



HAL
open science

Prédiction de la fibrillation atriale à l'aide des signaux physiologiques

Mohamed Mroueh, Farah Mourad-Chehade, Fahed Abdallah

► **To cite this version:**

Mohamed Mroueh, Farah Mourad-Chehade, Fahed Abdallah. Prédiction de la fibrillation atriale à l'aide des signaux physiologiques. Journées d'Etude sur la Télésanté, Sorbonne Universités, May 2019, Paris, France. hal-02161445

HAL Id: hal-02161445

<https://hal.science/hal-02161445>

Submitted on 20 Jun 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Prédiction de la fibrillation atriale à l'aide des signaux physiologiques

MROUEH Mohamed^(1,2), MOURAD-CHEHADE Farah⁽¹⁾, ABDALLAH Fahed^(1,2)

¹ *Institut Charles Delaunay, ROSAS, M2S, Université de Technologie de Troyes, Troyes, France*

² *EDST, Université Libanaise, Beyrouth, Liban*

mohamed.mroueh@utt.fr, farah.chehade@utt.fr, fahed.abdallah76@gmail.com

Abstract—La fibrillation atriale est une maladie du cœur qui se caractérise par une irrégularité des battements cardiaques pouvant causer d'autres complications comme un arrêt cardiaque ou la démence. Le but de cette étude est de prédire la survenue de la fibrillation atriale en se basant sur des signaux physiologiques mesurables à l'aide de capteurs non-invasifs. L'approche utilisée consiste à commencer par extraire des caractéristiques des signaux puis de construire une règle de décision en se basant sur la théorie des fonctions de croyances. Les données physiologiques ont été fournies par la base médicale MIMIC III. Le classifieur obtenu a prédit l'occurrence de la maladie avec une précision de 70,49%, une sensibilité de 77.07% et une spécificité de 63,9%.

Mots-clés—Fibrillation atriale, fonctions de croyance, fusion de données, apprentissage automatique, MIMIC III, prédiction.

I. INTRODUCTION

La Fibrillation Atriale ou Fibrillation Auriculaire (FA) est une irrégularité des battements du cœur, aussi appelée arythmie, émanant d'une mauvaise synchronisation des cellules myocardiques auriculaires qui résulte en des contractions irrégulières et rapides des oreillettes cardiaques. Les principales causes de la survenue de la fibrillation atriale sont une insuffisance mitrale, une hypertension artérielle et une insuffisance cardiaque. Cette maladie touche plus de 33 millions de personnes à travers le monde. Environ 10% des personnes âgées de plus de 80 ans en sont atteintes. La FA est un facteur de risque sérieux pour d'autres complications, telles que la thrombose, l'accident vasculaire cérébral ou l'altération de la fonction cardiaque.

La prédiction de cette maladie se trouve donc être une tâche importante pour permettre une intervention médicale anticipée [1]–[3]. Pour ce faire, on se sert de l'apprentissage automatique supervisé ou Machine Learning, qui consiste à traiter des données du patient dans le but d'en prédire l'état de santé. Cette prédiction passe par une branche spécifique du Machine Learning appelée la classification. Pour ne citer que quelques méthodes classiques de classification, on présente le SVM (Support Vector Machine) qui consiste à construire un classifieur à l'aide d'une tâche d'optimisation où il est question de maximiser un critère de marge, la méthode des k plus proches voisins qui attribue à une instance une classe en fonction de celles des k instances les plus proches de ce dernier selon un critère métrique précis [4], [5]. La théorie des fonctions de croyance, initialement introduite par Dempster puis développée ensuite par Shafer, est une extension

de la théorie probabiliste qui permet de prendre en compte l'incertitude ou l'incomplétude des données [6]. L'utilisation de cette théorie dans le cadre de la classification permet de prendre en considération différents niveaux de contributions pour chaque source d'information ce qui améliore la qualité de la prédiction.

Ce papier présente une approche originale pour la prédiction de la FA basée sur la théorie des fonctions de croyance. D'abord, des caractéristiques sont extraites des signaux physiologiques et une distribution est adaptée à chacune d'elles. Ces distributions sont ensuite utilisées pour construire les fonctions de masse, composants essentiels de la théorie des fonctions de croyance, qui seront par la suite combinées entre elles pour permettre de conclure le test en prenant une décision. Un patient est alors considéré à même de développer la maladie si à un instant donné le test est positif.

Ce document est ensuite structuré comme suit : Le matériel utilisé est présenté en Section II, puis la méthodologie suivie est détaillée en Section III, les résultats en Section IV et finalement la Section V conclue l'étude en ouvrant de nouvelles perspectives.

II. MATÉRIEL

A. Base de données MIMIC III

La base de données médicale MIMIC III (Medical Information Mart for Intensive Care III) est une ressource dédiée à la recherche médicale regroupant des informations sanitaires relevées dans l'Unité de Soins Intensifs (USI) d'un hôpital sur plus de 40.000 patients de 2001 à 2012 [7]. Le contenu de cette base de données est très varié puisqu'on y retrouve des détails précis sur le séjour des patients dans l'USI comme les diagnostics prononcés, les résultats des tests cliniques et du laboratoire, l'évolution des mesures médicales ou les signaux vitaux tels que les électrocardiogrammes. Cette base se divise en deux parties. La première est une collection de tables qui contiennent des informations cliniques sur les patients comme leurs admissions à l'hôpital, les codes ICD9 des diagnostics, les notes des médecins ou encore les procédures entreprises. La deuxième constitue les enregistrements numériques que ce soit à série temporelle (1 échantillon par minute), comme le rythme cardiaque, ou en continu (125 échantillons par seconde), comme les électrocardiogrammes. Bien que les deux parties soient distinctes, une correspondance a été établie entre les données cliniques et les enregistrements pour une partie des

patients. Étant donné que notre étude se base sur les signaux physiologiques pour une certaine maladie, nous avons besoin de cette correspondance.

B. Sélection des patients

La sélection des patients dans la base de données s'effectue grâce aux diagnostics prononcés par les médecins et qui figurent dans l'une des tables de la partie de la base réservée aux données cliniques. Cette dernière indexe les maladies par leur code ICD9. Dans le cas de la FA, le code ICD9 est 42731. Les durées des signaux diffèrent d'un patient à l'autre, ainsi, les patients atteints de FA ayant au minimum 2 heures d'enregistrement avant l'occurrence de la maladie sont au nombre de 733. Ce nombre décroît à 123 lorsque l'on sélectionne les patients ayant au moins 48 heures d'enregistrement avant la FA. L'occurrence de la maladie est détectable grâce à un critère médical précis qui se résume à localiser dans le signal du rythme cardiaque une séquence de mesures consécutives d'une durée de 15 min ayant des valeurs supérieures ou égales à 110 battements par minute [8]. La Figure 1 montre l'évolution du nombre de patients atteints de la FA en fonction de la durée d'enregistrement exigée avant l'occurrence de la maladie.

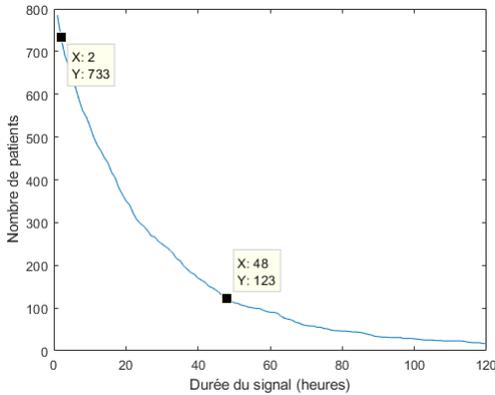


Fig. 1. Nombre de patients disponibles en fonction des durées requises

C. Choix des signaux physiologiques

Dans le but de choisir les signaux physiologiques à utiliser pour la prédiction, il est judicieux d'étudier la disponibilité de chacun d'entre eux car ceci permet de construire une base de données riches en informations utiles et éviter de stresser l'algorithme de classification plus tard. En sélectionnant uniquement les types de signaux physiologiques apparaissant dans plus de 40% de la population concernée, on aboutit à un ensemble de 5 signaux qui sont :

- Le Rythme Cardiaque ou RC (100 %),
- Le Rythme Respiratoire ou RR (92.15 %),
- Les Impulsions (81.77 %),
- La Saturation Pulsée en Oxygène ou SpO2 (46.13 %),
- La Pression Veineuse Centrale ou PVC (41.12 %).

La Figure 2 montre un exemple de ces 5 signaux pour un sujet atteint de la FA.

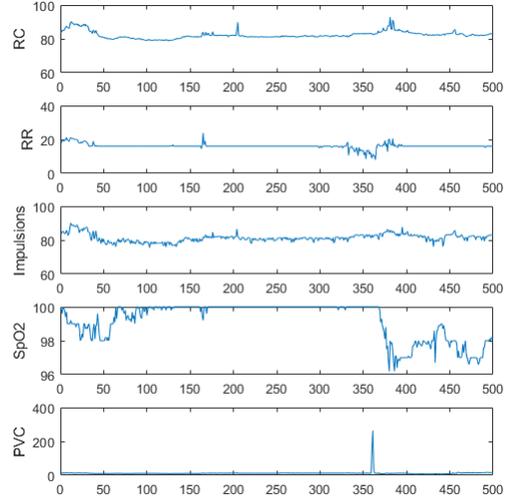


Fig. 2. Exemple des 5 signaux pour un patient

D. Sélection des patients de contrôle

Dans le but d'obtenir une base de données équilibrée, on a besoin de sujets non-atteints de la FA. Pour cela, on sélectionne dans la base de données des patients n'ayant aucun antécédent cardiaque et ayant été déchargés vivants de l'USI en prenant soin de choisir le même nombre que celui des patients atteints. Ainsi, on obtient deux groupes de patients que l'on nommera par la suite groupe sain et groupe malade, chaque sujet ayant un certain nombre de signaux physiologiques. Il est à noter que pour les sujets malades, seules les parties des signaux précédant l'occurrence de la maladie sont gardées pour que l'on reste dans la prédiction.

III. MÉTHODOLOGIE

L'objectif de cette étude est de proposer une approche de prédiction de la FA se basant sur les signaux physiologiques des sujets. Cette approche se base sur une extraction de caractéristiques des signaux physiologiques en temps réel, puis sur une conception d'un classifieur, dans le cadre de la Théorie des Fonctions de Croissance, qui prend en entrée les caractéristiques et rend en sortie une prédiction concernant l'occurrence de la FA dans un futur proche. Afin de construire le classifieur, le jeu de données obtenu de MIMIC III est divisé aléatoirement en deux parties, l'une pour l'apprentissage avec 2/3 des sujets et l'autre pour le test avec les sujets restants. Pour la suite, on note I_A l'ensemble des indices des sujets de la base d'apprentissage, I_T celui de la base de test et $I = I_A \cup I_T$ l'ensemble de tous les indices.

A. Environnement de la méthode

Soit $u_{i,j}(t)$ le signal de type j ($1 \leq j \leq 5$) du patient indexé par $i \in I$ mesuré à un instant t et soit $\mathbf{u}_{i,j}(t - N + 1 : t)$ le vecteur de taille N incluant les dernières valeurs du signal $u_{i,j}(\cdot)$ jusqu'à l'instant t . La longueur N pourra être

choisie égale à 1, 2, 3... heures ou optimisée sur la base d'apprentissage. Le but est de construire un classifieur qui prend en entrée le vecteur $\mathbf{u}_{(i,j)}(t-N+1:t)$ et donne en sortie à l'instant t une estimation $\hat{y}_i(t) \in \{0, 1\}$, avec $\hat{y}_i(t) = 1$ indiquant que le sujet développera dans l'avenir la FA et 0 sinon. Ce classifieur appliqué à toute la durée du signal constitue la courbe de risque $\hat{y}_i(\cdot)$.

B. Extraction des caractéristiques

Cette section décrit en détail les caractéristiques extraites des segments de signaux $\mathbf{u}_{i,j}(t-N+1:t)$ du sujet $i \in I$, à un instant t , et pour tout type de signal j . Les différentes caractéristiques extraites du vecteur $\mathbf{v}_{i,j,t} = \mathbf{u}_{i,j}(t-N+1:t)$ sont :

- La moyenne

$$f_1(\mathbf{v}_{i,j,t}) = \frac{1}{N} \sum_{k=1}^N u_{i,j}(t-N+k),$$

- L'écart-type

$$f_2(\mathbf{v}_{i,j,t}) = \sqrt{\frac{1}{N} \sum_{k=1}^N (u_{i,j}(t-N+k) - f_1(\mathbf{v}_{i,j,t}))^2},$$

- Le skewness

$$f_3(\mathbf{v}_{i,j,t}) = \frac{1}{N} \sum_{k=1}^N \left(\frac{u_{i,j}(t-N+k) - f_1(\mathbf{v}_{i,j,t})}{f_2(\mathbf{v}_{i,j,t})} \right)^3,$$

- Le kurtosis

$$f_4(\mathbf{v}_{i,j,t}) = \frac{1}{N} \sum_{k=1}^N \left(\frac{u_{i,j}(t-N+k) - f_1(\mathbf{v}_{i,j,t})}{f_2(\mathbf{v}_{i,j,t})} \right)^4.$$

Cette analyse conduit à l'obtention de 20 valeurs de caractéristiques à partir d'un segment de taille N comportant 5 types de signaux. Ces valeurs sont représentées par un vecteur $\mathbf{x}_i(t)$ de l'espace des caractéristiques, dont les éléments $x_{i,\ell}(t) = f_n(\mathbf{v}_{i,j,t})$ avec $\ell = (j-1) \times 4 + n$, $1 \leq j \leq 5$ et $1 \leq n \leq 4$.

Afin de construire le classifieur par la suite, il est judicieux de transformer les signaux de la base de données d'apprentissage en caractéristiques. Ainsi, pour les patients du groupe malade, les caractéristiques sont extraites du segment des N dernières mesures effectuées avant la manifestation de la maladie, et pour le groupe sain, elles sont prises du dernier segment avant la décharge de l'USI. Soit T_i la longueur des signaux du sujet i . Ainsi, pour chaque sujet $i \in I_A$ du jeu de données d'apprentissage, on obtient un vecteur de caractéristiques $\mathbf{x}_i = \mathbf{x}_i(T_i) = (x_{i,\ell}(T_i))_{1 \leq \ell \leq 20}$, auquel est associé un label $y_i = 1$ si le sujet i appartient au groupe malade et $y_i = 0$ sinon. Cela conduit à une nouvelle formulation de la base d'apprentissage du classifieur :

$$\mathcal{A} = \{(\mathbf{x}_i, y_i) \mid i \in I_A\},$$

avec I_A l'ensemble des indices des sujets de la base d'apprentissage.

C. Construction du classifieur utilisant la théorie des fonctions de croyance

L'idée principale de cette approche est de se servir de chaque caractéristique comme une source d'information concernant le groupe du patient dont l'état est à prédire. En utilisant la théorie des fonctions de croyance, on modélise les contributions des caractéristiques à la classification par le biais de fonctions de masse. Ces fonctions seront, par la suite, combinées grâce à une méthode de combinaison prédéfinie dans le but de réunir toutes les informations et de prendre une décision.

1) *Modélisation des données:* Dans le but de définir les fonctions de masse, on commence par poser le modèle qui se déduit de la base d'apprentissage \mathcal{A} . On suppose que les valeurs des caractéristiques sont réparties selon une loi normale dont les paramètres sont à estimer en maximisant la fonction de vraisemblance. Pour chacun des deux groupes, malade et sain, on adapte une loi normale en estimant son espérance et son écart-type. De ce fait, pour chaque caractéristique ℓ , on obtient les formules suivantes :

$$\hat{\mu}_\ell^{(1)} = \frac{\sum_{i \in I_A} x_{i,\ell}(T_i) \cdot y_i}{\sum_{i \in I_A} y_i},$$

$$\hat{\sigma}_\ell^{(1)} = \sqrt{\frac{\sum_{i \in I_A} (x_{i,\ell}(T_i) - \hat{\mu}_\ell^{(1)})^2 \cdot y_i}{\sum_{i \in I_A} y_i}},$$

$$\hat{\mu}_\ell^{(0)} = \frac{\sum_{i \in I_A} x_{i,\ell}(T_i) \cdot (1 - y_i)}{\sum_{i \in I_A} (1 - y_i)},$$

$$\hat{\sigma}_\ell^{(0)} = \sqrt{\frac{\sum_{i \in I_A} (x_{i,\ell}(T_i) - \hat{\mu}_\ell^{(0)})^2 \cdot (1 - y_i)}{\sum_{i \in I_A} (1 - y_i)}}.$$

La Figure 3 montre les distributions estimées pour les 20 caractéristiques pour une certaine séparation aléatoire de la base de données.

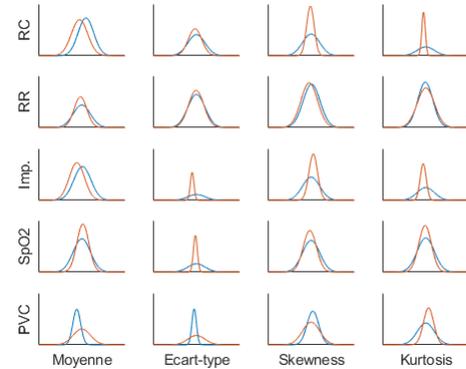


Fig. 3. Distributions des caractéristiques pour le groupe malade (bleu) et non-malade (rouge)

2) *Cadre de la théorie des fonctions de croyance*: La principale utilité de la théorie des fonctions de croyance est la prise en compte des lacunes informationnelles. La modélisation de cette imprécision se fait à travers l'attribution de masse à des ensembles d'évènements. Dans notre cas, les éventuelles possibilités pour un sujet sont Malade et Sain, en faisant correspondre les états comme suit :

- Malade : 1,
- Sain : 0.

On obtient donc un univers d'évènements constitué de ces 2 éléments :

$$\Omega = \{0, 1\}.$$

Les masses seront donc attribuées à tous les sous-ensembles de Ω .

3) *Attribution des masses*: L'attribution des masses est une étape très importante dans la théorie des fonctions de croyance car elle a un impact considérable sur les performances du classificateur. À chaque caractéristique ℓ , sera associée une fonction de masse $m_\ell(\omega, x)$ avec $\omega \in \{\emptyset, \{0\}, \{1\}, \{0, 1\}\}$ pour une certaine valeur x . D'abord, on modélise l'incertitude en calculant, pour chaque caractéristique, l'aire d'intersection entre les fonctions de densité de probabilité des lois normales estimées dans III-C1,

$$m_\ell(\{0, 1\}, x) = \int_{-\infty}^{+\infty} \min \left(\phi \left(\frac{x - \hat{\mu}_\ell^{(1)}}{\hat{\sigma}_\ell^{(1)}} \right), \phi \left(\frac{x - \hat{\mu}_\ell^{(0)}}{\hat{\sigma}_\ell^{(0)}} \right) \right) dx.$$

Ici, $\phi(\cdot)$ est la fonction de densité de probabilité de la loi normale centrée réduite. Évidemment, cette masse ne dépend pas de la variable x . Ceci permet de traduire la probabilité d'erreur en une valeur qui mesure l'incertitude. On attribue une masse nulle à l'ensemble vide car on est sûr que l'état sera malade ou sain,

$$m_\ell(\emptyset, x) = 0.$$

Pour le reste, la valeur de la masse dépend de la caractéristique x et elle est choisie proportionnellement à la valeur de la fonction de densité de probabilité en x . En plus, la somme de toutes les masses doit être égale à l'unité ce qui renvoie aux formules suivantes :

$$m_\ell(\{1\}, x) = \frac{(1 - m_\ell(\{0, 1\}, x)) \cdot \phi \left(\frac{x - \hat{\mu}_\ell^{(1)}}{\hat{\sigma}_\ell^{(1)}} \right)}{\phi \left(\frac{x - \hat{\mu}_\ell^{(0)}}{\hat{\sigma}_\ell^{(0)}} \right) + \phi \left(\frac{x - \hat{\mu}_\ell^{(1)}}{\hat{\sigma}_\ell^{(1)}} \right)},$$

$$m_\ell(\{0\}, x) = \frac{(1 - m_\ell(\{0, 1\}, x)) \cdot \phi \left(\frac{x - \hat{\mu}_\ell^{(0)}}{\hat{\sigma}_\ell^{(0)}} \right)}{\phi \left(\frac{x - \hat{\mu}_\ell^{(0)}}{\hat{\sigma}_\ell^{(0)}} \right) + \phi \left(\frac{x - \hat{\mu}_\ell^{(1)}}{\hat{\sigma}_\ell^{(1)}} \right)}.$$

Ceci garantit d'attribuer une masse en conformité avec la distribution des données dans la base d'apprentissage.

4) *Combinaison des fonctions de masses*: Après avoir défini les valeurs des fonctions de masses, on obtient une contribution par caractéristique à l'appartenance du sujet à l'un des deux groupes malade et sain. Il reste encore à combiner ces sources d'informations dans le but d'obtenir une décision finale. Dans le cadre de la théorie des fonctions de croyance, ceci est faisable grâce à la combinaison, méthode qui prend en entrée des fonctions de masses et qui en déduit une nouvelle plus informative. Plusieurs méthodes de combinaison existent dans la littérature comme la méthode conjonctive ou disjonctive. Mais dans le but d'éviter l'évanouissement de l'incertitude ($m(\{0, 1\}, x) \approx 0$), on choisit une méthode qui stabilise cette valeur au cours des combinaisons successives. Une des méthodes satisfaisant cette condition est la méthode de YAGER [9] décrite ci-dessous par les équations suivantes restreintes à notre cas pour $a, b \in \{1, 2, \dots, 20\}$:

$$m_{a \odot b}(\{1\}, x) = \sum_{A \cap B = \{1\}} m_a(A, x) \cdot m_b(B, x),$$

$$m_{a \odot b}(\{0\}, x) = \sum_{A \cap B = \{0\}} m_a(A, x) \cdot m_b(B, x),$$

$$m_{a \odot b}(\Omega, x) = \sum_{A \cap B \in \{\emptyset, \Omega\}} m_a(A, x) \cdot m_b(B, x).$$

La méthode est appliquée successivement à toutes les fonctions de masse pour obtenir une fonction de masse finale englobant les informations de toutes les caractéristiques du sujet à étudier,

$$m(\cdot, \mathbf{x}_i(t)) = m_1(\cdot, \mathbf{x}_{i,1}(t)) \odot m_2(\cdot, \mathbf{x}_{i,2}(t)) \dots m_{20}(\cdot, \mathbf{x}_{i,20}(t)).$$

5) *Transformation et décision*: Afin d'obtenir une valeur de probabilité pour chaque élément de Ω , il nous faut transmettre les masses des ensembles à plusieurs éléments aux singletons. Ceci est assuré par des méthodes de transformation visant à obtenir une distribution de probabilité. Dans notre étude, on a utilisé la méthode dite Pignistique pour sa simplicité et son efficacité. Cette dernière distribue les masses des ensembles de façon équitable sur les éléments qui les composent,

$$P(1, \mathbf{x}_i(t)) = m(\{1\}, \mathbf{x}_i(t)) + \frac{1}{2}m(\{0, 1\}, \mathbf{x}_i(t)),$$

$$P(0, \mathbf{x}_i(t)) = m(\{0\}, \mathbf{x}_i(t)) + \frac{1}{2}m(\{0, 1\}, \mathbf{x}_i(t)).$$

Une fois les masses transformées en probabilité, la décision peut se faire naturellement en comparant les probabilités entre elles. On décide donc que le sujet est malade si $P(1, \mathbf{x}_i(t)) > P(0, \mathbf{x}_i(t))$. Une autre approche serait de considérer un niveau de risque pour la prédiction. En effet, il s'agit d'un certain seuil γ au-delà duquel on décide d'attribuer le statut malade au sujet suivi, c'est-à-dire lorsque $P(1, \mathbf{x}_i(t)) > \gamma$. Ainsi, on applique la méthode détaillée en III-C sur la plus récente fenêtre de taille N disponible. Cela revient à glisser une fenêtre sur le signal du sujet en appliquant continuellement la méthode

de classification. Ceci conduit donc à une courbe de risque $C_i(t) = P(1, \mathbf{x}_i(t))$ pour le patient i . Ensuite, un sujet est considéré malade si, à un certain instant t , la courbe de risque dépasse le seuil γ . La décision ζ_i se définit donc comme suit

$$\zeta_i = \llbracket \max_t(C_i(t)) > \gamma \rrbracket,$$

avec $\llbracket \cdot \rrbracket$ la fonction qui prend la valeur 1 si la proposition \cdot est vraie et 0 sinon.

D. Illustration des performances de la méthode

1) *Définition de la base de test*: Dans le but d'évaluer les performances de la méthode, on se sert de la partie test de la base de données obtenue d'après la séparation expliquée en III. La base de test se présente comme suit :

$$\mathcal{T} = \{(\mathbf{X}_i, y_i) \mid i \in I_{\mathcal{T}}\},$$

avec $I_{\mathcal{T}}$ l'ensemble des indices de la base de test et \mathbf{X}_i la matrice constituée des lignes $\mathbf{x}_i(k\delta)$ avec δ une durée qui représente le pas de la fenêtre glissante et k l'indice de la ligne dans la matrice. On rappelle que $\mathbf{x}_i(\cdot)$ est le vecteur des caractéristiques extraites des signaux (voir III-B). Le nombre de lignes de cette matrice dépend donc de la longueur des signaux et de la valeur de δ et le nombre de colonnes est 20, 4 caractéristiques pour chacun des 5 signaux sélectionnés en II-C.

2) *Critères de performances*: Pour observer les évaluations numériquement, on se sert de certains critères de performances qui sont la précision $Pr(\mathcal{T})$, la sensibilité $Se(\mathcal{T})$ et la spécificité $Sp(\mathcal{T})$. Elles découlent des résultats ζ_i et des actuels états y_i ($i \in I_{\mathcal{T}}$) sont définies par les équations suivantes :

$$Pr(\mathcal{T}) = \frac{1}{\text{card}(I_{\mathcal{T}})} \sum_{i \in I_{\mathcal{T}}} \llbracket \zeta_i = y_i \rrbracket,$$

$$Se(\mathcal{T}) = \frac{\sum_{i \in I_{\mathcal{T}}} \llbracket \zeta_i = 1 \rrbracket \cdot \llbracket y_i = 1 \rrbracket}{\sum_{i \in I_{\mathcal{T}}} \llbracket y_i = 1 \rrbracket},$$

$$Sp(\mathcal{T}) = \frac{\sum_{i \in I_{\mathcal{T}}} \llbracket \zeta_i = 0 \rrbracket \cdot \llbracket y_i = 0 \rrbracket}{\sum_{i \in I_{\mathcal{T}}} \llbracket y_i = 0 \rrbracket}.$$

IV. RÉSULTATS

On expose dans cette section les résultats obtenus après avoir effectué quelques tests pour une durée de signal de 48h. Ce choix aboutit à 123 patients composant le groupe malade, ce qui pousse à sélectionner arbitrairement 123 patients composant le groupe non-malade. La base de données est donc constituée de 246 patients dont 164 pour l'apprentissage et 82 pour les tests. Le processus de sélection est répété 30 fois et les résultats affichés sont les moyennes des 30 obtenus. Afin de maximiser les valeurs des performances, on optimise les valeurs de γ et N en les faisant varier. On obtient une matrice de précisions illustrée par la Figure 4.

Dans la Figure 4, on remarque que le seuil a un impact beaucoup plus important que la taille de la fenêtre sur la valeur de la précision. D'ailleurs, les précisions maximales

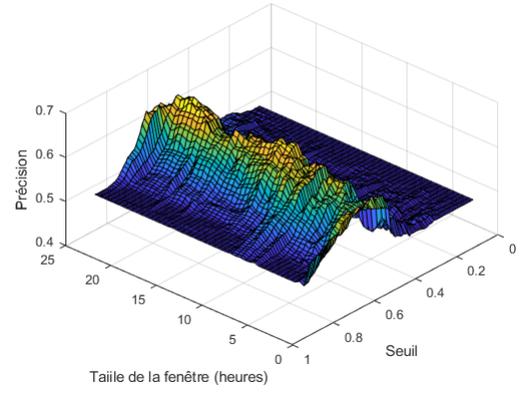


Fig. 4. Précision en fonction de la taille de la fenêtre et du seuil γ

TABLE I
PERFORMANCES EN FONCTION DU SEUIL POUR $N = 60$

Seuil	Précision	Sensibilité	Spécificité
0,43	0,5756	0,8829	0,2683
0,47	0,6073	0,8439	0,3707
0,51	0,6976	0,7756	0,6195
0,55	0,7049	0,7512	0,6585
0,59	0,6951	0,6878	0,7024
0,6	0,6805	0,6	0,761

pour chaque longueur de fenêtre varient avec un écart-type de 0,0185 ce qui est relativement faible. On a donc tendance à privilégier la précocité de la détection de la maladie ce qui se traduit par un choix de longueur de fenêtre plus réduit. En se focalisant sur $N = 60$, ce qui signifie une fenêtre de prédiction de durée d'une heure, on obtient une précision maximale pour un seuil $\gamma = 0,53$, seuil qui sera donc utilisé dans l'évaluation des performances de la méthode en utilisant la base de test. En considérant $N = 60$ et $\gamma = 0,53$, on obtient les résultats suivants sur la base de test :

- Précision : 70,49%
- Sensitivité : 77,07%
- Spécificité : 63,9%

Le Tableau I montre qu'il est possible de privilégier un critère de performance sur l'autre en faisant varier la valeur du seuil γ .

V. CONCLUSION

L'objectif du travail présenté est de se servir des prouesses de l'intelligence artificielle pour faire de la prédiction médicale. Dans ce contexte, aucune expertise médicale avancée est nécessaire, puisqu'il s'agit simplement et purement de la manipulation des données. Nous avons d'abord extrait des signaux physiologiques de la base de données MIMIC III. Ces signaux concernent les patients atteints de la fibrillation atriale. Puis, nous avons contrebalancé ceci avec des patients non-atteints d'une quelconque maladie cardiaque. Les signaux ont été ensuite considérés pour en extraire des caractéristiques diverses. Celles-ci sont offertes à une méthode de classification se basant sur la théorie des fonctions de croyance qui se

charge de sortir une courbe de risque pour un sujet dont l'état serait à prédire. Les tests montrent que la méthode a prédit correctement 70% des sujets de test dont 77% de malades (sensibilité) et 64% de non-malades (spécificité). Néanmoins, le contrôle d'un paramètre de seuillage permet le privilège d'un critère de performance sur l'autre. Plusieurs paramètres ont été évoqués dans l'exercice de cette méthode visant la prédiction médicale. Ces paramètres peuvent être donc optimisés tout en respectant les contraintes matérielles comme la disponibilité des données dans la base MIMIC III. Ensuite, de nouvelles caractéristiques peuvent être extraites à partir des signaux et la sélection de ces caractéristiques peut contribuer à l'amélioration de la prédiction. Dans cette étude, les signaux ont été sélectionnés en se basant sur leur abondance au sein de l'ensemble des patients atteints de la fibrillation atriale. Cependant, les signaux jugés médicalement liés à cette maladie peuvent être utilisés dans un cadre méthodologique adoptant les données partiellement fournies.

REMERCIEMENT

Les auteurs remercient l'Université Libanaise, l'Université de Technologie de Troyes en France et le FEDER pour le financement de cette recherche.

REFERENCES

- [1] Wang, Thomas J., et al. "A risk score for predicting stroke or death in individuals with new-onset atrial fibrillation in the community: the Framingham Heart Study." *Jama* 290.8 (2003): 1049-1056.
- [2] Stroke Prevention in Atrial Fibrillation Investigators. "Stroke prevention in atrial fibrillation study: final results." *Circulation* 84.2 (1991): 527-39.
- [3] Benjamin, Emelia J., et al. "Independent risk factors for atrial fibrillation in a population-based cohort: the Framingham Heart Study." *Jama* 271.11 (1994): 840-844.
- [4] Scholkopf, Bernhard, and Alexander J. Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001.
- [5] Fukunage, K., and Patrenahalli M. Narendra. "A branch and bound algorithm for computing k-nearest neighbors." *IEEE transactions on computers* 7 (1975): 750-753.
- [6] Dempster, Arthur P. "Upper and lower probabilities induced by a multivalued mapping." *Classic Works of the Dempster-Shafer Theory of Belief Functions*. Springer, Berlin, Heidelberg, 2008. 57-72.
- [7] Johnson, Alistair EW, et al. "MIMIC-III, a freely accessible critical care database." *Scientific data* 3 (2016): 160035.
- [8] Moskowitz, Ari, et al. "Management of Atrial Fibrillation with Rapid Ventricular Response in the Intensive Care Unit: A Secondary Analysis of Electronic Health Record Data." *Shock (Augusta, Ga.)* 48.4 (2017): 436-440.
- [9] Han, Deqiang, Chongzhao Han, and Yi Yang. "A modified evidence combination approach based on ambiguity measure." *2008 11th International Conference on Information Fusion*. IEEE, 2008.