République Algérienne Démocratique et Populaire
Ministère de l'enseignement supérieur et de la recherche scientifique
Université de 8 Mai 1945 – Guelma -
Faculté des Mathématiques, d'Informatique et des Sciences de la matière

Département d'Informatique

Mémoire de Fin d'études Master

Filière : Informatique

Option : Systèmes Informatiques

Thème

L'appariement des données ECG à base des séries chronologiques

Encadré Par :

Dr. Aggoune Aicha

Présenté par :

Hamici Loubna

Juin 2022

Remerciements

Je remercie tous d'abord Allah, tout puissant de m'avoir donné la force, la volonté et le courage pour réaliser ce modeste travail.

Je souhaite adresser mon remerciement la plus sincère à mon encadrante et enseignante Dr. **Aicha AGGOUNE**, votre modestie, vos qualités scientifiques et pédagogiques, votre rigueur et votre dynamisme font de vous un maitre tant apprécié.

Je remercie chaleureusement Docteur Abderrahmane TOUAIMIA médecin urgentiste à l'hôpital Dr. Okbi qui m'a toujours soutenu par son aide et ses conseils.

Mes remerciements les plus vifs s'adressent aussi aux membres de jury pour l'intérêt qu'ils ont porté à ma recherche tout en acceptant d'évaluer ce travail.

Mon plus grand merci s'adresse à tous mes enseignants du département d'informatique de l'université de Guelma.

Je formule également mes remerciements à toute personne qui m'aidé, de près ou de loin, au succès de ce mémoire surtout ma chère amie Amira.

Dédicaces

Je dédie ce modeste travail à ma joie et ma raison de vivre, ma chère mère, qui m'a accompagnée durant tout ce parcours.

A mon père, à qui je dois du respect, qu'il trouve ici l'expression de ma profonde reconnaissance pour tout l'effort et soutien incessant qui m'a toujours apporté.

À mon cher frère qui été toujours présente à mes côtés.

À tous ceux qui, par un mot, m'ont donné la force de continuer…

À tous ceux qui m'aiment et que j'aime…

Loubna

RÉSUMÉ

L'électrocardiogramme (ECG) est l'un des outils de diagnostic les plus utilisés en médecine et en soins de santé, facilitant le diagnostic de très nombreuses maladies cardiaques en association avec les données cliniques, biologiques ou échocardiographies. Le tracé ECG est vu comme une série de valeurs liées au temps qui peut être modélisée par les séries chronologiques.

Les méthodes d'apprentissage profond ont obtenu des résultats prometteurs dans les tâches de la détection intelligente des anomalies cardiaques à partir des données ECG.

L'objectif de ce travail est la proposition d'une approche d'appariement des données ECG modélisées par les séries chronologiques en utilisant les techniques de l'apprentissage profond. À cet effet, nous avons proposé un modèle de réseau de neurones récurrent RNN-LSTM. Ce modèle a été évalué et les résultats sont très satisfaisants.

Nous avons effectué une visualisation profonde des données ECG complexes de 12 dérivations stockées dans InfluxDB en utilisant le Framework Grafana.

Mots clés : Données ECG, Appariement des séries chronologiques, RNN LSTM, Apprentissage profond, Visualisation profonde, Grafana, InfluxDB.

ABSTRACT

The electrocardiogram (ECG) is one of the most commonly used diagnostic tools in medicine and healthcare, facilitating the diagnosis of a large number of cardiac diseases in combination with clinical, biological or echocardiographic data. The ECG trace is seen as a series of time-related values, which can be modeled by time series.
Deep learning methods have achieved promising results on predictive healthcare tasks using ECG signals.
The aim of this work is to propose a time series modeled ECG data matching approach by using deep learning techniques. For this purpose, we have proposed an RNN-LSTM recurrent neural network model. This model has been evaluated and the results are very satisfactory.
We performed deep visualization of complex 12-lead ECG data stored in InfluxDB using the Grafana Framework.

Keywords : ECG data, Time series matching, RNN LSTM, deep learning, deep visualization, Grafana, InfluxDB.

TABLE DES MATIÈRES

TABLE DES FIGURES

INTRODUCTION GÉNÉRALE

Les maladies cardiovasculaires sont la première cause de décès dans le monde, et l'électrocardiogramme (ECG) est un outil très important et majeur de leur diagnostic. Pour cela, le domaine médical nécessite actuellement de nouvelles techniques et technologies, afin d'évaluer les informations de manière objective.

En réalité, l'examen ECG est un outil non invasif effectué par le médecin en vue d'explorer le fonctionnement du cœur par l'emploi des électrodes externes mises en contact de la peau. Il s'agit d'un signal qui reflète l'activité électrique du cœur.

L'interprétation de signaux décrivant la progression de paramètres physiologiques générés par l'ECG peut conduire à des fausses et mauvaises orientations car il requiert une certaine expérience du clinicien qui doit lire intégralement le tracé, de haut en bas puis de gauche à droite. L'analyse automatique de telles données s'avère en effet nécessaire pour mieux aider le médecin à diagnostiquer profondément le patient, prévoir son évolution future et planifier les thérapies à adopter.

Dans le cadre de ce travail, nous avons adopté les séries chronologiques comme modèle de données pour la modélisation et la classification des données ECG qui sont devenues très évolutives en termes de volume et de fréquence d'utilisation par non seulement des médecins spécialistes mais aussi par les médecins généralistes, plus précisément les médecins urgentistes ou encore par les chercheurs des instituts de cardiologies.

INTRODUCTION GÉNÉRALE

Pour cette raison, nous avons demandé d'effectuer un stage au sein de l'hôpital Hakim Okbi-Guelma pour la compréhension de données ECG. Le stage a donc été effectué au niveau du service d'urgences avec le médecin maître de stage Dr Touaimia ABDERAHMANE.

Récemment, L'apprentissage profond (Deep Learning en anglais) qui fait partie de l'apprentissage automatique (Machine Learnig en anglais) est un domaine très prometteur aux applications de détection intelligente des anomalies cardiaques. Plusieurs travaux ont été proposés dans la littérature et qui sont basés sur des techniques d'apprentissage profond. Actuellement, deux modèles les plus populaires et largement utilisés dans la classification des séries chronologiques : CNN et RNN.

Dans notre étude, nous avons utilisé les RNN car ils capturent ces dépendances temporelles dans les données séquentielles plus efficacement que les autres types de réseaux neuronaux.

L'objectif de ce travail est la proposition d'une approche d'appariement des données ECG modélisées par les séries chronologiques pour l'analyse et la détection intelligentes des anomalies cardiaques à partir des données ECG.

Pour atteindre notre objectif, nous devons faire une étude d'une part, de données médicales d'ECG et d'autre part des travaux récents sur la classification de données

INTRODUCTION GÉNÉRALE

ECG par les techniques de deep learning. Assurer l'appariement de données ECG revient à classer des données dans des classes à partir de leurs caractéristiques. En effet, nous proposons un modèle de réseau de neurones profond récurrents RNN LSTM permettant l'analyse de données temporelles. Ce modèle doit être évalué.

Ce mémoire est organisé en 4 chapitres : trois premiers chapitres représentent l'état de l'art et le dernier chapitre décrit notre proposition avec implémentation et évaluation des performances.

Chapitre 01 : nous allons présenter une étude de cas sur les données médicales ECG. L'objectif de ce chapitre est de décrire le lien entre la fonction physiologique du cœur et les caractéristiques du signal ECG.

Chapitre 02 : Ce chapitre représente un état de l'art sur les techniques d'appariement de données ECG. Dans un premier lieu, nous présentons les techniques de l'apprentissage profond avec les différentes architectures de réseau de neurones profonds. Dans le second lieu, le chapitre termine par les travaux récents sur l'appariement de données ECG.

Chapitre 03 : Nous décrivons dans ce chapitre, les notions et les outils des séries chronologiques pour la représentation et l'exploration de données temporelles de l'électrocardiogramme.

Chapitre 04 : dans ce chapitre, nous évaluons la performance de notre approche pour

INTRODUCTION GÉNÉRALE

l'appariement de données ECG à base de séries chronologiques.

CHAPITRE 1

LES DONNÉES MÉDICALES : ETUDE DE CAS DE DONNÉES ECG

1.1 Introduction

Aux vues de l'importance de l'électrocardiogramme (ECG) dans le diagnostic rapide et la surveillance pour la prise en charge des maladies cardiovasculaires. Dans ce contexte Nous nous proposons d'étudier les indications de l'électrocardiogramme. L'objectif de ce chapitre est de décrire le lien entre la fonction physiologique du cœur et les caractéristiques du signal ECG, L'importance de l'électrocardiogramme dans la détection des anomalies cardiaques l'établissement du diagnostic médical. Pour cela, on va commencer par le fonctionnement du cœur humain. L'intérêt, les phases et les caractéristiques d'un signal ECG et les anomalies, en suite l'interprétation d'ECG...

1.2 Fonctionnement du cœur humain

Le cœur est un organe musculeux automatique qui permet une contraction et une relaxation périodiques régulières. Grace à ces deux activités, le cœur fonctionne comme une pompe qui pousse le sang portant d'oxygène dans tout le corps. Chaque jour, le cœur pompe environ 8 000 litres de sang.[6]

Le cœur est constitué de deux ventricules et de deux oreillettes (voir la figure 1.1). Chaque oreillette droite et gauche communique avec un ventricule correspondant.

Dans l'oreillette droite (OD), deux veines caves : la veine cave supérieure (VCS) et la veine cave inférieure (VCI).

Dans l'oreillette gauche (OG), quatre veines pulmonaires, les veines pulmonaires gauches (VPG) et droites (VPD) qui amènent le sang au cœur, le sang est drainé du ventricule gauche (LV) via l'aorte (Ao) et du ventricule droit (RV) via l'artère pulmonaire (AP). [5]

Le système responsable à l'excitation et la conduction électrique comprend : le nœud sinusal, les voies spécialisées internodales, le nœud auriculo-ventriculaire (NAV), le faisceau de His, les branches droite et gauche et les fibres de Purkinje.

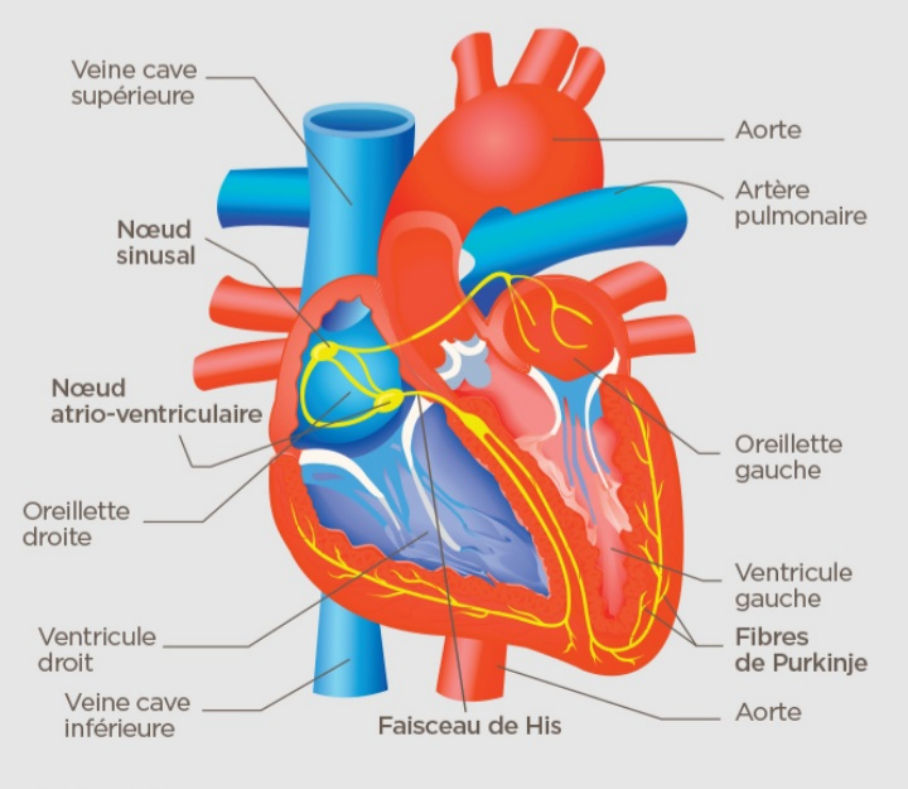


FIGURE 1.1 – Fonctionnement du coeur humain

Le nœud sinusal (NS) : est une minuscule région anatomique, en croissant, située au sommet de l'oreillette droite, près de l'abouchement de la veine cave supérieure.Il se dépolarise généralement plus vite que les autres cellules automatiques, aussi c'est le nœud sinusal qui donne généralement « le rythme sinusal ».

Le nœud auriculo-ventriculaire (AV) : est localisé près du septum interauriculaire, au-dessus de l'abouchement du sinus coronaire. Il reçoit et ralentit l'influx d'origine auriculaire pour permettre une synchronisation plus efficace sur le plan hémodynamique entre les oreillettes et les ventricules.Ainsi, les contractions auriculaires et ventriculaires sont décalées d'environ 120 à 200 msec.

Faisceau de His-Purkinje : ce réseau débute par le tronc du faisceau de His situé dans le septum membraneux interventriculaire, puis se divise rapidement en deux branches, G et D au niveau du septum musculaire. La branche droite reste homogène jusqu'à sa ramification à l'apex du ventricule droit, tandis que la branche gauche se

divise en un faisceau antérieur et un faisceau postérieur gauche avant de se ramifier.

Les fibres de Purkinje : Les branches du faisceau de His finissent dans un réseau de fibres qui arrivent dans les parois ventriculaires. Ils terminent en anastomoses avec les fibres myocardiques musculaires, facilitant leur excitation.

Réseau de Purkinje : spécialisé dans la conduction intraventriculaire (tissus nodal). Il associe le faisceau de His et les fibres de Purkinje qui le prolongent et se distribuent par arborescence dans le myocarde ventriculaire.

1.3 Electrocardiogramme

1.3.1 Définition

L'électrocardiogramme (ECG) est une représentation graphique de l'activité ´électrique cardiaque. Il mesure l'évolution d'une différence de potentiel (en millivolts). Il est enregistré à partir de nombreuses électrodes(capteurs) fixées sur la peau à l'aide de l'appareil appelé électrocardiographe. Cet enregistrement permet aux cardiologues de mesurer le rythme cardiaque et de détecter des anomalies cardiaques.[2]
L'intérêt : L'ECG est un examen complémentaire systématiquement réalisé en consultation de cardiologie et largement pratiqué en dehors de cette spécialité du fait :

De sa simplicité de réalisation.

De son innocuité

De son faible cout

Et surtout de sa grande rentabilité en matière de diagnostic et de surveillance.

1.3.2 Les phases de l'ECG

L'ECG est composé d'une phase d'activité appelée la systole qui alterne avec une phase de repos appelée diastole. Le rythme de ces deux phases est d'environ 70 fois par minute (70 bpm).[2]

La systole (dépolarisation) : C'est la phase du cycle cardiaque pendant laquelle les fibres du myocarde phase incomplète. La contraction se traduit par une diminution du volume de l'oreillette ou du ventricule et entraîne le phénomène d'éjection du sang qu'il contient.

La diastole (repolarisation) : C'est la phase de relâchement, une pause des oreillettes et des ventricules pendant laquelle les ventricules ou les oreillettes se remplissent de sang.

Généralement, la phase diastole est d'une durée supérieure à la phase systole pour un fonctionnement cardiaque normal.

Dans un enregistrement ECG, les impulsions électriques du nœud sino-auriculaire provoquent la première déviation (onde P) (voir figure 1.2 (a) (b)). Le passage de l'influx nerveux de l'oreillette au ventricule est défini par une ligne horizontale correspondant à l'espace PR puis activé électriquement le nœud AV, montrant une déviation rapide et adéquate (complexe QRS) (voir figure 1.2 (c) (d)).

Le pic R suivi d'un segment ST horizontal marquant la fin de la systole (voir figure 1.2 (e)).

Enfin, l'onde T correspondant à la récupération des propriétés électriques initiales, marquant la fin de la phase systolique de l'ECG. Cette série de déviations est suivie d'une période de repos cardiaque, qui est la période diastolique.[2]

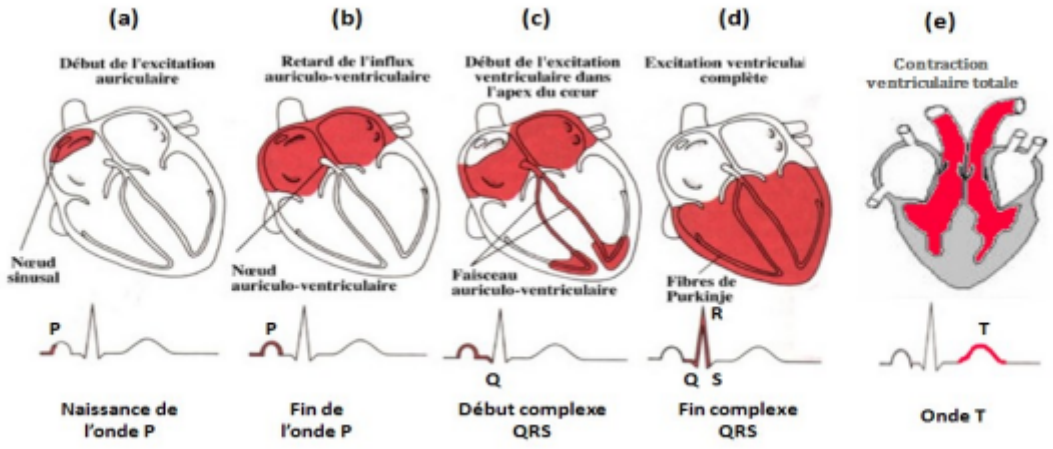


FIGURE 1.2 – les phases d'un éléctrocardiogramme [2]

a : la naissance de l'onde P commence avec le début de l'impulsion produite au nœud Sinusal.

b : La fin de l'onde P correspond au remplissage total des oreillettes.

c : L'onde Q correspond au début du passage du sang des oreillettes vers les ventricules.

d : Le complexe QRS traduit le remplissage total des ventricules avec un pic R qui correspond 'a l'ouverture totale des valves pour laisser passer le sang.

e : L'onde T correspond à la contraction du myocarde pour impulser le sang stocké dans les ventricules vers le corps.

1.4 L'électrocardiographe

C'est un appareil qui vise à mesurer et à enregistrer l'activité électrique du cœur d'un patient. Il est relié aux électrodes (des capteurs d'acquisition de données). Il capte le champ électrique créé par l'activité cardiaque entre 2 électrodes.[40]

L'ECG standard est enregistré sur 12 dérivations (6 précordiales, 6 dérivations des membres), s'effectue habituellement à une vitesse de déroulement du papier de 25mm/s et la détection d'une tension de 1mV provoque une déflexion verticale de 1cm.

1.4.1 Les électrodes :

Doivent être placées à un endroit bien défini sur le corps et directement sur la peau. Quatre électrodes sont placées sur les poignets et les chevilles, et six autres sont placées sur des points définis à la surface du thorax. Les enregistrements ECG montrent une tension positive lorsque l'onde de dépolarisation se déplace vers l'électrode et une tension négative lorsqu'elle s'éloigne de l'électrode. [2]

1.4.2 Les dérivations :

Une dérivation est un circuit électrique déterminé par un couple d'électrodes. Pour obtenir une topographie complète du cœur, il faut un enregistrement de 12 dérivations de l'ECG et qui sont réparties en deux catégories [15] :

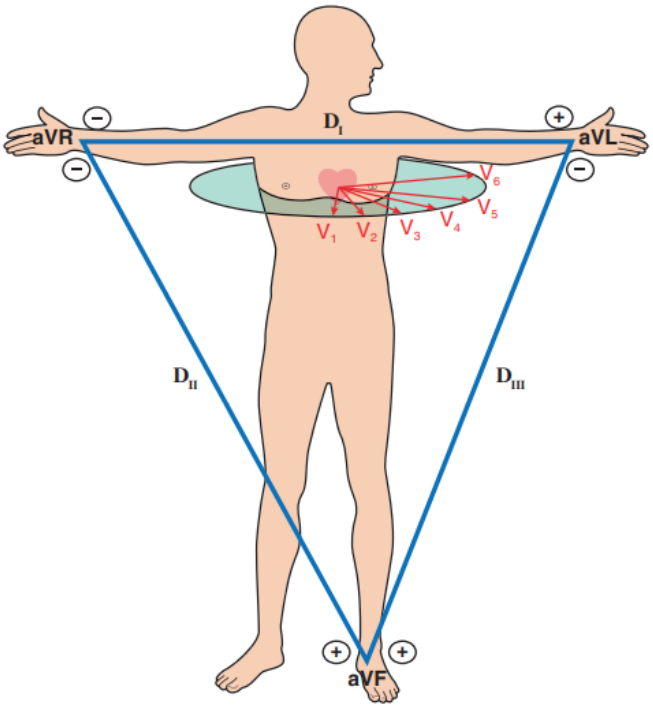


FIGURE 1.3 – Position de 12 dérivations d'un ECG

Dérivations périphériques : (appelées D) qui sont placées sur les quatre membres et qui explorent le plan frontal du cœur.

4 électrodes des membres permettent de déterminer les dérivations des membres.

Leur position respecte une nomenclature « couleur » précise :

R pour right (couleur rouge) : membre supérieur droit.

L pour left (couleur jaune) : membre supérieur gauche.

F pour foot (couleur verte) : membre inférieur gauche.

« Terre », électrode neutre (couleur noire) : membre inférieur gauche.

De ces 4 électrodes résultent 6 dérivations des membres, déterminées par Einthoven (DI, DII, DIII) et Goldberger (aVR, aVL, aVF) :

•DI : vecteur (L)-(R), positif dans le sens R vers L ; angle par rapport à l'horizontale=0°

•DII : vecteur (F)-(R), positif dans le sens R vers F ; angle=+60°

•DIII : vecteur (L)-(R), positif dans le sens R vers L ; angle=+120°

•aVR : vecteur (R)-(L+F), positif dans le sens L+F vers R ; angle=-150°

•aVL : vecteur (L)-(R+F), positif dans le sens R+F vers L ; angle=-30°

•aVF : vecteur (F)-(R+ L), positif dans le sens R+L vers F ; angle=+90°

Dérivations précordiales : (appelées V) qui sont placées sur le thorax du patient pour explorer le plan transversal du cœur

- V1 : 4ème espace intercostal, bord droit du sternum (ligne parasternale).
- V2 : 4ème espace intercostal, bord gauche du sternum (ligne parasternale).
- V3 : à mi-distance entre V2 et V4.
- V4 : 5ème espace intercostal, ligne médio-claviculaire gauche.
- V5 : à mi-distance entre V4 et V6, sur la ligne axillaire antérieure.
- V6 : même niveau horizontal que V4 et V5, ligne axillaire moyenne.

1.5 Les caractéristiques de l'ECG

Un ECG normal se compose des segments morphologiques et des intervalles, formant l'onde PQRST qui correspond à la conductivité électrique tout au long du cycle cardiaque [40].

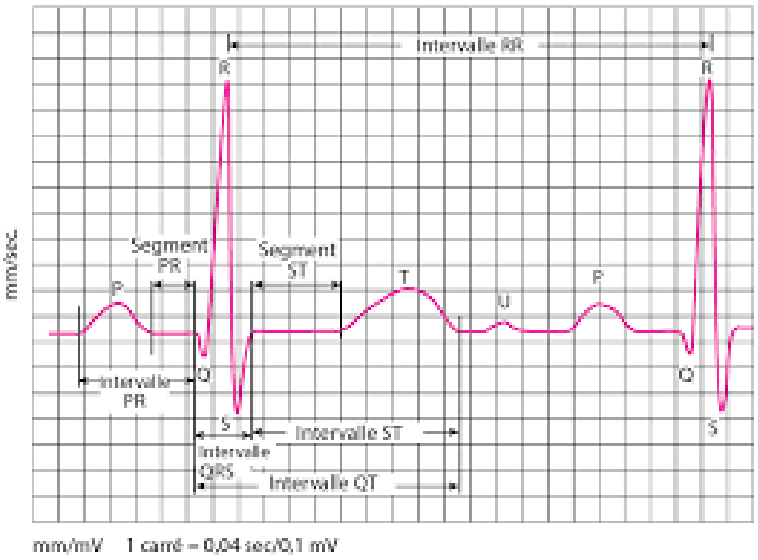


FIGURE 1.4 – Composants du signal d'ECG [7]

1.5.1 Les ondes P, T et U

L'onde P : la dépolarisation des oreillettes :

Caractéristiques de l'onde P normale chez l'adulte :

•Précède le complexe QRS, séparé d'un espace, l'intervalle PR.

•Positive dans les dérivations DI, DII, aVF, et de V4 à V6.

•Négative en en dérivation aVR

•Axe (de l'onde P) dans le plan frontal : 0° à +75°

•Amplitude (hauteur)<= 2,5 mm (ou 2,5 mvolt)

•Largeur <0,120 sec (ou 3 petits carreaux)

•Aspect biphasique habituel en V1

Anomalies de l'onde P :

•Onde P négative en DI : Electrodes inversées ou dextrocardie (situs invertus)

•Onde P négative en DII : Foyer auriculaire ectopique

•Onde P de morphologie différente : Extrasystole, fibrillation auriculaire

•Onde P amples pointues (>2,5 mm), notamment en DII : (Onde P pulmonaire) Hypertrophie auriculaire droite

•Onde P larges, (>=0,120 sec), bifides, notamment en DI (indice de Morris) : (Onde P ventrale) Hypertrophie auriculaire gauche.

L'onde T : repolarisation ventriculaire :

Caractéristiques de l'onde T normale chez l'adulte :

•Positive en DI, DII, et de V4 à V6 (comme l'onde P).

•Négative en aVR (comme l'onde P).

•Aspect asymétrique : pente ascendante lente et partie descendante raide.

•Amplitude (hauteur) < 5mm en périphérie, <10 mm en précordial.

Anomalies de l'onde T :

•Ondes T amples, pointues positives en « forme de tente ». Causes : hyperkaliémie, ischémie myocardique, bloc de branche, AVC, ...

•Ondes T plates (+/-ondes U diffuses) Causes : hypokaliémie (et hypomagnésémie).

•Ondes T négatives pathologique : Causes : infarctus du myocarde, bloc de branche, péricardite, embolie pulmonaire, hypertrophie ventriculaire, imprégnation digitalique, myocardites, rythme ventriculaire.

•Variante physiologique : l'onde T « juvénile » Onde T négative de V1 à V3 uniquement Chez l'enfant et la jeune femme.

L'onde U : repolarisation des fibres de Purkinje ou facteur mécanique de relaxation des ventricules :

Caractéristiques de l'onde U normale chez l'adulte :

•Visible essentiellement en V2 et V3

•Amplitude (hauteur)<=2 mm

•Onde positive

Anomalies de l'onde U :

•Ondes U positives diffuses : Causes : bradycardies importantes, hypothyroïdie, hypokaliémie (et hypomagnésémie), hypocalcémie, hypertrophie ventriculaire, ...

•Ondes U négative : Causes : ischémie myocardique (selon certains auteurs), cardiopathies gauches.

1.5.2 Le complexe QRS : dépolarisation des ventricules

Caractéristiques normales du complexe QRS chez l'adulte :

•Axe (dans le plan frontal) : de -30 à +110°

•Transition (ou rotation) normale : en V3 ou V4

•Aspect RS ou RSr' en V1, avec l'onde R< l'onde S

•Aspect RS ou QRS en V6, avec l'onde R > l'onde S

•Hauteur du complexe QRS>5 mm en périphérie, <25 mm en précordial

•Largeur des complexes <0,120 sec (ou 3 petits carreaux)

Anomalies des complexes QRS :

•L'onde Q non pathologique : Peu large (<1 mm), peu profonde (<2 mm), souvent isolée en DII Rapport onde Q / onde R <0,25

•L'onde Q pathologique : (séquelle d'infarctus du myocarde) Large (>1 mm), et profonde (>2 mm).

Dans au moins 2 dérivations concordantes (appartenant au même territoire vasculaire)

•Le micro voltage : Hauteur des complexes QRS <5 mm dans toutes les dérivations périphériques

Causes : Gene à la diffusion des ondes (tamponnade +++, emphysème, obésité) Faible dépolarisation des ventricules (ischémie coronaire diffuse, amylose, ...)

•Complexes QRS larges : (lenteur anormale de dépolarisation des ventricules) Largeur >0,120 sec (>3 petits carreaux).

Causes : Bloc de branche complet, rythme ventricule, toxiques (hyperkaliémie, antidépresseurs tricycliques, ...)

•Complexes QRS amples en précordial (>25 mm) : Recherche des signes électriques d'hypertrophie ventricules droite ou gauche par la mesure des indices de Cornell, Sokolow, de R1+S5, S6, ...

1.5.3 Les intervalles et les segments

L'intervalle PR (ou PQ) : le temps de conduction auriculo-ventriculaire :

Correspond au délai entre la dépolarisation des oreillettes et celle des ventricules.

Caractéristiques de l'intervalle PR normale :

•Il se mesure du début de l'onde P au début du complexe QRS.

•Intervalle PR normal : de 0,12 à 0,20 sec (3 à 5 petits carreaux)

•Ligne horizontale, isoélectrique

Anomalies de l'intervalle PR :

•Intervalle PR court : durée < 0,12 sec

Causes : syndromes de Wolff parkinson white, de Lown Ganone Levine, myopathies.

•Intervalle PR long : durée >0,20 sec

Causes : causes de bloc auriculo-ventriculaires du 1er, du 2ème ou du 3ème degré

•Le sus décalage du PR : Décalage vers le bas, de l'ordre du mm, par rapport au reste du tracé. Quasi pathognomonique d'une péricardite aigue débutante (présente dans 80% des cas)

L'intervalle QT : le temps de systole ventriculaire :

Correspond au temps de la systole ventriculaire, du début de l'excitation des ventricules jusqu'à la fin.

Caractéristiques de l'intervalle QT normale :

Il se mesure du début du complexe QRS à la fin de l'onde T.

Le QT mesuré est normal lorsqu'il se trouve dans un intervalle de 10% autour du QT calculé, c-à-d lorsqu'il est entre QT bas (QTc -10% du QTc) et QT haut (QTc $+10\%$ du QT c).

Anomalies de l'intervalle QT :

•Intervalle QT court : QTm<Qt bas (<QTc -10%)

Causes : hypercalcémie, digitaliques, maladies congénitales (syndrome de Lown-Ganone-Levine)

•Intervalle QT allongé : QTm> QT haut (>QTc $+10\%$)

Causes : hypocalcémie, maladies congénitales (syndrome de Jerbell et Lange Nielsen, syndrome de Romano Ward), médicaments (quinine, halofantrine, cisapride) ...

•Risque principal : torsade de points puis fibrillation ventriculaire, surtout si très allongé (>QTc +20%)

Le segment ST : le temps de stimulation complète des ventricules :

Correspond à la phase 2 de la repolarisation ventriculaire, phase laquelle les cellules ventriculaires sont toutes dépolarisées donc le segment est isoélectrique.

Caractéristiques du segment ST :

•Se mesure du point J, (fin du complexe QRS) au début de l'onde T.

•Segment horizontal.

•Sur la même ligne que l'intervalle PR qui précède (écart toléré < 1 mm).

Anomalies du segment ST :

•Variante physiologique : la repolarisation précoce sus-décalage ST de 1 à 3 mm de V2 à V4 Sujet jeune, vagotonique, souvent de sexe masculin

•Sus-décalage du segment ST (en-dehors de la repolarisation précoce) Décalage vers le haut du segment ST >1 mm, sur 1 ou plusieurs dérivations

Causes : infarctus du myocarde, bloc de branche, péricardite, syndrome de brugada

•Sous-décalage du segment ST

Décalage vers le bas du segment ST >1 mm

Causes : ischémie myocardique, bloc de branche, digitaliques, embolie pulmonaire

1.5.4 Rythme cardiaque et Fréquence cardiaque

Un rythme dit « sinusal » : l'activité cardiaque sous contrôle du nœud sinusal se caractérise par :

- Un rythme régulier avec un espace R-R constant.
- La présence d'une onde P avant chaque QRS et d'un QRS après chaque onde P.
- Des ondes P d'axe et de morphologie normales.

- Un intervalle PR constant.

Si rythme régulier : « fréquence cardiaque » :

En pratique, on peut la déterminer en divisant 300 par le nombre de petits carrés de 5mm séparant deux complexes QRS : la mémorisation de la séquence «300, 150, 100, 75, 60, 50 »

1.6 L'interprétation de l'ECG

L'électrocardiogramme permet de mettre en évidence diverses anomalies cardiaques et a une place importante dans les examens diagnostiques en cardiologie, comme pour la maladie coronarienne.

Elle ne peut être valide que si l'appareil est correctement étalonné et les électrodes correctement positionnées, l'analyse de l'ECG doit tenir compte de l'âge du patient

Les critères d'un ECG normal :

D'après Pierre Taboulet [44], Un ECG normal est caractérisé par les 12 critères suivants :

•L'onde P sinusale est positive en DI-DII (en dôme).

•Une seule onde P sinusale précède chaque QRS (DII < 2,5 mm et < 0,12 s).

•La fréquence sinusale est comprise entre 60-100/min (adulte).

•L'intervalle P-R (ou P-Q) a une durée constante (0,12 à 0,20 s).

•Les QRS ont un axe frontal entre -30 à 90° (les QRS DI-DII ont une polarité > 0).

•Les QRS sont tous fins (durée maximale d'un QRS < 0,11 s).

•Il n'y a pas d'onde Q (mais il doit exister une micro onde q en V5-V6 et il peut exister une micro onde q fine en frontales, voire QS en DIII).

•Les QRS ont un aspect RS en V1 et qR en V6 (l'onde R croit de V1 à V4(V5) puis décroit et l'onde S croît de V1 à V2 puis décroît).

•Les QRS ont des amplitudes de R et de S modérées (indices d'hypertrophie ventriculaire négatifs) et non microvoltées.

•Le segment ST est isoélectrique au segment PQ (mais sus-décalage de ST possible si variantes).

•L'onde T est asymétrique et positive (sauf en VR et V1 et parfois en DIII-VL) et son amplitude maximum < 2/3 du QRS et minimum > 10% de R.

•L'intervalle QT corrigé est normal (< 0,43 s homme et < 0,45 s femme)

1.7 Conclusion

Dans cette partie, nous avons décrire les signaux ECG qui jouent un rôle très important pour la détection des anomalies cardiaques et l'établissement du diagnostic médical. Pour cela, en passant par la présentation du fonctionnement du cœur humain, les phases de l'ECG et les principales caractéristiques d'ECG, et les diverses formes du tracé ECG (dérivation).

Dans le chapitre suivant, il devra nécessaire de faire une synthèse sur les techniques les plus récentes pour l'étude descriptive de ces données pour la classification des données ECG.

CHAPITRE 2

ETAT DE L'ART SUR LES TECHNIQUES D'APPARIEMENT DE DONNÉES ECG

2.1 Introduction :

Rappelons que notre travail a pour but d'effectuer l'appariement entre les données ECG en vue de classifier les ECG normaux de l'autres anormaux afin de détecter l'existence de pathologies cardiaques.

La classification de différents types de cardiopathies à partir de l'électrocardiogramme (ECG) représente une importance majeure dans le diagnostic des dysfonctionnements cardiaques.

Après avoir présenté les données médicales pour l'électrocardiogramme dans le chapitre 1, nous présenterons dans ce deuxième chapitre un état de l'art sur les techniques d'appariement de données ECG.

Dans la première partie nous allons présenter les techniques de l'apprentissage profond avec les différentes architectures de réseau de neurones profonds. Le chapitre termine par présenter les travaux récents sur l'appariement de données ECG.

2.2 Machine Learning et Deep Learning :

Machine Learning (ML) ou apprentissage automatique ou encore apprentissage artificiel, est une forme d'intelligence artificielle (IA) qui permet à un système d'apprendre à partir des données et non à l'aide d'un ensemble des règles explicites. Un modèle de machine Learning est le résultat généré après l'entraînement de l'algorithme d'apprentissage automatique avec des données. [3]. Les modèles de Machine Learning ML sont basés sur des exemples connus (des paires entrée/sortie). Cependant, ces algorithmes dépendent de type de données et qui ne sont pas toujours utiles pour la classification de données ECG, notamment que lorsque les données contiennent plusieurs caractéristiques (features) avec un volume très grand. Le problème réside dans le fait que les modèles de ML exigent une phase critique appelée l'ingénierie des caractéristiques (Feature Ingeneering) où l'expert de domaine sélectionne les caractéristiques importantes des données.[13] Ces modèles requièrent

beaucoup temps non seulement pour le prétraitement de données (ex. réduction dimensionnelle) mais aussi pour l'apprentissage, la résolution des problèmes, car ils nécessitent l'intervention des experts humains qui élaborent des caractéristiques utiles à partir des données brutes de l'ECG, appelées "caractéristiques d'expert", puis déploient des règles de décision ou d'autres méthodes d'apprentissage automatique pour générer les résultats finaux, comme le montre la partie supérieure de la figure 2.1 [20]

Le Deep Learning (DL), ou apprentissage profond, est un sous-ensemble du Machine Learning, basé sur des réseaux neuronaux artificiels [16]. Durant les dernières années, Deep Learning a fait l'objet de nombreuses études et a obtenu des résultats remarquables dans de nombreux domaines, y compris dans le domaine de la santé (par exemple la détection d'anomalies sur ECG, qui jouer un rôle vital dans le suivi des patients) [10]. Il a montré ses performances face aux différents problèmes de l'intelligence artificielle dépassant ainsi les algorithmes classiques de ML [10].

Le principal avantage des algorithmes d'apprentissage profond est qu'ils essaient d'apprendre les caractéristiques de haut niveau à partir de données de manière incrémentale, et donc les données peuvent être transmises directement au réseau de neurone profond pour obtenir de bons scores dès le départ [4]. Cela élimine totalement la phase d'extraction manuelle des caractéristiques du processus d'apprentissage, comme le montre le bas de la figure 2.1

Nous nous focalisons sur l'utilisation des techniques de l'apprentissage profond pour l'appariement de données ECG et la détection des anomalies cardiaques.

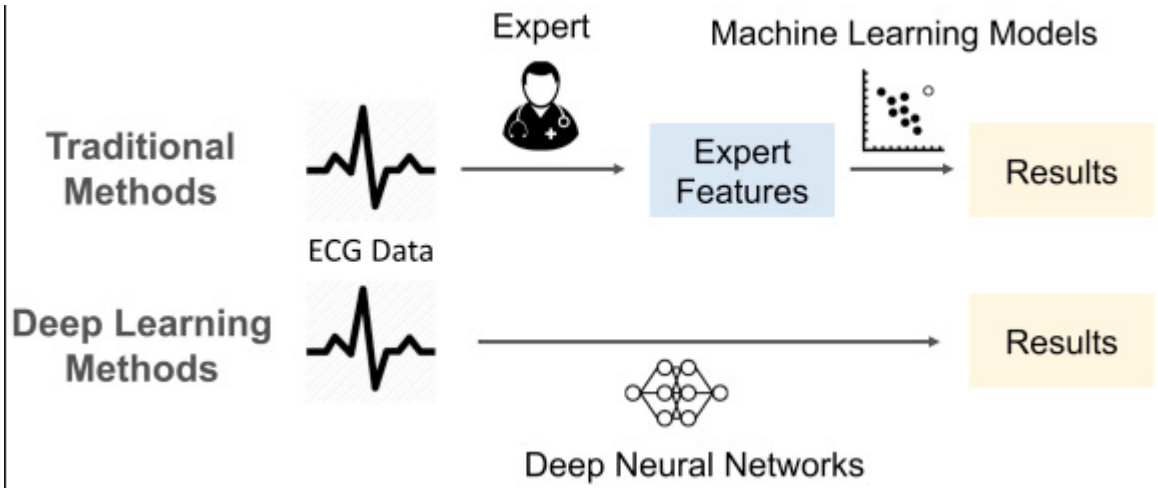


FIGURE 2.1 – Machine Learning et Deep Learning [20]

2.3 Les réseaux de neurones profonds DNN :

sont une classe d'algorithmes d'apprentissage automatique similaires au réseau de neurones artificiels et visent à imiter le traitement de l'information du cerveau humain.Un réseau de neurones profonds se compose de couche d'entrée (Input Layer), une ou plusieurs couches cachées (Hidden Layers) et une couche de sortie (Output Layer) [39].Chaque paire de couches voisines est connectée par des connexions appelées synapses qui sont associés par des poids (Weights). La figure 2.2 illustre une architecture standard d'un réseau de neurones profonds.

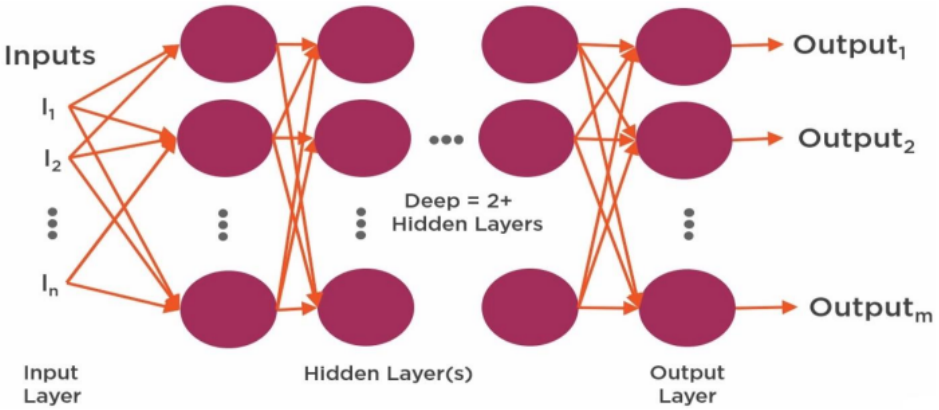


FIGURE 2.2 – architecture d'un DNN [39]

Pour comprendre le fonctionnement de ces réseaux et le processus d'apprentissage, nous devons d'abord comprendre le fonctionnement interne d'un seul neurone, illustré dans la figure 2.3.

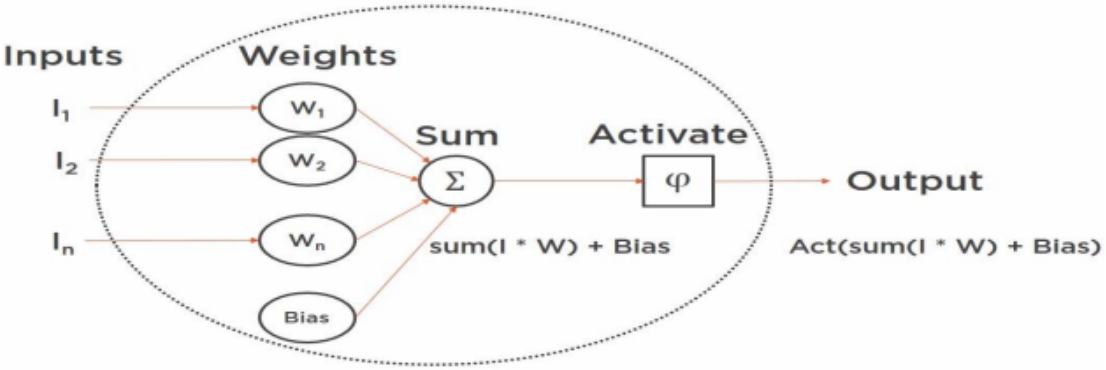


FIGURE 2.3 – fonctionnement interne d'un neurone [39]

Le neurone fait une simple somme mathématique des poids multipliés par les valeurs d'entrée et ajoute un biais. Le produit de ces opérations passe par une fonction d'activation non linéaire dont la sortie est la sortie du neurone [39]

Fonction d'activation :

$$\text{y =a(} \sum_{i=1}^{n} wx + b\text{)=a(wx+b)}$$

Avec :

- W : est le vecteur des poids.
- x : est le vecteur des entrées.
- b : désigne le biais.
- a : représente la fonction d'activation

L'une des principales caractéristiques du réseau neuronal est sa capacité à utiliser les données d'entrée pour former les poids et les biais afin que le signal sortant du neurone change en fonction des données d'entrée. Pour effectuer cette formation, nous exposons le réseau à des données, avec chaque ensemble de données, un algorithme est utilisé pour ajuster les poids et les biais afin de minimiser l'erreur du réseau dans la prédiction des valeurs des données. Cela se fait à travers des processus

appelé propagation avant et propagation arrière. Lorsque ces processus sont terminés, on dit que le réseau est formé et que les poids et les biais de tous les neurones ont été ajustés pour donner les meilleurs résultats sur les données d'entrainement. Cette opération de formation ou d'apprentissage peut prendre différentes formes selon le type de données dont nous disposons pour alimenter le réseau, elle peut être supervisée, non supervisée ou semi-supervisée (hybride). [39]

Les réseaux de neurones profonds (DNN) ont récemment remporté des succès remarquables dans des tâches telles que la classification d'images et la reconnaissance vocale [38].

2.4 Les réseaux de neurones convolutifs CNN :

Convolutional Neural Networks (CNN) sont des réseaux capables de reconnaître des modèles simples dans les données. Plus le nombre de couches est élevé, plus le modèle est complexe. Les CNN peuvent être appliqués à des séquences de données unidimensionnelles (comme l'ECG), bidimensionnelles (comme les images) ou même tridimensionnelles (comme la vidéo) [4].

CNNs sont l'architecture la plus utilisée pour les problèmes de classification des séries chronologiques, qui prend en entrée une série chronologique multivariée, est capable de capturer avec succès les modèles spatiaux et temporels à travers les filtres entraînables de l'application, et attribue une importance à ces modèles à l'aide de poids entraînables.Le prétraitement requis dans un réseau de neurones convolutifs est beaucoup plus faible que celui d'autres algorithmes de classification. Alors que dans de nombreuses méthodes, les filtres sont conçus à la main, le réseau neuronal convolutif a la capacité d'apprendre ces filtres [27].

Comme nous pouvons le voir dans la figure 2.4 qui illustre l'architecture d'un réseau de neurones convolutifs est composé de trois couches différentes [27] :

Couche convolutive (Convolutional Layer) : elle consiste à extraire les caractéristiques de haut niveau grâce à une opération de convolution avec une matrice de taille

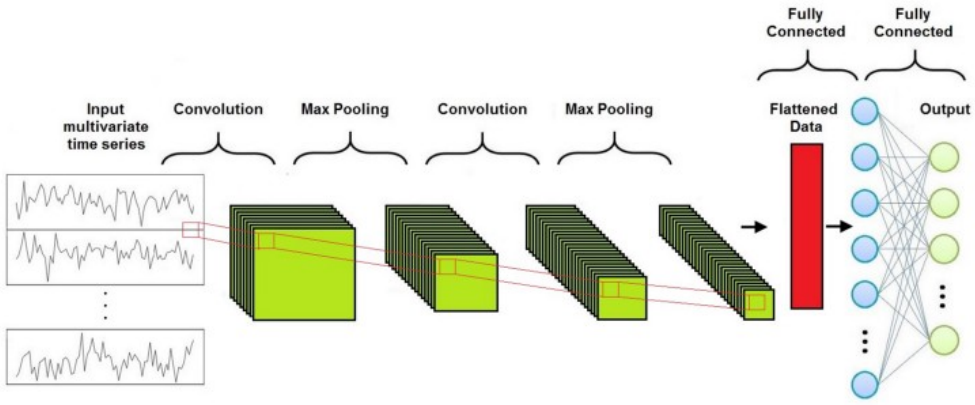


FIGURE 2.4 – L'architecture d'un CNN [27]

fixe « filtre » qui est appliqué à une sous-matrice de la carte d'entités en entrée pour donner par la somme du produit de chaque élément du filtre avec l'élément dans la même position de la sous-matrice (voir la figure 2.5) [27]

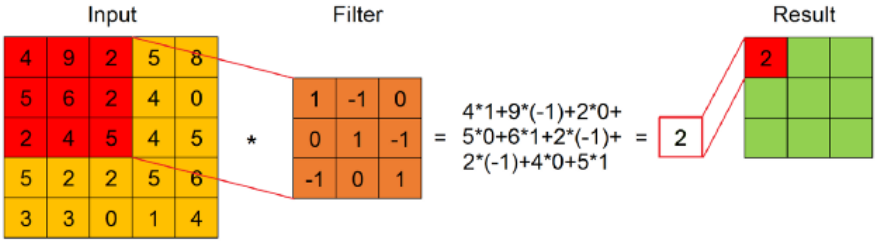


FIGURE 2.5 – Exemple d'application de filtre [27]

Les deux paramètres importants qui doivent être choisis pour la couche convolutive sont la foulée (stride) et le rembourrage (Padding).

• Stride : contrôle la manière dont le filtre s'articule autour de la carte des caractéristiques en entrée. En particulier, la valeur de stride indique combien d'unités doivent être décalées à la fois, comme indiqué sur cette figure 2.6 .

• Padding : indique le nombre de colonnes et de lignes supplémentaires à ajouter en dehors d'une carte de caractéristiques d'entrée, avant d'appliquer un filtre de convolution, comment nous pouvons voir dans cette figure 2.7.

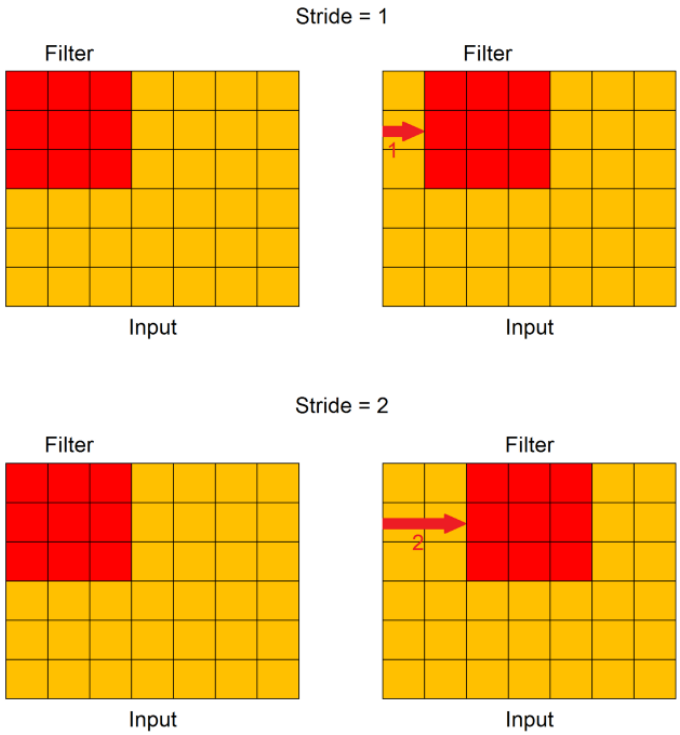


FIGURE 2.6 – Mécanisme de stride [27]

Généralement, toutes les unités des nouvelles colonnes et lignes ont une valeur fictive 0.

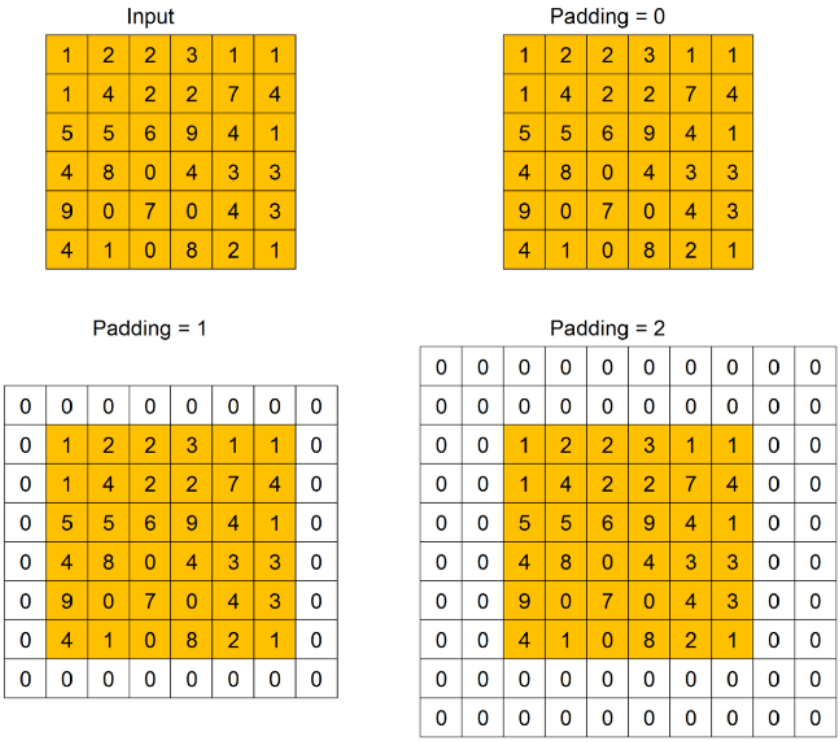


FIGURE 2.7 – Mécanisme de Padding [27]

Après avoir appliqué de nombreux filtres, la taille peut devenir trop petite à cause de padding. En ajoutant des lignes et des colonnes supplémentaires, nous pouvons conserver la taille d'origine ou la faire diminuer plus lentement. Lorsque la taille de sortie est égale ou supérieure à la taille d'entrée, nous appelons cette opération Same Padding [27].

Couche de regroupement (pooling layers) : c'est une couche de mise en commun pour obtenir une réduction de dimension des cartes des caractéristiques, en préservant autant d'informations que possible. Il est également utile pour extraire les caractéristiques dominantes qui sont invariantes en rotation et en position. Prend comme entrée une série de cartes d'entités et sa sortie est une série différente de cartes d'entités, avec une dimension inférieure [27]

Max Pooling le résultat de l'opération de regroupement est sa valeur maximale.

Average Pooling le résultat de l'opération de regroupement est sa valeur minimale.

La figure 2.8 montre un exemple d'opération de max pooling et Average pooling avec une taille de filtre 2x2.

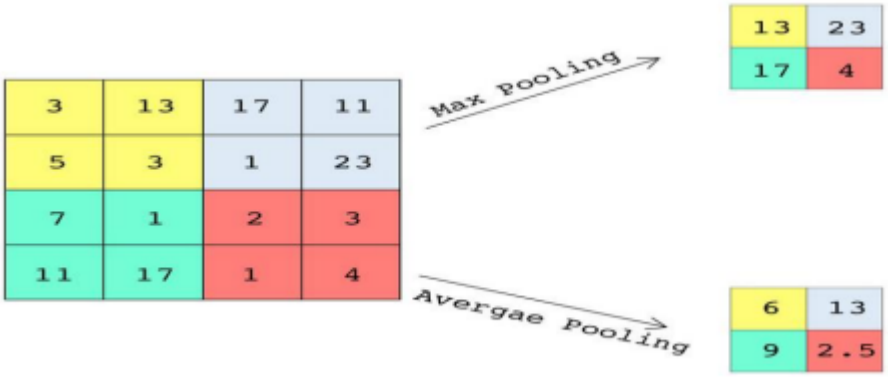


FIGURE 2.8 – Exemple de fonctionnement de Max pooling et Average pooling [27]

L'avantage de l'opération de mise en commun est de sous-échantillonner les bandes de sortie convolutionnelles, réduisant ainsi la variabilité des activations cachées [27].

Couche entièrement connectée (Fully-Connected Layer FCL) : Cette couche est à la fin du réseau pour apprendre des combinaisons non linéaires des fonctionnalités de haut niveau représentées par la sortie de la couche convolutive et de la couche de regroupement. Cette couche est implémentée avec un Perceptron multicouche. [27]

La rétro propagation est appliquée à chaque itération de l'entrainement. Sur une série d'époques, le modèle est capable de distinguer les séries temporelles comme type de données d'entrée grâce à leurs caractéristiques dominantes de haut niveau et de les classer [27].

2.5 Réseaux de neurones récurrents RNN :

Recurrent Neural Network (RNN) est considéré comme un réseau avec mémoire (Lai, Chen et Caraka 2019) [46].

Les réseaux neuronaux récurrents sont conçus pour interpréter des informations temporelles ou séquentielles. Ces réseaux utilisent d'autres points de données dans une séquence pour faire de meilleures prédictions. Pour ce faire, ils prennent des données en entrée et réutilisent les activations des nœuds précédents ou des nœuds ultérieurs dans la séquence pour influencer la sortie. RNNs sont capables de se souvenir d'éléments importants concernant les entrées qu'ils ont reçues, ce qui leur permet d'être très précis dans la prédiction de ce qui va suivre [12].

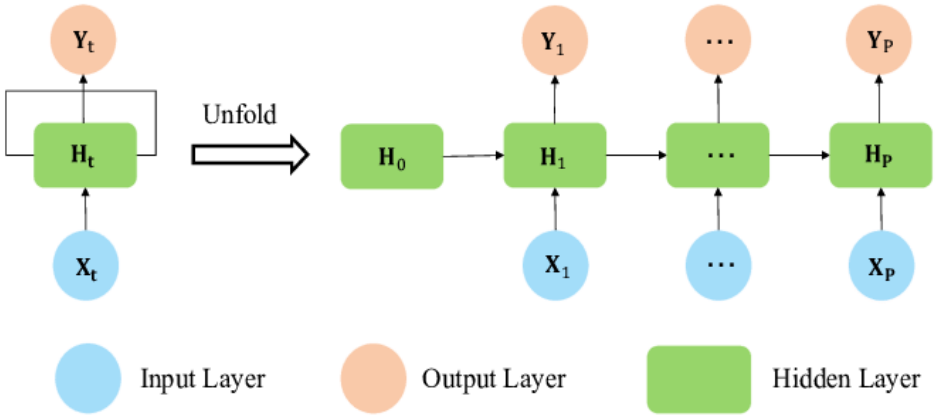


FIGURE 2.9 – architecture de RNN [37]

Dans la figure ci-dessus 2.9, "x" est la couche d'entrée, "H" est la couche cachée et "y" est la couche de sortie.

À n'importe quel moment donné l'entrée de courant est une combinaison des entrées à x (t) et x (t-1).

La sortie à un certain moment est récupérée du réseau pour optimiser la sortie.

Parmi les sous-types de réseaux de neurones récurrents (RNN), nous distinguons LSTM car il est très utilisé pour les séries chronologiques.

2.5.1 Réseau Long short-term memory (LSTM) :

Le réseau Long short-term memory (LSTM) est une architecture de réseau neuronal récurrent (RNN) artificielle, utilisée dans le domaine de l'apprentissage profond. Contrairement aux réseaux neuronaux à anticipation standard, le LSTM possède des connexions de rétroaction. Il peut non seulement traiter des points de données uniques (comme des images), mais aussi des séquences entières de données (comme la parole ou la vidéo) [19].

Les réseaux LSTM sont bien adaptés à la classification, au traitement et aux prédictions basés sur des données de séries temporelles, car il peut y avoir des décalages de durée inconnue entre des événements importants dans une série temporelle.

Les LSTM ont été développés pour résoudre le problème du gradient de fuite qui peut être rencontré lors de l'entraînement des RNN traditionnels. L'insensibilité relative à la longueur du décalage est un avantage des LSTM par rapport aux RNN, aux modèles de Markov cachés et à d'autres méthodes d'apprentissage de séquences dans de nombreuses applications [19].

2.6 Les travaux récents sur l'appariement de données ECG :

Nous présentons dans cette section les travaux les plus récents sur l'appariement de données ECG.

Jun, T. J. et al. (2018) [23] : Ils ont proposé une méthode efficace de classification des arythmies ECG en utilisant des réseaux neuronaux convolutifs bidimensionnels avec une image ECG en entrée. Des images en niveaux de gris 128 x 128 sont transformées à partir de l'enregistrement ECG de la base de données d'arythmie MIT-BIH. Plus de 100 000 images de battements d'ECG sont obtenues avec huit types de battements d'ECG, y compris les battements normaux et sept battements d'arythmie. Le modèle CNN optimisé est conçu en tenant compte de concepts importants tels que l'augmentation des données, la régularisation et la validation croisée K-fold.

En conséquence, le système proposé a obtenu une AUC de 0,989, une précision moyenne de 99, 05% , une spécificité de 99, 57%, une sensibilité moyenne de 97, 85% et une valeur prédictive positive moyenne de 98, 55%. Le résultat de classification de l'arythmie ECG indique que la détection de l'arythmie à l'aide d'images ECG et d'un modèle CNN peut être une approche efficace pour aider les experts à diagnostiquer les maladies cardiovasculaires qui peuvent être observées à partir des signaux ECG. En outre, la méthode de classification de l'arythmie ECG proposée peut être appliquée au robot médical ou au scanner qui peut surveiller les signaux ECG et aider les médecins.

Wang et al. (2019) [49] ont proposé un schéma de classification global et actualisable nommé Global Recurrent Neural Network (GRNN) avec quatre couches au total. Dans la partie morphologique, ils ont combiné la mémoire à long terme (LSTM) et un module de lancement pour la détection de CHF. Cinq bases de données open source ont été utilisées pour la formation et les tests, de plus, trois types de longueur de segment RR (N = 500, 1000 et 2000) ont été utilisés pour la comparaison avec d'autres études. La méthode proposée a atteint une précision de 99, 22%, 98, 85% et 98, 92% sur le dataset BIDMC.

Sajad Mousavi, Fatemeh Afghah. (2019) [30] : proposent une nouvelle approche pour la classification automatique des battements de cœur basée sur l'ECG en exploitant une méthode d'apprentissage profond de séquence à séquence et une méthode de suréchantillonnage appelée Synthetic Minority Over-sampling Technique (SMOTE) pour relever le défi mentionné avec les classes minoritaires, exploitant un modèle RNN de séquence à séquence avec un réseau de neurones convolutifs (CNN).

Cette étude utilise la base de données d'arythmie PhysioNet MIT-BIH pour évaluer les performances de la méthode proposée via les méthodes d'évaluation inter-patients et intra-patients. Elle est aussi basée sur l'utilisation de la technique de suréchantillonnage synthétique des minorités (SMOTE) qui génère les points de données synthétiques. Les auteurs ont entraîné le modèle proposé avec 80% de l'ensemble de données et l'avons évalué avec les 20% restants.

Le réseau proposé avec un faible nombre de paramètres (c'est-à-dire avec une taille maximale de 5,5 Mo) peut être utilisé avec des dispositifs portables.

Q. Yao, R. Wang and X. Fan et al. (2020) [52] : ont proposé un modèle de classification de L'ECG à 12 dérivations. Une génération d'une série temporelle prétraitée dans les couches entièrement convolutives. Cette série temporelle est ensuite introduite dans les cellules LSTM pour échanger des informations entre différents moments. Un module d'attention accepte la sortie des cellules LSTM, attribue des pondérations pour différents moments et génère un résultat final. Les données d'entraînement utilisées dans cette étude provenaient du 1er China Physiological Signal Challenge. Ils ont obtenu un F-score moyen de 81, 2% dans la classification de 8 types d'arythmies et de rythme sinusal, dépassant de 7, 7% le modèle CNN de référence. Le mécanisme d'attention aide le modèle à localiser la partie informative des signaux et améliore l'interprétabilité.

Klosowski, G et al.(2020) [25] : proposent une méthode de classification des signaux ECG basée sur l'extraction spectrale de caractéristiques à l'aide de la transformée logarithmique et l'utilisation du réseau de neurones RNN avec LSTM donne de bons résultats. Dans ce modèle, un seul signal ECG brut a été transformé en deux signaux de spectrogrammes, générés par diverses transformations temporelles, ce qui a augmenté l'efficacité de la prédiction. La précision moyenne du LSTM pour l'ensemble de tests était de 70,8

Ribeiro, Antônio H., et al. (2020) [38] : Ont utilisé un réseau neuronal convolutif similaire au réseau résiduel, adapté aux signaux unidimensionnels. Cette architecture permet d'entraîner efficacement les DNN en incluant des connexions de saut pour détecter : le bloc AV de 1er degré (1dAVb), le bloc de branche droit (RBBB), bloc de branche gauche (LBBB), bradycardie sinusale (SB), fibrillation auriculaire (AF) et tachycardie sinusale (ST).

Ils ont collecté un ensemble de données composé de 2 322 513 enregistrements d'ECG provenant de 1676 384 patients différents de 811 comtés de l'État de Minas Gerais/Brésil.

Ils ont divisé cet ensemble de données en un ensemble d'entraînement et un ensemble de validation. L'ensemble d'entraînement contient 98% des données. L'ensemble de validation est constitué des 2% restants (50 000 examens) de l'ensemble de données.

Ils ont développé un DNN de bout en bout capable de reconnaître avec précision six anomalies de l'ECG dans les examens S12L-ECG avec une performance diagnostique au moins aussi bonne que celle des résidents et des étudiants en médecine. Cette étude montre le potentiel de cette technologie, qui, lorsqu'elle sera pleinement développée, pourrait conduire à un diagnostic automatique plus fiable et à une meilleure pratique clinique.

Weimann, K., Conrad, T. O. (2021) [50] : Dans ce travail, ils ont utilisé l'apprentissage par transfert pour améliorer les réseaux de neurones convolutifs (CNN) formés pour classifier le rythme cardiaque à partir d'un court enregistrement ECG. Tout-d'abord, ont prétraîné les CNN sur un grand ensemble de données de signaux ECG bruts continus. Ensuite, ils ont entraîné les réseaux sur un petit ensemble de données pour la classification de la fibrillation auriculaire (FA). La performance des méthodes de pré-entraînement a été mesurée sur le jeu de données PhysioNet/CinC Challenge 2017.

Ils ont montré que le pré-entraînement des CNNs améliore la performance de la tâche cible, c'est-à-dire la classification de la fibrillation auriculaire (FA), jusqu'à $6,57\%$, réduisant efficacement le nombre d'annotations nécessaires pour atteindre la même performance que les CNNs qui ne sont pas prétraités. En outre, ils ont montré que le pré-entraînement non supervisé sur des données ECG améliore la performance sur la tâche cible, en particulier pour la classification de la FA. Bien que dans une moindre mesure que le pré-entraînement supervisé. Néanmoins, ont pensé que le pré-entraînement non supervisé deviendra plus pertinent car il ne repose pas sur les annotations, qui sont coûteuses à acquérir pour les données ECG.

2.7 Conclusion

La classification automatique des pathologies cardiaques est d'un grand intérêt dans le domaine médical, ce qui a intéressé plusieurs chercheurs dont nous avons cité les travaux concernant les différentes techniques utilisées pour l'appariement des données ECG. Nous avons présenté les techniques de Deep Learning les plus utilisés qui sont liées à notre travail comme les CNNs et RNNs.

Sur la base de ces travaux, nous proposons notre approche d'appariement d'ECG améliorant les approches existantes.

CHAPITRE 3

LES SÉRIES CHRONOLOGIQUES : NOTIONS ET OUTILS

3.1 Introduction

Dans notre travail, nous nous intéressons aux données de séries chronologiques pour la représentation et l'exploration de données médicales de l'électrocardiogramme de maladies cardiovasculaire.

Les données ECG sont vues comme une suite d'observations indexées par le temps. Cette description est fortement liée au modèle de données évolutives changent de valeurs dans des intervalles du temps. Ce modèle est appelé les séries chronologiques ou temporelles dans le sens où chaque changement de données est lié au temps. Il s'avère nécessaire de présenter ce modèle de données afin qu'on puisse modéliser, collecter et stocker les données ECG à analyser. C'est justement l'objet de ce que ce chapitre présent.

3.2 Les séries chronologiques

3.2.1 Définition :

Une série chronologique est une collection d'observations d'éléments de données bien définis résultant de mesures répétées au fil du temps. Les données de séries chronologiques sont indexées dans l'ordre temporel qui est une séquence de points de données [18].

Une série chronologique est vue comme un tableau de valeurs observées à un instant t données. La représentation formelle d'une série chronologique S est la suivante [24] :

$$S(tn)\ n \in\ N$$

Avec : si un élément de série chronologique S à un instant t i. Il est défini comme suit : si =S(ti). n est le nombre d'éléments de S.

3.2.2 Les composants d'une série chronologique :

On considère qu'une série chronologique est la résultante de différentes composantes fondamentales suivantes [41] :

•**La tendance** (ou trend) : représente l'évolution à long terme de la série étudiée. Elle traduit le comportement ″moyen″ de la série. Une tendance peut être positive ou négative selon que la série chronologique présente une tendance à long terme croissante ou une tendance à long terme décroissante. Par exemple la croissance de nombre de morts dus à la pandémie de coronavirus.

•**La composante saisonnière** (ou saisonnalité) : correspond à un phénomène qui se répète à des intervalles de temps réguliers (périodiques). En général, c'est un phénomène saisonnier d'où le terme de variations saisonnières. Elle présente des cycles réguliers au cours du temps et de même amplitude. Par exemple, les ventes au détail culminent au cours du mois de décembre.

•**La composante résiduelle** (ou bruit ou résidu) : correspond à des fluctuations irrégulières, en général de faible intensité mais de nature aléatoire. On parle aussi d'aléas.

•**Changement cyclique** : Tout modèle montrant changement qui se répète périodiquement est appelé un changement cyclique. La variation cyclique est une composante non saisonnière qui varie dans un cycle reconnaissable, par exemple les battements de cœur. Il s'agit d'un phénomène se répétant mais contrairement à la saisonnalité sur des durées qui ne sont pas fixes et généralement plus longues.

3.2.3 Types des séries chronologiques :

Il existe plusieurs classifications des séries chronologiques, la plus générale est celle liées à la nature de données en termes de valeurs connectées ou déconnectées.

Les séries chronologiques sont composé en deux types, Série temporelle continue et discrète [41] :

Série temporelle continue : Une série temporelle est dite continue lorsque l'observation est faite de manière continue dans le temps. Le terme continu est utilisé pour une série de ce type même lorsque la variable mesurée ne peut prendre qu'un ensemble discret de valeurs.

Série temporelle discrète : Une série temporelle est dite discrète lorsque les observations sont prises à des moments précis, généralement équidistants. Le terme discret est utilisé pour une série de ce type même lorsque la variable mesurée est une variable continue.

3.3 Les données de séries chronologiques

3.3.1 Définition :

les données de séries chronologiques, également appelées données horodatées, sont une séquence de points de données indexés dans l'ordre temporel. Les données horodatées sont collectées à différents moments dans le temps [21].

3.3.2 Types de données de séries chronologiques :

Les données de séries chronologiques peuvent être classées en deux types [21] :

Données de séries chronologiques linéaires : Une série chronologique linéaire est une série dans laquelle, pour chaque point de données X t, ce point de données peut être considéré comme une combinaison linéaire de valeurs ou de différences passées ou futures.

Dans le cas de notre étude, les données ECG sont vue comme des séries chronologiques linéaires et continues.

Données de séries chronologiques non linéaires : Elles sont générées par des équations dynamiques non linéaires. Elles ont des caractéristiques qui ne peuvent pas être modélisées par des processus linéaires : variance changeante dans le temps, cycles asymétriques, structures à moments plus élevés, seuils et ruptures.

3.3.3 Les domaines d'utilisation des données de séries chronologiques :

Les données de séries chronologiques sont collectées, stockées, visualisées et analysées à diverses fins dans divers domaines [21] :

•Dans l'exploration de données, la reconnaissance de formes et l'apprentissage automatique, la classification des données évolutives, la requête par contenu, la détection d'anomalies et la prévision.

•Dans le traitement du signal, l'ingénierie de contrôle et l'ingénierie des communications : les données de séries chronologiques sont utilisées pour la détection et l'estimation du signal.

•Dans les statistiques, l'économétrie, la finance quantitative, la sismologie, la météorologie et la géophysique : l'analyse des séries chronologiques est utilisée pour les prévisions.

3.4 Base de données de séries chronologiques

3.4.1 Définition :

Time Series database (TSDB) est une base de données optimisée pour les données horodatées ou de séries chronologiques. Les données de séries chronologiques sont simplement des mesures ou des événements qui sont suivis, surveillés, sous-échantillonnés et agrégés au fil du temps [21].Une TSDB permet à ses utilisateurs de créer, d'énumérer, de mettre à jour, de détruire et d'organiser diverses séries chronologiques de manière plus efficace [42].

3.4.2 Les Propriétés de données de séries chronologiques :

Les principales propriétés qui distinguent les données de séries temporelles des charges de travail de données ordinaires sont la compression, la gestion du cycle de vie des données et le balayage à grande échelle de nombreux enregistrements. Voici un aperçu de certaines des propriétés requises d'une TSDB [31].

Emplacement des données : Si les données liées ne sont pas situées ensemble dans le stockage physique, les requêtes de données peuvent être vraiment lentes et même entraîner des interruptions de service. Car les opérations d'E/S non séquentielles sont toujours très lentes par rapport aux E/S séquentielles, même en utilisant un SSD. Une TSDB colocalisé des blocs de données situées dans la même plage de temps sur la même partie physique du cluster de base de données et permet donc un accès rapide et plus efficaces.

Interrogation rapide et facile des plages : Comme une TSDB conserve les données liées entre elles, les requêtes sur les plages sont rapides.

Dans de nombreux cas, les bases de données ordinaires produisent une erreur d'indexation en raison du volume des données de séries chronologiques et affectent par la suite les performances des opérations de lecture et d'écriture.

d'écriture élevées : Les bases de données TSDB doivent garantir une haute disponibilité et des performances élevées pour les opérations de lecture et d'écriture pendant les pics de charge car elles sont généralement conçues pour rester disponibles même dans les conditions les plus exigeantes. Les données de séries chronologiques sont généralement enregistrées toutes les secondes, voire moins, de sorte que les opérations d'écriture doivent être rapides. Doivent être rapides.

La compression des données : Comme les données de séries temporelles sont le plus souvent enregistrées par seconde ou même avec une granularité moindre, elles nécessitent généralement une meilleure technique de compression des données. Et comme les données vieillissent, la granularité devient moins importante. Les bases de données TSDB doivent donc fournir une fonctionnalité permettant d'effectuer des roll-up dans de tels scénarios pour compacter les données.

Évolutivité : les données de séries chronologiques augmentent très rapidement et les bases de données classiques ne sont pas conçues pour gérer cette évolutivité. D'autre part, les bases de données de séries chronologiques sont conçues pour prendre

en charge l'échelle en introduisant des fonctionnalités qui ne sont possibles que lorsqu'on traite le temps. Cela peut entraîner des améliorations des performances, notamment : des taux d'insertion plus élevés, des requêtes plus rapides à grande échelle et une meilleure compression des données.

Convivialité : les TSDB incluent généralement des fonctions et des opérations pour l'analyse de données de séries chronologiques. Par exemple, ils utilisent des politiques de conservation des données, des requêtes continues, des agrégations de temps flexibles, des requêtes de plage, etc. Cela augmente donc la convivialité en améliorant l'expérience utilisateur en cas d'analyse liée au temps.

3.4.3 Principales bases de données sur les séries chronologiques :

De nos jours, de nombreuses bases de données de différents types existent et répondent chacune à des besoins bien précis. Nous pouvons citer [22] :

• **Les bases de données relationnelles :** propres pour le traitement de séries chronologiques et offrent des techniques de stockage et de manipulation de temps comme par exemple le système TokuDB fondé sur le SGBD MySQL.

•**Les entrepôts de données (Data Warehouse) :** sont conçues pour stocker des données liées à une période temporelle (des historiques), et pour gérer les gros volumes issus de multiples sources de données. Exemple la plate-forme SHAPE basée sur un entrepôt de données stockant la consommation quotidienne d'électricité.

•**Les bases de données NoSQL :** Hadoop TS permet de distribuer le stockage et la charge de calcul sur plusieurs machines, OpenTSDB basé sur le système HBASE et le paradigme MapReduce.

•**Les bases de données tourniquet (Round-Robin database RRD) :** un exemple de l'outil Round-Robin database Tool (RRDtool) pour la sauvegarde de données cycliques et le tracé de graphiques, de données chronologiques.

•**Les bases de données à grande échelle :** TimescaleDB qui rend SQL évolutif pour les données de séries chronologiques. Il est basé sur le SGBD relationnel PostgreSQL

qui est très puissant pour la manipulation des bases de données temporelles. Alibaba Cloud High-Performance Time Series Database (HiTSDB) prend en charge l'écriture fiable de données de séries chronologiques à grande échelle. InfluxDB qui représente le système le plus performant pour la gestion de séries temporelles d'après le site officiel de classement des systèmes de gestion de bases de données de séries temporelles SGBDST.

Pour voir les tendances au fil du temps, le graphique suivant 3.1 montre les 10 principales bases de données de séries chronologiques et leurs modifications historiques [21] :

RANK	DBMS	SCORE		
MAY 2022		MAY 2022	24 MOS ▲	12 MOS ▲
1	**InfluxDB**	**29.55**	**+8.37**	**+0.91**
2	Kdb+	8.98	+3.12	+0.66
3	Prometheus	6.13	+1.54	+0.21
4	Graphite	5.46	+1.83	+0.82
5	TimescaleDB	4.70	+2.52	+1.46
6	ApacheDruid	3.00	+1.05	+0.19
7	RRDtool	2.50	-0.40	-0.19
8	OpenTSDB	1.84	-0.19	+0.00
9	DolphinDB	1.65	+1.18	+0.69
10	Fauna	1.36	+0.17	-0.31

FIGURE 3.1 – Classement de SGBDST en mois d'Avril 2022TSDBs [21]

3.4.4 Les avantages de TSDB :

Une base de données de séries chronologiques est conçue spécifiquement pour [21] :

• Gérer les métriques et les événements ou les mesures horodatées. Les propriétés qui rendent les données de séries chronologiques très différentes des autres charges de travail de données sont la gestion du cycle de vie des données, la synthèse et les analyses à grande échelle de nombreux enregistrements.

• La gestion du cycle de vie des données, la synthèse et les analyses à grande échelle de nombreux enregistrements et des requêtes prenant en compte les séries chronologiques.

3.5 Stockage des séries chronologiques

Généralement, il n'existe pas de méthode standard bien définie pour modéliser et stocker les données de séries chronologiques. La plupart des ensembles de données des séries chronologiques sont stockées dans des fichiers aux formats CSV, JSON, XML.

Dans notre travail, nous utilisons le SGBD le plus populaire InfluxDB pour le stockage de données modélisées par les séries chronologiques.

3.6 Visualisation des séries chronologiques

C'est une fonction très importante pour l'analyse des séries chronologiques. Elle est effectuée à travers des graphiques. Ces graphiques mettent en évidence visuellement le comportement et les modèles des données voir la figure 3.2 [11].

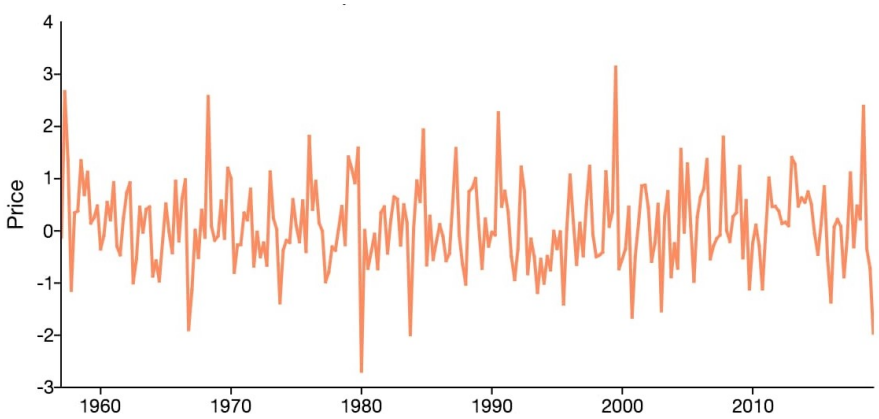


FIGURE 3.2 – Exemple d'une représentation graphique d'une série chronologique [11]

Dans le contexte de notre travail, la visualisation de données ECG est une étape primordiale pour aide à l'interprétation et à la prise de décision. La visualisation peut être établie par un programme en langage python en exploitant leurs librairies comme par exemple, Seaborn, Matplotlib. Il existe aussi des outils puissants pour la visualisation et l'analyse de ce type de données comme Chronograf de système InfluxDB et Grafana. Dans notre travail, nous utilisons Grafana comme outil de base pour la visualisation de données ECG.

L'objectif d'utiliser Grafana est de présenter facilement et de façon intuitive une grande quantité des données issues de sources différentes.

Elle est conçue pour des dashboards sur la base de métriques et de données temporelles de plusieurs sources de données comme InfluxDB, OpenTSDB, Graphite.

3.7 Types d'analyse des séries chronologiques

Il existe différents types et modèles d'analyse de données permettant d'explorer, de décrire, de classifier, de prédire des données de base de données [43]

•**Classification :** ou regroupement ou encore partitionnement (Clustering) qui permet d'identifier et d'attribuer des catégories aux données.

• **Ajustement de courbe :** trace les données le long d'une courbe pour étudier les relations des variables au sein des données.

• **Analyse descriptive :** identifie les modèles dans les données de séries chronologiques, comme les tendances, les cycles ou les variations saisonnières.

• **Analyse explicative :** tente de comprendre les données et les relations qui les composent, ainsi que la cause et l'effet.

•**Analyse exploratoire :** met en évidence les principales caractéristiques des données de séries chronologiques, généralement dans un format visuel.

•**Prévision :** prédit les données futures. Ce type d'analyse est basé sur des tendances historiques. Il utilise les données historiques comme modèle pour les données

futures, prédisant les scénarios qui pourraient se produire le long des futurs points de tracé.

•**Analyse d'intervention :** étudie comment un événement peut modifier les données.

•**Segmentation :** divise les données en segments pour afficher les propriétés sous-jacentes des informations source.

3.8 Conclusion

Comme on a vue dans les chapitres précédents la modélisation et l'étude de notre travail sont basées sur les bases de données à base des séries chronologiques pour la détection et le diagnostic des maladies cardiovasculaires. Malgré la plupart des travaux existants sont basés sur l'utilisation de fichiers CSV pour effectuer leur analyse et aussi la simplicité de ce type de format mais on ne peut pas assurer une liaison entre fichiers ni de présenter d'une façon structurée avec moins de redondance. L'utilisation de base de données des séries chronologiques s'avèrent nécessaire notamment pour les séries chronologies volumineuses. Dans le chapitre suivant, nous monterons clairement notre contribution qui vise à proposer une méthode d'appariement des données ECG à base de séries chronologiques.

CHAPITRE 4

MODÈLE LSTM POUR L'APPARIEMENT DES SÉRIES CHRONOLOGIES DE DONNÉES ECG

4.1 Introduction

Rappelons que notre travail sert à proposer une approche d'appariement des données ECG pour la détection des anomalies cardiaques.
Dans un premier temps, nous avons utilisé les séries chronologiques pour la modélisation de données ECG. Pour cela, nous avons choisi deux datasets : ECG5000 de données récentes proviennent de site "Time Series Classification" et PTB-XL d'ECG de 12 dérivations.
Dans un second temps, nous avons proposé un modèle LSTM de l'apprentissage profond pour atteindre notre objectif.
Pour valider notre proposition, nous avons utilisé Python comme langage de programmation le plus populaire en sciences de données. Nous avons également utilisé le SGBD des séries chronologiques InfluxDB et le framework Grafana pour stocker et visualiser des données ECG, respectivement.
Le chapitre termine par l'évaluation des performances par les mesures de l'exactitude (Accuracy), la précision (Precision), le rappel (Recall), et la F mesure ainsi que la visualisation des ECG dans Grafana.

4.2 Environnement d'exécution :

4.2.1 Google Colab :

Google Colaboratory (également connu sous le nom de Colab) est un service cloud basé sur Jupyter Notebook pour diffuser l'enseignement et la recherche sur l'apprentissage automatique [1]. Il permet d'écrire et d'exécuter du code en Python, Importation et l'enregistrement des blocs-notes depuis et vers Google Drive.
Google Collab permet également :
•D'améliorer les compétences de codage en langage de programmation Python.
•De développer des applications en Deep Learning en utilisant des bibliothèques Python populaires.

•D'utiliser un environnement de développement (Jupyter Notebook) qui ne nécessite aucune configuration.

4.2.2 Pourquoi les GPU ?

GPU (Graphics Processing Unit) signifie unité de traitement graphique, pour la plupart des approches de Machine Learning ML (apprentissage automatique) ou de Deep Learning DL (apprentissage profond). Les GPU sont essentiels en raison de la quantité de données sur laquelle le programme va opérer [47].
Les exécutions des algorithmes de DL sur un CPU peut prendre des mois ! Mais ces exécutions peuvent être assignés à des GPU pour un calcul plus rapide, mais le problème entendu c'est que les GPU sont trop chers, Colab vient à la rescousse ! Colab fournit un GPU Nvidia Tesla K80 gratuit [47] [9].

4.3 Langage de programmation, Framework et bibliothèques

4.3.1 Python

est un langage de programmation puissant et facile à apprendre, Créé par Guido van Rossum et sorti en 1991. Python est interprété, multi- paradigme et multi-plateformes, il est aussi un langage plus commun et plus populaire pour l'apprentissage automatique grâce à sa flexibilité et aussi parce qu'il a un nombre important de bibliothèques logicielles open source disponible, telles que Pandas, Numpy, Scikit-Learn, TensorFlow, Matplotlib et Keras . . .etc [48] [35].
Les bibliothèques python que nous avons utilisé dans notre travail sont :

4.3.2 TensorFlow :

est un Framework open source développé par des chercheurs de Google pour exécuter l'apprentissage automatique, l'apprentissage en profondeur et d'autres charges de travail d'analyse statistique et prédictive [45].
Nous avons utilisé cette bibliothèque pour effectuer des opérations numériques complexes et plusieurs autres tâches pour modéliser les architectures de Deep Learning. Il peut déployer facilement des calculs sur plusieurs plates-formes comme les CPU, les GPU.

4.3.3 Keras :

est une bibliothèque d'apprentissage en profondeur écrite en Python, s'exécutant sur la plate-forme d'apprentissage automatique TensorFlow. Il a été développé dans le but de permettre une expérimentation rapide. Pouvoir passer de l'idée au résultat le plus rapidement possible est essentiel pour faire de bonnes recherches [8].
Elle a été développée dans le but de permettre des expérimentations rapides. Elle supporte à la fois les réseaux convolutifs (CNN) et les réseaux récurrents (RNN) ainsi que la combinaison des deux méthodes.

4.3.4 NumPy :

est le package fondamental pour le calcul scientifique en Python [51].
La bibliothèque permet d'effectuer des calculs numériques avec Python. Elle introduit une gestion facilitée des tableaux de nombres [32].

4.3.5 Pandas :

est un package Python open source qui est le plus largement utilisé pour la science des données/l'analyse des données et les tâches d'apprentissage automatique. Il est

construit au-dessus d'un autre package nommer Numpy, qui prend en charge les tableaux multidimensionnels [33].
Pandas et Numpy sont utilisés pour la manipulation des données (le chargement, la réorganisation et le traitement des données).

4.3.6 Scikit-Learn :

C'est une bibliothèque en Python qui fournit de nombreux algorithmes d'apprentissage non supervisés et supervisés. Il s'appuie sur certaines technologies [36].
Scikit-Learn nous permet d'expérimenter différentes techniques et algorithmes d'apprentissage automatique et d'analyse de données prédéfinis rapidement et facile a utilisé.

4.3.7 Matplotlib :

C'est une bibliothèque complète pour créer des visualisations statiques, animées et interactives en Python [28].

Concernant le stockage et la visualisation de données sous formes des séries chronologiques, nous utilisons InfluxDB et Grafana, respectivement.

4.3.8 InfluxDB :

InfluxDB est un système de gestion de base de données de séries chronologiques développé par la société InfluxData, Inc. InfluxDB est un logiciel open source avec une communauté importante et dynamique [36].InfluxDB a été conçu à partir de zéro pour être une base de données de séries chronologiques spécialement conçue à cet effet ; c'est-à-dire qu'elle n'a pas été reconvertie en séries chronologiques. Le temps était intégré depuis le début. InfluxDB fait partie d'une plate-forme complète qui prend en charge la collecte, le stockage, la surveillance, la visualisation et l'alerte des données

de séries chronologiques. C'est bien plus qu'une simple base de données de séries chronologiques [21].

4.3.9 Grafana :

est une plateforme open source pour la surveillance, l'analyse et la visualisation des métriques. Conçu par Torkel Ödegaard (qui est toujours en charge de son développement et de sa maintenance) et créé en janvier 2014 [17].

4.4 Modélisation de Datasets utilisés :

Nous avons utilisé deux datasets : le premier ECG5000 de 5000 données ECG d'une seule dérivation et le deuxième PTB-XL qui est très complexe mais proche de données issues de l'électrocardiographe de 18885 patients où chaque patient est décrit par 12 dérivations dont chacune est composée de 1000 données ECG. Ces deux datasets vont être modélisés par les séries chronologiques selon le degré de complexité de données.

4.4.1 ECG5000 :

Le dataset ECG5000 provient de "Time Series Classification" [14]. Le jeu de données contient 5 000 exemples de séries temporelles (obtenues par ECG) avec 140 pas de temps. Chaque séquence correspond à un seul battement de cœur d'un seul patient souffrant d'insuffisance cardiaque congestive.
Le dataset se compose de cinq classes : 1. Normal (N)
2. Contraction ventriculaire prématurée R-on-T (PVC R-on-T)
3. Contraction ventriculaire prématurée (CVP)
4. Battement prématuré ou ectopique supra-ventriculaire (SP ou EB)

5. Battement non classifié (UB)

La figure suivante 4.1 illustre la répartition de données dans chaque classe.

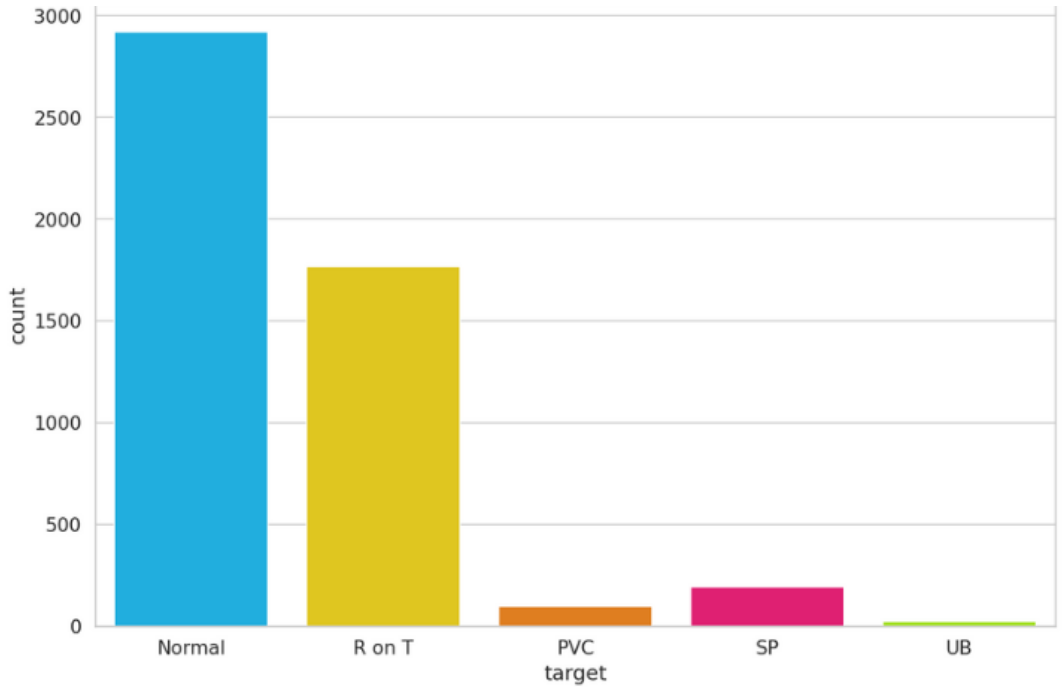


FIGURE 4.1 – Les données dans chaque classe

Chaque donnée du dataset ECG5000 est vue comme une suite des valeurs liées au temps représentant une des 12 dérivations ECG. Pour cette raison, nous utilisons la modélisation basique des séries chronologiques, décrit comme suit :

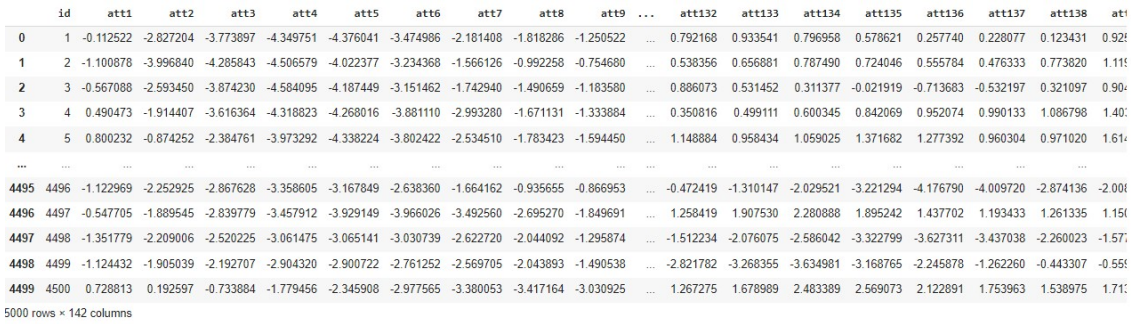
$$SCi = (S1, S2, \ldots Sn) \quad (1)$$

SCi une série chronologique (SC) du patient i avec un longueur n=140 pas de temps.
Si= S(ti) la valeur du tracé ECG à l'instant ti.
Notre dataset est alors représenté par un ensemble des séries chronologiques :

$$ECG5000=(SC1, SC2,\ldots, SCm) \quad (2)$$

Avec m=5000.

La figure suivante 4.2 illustre une partie de dataset ECG5000 :



	id	att1	att2	att3	att4	att5	att6	att7	att8	att9	...	att132	att133	att134	att135	att136	att137	att138	at
0	1	-0.112522	-2.827204	-3.773897	-4.349751	-4.376041	-3.474986	-2.181408	-1.818286	-1.250522	...	0.792168	0.933541	0.796958	0.578621	0.257740	0.228077	0.123431	0.92
1	2	-1.100878	-3.996840	-4.285843	-4.506579	-4.022377	-3.234368	-1.566126	-0.992258	-0.754680	...	0.538356	0.656881	0.787490	0.724046	0.555784	0.476333	0.773820	1.11
2	3	-0.567088	-2.593450	-3.874230	-4.584095	-4.187449	-3.151462	-1.742940	-1.490659	-1.183580	...	0.886073	0.531452	0.311377	-0.021919	-0.713683	-0.532197	0.321097	0.90
3	4	0.490473	-1.914407	-3.616364	-4.318823	-4.268016	-3.881110	-2.993280	-1.671131	-1.333884	...	0.350816	0.499111	0.600345	0.842069	0.952074	0.990133	1.086798	1.40
4	5	0.800232	-0.874252	-2.384761	-3.973292	-4.338224	-3.802422	-2.534510	-1.783423	-1.594450	...	1.148884	0.958434	1.059025	1.371682	1.277392	0.960304	0.971020	1.61
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
4495	4496	-1.122969	-2.252925	-2.867628	-3.358605	-3.167849	-2.638360	-1.664162	-0.935655	-0.866953	...	-0.472419	-1.310147	-2.029521	-3.221294	-4.176790	-4.009720	-2.874136	-2.00
4496	4497	-0.547705	-1.889545	-2.839779	-3.457912	-3.929149	-3.966026	-3.492560	-2.695270	-1.849691	...	1.258419	1.907530	2.280888	1.895242	1.437702	1.193433	1.261335	1.15
4497	4498	-1.351779	-2.209006	-2.520225	-3.061475	-3.065141	-3.030739	-2.622720	-2.044092	-1.295874	...	-1.512234	-2.076075	-2.586042	-3.322799	-3.627311	-3.437038	-2.260023	-1.57
4498	4499	-1.124432	-1.905039	-2.192707	-2.904320	-2.900722	-2.761252	-2.569705	-2.043893	-1.490538	...	-2.821782	-3.268355	-3.634981	-3.168765	-2.245878	-1.262260	-0.443307	-0.55
4499	4500	0.728813	0.192597	-0.733884	-1.779456	-2.345908	-2.977565	-3.380053	-3.417164	-3.030925	...	1.267275	1.678989	2.483389	2.569073	2.122891	1.753963	1.538975	1.71

5000 rows × 142 columns

FIGURE 4.2 – Une partie de Dataset ECG5000

4.4.2 PTB-XL :

Le dataset PTB-XL (Physikalisch-Technische Bundesanstalt en allemand (la fédération physio-technique)) provient de Physionet (PhysioNet, Research Resource for Complex Physiologic Signs) qui représente une banque de données médicales comprenant des Datasets pour les signaux physiologiques complexes[34].

Le jeu de données ECG PTB-XL est un grand jeu de données comprenant 21837 ECG cliniques à 12 dérivations provenant de 18885 patients et d'une durée de 10 secondes. Les données brutes du signal ont été annotées par deux cardiologues avec 71 déclarations ECG différentes et sont complétées par de riches métadonnées [34].

Le dataset PTB-XL se compose de cinq classes : NORM : ECG normal

CD : trouble de la conduction

IM : infarctus du myocarde

HYP : hypertrophie

STTC : changements ST/T

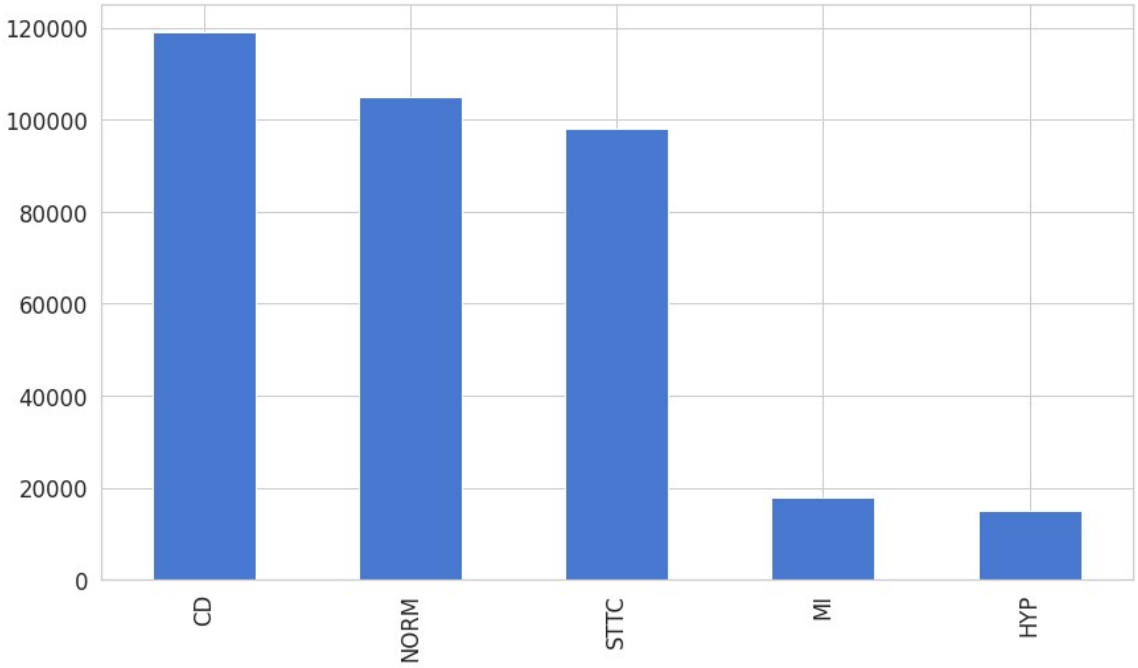


FIGURE 4.3 – Les types de pathologies dans PTB-XL

PTB-XL représente des données ECG des 12 dérivations qui permettent de mesurer fondamentalement toute l'activité cardiaque et de diagnostiquer des anomalies cardiaques sur la paroi antérieure du cœur. Les données PTB-XL sont tellement détaillées, il est difficile de les représenter par une simple modélisation par des séries chronologiques. La figure suivante 4.4 montre la structuration de données PTB-XL :

FIGURE 4.4 – La structure de données PTB-XL

Pour cette raison, il s'avère nécessaire d'utiliser la modélisation complexe des séries chronologiques sous le SGBD InfluxDB. Nous décrivons étape par étape notre modélisation.

Le dataset PTB-XL est représenté sous la forme des mesures (measurements) qui sont

proches des tables relationnelles SQL. Une mesure se décompose de plusieurs séries

```
> use ECG_12LEADS
Using database ECG_12LEADS
> show measurements
name: measurements
name
----
ECG
```

FIGURE 4.5 – Measurements ECG de PTB-XL

chronologiques (series) identifiées par des clés (keys) qui représentent les identifiants des patients.

```
> show series
key
---
ECG,patient-id=PATIENT0011
ECG,patient-id=PATIENT0012
ECG,patient-id=PATIENT0013
ECG,patient-id=PATIENT0014
ECG,patient-id=PATIENT0015
ECG,patient-id=PATIENT0016
ECG,patient-id=PATIENT0017
ECG,patient-id=PATIENT0018
ECG,patient-id=PATIENT0019
```

FIGURE 4.6 – Les séries chronologiques pour chaque patient

Une key est représentée par une collection des couples (field-key, field-value) qui indique les types de dérivations ECG.

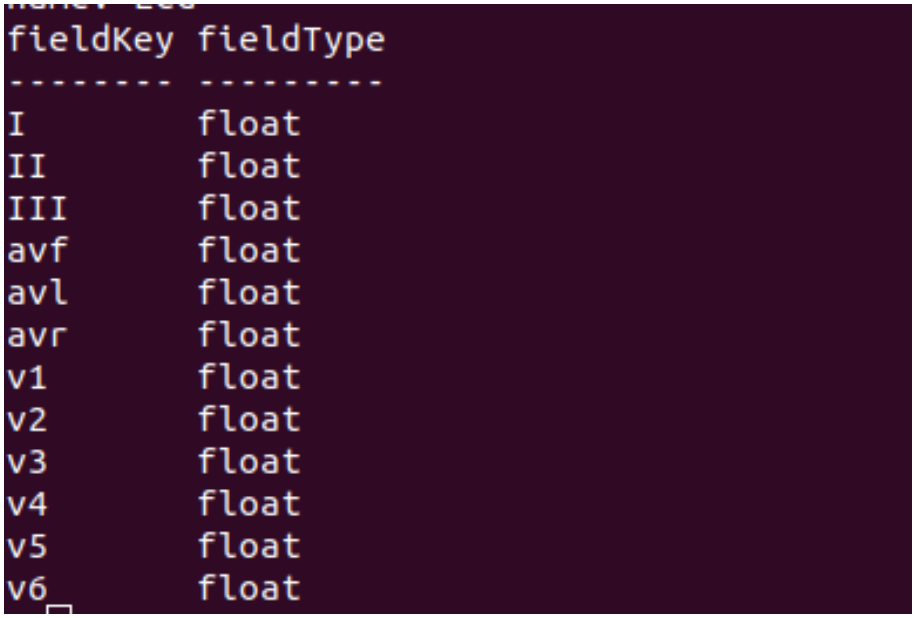


FIGURE 4.7 – Les keys de 12 dérivations de measurements ECG

De plus, chaque couple (field-key, field-value) représente un ensemble des points des séries chronologies sous la forme (field-key, field-value, timestamp). Timestamp est représenté au format yyyy-mm-dd T hh :mm :ssZ

La figure suivante 4.8 illustre un exemple des point d'une partie de série chronologique pour le patient 01 :

Time	I	II	III	aVf	aVl	aVr	V1	V2	V3	V4	V5	V6
0:00.000	-0.244	-0.229	0.015	0.237	-0.13	-0.107	-0.044	-0.12	-0.056	0.106	0.197	0.195
0:00.001	-0.242	-0.234	0.009	0.238	-0.126	-0.113	-0.042	-0.117	-0.051	0.11	0.202	0.198
0:00.002	-0.241	-0.234	0.007	0.238	-0.124	-0.114	-0.044	-0.118	-0.053	0.11	0.201	0.197

FIGURE 4.8 – Un exemple des point d'une partie de ECG pour le premier patient

4.5 Prétraitement des datasets :

4.5.1 Données ECG5000 :

Le dataset ECG5000 est dès le départ divisé en deux parties distinctes au format arff (Attribute-Relation File Format de l'outil data mining Weka) :

• TRAIN5000.arff : permet d'entraîner le modèle.

• TEST5000.arff : permet de tester le modèle sur des données qu'il n'a jamais vues pendant la phase d'apprentissage.

Ces deux fichiers sont de taille identique (division 50%). Pour cela, nous avons établi les opérations suivantes :

- Convertir les fichiers au format CSV. Le résultat sont deux fichiers TRAIN5000.csv et TEST5000.csv.

- Fusionner les deux fichiers pour diviser en 75% pour le training set et 25% pour le testing set.

- Nettoyage et normalisation de données.

Le nettoyage de ces deux sous-ensembles sert à détecter et à éliminer les données aberrantes, les données manquantes et les données dupliquées.

De plus, les cinq classes de datasets sont déséquilibrées (voir la figure 4.1). Pour cette raison, nous avons préparé les données de telle sorte qu'on distingue deux classes bien équilibrées : Classe 'Normal' et Abnormal pour représenter les ECG normaux et anormaux, respectivement.

La figure ci-dessous 4.9 montre la nouvelle division de notre dataset
Avant de commencer l'apprentissage, nous avons normalisé les données d'entrée, car cette étape a un impact sur la construction du modèle, et l'apprentissage du modèle converge rapidement. La normalisation permet de rendre les données toutes dans la même plage qui est généralement entre 0 et 1. Cela permet d'avoir moins souvent des gradients non nuls lors de l'entraînement et, par conséquent, les neurones de notre réseau apprendront plus rapidement.

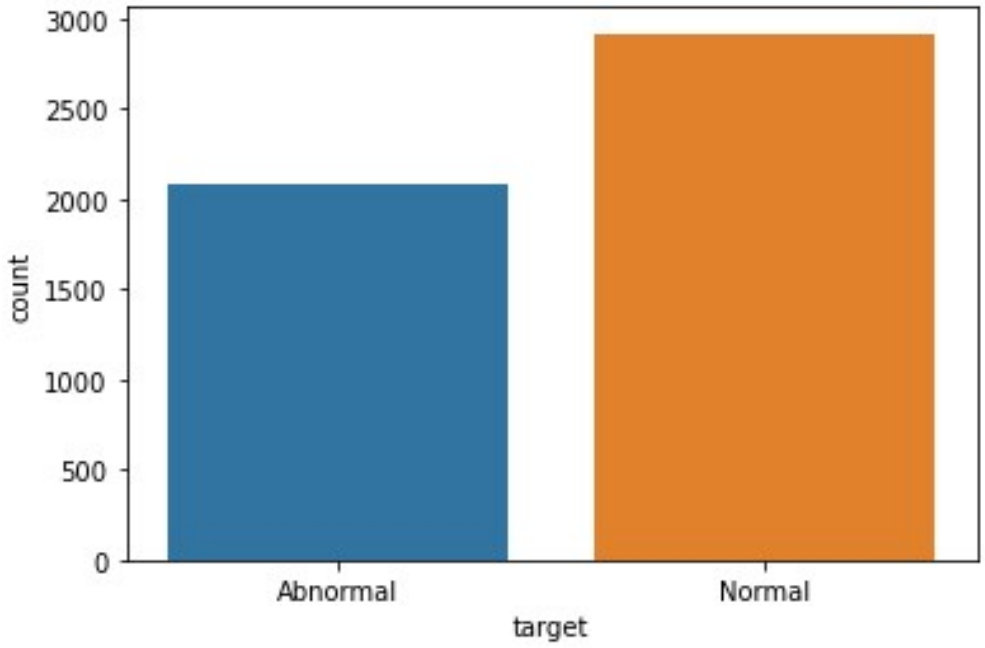


FIGURE 4.9 – Les deux classes principales de dataset ECG5000

De plus, nous avons divisé les données d'entraînement en deux : des données pour l'apprentissage et des données pour la validation de modèle.

4.5.2 Données PTB-XL :

Comme mentionné précédemment, le dataset PTB-XL contient des données complexes de 18885 patients. La première opération de prétraitement consiste à réduire la taille de données à 10 patients où chaque patient ayant 12 séries de longueur 1000.
La réduction de dimension traite d'autres problèmes de données comme les valeurs aberrantes, les valeurs manquantes, et les valeurs redondantes. Le problème de déséquilibrage des classes a été traité sur les données réduites, ce qui nous a donné quatre classes : ECG NORMAL, Perturbation de la conduction, Changement ST/T, et Infarctus du myocarde

Ensuite, à partir des datasets prétraités, nous avons proposé un modèle de réseau de neurones récurrents (Recurrent Neural Network RNN) à base de mémoire longue

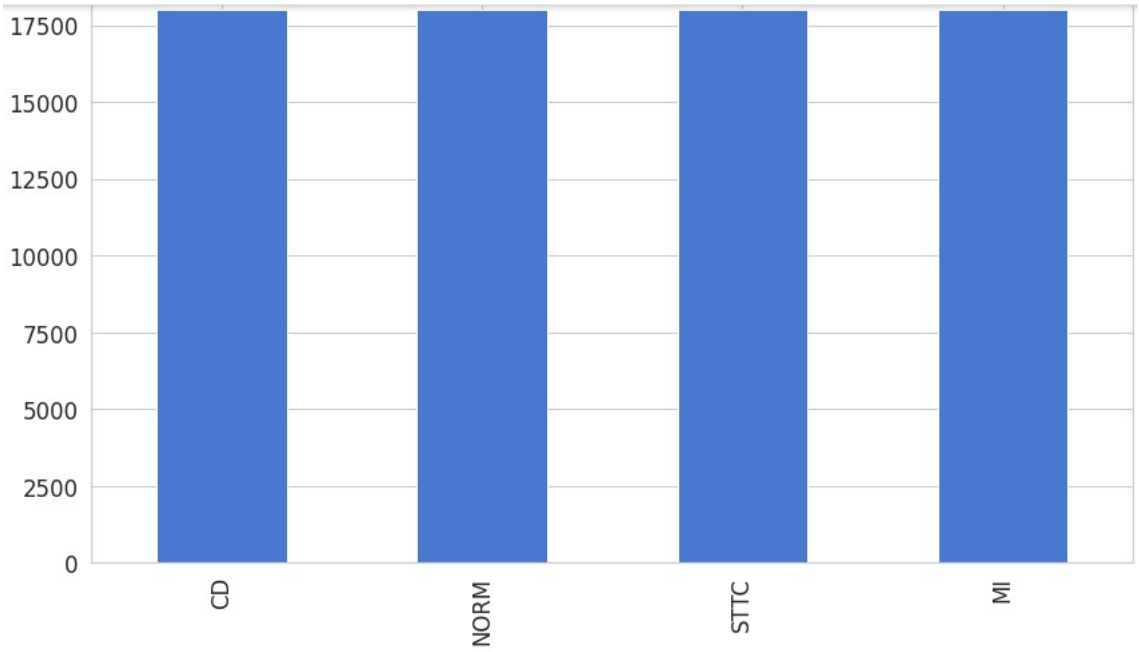


FIGURE 4.10 – Les types des pathologies dans PTB-XL equilibrer

et courte durée (LSTM Long short-term memory) qui est conçu pour modéliser des données séquentielles. En fait, le modèle proposé a été appliqué au dataset ECG5000.

Nous avons essayé d'appliquer le même modèle sur les données complexes de PTB-XL mais nous avons observé que plus les données sont complexes, plus la classification est difficile.

Par conséquent, il s'avère nécessaire d'utiliser le PTB-XL pour la visualisation profonde de données en utilisant le framework Grafana.

4.6 Architecture de notre modèle LSTM :

L'architecture de modèle proposé est illustrée dans la figure suivante 4.11 :
Les paramètres de notre modèle sont les suivants :

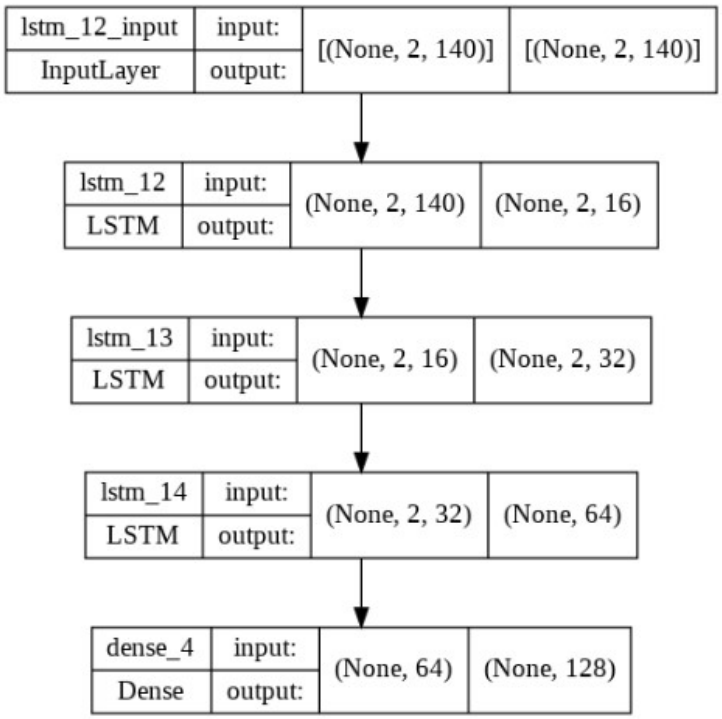


FIGURE 4.11 – l'architecture de modèle LSTM

• La couche d'entrée à la même dimension (nombre de neurones) que le nombre de caractéristiques (Features) dans le vecteur d'entrée.

• La fonction d'activation utilisée était ReLU.

• La couche de sortie à la même dimension que le nombre de classes.

• On est utilisé la technique dropout lorsqu'on tombe au problème de sur-apprentissage (Overfitting) pour obtenir un modèle généralisable.

• La fonction de perte (Loss function) sélectionnée était binary-cross-entropy.

• L'optimiseur " Adam " a été utilisé avec un taux d'apprentissage (Learning Rate) de 0.001.

4.7 Résultats et Discussion de modèle proposé :

Dans notre expérimentation les données d'entrainement ont été divisé sur deux : 60% pour l'apprentissage et 15% our la validation et 25% pour le test. Nous avons implémenté le modèle de Deep Learning RNN-LSTM en utilisant les données d'entraînement de dataset ECG5000 prétraité.

Les données de sous ensemble de validation ont été utilisées pour ajuster les hyper paramètres de modèle.

Nous avons établi plusieurs exécutions pour obtenir les meilleures hyperparamètres pour notre modèle. Ces paramètres ne peuvent pas être ajustés pendant la phase d'entraînement, mais ils ont un impact important sur les performances du modèle pendant l'entraînement. Ils comprennent les variables qui déterminent la structure du réseau (le nombre de neurones, nombre de couches, fonction d'activation, . . .), le lot d'échantillons (Batch Size) et le nombre d'itérations . . .etc.

Nous montrons dans la figure suivante 4.12 l'exactitude et la perte de modèle proposé par rapport aux époques d'apprentissage et de validation de notre modèle. Lorsqu'on arrive à un bon modèle avec le minimum de taux d'erreur et le maximum

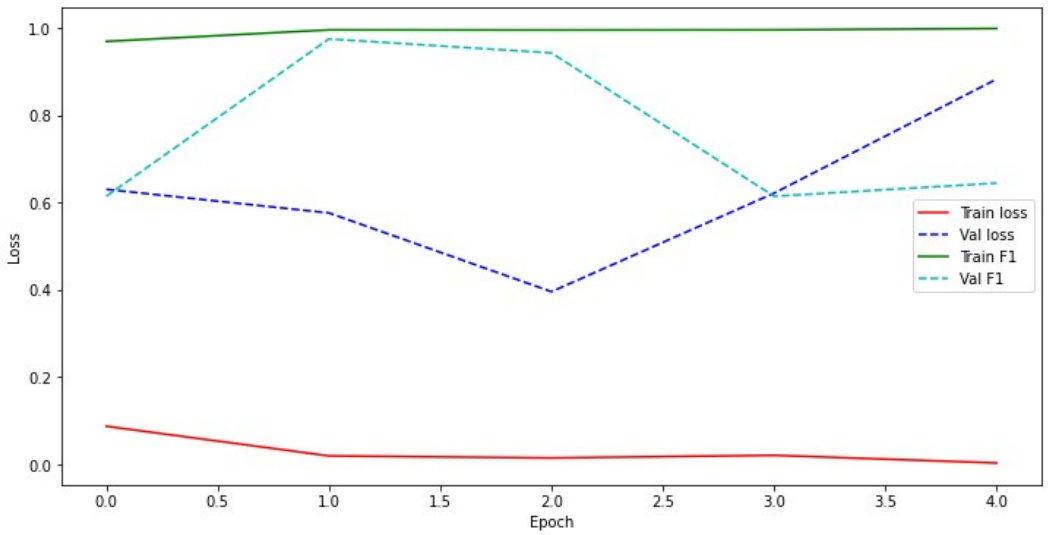


FIGURE 4.12 – l'exactitude et la pert de modèle LSTM

d'exactitude, nous effectuons ensuite un test de notre modèle finalisé sur le sous-ensemble de test.

L'apprentissage ne prend pas beaucoup de temps, le modèle a été formé des 15 époques. Il a obtenu une exactitude très bonne 96%.

Nous notons ici que notre modèle converge vers une valeur de perte minimale. Cela signifie que ce modèle apprenne mieux et effectue des meilleures prédictions après chaque époque d'optimisation pour les séries chronologiques.

Pour évaluer l'efficacité du modèle proposé, nous utilisons les mesures des performances suivantes [29] :

Précision : La précision et le pourcentage de détections correctes. IL met en évidence l'éxactitude des prédictions. Il se calcule parce ratio :

$$Prcision = \frac{vrai\ positif}{vrai\ positif + faux\ positif}$$

Rappel : le rappel est un indicateur qui mesure la capacité du modèle à prédire l'ensemble des résultats attendus. Calculer par :

$$Rappel = \frac{vrai\ positif}{vrai\ positif + faux\ positif}$$

Accuracy : il indique le pourcentage de bonnes prédictions :

$$Accuracy = \frac{vrai\ positif + vrai\ ngatif}{total}$$

F-measure : c'est la moyenne harmonique de la précision et du rappel. Calculer par :

$$F - measure = \frac{2}{recall^{-1} + precision^{-1}}$$

Les résultats sont présentés dans la figure suivante 4.13 :

	precision	recall	f1-score	support
Normal	1.00	0.91	0.95	516
	0.94	1.00	0.97	734
accuracy			0.96	1250
macro avg	0.97	0.95	0.96	1250
weighted avg	0.96	0.96	0.96	1250

FIGURE 4.13 – le rapport de classification des données

4.8 Visualisation profonde de données ECG modélisées par les séries chronologiques :

L'appariement automatique des données ECG modélisées par les séries chronologiques via le modèle RNN LSTM est utile pour la détection de la présence ou l'absence des pathologies cardiaques. Par ailleurs, la représentation visuelle de tracés ECG par des outils avancés améliore notablement l'interprétation et la compréhension des données ECG.

Dans ce contexte, pour représenter les données complexe d'ECG de 12 dérivations du dataset PTB-XL, nous avons utilisé le Framework Grafana le plus performant pour la visualisation profonde des séries chronologiques [26].

Pour une visualisation profonde de données ECG, la figure suivante 4.14 illustre une des fonctionnalités de Grafana.

FIGURE 4.14 – Les séries chronologiques des dérivations d'un seul patient

Après avoir visualisé les données ECG à 12 dérivations avec Grafana,le médecin maître de stage Dr Touaimia ABDERAHMANE les a interprétées pour s'assurer que nos résultats étaient corrects. Il l'a trouvé correct et peut être adopté pour faciliter la lecture de l'ECG à l'hôpital, au niveau du service d'urgences. Nous allons faire un article en commun pour développer notre travail.

4.9 Conclusion :

Nous avons mis en œuvre le modèle RNN LSTM d'appariement de données ECG modélisées par les séries chronologiques. Ce modèle ce compose de 3 Couches cachés. La définition de modèle RNN LSTM est basé sur l'utilisation de dataset ECG5000. Le modèle a été évalué et les résultats sont très satisfaisants. Il est caractérisé par une précision, un rappel et f score moyens qui sont identiques et vaut 96%. Nous avons également modélisé les données ECG complexe de 12 dérivations de dataset PTB-XL en utilisant InfluxDB. Pendant notre expérimentation, nous avons remarqué qu'il est difficile d'utiliser le même modèle RNN LSTM pour effectuer l'appariement sur ce

type complexe de données. En effet, nous avons utilisé un autre moyen d'interprétation de données complexes. Il s'agit de visualisation de données via le Framework Grafana.

CONCLUSION GÉNÉRALE

La classification des maladies cardiovasculaires par les techniques de l'intelligence artificielle et de l'apprentissage profond est aujourd'hui considérée comme une chose indispensable en raison des énormes développements technologiques dans le domaine de la médecine et de la disponibilité des moyens et équipements nécessaires. En raison de nombreux défis scientifiques, la plupart des chercheurs tentent d'imposer leurs recherches sur le terrain.

Dans notre recherche, nous sommes censés de comprendre les données médicales ECG. Pour cela, nous avons effectué un stage au sein de l'hôpital Hakim Okbi-Guelma au niveau du service d'urgences avec le médecin maître de stage Dr Touaimia ABDERAHMANE.

Dans un second temps, nous avons étudié et analysé les les travaux liés à la classification des données ECG, afin d'avoir une meilleure vue et d'obtenir de meilleurs résultats dans ce domaine. La nouveauté de notre travail par rapport les travaux existants est qu'il permet de faire l'appariement des données ECG à base des séries chronologiques en utilisant un modèle efficace de Deep Learning.

Nous avons commencé par le choix de l'ensemble de données où nous avons choisi de travailler avec le dataset PTB-XL provient de PhysioNet qui représentent les données de 12 dérivations. Ce dataset représente des données complexes multi-niveaux :

CONCLUSION GÉNÉRALE

le métadata des patients (Niveau 1) où chaque patient doit avoir 12 séries chronologiques (Niveau 1.1) présentant les 12 dérivations où chaque série a une longueur de 1000 valeurs (Niveau 1.1.1).
Pendant le prétraitement de ces données, nous avons rencontré des difficultés sur les niveaux de données qui rendent la classification très difficile et nécessite vraiment un autre travail pour préparer ces données.
Vue qu'on a modélisé ce type de données complexes à travers la structure des séries chronologiques de système InfluxDB, nous avons orienté notre vision vers un autre axe de recherche qui est la visualisation de données. A cet effet, nous avons utilisé le Framework Grafana qui offre une visualisation profonde de données complexes.
Dans le second temps, nous avons utilisé un autre dataset d'ECG appelé ECG5000 pour l'appariement de données ECG et donc la détection des anomalies cardiaques.
Ensuite, nous avons implémenté le modèle discriminatoire de Deep Learning (apprentissage supervisé) un réseau neuronal récurrent (RNN) avec une mémoire a long et court terme LSTM permettant de maintenir un état aussi longtemps que nécessaire.
Les résultats obtenus sont très satisfaisants, nous sommes arrivés à un bon modèle avec le minimum de taux d'erreur et une exactitude très bonne 96%.

A partir de ce travail, plusieurs perspectives ont été envisagés :

CONCLUSION GÉNÉRALE

- Traitement de données complexes de dataset PTB-XL pour être faciles à utiliser à des fins d'apprentissage automatique.

- Utilisation de dataset PTB-XL traité sur le modèle proposé LSTM pour l'appariement des données ECG.

- Améliorer si nécessaire le modèle LSTM pour l'appariement de données complexes modélisées par les séries chronologiques.

- Utilisation de modèle dans un système de telemonitoring muni des capteurs d'ECG pour un projet basé sur les outils IoT.

BIBLIOGRAPHIE

[1] Francisco Regis Vieira ALVES et Renata Passos Machado VIEIRA. « The Newton fractal's Leonardo sequence study with the Google Colab ». In : *International Electronic Journal of Mathematics Education* 15.2 (2019), em0575.

[2] Ouadi BEYA. « Analyse et reconnaissance de signaux vibratoires : contribution au traitement et à l'analyse de signaux cardiaques pour la télémédecine ». Thèse de doct. Dijon, 2014.

[3] Giuseppe BONACCORSO. *Machine learning algorithms*. Packt Publishing Ltd, 2017.

[4] CARDIO. `https://www.tvcjdc.be/nl/article/23306019/cardiologue`. (accéder le 26/01/2022).

[5] CARDIOLOGIE. `http://eprints.univ-batna2.dz/1607/1/Main-th`. (accéder le 22/03/2022).

[6] CARDIOLOGIE. `https://www.chuv.ch/fr/cardiologie/car-home/patients-et-famille/fonctionnement-du-coeur`. (accéder le 27/01/2022).

[7] CARDIOVASCULAIRES. `https://www.msdmanuals.com/fr/professional/troubles-cardiovasculaires/tests-et-procedures-cardiovasculaires/electrocardiographie`. (accéder le 30/01/2022).

[8] others. (2015). Keras. GitHub. Retrieved CHOLLET F. https://github.com/fchollet/keras. (accéder le 26/05/2022).

[9] COLABORATORY. https://research.google.com/colaboratory/faq.html. (accéder le 25/05/2022).

[10] Khouloud DAHMANE. « Analyse d'images par méthode de Deep Learning appliquée au contexte routier en conditions météorologiques dégradées ». Thèse de doct. Université Clermont Auvergne, 2020.

[11] Time Series DATA et ANALYSIS. https://www.aptech.com/blog/introduction-to-the-fundamentals-of-time-series-data-and-analysis/. (accéder le 22/03/2022).

[12] DIFFERENCE.BETWEEN.CNN.RNN. https://www.telusinternational.com/articles/difference-between-cnn-and-rnn.

[13] Zahra EBRAHIMI et al. « A review on deep learning methods for ECG arrhythmia classification ». In : *Expert Systems with Applications : X* 7 (2020), p. 100033.

[14] ECG5000. https://timeseriesclassification.com/description.php?Dataset=ECG5000. (accéder le 24/05/2022).

[15] Ardalan Sharifzadehgan ELOI MARIJON. *comprendre l'ECG*. Elsevier Masson SAS, 2020.

[16] Ian GOODFELLOW, Yoshua BENGIO et Aaron COURVILLE. *Deep learning*. MIT press, 2016.

[17] GRAFANA. https://pandorafms.com/blog/what-is-grafana/. (accéder le 21/05/2022).

[18] india-java-user GROUP. https://wearecommunity.io/communities/india-java-user-group/articles/891. (accéder le 26/03/2022).

[19] Sepp HOCHREITER et Jürgen SCHMIDHUBER. « Long short-term memory ». In : *Neural computation* 9.8 (1997), p. 1735-1780.

[20] Shenda HONG et al. « Opportunities and challenges of deep learning methods for electrocardiogram data : A systematic review ». In : *Computers in Biology and Medicine* 122 (2020), p. 103801.

[21] INFLUXDATA. `https://www.influxdata.com/time-series-database/`. (accéder le 26/01/2022).

[22] Søren Kejser JENSEN, Torben Bach PEDERSEN et Christian THOMSEN. « Time series management systems : A survey ». In : *IEEE Transactions on Knowledge and Data Engineering* 29.11 (2017), p. 2581-2600.

[23] Tae Joon JUN et al. « ECG arrhythmia classification using a 2-D convolutional neural network ». In : *arXiv preprint arXiv :1804.06812* (2018).

[24] Eamonn KEOGH et al. « Segmenting time series : A survey and novel approach ». In : *Data mining in time series databases*. World Scientific, 2004, p. 1-21.

[25] Grzegorz KŁOSOWSKI et al. « The use of time-frequency moments as inputs of lstm network for ecg signal classification ». In : *Electronics* 9.9 (2020), p. 1452.

[26] Kyriakos KRITIKOS et Paweł SKRZYPEK. « A review of serverless frameworks ». In : *2018 IEEE/ACM International Conference on Utility and Cloud Computing Companion (UCC Companion)*. IEEE. 2018, p. 161-168.

[27] time-series-classification-with-deep LEARNING. `https://towardsdatascience.com/time-series-classification-with-deep-learning-d238f0147d6f`. (accéder le 16/02/2022).

[28] MATPLOTLIB. `https://matplotlib.org/`. (accéder le 25/05/2022).

[29] Jiaju MIAO et Wei ZHU. « Precision–recall curve (PRC) classification trees ». In : *Evolutionary intelligence* (2021), p. 1-25.

[30] Sajad MOUSAVI et Fatemeh AFGHAH. « Inter-and intra-patient ecg heartbeat classification for arrhythmia detection : a sequence to sequence deep learning approach ». In : *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2019, p. 1308-1312.

[31] Syeda Noor Zehra NAQVI, Sofia YFANTIDOU et Esteban ZIMÁNYI. « Time series databases and influxdb ». In : *Studienarbeit, Université Libre de Bruxelles* 12 (2017).

[32] apprendre NUMPY. https://courspython.com/apprendre-numpy.html. (accéder le 26/05/2022).

[33] what-is PANDAS. https://www.activestate.com/resources/quick-reads/what-is-pandas-in-python-everything-you-need-to-know/. (accéder le 26/05/2022).

[34] PYSIONET. https://physionet.org/content/ptb-xl/1.0.1/. (accéder le 20/05/2022).

[35] PYTHON. https://www.python.org. (accéder le 25/05/2022).

[36] QUINFLUXDB. https://www.ionos.fr/digitalguide/hebergement/aspects-techniques/quest-ce-quinfluxdb/. (accéder le 21/05/2022).

[37] RESEARCHGATE. https://www.researchgate.net/publication/341639694_How_to_Build_a_Graph-Based_Deep_Learning_Architecture_in_Traffic_Domain_A_Survey. (accéder le 20/02/2022).

[38] Antonio H RIBEIRO et al. « Automatic diagnosis of the 12-lead ECG using a deep neural network ». In : *Nature communications* 11.1 (2020), p. 1-9.

[39] MEBROUKI Mahmoud SEKKIL HICHAM MOHAMED. « Etude comparative entre les différentes architectures des réseaux de neurones convolutifs (CNNs) pour la détection de la fatigue du conducteu ». In : (2021).

[40] Jean SENDE. *Guide pratique de l'ECG*. De Boeck Secundair, 2003.

[41] analysis-time SERIES. https://itfeature.com/time-series-analysis-and-forecasting/components-of-time-series. (accéder le 26/03/2022).

[42] deep-learning-time SERIES. https://towardsdatascience.com/time-series-classification-with-deep-learning-d238f0147d6f. (accéder le 05/04/2022).

[43] Types Techniques TABLEAU SOFTWARE . (n.d.). Time Series Analysis : Definition et When It's USED. https://www.tableau.com/learn/articles/time-series-analysis/. (accéder le 22/03/2022).

[44] Pierre TABOULET. *l'ECG de A a Z*. MALOINE, 2009.

[45] TENSOFLOW. https://www.techtarget.com/searchdatamanagement/definition/TensorFlow. (accéder le 26/05/2022).

[46] Toni TOHARUDIN et al. « Employing long short-term memory and Facebook prophet model in air temperature forecasting ». In : *Communications in Statistics-Simulation and Computation* (2020), p. 1-24.

[47] deep-learning TURKEY. https://medium.com/deep-learning-turkey/google-colab-free-gpu-tutorial-e113627b9f5d. (accéder le 26/02/2022).

[48] Guido VAN ROSSUM et Fred L DRAKE. *An introduction to Python*. Network Theory Ltd. Bristol, 2003.

[49] Ludi WANG et Xiaoguang ZHOU. « Detection of congestive heart failure based on LSTM-based deep network via short-term RR intervals ». In : *Sensors* 19.7 (2019), p. 1502.

[50] Kuba WEIMANN et Tim OF CONRAD. « Transfer learning for ECG classification ». In : *Scientific reports* 11.1 (2021), p. 1-12.

[51] WHATISNUMPY. https://numpy.org/doc/stable/user/whatisnumpy.html.

[52] Qihang YAO et al. « Multi-class arrhythmia detection from 12-lead varied-length ECG using attention-based time-incremental convolutional neural network ». In : *Information Fusion* 53 (2020), p. 174-182.