

Année 2023/2024

N°

## Thèse

Pour le

### DOCTORAT EN MEDECINE

Diplôme d'État

par

**Yassine LEMRINI**

Né le 18/06/1996 à Casablanca - Maroc

---

#### TITRE

Prédiction de la fibrillation atriale chez les patients ayant présenté un accident vasculaire  
ischémique en utilisant le Machine Learning : une étude à l'échelle nationale.

---

Présentée et soutenue publiquement le 22/01/2024 devant un jury composé de :

Président du Jury : Professeur Laurent FAUCHIER, Cardiologie, Faculté de Médecine -Tours

Membres du Jury :

Professeur Anne BERNARD, Cardiologie, Faculté de Médecine – Tours

Docteur Marco PASI, Neurologie, PH, CHU – Tours

Docteur Alexandre BODIN, Cardiologie, PH, CHU– Tours

**Directeur de thèse : Docteur Arnaud BISSON, Cardiologie, MCU-PH, Faculté de Médecine – Tours**

UNIVERSITE DE TOURS  
**FACULTE DE MEDECINE DE TOURS**

**DOYEN**

**Pr Patrice DIOT**

**VICE-DOYEN**

Pr Henri MARRET

**ASSESEURS**

Pr Denis ANGOULVANT, *Pédagogie*

Pr Mathias BUCHLER, *Relations internationales*

Pr Theodora BEJAN-ANGOULVANT, *Moyens – relations avec l'Université*

Pr Clarisse DIBAO-DINA, *Médecine générale*

Pr François MAILLOT, *Formation Médicale Continue*

Pr Patrick VOUREC'H, *Recherche*

**RESPONSABLE ADMINISTRATIVE**

Mme Carole ACCOLAS

\*\*\*\*\*

**DOYENS HONORAIRES**

Pr Emile ARON (†) – 1962-1966

*Directeur de l'Ecole de Médecine - 1947-1962*

Pr Georges DESBUQUOIS (†) – 1966-1972

Pr André GOUAZE (†) – 1972-1994

Pr Jean-Claude ROLLAND – 1994-2004

Pr Dominique PERROTIN – 2004-2014

**PROFESSEURS EMERITES**

Pr Daniel ALISON

Pr Gilles BODY

Pr Philippe COLOMBAT

Pr Etienne DANQUECHIN-DORVAL

Pr Luc FAVARD

Pr Bernard FOUQUET

Pr Yves GRUEL

Pr Gérard LORETTE

Pr Loïc VAILLANT

**PROFESSEURS HONORAIRES**

P. ANTHONIOZ – P. ARBEILLE – A. AUDURIER – A. AUTRET – C. BARTHELEMY – J.L. BAULIEU – C. BERGER – JC. BESNARD – P. BEUTTER – C. BONNARD – P. BONNET – P. BOUGNOUX – P. BURDIN – L. CASTELLANI – J. CHANDENIER – A. CHANTEPIE – B. CHARBONNIER – P. CHOUTET – T. CONSTANS – C. COUET – L. DE LA LANDE DE CALAN – P. DUMONT – J.P. FAUCHIER – F. FETISSOF – J. FUSCIARDI – P. GAILLARD – G. GINIES – D. GOGA – A. GOUDEAU – J.L. GUILMOT – O. HAILLOT – N. HUTEN – M. JAN – J.P. LAMAGNERE – F. LAMISSE – Y. LANSON – O. LE FLOCH – Y. LEBRANCHU – E. LECA – P. LECOMTE – AM. LEHR-DRYLEWICZ – E. LEMARIE – G. LEROY – M. MARCHAND – C. MAURAGE – C. MERCIER – J. MOLINE – C. MORAINÉ – J.P. MUH – J. MURAT – H. NIVET – D. PERROTIN – L. POURCELOT – R. QUENTIN – P. RAYNAUD – D. RICHARD-LENOBLE – A. ROBIER – J.C. ROLLAND – P. ROSSET – D. ROYERE – A. SAINDELLE – E. SALIBA – J.J. SANTINI – D. SAUVAGE – D. SIRINELLI – J. WEILL

## PROFESSEURS DES UNIVERSITES - PRATICIENS HOSPITALIERS

---

ANDRES Christian.....	Biochimie et biologie moléculaire
ANGOULVANT Denis .....	Cardiologie
APETOH Lionel .....	Immunologie
AUPART Michel.....	Chirurgie thoracique et cardiovasculaire
BABUTY Dominique .....	Cardiologie
BACLE Guillaume.....	Chirurgie orthopédique et traumatologique
BAKHOS David.....	Oto-rhino-laryngologie
BALLON Nicolas.....	Psychiatrie ; addictologie
BARBIER François.....	Médecine intensive et réanimation
BARILLOT Isabelle.....	Cancérologie ; radiothérapie
BARON Christophe .....	Immunologie
BEJAN-ANGOULVANT Théodora .....	Pharmacologie clinique
BERHOUEZ Julien .....	Chirurgie orthopédique et traumatologique
BERNARD Anne .....	Cardiologie
BERNARD Louis .....	Maladies infectieuses et maladies tropicales
BLANCHARD-LAUMONNIER Emmanuelle .....	Biologie cellulaire
BLASCO Hélène.....	Biochimie et biologie moléculaire
BONNET-BRILHAULT Frédérique .....	Physiologie
BOURGUIGNON Thierry .....	Chirurgie thoracique et cardiovasculaire
BRILHAULT Jean.....	Chirurgie orthopédique et traumatologique
BRUNEREAU Laurent .....	Radiologie et imagerie médicale
BRUYERE Franck.....	Urologie
BUCHLER Matthias.....	Néphrologie
CAILLE Agnès .....	Biostat., informatique médical et technologies de communication
CALAIS Gilles.....	Cancérologie, radiothérapie
CAMUS Vincent.....	Psychiatrie d'adultes
CORCIA Philippe.....	Neurologie
COTTIER Jean-Philippe .....	Radiologie et imagerie médicale
DEQUIN Pierre-François.....	Thérapeutique
DESMIDT Thomas .....	Psychiatrie
DESOUBEAUX Guillaume.....	Parasitologie et mycologie
DESTRIEUX Christophe .....	Anatomie
DI GUISTO Caroline .....	Gynécologie obstétrique
DIOT Patrice.....	Pneumologie
DU BOUEXIC de PINIEUX Gonzague .....	Anatomie & cytologie pathologiques
DUCLUZEAU Pierre-Henri.....	Endocrinologie, diabétologie, et nutrition
EHRMANN Stephan .....	Médecine intensive – réanimation
EL HAGE Wissam.....	Psychiatrie adultes
ELKRIEF Laure.....	Hépatologie – gastroentérologie
ESPITALIER Fabien.....	Anesthésiologie et réanimation, médecine d'urgence
FAUCHIER Laurent .....	Cardiologie
FOUGERE Bertrand .....	Gériatrie
FRANCOIS Patrick.....	Neurochirurgie
FROMONT-HANKARD Gaëlle .....	Anatomie & cytologie pathologiques
GATAULT Philippe.....	Néphrologie
GAUDY-GRAFFIN Catherine.....	Bactériologie-virologie, hygiène hospitalière
GOUPILLE Philippe .....	Rhumatologie
GUERIF Fabrice .....	Biologie et médecine du développement et de la reproduction
GUILLOIN Antoine.....	Médecine intensive – réanimation
GUILLOIN-GRAMMATICO Leslie.....	Epidémiologie, économie de la santé et prévention
GUYETANT Serge .....	Anatomie et cytologie pathologiques
GYAN Emmanuel.....	Hématologie, transfusion
HALIMI Jean-Michel.....	Thérapeutique
HANKARD Régis.....	Pédiatrie
HERAULT Olivier .....	Hématologie, transfusion
HERBRETEAU Denis .....	Radiologie et imagerie médicale
HOURIOUX Christophe.....	Biologie cellulaire
IVANES Fabrice .....	Physiologie
LABARTHE François .....	Pédiatrie
LAFFON Marc .....	Anesthésiologie et réanimation chirurgicale, médecine d'urgence
LARDY Hubert.....	Chirurgie infantile
LARIBI Saïd.....	Médecine d'urgence
LARTIGUE Marie-Frédérique .....	Bactériologie-virologie
LAURE Boris.....	Chirurgie maxillo-faciale et stomatologie
LECOMTE Thierry.....	Gastroentérologie, hépatologie

LEGRAS Antoine.....	Chirurgie thoracique
LESCANNE Emmanuel.....	Oto-rhino-laryngologie
LEVESQUE Éric.....	Anesthésiologie et réanimation chirurgicale, médecine d'urgence
LINASSIER Claude .....	Cancérologie, radiothérapie
MACHET Laurent .....	Dermato-vénérologie
MAILLOT François .....	Médecine interne
MARCHAND-ADAM Sylvain .....	Pneumologie
MARRET Henri .....	Gynécologie-obstétrique
MARUANI Annabel.....	Dermatologie-vénérologie
MEREGHETTI Laurent.....	Bactériologie-virologie ; hygiène hospitalière
MITANCHEZ Delphine .....	Pédiatrie
MOREL Baptiste.....	Radiologie pédiatrique
MORINIERE Sylvain.....	Oto-rhino-laryngologie
MOUSSATA Driffa .....	Gastro-entérologie
MULLEMAN Denis.....	Rhumatologie
ODENT Thierry.....	Chirurgie infantile
OUAISSI Mehdi .....	Chirurgie digestive
OULDAMER Lobna.....	Gynécologie-obstétrique
PAINTAUD Gilles .....	Pharmacologie fondamentale, pharmacologie clinique
PATAT Frédéric .....	Biophysique et médecine nucléaire
PERROTIN Franck .....	Gynécologie-obstétrique
PISELLA Pierre-Jean.....	Ophtalmologie
PLANTIER Laurent.....	Physiologie
REMERAND Francis .....	Anesthésiologie et réanimation, médecine d'urgence
ROINGEARD Philippe.....	Biologie cellulaire
RUSCH Emmanuel.....	Epidémiologie, économie de la santé et prévention
SAINT-MARTIN Pauline.....	Médecine légale et droit de la santé
SALAME Ephrem.....	Chirurgie digestive
SAMIMI Mahtab .....	Dermatologie-vénérologie
SANTIAGO-RIBEIRO Maria .....	Biophysique et médecine nucléaire
SAUTENET-BIGOT Bénédicte .....	Thérapeutique
THOMAS-CASTELNAU Pierre .....	Pédiatrie
TOUTAIN Annick.....	Génétique
VELUT Stéphane.....	Anatomie
VOURC'H Patrick.....	Biochimie et biologie moléculaire
WATIER Hervé .....	Immunologie
ZEMMOURA Ilyess .....	Neurochirurgie

## **PROFESSEUR DES UNIVERSITES DE MEDECINE GENERALE**

---

DIBAO-DINA Clarisse  
LEBEAU Jean-Pierre

## **PROFESSEURS ASSOCIES**

---

MALLET Donatien .....

## **PROFESSEUR CERTIFIE DU 2<sup>ND</sup> DEGRE**

---

MC CARTHY Catherine.....Anglais

## **MAITRES DE CONFERENCES DES UNIVERSITES - PRATICIENS HOSPITALIERS**

---

AUDEMARD-VERGER Alexandra.....	Médecine interne
BISSON Arnaud.....	Cardiologie (CHRO)
BRUNAUT Paul.....	Psychiatrie d'adultes, addictologie
CARVAJAL-ALLEGRIA Guillermo.....	Rhumatologie (au 01/10/2021)
CLEMENTY Nicolas.....	Cardiologie
DOMELIER Anne-Sophie.....	Bactériologie-virologie, hygiène hospitalière
DUFOUR Diane.....	Biophysique et médecine nucléaire
FOUQUET-BERGEMER Anne-Marie.....	Anatomie et cytologie pathologiques
GARGOT Thomas.....	Pédopsychiatrie
GOUILLEUX Valérie.....	Immunologie
HOARAU Cyrille.....	Immunologie
KERVARREC Thibault.....	Anatomie et cytologie pathologiques
LE GUELLEC Chantal.....	Pharmacologie fondamentale, pharmacologie clinique
LEDUCQ Sophie.....	Dermatologie
LEFORT Bruno.....	Pédiatrie
LEJEUNE Julien.....	Hématologie, transfusion
LEMAIGNEN Adrien.....	Maladies infectieuses
MACHET Marie-Christine.....	Anatomie et cytologie pathologiques
MOUMNEH Thomas.....	Médecine d'urgence
PARE Arnaud.....	Chirurgie maxillo-faciale et stomatologie
PIVER Éric.....	Biochimie et biologie moléculaire
ROUMY Jérôme.....	Biophysique et médecine nucléaire
STANDLEY-MIQUELESTORENA Elodie.....	Anatomie et cytologie pathologiques
STEFIC Karl.....	Bactériologie
TERNANT David.....	Pharmacologie fondamentale, pharmacologie clinique
VAYNE Caroline.....	Hématologie, transfusion
VUILLAUME-WINTER Marie-Laure.....	Génétique

## **MAITRES DE CONFERENCES DES UNIVERSITES**

---

AGUILLON-HERNANDEZ Nadia.....	Neurosciences
NICOGLU Antonine.....	Philosophie – histoire des sciences et des techniques
PATIENT Romuald.....	Biologie cellulaire
RENOUX-JACQUET Cécile.....	Médecine Générale

## **MAITRES DE CONFERENCES ASSOCIES**

---

AUMARECHAL Alain.....	Médecine Générale
BARBEAU Ludivine.....	Médecine Générale
CHAMANT Christelle.....	Médecine Générale
ETTORI Isabelle.....	Médecine Générale
MOLINA Valérie.....	Médecine Générale
PAUTRAT Maxime.....	Médecine Générale
PHILIPPE Laurence.....	Médecine Générale
RUIZ Christophe.....	Médecine Générale
SAMKO Boris.....	Médecine Générale

## CHERCHEURS INSERM - CNRS - INRAE

---

BECKER Jérôme.....	Chargé de Recherche Inserm – UMR Inserm 1253
BOUAKAZ Ayache .....	Directeur de Recherche Inserm – UMR Inserm 1253
BOUTIN Hervé.....	Directeur de Recherche Inserm – UMR Inserm 1253
BRIARD Benoit.....	Chargé de Recherche Inserm – UMR Inserm 1100
CHALON Sylvie .....	Directrice de Recherche Inserm – UMR Inserm 1253
DE ROCQUIGNY Hugues .....	Chargé de Recherche Inserm – UMR Inserm 1259
ESCOFFRE Jean-Michel.....	Chargé de Recherche Inserm – UMR Inserm 1253
GILLOT Philippe.....	Chargé de Recherche Inrae – UMR Inrae 1282
GOMOT Marie.....	Chargée de Recherche Inserm – UMR Inserm 1253
GOUILLEUX Fabrice .....	Directeur de Recherche CNRS – EA 7501 - ERL CNRS 7001
GUEGUINOU Maxime.....	Chargé de Recherche Inserm – UMR Inserm 1069
HEUZE-VOURCH Nathalie.....	Directrice de Recherche Inserm – UMR Inserm 1100
KORKMAZ Brice.....	Chargé de Recherche Inserm – UMR Inserm 1100
LATINUS Marianne.....	Chargée de Recherche Inserm – UMR Inserm 1253
LAUMONNIER Frédéric .....	Directeur de Recherche Inserm - UMR Inserm 1253
LE MERRER Julie .....	Directrice de Recherche CNRS – UMR Inserm 1253
MAMMANO Fabrizio.....	Directeur de Recherche Inserm – UMR Inserm 1259
MEUNIER Jean-Christophe .....	Chargé de Recherche Inserm – UMR Inserm 1259
PAGET Christophe .....	Directeur de Recherche Inserm – UMR Inserm 1100
RAOUL William .....	Chargé de Recherche Inserm – UMR Inserm 1069
SECHER Thomas.....	Chargé de Recherche Inserm – UMR Inserm 1100
SI TAHAR Mustapha .....	Directeur de Recherche Inserm – UMR Inserm 1100
SUREAU Camille .....	Directrice de Recherche émérite CNRS – UMR Inserm 1259
TANTI Arnaud .....	Chargé de Recherche Inserm - UMR Inserm 1253
WARDAK Claire.....	Chargée de Recherche Inserm – UMR Inserm 1253

## CHARGES D'ENSEIGNEMENT

---

### *Pour l'éthique médicale*

BIRMELE Béatrice.....Praticien Hospitalier

### *Pour la médecine manuelle et l'ostéopathie médicale*

LAMANDE Marc .....

### *Pour l'orthophonie*

BATAILLE Magalie.....Orthophoniste  
CLOUTOUR Nathalie.....Orthophoniste  
CORBINEAU Mathilde.....Orthophoniste  
EL AKIKI Carole .....

### *Pour l'orthoptie*

BOULNOIS Sandrine.....Orthoptiste

# SERMENT D'HIPPOCRATE

En présence des enseignants et enseignantes  
de cette Faculté,  
de mes chers condisciples  
et selon la tradition d'Hippocrate,  
je promets et je jure d'être fidèle aux lois de l'honneur  
et de la probité dans l'exercice de la Médecine.

Je donnerai mes soins gratuits aux indigents,  
et n'exigerai jamais un salaire au-dessus de mon travail.

Admis(e) dans l'intérieur des maisons, mes yeux  
ne verront pas ce qui s'y passe, ma langue taira  
les secrets qui me seront confiés et mon état ne servira pas  
à corrompre les mœurs ni à favoriser le crime.

Respectueux(euse) et reconnaissant(e) envers mes Maîtres,  
je rendrai à leurs enfants  
l'instruction que j'ai reçue de leurs parents.

Que les hommes et les femmes m'accordent leur estime  
si je suis fidèle à mes promesses.

Que je sois couvert(e) d'opprobre  
et méprisé(e) de mes confrères et consœurs  
si j'y manque.

# RESUME

## Introduction.

L'identification d'une population à risque de développer une fibrillation atriale (FA) parmi les patients présentant un accident vasculaire cérébral ischémique (AVCi) est un enjeu majeur de santé afin de sélectionner le profil de patient qui pourrait bénéficier d'une étroite surveillance. Plusieurs scores cliniques ont été développés dans ce but mais leurs performances restent insuffisantes. Récemment, des modèles de Machine Learning (ML) ont été développés dans plusieurs domaines médicaux dont celui de la FA et pourraient apporter une aide quant à la sélection des patients les plus à risques de développer une FA parmi la population d'AVCi.

## Objectif.

Cette étude a pour objectif de développer puis d'évaluer la performance de modèles de ML dans la prédiction de la survenue d'une FA parmi les patients ayant présenté un AVCi puis de les comparer aux performances des scores cliniques existants.

## Matériels et méthodes.

Nous avons utilisé une base de données nationale de 240,459 patients ayant été hospitalisés pour un AVCi sans antécédent de FA dans tous les hôpitaux de France de 2009 à 2012. Cinq modèles de ML ont été développés, puis entraînés au sein d'une cohorte d'entraînement. Nous avons sélectionné le modèle de ML le plus performant en se basant sur l'aire sous la courbe ROC (AUC) puis nous l'avons comparé aux scores CHA2DS2-VASc, Framingham, HAVOC et C2HEST au sein d'une cohorte de validation.

## Résultats.

Après un suivi de  $7,9 \pm 11.5$  mois, 14095 patients ont développé une FA (âge moyen  $77.6 \pm 10.6$ ). Après entraînement, le meilleur modèle de ML sélectionné était celui basé sur un réseau de neurones avec un C index de 0,77 (95% CI, 0.76-0.78).

En comparaison, avec les scores cliniques standards, le modèle de ML était supérieur au score CHA2DS2-VASc (AUC: 0.70), au Framingham risk score (AUC: 0.72), au score HAVOC (AUC: 0.738) et enfin au score C2HEST (AUC: 0.73) ( $P < 0.0001$ ). L'index de reclassification était amélioré par le modèle de réseau de neurones.

## Conclusion.

Le ML permet d'améliorer la prédiction de la FA après un AVC i par rapport aux scores de risque cliniques classiques.

**Mots clés :** Fibrillation atriale, Accident vasculaire cérébral, Machine Learning, Intelligence artificielle.

# ABSTRACT

## Introduction

Efficiently targeting ischemic strokes patients at risk of incident atrial fibrillation (AF) for prolonged cardiac monitoring and oral anticoagulation introduction remains a challenge. Clinical risk scores have been developed to predict post stroke AF with suboptimal performances. Machine learning models are developing in the field of AF prediction and may be used to discriminate post-stroke patients at risk of new onset AF.

## Objective

This study aimed to train, evaluate machine learning models for the prediction of AF and to compare predictive ability to usual clinical scores.

## Method

We used for this purpose a nationwide cohort of 240,459 ischemic stroke patients without AF at baseline seen in all the French hospitals from 2009 to 2012. Five ML models were trained on a train set and the best model was selected to be evaluate on the test set. Discrimination of the best model was evaluated using the C index. We finally compared our best model with the CHA2DS2-VASc, Framingham, HAVOC and C2HEST scores.

## Results

During a mean follow-up of  $7.9 \pm 11.5$  months, 14,095 patients (mean age  $77.6 \pm 10.6$ ; 50.3% female) developed incident AF. After training, the best machine learning model selected was a deep neural network with a C index of 0.77 (95% CI, 0.76-0.78) on the test set. The incidence of AF rose in a stepwise fashion from 19.9 per 1000 patient-years for the first quartile to 175.8 per 1000 for the fourth quartile.

Compared to traditional clinical scores, the selected model was statistically significantly superior to the CHA2DS2-VASc score (C index: 0.70), Framingham risk score (C index: 0.72), HAVOC score C index: 0.738) and C2HEST score (C index: 0.73) ( $P < 0.0001$ ). The ability to predict AF was improved as shown by net reclassification index increase ( $P < 0.0001$ ) and decision curve analysis.

## Conclusion

Machine learning algorithms predict incident AF post stroke with a better ability than previously developed clinical risk scores.

**Keywords:** Atrial fibrillation, Ischemic stroke, Machine Learning, Artificial intelligence.

# Remerciements

---

*Je vous prie d'accepter ce travail comme témoin de ma gratitude,*

*Au président du Jury,*

**Mr le professeur Laurent FAUCHIER**, Merci de juger ce travail, je vous remercie pour votre enseignement en rythmologie et en insuffisance cardiaque. Je vous remercie de m'avoir inculqué la rigueur de travail dont vous faites preuve.

*A mon directeur de thèse,*

**Dr BISSON**, je te remercie de m'avoir permis de travailler avec toi sur ce sujet très intéressant. Je te remercie d'avoir encadré ce travail. Nous nous sommes lancés dans une discipline informatique très lointaine de la médecine que nous avons l'habitude d'exercer. Par ailleurs, ton raisonnement, ta clairvoyance et ton analyse des situations cliniques ont toujours été un exemple que j'essaie de suivre.

*Aux membres du Jury,*

**A Mme le professeur Anne BERNARD**, Merci de juger ce travail je ne sais pas comment cela a pu arriver mais vous avez fait naître en moi une grande curiosité pour les valvulopathies complexes, entité qui était très sombre pour moi au début de mon internat. Votre enseignement, votre rigueur de travail et la bienveillance dont vous faites preuves sont un exemple.

**Au Dr BODIN**, j'ai commencé par être ton externe lors de mes premiers pas en médecine. Déjà à ce moment, tu m'avais montré toute ta bienveillance qui a ensuite rythmée notre relation. Ta rigueur de travail et ta clairvoyance sont un exemple que j'essaie de suivre tous les jours.

**Au Dr PASI**, je vous remercie d'avoir lu et jugé ce travail en tant que neurologue expérimenté.

*A mes maitres,*

**A Mr le professeur Dominique BABUTY**, vous avez été la première personne de la cardiologie tourangelle avec qui j'ai eu un contact. Vous avez été mon mentor. Vous m'avez soutenu depuis la 2<sup>e</sup> année de médecine. Grâce à votre enseignement (à l'ancienne mais indémodable), j'ai pu intégrer l'équipe de cardiologie tourangelle que vous avez participé à construire. Je vous souhaite une belle retraite.

**A Mr le professeur Denis ANGOULVANT**, je vous remercie pour votre enseignement de très grande qualité et le partage de vos vastes connaissances dans beaucoup de domaines. Je vous remercie également pour votre bienveillance sans égale envers les internes.

**A Mr le professeur Fabrice IVANES**, je vous remercie de votre enseignement dans le domaine de la cardiologie interventionnelle et plus largement dans le domaine de la cardiologie et de la physiologie.

**Au Dr CLEMENTY**, je te remercie pour le temps que tu as consacré à nous enseigner à travers ton célèbre staff ECG. Tu as toujours été accessible et je suis très heureux d'avoir été ton élève.

**Au Dr DION**, ta bonne humeur constante, ta rigueur de travail et la facilité avec laquelle tu manies la sonde d'échographie sont sources de motivation. J'espère qu'un jour tu valideras ma mesure de l'oreillette gauche.

**Au Dr PIERRE**, pour votre enseignement et le partage de votre colossale expérience en rythmologie.

**Au Dr GENET**, pour ton enseignement dans le domaine des soins intensifs et ta bienveillance en stage.

**Aux Dr SAINT ETIENNE et au Dr CLERC**, je vous remercie pour votre disponibilité, vos explications et votre bienveillance.

*A ceux qui ont participé à former le médecin que je suis devenu,*

**Au Dr BOUTEAU**, je te remercie pour ta bienveillance et ton enseignement. J'ai beaucoup appris auprès de toi. J'ai trouvé en toi quelqu'un vers qui confier mes craintes dès le début de mon internat.

**Au Dr GOURRAUD**, je te remercie de m'avoir permis d'apprendre auprès de toi, tu as toujours su te montrer disponible.

**Au Dr DARWICHE**, ta vision pragmatique de la cardiologie et tes compétences cliniques, techniques ont été d'un grand secours. Tu réussis toujours à trouver une solution aux problèmes aussi insolubles qu'ils soient.

**Au Dr SEMAAN**, je te remercie pour ton enseignement en hémodynamique, ta bienveillance et ta constante bonne humeur. Tu as toujours su te rendre disponible.

**Au Dr GAMET**, pour ton enseignement aussi bien théorique que technique. Tu m'as guidé sur mes tous premiers gestes techniques. **Aux Dr GORALSKI, Dr RAMOUL, NGUYEN, MALMARE et à toute l'équipe de cardiologie du CHU d'Orléans** pour m'avoir permis d'apprendre à vos côtés et de la confiance que vous m'avez accordée.

**A Mr le Professeur François BARBIER** pour votre enseignement en réanimation médicale et en infectiologie, pour votre bienveillance ainsi que pour la confiance que vous m'avez accordée. **Aux Dr BOULAIN, MULLER, SKARZYNSKI, NAY, MATHONNET, BETAGNOL, KAMEL et LE PAPE** pour m'avoir enseigné les finesses de la réanimation médicale.

**A Vincent AH-FAT**, l'un des premiers à m'avoir montré la cardiologie et la rythmologie dès le début de l'externat. Ça a été un plaisir d'avoir été ton élève.

**A Mathieu N.**, tu es un exemple pour moi. Ta rigueur de travail, tes capacités et ton esprit de décision représentent un modèle pour moi. Je te souhaite un merveilleux avenir en rythmologie et j'espère pouvoir compter un jour parmi tes collègues rythmologues.

**A Nicolas C.S.**, ta vision unique et claire de la cardiologie, ton humanité ont marqué le milieu et la fin de mon internat. Ton humour et tes associations médicamenteuses aussi originales qu'efficaces m'inspirent tous les jours. J'ai trouvé en toi un ami qui a su trouver les bons mots au bon moment.

**A Marie L.**, je te remercie pour tes conseils avisés en échographie mais également dans le reste des domaines, pour ta bienveillance et ta disponibilité

**A Alexis B.**, ton calme et ta résilience ont toujours été un exemple. A nos promenades. Tu as toujours été un soutien indéfectible et je t'en remercie.

**A Romain A.**, je te remercie pour les balades qu'on a pu réaliser, ta patience et surtout le partage de ta grande curiosité intellectuelle.

**A Antonin F.**, je te remercie d'avoir partagé avec moi ton savoir-faire et tes connaissances.

**A Hannah B.**, je te remercie pour tes conseils ainsi que pour tous les fou-rires qu'on a pu partager.

*Aux équipes paramédicales,*

**A Ilona**, je te remercie pour ton soutien sans faille depuis la première année d'internat.

**A l'équipe paramédicale du CHRU de Tours, à Sandra, Guillaume, Antoine, Julien, Christophe, Marion et tous les autres**, merci de votre aide sans laquelle je ne serai pas le médecin que je suis devenu.

**A l'équipe paramédicale du service de cardiologie et de réanimation médicale CHU d'Orleans** pour leur aide et de m'avoir permis de travailler à leurs côtés.

**A l'équipe du secrétariat, Valérie, Roselyne, Vanessa, Marie Pierre** pour votre patience, votre bienveillance et tout le travail que vous réalisez.

**A l'équipe de recherche clinique, Céleste, Djedjiga, Mohammed, Emilie, Mathilde, Katia, Alexandra** pour m'avoir accueilli au sein du laboratoire, pour votre disponibilité sans égal ainsi que votre enseignement et surtout pour votre bonne humeur constante.

*A mes co-internes devenus amis*

**A Ahmad**, peu de mot peuvent décrire la confiance et l'estime que je te porte. Tu as été présent dès le début de nos études de médecine et tu m'as soutenu dans les moments les plus difficiles. Nous avons choisi ensemble de nous aventurer en cardiologie à Tours. Je vous souhaite tout le bonheur du monde avec ta moitié.

**A Ivann**, à nos visites « professorales » alors que nous étions externes, à toutes nos discussions articulées autour d'un diner/McDo. J'ai trouvé en toi un très bel ami.

**A Roxane et Mickael**, mes 2 premiers co-internes avec qui j'ai partagé la joie du premier stage en HC. Je n'aurai pas pu espérer mieux comme co-internes, vous avez toujours été là et votre rigueur exemplaire a été exemple que j'ai essayé de suivre dès le début de mon internat.

**A Claire et Lucile**. Vous êtes deux merveilleuses personnes. Votre humanité n'a d'égale que votre rigueur de travail et ça a été un plaisir pour moi de partager avec vous cet internat.

**A Charlène et Inès**, je vous remercie pour votre soutien tout au long de l'internat. J'ai trouvé en vous de merveilleuses amies avec qui partager les bons comme les pires moments sans crainte.

**A Mathieu J.**, pour tout ce que tu as fait pour nous tes co-interne. Pour ton soutien sans faille, ta bonne humeur.

**A Sihame**, pour m'avoir supporté pendant 1 an en tant que co-interne. Pour nos discussions sérieuses et surtout moins sérieuses....

**A Marine**, pour tes conseils en course à pied. Pour les parcours que tu m'as montré. Evidemment pour ta joie de vivre à toutes épreuves.

**A Augustin**, le meilleur réanimateur. Tes conseils avisés, tes connaissances et ton sens clinique sont sources d'inspiration.

**A Etienne**, mon acolyte aux urgences cardiaques. Les boxs ne sont jamais allé aussi « mal » qu'avec nous.

**A Laura**, la meilleure des gériatres !

**Au Dr PEYRAFORT** dont la soif de connaissance et la culture ne cessent de m'impressionner.

**A Alexandra, Clément, Julien et Alexis**, sans qui le stage de réa med n'aurait pas été le même.

**A Paul Morra**, Je tenais particulièrement à te remercier. Nous avons partagé tous les deux un stage assez difficile. J'ai toujours pu compter sur toi.

**A Constance, Alexandre, Dylan, Lisa, Lauriane, Pierre et Sami**, pour le chemin que nous avons parcouru ensemble et les discussions plus ou moins sérieuses que nous avons pu avoir.

**A Amélie, Kassem, Jean Baptiste, Thibault**, aux côtés desquels j'ai beaucoup appris.

**A Louise, Kélio, Paul et Matthieu**, cela a été très agréable de travailler avec vous. Je vous souhaite d'être épanouis durant le reste de votre internat.

*A mes amis avec qui j'ai partagé mes études,*

**A Kamil**, nous nous sommes connus sur les bancs du collège puis nous avons choisi tous les 2 de nous éloigner de nos habitudes pour atterrir à 2000 km de chez nous. Tu as toujours été là pour moi, dans les meilleurs comme dans les pires moments. Même si l'internat a fait que nous nous sommes éloignés, les liens que l'on a tissés au fil des années sont aussi solides qu'une liaison covalente.

**A Mamoune**, pour les merveilleux moments que nous avons pu partager, nos fou-rire et nos discussions pas toujours très sérieuses.

**A Nacer et Ghita**, pour votre soutien à toute épreuve depuis le début de nos études.

**A Chadia**, pour ce premier message en P1 qui a fait naître notre amitié. J'ai trouvé en toi une amie qui a toujours été à l'écoute.

**A Iyad**, pour le calme dont tu fais preuve. Ta rigueur est un exemple. Ainsi que pour toutes ses séances de cinéma que nous avons partagé et que nous continuerons à partager qui sont toujours les bienvenus.

**A Samia**, pour ta constante bonne humeur, ta résilience et ton soutien sans faille.

**A Audrey**, pour tous les bons moments passés ensemble. Pour nos discussions toujours très enrichissantes.

**A Axel**, pour ta sérénité légendaire et inébranlable. Pour ton humour parfois décalé mais toujours bienvenu.

*A ma famille,*

**A mes parents Mounir et Bouchra,** le dictionnaire ne comporte pas de mot assez fort pour décrire l'attachement et l'amour que j'ai pour vous. Je n'ai peut-être pas toujours manifesté cet amour de manière explicite mais ce n'est pas mon fort. Vous m'avez encouragé dès le début. Nous avons partagé ensemble beaucoup de choses. Vous m'avez permis de devenir la personne que je suis actuellement mais j'aurai aimé plus vous ressembler. J'espère pouvoir vous rendre heureux par ce travail.

**A ma sœur Malak,** je te remercie pour ton soutien depuis plus de 27 ans. Je ne manifeste pas très bien ma reconnaissance envers toi par des mots mais permet moi de le manifester par le biais de ce travail. Et A son époux **Reda,** tu es entré dans notre famille récemment mais tu y as trouvé ta place très rapidement. Ton humour parfois décalé et ta bonne humeur inébranlable seront toujours accueillis avec enthousiasme.

**A mon grand-père maternel Dr Abdelwahab Lemseffer, A mon grand-père paternel Abdelkader Lemrini parti trop tôt, A mes grands-mères Latifa et Naima,** je vous remercie pour tout ce que vous avez pu faire pour vos enfants et petits-enfants. Tout l'amour que vous avez pu nous manifester et notre éducation à laquelle vous avez participé grandement. Je vous serai éternellement reconnaissant.

**A mes oncles et tantes,** je vous dédie cette thèse en guise de gratitude, pour le sens de la famille dont vous faites preuves et votre soutien.

**A mes cousins Kamil et Moulay Ahmed,** avec qui nous avons partagé tant de bons moments dans l'enfance.

**A Yassir et Walid,** vous qui m'avez guidé dans mes premiers pas à Tours, loin de ma terre natale et sans qui je n'aurai sans doute pas pu parcourir tout ce chemin.

**A mes cousins et cousines,** trouvez en ce travail le moyen de manifester ma gratitude.

*“Every great and deep difficulty bears in itself its own solution. It forces us to change our thinking in order to find it.”*

*« Chaque grande difficulté porte en elle sa propre solution.  
Elle nous oblige à changer notre façon de penser afin de la trouver. »*

Niels Bohr

## Table des matières

Table des matières .....	16
Table des tableaux et figures .....	17
Abréviations .....	18
Introduction .....	19
Fibrillation atriale et Accident vasculaire cérébral ischémique .....	19
Intelligence artificielle et Machine Learning .....	21
Objectif de l'étude .....	22
Matériel et méthodes .....	23
Population d'étude .....	23
Sélection des variables d'intérêt et développement du modèle .....	23
Scores cliniques .....	25
Évaluation des modèles et comparaisons.....	26
Analyses statistiques .....	26
Résultats .....	27
Discussion.....	29
Utilisation potentielle en pratique clinique .....	30
Forces et limites .....	31
Conclusion.....	32
Références .....	33
Tableaux .....	40
Figures .....	46
Annexes .....	53

## Table des tableaux et figures

<b>Tableau 1.</b> Caractéristiques initiales des patients à l'inclusion.....	40
<b>Tableau 1bis.</b> Comparaison de la cohorte d'entraînement et de la cohorte de validation.....	41
<b>Tableau 2.</b> Score de Chi2 d'association avec la survenue d'une FA pour chaque variable.....	42
<b>Tableau 3.</b> Liste des hyperparamètres définis après optimisation.....	43
<b>Tableau 4.</b> Rapports de côtes et prévalence de survenue de FA après AVCi en fonction des scores...44	
<b>Tableau 5.</b> Comparaison de la performance des scores cliniques et du DNN.....	45
<b>Figure 1.</b> Représentation des différentes étiologies d'AVC.....	46
<b>Figure 2.</b> Courbe du nombre de transistors par microprocesseur prédite par la loi de Moore comparée avec le nombre réel de transistor sur les microprocesseurs développés en fonction du temps.....	47
<b>Figure 3.</b> Nombre de publication concernant l'IA référencées sur PubMed® en fonction du temps...47	
<b>Figure 4.</b> Représentation de la cross-validation à 5 plis.....	48
<b>Figure 5.</b> Modélisation de l'overfitting.....	48
<b>Figure 6.</b> Modélisation d'un modèle de Random Forrest Classifier.....	49
<b>Figure 7.</b> Flow chart.....	50
<b>Figure 8.</b> Distribution de la population et de l'incidence de survenue de FA en fonction du DNN score.....	51
<b>Figure 9. A,</b> Courbe ROC modélisant la survenue de la fibrillation atriale en fonction des scores utilisés.....	52
<b>Figure 9. B,</b> Courbe "Decision curve analyses" pour l'algorithme de Deep Neural Network et des scores traditionnels.....	52
<b>Tableau supplémentaire 1.</b> Classification TOAST des étiologies d'AVC ischémique.....	53
<b>Tableau supplémentaire 2.</b> Caractéristiques du score C2HEST.....	53
<b>Tableau supplémentaire 3.</b> Caractéristiques du score CHA2DS2VASc.....	53
<b>Tableau supplémentaire 4.</b> Caractéristiques du score HAVOC.....	53
<b>Figure supplémentaire 1.</b> Résumé visuel de la méthode de l'étude. ....	54

## **Abréviations**

AVCi : Accident vasculaire cérébral ischémique

AUC : Area under the curve

BPCO : Bronchopneumopathie chronique obstructive

ECG : Electrocardiogramme

ESC : European Society of Cardiology

ESUS : Embolic strokes of undetermined source

FA : Fibrillation atriale

IA : Intelligence artificielle

ML : Machine Learning

## Introduction

### *Fibrillation atriale et Accident vasculaire cérébral ischémique*

La fibrillation atriale (FA) est l'arythmie cardiaque la plus fréquente. En effet elle concerne environ 1 million de personne en France<sup>1</sup> et est associée à une morbidité importante notamment du fait de son risque embolique. Son diagnostic repose sur un enregistrement électrocardiographique montrant l'absence d'onde P et des intervalles RR irréguliers durant plus de 30 secondes<sup>2</sup>. La FA multiplie par 5 le risque d'accident vasculaire ischémique (AVCi)<sup>3</sup>. L'introduction d'un traitement anticoagulant permet de réduire ce risque de 64%<sup>4</sup>.

Par ailleurs, l'accident vasculaire cérébral ischémique (AVCi) est aussi un enjeu de santé publique. On compte environ 68 000 AVCi par an en France<sup>5</sup>. La mortalité et la morbidité qui en découlent, du fait du handicap moteur et des troubles cognitifs, sont importantes. Mais cette morbi-mortalité est variable selon la cause de l'AVCi. C'est pourquoi il est important de classer les AVCi selon leur étiologie en utilisant la classification TOAST<sup>6</sup> résumant les diagnostics étiologiques retenus à l'issue de la prise en charge initiale (**Tableau supplémentaire 1**).

25% des AVCi environ sont dus à une cardiopathie emboligène dont la principale cause est la FA<sup>7</sup>. Ces AVCi cardio-emboliques sont plus graves avec une mortalité au sein de cette population très importante de 50% à 2 ans<sup>8</sup>, et ont tendance à davantage récidiver (jusqu'à 22% à 2 ans)<sup>9</sup>. Cependant, l'étiologie de l'AVCi demeure indéterminée dans 25-30% des cas : ce sont les AVCi dits cryptogéniques. Or un AVCi sans cause n'est pas sans gravité car les AVCi cryptogéniques récidivent dans 32% des cas à 10 ans et entraînent un taux de dépendance et de décès de 23% à 6 mois<sup>10</sup>. Parmi ces AVCi cryptogéniques, il est nécessaire de repérer les AVCi cardio-emboliques méconnus. C'est dans cette optique qu'en 2014 est apparu un nouveau concept : celui de l'ESUS<sup>11</sup> (ie. Embolic strokes of undetermined source). L'ESUS est une sous-catégorie plus homogène d'AVCi cryptogéniques dont on suspecte fortement une cause embolique et qui est définie comme un infarctus cérébral non lacunaire dont on ne retrouve pas de sténose carotidienne de plus de 50%, ni de cause cardio-embolique manifeste sur les examens initiaux. Les ESUS représentent 50% des AVCi cryptogéniques (**Figure 1**).

Le dépistage de la FA après un AVCi est crucial car le traitement en prévention secondaire d'un AVCi d'origine cardio-embolique repose sur un traitement spécifique par anticoagulant et non par antiagrégant plaquettaire, comme c'est le cas pour les autres types d'AVCi, avec une réduction du risque relatif d'évènement cardio-vasculaires majeurs de 40%

par rapport à l'aspirine<sup>12</sup>. Les dernières recommandations de l'ESC de 2020 proposent donc de réaliser systématiquement un monitoring de 72h de l'électrocardiogramme en continu à l'arrivée du patient en unité de neurovasculaire (recommandation de grade I)<sup>2</sup>. En effet, la surveillance télémétrique a fait ses preuves et se révèle plus efficace qu'un simple holter ECG des 24h<sup>13</sup> faisant actuellement partie de la pratique courante dans les unités de neurovasculaire. Cependant, celle-ci n'est pas suffisante puisque certains patients ont une faible charge de FA avec de rares événements et de courte durée tout en restant néanmoins à risque embolique élevé mettant en défaut la surveillance télémétrique des premières 48-72h. *Sposato et al.* ont montré à travers une méta-analyse que plus la surveillance rythmique était longue plus on diagnostiquait de FA<sup>7</sup>. En effet, cette étude montrait une prévalence de FA de 16% parmi les patients dont le diagnostic n'a pas pu être posé au cours de l'hospitalisation ainsi qu'au cours du suivi ambulatoire classique (holter courte durée)<sup>7</sup>. Les études CRYSTAL-AF<sup>14</sup>, EMBRACE<sup>15</sup> et FIND-AF<sup>16</sup> confirment ceci en montrant que, sur une population d'AVCi cryptogénique implantée d'un holter implantable, on pouvait retrouver plus de 12% de FA après 1 an de suivi et jusqu'à 30% à 3 ans<sup>14</sup>. D'autant plus qu'on estime que 90% des épisodes de FA paroxystique sont asymptomatiques et que les symptômes, s'il sont présents, ne sont pas forcément corrélés à un épisode de FA, et ce avec une valeur prédictive positive à 17%<sup>17, 18</sup>.

Il existe des moyens de surveillance du rythme cardiaque de longue voire de très longue durée (plusieurs années) qui peuvent s'avérer très sensibles quant à la détection de FA lorsque les événements sont rares<sup>14-16</sup> mais se révèlent être contraignant physiquement pour le patient et coûteux en temps médical pour leur interprétation et économiquement pour la collectivité s'ils étaient utilisés de manière systématique en comptant le prix du moniteur et des visites régulières pour le suivi des événements. De ce fait, une sélection de patients les plus à risque d'être atteint de FA s'impose afin de mieux cibler les profils pour qui une stratégie intensive de monitoring cardiaque pourrait être efficace en terme diagnostique<sup>2</sup>. Plusieurs outils statistiques ont été étudiés pour réaliser cette sélection. Ils se basent sur le terrain du patient (âge, sexe, facteurs de risque cardiovasculaire, comorbidités, cardiopathie...). On peut citer par exemple le C2HEST score<sup>19</sup>, initialement développé pour prédire la survenue d'une FA dans la population générale asiatique et testé pour prédire la survenue de FA après un AVCi avec une bonne performance prédictive (AUC = 0.734) sur une population européenne<sup>20</sup>. D'autres scores ont été proposés comme le HAVOC<sup>21</sup> avec un C-index à 0.77, le CHASE-LESS<sup>22</sup> avec un AUC à 0.73. Alors que les précédents scores utilisent des caractéristiques cliniques, certaines équipes se sont intéressées à une approche multiparamétrique intégrant des données paracliniques

comme l'électrocardiogramme, l'échographie cardiaque ou encore la biologie pour rechercher des marqueurs de cardiomyopathie atriale<sup>23</sup> faisant le lit de la FA et des complications emboliques. L'intérêt a notamment été porté aux anomalies intéressant l'onde P sur l'ECG de surface en rythme sinusal<sup>24</sup> ainsi qu'à des paramètres échographiques tels que la dilatation atriale gauche ou l'étude de la déformation longitudinale ('strain') de l'oreillette gauche<sup>25</sup>. Les recommandations européennes actuelles proposent un monitoring par holter implantable ou autre dispositif de longue durée pour une population sélectionnée (personnes âgées, avec facteur de risque cardiovasculaire, signes de cardiomyopathie atriale, score C2HEST élevé ou si les caractéristiques à l'imagerie sont en faveur d'un AVC embolique) (grade IIa)<sup>2</sup> tandis que les recommandations américaines proposent ce type de suivi pour tout AVCi cryptogénique (grade IIa)<sup>26</sup>.

### *Intelligence artificielle et Machine Learning*

Plus récemment, en médecine et plus particulièrement en cardiologie, des modélisations informatiques utilisant le concept de Machine Learning (ML) commencent à se développer. Le ML (ou « apprentissage automatique) est une sous partie de l'intelligence artificielle (IA) qui consiste en l'apprentissage autonome des caractéristiques d'un jeu donnée afin d'effectuer une tâche donnée ou une prédiction<sup>27</sup>. C'est en 1956 qu'est imaginé le concept d'intelligence artificielle après une conférence organisée par 10 ingénieurs et chercheurs américains dont J. McCarthy (Dartmouth) considéré depuis comme le père de l'IA. L'IA est d'abord conçue afin d'imiter le raisonnement d'un humain avec un sens commun<sup>28</sup>. Cependant, l'IA s'est assez vite heurtée aux limites technologiques physiques de l'époque (capacité de mémoire, capacité de calcul des ordinateurs). C'est avec le développement de micro-processeurs informatiques de plus en plus puissants que l'IA a pu se développer au travers de « 3 vagues » d'avancées technologiques ponctuées de deux périodes de stagnation que les experts appellent « hiver technologiques ». La puissance croissante des ordinateurs est basée sur la Loi de Moore<sup>29</sup>. Il s'agit d'un modèle technico économique modélisant la progression des performances des outils informatiques à prix constant. Cette règle stipule que le nombre de transistor d'un micro-processeur double chaque 18-24 mois permettant une croissance exponentielle des performances des ordinateurs qu'ils équipent (**Figure 2**). Dans les années 80, un modèle de systèmes experts commence à émerger en tant que sous partie de l'IA et dont l'objectif est de réaliser une tâche précise prédéfinie par l'utilisateur avec des règles également prédéfinies qui étaient de plus en plus nombreuses jusqu'au point où ce type de système expert était trop coûteux en temps et trop complexe pour être utilisé. C'est avec l'arrivée d'internet puis du Big

Data dans les années 2010 que le ML connaît un grand essor et se démocratise du fait de sa capacité à synthétiser une importante quantité de données. Les applications actuellement les plus populaires du ML sont des problèmes de classification : à partir d'une entrée (input : image, email...), l'algorithme est censé trouver une sortie lui correspondant (output : chat/chien, email important/spam...). Il existe 2 principaux types d'algorithmes : les algorithmes supervisés et non supervisés. Les algorithmes supervisés nécessitent une base de données pour laquelle l'entrée et la sortie sont connues et annotées (labellisées) permettant à l'algorithme de reconnaître la 'bonne réponse' et donc de s'entraîner. Les algorithmes non supervisés ont pour objectif de catégoriser eux même les observations en plusieurs groupes (clusters) ayant des caractéristiques similaires sans que l'utilisateur n'interfère.

Cette capacité des algorithmes à exploiter en profondeur des données de grand volume et de dimensionnalité importante et complexe (grand nombre de variables) rend l'IA attrayante dans certains domaines de la médecine. Ces techniques se sont considérablement démocratisées grâce à l'accessibilité de bibliothèques, de packages de codes et de software open-source tels que Python® ou R®. Dans la **Figure 3**, nous pouvons voir l'augmentation du nombre de publications en lien avec le ML sur la base de données PubMed®.

Une première application publiée dans la revue JAMA a fait grand bruit en 2016 dans le domaine de la détection de la rétinopathie diabétique à partir d'images de fond d'œil à l'aide d'algorithmes<sup>30</sup>. Dans le papier de *Gulshan et al*, les algorithmes développés atteignent des AUC autour de 0,99 sur les cohortes de validation dans la détection de la rétinopathie diabétique<sup>30</sup>. Depuis, en cardiologie et en épidémiologie<sup>31</sup>, plusieurs auteurs ont utilisé l'IA afin de prédire la survenue de FA dans la population générale (et pas seulement après un AVCi) en se basant essentiellement sur des tracés ECG 12 dérivations de repos en rythme sinusal et ce avec de très bonnes performances<sup>32-34</sup>.

### *Objectif de l'étude*

Par conséquent, l'objectif principal de cette étude est de développer un modèle prédictif de FA survenant après un AVCi en utilisant une approche de type ML et de comparer ses performances avec celles des scores cliniques développés précédemment à l'aide d'outils statistiques standards.

## **Matériel et méthodes**

### *Population d'étude*

Cette étude repose sur la base de données nationale française d'hospitalisation : le « Programme de Médicalisation des Systèmes d'Information » (PMSI). A chaque fin d'hospitalisation dans un des 1546 établissements de santé français public ou privé, un codage du diagnostic principal à l'hospitalisation ainsi que des diagnostics associés est effectué manuellement. Ce codage est basé sur la Classification Internationale des Maladies dans sa 10<sup>e</sup> version (CIM-10). Chaque personne possède un identifiant unique permettant de synthétiser tous les séjours hospitaliers dont il a fait l'objet sans rompre l'anonymat. La fiabilité de cette base de données a déjà été éprouvée et a été utilisée dans de nombreuses études de grande envergure<sup>20, 35, 36</sup>.

Les données de tous les patients admis pour AVCi en France de janvier 2008 à Décembre 2012 ont été collectées à partir du PMSI. L'étude incluait les adultes de plus de 18 ans avec un diagnostic d'AVCi (code I63 et divisions) codé comme diagnostic principal, ou diagnostic associé hospitalisés entre le 1<sup>er</sup> janvier 2008 et le 31 Décembre 2012. Nous avons exclu tous les patients dont l'hospitalisation pour AVCi était avant janvier 2009 afin de garder un recul d'au moins un an sur les comorbidités et antécédents avant AVCi. Les patients sans diagnostic de FA posé antérieurement (code I48 et divisions) ont été considérés comme sans antécédent de FA.

Toutes les données contenues dans cette base de données sont stockées de manière confidentielle et sécurisée. Du fait du caractère épidémiologique non interventionnel de l'étude, aucun consentement n'a été recueilli. Il s'agissait d'une étude rétrospective sans conséquence sur la prise en charge des patients. Ce type d'étude a été approuvée par le comité institutionnel du Pole Cœur Thorax Vaisseaux du CHRU de Tours en décembre 2015 et a été enregistré comme un audit clinique. La collection des données et leur analyse a été approuvée par le Conseil National de l'Informatique et des Libertés (CNIL), un comité national d'éthique indépendant dont le but est de protéger les droits de l'Homme en France et de s'assurer de la confidentialité et de l'anonymisation des données (autorisation n° 1749007).

### *Sélection des variables d'intérêt et développement du modèle*

L'âge (seule variable continue) a été normalisé tandis que les variables catégorielles ont été traduites sous la forme d'un code binaire (0 et 1). Puis, la base de données a été séparée

aléatoirement en deux cohortes : une cohorte d'entraînement ou train set (70%) et une cohorte de validation ou test set (30%). Dans un souci de précision, seules les variables les plus importantes de la cohorte d'entraînement ont été retenues pour analyse à l'aide d'un test du Chi<sup>2</sup>. Les 12 variables les plus influentes ont été appliquées à cinq modèles de ML bien connus afin de prédire le risque de survenue d'une FA après un AVCi. Ces modèles sont : La régression logistique (LR), le Random Forest Classifier (RFC), le K-nearest neighbors (KNN), l'eXtreme gradient boosting (XGBoost) et le deep neural network (DNN). Ces modèles ont été écrits et compilés en Python en utilisant des bibliothèques open-source : Scikit-learn version 1.1.1 et XGboost version 1.6.1.

Chaque modèle a été entraîné et optimisé sur la cohorte d'entraînement en utilisant une validation croisée à 5 plis (5-fold cross validation) (**Figure 4**). Il s'agit d'une méthode cherchant à séparer le train set en 2 sous cohortes de manière aléatoire : l'une pour l'entraînement et l'autre pour la validation et ce à plusieurs reprises. Par exemple, pour une cross validation en 5 plis, cette opération est réalisée aléatoirement 5 fois. Pour chaque ensemble (entraînement + validation), l'algorithme sera entraîné puis testé (ce qui fait 5 entraînements). La « cross-validation » permet de se prémunir du risque de surentraînement ou « overfitting ». L'overfitting est un phénomène bien connu en ML qui doit absolument être contrôlé. Un algorithme d'IA essaye d'être le plus performant dans la cohorte d'entraînement en utilisant des détails parfois futiles issues d'une simple fluctuation d'échantillonnage (« bruit de fond »), ce qui aboutit à la construction d'un modèle complexe très performant pour les données sur lesquelles il s'est entraîné (données « vues ») mais peu performant sur les données « non vues » (**Figure 5**). Le risque étant de construire un algorithme très performant sur la cohorte d'entraînement mais très peu performant lorsqu'il est appliqué à la population générale.

Concernant les cinq algorithmes utilisés à savoir le Random Forrest Classifier (RFC), la Logistic Regression (LR), l'eXtreme Gradient Boosting (XGBoost), le KNN et le Deep neural network (DNN), leur nature est différente.

La régression logistique est le modèle le plus simple pour la prédiction de données qualitatives en ML et donne une probabilité pour la variable à expliquer en fonction de coefficients données pour les valeurs à expliquer.

Le RFC synthétise les données des variables explicatives en plusieurs arbres de décisions binaires avec à la clé de chacun la prédiction de la variable à expliquer. La prédiction finale sera la prédiction majoritaire issue de chaque arbre (**Figure 6**).

Le KNN est une forme d'apprentissage supervisée dont l'objectif est de classifier le patient dans le groupe avec lequel il partage le plus de similitudes en termes de variables.

Le XGboost est un algorithme reposant également comme le RFC sur des arbres décisionnels mais il s'agit d'un modèle optimisé.

Le DNN est un modèle plus complexe qui se base sur un réseau de neurones à plusieurs couches inspiré du mode de fonctionnement des réseaux neuronaux biologiques<sup>31</sup>.

Chaque algorithme nécessite le réglage de certains hyperparamètres afin d'optimiser son fonctionnement (nombre de variables à utiliser, nombre d'arbre maximal pour le RFC par exemple...). Les valeurs accordées à chaque hyperparamètre ont été optimisées en utilisant un algorithme de RandomizedSearchCV dont l'objectif est de tester aléatoirement un grand nombre combinaisons différentes d'hyperparamètres.

Un premier temps de cette étude était de comparer les cinq modèles de ML entre eux en utilisant leur C-index respectif puis de garder le plus performant afin de l'évaluer sur le test set de le comparer aux scores cliniques pré existants.

### *Scores cliniques*

Pour chaque patient nous avons calculé le score C<sub>2</sub>HEST, CHA<sub>2</sub>DS<sub>2</sub>VASc, le HAVOC et enfin le score Framingham.

Le score C<sub>2</sub>HEST a récemment été étudié pour la prédiction de la FA après AVCI sur cette même base de donnée<sup>20</sup>. Il est basé sur des caractéristiques cliniques à savoir l'existence d'une coronaropathie ou d'une maladie respiratoire obstructive chronique, d'une hypertension, de l'âge, d'une altération de la fonction cardiaque systolique et de l'existence d'une maladie thyroïdienne (**Tableau supplémentaire 1**).

Le score CHA<sub>2</sub>DS<sub>2</sub>VASc<sup>37</sup>, développé pour prédire le risque embolique lié à la fibrillation atriale combine les critères suivant : insuffisance cardiaque, hypertension artérielle, âge, diabète, antécédent embolique artériel, atteinte athéromateuse, sexe (**Tableau supplémentaire 2**). Ce score a déjà démontré son efficacité dans la prédiction de la FA post-AVCI<sup>20</sup>.

Le score Framingham<sup>38</sup> est utilisé pour prédire la survenue d'un événement cardiovasculaire dans les 10 prochaines années chez un patient sans antécédent cardiovasculaire connu. Il s'agit d'un score qui intègre des données cliniques et biologiques à savoir : l'âge, le

taux sanguin de cholestérol total et d'HDL, le niveau de pression artérielle systolique ou diastolique, l'existence d'un tabagisme et d'un diabète. L'absence de certaines de ces variables ne permet qu'une estimation de ce score via d'autres variables de substitution.

Le score HAVOC<sup>21</sup> dont le but est de prédire la survenue d'une FA dans une population d'AVCi cryptogénique incluant : hypertension, âge, valvulopathie, artériopathie, obésité, insuffisance cardiaque et coronaropathie (**Tableau supplémentaire 3**).

#### *Évaluation des modèles et comparaisons*

Nous avons fait le choix d'utiliser l'aire sous la courbe ROC (Receiver operating characteristic curves) qui paraît être la métrique la plus adaptée aux échantillons déséquilibrés. L'aire sous la courbe ROC de la prédiction de la FA a donc été calculée pour le meilleur algorithme de ML ainsi que pour les trois scores cliniques décrits plus haut.

#### *Analyses statistiques*

Les variables qualitatives sont présentées en utilisant l'effectif en valeur absolue associé au pourcentage entre parenthèses. Les variables quantitatives sont présentées en utilisant la moyenne +/- écart type ou la médiane [intervalle interquartile]. Les variables qualitatives ont été comparées en utilisant un test du Chi<sup>2</sup> et les variables quantitatives ont été comparées en utilisant les tests de Kruskal Wallis et Wilcoxon.

Afin de comparer les AUC, nous avons utilisé un test de DeLong<sup>39</sup>. Nous avons également calculé le Net Reclassification Index (NRI) du meilleur modèle ML par rapport aux scores<sup>40</sup>. Le bénéfice clinique du modèle de ML en comparaison avec les scores cliniques a été estimé en utilisant une « decision curve »<sup>41</sup>. Le risque alpha utilisé est de 0,05.

Les algorithmes de ML ont été codés en utilisant le langage Python version 3.09 avec la bibliothèque scikit-learn version 1.1.2. La manipulation des bases de données était réalisée à l'aide des bibliothèques pandas version 1.4.1 et numpy version 1.22.3. La compilation des algorithmes a été réalisée à l'aide de la plateforme Jupyter 1.0.0. Les analyses statistiques (hors DeLong) ont été réalisées à l'aide de la bibliothèque tableone sur Python. Le logiciel STATA version 16.0 (StataCorp) a été utilisé pour la modélisation des decision curves, le calcul des AUC, et du test de DeLong.

Un résumé de la méthode est disponible dans la **figure supplémentaire 1**.

## Résultats

Un total de 240,459 patients a répondu aux critères de sélection. Après une étape de preprocessing (mise en forme et nettoyage des données), nous avons pu inclure 240,306 patients pour analyse. Parmi les patients sélectionnés, 14,095 (5,9%) ont présenté une FA dans les suites après un suivi moyen de 7,9 +/- 11,5 mois.

Les caractéristiques des patients sont présentées sur le **Tableau 1**. Le groupe comprenant les patients ayant présenté une FA dans les suites de leur AVCi avaient plus d'antécédent d'accident ischémique transitoire, d'hypertension artérielle, de diabète, d'embolie artérielle, d'obésité, de maladie rénale chronique, de maladie hépatique, d'anémie, de bronchopathie chronique obstructive, de cancer, de maladie inflammatoire chronique, de dysthyroïdie, de dyslipidémie, d'insuffisance cardiaque, de coronaropathie et d'artériopathie oblitérante des membres inférieurs, et de valvulopathies. D'autre part, les femmes étaient légèrement plus représentées dans le groupe FA que dans le groupe sans FA ( $p < 0,0001$ ). En conséquence, les patients sujets à de la FA après l'AVCi présentait des scores CHA<sub>2</sub>DS<sub>2</sub>VASc, HAVOC et C2HEST plus élevées que dans le groupe sans FA ( $p < 0,0001$ ). Cependant, le groupe sans FA comprenait significativement plus de fumeur et plus de trouble lié à l'usage de l'alcool ( $p < 0,0001$ ).

La cohorte principale a ensuite été divisée aléatoirement en deux cohortes, une cohorte d'entraînement ou dérivation (train set) comprenant 168,214 patients et une cohorte d'évaluation (test set) comprenant 72,092 patients. Les deux cohortes étaient stratifiées selon le diagnostic de FA ou non permettant le respect de la proportion initiale de FA dans chacun des deux groupes.

Les caractéristiques des train et test sets sont résumées dans le **Tableau 1bis**. Il n'y avait pas de différence significative entre les deux groupes pour les variables étudiées sauf pour le sexe puisqu'il y avait légèrement plus de femme dans la cohorte d'évaluation.

Nous avons ensuite réalisé une sélection des variables d'intérêt (« feature selection ») via un test de Chi<sup>2</sup> où nous avons sélectionné les 12 variables les plus associées à la survenue de FA post AVCi (**Tableau 2**). Les 12 variables qui ont été retenues pour l'analyse finale pour les cinq algorithmes de ML étaient par ordre d'importance : l'âge, l'insuffisance cardiaque, la maladie rénale chronique, la présence d'une valvulopathie, la présence d'un stimulateur/défibrillateur cardiaque implantable, la coronaropathie, l'anémie, l'artériopathie périphérique, la BPCO, l'hypertension artérielle, la dysthyroïdie et la dyslipidémie.

Les algorithmes ont donc été entraînés sur la cohorte d'entraînement et une optimisation des hyperparamètres a été réalisées sur cette même cohorte à l'aide d'une validation croisée permettant d'obtenir la version la plus performante de chacun des modèles testés. Les hyperparamètres retenus pour chacun des algorithmes sont disponibles sur le **Tableau 3**.

Les AUC ont été calculées pour chaque score sur le set de validation : LR 0.761 (95% CI, 0.757- 0.769), KNN 0.741 (95% CI, 0.736 - 0.752), RF 0.763 (95% CI, 0.760 - 0.769), XGBoost 0.758 (95% CI, 0.750 – 0.766), DNN 0.766 (95% CI, 0.763 - 0.771). Le modèle par DNN était le plus performant (P=0.002 vs LR, P<0.0001 vs KNN, P=0.02 vs RF, P=0.002 vs XGBoost) et a été sélectionné pour la comparaison avec les scores cliniques sur la cohorte d'évaluation.

Le modèle DNN a donc été ensuite appliqué sur la cohorte d'évaluation constitué de 72,092 patients et avait un AUC similaire à 0,766 (95% CI 0.759–0.773). La prévalence de survenue de FA augmentait de manière linéaire avec le DNN score et avec une association plus importante qu'avec les scores traditionnels (**Figure 8** et **Tableau 4**).

En comparaison avec les scores cliniques, le DNN était plus performant : CHA<sub>2</sub>DS<sub>2</sub>-VASc score 0.702 (95% CI, 0.694 - 0.709 [P<0.0001]), C<sub>2</sub>HEST scores 0.735 (95% CI, 0.727 - 0.742 [P<0.0001]), HAVOC score 0.738 (95% CI, 0.730 - 0.745 [P<0.0001]) et Framingham risk score 0.720 (95% CI, 0.712 - 0.727 [P<0.0001]) (**Tableau 5** et **Figure 9A**). Le DNN avait également un meilleur NRI (net reclassification improvement) par rapport au score CHA<sub>2</sub>DS<sub>2</sub>-VASc (64% [P<0.0001]), score C2HEST (52 % [P<0.0001]), score HAVOC (38% [P<0.0001]) et score de Framingham (48% [P<0.0001]). Enfin, l'analyse de la 'decision curve' montrait un taux de reclassification supérieur du DNN par rapport aux scores cliniques traditionnels (**Figure 9B**).

## Discussion

A notre connaissance, il s'agit de la première étude testant la performance d'un algorithme de ML dans la prédiction de l'arythmie atriale sur la base de données du PMSI. Nous avons montré que la performance de prédiction de la FA après un AVCi des algorithmes de ML était meilleure que celle des scores cliniques ayant fait l'objet de publications. Parmi les algorithmes de ML, le plus efficace était le DNN.

Les comorbidités retenues comme étant les plus associées au risque de FA post AVCi sont pour la plupart également associées au développement d'un remodelage atrial pouvant faire le lit d'une arythmie atriale<sup>42</sup>. En effet, l'hypertension, l'insuffisance cardiaque, l'âge ainsi que la présence d'une valvulopathie gauche ont été décrits comme associés à la cardiomyopathie atriale<sup>42</sup>. D'autres variables moins attendues ont été retrouvées comme associées au risque de survenue d'une FA après un AVCi comme l'antécédent de stimulateur cardiaque et défibrillateur cardiaque mais ceci pourrait être lié à l'indication de la prothèse, aux conséquences de la stimulation chronique<sup>43</sup> et au fait que la présence de ce type de prothèse intra cardiaque permet de monitorer en continu la survenue d'une FA après un AVCi. L'association entre la maladie pulmonaire obstructive et la survenue de FA a fait l'objet de publications ces dernières années. En effet il a été montré une association importante entre les deux maladies<sup>44</sup>, ce qui pourrait être expliqué par une fibrose atriale liée à l'hypoxémie et la fréquence de l'association au syndrome d'apnée obstructive du sommeil qui lui est clairement identifié comme un facteur de risque de survenue de FA<sup>45</sup>.

La puissance du ML réside dans sa capacité à exploiter une grande quantité de variables explicatives de nature différentes dont l'intrication est parfois complexe à impossible à modéliser par les modèles statistiques traditionnels. L'optimisation automatisée des algorithmes permet de simplifier la construction des modèles. Le ML est de plus en plus utilisé dans la prise en charge de la FA, de la prédiction à la prise en charge en passant par son diagnostic sur des tracés ECG<sup>46</sup>. Notre étude concorde avec les études déjà publiées sur la prédiction de la FA par le biais d'algorithmes de ML utilisant les comorbidités associées comme variable<sup>47</sup>. De plus, les performances des scores traditionnels retrouvées dans notre études sont similaires aux précédentes publications dont ils ont fait l'objet<sup>20, 35</sup> permettant de valider la supériorité de notre algorithme de DNN par rapport à ces scores.

Ici nous n'avons utilisé que des variables binaires « présence ou non de telle comorbidité » (sauf l'âge) puisque notre base de données est construite comme telle à partir de

codes CIM-10. Mais afin de pleinement exploiter les capacités du ML, nous aurions tout intérêt à implémenter des variables plus complexes (images, variables continues...). *Iwakawa et al.* ont montré que certaines caractéristiques électrocardiographiques comme les anomalies d'axe de l'onde P étaient significativement associées à l'origine cardio-embolique des AVCi<sup>24</sup>. L'étude PROACTIA<sup>25</sup> allait aussi dans le même sens en y ajoutant le volume de l'OG ou la charge d'extrasystole atriale comme élément déterminant. Plus récemment, certaines équipes se sont intéressées à la fonction atriale gauche par le biais du *Speckle Tracking* («strain») et ont retrouvé une association importante entre l'altération du strain atrial et notamment de sa fonction réservoir avec la survenue d'une FA même infraclinique dans les suites d'un AVCi<sup>48</sup>.<sup>49</sup>. On pourrait donc supposer qu'implémenter nos bases des données avec des variables provenant de mesures ECG ou échographiques pourrait faire pleinement profiter de la puissance du ML. D'autres travaux ont développé des modèles de ML de prédiction de FA basés sur l'ECG en rythme sinusal avec de très bonnes performances<sup>32-34</sup>. *Zheng et al* ont développé un algorithme de ML comprenant le NIHSS (score de sévérité de l'AVCi) parmi les variables explicatives et qui se trouve être très performant (C index = 0,922 pour le DNN)<sup>50</sup>.

#### *Utilisation potentielle en pratique clinique*

Il est absolument nécessaire de différencier les patients qui vont bénéficier d'un traitement anticoagulant de ceux qui n'en bénéficieront pas parmi les victimes d'un AVCi cryptogénique. Les données actuelles ne sont pas favorables à l'utilisation d'anticoagulants chez tous les patients souffrant d'ESUS sans diagnostic de la FA au préalable (NAVIGATE-ESUS<sup>51</sup> et RE-SPECT ESUS<sup>52</sup>). Il est également établi que, plus la surveillance rythmique dure longtemps, plus la sensibilité pour le diagnostic de FA est élevée. Il existe des moyens de surveillance rythmique de longue durée (holter implantable) mais dont le coût pour la société serait important en cas d'utilisation systématique (augmentation du nombre de consultation, d'hospitalisation, coût de l'appareil). Une stratification des patients les plus à risque de FA est nécessaire afin de leur proposer un monitoring plus long. Un algorithme de ML comme celui décrit ici pourrait représenter un outil permettant de réaliser cette sélection. Des études supplémentaires, avec notamment une validation externe, sont cependant nécessaires. Afin de pouvoir utiliser le modèle développé en pratique clinique courante, il faut qu'il soit facile d'utilisation pour tous les médecins prenant en charge les patients atteints d'AVCi. L'intégration d'un tel outil dans une application mobile ou directement dans les logiciels informatiques de dossier patients pourraient faciliter leur diffusion.

### *Forces et limites*

Il s'agit d'une des premières études en France utilisant une base de données nationale de grande envergure pour la construction de modèle de ML. Le ML permet de synthétiser rapidement et efficacement les données afin de réaliser des prédictions plus précises que les outils de statistique standards. Cependant, malgré le nombre important de données, le manque de granularité des variables qui sont quasiment toutes catégorielles limite les performances du modèle.

D'autre part, les données issues des consultations externes ne sont pas disponibles dans la base données PMSI sous-estimant donc l'incidence de FA après un AVCi et induisant un biais de catégorisation. En effet, il est possible que certains patients étiquetés « non-FA » aient en réalité une FA n'ayant pas fait l'objet d'une hospitalisation ultérieure à l'AVCi. Des variables supplémentaires potentiellement utiles comme l'imagerie cérébrale, l'échographie cardiaque, la sévérité de l'AVCi n'étaient pas disponibles.

## **Conclusion**

A partir d'une cohorte nationale d'AVCi, le développement et l'utilisation d'une approche de type ML a montré une prédiction efficace du risque de FA post-AVCi. Un modèle de DNN a obtenu la meilleure performance de prédiction par rapport à d'autres algorithmes de ML et par rapport aux scores de risque clinique traditionnels. Cette méthode prometteuse pourrait aider à la stratification du risque pour la prise de décision en relation avec une stratégie de dépistage de la FA chez les patients ayant subi un AVCi.

## Références

1. Charlemagne A, Blacher J, Cohen A, et al. Epidemiology of atrial fibrillation in France: Extrapolation of international epidemiological data to France and analysis of French hospitalization data. *Archives of Cardiovascular Diseases*. 2011;104:115–124.
2. Hindricks G, Potpara T, Dagres N, et al. 2020 ESC Guidelines for the diagnosis and management of atrial fibrillation developed in collaboration with the European Association for Cardio-Thoracic Surgery (EACTS). *European Heart Journal*. 2021;42:373–498.
3. Lip GYH, Gue Y, Zhang J, Chao T-F, Calkins H, Potpara T. Stroke prevention in atrial fibrillation. *Trends in Cardiovascular Medicine*. 2022;32:501–510.
4. Kishore A, Vail A, Majid A, et al. Detection of Atrial Fibrillation After Ischemic Stroke or Transient Ischemic Attack: A Systematic Review and Meta-Analysis. *Stroke*. 2014;45:520–526.
5. De Peretti C, Nicolau J, Tuppin P, Schnitzler A, Woimant F. Évolutions de la prise en charge hospitalière des accidents vasculaires cérébraux en court séjour et en soins de suite et de réadaptation entre 2007 et 2009 en France. *La Presse Médicale*. 2012;41:491–503.
6. Adams HP, Bendixen BH, Kappelle LJ, et al. Classification of subtype of acute ischemic stroke. Definitions for use in a multicenter clinical trial. TOAST. Trial of Org 10172 in Acute Stroke Treatment. *Stroke*. 1993;24:35–41.
7. Sposato LA, Cipriano LE, Saposnik G, Vargas ER, Riccio PM, Hachinski V. Diagnosis of atrial fibrillation after stroke and transient ischaemic attack: a systematic review and meta-analysis. *The Lancet Neurology*. 2015;14:377–387.

8. Béjot Y, Bailly H, Durier J, Giroud M. Epidemiology of stroke in Europe and trends for the 21st century. *La Presse Médicale*. 2016;45:e391–e398.
9. Kolominsky-Rabas PL, Weber M, Gefeller O, Neundoerfer B, Heuschmann PU. Epidemiology of Ischemic Stroke Subtypes According to TOAST Criteria: Incidence, Recurrence, and Long-Term Survival in Ischemic Stroke Subtypes: A Population-Based Study. *Stroke*. 2001;32:2735–2740.
10. Li L, Yiin GS, Geraghty OC, et al. Incidence, outcome, risk factors, and long-term prognosis of cryptogenic transient ischaemic attack and ischaemic stroke: a population-based study. *The Lancet Neurology*. 2015;14:903–913.
11. Hart RG, Diener H-C, Coutts SB, et al. Embolic strokes of undetermined source: the case for a new clinical construct. *The Lancet Neurology*. 2014;13:429–438.
12. Eaft (European Atrial Fibrillation Trial) Study Group. Secondary prevention in non-rheumatic atrial fibrillation after transient ischaemic attack or minor stroke. *The Lancet*. 1993;342:1255–1262.
13. Rizos T, Güntner J, Jenetzky E, et al. Continuous Stroke Unit Electrocardiographic Monitoring Versus 24-Hour Holter Electrocardiography for Detection of Paroxysmal Atrial Fibrillation After Stroke. *Stroke*. 2012;43:2689–2694.
14. Sanna T, Diener H-C, Passman RS, et al. Cryptogenic Stroke and Underlying Atrial Fibrillation. *N Engl J Med*. 2014;370:2478–2486.
15. Gladstone DJ, Spring M, Dorian P, et al. Atrial Fibrillation in Patients with Cryptogenic Stroke. *N Engl J Med*. 2014;370:2467–2477.

16. Wachter R, Gröschel K, Gelbrich G, et al. Holter-electrocardiogram-monitoring in patients with acute ischaemic stroke (Find-AF RANDOMISED ): an open-label randomised controlled trial. *The Lancet Neurology*. 2017;16:282–290.
17. Page RL, Wilkinson WE, Clair WK, McCarthy EA, Pritchett EL. Asymptomatic arrhythmias in patients with symptomatic paroxysmal atrial fibrillation and paroxysmal supraventricular tachycardia. *Circulation*. 1994;89:224–227.
18. Strickberger SA, Ip J, Saksena S, Curry K, Bahnson TD, Ziegler PD. Relationship between atrial tachyarrhythmias and symptoms. *Heart Rhythm*. 2005;2:125–131.
19. Li Y-G, Pastori D, Farcomeni A, et al. A Simple Clinical Risk Score (C2HEST) for Predicting Incident Atrial Fibrillation in Asian Subjects. *Chest*. 2019;155:510–518.
20. Li Y, Bisson A, Bodin A, et al. C<sub>2</sub> HEST Score and Prediction of Incident Atrial Fibrillation in Poststroke Patients: A French Nationwide Study. *JAMA*. 2019;8:e012546.
21. Kwong C, Ling AY, Crawford MH, Zhao SX, Shah NH. A Clinical Score for Predicting Atrial Fibrillation in Patients with Cryptogenic Stroke or Transient Ischemic Attack. *Cardiology*. 2017;138:133–140.
22. Hsieh C-Y, Lee C-H, Sung S-F. Development of a novel score to predict newly diagnosed atrial fibrillation after ischemic stroke: The CHASE-LESS score. *Atherosclerosis*. 2020;295:1–7.
23. Kreimer F, Gotzmann M. Left Atrial Cardiomyopathy – A Challenging Diagnosis. *Front Cardiovasc Med*. 2022;9:942385.

24. Iwakawa H, Terata K, Kato R, et al. Utility of P-wave abnormalities for distinguishing embolic stroke from non-embolic stroke. *International Journal of Cardiology*. 2022;368:72–77.
25. Skrebelyte-Strøm L, Rønning OM, Dahl FA, Steine K, Kjekshus H. Prediction of occult atrial fibrillation in patients after cryptogenic stroke and transient ischaemic attack: PROACTIA. *EP Europace*. 2022;24:1881–1888.
26. Kleindorfer DO, Towfighi A, Chaturvedi S, et al. 2021 Guideline for the Prevention of Stroke in Patients With Stroke and Transient Ischemic Attack: A Guideline From the American Heart Association/American Stroke Association. *Stroke*. 2021;52.
27. Wiemken TL, Kelley RR. Machine Learning in Epidemiology and Health Outcomes Research. *Annu Rev Public Health*. 2020;41:21–36.
28. McCarthy, John. PROGRAMS WITH COMMON SENSE. *Proceedings of the Teddington Conference on the Mechanization of Thought Processes*. 1959:75–91.
29. Shalf J. The future of computing beyond Moore’s Law. *Phil Trans R Soc A*. 2020;378:20190061.
30. Gulshan V, Peng L, Coram M, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA*. 2016;316:2402.
31. Bi Q, Goodman KE, Kaminsky J, Lessler J. What is Machine Learning? A Primer for the Epidemiologist. *American Journal of Epidemiology*. 2019:kwz189.

32. Attia ZI, Noseworthy PA, Lopez-Jimenez F, et al. An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. *The Lancet*. 2019;394:861–867.
33. Raghunath S, Pfeifer JM, Ulloa-Cerna AE, et al. Deep Neural Networks Can Predict New-Onset Atrial Fibrillation From the 12-Lead ECG and Help Identify Those at Risk of Atrial Fibrillation–Related Stroke. *Circulation*. 2021;143:1287–1298.
34. Christopoulos G, Graff-Radford J, Lopez CL, et al. Artificial Intelligence–Electrocardiography to Predict Incident Atrial Fibrillation: A Population-Based Study. *Circ: Arrhythmia and Electrophysiology*. 2020;13:e009355.
35. Bisson A, Bodin A, Clementy N, Babuty D, Lip GYH, Fauchier L. Prediction of Incident Atrial Fibrillation According to Gender in Patients With Ischemic Stroke From a Nationwide Cohort. *The American Journal of Cardiology*. 2018;121:437–444.
36. Couris CM, Polazzi S, Olive F, et al. Breast cancer incidence using administrative data: correction with sensitivity and specificity. *Journal of Clinical Epidemiology*. 2009;62:660–666.
37. Lip GYH, Nieuwlaat R, Pisters R, Lane DA, Crijns HJGM. Refining Clinical Risk Stratification for Predicting Stroke and Thromboembolism in Atrial Fibrillation Using a Novel Risk Factor-Based Approach. *Chest*. 2010;137:263–272.
38. Wilson PWF, D’Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of Coronary Heart Disease Using Risk Factor Categories. *Circulation*. 1998;97:1837–1847.

39. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988;44:837–845.
40. Pencina MJ, D’Agostino RB, D’Agostino RB, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med*. 2008;27:157–172; discussion 207-212.
41. Vickers AJ, Cronin AM, Elkin EB, Gonen M. Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers. *BMC Medical Informatics and Decision Making*. 2008;8:53.
42. Goette A, Kalman JM, Aguinaga L, et al. EHRA/HRS/APHRS/SOLAECE expert consensus on atrial cardiomyopathies: definition, characterization, and clinical implication. *Europace*. 2016;18:1455–1490.
43. Sweeney MO, Hellkamp AS, Ellenbogen KA, et al. Adverse Effect of Ventricular Pacing on Heart Failure and Atrial Fibrillation Among Patients With Normal Baseline QRS Duration in a Clinical Trial of Pacemaker Therapy for Sinus Node Dysfunction. *Circulation*. 2003;107:2932–2937.
44. Simons SO, Elliott A, Sastry M, et al. Chronic obstructive pulmonary disease and atrial fibrillation: an interdisciplinary perspective. *European Heart Journal*. 2021;42:532–540.
45. Linz D, McEvoy RD, Cowie MR, et al. Associations of Obstructive Sleep Apnea With Atrial Fibrillation and Continuous Positive Airway Pressure Treatment: A Review. *JAMA Cardiol*. 2018;3:532.

46. Olier I, Ortega-Martorell S, Pieroni M, Lip GYH. How machine learning is impacting research in atrial fibrillation: implications for risk prediction and future management. *Cardiovascular Research*. 2021;117:1700–1717.
47. Sekelj S, Sandler B, Johnston E, et al. Detecting undiagnosed atrial fibrillation in UK primary care: Validation of a machine learning prediction algorithm in a retrospective cohort study. *European Journal of Preventive Cardiology*. 2021;28:598–605.
48. Deferm S, Bertrand PB, Churchill TW, et al. Left Atrial Mechanics Assessed Early during Hospitalization for Cryptogenic Stroke Are Associated with Occult Atrial Fibrillation: A Speckle-Tracking Strain Echocardiography Study. *Journal of the American Society of Echocardiography*. 2021;34:156–165.
49. Kusunose K, Takahashi H, Nishio S, et al. Predictive value of left atrial function for latent paroxysmal atrial fibrillation as the cause of embolic stroke of undetermined source. *Journal of Cardiology*. 2021;78:355–361.
50. Zheng X, Wang F, Zhang J, et al. Using machine learning to predict atrial fibrillation diagnosed after ischemic stroke. *International Journal of Cardiology*. 2022;347:21–27.
51. Hart RG, Sharma M, Mundl H, et al. Rivaroxaban for Stroke Prevention after Embolic Stroke of Undetermined Source. *N Engl J Med*. 2018;378:2191–2201.
52. Diener H-C, Sacco RL, Easton JD, et al. Dabigatran for Prevention of Stroke after Embolic Stroke of Undetermined Source. *N Engl J Med*. 2019;380:1906–1917.

## Tableaux

**Tableau 1.** Caractéristiques initiales des patients à l'inclusion

Variables	FA durant suivi n=14,095	Pas de FA durant suivi n=226,364	p value
Age, années	77.6±10.6	70.8±15.7	<0.0001
Age ≥75 ans, n (%)	9,755(69,2%)	110,782(49%)	<0.0001
Sexe (Féminin), n (%)	7,082(50.3%)	107,266(47.4%)	<0.0001
Diagnostics associés, n (%)			
Hypertension	11,745(83.3%)	141,045(62.3%)	<0.0001
Diabète	4,083(29%)	50,977(22.5%)	<0.0001
Insuffisance cardiaque	6,261(44.4%)	33,162(14.7%)	<0.0001
Artériopathie périphérique	6,907(49%)	70,636(31.2%)	<0.0001
Score CHA2DS2-VASc, n (%)			
Score=2	161(1.1%)	16,992(7.5%)	<0.0001
Score=3	523(3.7%)	33,266(14.7%)	<0.0001
Score=4	1,192(8.5%)	39,300(17.4%)	<0.0001
Score=5	2,596(18.4%)	50,585(22.4%)	<0.0001
Score=6	3,903(27.7%)	48,443(21.4%)	<0.0001
Score=7	3,441(24.4%)	26,634(11.8%)	<0.0001
Score=8	1,881(13.4%)	9,214(4.1%)	<0.0001
Score=9	398(2.8%)	1,777(0.8%)	<0.0001
Comorbidités			
Embolisation systémique	807(5.7%)	6,261(2.8%)	<0.0001
Maladie coronaire	4,969(35.3%)	39,652(17.5%)	<0.0001
Obésité	2,071(14.7%)	22,901(10.1%)	<0.0001
Maladie rénale chronique	5,393(38.3%)	38,618(17.1%)	<0.0001
Hépatopathie	593(4.2%)	6,705(3%)	<0.0001
Anémie	3,980(28.2%)	31,165(13.8%)	<0.0001
Bronchopathie chronique	3,661(26%)	35,320(15.6%)	<0.0001
Cancer dans les 5 ans	3,056(21.7%)	37,400(16.5%)	<0.0001
Maladie inflammatoire chronique	1,470(10.4%)	14,186(6.3%)	<0.0001
Trouble de l'usage de l'alcool	954(6.8%)	17,680(7.8%)	<0.0001
Dysthyroïdie	6,040(42.9%)	68,247(30.2%)	<0.0001
Dyslipidémie	5,793(41.1%)	69,428(30.7%)	<0.0001
Dispositif intracardiaque (Pace-maker ou défibrillateur)	1,643(11.7%)	7,201(3.2%)	<0.0001
Valvulopathie	2,780(19.7%)	15,121(6.7%)	<0.0001
Fumeur	1,415(10%)	28,840(12.7%)	<0.0001

**Tableau 1bis.** Comparaison de la cohorte d'entraînement et de la cohorte de validation.

Variables	Training set n = 168214	Testing set n = 72092	P-Value
Age, moyenne (DS)	71.1 (15.5)	71.2 (15.5)	0.688
Sexe F, n (%)	79768 (47.4)	34515 (47.9)	0,041
Hypertension, n (%)	106788 (63.5)	45999 (63.8)	0,133
Insuffisance cardiaque, n (%)	27564 (16.4)	11839 (16.4)	0,833
Diabète, n (%)	38517 (22.9)	16543 (22.9)	0,796
Artériopathie périphérique, n (%)	54156 (32.2)	23385 (32.4)	0,245
Embolie systémique, n (%)	4921 (2.9)	2146 (3.0)	0,503
Coronaropathie, n (%)	31203 (18.5)	13416 (18.6)	0,733
Obésité, n (%)	17462 (10.4)	7510 (10.4)	0,794
Maladie rénale chronique, n (%)	30738 (18.3)	13255 (18.4)	0,515
Hépatopathie n (%)	5145 (3.1)	2150 (3.0)	0,324
Anémie, n (%)	24594 (14.6)	10531 (14.6)	0,939
Bronchopathie chronique, n (%)	27259 (16.2)	11719 (16.3)	0,762
Cancer, n (%)	28477 (16.9)	11978 (16.6)	0,06
Maladie inflammatoire, n (%)	10963 (6.5)	4693 (6.5)	0,952
Trouble de l'usage de l'alcool, n (%)	12985 (7.7)	5649 (7.8)	0,332
Dysthyroïdie, n (%)	52040 (30.9)	22233 (30.8)	0,641
Dyslipidémie, n (%)	52669 (31.3)	22551 (31.3)	0,889
Valvulopathie, n (%)	12403 (7.4)	5488 (7.6)	0,042
Fumeur, n (%)	21062 (12.5)	9193 (12.8)	0,12
Survenue d'une FA après AVCi moyenne (SD)	0.1 (0.2)	0.1 (0.2)	0.993
Score C2HEST, médiane [Q1,Q3]	2.0 [1.0,4.0]	2.0 [1.0,4.0]	0.342
Score HAVOC, médiane [Q1,Q3]	4.0 [2.0,5.0]	4.0 [2.0,5.0]	0.139
Score CHA2DS2VASc, médiane [Q1,Q3]	5.0 [4.0,6.0]	5.0 [4.0,6.0]	0.100
Framingham, moyenne (SD)	6.3 (3.4)	6.3 (3.4)	0.628

**Tableau 2.** Score de Chi2 d'association avec la survenue d'une FA pour chaque variable.

Variables	Score Chi2
Age	6244,65700
Insuffisance cardiaque	5101,45001
Maladie rénale chronique	2366,81572
Valvulopathie	2164,78306
Stimulateur/Défibrillateur cardiaque	1977,46912
Coronaropathie	1549,61585
Anémie	1367,19574
Artériopathie périphérique	903,55274
Maladie pulmonaire	659,43285
Hypertension	652,05143
Dysthyroïdie	493,50797
Dyslipidémie	334,92595
Maladie inflammatoire	266,99403
Obésité	203,39791
Diabète	191,36031
Accident ischémique transitoire	175,54576
Cancer	145,79802
Hépatopathie	51,74444
Fumeur	50,07817
Trouble liée à l'usage de l'alcool	12,21751
Sexe	4,65623

**Tableau 3.** Liste des hyperparamètres définis après optimisation

Abbreviations: ML, machine learning; LR, logistic regression with L2 regularization; RFC, random forest classifier; SVM, support vector machine; XGBoost, extreme gradient boosting; DNN, deep neural network. Les autres options ont été laissés à leur valeur par défaut.

Modèle	Hyper-paramètres	Valeur
LR	C	0.01
	Class weight	0:1; 1:16
RFC	N estimators	245
	Max depth	6
KNN	N_neighbors	171
	Leaf_size	49
XGBoost	Learning rate	0.8
	Min child weight	10
	Colsample bytree	0.04
DNN	Hidden layers	50-100-50
	Learning rate	adaptive
	Activation	relu
	Alpha	0.0001

**Tableau 4.** Rapports de côtes et prévalence de survenue de FA après AVCi en fonction des scores

Scores	OR (95% CI)	AF prevalence per quartile (%)			
		Q1	Q2	Q3	Q4
C <sub>2</sub> HES <sup>*</sup>	2.32 (2.24- 2.4)	1.7	3.4	6.9	16.4
CHA <sub>2</sub> DS <sub>2</sub> VASc <sup>*</sup>	1.87 (1.82 - 1.92)	2.1	5	7.7	13.3
HAVOC <sup>*</sup>	1.94 (1.89 – 2)	1.9	5.1	6.1	14
Framingham <sup>*</sup>	2.24 (2.16 - 2.32)	1.6	4.2	6.6	17
DNN	2.54 (2.45 - 2.63)	1	2.7	5.9	14.3

\* P<0.0001 vs DNN. AF: Atrial Fibrillation; CI: Confidence interval; DNN: Deep Neural Network OR: Odds ratio;

**Tableau 5.** Comparaison de la performance des scores cliniques et du DNN

	C-Index (95% CI)	Comparaison				
		C <sub>2</sub> HEST	CHA <sub>2</sub> DS <sub>2</sub> -VASc	HAVOC	Framingham	DNN
C <sub>2</sub> HEST	0.735 (0.727 - 0.742)	-	<0.0001	0.1537	<0.0001	<0.0001
CHA <sub>2</sub> DS <sub>2</sub> VASc	0.714 (0.713 - 0.714)	<0.0001	-	<0.0001	<0.0001	<0.0001
HAVOC	0.738 (0.730 - 0.745)	0.1537	<0.0001	-	<0.0001	<0.0001
Framingham	0.720 (0.712 - 0.727)	<0.0001	<0.0001	<0.0001	-	<0.0001
DNN	0.766 (0.759 - 0.773)	<0.0001	<0.0001	<0.0001	<0.0001	-

Figures

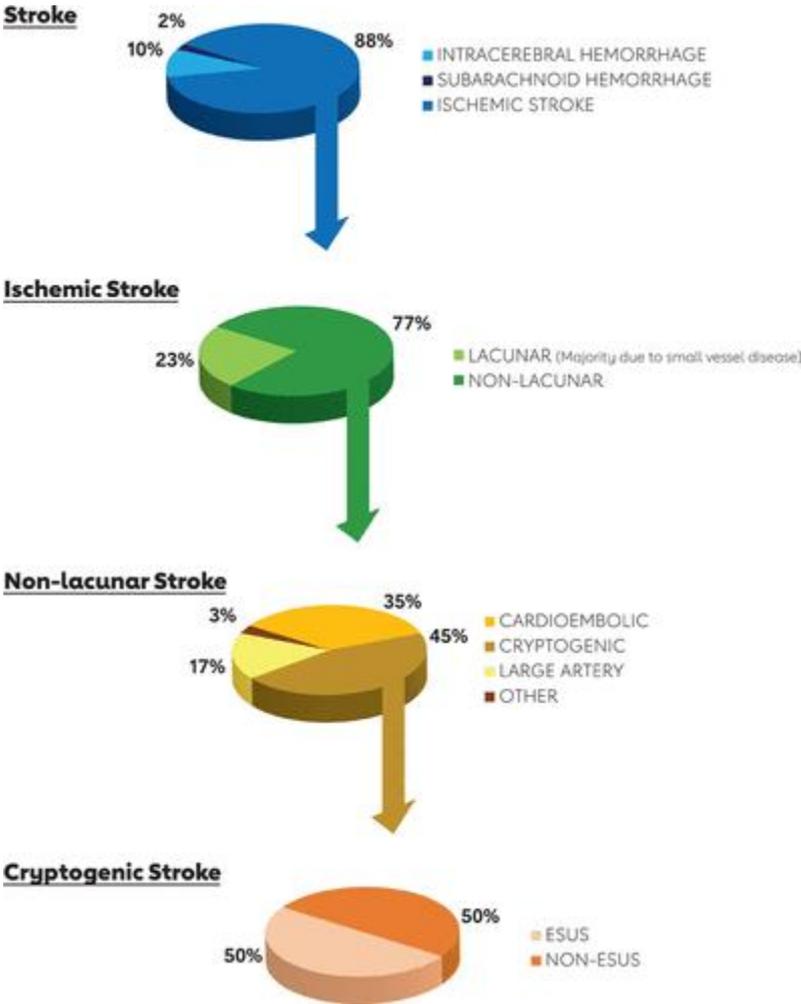
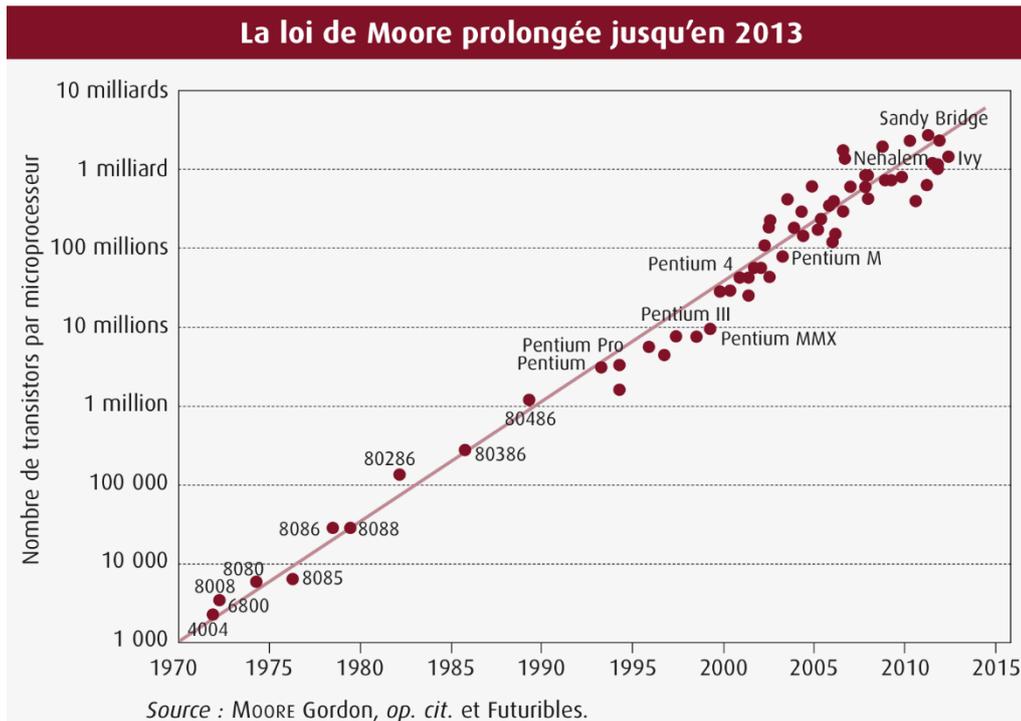
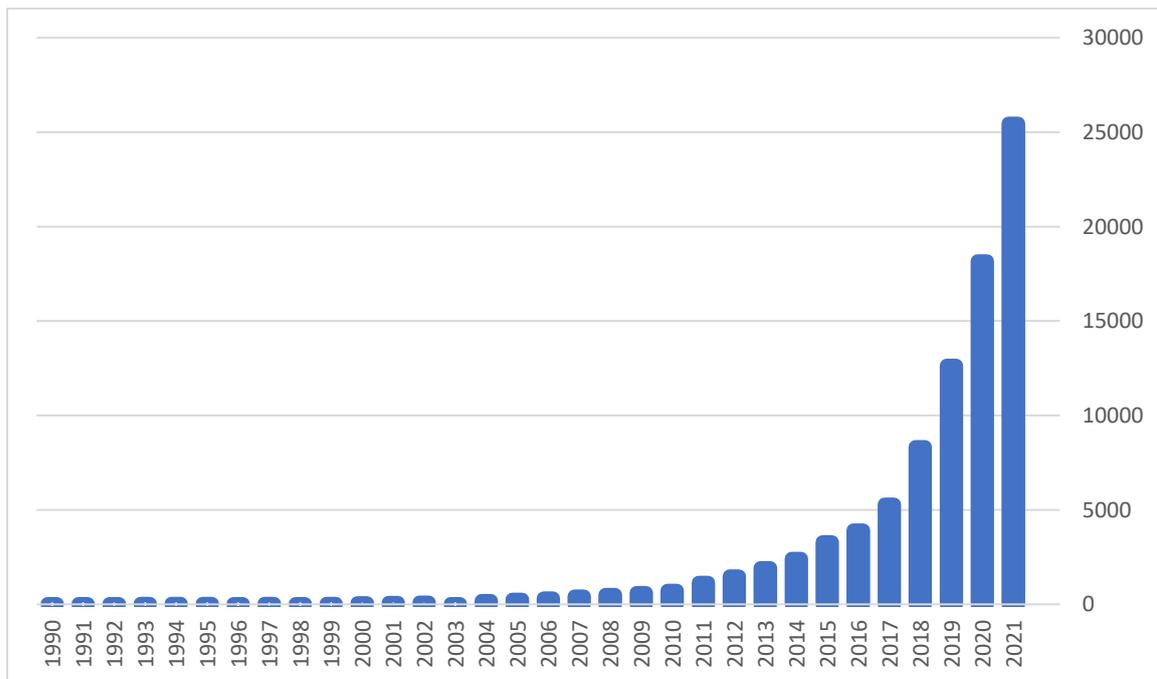


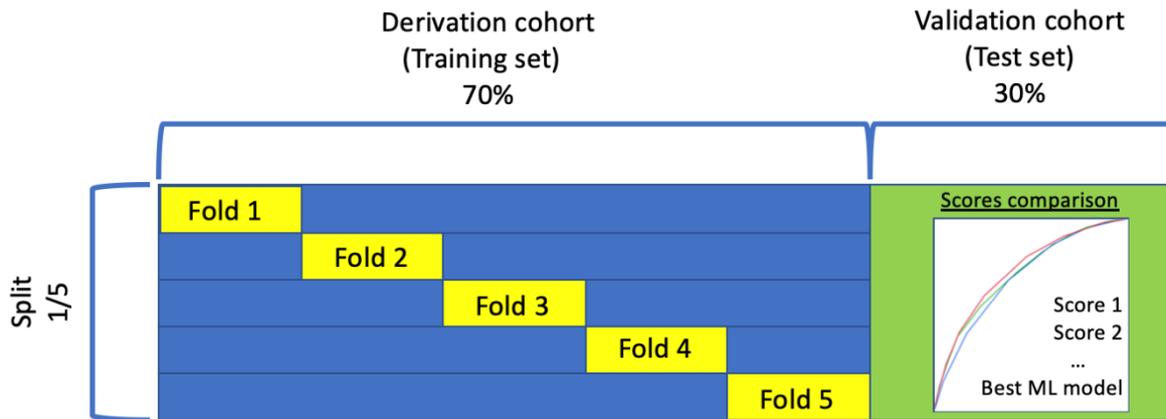
Figure 1. Représentation des différentes étiologies d'AVC.



**Figure 2.** Courbe du nombre de transistors par microprocesseur prédite par la loi de Moore comparée avec le nombre réel de transistor sur les microprocesseurs développés en fonction du temps.



**Figure 3.** Nombre de publication concernant l'IA référencées sur PubMed® en fonction du temps.

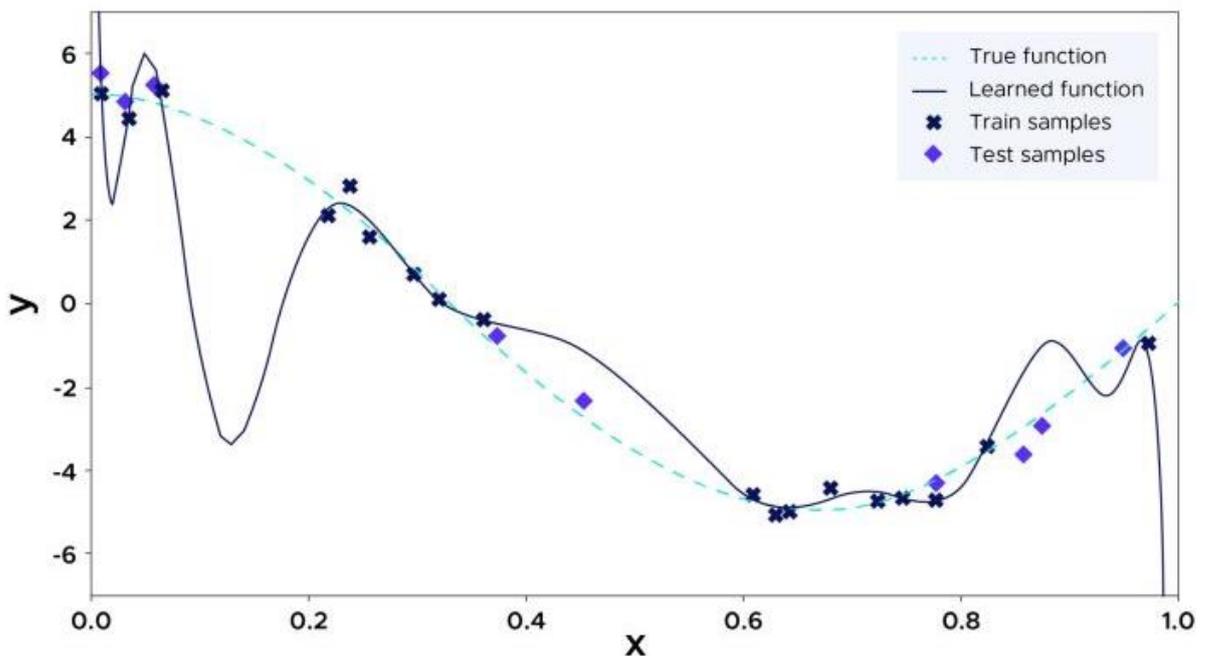


**Figure 4.** Représentation de la cross-validation à 5 pli.

*En bleu : la partie des données qui sert à l'entraînement.*

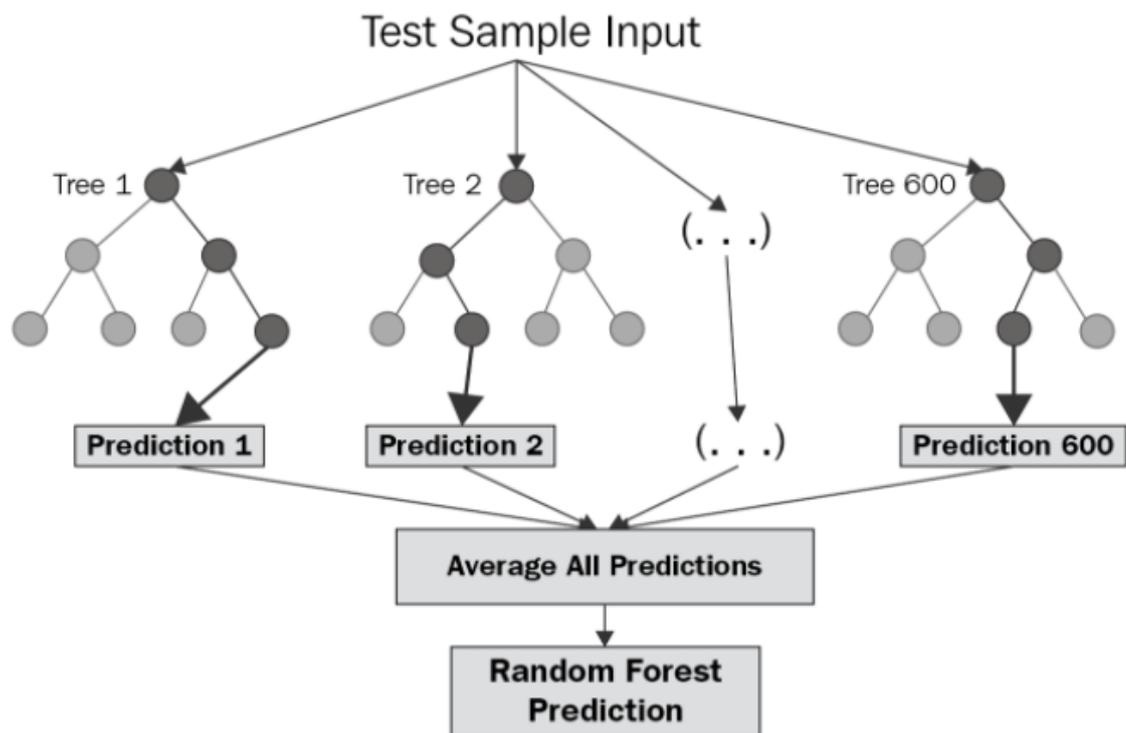
*En jaune : la partie des données qui sert à la sélection des hyperparamètres et la comparaison des modèles de ML entre eux.*

*En vert : la partie des données qui va servir à évaluer le meilleur modèle de ML avec les scores cliniques existants.*

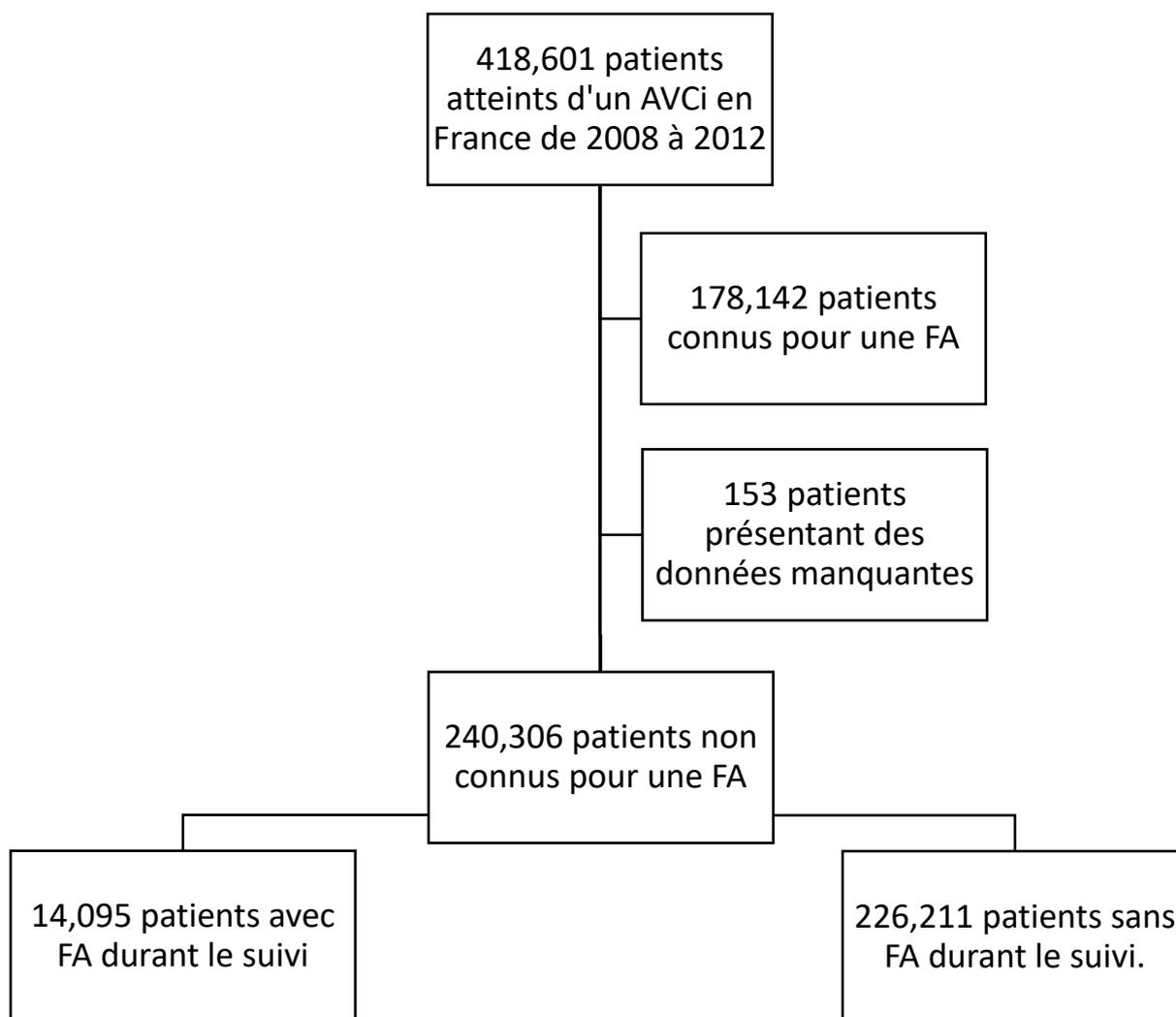


**Figure 5.** Modélisation de l'overfitting.

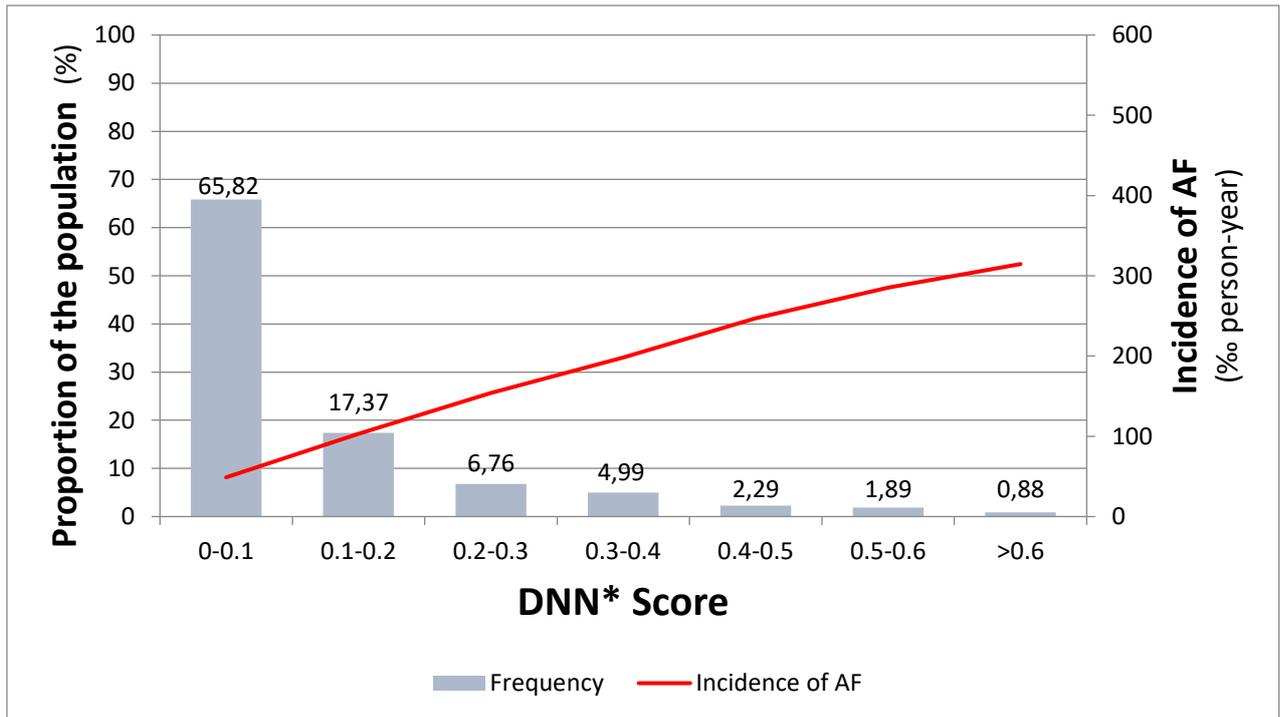
*X représente les caractéristiques explicatives et Y la caractéristique à expliquer. Les croix représentent les données d'entraînement et les points losangiques les données de la cohorte de validation. Le trait plein représente les prédictions de l'algorithme, le trait pointillé représente la « vraie » loi régissant l'association des variables X et Y. Source : <https://datascientest.com>*



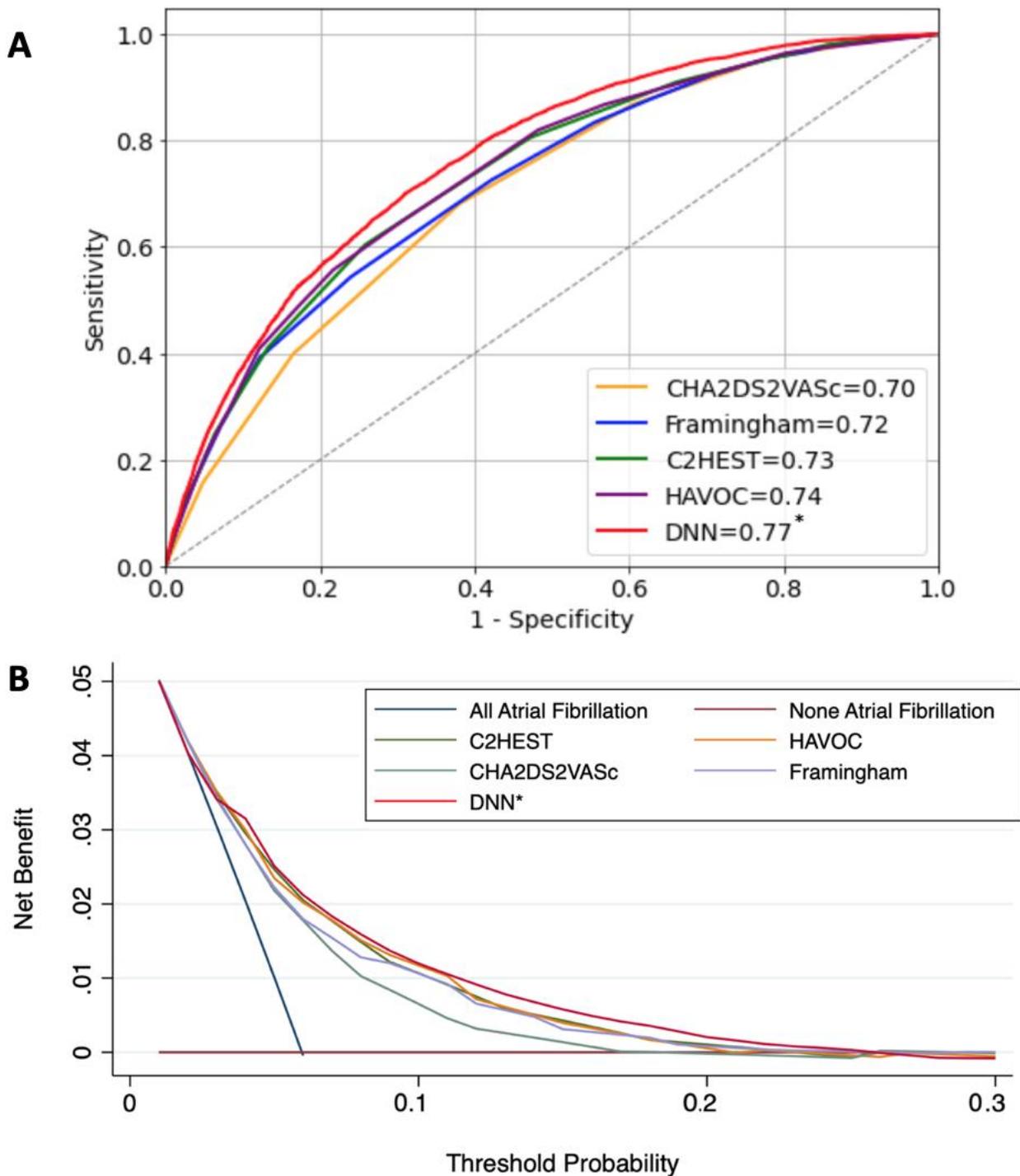
**Figure 6.** Modélisation d'un modèle de Random Forrest Classifier



**Figure 7.** Flow chart



**Figure 8.** Distribution de la population et de l'incidence de survenue de FA en fonction du DNN score. \* DNN: Deep Neural Network



**Figure 9. A,** Courbe ROC modélisant la survenue de la fibrillation atriale en fonction des scores utilisés. \* DNN: Deep Neural Network,  $P < 0.0001$  vs Framingham risk score, CHA<sub>2</sub>DS<sub>2</sub>VASc, HAVOC and C<sub>2</sub>HEST.

**B,** Courbe “Decision curve analyses” pour l’algorithme de Deep Neural Network et des scores traditionnels C<sub>2</sub>HEST, CHA<sub>2</sub>DS<sub>2</sub>-VASc, HAVOC et Framingham. DNN: Deep Neural Network

## Annexes

**Tableau supplémentaire 1.** Classification TOAST des étiologies d'AVC ischémique

I.	Athérosclérose
II.	Cause cardio-embolique
III.	Maladie des petits vaisseaux (AVC lacunaire)
IV.	Autres causes identifiées
V.	Absence de cause identifiée (AVC cryptogénique)

**Tableau supplémentaire 2.** Caractéristiques du score C<sub>2</sub>HEST

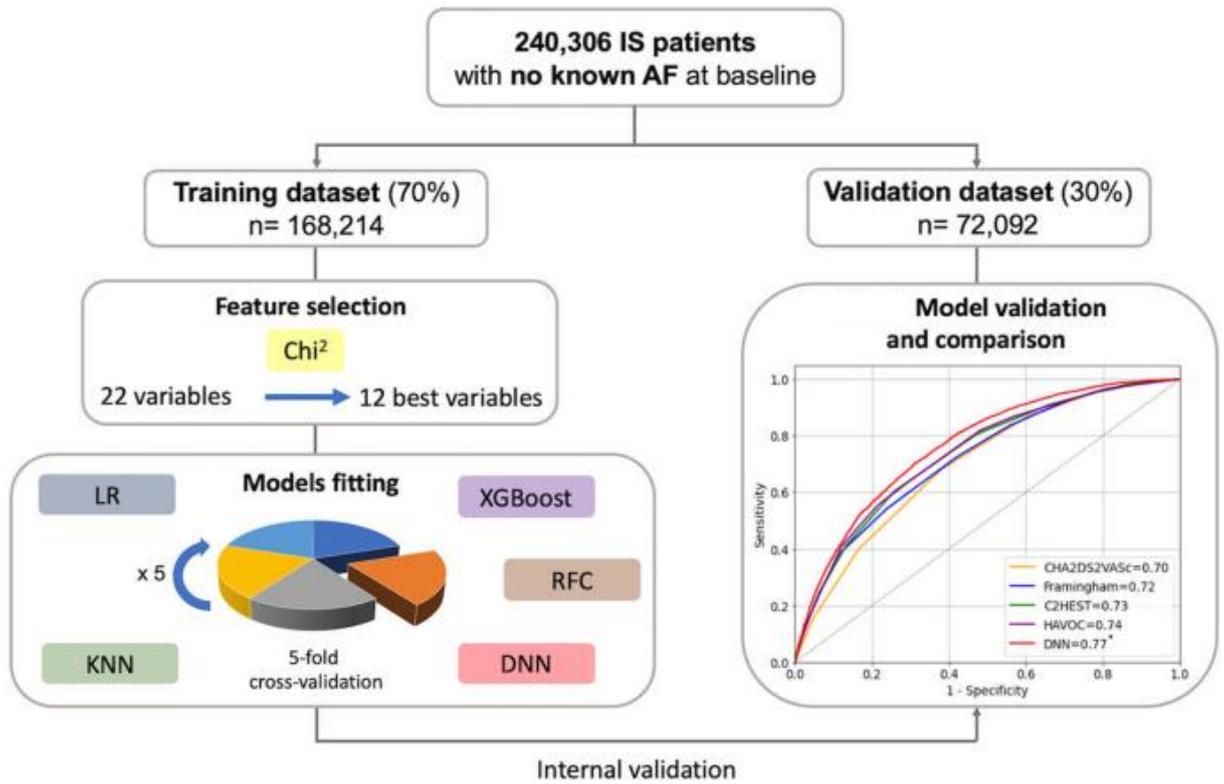
C = Coronaropathie ou bronchopathie chronique obstructive	1 ou 2 points
H = Hypertension	1 point
E = Age $\geq 75$ ans	2 points
S = Insuffisance cardiaque systolique	2 points
T = Thyroïdopathie	1 point

**Tableau supplémentaire 3.** Caractéristiques du score CHA<sub>2</sub>DS<sub>2</sub>VASc

C = Insuffisance cardiaque	1 point
H = Hypertension	1 point
A = Age	1 point si 65-74 ans 2 points si $\geq 75$ ans
D = Diabète	1 point
S = AVC ou embolie artérielle	2 points
VA = atteinte vasculaire athéromateuse	1 point
Sc = Sexe	1 point si femme

**Tableau supplémentaire 4.** Caractéristiques du score HAVOC

H = Hypertension	2 points
A = Age	2 points si $\geq 75$ ans
V = Valvulopathie	2 points
Artériopathie périphérique	1 point
O = Obésité	1 point
C = Insuffisance cardiaque	4 points
Coronaropathie	2 points



**Figure supplémentaire 1.** Résumé visuel de la méthode de l'étude.

**Vu, le Directeur de Thèse**

A handwritten signature in black ink, consisting of a long horizontal stroke with a small loop at the end.

**Vu, le Doyen  
De la Faculté de Médecine de Tours  
Tours, le**

## **LEMIRINI Yassine**

57 pages – 9 tableaux – 10 figures

### **Résumé :**

L'identification d'une population à risque de développer une fibrillation atriale (FA) parmi les patients présentant un accident vasculaire cérébral ischémique (AVCi) est un enjeu majeur de santé afin de sélectionner le profil de patient qui pourraient bénéficier d'une étroite surveillance. Plusieurs scores cliniques ont été développés dans ce but mais leurs performances restent insuffisantes. Récemment, des modèles de Machine Learning (ML) ont été développés dans plusieurs domaines médicaux dont celui de la FA et pourraient apporter une aide quant à la sélection des patients les plus à risques de développer une FA parmi la population d'AVCi. Nous avons utilisé une base de données nationale de 240,459 patients ayant été hospitalisé pour un AVCi sans antécédent de FA dans tous les hôpitaux de France de 2009 à 2012. Cinq modèles de ML ont été développés, puis entraînés au sein d'une cohorte d'entraînement. Nous avons sélectionné le modèle de ML le plus performant en se basant sur l'aire sous la courbe ROC (AUC) puis nous l'avons comparé aux scores CHA2DS2-VASc, Framingham, HAVOC et C2HEST au sein d'une cohorte de validation. Après un suivi de  $7,9 \pm 11.5$  mois, 14095 patients ont développé une FA. Après entraînement, le meilleur modèle de ML sélectionné était celui basé sur un réseau de neurones avec un C index de 0,77. En comparaison, avec les scores cliniques standards, le modèle de ML était supérieur au score CHA2DS2-VASc, au Framingham risk score, au score HAVOC et enfin au score C2HEST. En sommes, le ML permet d'améliorer la prédiction de la FA après un AVCi par rapport aux scores de risque cliniques classiques.

**Mots clés :** Fibrillation atriale, Accident vasculaire cérébral, Machine Learning, Intelligence artificielle.

### **Jury :**

Président du Jury : Professeur Laurent FAUCHIER

Directeur de thèse : Docteur Arnaud BISSON

Membres du Jury : Professeur Anne BERNARD

Docteur Marco PASI

Docteur Alexandre BODIN

Date de soutenance : 22/01/2024