réseau de neurones déjà appris (approche dite *post-hoc*). Quelle que soit la méthode choisie, l'explication fournie consiste généralement en l'identification des variables d'entrée et des neurones importants pour la prédiction. Or, dans le cas d'une application notamment sur les données de santé comme les ECG, cela n'est pas suffisant. Une des pistes d'amélioration pourrait être d'avoir recours à l'utilisation de concepts de plus haut niveau sémantique pour fournir des explications qui utilisent le même langage que celui des médecins. Un premier travail a été réalisé avec une méthode d'occlusion en croisant les explications obtenues avec les connaissances du domaine (Prifti et al., 2021). Une solution alternative est d'intégrer directement ces concepts dans l'approche post-hoc. Ainsi, dans la continuité des travaux précédents sur les données d'image (Kim et al., 2018; Zaeem et Komeili, 2021; Crabbé et van der Schaar, 2022), nous proposons une nouvelle approche post-hoc qui permet d'interpréter n'importe quel type de réseau de neurones boîte noire déjà appris sur des données ECG en intégrant des concepts médicaux ayant un sens sémantique pour les cardiologues. Cette approche se nomme ConceptECGxAI.

## 2 Méthode

Un ECG enregistre l'activité électrique du cœur à l'aide de plusieurs dérivations, obtenues à partir d'électrodes. La Fig. 1a illustre le tracé normal d'un battement cardiaque, où il est possible d'identifier la position des ondes (P,T) et des pics du complexe QRS (correspondant à l'onde de dépolarisation des ventricules cardiaques). À partir de ces positions, diverses mesures peuvent être réalisées, telles que la durée du complexe QRS, l'intervalle inter-battement R-R, la distance QT, ainsi que les amplitudes des différentes ondes P, T, la tangente de l'onde T, etc. Ces informations permettent aux cardiologues d'évaluer la normalité de l'ECG ou d'identifier d'éventuelles anomalies, comme le QT-long.

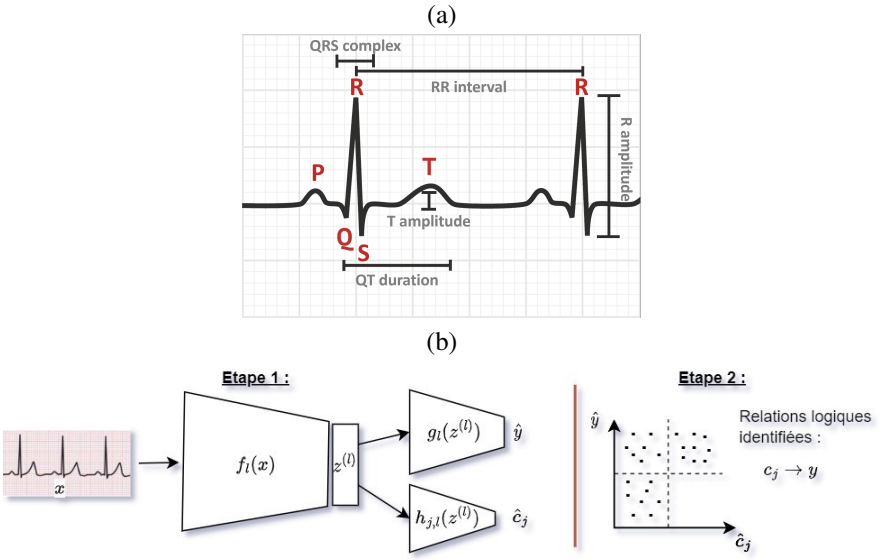



FIG. 1 – (a) Exemple d'un électrocardiogramme de patient annoté. (b) Illustration de ConceptECGxAI pour l'identification d'un concept $c_j$ donné (tel que <qt-duration-long>).

Supposons que nous disposons d'un réseau de neurones prédictif déjà entraîné et qui prend en entrée un électrocardiogramme. ConceptECGxAI vise à accomplir deux objectifs principaux : (i) identifier la présence de concepts significatifs pour le domaine d'application dans les couches cachées du modèle prédictif, puis (ii) établir les relations logiques entre ces concepts et les prédictions du modèle en question. Pour ce faire, nous proposons un pipeline en deux étapes comme illustré dans la Fig. 1b. Une première étape consiste à apprendre par concept une méthode tierce (un interpréteur) permettant d'évaluer à quel point le concept en question est représenté dans une des couches latentes du modèle prédictif. Les concepts sont définis de manière à être compréhensibles pour les utilisateurs finaux, c.-à-d. suffisamment abstraits et représentatifs dans le domaine de compétences de ce dernier. Noter que le modèle prédictif n'a pas été contraint dans son apprentissage à extraire ces concepts. Dans la deuxième étape, nous cherchons à identifier l'existence de relations logiques entre les concepts les plus capturés et les classes de prédiction du modèle prédictif pour construire des règles du type : *"si le concept <QT-long> est capturé dans une des couches latentes du modèle prédictif, alors il est très probable que celui-ci classifie le patient comme étant à risque de <TdP>"*. Les règles ainsi construites pourraient permettre de fournir une interprétation globale du modèle et les interpréteurs permettront de vérifier la présence de ces concepts sur de nouvelles données.

## 2.1 Description du pipeline

Soit un ensemble de concepts binaires $\{c_j\}_{j=1,\dots,C}$ où $C$ correspond au nombre de concepts. On désigne respectivement par $(x, y, \{c_{x,j}\}_{j=1,\dots,C})$ un profil d'électrocardiogramme d'une classe, l'indicateur booléen d'appartenance à la classe positive à prédire et les indicateurs boo-

léens de présence des concepts $\{c_j\}_{j=1,\ldots,C}$. L'approche ConceptECGxAI est illustrée dans la Fig. 1b.

**Étape 1 :** On étudie la présence d'un concept $c_j$ dans une des représentations latentes $\{z^{(l)}\}_{l=1,\cdots,L}$ du réseau de neurones avec $L$ la profondeur du réseau de neurones. Le réseau est découpé en deux parties distinctes : la partie avant la couche cachée $(l)$ étudiée ($f_l(x)$), et celle après ($g_l(z^{(l)})$). Le réseau entier est le résultat de la composition de fonctions telles que $\hat{y} = g_l(f_l(x))$ où la notation $\hat{}$ signifie la probabilité de prédiction retournée par un algorithme de classification. $f_l$ et $g_l$ varient en fonction du point de découpe, c.-à-d. la couche $l$. Le choix de la couche cachée est un hyperparamètre de l'approche. Pour déterminer la présence de concepts, on transforme le problème sous la forme d'une tâche de classification en utilisant une méthode tierce, basée sur un interpréteur $h$ qui prend en entrée une représentation latente $z^{(l)}$. Cet interpréteur peut être un modèle d'apprentissage automatique simple tel qu'un modèle linéaire ou encore plus complexe, tel qu'une machine à vecteurs de support (SVM) ou un réseau de neurones. Ce choix dépendra de la complexité du concept et de la manière dont il est encodé dans les couches cachées. Ainsi, pour chaque concept $c_j$ et chaque représentation latente $z^{(l)}$, on apprend un interpréteur $h_{j,l}$ différent qui retourne une probabilité de prédiction du concept $\hat{c}_j$. Pour entraîner les interpréteurs, on dispose d'une base de concepts provenant d'une base de données, utilisée ou pas dans l'apprentissage du modèle prédictif. On utilise les métriques usuelles d'évaluation des performances en classification (taux d'erreur, F1-score, AUC...) pour déterminer la présence des concepts dans les différentes couches cachées ainsi que le type d'interpréteur le plus performant. Il est possible que les concepts ne soient ni capturés sur la même couche cachée, ni par le même type d'interpréteur. À la fin de cette étape, on dispose d'une liste de taille $K$ de concepts capturés avec $K <= C$.

**Étape 2 :** On examine ensuite les relations logiques entre les concepts capturés et les prédictions du modèle prédictif. Pour cela, nous nous inspirons des règles d'association afin de construire un ensemble de règles logiques que les cardiologues pourront facilement utiliser. Pour ce faire, étant donné un concept $c_j$, on peut étudier la répartition dans l'espace des paires de point $(\hat{y}, \hat{c}_{j,x})_{1,\cdots,N}$ (avec $N$ le nombre d'exemples) et calculer différentes mesures d'intérêt utilisées dans les règles d'association telles que la confiance et le support (Agrawal et al., 1994). Le support est la proportion d'exemples dans l'ensemble de données contenant une occurrence particulière, telle que le concept $c_j$, qu'on pourra exprimer par la probabilité $p(c_j)$. La confiance, quant à elle, se définit comme la probabilité conditionnelle $p(y|c_j)$. On compare généralement ces mesures à un seuil de décision défini par l'utilisateur pour valider ou invalider les règles. Seules les règles respectant ce seuil sont retenues.

## 2.2 Protocole expérimental

**Jeu de données réel** Les données proviennent de la cohorte Generepol (NCT00773201) généré par le centre de recherche d'investigation de la Pitié-Salpétrière à Paris (Salem et al., 2017). Les ECGs de 990 sujets sains ont été enregistrés avant et 1, 2, et 3 h après la prise orale d'une dose de 80-mg de Sotalol (connu pour présenter un risque d'induire la TdP). Dans le cadre de ce travail, nous nous sommes concentrés sur la dérivation II, offrant généralement à elle seule un bon aperçu du signal électrique pour la caractérisation du QT-long (Prifti et al.,

2021). Les ECGs sont enregistrés sur une fenêtre de 10 secondes avec une fréquence d'échantillonnage de 500Hz. Après pré-traitement et normalisation, nous disposions de 10292 ECGs répartis en deux classes : Sot- (avant la prise du médicament Sotalol) et Sot+ (après la prise du médicament). Le jeu est découpé de telle sorte que 10% du jeu est réservé pour une évaluation indépendante (*holdout*) et les 90% restants pour l'entraînement général (avec une répartition en 75% apprentissage, 10% validation et 15% test).

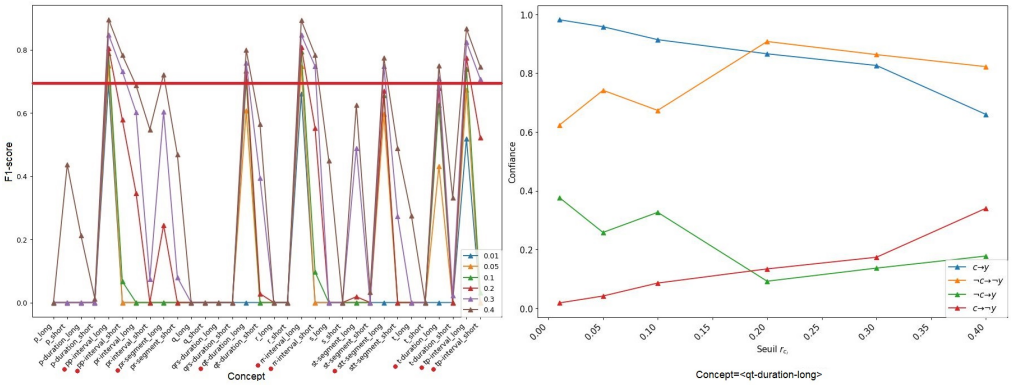
**Modèle prédictif** Le modèle utilisé (Prifti et al., 2021) est un réseau de neurones convolutif densément connecté (DenseNet) qui prend en entrée un électrocardiogramme de patient de 10 secondes sur 5000 points et retourne une prédiction Sot+ ($\hat{y} > 0.5$) ou Sot- ($\hat{y} \leq 0.5$). Ce réseau est formé de blocs convolutifs denses (DenseBlock) reliés par des transitions. Chaque bloc contient plusieurs couches de convolution qui sont toutes connectées directement les unes aux autres. Le dernier bloc du modèle correspond à un perceptron multicouche totalement connecté. Ce modèle présente un taux d'erreur de 2% sur le jeu *holdout*. Les performances complètes du modèle ainsi que son architecture sont décrites et discutées en détail dans l'article d'origine (Prifti et al., 2021).

**Acquisition des concepts** La banque de concepts provient du même jeu de données, mais pourrait venir d'une source extérieure. Les concepts $c_j$ sont établis à partir des positions des ondes et des pics sur l'ensemble du jeu de données. À partir de celles-ci, les amplitudes, ainsi que la durée d'intervalles entre deux positions intra ou inter-battement, sont mesurées pour former une base de seize concepts. Pour un ECG et un concept donnés, le concept est mesuré sur tous les battements de l'ECG et on en calcule la médiane pour n'avoir qu'une seule mesure représentative du concept. Pour chaque concept, on trace ensuite la distribution des médianes obtenues sur l'ensemble des ECGs et on seuille pour déterminer les sous-concepts <long> (vs <normal>) ou <court> (vs <normal>). Les sous-concepts sont binarisés de sorte que la valeur 1 encode le sous-concept <long> (resp. <court>) et 0 <normal>. La valeur du seuil $r_{c_j}$ est déterminée à partir de la distribution du concept issu des ECGs de classe Sot- (groupe témoin). Ce seuil peut être fixé a priori ou être considéré comme un hyper-paramètre dont l'impact sur les résultats sera étudié. Chaque ECG du jeu de données Generepol (Sot- ou Sot+) prend donc deux valeurs de sous-concept par concept en fonction du positionnement de la valeur médiane du concept vis-à-vis du seuil choisi. Nous avons ainsi constitué une base de 32 concepts en incluant les sous-concepts. Nous recherchons la présence de ces concepts dans les couches cachées de la partie prédictive du modèle (le perceptron multicouche), après les blocs denses de convolution qui sont en charge d'extraire des motifs abstraits des données.

**Interpréteur** Rappelons que pour chaque concept étudié $c_j$, l'interpréteur $h_{j,l}$ prend en entrée la représentation latente choisie $z^{(l)}$ d'un ECG. Ici, chaque interpréteur est un réseau de neurones à une seule couche cachée dont le nombre de neurones ($n_{hidden}$) dépend du nombre de variables d'entrée ($n_{in}$) de $z^{(l)}$ selon la règle suivante : $n_{hidden} = \frac{n_{in}}{4}$. Pour apprendre les interpréteurs, nous avons utilisé le même jeu d'entraînement que celui utilisé par le modèle prédictif, sachant que le nombre d'exemples peut légèrement différer dans chacun des sets du fait de l'existence de valeurs manquantes ou aberrantes dans les annotations. D'autres types d'interpréteurs peuvent être considérés. Néanmoins, les premiers tests ont montré qu'un mo-

dèle linéaire retrouvait les concepts bien moins efficacement qu'un modèle non-linéaire (MLP ou SVM).

# 3 Résultats




(a) (b)

| Concept | Règles |
|---|---|
| pp-interval_long | $c \rightarrow y$ (0,1) |
| pp-interval_short | $c \rightarrow \neg y$ |
| rr-interval_long | $c \rightarrow y$ |
| rr-interval_short | $c \rightarrow \neg y$ (0,1) |
| stt-segment_long | $c \rightarrow y$ |

| Concept | Règles |
|---|---|
| qt-duration_long | $c \rightarrow y$<br>$y \rightarrow c$ |
| tp-interval_long | $c \rightarrow y$ |
| tp-interval_short | $c \rightarrow \neg y$ |
| t-duration_long | $c \rightarrow y$ |

(c)

FIG. 2 – (a) Évaluation des performances des interpréteurs de type réseau de neurones sur le jeu de test selon le F1-score en fonction de l'ensemble des 32 concepts et des seuils $r_{c_j}$ (b) Évaluation des différentes relations logiques entre le concept $c$ = <qt-duration-long> et la prédiction $y$ = Sot+ selon le score de confiance en fonction du seuil $r_c$. $\neg c$ représente le concept <normal> et $\neg y$ représente la classe Sot-. (c) Ensemble des relations logiques extraites avec $r_{c_j} = 0.2$ dans la majorité des cas, sauf à 0.1 pour les cas précisés entre parenthèses.

Concernant les résultats de l'étape 1, rappelons qu'en faisant varier à la fois le concept $c_j$, la couche $(l)$ et le seuil $r_{c_j}$, cela emmène à réapprendre systématiquement un nouvel interpréteur. Nous exposons en Fig. 2 les résultats obtenus avec $l$ correspondant à la couche d'entrée du perceptron multicouche du modèle prédictif. En effet, des tests préliminaires ont montré que cette couche permettait de mieux prédire les concepts contrairement aux couches plus profondes du perceptron multicouche. La Fig. 2a présente les performances des interpréteurs de type réseau de neurones pour chaque concept en faisant varier également le seuil $r_{c_j}$. On remarque que les concepts entre deux battements <pp-interval-{long,short}, rr-interval-{long,short}, tp-interval-{long,short}> et intra-battement <qt-duration-long, stt-segment-long, t-duration-long> sont capturés avec un F1-score supérieur à 0.7. Aucun concept sur les ampli-

tudes semble être présent. Les performances varient également avec le seuil $r_{c_j}$. On peut par exemple noter une différence de performances de 0.2 sur le concept <rr-interval-long> entre deux valeurs de seuil différent (0.4 et 0.01). Concernant les résultats de l'étape 2, l'objectif est d'établir les relations logiques sur la liste $K$ de concepts retenus à l'étape précédente (F1-score > 0.7). Les résultats préliminaires sur le concept $j =$ <qt-duration-long> sont présentés dans la Fig.2b en utilisant la confiance comme métrique. Cela permet d'avoir un premier aperçu des implications logiques intéressantes. Cette figure illustre l'évolution du score de confiance en fonction du seuil $r_c$ de la règle inspectée. Ce score, variant entre 0 et 1, indique la force de l'association, où un score plus élevé signifie une association plus forte. On observe que les règles logiques $(c \rightarrow y)$ et $(\neg c \rightarrow \neg y)$ se démarquent avec un score supérieur à 0.6. Deux tendances émergent également : avec un seuil plus élevé, le score de confiance tend à augmenter pour les règles $\{(\neg c \rightarrow \neg y), (c \rightarrow \neg y)\}$, tandis qu'il a tendance à diminuer pour les règles $\{(c \rightarrow y), (\neg c \rightarrow y)\}$. Il semble y avoir un compromis à trouver dans le choix du seuil. Un seuil élevé signifie un équilibre entre les données d'entraînement présentant et ne présentant pas le concept. En revanche, un seuil bas signifie que peu de données le présentent. Dans le cas exposé, le seuil 0.2 semble être le plus approprié, où les règles $(c \rightarrow y)$ et $(\neg c \rightarrow \neg y)$ ont toutes deux un score de confiance entre 0.8 et 0.9. Cette analyse peut être répétée pour tous les concepts retenus à l'étape précédente, permettant ainsi de définir un ensemble de règles logiques pour fournir une interprétation globale du modèle. Une première estimation des règles est présentée dans le tableau 2c.

# 4 Conclusion et perspectives

Ces premières expériences ont démontré que notre approche est efficace pour détecter la présence de concepts dans les couches cachées du modèle prédictif et établir des relations logiques entre les concepts et les prédictions. Sur le plan méthodologique, des expériences approfondies sont nécessaires pour valider l'approche, notamment en explorant différents types d'interpréteurs non-linéaires tels qu'un SVM. Ensuite, il existe plus d'une vingtaine de mesures dans la littérature (Lenca et al., 2007) pour évaluer les règles d'association, comme l'indépendance et l'absence de contre-exemples. Celles-ci pourraient être utilisées pour choisir les relations logiques les plus adaptées. L'utilisation d'algorithmes d'extraction d'items-sets fréquents [2], tels qu'Apriori (Agrawal et al., 1994), permettrait également de déduire des règles de combinaison de concepts. La validation de l'approche sur d'autres ensembles de données et types de réseaux neuronaux est aussi essentielle. Enfin, il est important d'évaluer et de valider l'interprétation construite de manière qualitative et quantitative par des critères spécifiques (Islam et al., 2020), mais aussi auprès des cardiologues pour vérifier l'alignement des règles logiques avec les connaissances du domaine. Notons déjà que la présence du concept <qt-duration-long> dans les règles est cohérente avec les connaissances du domaine.

# Financement

Cette étude a été soutenue par le financement ANR-20-CE17-0022 DeepECG4U de l'Agence Nationale de la Recherche.

2. Par définition, un item-set est un ensemble d'items correspondant dans notre cas aux concepts.

# Références

Agrawal, R., R. Srikant, et al. (1994). Fast algorithms for mining association rules. In *Proceedings of 20th International Conference on Very Large Data Bases*, Volume 1215.

Crabbé, J. et M. van der Schaar (2022). Concept activation regions : A generalized framework for concept-based explanations. In *Advances in Neural Information Processing Systems*, Volume 35, pp. 2590–2607.

Goodman, B. et S. Flaxman (2017). European union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine 38*(3), 50–57.

Islam, S. R., W. Eberle, et S. K. Ghafoor (2020). Towards quantification of explainability in explainable artificial intelligence methods. In *Proceedings of the Thirty-Third International Florida Artificial Intelligence Research Society Conference*, pp. 75–81. AAAI Press.

Kim, B., M. Wattenberg, J. Gilmer, C. J. Cai, J. Wexler, F. B. Viégas, et R. Sayres (2018). Interpretability beyond feature attribution : Quantitative testing with concept activation vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning*, Volume 80, pp. 2673–2682. PMLR.

Lenca, P., B. Vaillant, P. Meyer, et S. Lallich (2007). Association rule interestingness measures : Experimental and theoretical studies. In *Quality Measures in Data Mining*, Studies in Computational Intelligence, pp. 51–76. Springer.

Prifti, E., A. Fall, G. Davogustto, A. Pulini, I. Denjoy, C. Funck-Brentano, Y. Khan, A. Durand-Salmon, F. Badilini, Q. S. Wells, A. Leenhardt, J.-D. Zucker, D. M. Roden, F. Extramiana, et J.-E. Salem (2021). Deep learning analysis of electrocardiogram for risk prediction of drug-induced arrhythmias and diagnosis of long QT syndrome. *European Heart Journal 42*(38), 3948–3961.

Salem, J.-E., M. Germain, J.-S. Hulot, P. Voiriot, B. Lebourgeois, J. Waldura, D.-A. Tregouet, B. Charbit, et C. Funck-Brentano (2017). Genome wide analysis of sotalol-induced ikr inhibition during ventricular repolarization,"generepol study" : Lack of common variants with large effect sizes. *PLoS One 12*(8), e0181875.

Zaeem, M. N. et M. Komeili (2021). Cause and effect : Concept-based explanation of neural networks. In *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 2730–2736.

# Summary

The development of artificial intelligence contributes to the emergence of a new form of personalized medicine, aiming to better consider patients' characteristics. In this context, we focus on the application of deep learning to detect cardiovascular diseases, such as Long QT syndrome, from electrocardiograms. However, a major concern lies in the often opaque nature of these so-called "black box" AI models. This article aims to make these models more interpretable by integrating medical concepts into a new post-hoc approach, ConceptECGxAI, providing more understandable explanations to physicians. Experiments have shown that the approach can identify the presence of meaningful concepts in the hidden layers of the predictive model and establish logical relationships between these concepts and predictions.