N° d'ordre :

الجمهورية الجزائرية الديمقراطية الشعبية

République Algérienne Démocratique et Populaire

وزارة التعليم العالي والبحث العلمي

Ministère de L'Enseignement Supérieur et de la Recherche Scientifique

جامعة عين تموشنت بلحاج بوشعيب

Université Ain-Temouchent Belhadj BOUCHAIB

Faculté : Sciences et technologie
Département : Électronique et Télécommunications
Laboratoire : Structures intelligentes

THESE

Présentée pour l'obtention du **diplôme** de **DOCTORAT**
Domaine : Sciences et Technologies
Filière : Télécommunications
Spécialité : Réseaux de Télécommunications

Par : KORTI Djazila Souhila

Intitulé

Réseaux de capteurs UWB pour la reconnaissance d'activités humaines basée sur l'IoT et l'intelligence artficielle

Soutenue publiquement, le 14/07/2024, devant le jury composé de :

Nom & Prénom(s)	Grade	Qualité	Etablissement de rattachement
M. MERADI Abdelhafid	MCA	Président	Université de Ain-Témouchent
Mme. SLIMANE Zohra	MCA	Directrice de thèse	Université de Tlemcen
M. MERZOUGUI Rachid	Pr	Examinateur	Université de Tlemcen
M. ZERROUKI Hadj	MCA	Examinateur	Université de Sidi Bel Abbès
Mme. MOULESSEHOUL Wassila	MCA	Examinatrice	Université de Ain-Témouchent

Année universitaire : 2023-2024

Résumé

Les soins post-Accident Vasculaire Cérébral (AVC) impliquent une thérapie physique intensive sur plusieurs mois, créant une charge financière importante qui pèse sur les patients et la société, ainsi que de la grande dépendance à l'égard des professionnels et des établissements médicaux. La rééducation à domicile est une approche prometteuse et rentable, permettant un suivi à long terme avec un encadrement minimal. L'environnement intelligent constitue une solution encourageante pour répondre à cette préoccupation, permettant d'assister les patients depuis chez eux grâce à l'intégration de capteurs et d'intelligence artificielle. Au cœur de cette avancée se trouve la reconnaissance d'activités, une tâche complexe visant à identifier les gestes et actions entreprises par les patients. La difficulté de ce défi est liée à la nature des activités à reconnaître, des capteurs disponibles, et de l'incidence sur la confidentialité. Parmi les choix de capteurs pour la reconnaissance d'activités, les radars Ultra-Wideband (IR-UWB) qui ont suscité un vif intérêt, capables d'offrir un équilibre favorable entre précision et préservation de la vie privée. Les travaux couverts par cette thèse visent à mettre en place un système de reconnaissance d'activités qui englobe les Gestes de la Main (GM) et les Actions Humaines (AH) pour assister les patients post-AVC dans leur thérapie à domicile. Initialement, un modèle d'apprentissage profond a été proposé, combinant un réseau de neurones convolutif (Convolutional Neural Network : CNN) et un réseau de mémoire à long terme (Long Short-Term Memory : LSTM) pour la reconnaissance d'actions et de gestes de la main à partir des données IR-UWB. Ce modèle présente l'avantage d'une architecture parallèle et hybride, en maintenant une structure optimisée avec un nombre minimal de paramètres entraînables. Par la suite, nous avons proposé un nouveau modèle de classification conçu spécifiquement pour les systèmes multi-capteurs, appelé Multi-Input Multi-Output Convolutional Extra Trees (MIMO-CxT). Ce modèle offre l'avantage de fusionner des données provenant de diverses sources avec un prétraitement minimale. Il se distingue par ses performances élevées, renforçant la précision, la fiabilité et la robustesse des systèmes de reconnaissance.

Mots clés : radar ultra large bande, reconnaissance de gestes de la main, reconnaissance d'actions humaines, apprentissage profond, modèle hybride.

Abstract

Post-stroke care involves intensive physical therapy over several months, creating a significant financial burden on both patients and society, as well as a heavy reliance on professionals and medical facilities. Home rehabilitation is a promising and cost-effective approach, allowing long-term monitoring with minimal supervision. The smart environment provides an encouraging solution to address this concern by assisting patients at home through the integration of sensors and artificial intelligence. At the core of this advancement is activity recognition, a complex task aimed at identifying gestures and actions performed by patients. The challenge is linked to the nature of activities to be recognized, available sensors, and the impact on privacy. Among the sensor choices for human activity recognition, Impulse Radio Ultra-Wideband (IR-UWB) have garnered significant interest, capable of striking a favorable balance between precision and privacy preservation. The research covered in this thesis aims to establish an activity recognition system that encompasses Hand Gestures (HG) and Human Actions (HA) to assist post-stroke patients in their home therapy. Initially, a deep learning model was proposed, combining a Convolutional Neural Network (CNN) and a Long Short-Term Memory (LSTM) network for recognizing HG and HA from IR-UWB data. This model has the advantage of a parallel and hybrid architecture, maintaining an optimized structure with a minimal number of trainable parameters. Subsequently, we proposed a new classification model specifically designed for multi-sensor systems, called Multi-Input Multi-Output Convolutional Extra Trees (MIMO-CxT). This model excels in merging data from various sources with minimal preprocessing, standing out for its high performance and reinforcing the accuracy, reliability, and robustness of recognition systems.

Keywords : ultra-wideband radar, hand gesture recognition, human action recognition, deep learning, hybrid model

ملخص

الرعاية بعد السكتة الدماغية تتطلب علاجًا فيزيائيًا مكثفًا على مدى عدة أشهر، مما يؤدي إلى تكلفة مالية كبيرة على المرضى والمجتمع، وإلى اعتماد كبير على المحترفين والمؤسسات الطبية. إعادة التأهيل في المنزل تمثل نهجًا واعدًا واقتصاديًا، مما يسمح بمتابعة طويلة الأمد مع إشراف أدنى. تقدم البيئة الذكية حلاً واعدًا لمواجهة هذه القضية من خلال مساعدة المرضى في منازلهم باستخدام تكنولوجيا الاستشعار والذكاء الاصطناعي. في قلب هذه التقدمات يكمن التركيز على التعرف على الأنشطة، وهي مهمة معقدة تهدف إلى تحديد حركات وأفعال المرضى. تترتب صعوبة هذا التحدي على طبيعة الأنشطة التي يجب تعرفها، وعلى الأجهزة المتاحة، وعلى التأثير على الخصوصية. بين اختيارات أجهزة التعرف على الأنشطة، أثارت رادارات فائقة النطاق (IR-UWB) اهتمامًا كبيرًا، حيث يمكنها تحقيق توازن ملائم بين الدقة والحفاظ على الخصوصية. تهدف الأبحاث المشمولة في هذه الرسالة إلى إقامة نظام للتعرف على الأنشطة يشمل أفعال اليد والجسم لمساعدة المرضى بعد السكتة الدماغية في علاجهم في المنزل. في البداية، تم اقتراح نموذج للتعلم العميق يجمع بين شبكة العصب الاصطناعي (CNN) وشبكة الذاكرة الطويلة الأمد (LSTM) للتعرف على حركات اليد وأفعال الإنسان من خلال بيانات الردار فائق النطاق. يتميز هذا النموذج بهيكله المتوازي والهجين، مع الحفاظ على هيكل محسن بعدد أدنى من المعلمات التي يمكن تدريبها. في وقت لاحق، قدمنا نموذج تصنيف جديد مصمم خصيصًا لأنظمة متعددة الاستشعار (MIMO-CxT) . يقدم هذا النموذج فوائد دمج البيانات من مصادر متنوعة مع معالجة أدنى، ويتميز بأداء متفوق يعزز الدقة والموثوقية والصلابة في أنظمة التعرف.

الكلمات المفتاحية : رادار فائق النطاق، التعرف على حركات اليد، التعرف على حركات الجسم، التعلم العميق، النموذج الهجين.

Remerciements

Avant tout, je remercie ALLAH, mon Créateur, qui m'a accordé la patience et le courage nécessaires pour réaliser ce travail.

Je souhaite exprimer ma sincère gratitude envers Madame SLIMANE Zohra, ma directrice de thèse, pour sa présence constante et son soutien inestimable. Ses conseils éclairés ont grandement contribué à l'avancement de mes recherches.

Je tiens également à remercier chaleureusement Monsieur MERADI Abdelhafid pour avoir présidé le jury de ma thèse avec un intérêt marqué pour mon travail.

Par ailleurs, je souhaite exprimer ma profonde reconnaissance envers les membres du jury, Monsieur MERZOUGUI Rachid, Monsieur ZERROUKI Hadj, et Madame MOULESSEHOUL Wassila, pour l'honneur qu'ils m'ont fait en évaluant mes travaux de thèse. Je les remercie sincèrement pour leur contribution précieuse à la version finale de ce manuscrit.

Enfin, je conclus ces remerciements en exprimant toute ma gratitude envers ma famille, mes parents et mes frères, pour leur soutien indéfectible. Leurs encouragements et leur présence ont été des piliers essentiels tout au long de ce parcours.

Table des matières

Table des figures

Liste des tableaux

Acronymes

2D2D-PCA two-dimensional two-directional principal component analysis
ACGAN Auxiliary Classifier Generative Adversarial Networks
ACP Analyse en Composantes Principales
ADL Analyse Discriminante Linéaire
AEN AutoENcoder
AH Actions Humaines
AVC Accident Vasculaire Cérébrale
AVQ Activités de la Vie Quotidienne
Bi-LSTM BidirectionalLSTM
CAE Convolutional Auto Encoder
CIR Channel Impulse Response
CNN Convolutional Neural Network
CW Continuous Wave
DA Domain Adaptation
DC Double Click
DCGAN Deep Convolution Generative Adversarial Network
DCV Diagramme cadence-vitesse
diag-LR-DU Diagonal-Left Right-Down Up
diag-LR-UD Diagonal-Left Right-Up Down
diag-RL-DU Diagonal-Right Left-Down Up
diag-RL-UD Diagonal-Right Left-Up Down
DL Deep Learning
DT Decision Tree
DU Down-Up
ETs Extra Trees
FCNN Fully Connected Neural Network
FFT Fast Fourier Transform

FMCW Frequency Modulated Continuous Wave
GAN Generative Adversarial Networks
GM Gestes de la Main
GRU Gated Reccurent Unit
IA Intelligence Artificielle
IHM Interaction Homme-Machine
IoT Internet Of Things
IR-UWB Impulse Radio Ultra Wide Band
kNN K-Neirest Neighbors
LNA Low Noise Amplifier
LR Left-Right
LS Left Swipe
LSTM Long Short-Term Memory
MDS Micro-Doppler Signature
MIMO-CxT Multi-Input Multi Output Convolutional Extra Trees
ML Machine Learning
MOCAP Motion Capture
MTI Moving Target Identification
NB NaÏve Bayes
PRF Pulse Repetition Frequency
RAH Reconnaissance des Actions Humaines
RF Random Forest
RGM Reconnaissance des Gestes de la Main
RL Right-Left
RNN Reccurent Neural Network
RS Right Swipe
RT Regression Tree
RVB Rouge Vert Bleu
SC Simple Click
SCGRNN Segmented Convolutional Gated Recurrent Neural Networks
STFT Short Time Fourier Transform
SVM Support Vector Machine
TL Transfert Learning
TRDI Time-varying Range–Doppler Images
UD Up-Down
VGG Visual Geometry Group
WRTFT Weighted Range-Time-Frequency Transform

Introduction générale

Contexte de la recherche

L'Accident Vasculaire Cérébral (AVC) est considéré comme l'un des problèmes de santé publique les plus préoccupants au monde. Il s'agit d'un handicap grave et courant dans le monde entier, avec des conséquences personnelles et sociétales importantes. Selon L'organisation mondiale de la santé, près de 15 millions de personnes chaque année dans le monde sont victimes d'un AVC, soit environ une personne toutes les deux secondes [1]. Les prévisions estimnte qu'en 2050, le nombre de patients de plus de 75 ans victimes d'AVC sera de 75% contre 55% en 2005 [2]. De même, Il est important de souligner que l'AVC ne touche pas que la population âgée, puisque 25% des patients victimes d'AVC ont moins de 65 ans [3]. Grâce à l'amélioration des interventions médicamenteuses précoces et des soins aigus, la mortalité due aux AVCs a progressivement diminué [4]. Cependant, 5 millions des victimes survivants d'AVC représentent des séquelles, qu'elles soient neuro-locomotrices, cognitives ou encore psycho-affectives, et nécessitent des soins importants [1]. Cela se traduit par des coûts importants associés aux traitements et aux soins post-AVC pour les patients, leurs familles et les services de santé [5].

Les complications courantes qui subsistent après un AVC se manifestent généralement par un handicap sévère avec une paralysie transitoire de longue durée des membres inférieurs et supérieurs [6]. Ces déficiences se traduisent généralement par des difficultés à bouger les pieds et à coordonner les bras, les mains et les doigts. Elle affecte jusqu'à 87% des survivants d'un AVC et les empêche d'accomplir 80% des Activités de la Vie Quotidienne (AVQ) [7,8]. La fonction des membres inférieurs et supérieurs après un AVC n'est souvent pas récupérée rapidement et nécessite un traitement de rééducation physique continu et récurrent pendant au moins six mois [9, 10]. La rééducation implique souvent plusieurs interventions différentes et nécessite généralement la coopération du patient, des

soignants et de l'équipe de rééducation.

L'approche conventionnelle implique l'élaboration d'un plan de rééducation et une évaluation périodique par un spécialiste en physiothérapie. La réadaptation s'effectue par des visites au centre de rééducation, où les patients se rendent en moyenne trois fois par semaine et effectuent environ une heure d'exercice. Les séances de rééducation se concentrent généralement sur des exercices et des actions spécifiquement ciblés et répétitifs. Il est suggéré qu'un grand nombre de répétitions de l'ordre de dizaines de milliers est nécessaire pour aider à réapprendre les mouvements et la coordination des membres [11, 12].

Des recherches suggèrent que le temps disponible pour les séances de réadaptation ne permet pas aux patients d'effectuer le nombre de répétitions requis pour une récupération optimale. Une étude a conclu qu'une séance horaire typique ne permettait en moyenne que 32 répétitions [13]. Cela contraste fortement avec les milliers de répétitions nécessaires pour induire une amélioration de la motricité. La nécessité d'augmenter le nombre de répétitions est donc évidente. Cependant une augmentation du nombre de séances ou de la durée de ces dernières est très improbable en particulier avec la charge élevée sur les systèmes de soin de santé ainsi que la limitation des ressources disponibles. De plus, une séance de thérapie tente de maximiser l'implication du patient, mais néanmoins est limitée par les contraintes physiques de ce dernier. Les exercices de rééducation sont difficiles, longs et fatigants. Augmenter le nombre de répétitions de manière significative est n'est donc pas réalisable dans le cadre clinique.

Les limites de la rééducation clinique soulignent donc l'importance d'explorer de nouvelles voies qui permettent aux patients de se rétablir en dehors du cadre hospitalier. À cette fin, la rééducation à domicile est proposée, où les patients accomplissent des tâches de manière indépendante pour améliorer ou maximiser le processus de récupération [14]. Elle consiste en un programme d'exercice conçu par une équipe de thérapeutes et fourni avec les instructions nécessaires. Compte tenu de la charge financière importante qui pèse sur les patients et la société, ainsi que de la grande dépendance à l'égard des professionnels et des établissements médicaux, la thérapie à domicile peut aider les personnes à atteindre leurs objectifs de rétablissement grâce à ses avantages, notamment : réduction de la durée d'hospitalisation, séances de rééducation individualisées, économies de temps et d'argent tant pour le traitement que pour le transport. Cependant, comme les thérapeutes ne sont pas toujours physiquement présent pour suivre et guider la thérapie de rééducation à domicile, ils évaluent les progrès réalisés à l'aide de questionnaires, d'entretiens ou même de mesures d'auto-évaluation, où le patient doit consigner l'exécution du pro-

gramme d'exercices quotidien [10]. Bien que ces méthodes traditionnelles soient utilisées depuis près de 50 ans [9], elles ne sont pas suffisamment précises pour l'évaluation des patients. Ces derniers peuvent soit surestimer, soit sous-estimer de manière significative leur réelle dépenses énergétiques et leur taux d'inactivité. L'outil d'évaluation idéal devrait être capable d'identifier et de fournir des estimations précises de l'intensité, du volume, de la durée et de la fréquence des différentes activités prescrites [10].

Pour résoudre ce problème, et fournir une évaluation et un suivi objectif de l'état du patient, de nombreuses études ont examiné la faisabilité et l'efficacité des approches de suivi autonome [15–17]. Cela vise à mettre au point des solutions de rééducation qui peuvent être utilisés facilement et en toute sécurité à domicile et qui sont agréables et abordables. Ces solutions reposent sur l'incorporation de capteurs et l'intelligence artificielle dans les espaces de vie, souvent appelés environnements intelligents [18, 19].

Par environnement intelligent, on entend un environnement capable de collectées, traitées et analysées des données afin de rendre la vie des gens considérablement plus accessible dans le but d'offrir des services adaptés [20]. Un tel environnement consiste généralement en un réseau de capteurs composé d'une série de dispositifs intégrés, d'algorithmes intelligents et de diverses Interactions Homme-Machine (IHM). Les données peuvent être des données environnementales ou de localisation, ou des données produites par les utilisateurs tel que des informations comportementales et physiologiques [21]. Deux concepts fondamentaux sont alors élaborés : la collecte de données à partir d'un réseau de capteurs et l'intelligence ambiante [21].

Les réseaux de capteurs permettent une surveillance environnementale omniprésente, en offrant de nombreux avantages importants tels que la couverture de vastes zones, la surveillance à long terme et l'évolutivité du système. Les réseaux de capteurs se composent de plusieurs dispositifs distribués et autonomes, qui interagissent avec l'environnement et capables de détecter des paramètres physiques tel que les mouvements, les gestes, les chutes, etc.

En matière de collecte de données, un large éventail de capteurs disponibles peut être envisagé pour le déploiement d'un réseau de capteurs. En fonction de certains facteurs tels que la mise en œuvre, le type de données collectées ainsi que l'impact sur la vie privée, les capteurs peuvent être regroupés en différentes catégories. Il s'agit par exemple de capteurs intégrés dans des objets de la vie quotidienne comme une montre, un smartphone, de capteurs ambiants dispersés dans l'environnement audiovisuel comme les caméras ou utilisant la technologie des radiofréquences comme les radars.

La combinaison des réseaux de capteurs avec la recherche en informatique et en intelligence artificielle a permis de créer un concept interdisciplinaire d'intelligence ambiante. L'intelligence ambiante a été présentée pour la première fois par Eli Zhelkalors d'un atelier de recherche de Philips en 1998 [22]. Selon la littérature [23, 24], l'intelligence ambiante a pour objectif d'exploiter les capacités perceptives offertes par tous les capteurs pour analyser l'environnement, les utilisateurs et leurs activités, et permettre au système de réagir en fonction du contexte. En effet, l'intelligence ambiante est le cœur du problème de l'assistance. Pour qu'un patient post-AVC puisse être assisté lors de séances de thérapie à domicile, il est impératif de mettre en œuvre un système d'évaluation holistique pour déterminer l'intensité et la dépense énergétique associée à l'activité. Cela peut se faire en combinant plusieurs catégories d'évaluation lors des séances de thérapie comme une reconnaissance des activités qui englobe les actions, gestes, etc.

Problématiques et motivations de la recherche

La Reconnaissance des Gestes de la Main (RGM) et des Actions Humaines (AH) sont au cœur des solutions qui peuvent être déployées pour l'assistance des patients post-AVC lors des séances de thérapie à domicile. Cependant, plusieurs défis entravent la mise en œuvre de ces technologies. L'un des principaux obstacles réside dans l'acceptation des utilisateurs. L'introduction de ces solutions dans le cadre domestique peut susciter des préoccupations quant à l'intrusion dans la vie privée des patients, ce qui peut entraîner un certain inconfort. Pour garantir l'adoption et l'acceptation par les patients, il est impératif de concevoir des solutions d'assistance conviviales qui mettent un fort accent sur la protection de la vie privée et la confidentialité des individus. Cela permet aux patients de bénéficier pleinement de ces technologies tout en se sentant en sécurité chez eux. De plus, les environnements domestiques sont notoirement complexes et sujets à des variations constantes. Les conditions environnementales, telles que la luminosité changeante et l'obscurité, ou bien la présence d'objets, rendent la détection des activités plus ardue. Les systèmes de reconnaissance doivent être capables de s'adapter à ces variations pour fournir des résultats fiables. En outre, la diversité des gestes de la main et des actions humaines nécessite une quantité considérable de données pour former des systèmes fiables, ce qui représente un défi majeur. La collecte de données est effectivement une tâche qui peut s'avérer difficile et exigeante en termes de temps et d'efforts. La conception d'un processus de collecte de données efficace nécessite une planification minutieuse. Cette phase prélimi-

naire demande des compétences pour assurer que les données recueillies seront pertinentes et utiles.

L'objectif principal de cette thèse est de proposer un système d'assistance à domicile, le plus intuitif et le plus discret possible. L'évolution récente des technologies de détection a conduit au développement de réseaux de capteurs sans fil qui offrent une solution intéressante, peu coûteuse. Dans la panoplie des approches disponibles, l'utilisation des radars impulsionnels ultra large bande (Impulse radio ultra-wideband : IR-UWB) semble être la solution la plus adéquate. Ils présentent des caractéristiques uniques qui les rendent exclusifs par rapport à d'autres approches de détection. Les IR-UWBs sont simples à installer et recueillent des données en dehors du visible, offrant plus d'intimité et ne posant aucun problème de confidentialité. De plus, étant des capteurs ambiants dispersés dans l'environnement, ils n'entraînent pas d'encombrement ni de gêne pour les utilisateurs.

Contribution de la recherche

Les travaux antérieurs ne se sont concentrés sur des problèmes individuels, soit la RGM soit des AH. Ce travail de thèse a fait un pas de plus pour combiner les deux taches.

Plusieurs algorithmes de classification ont été proposés dans la littérature. La plupart d'entre eux se concentrent sur l'obtention de performances élevées. Cela s'accompagne généralement par des traitements de données intensifs et des algorithmes complexes avec de très grands paramètres d'apprentissage, des calculs importants et une consommation de mémoire élevée. Outre la motivation de concevoir des algorithmes avec de meilleurs résultats de classification, il est nécessaire de trouver des solutions plus légères et moins coûteuses en termes de calcul, notamment pour justifier l'application de ces algorithmes dans des cas d'utilisation réels. Notre première contribution consiste en une solution de classification de bout en bout pour la RGM et des AH. Le modèle combine deux classes d'architecture : les réseaux de neurones convolutifs (Convolutional Neural Network : CNN) et les cellules de longue mémoire à court terme (Long Short-Term Memory : LSTM). Contrairement aux architectures profondes, notre modèle consiste en une architecture parallèle et hybride qui exploite la concaténation de données, les caractéristiques spatio-temporelles et consiste en une structure optimisée avec un nombre réduit de paramètres entrainable.

Notre seconde contribution se concentre sur la mise en place d'une méthode d'augmentation de données pour les actions humaines, sans nécessiter de collecte supplémentaire. Cette

approche repose entièrement sur la manipulation des données au moyen du traitement d'images. Au départ, les images subissent une décomposition en coefficients à l'aide de la transformation en ondelette. Ces coefficients sont ensuite employés dans diverses configurations pour la création de nouvelles données. À partir d'un seul échantillon, trois nouvelles images sont générées et utilisées dans le processus d'entraînement du modèle de classification. L'application de cette technique d'augmentation de données a révélé une nette amélioration des performances par rapport aux travaux précédents.

Notre troisième contribution repose sur la motivation principale d'exploiter un réseau de capteurs visant à améliorer la fiabilité du système de reconnaissance, particulièrement en cas de panne ou de défaillance d'un capteur. Dans cette perspective, nous avons élaboré un modèle de classification spécialement conçu pour les systèmes multi-capteurs dédiés à la RGM . Nous avons adopté une approche d'entrées-sorties multiples, permettant de traiter simultanément les données provenant de divers capteurs de manière autonome et indépendante. Cette approche confère au modèle la capacité de mieux appréhender les nuances et les spécificités des informations receuillis par chaque capteur, ce qui, en fin de compte, renforce la précision, la fiabilité et la robustesse de la reconnaissance des gestes et des actions.

Organisation de la thèse

Le manuscrit est organisé et chapitré en 5 parties :

Chapitre 1: Ce chapitre est divisé en deux grandes parties. La première partie vise à présenter une vue d'ensemble du domaine de la reconnaissance des activités humaines. Nous commençons par examiner la diversité des catégories et types d'activités humaines. Ensuite, nous explorons le processus de reconnaissance en décrivant comment les systèmes sont conçus pour identifier et interpréter les activités. Cette première partie se conclut par un examen détaillé des différentes catégories de capteurs utilisés dans la reconnaissance des activités humaines, avec un accent particulier sur la technologie IR-UWB. Dans la seconde partie, nous expliquons le fonctionnement des IR-UWBs et explorons les concepts fondamentaux liés à la collecte et au traitement des données IR-UWB, notamment leur utilisation dans la RGM et AH. Enfin, nous présentons les IR-UWBs disponibles sur le marché et utilisés dans ce contexte.

Chapitre 2 : Ce chapitre procède à une étude approfondie de l'état de l'art dans

le domaine de la RGM et des AH en utilisant la technologie IR-UWB. Dans un premier temps, nous explorons en détail les concepts fondamentaux, notamment les algorithmes d'apprentissage automatique et les algorithmes d'apprentissage profond, en mettant l'accent sur leurs architectures et leur fonctionnement. Ensuite, nous examinons de manière détaillée les travaux antérieurs, en mettant en lumière les différents modèles proposés et les stratégies d'augmentation des données utilisées.

Chapitre 3 : Ce chapitre se compose de trois sections. Dans la première section, nous décrivons en détail le système de RGM et des AH. Cela englobe une exploration approfondie du processus de prétraitement des données, ainsi que la mise en place du modèle de classification. La deuxième section se concentre sur l'étape de RGM, couvrant divers aspects tels que la présentation de la base de données, la mise en œuvre du modèle, et l'analyse des résultats expérimentaux que nous avons obtenus. La troisième section est dédiée à la reconnaissance des actions humaines. Nous débutons par une description complète de la base de données que nous utilisons. Ensuite, nous introduisons une nouvelle stratégie d'augmentation des données spécifiquement appliquée aux signatures micro-Doppler. Enfin, nous abordons la phase d'entraînement et d'évaluation du modèle de classification afin de démontrer l'efficacité de notre approche.

Chapitre 4 : Ce chapitre présente un système de RGM basé sur l'utilisation de multiples capteurs. Nous soulignons son importance par rapport aux systèmes à capteur unique et nous abordons également ses caractéristiques distinctives. Nous détaillons par la suite l'implémentation du modèle, en décrivant le processus d'apprentissage des gestes de la main, le réglage des paramètres pour optimiser les performances du modèle, et enfin, l'évaluation pour mesurer l'efficacité et la précision de la reconnaissance des gestes.

Chapitre 1

Etat de l'art sur la reconnaissance des activités humaines par capteurs UWB

1.1 Notions d'activité

Une activité fait référence aux mouvements du corps entier et/ou de ses membres au moyen de la contraction des muscles squelettiques, ce qui entraîne une dépense énergétique [25, 26]. L'activité est un paramètre complexe variant dans le temps qui peut être classé quantitativement par sa fréquence (nombre d'activité dans une période de temps spécifique), sa durée (temps passé à effectuer une activité) et son intensité (effort physiologique ou taux de dépense énergétique), mais aussi qualitativement en grandes catégories de comportement sédentaire, de locomotion, de travail, de loisirs, etc. Le tableau 1.1 éxplique en détails les paramètres décrivant une activité.

1.2 Catégories d'activités

On distingue deux catégories d'activités [27] :

— Activité structurée : ou activité guidée, implique un ensemble de règles ou d'instructions pour effectuer des mouvements planifiés afin de promouvoir des bénéfices pour la santé. En général, un temps est réservé à l'exécution d'activités structurées par

Tableau 1.1 – Paramètres décrivant une activité : Mode, Fréquence, durée et intensité.

Paramètres	**Définition et contexte**
Mode	Le type d'activité effectuée réalisée par le membre impliqué dans son exécution. Le mode peut également être défini selon le contexte d'exigences par exemple, entraînement de résistance ou de force, entraînement d'équilibre et de stabilité.
Fréquence	Le nombre d'activité effectuée dans une période de temps spécifique, par exemple, par jour ou par semaine. Dans le contexte de l'activité favorable à la santé, la fréquence est souvent désignée par le nombre de séances.
Duration	Temps par minutes ou heures passées à effectuer une activité au cours d'une période donnée par exemple, jour, semaine, mois.
Intensité	Taux de dépense énergétique ou effort physiologique. L'intensité peut être quantifiée par des mesures physiologiques telles que le rapport d'échange respiratoire, la fréquence cardiaque ou évaluée par des caractéristiques perceptives, par exemple l'évaluation de l'effort perçu, le test de la marche et de la parole, ou encore quantifiée par des mouvements corporels, par exemple la fréquence des pas.

exemple le sport, la rééducation.

— Activité non structurée : implique des mouvements non planifiés, effectués tout au long de la journée, généralement à une intensité moindre que celle prévue. Il peut s'agir d'activités quotidiennes au travail, à la maison ou pendant les déplacements.

1.3 Types d'activités

En fonction de la complexité, de la durée et des parties du corps impliquées dans le mouvement, on peut distinguer différents types d'activités [28] (voir Figure 1.1) :

— Geste : défini comme le mouvement élémentaire d'un seul membre du corps humain. Il est généralement caractérisé par une faible complexité et une faible durée temporelle. Hocher la tête, lever la main, étirer le bras sont de bons exemples de gestes.

— Action : définie comme une activité simple, qui combine une série de gestes, organisés dans le temps, et qui inclut le mouvement de plus d'une partie du corps humain. En d'autres termes une action représente un mouvement particulier qui peut faire partie d'une activité plus complexe. On peut citer quelques exemples

d'actions comme "marcher", "parler", "boire", "s'asseoir", etc.

— Intéraction : implique la présence de deux sujets ; l'un doit être un humain, tandis que l'autre peut être un humain ou un objet. Il peut s'agir d'une interaction homme-homme ou une interaction homme-objet. Par exemple : deux personnes qui se serrent la main.

— Activité : souvent appelé événement, c'est un mouvement corporel plus complexe, qui décrit les gestes ou les actions d'une ou plusieurs personnes. Une activité peut inclure :

  — Un ensemble d'actions d'une seule personne. Par exemple : "marcher et compter en même temps" est une activité composée de deux actions qui se déroulent en parallèle.

  — Une interaction ou un ensemble d'interactions entre une personne et un ou plusieurs objets. Par exemple : "parler au téléphone" ou "prendre un médicament".

  — Une interaction ou un ensemble d'interactions entre plusieurs personnes et/ou objets.Par exemple : "une personne vole le sac d'une autre personne".

  — Une activité de groupe qui s'agit d'une combinaison de gestes, d'actions et d'interactions. Il implique plus de deux humains et un ou plusieurs objets. Par exemple : "un groupe de personnes manifestant", " deux équipes jouant à un jeu".

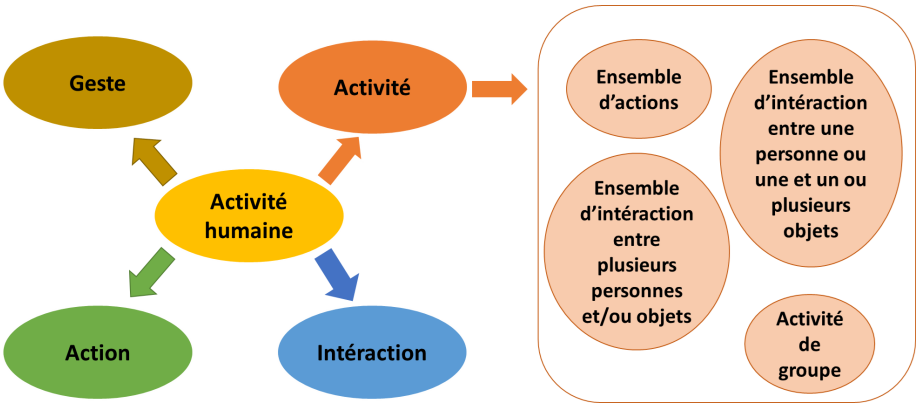


Figure 1.1 – Types d'activités humaines.

1.4 Processus de reconnaissance des activités humaines

La reconnaissance des activités est traitée comme un problème de classification. C'est un processus qui vise l'identification et l'analyse automatique des activités effectuées par une ou plusieurs personnes à partir d'une séquence d'événements observés par différents types de capteurs. Le but ultime est la compréhension des comportements des personnes et de leurs interactions avec l'environnement réel. Il s'agit d'un domaine de recherche actif et émergent, impliquant une grande variété d'applications potentielles dans divers domaines, notamment la surveillance des soins de santé, les sports, les maisons intelligentes, l'IHM, la réalité virtuelle, la surveillance et la sécurité [29–31]. Le processus de reconnaissance des activités comporte principalement quatre étapes : l'acquisition des données, prétraitement, extraction et séléction des caractéristiques et classification.

1.4.1 L'acquisition des données

Le développement d'un système de reconnaissance d'activité nécessite d'abord une collection correcte des données. Alors que certaines applications nécessitent le couplage de plusieurs capteurs à savoir des capteurs homogènes ou même hétérogènes [32–35], d'autres peuvent n'adopter qu'un seul capteur capable de capturer plusieurs paramètres [36]. Cette première phase est très cruciale et doit être réalisée de manière appropriée, car l'ensemble du processus de reconnaissance dépend des données recueillies et de leur qualité. Trois points importants doivent être pris en considération : le type, l'emplacement et le nombre de capteurs à utiliser. Une sélection incorrecte de ces derniers peut nuire aux performances de reconnaissance. Dans le cadre du suivi des patients post-AVC à domicile, le choix du capteur se fait selon cinq critères comme indiqué sur la Figure 1.2. Notre préoccupation majeure est la protection de la vie privée des patients. L'ensemble de données recueillies doivent être préservées ou anonymisées. En outre le capteur déployé dans le contexte du suivi à domicile doit être peu coûteux et opérable sans discontinuité. En d'autres termes, il s'agit d'un capteur robuste qui fonctionne avec une faible consommation d'énergie, quelles que soient les conditions environnementales. Par exemple, il doit fonctionner aussi bien dans des environnements très éclairés que dans des environnements sombres. De plus le capteur doit être non invasif. Il tend à collecter des informations et à déduire l'activité sans interrompre ou déranger les sujets.

Dans ce cadre, l'Internet des objets (Internet Of Things : IoT) joue un rôle crucial. Les capteurs utilisés pour l'acquisition des données peuvent communiquer entre eux et

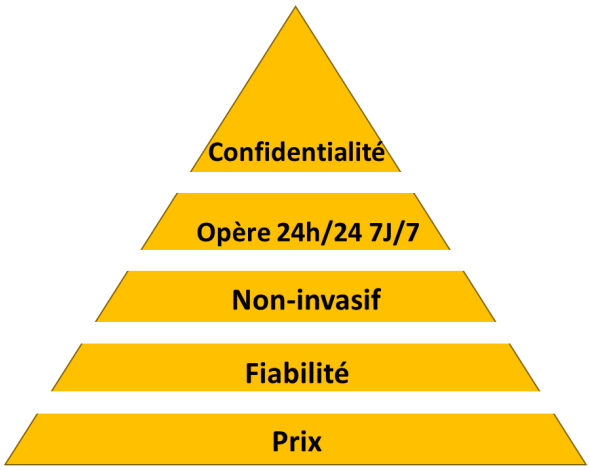


Figure 1.2 – Critères de selection de capteur.

avec l'unité centrale de traitement via des réseaux sans fil tels que le WiFi ou les technologies UWB. Cette communication IP permet une transmission efficace et en temps réel des données collectées vers l'unité de prétraitement, garantissant ainsi une analyse rapide et une réaction immédiate si nécessaire. L'IoT facilite également l'interconnexion de multiples capteurs déployés dans différents endroits, assurant une couverture complète et une surveillance continue des activités des patients.

1.4.2 Prétraitement

Afin d'extraire efficacement les caractéristiques des données des capteurs, une étape de prétraitement est tout d'abord nécessaire. Elle correspond à l'élimination d'artefacts pouvant contaminer les données pour diverses raisons, telles qu'un capteur mal fixé ou des artefacts externes, des bruits et des vibrations. Dans cette étape le filtrage est requis, pour minimiser la quantité de redondance et données indésirables. Cette étape de prétraitement englobe aussi la phase de normalisation et d'étiquetage des données.

1.4.3 Extraction et sélection des caractéristiques

L'application des techniques d'extraction des caractéristiques vise à caractériser les données des capteurs avant leur classification en modèles d'activité. Les caractéristiques peuvent être de différents types a savoir : temporel, fréquentiel, etc. Cependant, un grand ensemble de caractéristiques contenant des informations redondantes ou non pertinentes peut augmenter la charge de calcul, ralentir le processus de classification et nuire à la

précision de la reconnaissance. Ainsi, la mise en œuvre de techniques de sélection des caractéristiques les plus appropriées est un facteur important pour une classification efficace. Cette sélection vise à filtrer les informations pertinentes qui sont par la suite utilisées comme données d'entrée pour les algorithmes de classification. Les techniques d'extraction et sélection des caractéristiques se divisent en trois grandes catégories à savoir les approches basées sur : l'extraction manuelles, l'apprentissage profond et les approches hybrides (voir Figure 1.3).

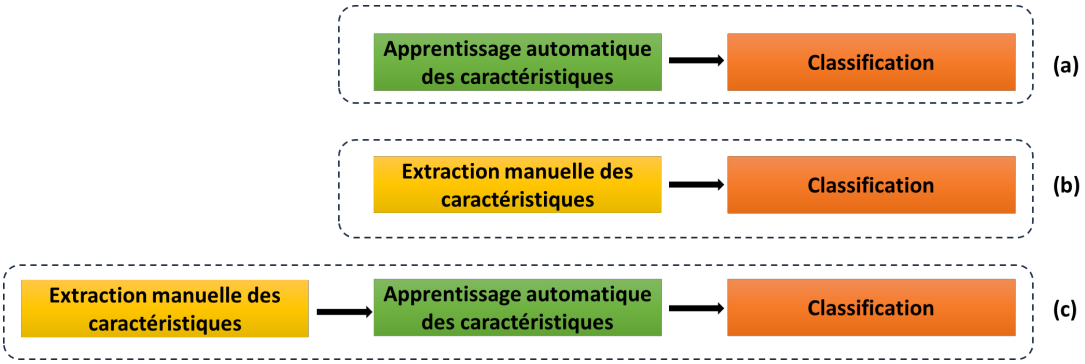


Figure 1.3 – Approches d'extraction des caractérisitques : (a) automatique, (b) manuelle, (c) hybride.

1.4.4 Classification

La classification consiste à attribuer une étiquette de classe d'activité à chaque instance de l'ensemble des caractéristiques extraites. Pour ce faire, un classifieur est entrainé à l'aide d'un ensemble de données d'apprentissage, puis évalué sur un ensemble de données test. Le principe de la plupart des algorithmes de classification consiste à déterminer des frontières de décision capables de séparer les classes dans l'espace des caractéristiques. Les approches d'apprentissage peuvent être divisées en trois catégories: les approches supervisées, non-supervisées et semi-supervisées, comme indiqué sur la Figure 1.4 . Pour l'apprentissage supervisé, l'ensemble des données d'apprentissage doit être étiqueté avant d'entraîner le classificateur. En d'autres termes, cela veut dire que les différentes classes possibles sont déjà connues et que pour fonctionner, l'algorithme reçois comme entrée une paire de données et son étiquette associée. Tandis que pour l'apprentissage non supervisé les classes ne sont pas connues à l'avance. Le classifieur utilise des données non étiquetées et identifie automatiquement un certain nombre de groupes, chacun correspondant à une certaine activité. En ce qui concerne l'apprentissage semi-supervisé, cette approche est

peu utilisée. Elle est à mi-chemin entre les approches supervisées et non-supervisées. Elle consiste en l'utilisation des données étiquetées et non étiquetées. Les classes sont connues lors de l'apprentissage, mais l'algorithme accepte en entrée des données qui ne sont pas forcément étiquetés.

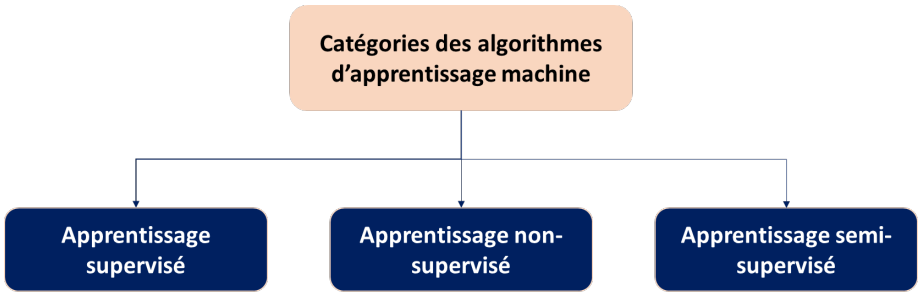


Figure 1.4 – Types d'algorithmes de classification.

1.5 Principaux capteurs pour la reconnaissance des activités humaines

Diverses techniques pour la reconnaissance des activités humaines ont été proposées. Habituellement, ces techniques sont divisées en trois catégories, à savoir les méthodes basées sur la vision, les méthodes basées sur les capteurs portables et les méthodes basées sur les capteurs ambiants [37]. Ces méthodes présentent une série de variations en termes de précision, de complexité, de fonctionnement, de confort de l'utilisateur et de coût.

1.5.1 Les capteurs portatifs

Les méthodes basées sur le contact nécessitent une interaction physique entre l'utilisateur et le dispositif d'acquisition comme les capteurs montés sur le corps tels que l'accéléromètre ou les capteurs portables tels que les gants [38, 39]. Ces méthodes génèrent une quantité limitée de données qui peuvent être traitées facilement par rapport aux données vidéo et fournissent des résultats précis. Néanmoins, elles sont de plus en plus abandonnées. Elles nécessitent le port d'un équipement ou le reliement avec une multitude de câbles à un ordinateur ce qui peut s'avérer peu pratique dans de nombreux scénarios. Tout cela dépend de l'acceptation de l'utilisateur et de sa volonté d'effectuer des tâches de surveillance continue. En outre, le contact physique requiert un certain niveau de compétence

et un équipement sophistiqué et ne serai destiné à l'utilisation que par des utilisateurs expérimentés.

1.5.2 Les capteurs de vision

Les méthodes basées sur la vision consistent à utiliser différents types de caméras tels que: caméra couleur, caméra de profondeur, caméra infrarouge et caméra thermique [40,41]. Les activités sont capturées sous forme de vidéos et traitées image par image pour supprimer les parties inutiles et les objets sans rapport et ne conserver que les informations importantes. Un ensemble de caractéristiques décrivant au mieux les activités est extrait et fourni à un classifieur. Ce dernier détermine quel type de posture est lié à quelle activité, et attribue l'étiquette appropriée à l'activité effectuée. Selon [42], Il existe principalement deux approches de reconnaissance par vision : les approches à couche unique et les approches hiérarchiques. Les approches à couche unique reconnaissent les activités simples à partir d'une séquence d'images et conviennent mieux à la reconnaissance de gestes ou d'actions. Les approches hiérarchiques reconnaissent les activités plus complexes en les décomposant en activités ou en actions simples. Bien que les méthodes basées sur la vision s'avèrent être une option attrayante pour la reconnaissance des activités humaines en raison de leur facilité de mise en œuvre, de leur capacité de surveillance à distance et de la grande quantité d'informations capturées, elles peuvent parfois être jugées inacceptables pour de multiples raisons. Une préoccupation majeure est le risque de violation de la vie privée lorsqu'elles sont utilisées dans des environnements personnels. Les caméras peuvent être affectées négativement par les conditions météorologiques, les changements soudains de l'éclairage environnemental et d'arrière-plan, la sensibilité aux variations de taille et de vitesse, ou la présence d'occlusions sévères. En outre, les données issues de la vision sont plus volumineuses et nécessitent un traitement plus important.

1.5.3 Les capteurs ambiants

Cette approche repose sur l'utilisation de capteurs incorporés dans des lieux souvent qualifiés d'environnements intelligents. Les capteurs produisent des données sous forme de séries temporelles de changements d'état ou de valeurs de paramètres. Un large éventail de capteurs ambiants peut être utilisé pour surveiller l'ensemble d'une pièce ou d'une maison. Nous constatons trois catégories de capteurs ambiants notamment : Les capteurs qui détectent la pression et sont installés sur les sols [43], les capteurs qui détectent les

sons ambiants [44] et les capteurs qui détectent les mouvements (WiFi [45], Radar [46], ultrasons [47]). Les systèmes de reconnaissance d'activités basés sur des capteurs ambiants semblent dominer le domaine de la recherche sur les maisons intelligentes [48,49]. En effet, les capteurs ambiants sont généralement considérés comme moins intrusifs et sont donc mieux acceptés.

Le tableau 1.2 résume les avantages et les inconvénients des technologies de capteurs sur la base de certains critères.

Tableau 1.2 – Comparaison entre différents capteurs utilisés pour la reconnaissance des activités humaines.

Critère de comparaison	**Capteur portatif**	**Capteur de vision**	**Capteur ambiant**
Confidentialité	Aucun problème de confidentialité. Aucune image visuelle n'est capturée.	Une moindre acceptabilité par les utilisateurs en vue de la violation de leurs vie privée.	Aucun problème de confidentialité. Aucune image visuelle n'est capturée.
Sensibilité au conditions environnementales	Moins sensible.	Plus sensible.	Moins sensible.
Problème d'occlusion	Non.	Oui.	Non.
Problème de santé	Peut provoquer une gêne pour les utilisateurs et causer des problèmes de sensibilité de la peau.	Ne cause aucun problème de santé lié à la peau	
Confort	Moins pratique. Les utilisateurs sont toujours obligés de porter un capteur ou une charge de câbles qui relient l'appareil à un ordinateur.	Plus pratiques. Les utilisateurs ne sont pas obligés de porter ou attacher un capteur au corps.	
Plage de détection	Généralement élevée.	Fonctionne dans un environnement confiné (limité).	
Fonctionnement	Capacité à capturer des données d'activité à l'intérieur comme à l'extérieur. La position d'un capteur portable affecte grandement la précision de détection.	Capacité à capturer des données d'activité qu'en milieu intérieur. Ligne de mire (Line of sight) généralement requise entre la cible et les capteurs.	
Problème lié au capteur	Peut-être perdu ou oublié. Susceptible d'être endommagé ou cassé.	Le capteur est généralement fabriqué à l'intérieur d'un appareil ou installé à un certain endroit.	

1.6 Les capteurs UWB pour la reconnaissance des gestes de la main et des actions humaines

L'avènement des technologies radar miniaturisé et à faible coût, a rendu possible de nouvelles applications radar notamment dans les domaines de la médecine et des environnements intelligents. La sécurité, la fiabilité, la portabilité et le prix abordable des appareils radar en font des candidats de choix pour une utilisation en environnement in-

térieur. Le concept général du radar repose sur l'émission des ondes électromagnétiques sur une certaine gamme de fréquences et l'analyse des signaux réfléchis par les objets qu'elles rencontrent (voir Figure 1.5). Grâce à ces réflexions, il est capable d'estimer des paramètres tel que la vitesse, la distance, ou l'angle d'arrivée des signaux reçus et permettent de réaliser la RGM et des AH.

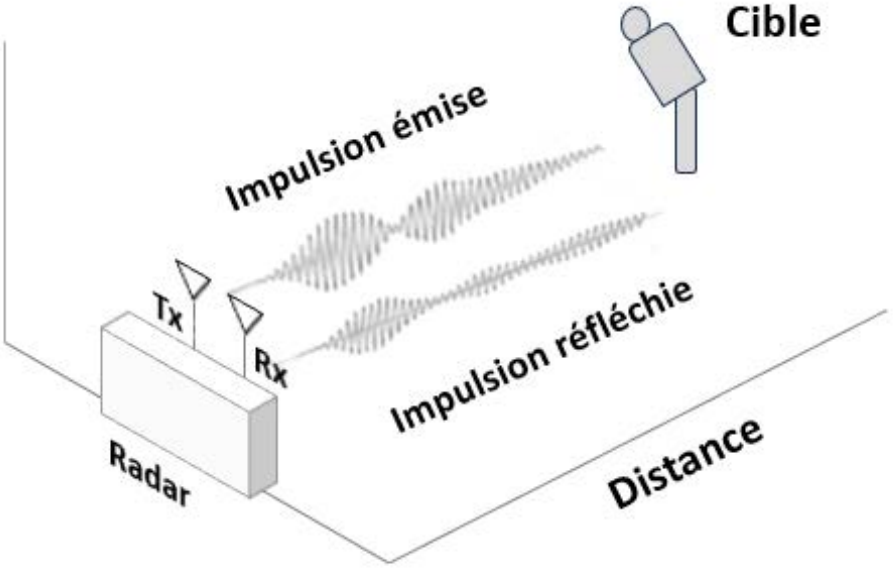


Figure 1.5 – Principe de fonctionnement du radar.

Le radar possède des avantages uniques qui complètent d'autres capteurs, tels que les capteurs visuels et les capteurs portables. Par rapport aux capteurs visuels, le radar fonctionne en dehors de la zone visible. Les signaux radar réfléchis peuvent donc révéler les mouvements humains sans capturer aucune image visuelle. Cela garantit une plus grande confidentialité pour l'utilisateur. Cela en fait du radar la solution idéale de surveillance régulière pour les environnements sensibles, tels que les habitations. En outre, par rapport aux capteurs portables, le radar ne nécessite pas la fixation d'un dispositif sur le corps de l'utilisateur. Il s'agit d'un dispositif sans contact, ne cause pas d'inconfort ou de gêne et offre une précision et une robustesse accrues. De plus, Il s'agit d'un dispositif insensible aux conditions environnementales, capable de pénétrer des objets opaques, tels que des tables ou des murs. Les radars les plus couramment utilisés pour la RGM et des AH sont les IR-UWBs [50, 51] et les radars à ondes continues (Continuous Wave : CW), à voir les radars Doppler, les radars à ondes continues modulées en fréquence (Frequency Modulated

Continuous Wave : FMCW) [52, 53].

Au cours de cette thèse nous nous intéressons au IR-UWB. Ce dernier est bien adapté à la surveillance en intérieur. Grâce à sa grande largeur de bande opérationnelle, le IR-UWB peut fournir une résolution de portée et une capacité de localisation élevées. De plus, il est résistant aux effets multi-trajets et d'évanouissement et offre des débits de données élevés sur de courtes distances. Compte tenu de l'aspect lié au déploiement du système radar, le IR-UWB présente l'avantage d'une faible consommation d'énergie et d'un prix abordable.

1.7 Radar Implusionnel Ultra large bande (Impulse Radio Ultra-Wideband : IR-UWB)

L'ultra large bande est une technologie de communication sans fil qui utilise des impulsions étroites non sinusoïdales pour transmettre des données. Le signal IR-UWB occupe un large domaine de fréquences, c'est pourquoi il est appelé ultra-large bande (voir Figure 1.6). L'impulsion du signal ultra large bande dans le domaine temporel étant relativement étroite, elle présente une bonne résolution dans le temps et dans l'espace, ce qui permet une forte résistance à l'influence de l'effet de multi trajet. Bien que l'IR-UWB utilise la communication sans fil, son taux de transfert de données peut atteindre jusqu'à quelques centaines de mégabits par seconde, ce qui permet une synchronisation avec précision inférieure au nanomètre.

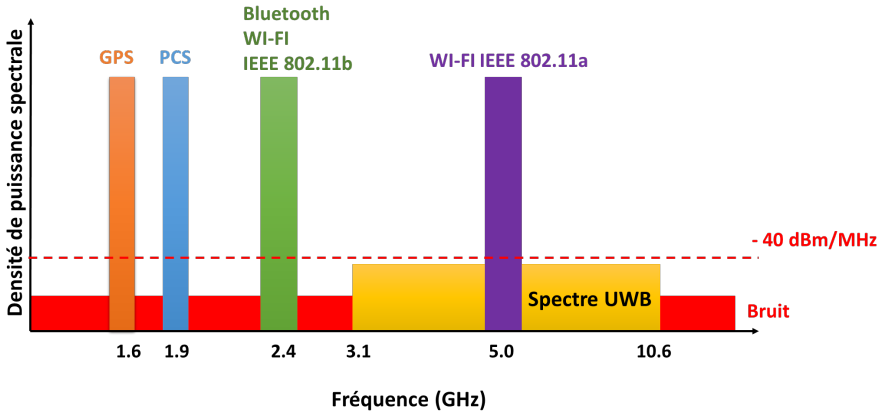


Figure 1.6 – Spectres fréquentiels pour différents types de systèmes radio [51].

Le IR-UWB utilisent une large gamme de fréquences par rapport aux systèmes radar conventionnels allant de 3.1 à 10.6 GHz et reste extrêmement difficile à détecter. La

largeur de bande relative (Bandwidth : B_w) des signaux UWB est approximativement de 500MHz, supérieure à 20% de la fréquence centrale (f_C). La largeur de bande relative est déduite à partir de la largeur de bande absolue. Cette dernière est calculée comme la différence entre la fréquence supérieure (f_H) du point de transmission à -10 dB et la fréquence inférieure (f_L) du point de transmission à -10 dB :

$$B_{abs} = f_H - f_L \tag{1.1}$$

La largeur de bande absolue est également appelée largeur de bande de 10 dB, comme le montre la Figure 1.7.

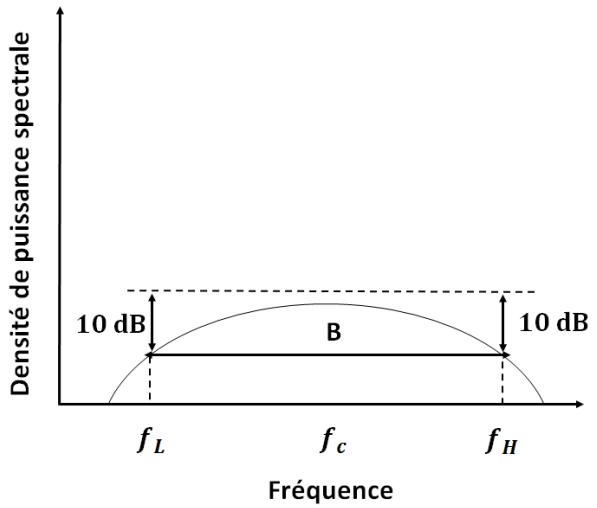


Figure 1.7 – Densité Spectrale de Puissance d'un signal UWB [50].

D'autre part, la largeur de bande relative est définie comme suit:

$$B_w = \frac{B_{abs}}{f_C} \tag{1.2}$$

où f_C est la fréquence centrale donnée par:

$$f_C = \frac{1}{2}(f_H + f_L) \tag{1.3}$$

D'après 1.1 et 1.3, la largeur de bande relative B_w dans 1.2 peut être exprimée comme suit :

$$B_w = 2(\frac{f_H - f_L}{f_H + f_L}) \tag{1.4}$$

En raison du large spectre disponible pour cette technologie il est possible d'émettre et de

recevoir plus rapidement des impulsions de courte durée, et à faible énergie, réfléchies par des objets ciblés. Il faut savoir que la largeur d'impulsion du signal UWB dans le domaine temporel est inversement proportionnelle à la largeur de bande du signal. Cela signifie que plus la largeur d'impulsion est étroite, plus la largeur de bande du signal est grande. Les impulsions émises sont de durée τ très brève généralement de l'ordre de quelques nanosecondes à quelques centaines de picosecondes et se propagent dans l'atmosphère à la vitesse de la lumière. Ces impulsions extrêmement courtes offrent plusieurs avantages :

— Débit important.
— Large couverture.
— Robustesse au brouillage.
— Faible puissance.
— Pénétration à travers une grande variété de matériaux.

1.8 Fonctionnement du IR-UWB

Le IR-UWB se compose de deux sections de base : la transmission et la réception. La section de transmission comprend un générateur d'impulsions, un amplificateur de puissance, un modulateur, un mélangeur et une antenne d'émission. La section de réception comprend une antenne de réception, un amplificateur à faible bruit (Low Noise Amplifier : LNA), un corrélateur (composé d'un intégrateur et d'un mélangeur) et un filtre passe-bande. La Figure 1.8 montre le schéma d'un IR-UWB.

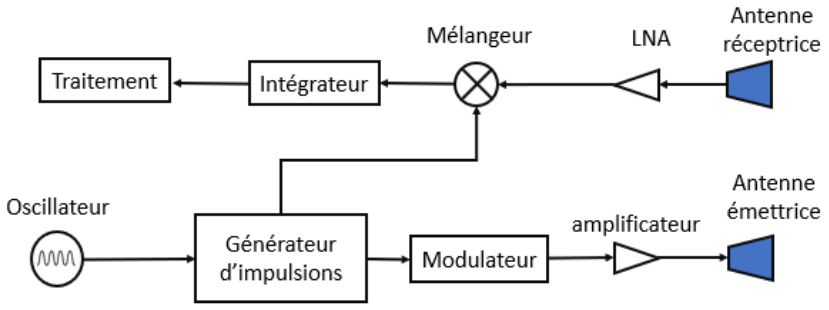


Figure 1.8 – Schéma d'un IR-UWB.

Le générateur d'impulsions est commandé par l'oscillateur et génère une impulsion d'une largeur très étroite pour la transmission. L'oscillateur est chargé de piloter le généra-

teur d'impulsions avec une forme d'onde souhaitée et détermine également la période de répétition des impulsions (Temps de répétition : T_r). Le signal est ensuite amplifié par un amplificateur de puissance dans une certaine mesure tout en respectant les conditions de faible densité spectrale. Le signal est ensuite transmis à l'antenne via une ligne d'alimentation et est ensuite rayonné dans l'espace vers la cible. Le nombre d'impulsions transmises par seconde est appelé fréquence de répétition des impulsions ou taux de répétition des impulsions (Pulse Repetition Frequency : PRF). Le temps qui s'écoule entre le début d'une impulsion et le début de l'impulsion suivante est le Tr et est défini par l'inverse de la fréquence de répétition $\frac{1}{PRF}$. Chaque impulsion est suivie d'un temps de silence afin qu'elle puisse atteindre la cible, puis être réfléchie et captée par l'antenne de réception. La fraction de la durée totale pendant laquelle l'émetteur est en marche s'appelle le rapport cyclique (duty cycle : d_c). Pour une T_r fixe, le rapport cyclique est donné par :

$$d_c = \tau T_r \tag{1.5}$$

avec τ : durée de l'impulsion.
A la réception, le filtre passe bande est conçu pour rejeter les interférences hors bande. Le signal passe par le LNA pour produire un signal avec un faible niveau de bruit par rapport au signal reçu. Il convient de noter que la quantité de réflexion dépend de la distance entre la cible et le radar, de la forme, de la taille et du type de matériau constituant la surface réfléchissante.
L'IR-UWB capte la réflexion émise par chaque objet dans son champ de vision, et la distance entre le radar et la cible peut être facilement déterminée à partir de la différence de temps d'arrivée Δt des instants d'émission d'impulsion, de réception d'écho et de la vitesse de la lumière dans le milieu. La distance est calculée à l'aide de l'équation 1.6.

$$d = \frac{c\Delta t}{2} \tag{1.6}$$

Lorsqu'une cible est située à une distance assez éloignée de sorte que son écho est reçu après l'émission d'une nouvelle impulsion, elle apparait comme proche du radar (voir Figure 1.9). Ces échos de second retour produisent une ambigüité sur la mesure de distance qui se manifeste lorsque le temps aller–retour est supérieur au temps d'écoute entre deux impulsions Δt. Pour lever cette ambiguïté, on définit une distance maximale non ambiguë

d_{msa}, au-delà de laquelle les cibles sont considérées comme indétectables par le radar :

$$d_{msa} = \frac{cT_r}{2} \tag{1.7}$$

L'IR-UWB ne peut donc pas fonctionner à une distance inférieure à la moitié de la longueur d'impulsion dans l'espace. Une telle distance est définie comme la distance de détection minimale d_{min}, ou distance aveugle, comme indiqué sur l'équation 1.8. Cela est dû au fait que nous devons dégager l'antenne de l'impulsion émise avant d'être prête à recevoir un écho.

$$d_{min} = \frac{c\tau}{2} \tag{1.8}$$

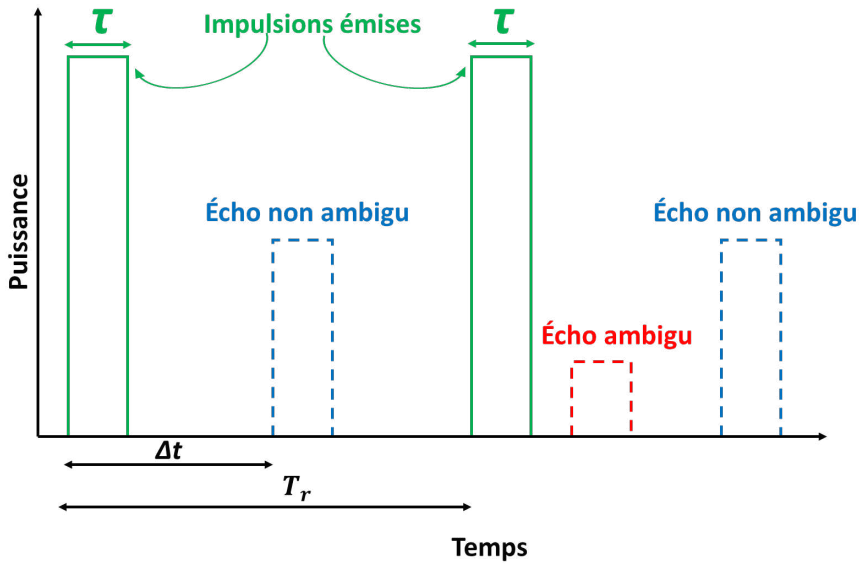


Figure 1.9 – Schéma fonctionnel de l'IR-UWB.

1.8.1 Acquisition et traitement des signaux IR-UWB

Le signal s(t), émis par l'IR-UWB, est considéré comme étant réfléchi par plusieurs chemins différents et reçu par le récepteur IR-UWB désigné par r(t). Les échos reçus de "N" chemins différents sont numérisés sous la forme r[n,k] et représentés par la formule suivante :

$$r[n,k] = \sum_{m=1}^{N_{chemin}} s[n,k-m] + bruit \tag{1.9}$$

s[n,k] représente le signal d'impulsion UWB transmis, et le bruit représente les réflexions par trajets multiples non-désirées. "n" représente les données radar reçues séquencées dans le temps, généralement appelées valeur en temps lent, et "k" représente la distance des réflexions reçues, généralement appelées valeur en temps rapide. La valeur en temps lent est définie par la fréquence de répétition du radar, et l'indice en temps rapide représente le temps d'arrivée du signal. Le signal reçu est stocké sous la forme d'une matrice de données 2D comprenant n lignes et k colonnes, comme le montre la Figure 1.10.

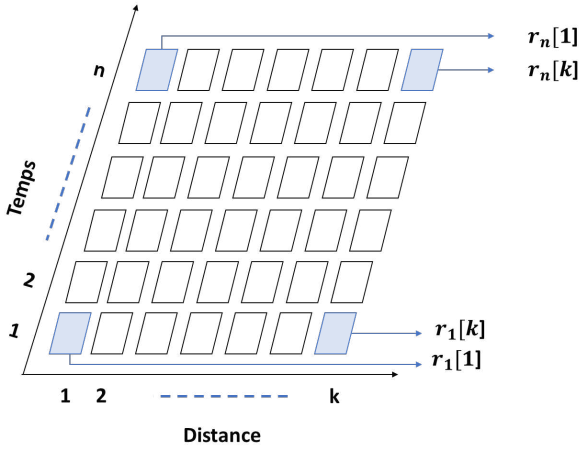


Figure 1.10 – Représentation matricielle des données IR-UWB.

La forme matricielle correspondante, appelée la matrice de données, peut être écrite comme suit :

$$\vec{r} = \vec{s}_{n,k} + bruit \tag{1.10}$$

Sous forme mathématique, la version numérisée des retours radar r[n,k] sous forme 2D ayant respectivement n et k indices de temps lent et rapide, peut être exprimée comme suit:

$$\vec{r} = \begin{bmatrix} r_{n,1} & \dots & r_{n,k} \\ \vdots & \ddots & \vdots \\ r_{1,1} & \dots & r_{1,k} \end{bmatrix} + \vec{C}_{n,k} \tag{1.11}$$

"C" représente les réflexions de l'environnement, généralement des objets statiques dans la portée opérationnelle du radar. Ces réflexions indésirables sont communément appelées fouillis ou clutter.

1.8.2 Filtrage

Le fouillis est un signal de retour radar non désiré. Il contient des interférences dues aux réflexions des objets dans l'environnement qui affectent fortement la probabilité de détection et la précision. Un filtrage est donc nécessaire pour effectuer une opération d'élimination du fouillis avant de poursuivre le traitement. Les techniques de filtrages sont plus ou moins complexes en fonction de l'application, de la vitesse, de la précision et de la mémoire requise. Différentes méthodes sont proposées, comme le filtre en boucle [54], décomposition en valeurs singulières [55], et la méthode du calcul de la moyenne [56, 57], L'indicateur de cible mobile (Moving Target Identification : MTI) [58, 59]. Cependant, le filtre en boucle reste le choix le plus courant [60–62], en raison de sa structure simple et de son faible coût de calcul. La structure du filtre de suppression du fouillis est illustrée à la Figure 1.11. Le filtre comprend une structure récursive avec un seul terme de rétroaction à retard. Le principe de fonctionnement du filtre peut être représenté comme suit:

$$c_n[k] = \alpha c_n[k-1] + (1-\alpha) r_n[k] \tag{1.12}$$

Le terme de fouillis c_n[k] est dérivé en utilisant le terme de fouillis estimé précédemment c_n[k-1] et l'échantillon reçu r_n[k]. α est le facteur de pondération entre 0 et 1. Le fouillis estimé est ensuite soustrait du signal de l'équation 1.11 pour obtenir un signal sans réflexions indésirables. Nous pouvons définir la version sans fouillis de la matrice Z du signal reçu comme suit :

$$\vec{Z} = \vec{R} - \vec{C}_{n,k} \tag{1.13}$$

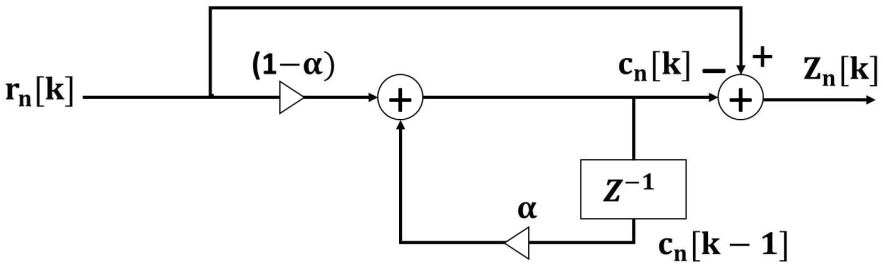


Figure 1.11 – Structure du filtre de suppression du fouillis.

1.9 Domaines de représentation des données

Les formes d'onde organisées sous forme de matrice sont représentées comme une séquence d'intervalles à Temps rapide, chacun étant associé à un temps de mesure spécifique (Temps lent). L'objectif du traitement de ces matrices est de les convertir en de nouveaux domaines dans différents formats de représentation où des caractéristiques utiles peuvent être facilement extraites pour la classification [53].

1.9.1 La transformée de Fourier rapide

La Transformée de Fourier Rapide (Fast Fourier Transform : FFT) est une technique couramment utilisée dans le traitement des données radar et contribue à la conversion d'un domaine à un autre. On appliquons la FFT sur les lignes et les colonnes de la matrice Temps Lent-Temps Rapide, il est possible de générer différentes représentations contenant des informatiques plus riches. La transformée de Fourier rapide est un algorithme qui repose sur la stratégie de diviser pour régner. L'astuce au cœur de l'algorithme de la FFT consiste à présenter le signal sous forme de somme de composante fréquentielles pairs et impaires comme indiqué sur l'équation 1.14, et qui peuvent être calculées simultanément, ce qui conduit à une exécution plus rapide de l'algorithme.

$$x[k] = \sum_{n=0}^{\frac{N}{2}-1} x[n]^{pair} e^{\frac{-2\pi i}{N/2}nt} + \sum_{n=0}^{\frac{N}{2}-1} x[n]^{impair} e^{\frac{-2\pi i}{N/2}nt} \tag{1.14}$$

L'algorithme FFT est présenté ci-dessous:
FFT (x - signal à une dimension).
Étape 1. Calculer N qui représente le nombre d'échantillons du signal.
Étape 2. S'assurer que N est multiple de 2.

Si N%2 > 0 :

Lever l'exception, N doit être une puissance de 2.

Sinon Si N<=0 :

Calculer la transformée de Fourier discrète x[k] = $\sum_{n=0}^{N-1} x[n]e^{\frac{-2\pi i}{N}nt}$.

Sinon :

Pair$_x$ = FFT(Obtenir tous les termes pairs de x).

Impair$_x$ = FFT(Obtenir tous les termes impairs de x).

Calculer E les facteurs exponentiels.

Renvoie les valeurs ajoutées concaténées des sous-séquences calculées en multipliant avec la valeur E.

Fin

La Figure 1.12 synthétise les diverses représentations des données IR-UWB employées dans la RGM et des AH générées par l'application de la FFT. Ces éléments seront détaillés dans les sections suivantes.

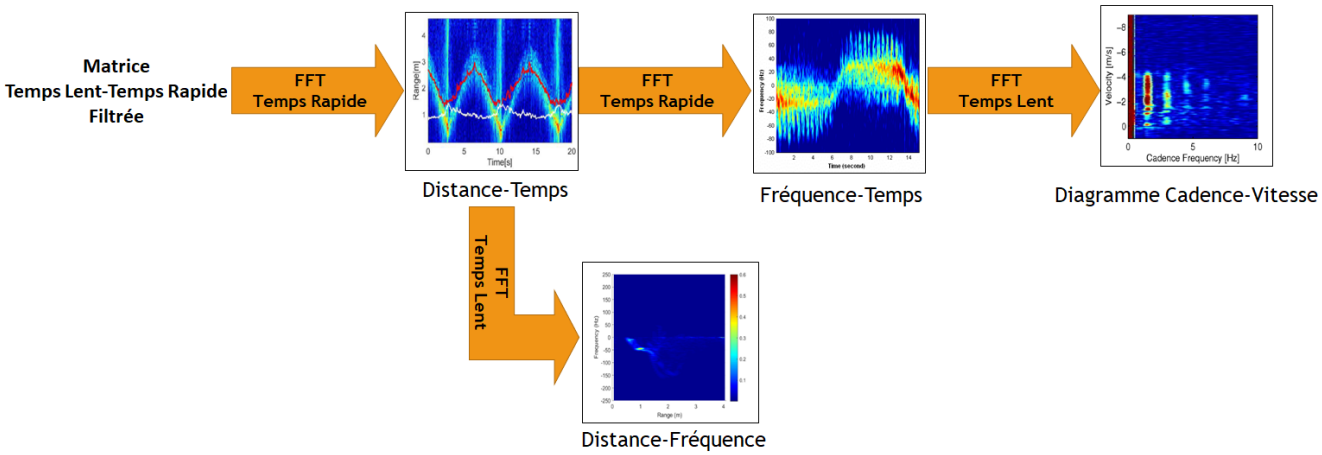


Figure 1.12 – Les différentes représentations des données IR-UWB.

1.9.2 Représentation Amplitude-Temps

La représentation Amplitude-Temps est une représentation 1D brute des données radars sous forme de séries temporelles complexes, dont l'amplitude et la phase peuvent être liées à la diffusion électromagnétique et à la cinématique de la cible observée [53]. Une telle représentation n'est souvent pas utilisée directement pour la classification et est généralement convertie en format 2D à l'aide d'un traitement du signal et l'analyse Fréquence-Temps.

1.9.3 Représentation Temps lent-Temps rapide

Les échantillons des données radar sont organisés dans une matrice 2D, où chaque ligne est un profil de gamme associé à une impulsion transmise [61]. Ainsi, l'index de la ligne correspond au numéro d'impulsion "n", tandis que l'indice de la colonne correspond à un numéro d'échantillon dans un profil de portée. Le temps correspondant à la portée sur l'axe horizontal est appelé Temps rapide (Fast-time) car il est mesuré sur une échelle

comparable à la durée de l'impulsion individuelle. Le temps correspondant au temps sur l'axe vertical est appelé Temps lent (Slow-time), car il est mesuré sur une échelle comparable à l'intervalle de répétition des impulsions T_r. C'est la représentation la plus simple en raison du traitement minimal qu'elle nécessite.

1.9.4 Représentation Distance-Temps

La représentation Distance-Temps est une matrice 2D, obtenue en effectuant la FFT le long de l'axe du temps rapide de la représentation (Temps lent- Temps rapide) [63]. Par l'accumulation des profils de distance dans le temps, il en résulte une représentation 2D qui offre des informations sur la distance variant dans le temps entre la cible et le radar.

1.9.5 Représentation Fréquence-Temps

La représentation Fréquence-Temps également connue sous le nom de spectrogrammes micro-Doppler (Micro-Doppler Signature : MDS), est obtenue en effectuant une FFT avec des fenêtres temporelles courtes et superposées sur la matrice Distance-Temps [64]. Le décalage Doppler variant dans le temps est utilisé pour extraire des informations discriminantes sur le mouvement et les micro-mouvements, tels que la vitesse des différentes parties du corps comme les jambes et les mains.

1.9.6 Représentation Fréquence-Distance

La représentation Doppler-distance est une matrice 2D, obtenue en effectuant la FFT le long de l'axe du temps lent capable de séparer les différents composants du corps humain en mouvement et de localiser la cible avec précision [65]. Une image Fréquence-Distance est clairsemée et contient un espace vacant important à la fois en distance et en vitesse d'une cible mobile à un moment précis. Elle offre également la possibilité de mesurer la distance entre la cible et le radar. En outre, les images Fréquence-Distance sont capables de détecter plusieurs cibles simultanément.

1.9.7 Diagramme Cadence-Vitesse

Le Diagramme Cadence-Vitesse (DCV) est obtenue en effectuant une FFT le long de l'axe temporel des MDSs [66, 67]. Une telle représentation permet d'analyser la nature cyclique et les périodicités inhérentes aux mouvements. Par exemple, le DCV permet de

mesurer la fréquence de répétition des différents mouvements périodiques des membres pendant la marche.

1.9.8 Représentation Temps-Distance-Fréquence

La représentation Temps-Distance-Fréquence est une représentation 3D combinant les trois variables : temps, distance et fréquence dans un format cubique (voir Figure 1.13). Elle peut être obtenue en rassemblant des images Fréquence-Temps qui varient en fonction de la distance [68, 69], ou en empilant des images Fréquence-Temps en fonction du temps [70]. Malgré le fait qu'elle offre de nombreuses informations sur l'activité, elle n'est pas utilisée largement en raison du faible nombre d'algorithmes qui traitent directement les données 3D, et de la complexité du processus.

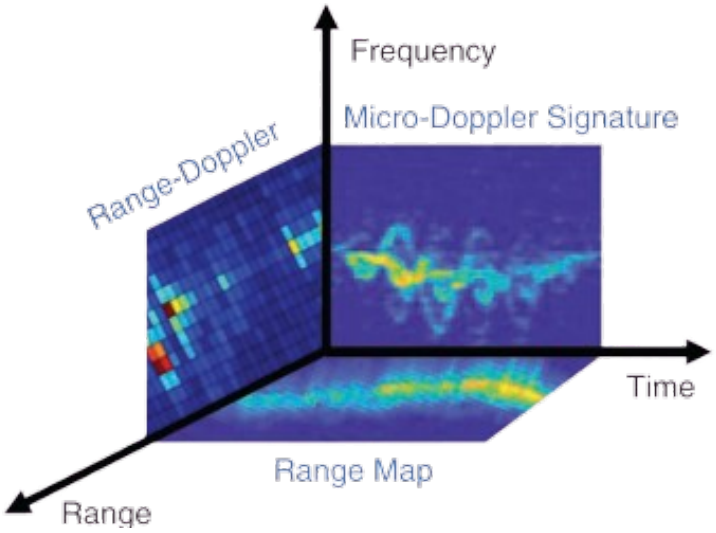


Figure 1.13 – Exemple de représentation Temps-Distance-Fréquence [69].

Le tableau 1.3 résume les différentes approches de représentation des données radars utilisés dans la littérature pour la RGM et des AH.

Tableau 1.3 – Les différentes représentations des données IR-UWB.

Forme d'échos	Domaine de représentation	Références
1D	Amplitude-Temps	[71–73]
2D	Distance-Temps	[60, 62, 63, 74–80]
2D	Fréquence-Temps	[33, 57, 64, 65, 81–95]
3D	Distance-Temps (images Rouge Vert Bleu (RVB)) Temps-Distance-Fréquence	[61] [55, 56, 59, 96–98]

1.10 IR-UWB disponibles dans le commerce pour la reconnaissance des gestes de la main et des actions humaines

De nombreuses études ont utilisé des radars disponibles dans le commerce. Cette section fournit des détails supplémentaires sur ces solutions commerciales. Le tableau 1.4 présente quelques exemples de radars disponibles dans le commerce qui peuvent être utilisés pour la RGM et des AH. Pour tous ces radars, un exemple de référence est également inclus.

Tableau 1.4 – Radars disponibles dans le commerce pour la RGM et RAH.

Modèle radar	Entreprise	Application	Références
XeThru X4	Novelda, Oslo, Norvège	RGM/RAH	[55, 60, 61, 66, 73, 77–79, 81, 99–104]
XeThru X2	Novelda, Oslo, Norvège	RGM	[76]
NVA6100	Novelda, Oslo, Norvège	RGM/RAH	[59, 71, 98]
NVA6201	Novelda, Oslo, Norvège	RGM	[74, 105]
RS640	Novelda, Oslo, Norvège	RAH	[106]
P220	Time domaine part of Humatics huntsville, États unis	RAH	[107]
P410	Time domaine part of Humatics huntsville, États unis	RAH	[108]
P440	Time domaine part of Humatics huntsville, États unis	RAH	[63, 109]
DW1000	DecaWave, Dublin, Irelande	RAH	[110, 111]
DWM1000	DecaWave, Dublin, Irelande	RGM	[112]
NVA-R661	Novelda, Oslo, Norvège	RGM	[113]

1.11 Conclusion

Dans ce chapitre, nous avons présenté les concepts clés liés aux activités humaines, en mettant en lumière les différents types d'activités comme les gestes et les actions, ainsi que leur processus de reconnaissance, essentiel pour les applications de rééducation. Nous avons également examiné les divers types de capteurs utilisés pour cette reconnaissance, en soulignant les avantages des capteurs ambiants, notamment les IR-UWBs, par rapport à d'autres technologies. Ensuite, nous avons détaillé la théorie des IR-UWBs, en expliquant la technologie ultra large bande, ses composants fondamentaux, son fonctionnement, et le processus de traitement des données IR-UWB. Enfin, nous avons décrit les différentes représentations des données utilisables pour la reconnaissance des gestes de la main et des actions humaines (RGM et AH). Le chapitre suivant présentera une vue d'ensemble des principales approches utilisées pour la reconnaissance, notamment celles basées sur l'apprentissage automatique et l'apprentissage profond. La deuxième partie offrira un état de l'art détaillé des travaux antérieurs réalisés pour la RGM et les AH en utilisant les IR-UWBs.

Chapitre 2

Background des techniques pour la reconnaissance des gestes de la main et des actions humaines

2.1 Introduction

Diverses techniques ont été proposées pour la RGM et des AH à partir des données radar ultra large bande. Le schéma généralement adopté se compose d'une étape d'extraction des primitives et d'une étape de classification. L'extraction des primitives consiste à identifier des caractéristiques distinctes à partir des données tout en étant robuste. Elle peut être réalisée de deux manières : manuellement ou automatiquement. Dans ce contexte, le chapitre suivant se divise en deux parties. La première partie présente une vue d'ensemble des principales approches utilisées pour la reconnaissance, notamment les approches basées sur l'apprentissage automatique et l'apprentissage profond. La deuxième partie expose un état de l'art détaillé des travaux antérieurs réalisés pour la RGM et des AH en utilisant le IR-UWB.

2.2 Apprentissage machine

Les approches conventionnelles basées sur l'apprentissage machine (Machine Learning : ML) impliquent une ingénierie manuelle des caractéristiques et requièrent des connaissances d'expert. Ces approches s'appuient sur des caractéristiques créées à la main qui

nécessitent un prétraitement assez important des données pour réduire la dimensionnalité ou extraire les caractéristiques pertinentes (voir Figure 2.1). Ces caractéristiques peuvent être de type fréquentiel ou de type temporel [114]. Quelques exemples de caractéristiques sont indiquées sur le tableau 2.1.

Tableau 2.1 – Type de caractéristiques.

Caractéristiques temporelles	Caractéristiques fréquentielles
Moyenne	Fréquence dominante
Écart-type	Centroïde spectral
Intervalle interquartile	Maximum
Autocorrélation	Moyenne
Percentiles	Médiane
Amplitude crête à crête	Écart-type
Puissance	
Skewness (asymétrie)	
Kurtosis	
Energie	
Moyenne quadratique	

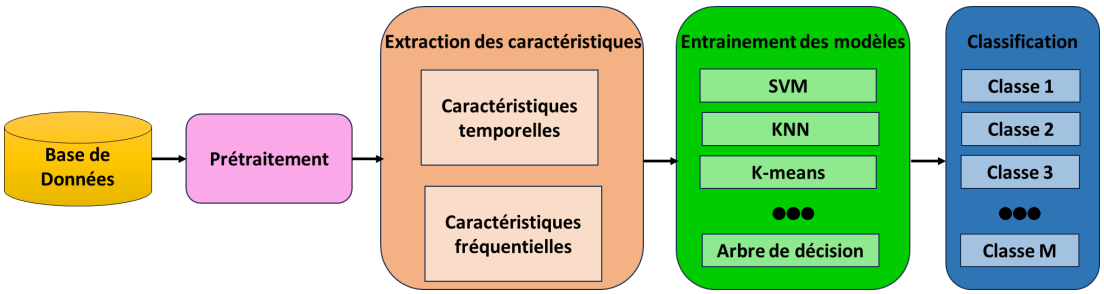


Figure 2.1 – Reconnaissance des gestes de la main/actions humaines par les méthodes conventionnelles d'apprentissage machine.

La section suivante résume les différents algorithmes ML, leurs caractéristiques phares ainsi que leurs avantages et inconvénients permettant de les mettre en contraste les uns par rapport aux autres.

2.3 Les algorithmes d'apprentissage machine

2.3.1 Les machines à vecteurs de support

2.3.1.1 Classification binaire linéairement séparable

Les machines à vecteurs de support (Support Vector Machine : SVM) est un algorithme qui vise à trouver le meilleur hyperplan de séparation entre les échantillons d'apprentissage de différentes classes, celui dont les points sont les plus éloignés : cette distance a un impact direct sur la capacité de généralisation du modèle.

— **Classification binaire pour les données linéairement séparable**

Nous disposons de L points d'apprentissage, où chaque entrée x_i à D attributs (de dimensionnalité D) et appartient à l'une des deux classes y_i= -1 ou +1. Les données d'apprentissage sont de la forme :

$$\{x_i\ , y_i\}\ \text{,oú}\ i = 1, ..., L \in \{\text{-1,1}\}, x \in R^D \tag{2.1}$$

Nous supposons ici que les données sont linéairement séparables, ce qui signifie que nous pouvons tracer une ligne sur un graphique de x_1 vs x_2 séparant les deux classes lorsque D = 2 et un hyperplan sur les graphiques de x_1, x_2, . . .,x_D lorsque D > 2. Cet hyperplan peut être décrit par $w.x + b = 0$ où :

— w est la normale à l'hyperplan.

— $\frac{b}{||w||}$ est la distance perpendiculaire de l'hyperplan à l'origine.

Les vecteurs de support sont les exemples les plus proches de l'hyperplan de séparation et le but des SVM est d'orienter cet hyperplan de manière à ce qu'il soit le plus éloigné possible des membres les plus proches des deux classes.

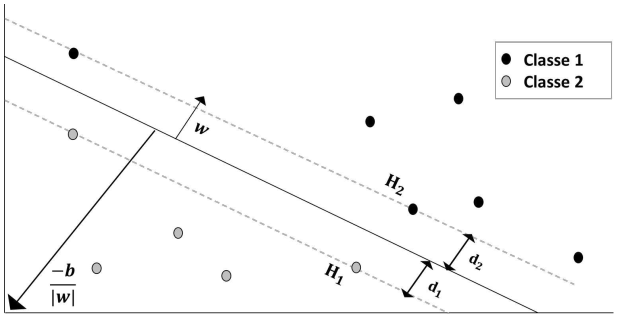


Figure 2.2 – Hyperplan passant par deux classes linéairement séparables

En se référant à la Figure 2.2, la mise en œuvre d'un SVM se résume à sélectionner

les variables w et b de manière à ce que nos données d'apprentissage puissent être décrites comme suit :

$$x_i.w + b \geq +1 \text{ pour } y_i = +1 \tag{2.2}$$

$$x_i.w + b \leq -1 \text{ pour } y_i = -1 \tag{2.3}$$

ces équations peuvent être combinées en :

$$y_i(x_i.w + b) - 1 \geq 0 \ \forall_i \tag{2.4}$$

Si nous considérons les points les plus proches de l'hyperplan de séparation, c'est-à-dire les vecteurs de support (représentés par des points sur la Figure 2.2), alors les deux plans H_1 et H_2 sur lesquels ces points se trouvent peuvent être décrits par :

$$x_i.w + b = -1 \text{ pour } H_1 \tag{2.5}$$

$$x_i.w + b = +1 \text{ pour } H_2 \tag{2.6}$$

En nous référant à la Figure 2.2, nous définissons d_1 comme étant la distance entre H_1 et l'hyperplan et d_2 comme étant la distance entre H_2 et l'hyperplan. L'équidistance de l'hyperplan par rapport à H_1 et H_2 signifie que $d_1 = d_2$ – (une quantité connue sous le nom de marge du SVM). Afin d'orienter l'hyperplan de manière à ce qu'il soit le plus éloigné possible des vecteurs de support, nous devons maximiser cette marge. La marge est égale à $\frac{1}{||w||}$ et la maximiser sous la contrainte de 2.4 est équivalent à trouver :

$$min||w|| \text{ tel que } y_i(x_i.w + b) - 1 \geq 0 \ \forall_i \tag{2.7}$$

Minimiser ||w|| est équivalent à minimiser $\frac{1}{2}$ ||w||^2. Il s'agit donc de trouver :

$$min\frac{1}{2}||w||^2 \text{ tel que } y_i(x_i.w + b) - 1 \geq 0 \ \forall_i \tag{2.8}$$

— **Classification binaire pour les données non entièrement linéairement séparables**

Afin d'étendre la méthodologie SVM au traitement de données qui ne sont pas entièrement séparables linéairement (voir Figure 2.3), nous assouplissons légèrement les contraintes pour 2.2 et 2.3 afin de tenir compte des points mal classés. Pour ce faire, nous

introduisons une variable positive ξ_i , i = 1,..., L :

$$x_i.w + b \geq +1 - \xi_i \text{ pour } y_i = +1 \tag{2.9}$$

$$x_i.w + b \leq -1 + \xi_i \text{ pour } y_i = -1 \tag{2.10}$$

$$\xi_i \geq 0 \ \forall_i \tag{2.11}$$

qui peuvent être combinés en :

$$y_i(x_i.w + b) - 1 + \xi_i \text{ pour } \xi_i \geq 0 \tag{2.12}$$

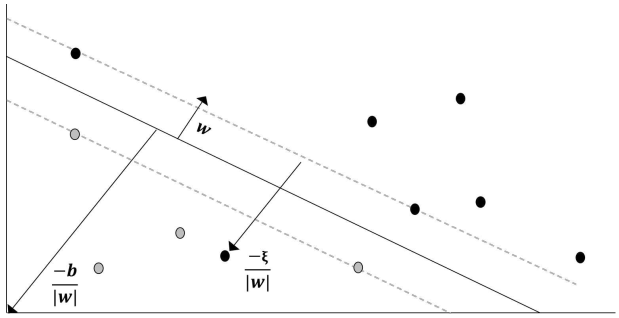


Figure 2.3 – Hyperplan passant par deux classes non-linéairement séparables.

Les points de données situés du côté incorrect de la limite de la marge sont pénalisés en fonction de la distance qui les sépare de cette limite. Comme nous essayons de réduire le nombre de classifications erronées, une façon judicieuse d'adapter notre fonction objective 2.8 consiste à trouver :

$$min\frac{1}{2}||w||^2 + C\sum_{i=1}^{L}\xi_i \text{ tel que } x_i.w + b \leq -1 + \xi_i \geq 0 \ \forall_i \tag{2.13}$$

où le paramètre C contrôle le compromis entre la pénalité de la variable d'étalement et la taille de la marge.

2.3.1.2 Les machines à vecteurs de support non linéaires

Lorsque nous appliquons le SVM à des données linéairement séparables, nous commençons par en créer une matrice H à partir du produit de points des variables d'entrée :

$$H_{ij} = y_i y_j (x_i.x_j) = x_i^T x_j \tag{2.14}$$

k(x_i,x_j) est une fonction noyau k(x_i,x_j) = ${x_i}^T$x$_j$ connue sous le nom de noyau linéaire. La fonction noyau a pour but de transformer les données de l'espace des caractéristiques d'origine en un espace de dimension supérieure. Elle prend les données d'entrée et calcule le produit intérieur entre les paires de points de données dans l'espace des caractéristiques. Cette transformation permet aux SVM implicitement d'opérer dans un espace de dimension supérieure de trouver une limite de décision qui sépare au maximum les points de données des différentes classes. C'est ce qu'on appelle "l'astuce du noyau". Parmi les fonctions noyaux couramment utilisées, on peut citer:

— Noyau linéaire : k(x_i,x_j) = $x_i^T x_j$.

— Noyau polynomiale : k(x_i,x_j) = $(\gamma x_i^T x_j + C)^d$ *avec* C$\geq 0, d \in$N*.

— Noyau sigmoïde : k(x_i,x_j) = Tanh $(x_i^T\ x_j + C)$ avec $C \geq 0$.

— Noyau RBF :k(x_i,x_j) = exp(-$\gamma||x_i$-$x_j||_2^2 + C)$ avec $\gamma > 0$.

C est une variable de pénalisation des points mal classés. Le gamma γ détermine le degré de courbure souhaité pour une frontière de décision.

Dans sa forme la plus simple, le SVM ne prend pas en charge la classification multi-classes de manière native. Il prend en charge la classification binaire et la séparation des points de données en deux classes. Cependant, dans le cas où les données se situent dans plusieurs classes, une classification en k classes est réalisée en construisant un ensemble de classifieurs binaires. Chaque classifieur binaire est entraîné à séparer une classe des autres et à les combiner en effectuant la classification multi-classes (en appliquant un système de vote).

2.3.2 Les arbres de décision

Les arbres de décision (Decision Trees : DT) sont construits en analysant un ensemble de données d'apprentissage pour lesquels les étiquettes de classe sont connues. Chaque arbre de décision se compose de nœuds, de branches et de feuilles (voir Figure 2.4). Chaque nœud de l'arbre agit comme un cas de test pour un attribut, et chaque nœud enfant descendant du nœud correspond aux réponses possibles au cas de test. Un échantillon est classé dans une classe en suivant le chemin qui mène du nœud racine au nœud feuille, en fonction des réponses qui s'appliquent à l'échantillon. Un élément est assigné à la classe qui a été associée à la feuille qu'il atteint.

Bien que les DTs ne soient pas l'algorithme le plus compétitif, il est souvent choisi parmi tant d'autres lorsqu'il s'agit de construire un modèle avec des décisions de classification

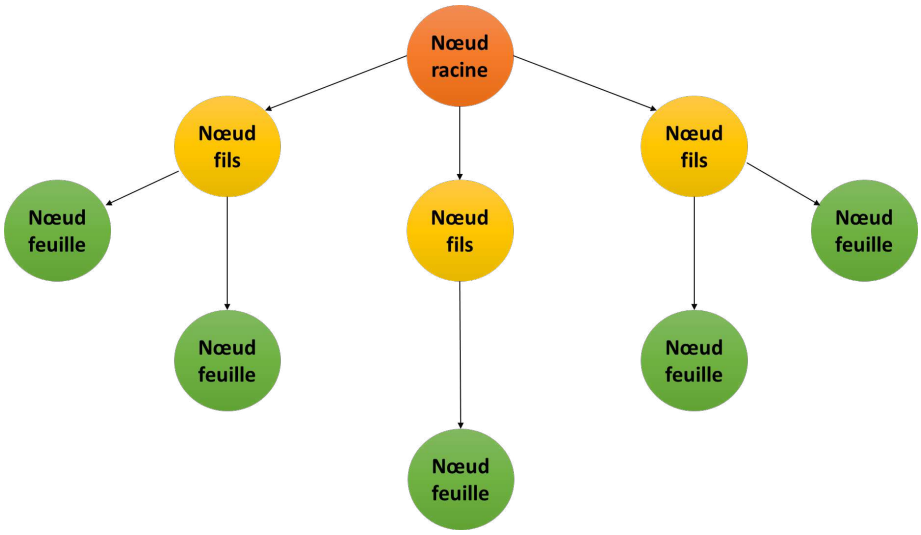


Figure 2.4 – Structure d'arbre de décision.

logiques. Il offre la possibilité d'identifier clairement ce qui a permis à l'algorithme de prédire un résultat spécifique. Cependant, une préoccupation majeure dans la construction des DTs est le problème de la variance élevée, qui se traduit généralement par une faible précision, ce qui rend l'algorithme sujet aux erreurs.

Un moyen très efficace de contourner ce problème est d'utiliser les DTs dans le contexte des méthodes d'ensemble. Les méthodes d'ensemble basées sur les DTs cherchent à améliorer les performances prédictives en combinant les résultats des multiples arbres qui les composent. Cependant, il est nécessaire de prendre en compte que les arbres individuels doivent être précis et différents les uns des autres pour obtenir un classifieur plus stable et plus robuste [115]. Cela peut se faire en utilisant le processus de randomisation (répartition aléatoire) qui aide à réduire la corrélation, permettant aux arbres de croître avec une plus grande diversité.

2.3.3 Les arbres extrêmement aléatoires

Les arbres extrêmement aléatoires (Extra-Trees : ET) sont une approche consistant à former un ensemble d'arbre de décision totalement aléatoires et indépendants. Ces arbres sont construits en utilisant des sous-ensembles aléatoires de caractéristiques et effectuent des prédictions sur des échantillons en se basant sur une séquence de règles définies dans une structure de type forêt. Les étapes de mise en œuvre des ET sont résumées ci-dessous : Étant donné un ensemble de données d'apprentissage x= x_1,x_2,. . . ,x_i, où un échantillon

x_1= f_1,f_2,. . . ,f_D est un vecteur à D dimensions avec f_j comme caractéristique et $j \in$ 1,2,. . . ,D. Les trois paramètres importants requis pour ET sont : le nombre d'arbres M, le nombre d'attributs K sélectionnés au hasard à chaque nœud, et la taille minimale de l'échantillon n_{min}, pour diviser un nœud.

1. En commençant par le nœud racine, le DT utilise l'ensemble des échantillons d'apprentissage pour se développer de manière descendante.
2. Dans chaque nœud interne, le DT sélectionne au hasard K caractéristiques f_1,f_2,. . . ,f_k, parmi tous les candidats. Pour chaque caractéristique k dans le sous-ensemble, ses valeurs maximales et minimales, f_{kmax} et f_{kmin} sont calculées et la valeur de division optimale (point de coupure) f_k avec des capacités maximales de réduction de la variance est sélectionnée dans l'intervalle [f_{kmin}, f_{kmax}]. En détail, l'entropie est utilisée comme fonction de score, où la meilleure division est choisie par la caractéristique ayant le moins d'entropie et est maintenue constante pendant la croissance de l'arbre. L'entropie est calculée à l'aide de la formule suivante:

$$Entropie = -\sum_t p_t \; log_2 p_t \tag{2.15}$$

   où p_t est la probabilité de la classe t.
3. Itérativement, les sous-ensembles sont divisés et les arbres sont étendus jusqu'à l'obtention des nœuds purs en termes de sorties ou qu'un nombre minimum d'échantillons d'apprentissage nécessaires pour la division (n_{min}) soit atteint. Cela met fin au processus de partitionnement et crée une feuille qui prédit l'étiquette de la classe.
4. Les étapes (1), (2) et (3) sont répétées M fois, et un modèle arbres extrêmement aléatoires composé de M DT indépendants est généré.
5. Enfin, en agrégeant les prédictions des M arbres, le résultat final de la classification est obtenu par un vote majoritaire de chaque classe aux nœuds feuilles.

Bien que les caractéristiques créées à la main aient bien fonctionné pour la RGM et AH, elles présentent un inconvénient de taille, à savoir qu'elles sont spécifiques à un domaine. Un ensemble différent de caractéristiques doit être défini pour chaque type de données d'entrée différents, le domaine temporel et le domaine fréquentiel. De plus, il n'est pas toujours évident que de telles caractéristiques soient susceptibles de fonctionner le mieux. Elles manquent encore de capacité de généralisation et leurs performances varient en fonction des caractéristiques sélectionnées. Le choix des caractéristiques se fait généralement

par le biais d'une évaluation empirique de différentes combinaisons de caractéristiques ou à l'aide d'algorithmes de sélection de caractéristiques.

2.4 Apprentissage profond

Les approches automatiques basées sur l'apprentissage profond (Deep Learning : DL), permettent de surmonter les restrictions des approches conventuelles. Elles utilisent des modèles dont leur structures complexes impliquent de multiples couches empilées. Elles leur permettent de réaliser de nombreuses transformations sur les données, et effectuer l'apprentissage de caractéristiques abstraites à plusieurs niveaux. Cet apprentissage englobe un ensemble de méthodes qui permettent à la machine de traiter les données sous forme brute et de les transformer automatiquement en une représentation appropriée nécessaire à la classification. C'est ce que nous appelons des extracteurs de caractéristiques entraînables (voir Figure 2.5). Les algorithmes d'apprentissage profond largement utilisés pour la RGM et des AH comprennent les CNNs, les réseaux de neurones récurrent (Recurrent Neural Network : RNN) et les auto-encodeurs (AutoENcoder : AEN).

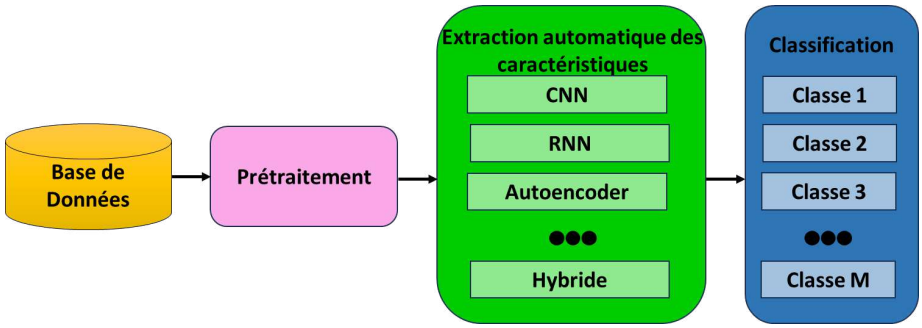


Figure 2.5 – Reconnaissance d'actions/gestes par apprentissage profond.

Le tableau 2.2 résume les algorithmes DL en citant les avantages et inconvénient de chacun.

Tableau 2.2 – Comparaison entre les algorithmes d'apprentissage profond.

Modèle	**Déscription**	**Avantages**	**Inconvénients**
CNN	Capturer la relation spatiale par de multiples couches convolutionnelles. Souvent utilisé comme un excellent extracteur de caractéristiques localisées.	Extraction automatique des caractéristiques. Invariant à l'orientation des données des capteurs et les changements dans les détails d'une activité.	Nécessite un grand ensemble de données et le paramétrage des hyperparamètres pour atteindre l'optimum. Peut être difficile à embarquer.
RNN	L'exploration de la relation temporelle dans les données, des variantes sont souvent utilisées, comme les LSTM.	Utiliser pour modéliser les dépendances temporelles qui existent entre les données.	Apprentissage difficile liés à la des cente du gradient (explosion et disparition) ainsi que la mise à jours des différents paramètres (cas du LSTM).
AEN	Un réseau de neurones à action directe qui apprend des caractéristiques profondes de manière non supervisée.	Denoising : Robuste dans le cas de données corrompues. Sparse : produit des caractéristiques plus distinctes les unes des autres. Contractive : Rend les caractéristiques invariantes aux changements en réduisant les dimensions de leurs espaces.	Denosing et sparse : Temps de calculs élevé. Contractive : difficile à optimiser.
Modèle hybride	La combinaison de certains modèles profonds, en s'appuyant sur la force de chaque modèle pour obtenir de meilleures performances.	Meilleures performances. Converge rapidement.	Modèle assez complexe nécessite de l'espace mémoire élevé.

2.5 Les algorithmes d'apprentissage profond

2.5.1 Les réseaux de neurones convolutifs

Les CNNs sont à ce jour les modèles les plus performants et les plus répandus. Ils sont conçus en particulier pour traiter des données avec une structure spatiale, telles que des images. Comme leur nom l'indique, les CNNs utilisent les opérations de convolution pour l'extraction des primitives. La couche de convolution permet de détecter la présence de caractéristiques dans les images. Elle se compose d'un ensemble de filtre, où chaque filtre est une matrice d'entiers. Outre le nombre de filtre, une couche convolutive possède trois autres hyperparamètres :

- La taille du filtre k définit la taille du champ réceptif.
- Le stride s représente le nombre de pixels par lesquels le champ réceptif glisse après chaque opération.
- L'opération de remplissage (padding: p) augmente la hauteur et la largeur de l'image d'entrée en ajoutant des zéros aux côtés gauche/droit ou haut/bas.

La couche de convolution reçoit en son entrée une image découpée en petites zones appelées champ récepteurs. Ce découpage permet l'extraction de pixels corrélés localement. Les filtres de convolution sont glissés sur l'image d'entrée et calculent le produit de convolution entre le filtre et chaque portion de l'image balayée. A chaque convolution on obtient à la sortie une carte de caractéristique. La Figure 2.6 présente un exemple de couche convolutive. Plusieurs convolutions peuvent être calculées en parallèle. Les dimensions de la sortie sont déterminées par les hyperparamètres mentionnés ci-dessus. Si une couche convolutive possède d_l filtre, la taille de la sortie : $Sortie = h_{sortie} \times w_{sortie} \times c_{sortie}$ peut être calculée comme suit :

$$h_{out} = \frac{h_{in} - k + 2p}{s + 1} \tag{2.16}$$

$$w_{out} = \frac{w_{in} - k + 2p}{s + 1} \tag{2.17}$$

$$w_{out} = d_l \tag{2.18}$$

où h_{in} et h_{out} la hauteur de l'image d'entrée et de sortie, respectivement, w_{in} et w_{out} la largeur de l'image d'entrée et de sortie, respectivement et c_{out} le nombre de carte de caractéristiques obtenue après l'opération de convolution.

Les couches convolutives multiples permettent au CNN d'extraire hiérarchiquement

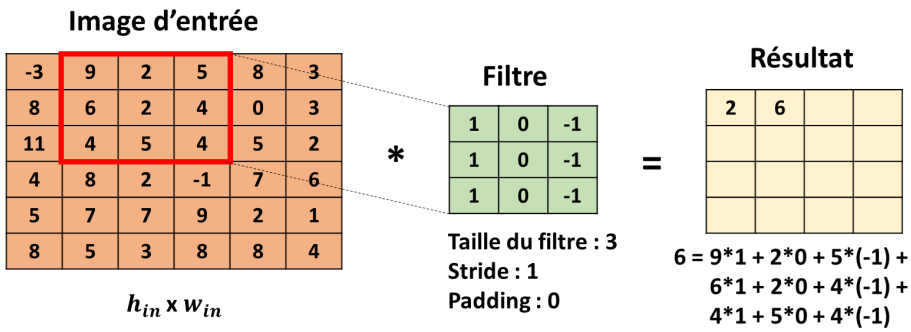


Figure 2.6 – Exemple d'opération de convolution.

des caractéristiques spatiales de niveau supérieur à partir des caractéristiques de niveau inférieur, ce qui élimine carrément la procédure d'extraction manuelle des caractéristiques. Grâce à l'excellente capacité d'apprentissage des caractéristiques profondes, le CNN est souvent utilisé comme extracteur automatique de caractéristiques pour une variété de tâches. Après les couches convolutives, les couches de mise en commun (Pooling) et les couches entièrement connectées sont généralement utilisées.

La couche de mise en commun est une forme de sous-échantillonnage des cartes de caractéristiques. Elle vise à réduire leurs dimensionnalités tout en préservant les caractéristiques les plus importantes. Il s'agit d'appliquer un filtre par fenêtre de taille fixe glissante sur les données d'entrée. Il existe trois techniques de redimensionnement utilisées dans les modèles CNN, comme illustré sur Figure 2.7

— La mise en commun maximale : appelé maximum pooling, qui consiste à récupérer la valeur maximale dans la fenêtre d'observation.

— La mise en commun moyenne : appelé averagepooling, qui consiste à calculer la moyenne des valeurs dans la fenêtre d'observation.

— La mise en commun par somme des valeurs : appelé sumpooling, qui consiste à calculer la somme des valeurs dans la fenêtre d'observation.

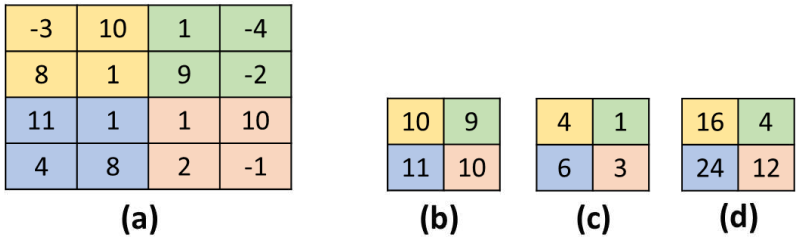


Figure 2.7 – Exemple d'opération de mise en commun. (a) image d'entrée, mise en commun (b) maximale, (c) moyenne, (d) par somme des valeurs.

2.5.2 Les réseaux de neurones récurrents

Les RNNs ont une topologie de connexions quelconque, comprenant notamment des boucles dont le sens de propagation du flux de données reste aléatoire. Il s'agit de réseaux de neurones avec une contre réaction (voir Figure 2.8). Les RNNs ont mis en lumière la modélisation des séquences temporelles en raison de leurs capacités à extraire des informations temporelles. Du point de vue de la structure du réseau, les RNNs se souviennent des informations précédentes et les utilisent pour influencer la sortie des nœuds suivants. Le RNN comporte des boucles dans la dimension temporelle. Comme le montre la Figure 2.8, à l'instant t, le RNN prend la séquence de données d'entrée x_t et la concatène avec l'état caché précédent h_{t-1}. Ensuite, l'état caché h_t et le vecteur de prédiction y_t sont générés par une combinaison de transformations linéaires et non linéaires. La boucle se poursuit et l'état caché se propage jusqu'à ce que la dernière séquence de données soit traitée. Ainsi, pour faire une prédiction à un moment donné, le RNN utilise non seulement l'entrée actuelle mais aussi l'état caché précédent qui contient les informations du passé. La description mathématique du processus est la suivante :

$$h_t = tanh(w_{hh}.\ h_{(t-1)} + w_{xh}.\ x_t + b_h) \tag{2.19}$$

— w_{hh} la matrice de poids pour l'état caché précédent h_{t-1}.
— w_{xh} est la matrice de poids pour l'état actuel x_t.
— x_t est l'entrée au même instant t.
— h_t est l'état caché au moment t.
— h_{t-1} est l'état caché au moment $t-1$.
— b_h biais de mise à jour.
— tanh est une fonction d'activation.

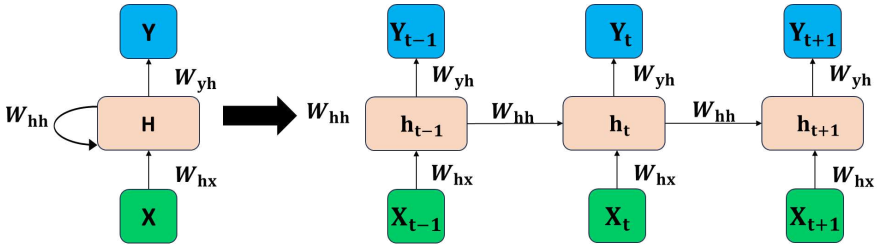


Figure 2.8 – Structure d'un réseau de neurones récurent.

Un inconvénient majeur du RNN est la non capture des dépendances à long terme en

raison du problème de l'évanouissement du gradient. Une variante du RNN efficace pour gérer les corrélations temporelles à long terme est le LSTM.

2.5.2.1 Les cellules de longue mémoire à court terme

Le LSTM a un flux de contrôle similaire à celui du RNN. Il traite les données d'entrée de manière séquentielle et propage les informations le long de la dimension temporelle. Pour les différences, le LSTM introduit trois types de portes, à savoir la porte d'oubli, la porte de sortie et la porte d'entrée. En outre, le LSTM utilise un état cellulaire distinct ct pour transporter les informations relatives dans le temps. La Figure 2.9 présente l'architecture du LSTM.

— La porte d'oubli f_t est utilisée pour ignorer les informations non utilisées auparavant et décider quelles données seront transmises à la porte d'entrée $t-1$ se réfèrant au temps précédent. Les informations provenant de l'entrée actuelle x_t et de l'état caché h_{t-1} sont transmises à la porte et multipliées. Le résultat est transmis à une fonction d'activation qui donne une sortie binaire. Si, pour un état particulier de la cellule, la sortie est 0, l'information est oubliée et si la sortie est 1, l'information est conservée pour un usage ultérieur.

— La porte d'entrée i_t décide quelle information est importante et l'utilise pour mettre à jour l'état de la cellule au moment t. Tout d'abord, l'information est normalisée à l'aide de la fonction sigmoïde σ et filtre les valeurs à stocker de manière similaire à la porte d'oubli en utilisant les entrées h_{t-1}. Ensuite, un vecteur C est créé à l'aide de la fonction tanh qui donne une sortie de -1 à +1 contenant toutes les valeurs possibles de h_{t-1} et x_t. Les valeurs de sortie générées par les fonctions d'activation sont prêtes à être multipliées point par point pour obtenir les informations utiles.

— La porte de sortie o_t spécifie la valeur du prochain état caché. Les valeurs de l'état actuel x_t et de l'état caché précédent h_{t-1} sont introduites dans la troisième fonction sigmoïde . Ensuite, le nouvel état de la cellule généré à partir de l'état de la cellule est passé dans la fonction tanh. Ces deux sorties sont multipliées point par point. En fonction de la valeur finale, le réseau décide de l'information quel état caché doit contenir. Cet état caché est utilisé pour la prédiction.

Les formules de calcul pour la cellule LSTM sont illustrées par les équations ci-dessous :

$$f_t = \sigma g(f)x_t + U(f)h_{t-1} + b(f) \tag{2.20}$$

$$i_t = \sigma g(i)x_t + U(i)h_{t-1} + b(i) \tag{2.21}$$

$$o_t = \sigma g(o)x_t + U(o)h_{t-1} + b(o) \tag{2.22}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \sigma c(W(c)x_t + b(c)) \tag{2.23}$$

$$h_t = o_t \odot \sigma hc(t) \tag{2.24}$$

x_t est le vecteur d'entrée, c_t est le vecteur d'état de la cellule, g désigne la fonction sigmoïde, (i) représente la matrice de poids et b est le vecteur de biais. En outre, $\odot$ désigne la multiplication élément par élément. La sortie de σ, crée f_t, i_t et o_t qui sont respectivement la sortie de la porte d'oubli, la porte d'entrée et la porte de sortie. Pour créer le vecteur d'état de la cellule, le vecteur d'entrée est d'abord multiplié par la métrique de poids $W(c)$ et additionné avec le biais $b(c)$, puis la fonction d'activation $tanh$ (σc) est appliquée. L'équation 3.23 réalise plusieurs opérations pour calculer différents vecteurs dans le contexte du modèle. Tout d'abord, elle effectue la multiplication par élément ($\odot$) entre la porte d'entrée et la sortie de σh, ce qui donne lieu à l'addition de ce résultat avec le produit par élément de la porte d'oubli et le vecteur d'état de la cellule précédente c_{t-1}, créant ainsi le nouveau vecteur d'état de la cellule. Enfin, la même équation permet d'obtenir le vecteur d'état caché h_t en réalisant la multiplication par élément de la porte de sortie et la sortie de la fonction d'activation appliquée à l'état de la cellule h_t.

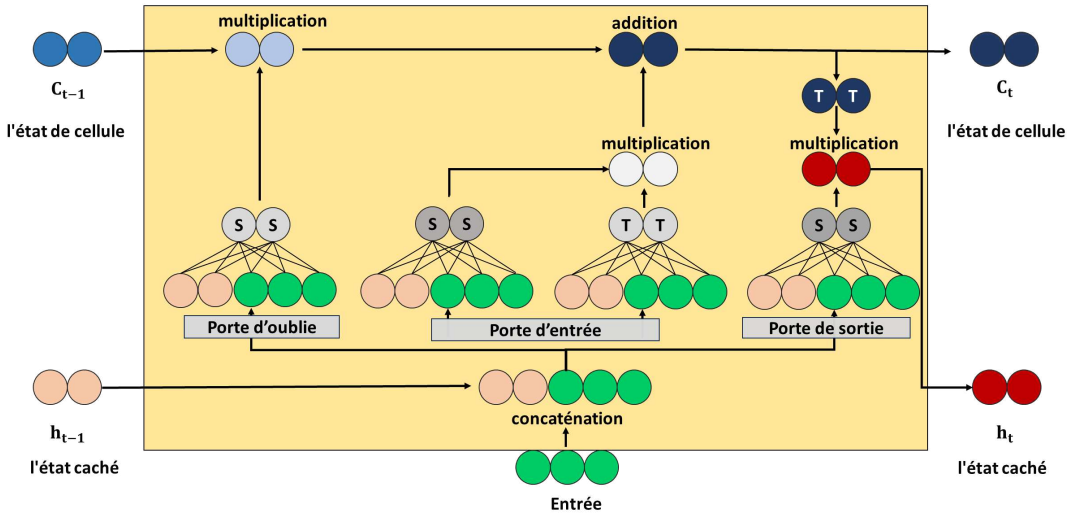


Figure 2.9 – Structure du LSTM.

2.5.3 Auto-encodeur

Un AEN est un type particulier de réseau neuronal à action directe. Il vise à reconstruire des données fournies à son entrée par le biais de plusieurs cycles d'encodage-décodage de manière non supervisée. Un AEN basique comporte une branche codeur, une branche décodeur et des couches cachées composées de couches denses ou convolutives (voir Figure 2.10). La taille des couches cachées de la branche codeur diminue progressivement jusqu'à assurer une représentation de moindre dimension des données d'origine (Bottleneck), puis commence à augmenter symétriquement pour atteindre la taille d'entrée originale à la sortie. Chaque couche cachée du codeur extrait une hiérarchie de caractéristiques et les projette dans un espace de dimension inférieure. Au niveau du Bottleneck, on obtient une représentation compressée de l'entrée, appelée caractéristique latente. Le décodeur reconstruit ensuite l'entrée en décodant les caractéristiques latentes. L'apprentissage des AEN se fait de manière non-supervisé en minimisant l'erreur de reconstruction.

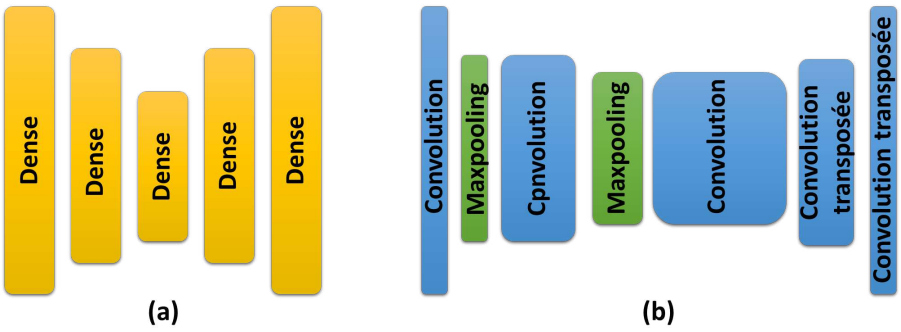


Figure 2.10 – (a) autoencodeur, (b) autoencodeur convolutif.

2.5.3.1 Encodeur convolutif

Un auto-encodeur convolutif (Convolutional Auto Encoder: CAE) dispose d'une architecture similaire à celle d'un AEN basique, bien qu'il ne soit pas symétrique en termes d'opérations effectuées dans les branches du codeur et du décodeur. Au niveau de la branche codeur, les couches cachées sont des couches convolutives et des couches mise en commun, comme dans les CNN standard. Au niveau de la branche décodeur, des opérations de suréchantillonnage sont utilisées pour restaurer les dimensions originales de l'entrée. Pour ce faire, il utilise une couche convolutive transposée, qui effectue un suréchantillonnage et une convolution simultanément (voir Figure 2.11). La couche convolutive transposée effectue l'opération de convolution dans la direction opposée. L'entrée se

fait glisser sur le filtre et effectue une multiplication et une sommation par élément. Il en résulte une sortie plus grande que l'entrée, et la taille de la sortie peut être contrôlée par les paramètres stride et padding de la couche. Une couche convolutive transposée possède 4 hyperparamètres : le nombre de filtre c, la taille du filtre k, le stride s et la taille de remplissage p. L'entrée de couche convolutive est une carte de caractéristiques de taille I_hxI_w, où I_h et I_w sont la hauteur et la largeur de l'entrée, et la taille du filtre est K_hxK_w, où K_h et K_w sont la hauteur et la largeur du filtre. La convolution transposée effectue les étapes suivantes : tout d'abord, $s-1$ zéros sont insérés entre deux pixels voisins, ce qui suréchantillonne l'entrée d'un facteur s. Ensuite, un remplissage suivit d'une convolution normale en utilisant un pas fixe de 1 sont effectués. La sortie de la couche convolutive transposée est alors la suivante :

$$O_h = ((I_{h-1})s_h + K_h - 2p) \tag{2.25}$$

$$O_w = ((I_{w-1})s_w + K_h - 2p) \tag{2.26}$$

où O_h et O_w sont la hauteur et la largeur de la sortie.

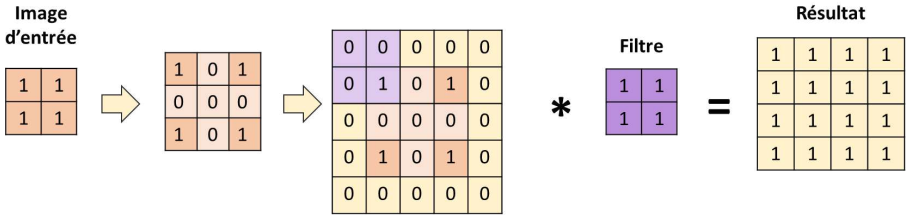


Figure 2.11 – Exemple de convolution transposée.

2.5.4 Modèle hybride profond

Chaque modèle possède ces propres avantages et inconvénients et n'est pas performent pour toutes les tâches. Les modèles profonds hybrides intègrent plusieurs réseaux ensemble et tirent parti de tous ces réseaux. Cette coopération s'appuie sur les points forts de chaque modèle afin d'obtenir de meilleures performances. Jusqu'à présent, la combinaison des CNN et RNN est la plus répandue pour la RGM et des AH. Cette combinaison permet l'abstraction de différentes caractéristiques d'un domaine : Le CNN capture les caractéristiques spatiales tandis que le RNN capture les caractéristiques temporelles. Multiples travaux ont démontré que la combinaison du CNN et du RNN tend à renforcer le pouvoir

de RGM et des AH qui varient dans le temps et l'espace [99,116,117]. En outre, l'AEN est souvent combiné avec le CNN ou le RNN en raison de sa capacité à extraire de manière non supervisée des caractéristiques de haute dimension [118].

2.6 Etat de l'art

La section suivante fournit des détails concernant l'état de l'art dans le domaine de la RGM et des AH en utilisant la technologie IR-UWB en conjonction avec des méthodes ML/DL qui ont été appliquées à des fins de classification.

2.6.1 La reconnaissance des gestes de la main

Les gestes qui s'expriment par le mouvement de diverses parties du corps humain sont considérés comme un aspect du langage corporel et sont naturellement utilisés pour transmettre des informations entre humains. Ce langage corporel considéré comme échange non verbale constitue jusqu'à deux tiers de toute la communication entre les humains [119] et se compose de l'expression faciale, de la posture, des GM et de plusieurs autres mouvements du corps comme indiquer sur la Figure 2.12. Toutefois, les GM sont les

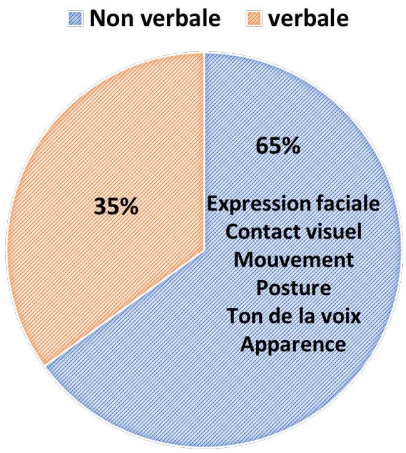


Figure 2.12 – Répartition des caractéristiques verbales et non verbales [116].

plus couramment utilisés. Ils offrent un moyen plus expressif et naturel d'interagir avec l'environnement grâce à une variété de mouvements, par exemple en utilisant la paume de la main, en changeant la position des doigts ou même en percevant une forme subtile de la main. D'après la littérature, il est indiqué que la main est la plus largement utilisée

par rapport à d'autres parties du corps [53]. Elle est considérée la mieux adaptée à l'IHM, comme le montre la Figure 2.13.

Il est important de faire une distinction claire entre les gestes et les poses. Une pose correspond à un geste statique (micro geste) qui fait référence à la forme stable de la main, et reste presque inchangé dans le temps. En revanche, le geste (macro-geste) est un acte dynamique formé par une séquence consécutive de poses pendant une courte période de temps.

Outre cette différenciation, l'interprétation des gestes statiques ainsi que dynamiques reste une tache assez compliquée en raison de la variété des concepts associés à un même geste. Ce processus est appelé reconnaissance des gestes.

Le terme de reconnaissance des gestes est apparu pour la première fois au milieu des années 1970 et a été proposé par Myron Krueger comme une méthode d'interaction entre les humains et les machines [120]. L'objectif principal derrière cette proposition est l'élimination d'interaction physique par la génération d'interfaces pratiques et hautement adaptables entre les appareils et les utilisateurs. Aujourd'hui, en raison des progrès rapides du monde numérique et de l'émergence de matériels puissants, la RGM est devenu sujet majeur dans les domaines de l'IHM. La RGM joue un rôle clé et fournit un support adéquat pour de nombreuses applications de santé telles que le diagnostic de plusieurs troubles neurologiques, la rééducation de patients souffrant de déficiences motrices, la surveillance à distance de personnes âgées, entre autres.

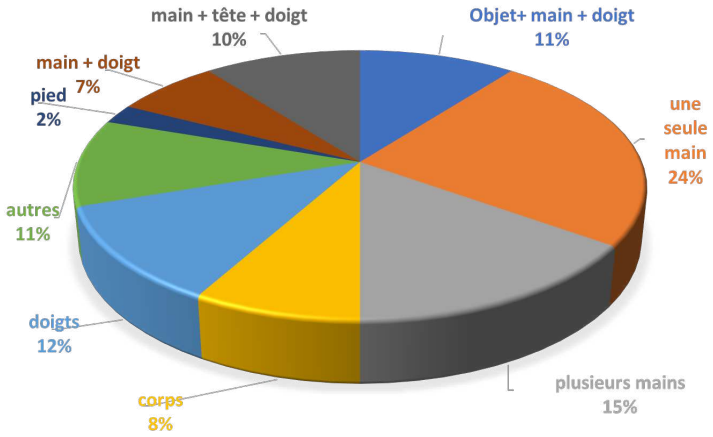


Figure 2.13 – Utilisation de différentes parties du corps humain pour créer une IHM [51].

2.6.1.1 Apprentissage machine pour la reconnaissance des gestes de la main

Dans l'étude de Ren *et al.* [71], six gestes différents ont été capturés et un taux de reconnaissance global de 86% a été obtenu. L'approche proposée a consisté en une simple technique de seuillage 1D basée sur la différence de position finale pour différencier les gestes. Bien que cette technique ait fonctionné pour des gestes avec des directions définies de la main, il se peut qu'elle ne fournît pas de solution pour des gestes complexes.

Park *et al.* [62] ont réalisé la classification de six GM en exploitant des caractéristiques extraites du signal radar brute et traité par Analyse en Composantes Principales (ACP). Un total de seize caractéristiques extraites dont huit du signal brute et huit du signal filtré, ont été utilisées comme données d'entrée pour entraîner un réseau neuronal à l'aide de la méthode du gradient conjugué gradué et de l'algorithme de rétropropagation. Le modèle a été capable d'atteindre une précision de 100 % et de différencier les gestes. Cependant, leur approche n'a été validé que sur des gestes simples.

Khan et Cho [75] ont effectué une classification de six gestes dont quatre stationnaires et deux non-stationnaires à l'aide d'un réseau neuronal piloté par trois caractéristiques, à savoir les variations de distance, d'amplitude et de surface. Bien que cette approche ait bien fonctionnée sur des gestes bien définis et fournies un taux d'erreur faible, elle pourrait ne pas offrir les mêmes performances pour les mini gestes comme les mouvements des doigts.

Khan *et al.* [76] en mis en œuvre un système de RGM à l'intérieur d'une voiture à l'aide d'un IR-UWB. Les cinq gestes utilisés dans cette étude consistent en trois petits gestes où seuls les doigts sont déplacés, et deux grands gestes où la main entière est déplacée. Un geste vide supplémentaire a été également ajouté. Ils ont obtenu une précision moyenne de 98% en utilisant seulement trois caractéristiques, à savoir : la variance de la fonction de densité de probabilité de l'histogramme d'amplitude, la variation du temps d'arrivée et la fréquence du signal réfléchi comme paramètre d'entrée à l'algorithme K-means.

Khan *et al.* [121] ont proposé un algorithme de détection des gestes basé sur le comptage des doigts. Pour ce faire, cinq gestes ont été implémentés correspondant au nombre de doigts de la main qui changent de position de bas en haut de manière séquentielle pendant le geste. Leur approche est basée sur l'utilisation de la magnitude du signal extraite à chaque petite portion du geste comme entrée pour l'entraînement d'un FCNN. Un taux de reconnaissance de 97.7% a été obtenu.

Ghaffar *et al.* [122] ont mis en œuvre un panneau de menu numérique basé sur le

pointage manuel. Les gestes, à savoir le clic simple (SC), le clic double (DC), le balayage droit (RS) et le balayage gauche (LS), ont été exécutés dans un plan 2D divisé en plusieurs grilles et enregistrés à l'aide de quatre IR-UWB. Selon les résultats rapportés, une précision de classification de 96 % a été atteinte en utilisant un descripteur d'histogramme de gradient orienté basé sur un classifieur SVM multi classes. Cependant, il a été remarqué une confusion entre les gestes effectués dans des grilles proches.

Li *et al.* [123] ont développé un système de reconnaissance du langage des signes et des GM basé sur des caractéristiques de densité de distribution cumulée extraites à partir de spectrogrammes radar. Bien que leur système ait montré de bons résultats contre la variation de la distance radiale par rapport au radar, une précision d'uniquement de 82,4% a été obtenue en utilisant l'algorithme k-Nearest Neighbors (kNN). De plus, il a été nécessaire de tester plusieurs combinaisons de caractéristiques pour trouver celle qui donne les meilleures performances.

2.6.1.2 Apprentissage profond pour la reconnaissance des gestes de la main

Kim *et al.* [74] ont appliqué un 1D-CNN aux signaux du domaine temporel (Amplitude-Temps) générer par un IR-UWB pour classer six gestes dynamiques de la langue des signes américaine (S,E,V,W,B et C). Leur modèle a obtenu un taux de précision supérieure à 90%.

Ahmed *et al.* [101] ont proposé une méthode basée sur 2D-CNN pour reconnaître les gestes dynamiques de la main. Les douze gestes à savoir : un balayage gauche-droite (LR), un balayage droite-gauche (RL), un balayage haut-bas (UD), un balayage bas-haut (DU), un balayage diagonal (diag)-LR-UD, un balayage diag-LR-DU, un balayage diag-RL-UD, un balayage diag-RL-DU dans le sens des aiguilles d'une montre, dans le sens inverse des aiguilles d'une montre, un geste d'enfoncement et un geste vide sont recueillies à l'aide d'un IR-UWB. Les données radar sous forme de matrices de temps lent-temps rapide sont converties en images et utilisées comme entrée pour l'apprentissage du 2D-CNN à quatre couches. Les performances de RGM de la méthode basée sur 2D-CNN a montré une précision de 94%.

Li *et al.* [124] ont utilisé des matrices de Fréquence-Distance pour représenter sept GM, à savoir (1) claquer les doigts (2) retourner la paume vers la gauche (3) retourner la paume vers la droite (4) la rotation vers le haut (5) la rotation vers le bas (6) la rotation de la paume gauche (7) la rotation de la paume droite. Trois algorithmes ont été mis en œuvre, notamment : Visual Geometry Group-16 (VGG-16), MobileNetV2, et ShuffleNet V2 pour

classer les données. ShuffleNet V2 a montré une précision de 98%, surpassant les approches VGG-16 et MobileNetV2 avec une précision de 94,28% et 95,84%, respectivement.

Park *et al.* [113] ont proposé de combiner la FFT avec un CNN pour classifier cinq gestes de la langue des signes américaine à savoir : "Tout fait", "Manger", "Plus", "Désolé" et "Merci". L'idée est de convertir les données du domaine temporel vers le domaine fréquentiel et les utilisés pour entrainer et tester différents modèles à savoir : CNN en double parallèle, CNN en série à deux étages, GoogLeNet, ResNet-50, VGG-19 et AlexNet. La précision la plus élevée est obtenue par GoogLeNet et est égale à 97.97%. Les résultats expérimentaux montrent que l'approche proposée est généralisable et permet d'augmenter la précision de classification.

Après l'énorme succès des CNNs, les RNNs ont montré un grand potentiel pour la RGM à partir des signaux IR-UWB. Park *et al.* [125] ont fourni un matériel radar comprenant un module d'échantillonnage à haute vitesse capable d'échantillonner les signaux UWB du domaine temporel. Le système a été conçu pour reconnaître les gestes statiques et dynamiques. Un algorithme de reconnaissance basé sur LSTM a été employé pour la classification de six gestes dynamiques de la main, à savoir l'ouverture, le serrement, l'avance, le recul, le balancement et le pointage, et a atteint une précision de 90,5

Noori *et al.* [100] ont introduit l'ACP et l'Analyse Discriminante Linéaire (ADL) pour l'extraction des caractéristiques des gestes dynamiques de la main. Les auteurs ont utilisé une architecture LSTM qui a montré des performances supérieures en atteignant une précision de 97%.

Un autre paradigme de conception dans le domaine de la RGM consiste à développer des modèles hybrides. Ces derniers impliquent l'utilisation d'une combinaison de couches CNN et RNN afin de tirer parti des avantages qu'offre chacune des méthodes, à savoir une extraction automatique des caractéristiques et la modélisation des dépendances temporelles.

Skaria *et al.* [126] ont présenté les données d'un IR-UWB sous la forme d'une séquence de trames de Fréquence-Distance et alimenté en entrée d'un modèle pour la classification de quatorze GM. Quatre classifieurs ont été utilisés à savoir 3D-CNN-Fully Connected Neural Network (FCNN), 3D-CNN-kNN, 3D-CNN–SVM, 2D-CNN-LSTM et obtenue une précision de 93,33%, 92,02%, 94,08% et 96,15%.

Hendy *et al.* [117] ont exploité deux types de représentation de données à savoir les images 2D et 3D de Fréquence-Distance pour classifier les gestes d'écriture dans l'air des chiffres de 0 à 9. Les auteurs ont notamment introduit cinq modèles de réseaux neuronaux,

à savoir, FCNN, 2D-CNN, 3D-CNN, 2D-CNN-LSTM, et 3D-CNN-LSTM. Les résultats ont démontré que le modèle le plus performant est le 3D-CNN-LSTM avec une précision de reconnaissance de 98.5%.

2.6.1.3 Fusion de données pour la reconnaissance des gestes de la main

Skaria *et al.* [99] ont exploité les avantages de l'utilisation de deux capteurs pour une classification robuste des gestes, à savoir un IR-UWB et un capteur thermique. Les auteurs ont utilisé un CNN-LSTM et ont obtenu une précision de 99% pour quatorze GM.

Khan *et al.* [79] ont mis en œuvre un système basé sur de multiples IR-UWB, capable de reconnaître des chiffres et des caractères dessinés dans l'air. Ils ont utilisé un CNN à cinq couches pour classifier les gestes en se basant sur le modèle de trajectoire de l'image.

Ahmed et Cho [61] ont présenté un classifieur basé sur le modèle GoogLeNet pour la reconnaissance de huit GM à savoir : balayage gauche-droite (Left Right : LR-swipe), balayage droite-gauche (Right Left : RL-swipe), balayage haut-bas (Up Down : UD-swipe), balayage bas-haut (Down Up : DU-swipe), balayage diagonal de gauche à droite (diag-LR-UD-swipe), balayage diagonal de gauche à droite (diag-LR-DU swipe), rotation dans le sens des aiguilles d'une montre (ClockWise : CW-rotation) et rotation dans le sens inverse des aiguilles d'une montre (CounterClockWise :CCW-rotation). Différentes variations structurelles sur le bloc d'extraction des caractéristiques ont été réalisées pour former un algorithme CNN très profond. Les données Temps Lent-Temps Rapide acquises par deux IR-UWB, fusionnées, converties en images 3D-RVB sont utilisées pour l'apprentissage du modèle. Une précision de 95% a été obtenue.

Leem *et al.* [78] ont utilisé un réseau de capteurs composé de trois IR-UWB et un CNN pour classifier le mouvement de la main utilisé pour écrire les chiffres de 0 à 9 dans l'air. Tout d'abord, le mouvement de la main a été suivi dans un plan 2D générant une image du chiffre ciblé. Ensuite les données de la trajectoire suivie ont été utilisées comme entrée pour un CNN. La méthode proposée surpasse les approches conventionnelles et démontre sa robustesse aux changements d'orientation, de distance, de forme et de taille de la main.

Afin d'extraire pleinement les caractéristiques spatiales discriminantes locales et d'éviter la perte d'informations, Lihang *et al.* [127] ont proposés d'appliquer les encodeurs de motifs binaires locaux sur les données de trois IR-UWB pour extraire les informations spatiales locales. Parallèlement, le ShuffleNet multicouches avec convolution séparable en profondeur est utilisé pour exploiter progressivement les caractéristiques spatiales de haut

niveau et classifier les différents gestes. L'approche proposée a fourni des performances de reconnaissance et une efficacité plus satisfaisante avec une structure légère. Le système de reconnaissance des gestes a pu atteindre une précision prometteuse de 96,52% sur le réseau IR-UWB.

Sur la même base de données, Sharma *et al.* [128] ont suggéré de calculer différentes caractéristiques à partir des données Distance-Temps pour l'identification des GM. Les caractéristiques basées sur la fonction de densité moyenne, à savoir la densité de réflexion par unité de temps, la densité de réflexion par unité de distance et les moments mixtes de temps et de distance sont calculées dans l'ensemble de la région de distance temporelle, puis limitées à des plages de distance temporelle locales. Les caractéristiques calculées sont classées et transmises aux modèles de classification à savoir SVM, kNN et RF. La précision élevée de 97.41% pour le kNN obtenue justifie la primauté des caractéristiques et du système développé.

2.6.2 La reconnaissance des actions humaines

Les termes "action" et "activité" sont souvent utilisés de manière interchangeable dans la littérature [129, 130]. Nous entendons par "action" des mouvements simples généralement exécutés par une seule personne et qui durent généralement peu de temps, de l'ordre de quelques dizaines de secondes. Parmi les exemples d'actions, on peut citer se pencher, marcher, courir. D'autre part, par "activité", nous entendons la séquence complexe d'actions exécutées par plusieurs personnes susceptibles d'interagir les unes avec les autres de manière contraignante. Ces activités sont généralement caractérisées par des durées temporelles beaucoup plus longues. Par exemple deux personnes qui se serrent la main. Toutefois, une action peut être défini comme une activité simple. Cette catégorisation fournit un point de départ pour organiser les nombreuses approches qui ont été proposées pour résoudre le problème.

La majorité des travaux de recherches réalisés dans ce contexte en utilisant la technologie IR-UWB se sont focalisés sur la RAH d'une personne individuelle. Les études examinées peuvent être classées en deux grandes catégories : reconnaissance des actions de la vie quotidienne et des actions en temps réel (voir Figure 2.14). Les actions de la vie quotidienne sont classées en actions statiques, comme se tenir debout ou s'asseoir, et en actions dynamiques, comme marcher ou courir. Les études portant sur les actions en temps réel sont regroupées dans les catégories "soins de santé" et "surveillance".

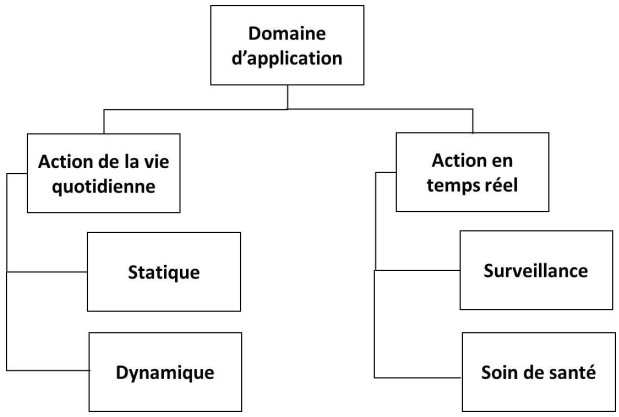


Figure 2.14 – Domaine de la reconnaissance des actions humaines.

2.6.2.1 Apprentissage machine pour la reconnaissance des actions humaines

Bryan *et al.* [102, 131] ont effectués la classification de plusieurs AH en exploitant la variation des signatures IR-UWB dû aux mouvements des différents parties du corps (torse, mains, pieds, bras). Différentes caractéristiques ont été extraites pour entrainer un SVM.

Dans l'étude menée par Ota *et al.* [132], un système d'assistance pour les gens âgés a été proposé. A partir du calcul du changement de puissance du signal IR-UWB reçu, le système arrive à reconnaitre divers types de mouvements au lit (assis, allongé) et mouvements à l'extérieur du lit mais à l'intérieur de la pièce (marcher, tomber). Un taux de reconnaissance de 95% a été obtenu.

Saho *et al.* [133] ont proposé un algorithme capable de classifier en temps réel les mouvements de différents piétons en utilisant des caractéristiques basées sur des images Fréquence-Temps et la vitesse radiale de déplacement des piétons. L'algorithme kNN a été utilisé pour la classification et une précision de 96% a été obtenue avec un temps de calcul de 0.55 secondes.

Le travail de Kiasari *et al.* [134] a porté sur l'utilisation de quarante caractéristiques extraites des dix premières ACP des échos radar obtenus avec un IR-UWB. Ces caractéristiques correspondent à la moyenne, la variance, l'asymétrie et l'aplatissement de chaque vecteur de composantes principales. Les auteurs ont utilisé un FCNN pour classifier les postures : assise, debout et allongé avec respectivement 83%, 86% et 80% de précision.

Dans l'étude de Diraco *et al.* [108], il a été démontré qu'avec un seul IR-UWB monostatique, des performances de reconnaissance satisfaisantes peuvent être obtenues. La détection des chutes en tant qu'anomalies a été effectuée en exploitant des caractéristiques

Fréquence-Temps et en utilisant une approche de clustering non supervisée.

Erol *et al.* [135] ce sont focalisés sur l'utilisation des informations de distance intégrées à un algorithme de détection de chute pour distinguer une chute réelle d'une position d'assise afin de réduire le taux de fausses alarmes. Il a été conclu qu'une chute réelle présente une portée deux fois plus étendue que celle d'une position assise.

C'est depuis ces deux études [108, 135] que Erol et Amin [96] ont proposé d'utiliser la représentation Fréquence-Distance contenant à la fois les informations de distance et de vitesse variant au cours du temps pour la reconnaissance des chutes. Les résultats ont montré que les caractéristiques extraites à partir de la représentation distance-fréquence permettent d'atteindre un taux de précision élevé de 99.6% et surpasse les approches basées sur les représentations Fréquence-Temps et Distance-Temps.

Mokhtari *et al.* [106] ont adopté une méthode de détection de chute basée sur les caractéristiques du temps d'arrivée qui modélise le mouvement et la vitesse de l'individu. L'algorithme RF a été utilisé pour la classification. Les résultats obtenus ont montré que dans le cas des approches supervisées, la dépendance de l'algorithme de classification aux données d'entraînement et de test spécifiques au sujet est très élevée. Cela engendre une dégradation des performances de reconnaissance en termes de précision. Motivé par les limitations des approches basées sur l'apprentissage supervisé, les mêmes auteurs proposent une nouvelle solution non supervisée pour la détection de chutes à l'aide de IR-UWB [136]. L'un des avantages de l'apprentissage non supervisé est qu'aucune exigence n'est faite pour fournir des données d'apprentissage étiquetées. Cela offre un avantage significatif dans les cas où, l'étiquetage des données est difficile, prend du temps, et coûteux.

L'approche proposée dans [136], est aussi précise que la méthode supervisée et est capable de détecter les chutes beaucoup plus rapidement.

Seifert *et al.* [137] se sont concentrés sur l'étude de l'influence des appareils d'aide à la marche telle qu'une canne sur les signaux radars rétro-diffusés. La reconnaissance de la marche assistée et non assistée a été effectuée sur la base des caractéristiques obtenues à partir du DCV.

Une approche similaire a été proposée par Baird *et al.* [138] utilisant le premier composant principal et un nombre de caractéristiques réduit égale à trente-trois caractéristiques. Ces dernières ont été extraites à partir des échos IR-UWB recueillis à trois distances différentes 3m, 4.5m et 6m. L'algorithme DT a permis d'atteindre la précision la plus élevée de 84.94%.

Seifert *et al.* [139] ont proposé une approche qui permet de classifier avec un degré de

fiabilité élevé les différents styles de marche à savoir la marche normale, pathologique et assistée par une canne en utilisant des caractéristiques extraites à partir de la représentation Fréquence-Temps des signatures radar. Les résultats obtenus démontrent la faisabilité d'utiliser le IR-UWB comme outil de diagnostic et d'analyse médicale ambulatoire de la marche.

Ding *et al.* [59] ont traité la reconnaissance d'activités comme étant un problème de classification multi-étapes, ou différentes caractéristiques ont été extraites en fonction de la nature de l'activité. Dans un premier temps les informations de distance ont été extraites à partir de la représentation Distance-Temps pour effectuer une classification binaire des activités lorsque la cible est statique ou mobile avec une précision de 99,9%. Ensuite la méthode de transformation Temps-Distance-Fréquence pondérée (Weighted Range-Time-Frequency Transform : WRTFT) a été utilisée pour générer les spectrogrammes. Les caractéristiques physiques empiriques et l'ACP ont été utilisées pour classifier les activités sur douze classes, donnant respectivement une précision de 98,1% (classifieur bagged tree, cible statique) et de 98,5% (classifieur kNN, cible mobile). Les auteurs ont pu conclure que les caractéristiques empiriques physique sont moins précises et peuvent être perturbés par la diversité individuelle.

Wang *et al.* [140] ont proposé de réaliser la reconnaissance d'AH uniquement sur la base des caractéristiques temporelles à savoir : l'énergie, la moyenne, la moyenne carrée, la marge, l'aplatissement, l'asymétrie, et la variance, permettant une réduction significative de la dimension des données utilisées pour l'entraînement de l'algorithme DT. Les résultats de classification montrent que le modèle arrive très bien à reconnaitre la cible debout, tandis qu'un taux de reconnaissance faible et déroutant est obtenu lorsque la cible se déplace avec un mouvement tangentiel et radial.

Kumar *et al.* [81] ont proposé la détection et la classification des AH basée sur le schéma de respiration. L'ensemble de données du signal de respiration généré à l'aide d'un IR-UWB a été créé en effectuant quatre activités à savoir : assis inactif, yoga, parler fort, méditation sur la respiration profonde. Les auteurs ont utilisé la Transformée de Fourier de courte durée (Short Time Fourier Transform : STFT) converties par la suite en un vecteur de caractéristiques, et utilisée comme entrée fournies à un SVM. Une précision de 85.25% a été obtenue.

Dans l'étude de Tsuchiyama *et al.* [141] un système de reconnaissance de chutes dans les sanitaires a été proposé. Un seul IR-UWB a était utilisé et fixé à l'arrière du couvercle des toilettes. En considérant le fait que le signal s'affaibli lorsqu'il y a un humain sur

le siège des toilettes ou à l'extérieur du sanitaire, le système détecte efficacement le type d'activités. Cependant, les mesures ont été effectuées sur un espace très petit. Il pourrait ainsi être difficile d'atteindre une précision de 95% lorsque la chute se produit dans un espace plus large.

Hämäläinen *et al.* [142] ont proposé un système permettant d'identifier les postures dynamiques et statiques à partir d'un IR-UWB monté sur le mur à une hauteur de 1 m. Le système proposé peut également extraire les taux de respiration et de toux lorsque l'utilisateur est dans une posture statique. Pour ce faire, ils n'ont calculé que trois paramètres statistiques à savoir : aléas (Skewness), aplatissement et puissance et ont utilisé l'algorithme k-NN pour discriminer les postures statiques et dynamiques. Les postures à reconnaître sont la position se mettre debout, s'assoir et se coucher. Les performances obtenues sont supérieures à 99% de précision.

Fugui *et al.* [143] ont utilisé un système radar à entrées et sorties multiples pour collecter sept activités à partir de huit volontaires dans un milieu intérieur et traversant un mur. La collecte des données s'est faites sous un angle de vue de $\pm 20°$ et une distance qui varis jusqu'à 6 m. Différents algorithmes ont été impliqués, et une précision de 79,34%, 72,98% et 75,83% a été obtenue pour le SVM, kNN et RF, respectivement. Les résultats expérimentaux ont montré que la méthode proposée a une forte tolérance aux interférences, même pour des distances de détection, des angles, des activités et des cibles différentes.

2.6.2.2 Apprentissage profond pour la reconnaissance des actions

Shao *et al.* [80] ont utilisé des images de Distance-Temps pour représenter sept actions à savoir : marcher, courir, marcher en tenant un bâton, boxer, boxer en avançant, tomber et s'asseoir. Pour la classification, les auteurs ont opté pour un CNN. Le modèle a donné de bons résultats et a atteint une précision moyenne de 95.24%. Cependant, les différentes expérimentations ont démontré que la précision diminue lorsque la cible n'est pas sur la même ligne de visé du radar.

Sadreazami *et al.* [72, 73] ont proposé d'utiliser des séries temporelles représentant la variation de l'amplitude du signal au cours du temps pour différencier entre les chutes et non-chutes. À cette fin, les auteurs proposent d'utiliser un LSTM empilé pour une extraction automatique de caractéristiques et une classification robuste des données. Bien que leur approche soit facile à implémenté, le modèle a nécessité un large nombre d'itérations durant la phase d'apprentissage afin d'atteindre les résultats désirés. De plus il a été con-

clu que leur approche est vulnérable aux variations d'angles. Plus l'angle entre la cible et le radar s'élargie, plus les performances du modèle se dégradent.

Chen *et al.* [98] ont utilisé la représentation WRTFT comme entrée à un CNN pour la reconnaissance de six AH. Les résultats obtenus prouvent que le CNN offre une bonne robustesse faces à la diversité individuelle et atteint une précision de 92,8%.

Wang *et al.* [55] ont proposé une nouvelle représentation de données qui est la Distance-Doppler variable dans le temps (Time-varying Range–Doppler Images : TRDI) représentant les caractéristiques du mouvement variable dans le temps en effectuant une analyse Fréquence-Temps sur le signal IR-UWB. Il a été ensuite utilisé l'algorithme d'ACP ou un CAE pré-entraîné pour l'extraction des caractéristiques à partir de la TRDI. Enfin un réseau récurrent à portes (Gated Reccurent Units : GRU) est employé pour modéliser dynamiquement les caractéristiques et obtenir le résultat de la classification. Huit actions ont été impliqué à savoir : ramper, tomber, courir, sauter, s'asseoir, se pencher, s'accroupir et marcher. La précision la plus élevée a été obtenue par le CAE et est égale à 96.80%.

Du *et al.* [32] ont proposé un Segmented Convolutional Gated Recurrent Neural Networks (SCGRNN), qui est une combinaison de CNN et de GRU. Les couches CNN ont été utilisées pour extraire les caractéristiques des spectrogrammes Fréquence-Temps. Sur la base de ces dernières, les GRU ont été utilisés pour calculer les scores de classification d'activité. Le modèle a été capable d'atteindre une précision de 88,19%.

Han *et al.* [144] ont utilisé un IR-UWB avec un 2D-CNN pour distinguer entre une chute et les AVQ à savoir : marcher, se tenir debout, s'asseoir et se lever. L'approche consiste a utilisé une classification binaire et s'est concentré sur l'utilisation des images Distance-Temps. L'algorithme CNN a permis d'obtenir un score de précision de 96,35%.

Yang *et al.* [118] ont montré la faisabilité de l'utilisation des informations de distance pour la reconnaissance des mouvements humains à travers le mur. Ils ont combiné un AEN à trois couches denses avec un GRU à deux couches cachées. Le modèle a atteint une précision de reconnaissance de 93% dans les 20% de la durée initiale des activités.

Sadreazami *et al.* [145] ont conçu une méthode de détection des chutes à partir de l'analyse Fréquence-Temps des signaux IR-UWB. Les spectrogrammes obtenus sont convertis en images binaires et sont introduites comme entrées d'un CNN. Le modèle a pu atteindre une précision de 98.37%.

Noori *et al.* [100] ont classifié cinq activités à savoir allongé, assis sur le lit avec les jambes sur le lit, assis sur le lit avec les jambes sur le sol, debout et en marchant, obtenues auprès de treize participants à l'aide d'un IR-UWB. En utilisant une analyse discriminante

améliorée avec LSTM, ils ont obtenu une précision de classification moyenne de 99,6%.

Pour assurer la sécurité tant du conducteur que des passagers due à la distraction ou l'utilisation à long terme du pilote automatique, Brishtel *et al.* [146] ont étudié la faisabilité d'un IR-UWB pour la reconnaissance des activités de conduite. Les auteurs ont utilisé les données Fréquence-Temps de six activités associées à la conduite conventionnelle, autonome et distraite pour entrainer et évaluer différents modèles. Une précision de 100% a été atteinte par le modèle ResNet-18.

Xikang *et al.* [147] ont proposé un système d'apprentissage multi-tâches : un réseau d'identification personnelle et de détection de chute basé sur un IR-UWB. Les résultats expérimentaux obtenus à l'aide d'un ensemble de données provenant de onze personnes se trouvant dans un milieu intérieur démontrent des performances exceptionnelles. Le système à apprentissage multi-tâches a atteint une précision d'identification personnelle moyenne de 98,7%, et une précision de détection de chutes de 96,5%.

2.6.2.3 Fusion de données pour la reconnaissance des actions humaines

Les informations générées à partir d'un seul domaine/capteur sont généralement insuffisantes pour la reconnaissance d'AH par exemple lorsque la cible change de distance ou bien d'orientation par rapport à la ligne de visée du radar. La fusion de données est alors utilisée pour améliorer les performances globales de classification en combinant des informations complémentaires provenant de différents domaines/capteurs à différents niveaux d'abstraction : signal, caractéristique et/ou décision.

Jokanovic *et al.* [36] ont adopté une représentation multi-domaine pour la détection des chutes. Les informations Distance-Temps, Fréquence-Temps et Distance-Fréquence ont été utilisées et traitées avec un AEN pour l'extraction des caractéristiques. Un vote est ensuite effectué pour définir l'étiquette de l'activité à savoir marcher, tomber, s'asseoir, se pencher, classifier par une couche Softmax pour chacune des trois représentations d'entrées. Le résultat obtenu montre que la sensibilité dans la détection des chutes a été réduite par rapport aux entrées individuelles. Cela indique que la stratégie de vote devrait être révisée pour maximiser la détection des chutes.

Piriyajitakonkij *et al.* [56] ont développé Sleep PoseNet, un CNN pour la reconnaissance de différentes postures pendant le sommeil. Le modèle a été entraîné à la fois sur la représentation Distance-Temps et sur la représentation WRTFT et a atteint une précision de 73,4%. Les auteurs ont suggéré de l'utiliser pour la détection des troubles du sommeil.

Cependant une limitation de l'approche proposée est la confusion entre les mouvements lorsque plusieurs cibles se trouvent sur la même distance avec différents angles par rapport à la ligne de visée du radar.

Mostafa *et al.* [104] ont proposé une approche de reconnaissance d'activité en deux étapes fondée sur une représentation à domaines multiples, à savoir la combinaison des domaines Fréquence-Temps et Distance-Temps. La première étape consiste à effectuer une classification binaire des mouvements pour chute/non chute sur la base des caractéristiques extraites des différents domaines. Une précision de 96.9% a été obtenue. La deuxième étape utilise le résultat de classification obtenu à partir de la première pour reconnaitre les huit types d'activités avec une précision de 95.8%.

L'étude menée par Qi *et al.* [148] a consisté en l'implémentation d'un algorithme multi-classification à trois étapes pour la reconnaissance de douze activités basée sur un IR-UWB. La première étape consiste à extraire les caractéristiques radiales, puis les classifier par l'algorithme kNN afin de déterminer la direction de déplacement de l'humain à savoir : stationnaire, se déplace vers le radar, s'éloigne du radar. La seconde étape consiste à utiliser l'algorithme d'extraction des caractéristiques du spectre de puissance et l'algorithme d'extraction des caractéristiques des décalages Doppler pour extraire et visualiser les caractéristiques des différentes catégories classifiées dans la phase. La troisième étape consiste à fournir les spectrogrammes obtenus au modèle Google-Net pour effectuer la classification finale et réaliser la reconnaissance des mouvements humains.

Le corps humain est considéré comme un bon réflecteur de signaux. Ces derniers peuvent suivre plusieurs chemins de réflexion et peuvent interférer de manière constructive ou destructive. L'effet de trajets multiples peut être capturé à partir des informations de réponse impulsionnelle du canal (Channel Impulse Réponse : CIR) fournissant des informations sur la propagation des signaux dans l'environnement. Il a été démontrer dans [110,111] la faisabilité d'utiliser ces données CIR comme caractéristiques pour entrainer des modèles ML et DL dans le but de réaliser la RAH. Les résultats ont montré que des activités simples telles que se tenir debout, s'asseoir et s'allonger peuvent être reconnues avec une précision élevée grâce à la quantité d'information que les données CIR contiennent, ainsi que le faible pré-traitement qu'elles nécessitent.

Bouchard *et al.* [102] ont suggéré de simuler un environnement de maison réaliste. Les données ont été collectées à partir de trois IR-UWB en présence de plusieurs individus dans différentes pièces. La reconnaissance d'activité basée sur kNN, arbre de régression (Regression Tree : RT), AdaBoost et RF a été réalisée dans un appartement de quarante mètres

carrés. Bien que la précision ait atteint 80% en utilisant l'algorithme RF, avec une bonne sélection de caractéristiques, un petit ensemble de données et des efforts de configuration négligeables, plusieurs radars étaient nécessaires pour atteindre cette performance.

Maitre *et al.* [149] ont proposé un système pour détecter les chutes dans un milieu intérieur en utilisant trois IR-UWB et un modèle hybride composé d'un 1D-CNN et un LSTM. Les auteurs ont suggéré d'utiliser les séries temporelles divisées sur cinq séquences et fournissent à l'entrée du modèle pour réaliser l'extraction des caractéristiques. Il en résulte cinq séries de caractéristiques qui alimentent le classifieur final. Bien que l'étude ait impliqué quatre types de chutes, à savoir la chute vers l'avant, la chute latérale (gauche ou droite), la chute vers l'arrière et la chute après avoir essayé de s'asseoir, le problème abordé est traité comme une classification binaire visant à différencier les événements avec ou sans chute. D'après les résultats obtenus, il a été conclu que l'emplacement du radar affecte fortement les performances de reconnaissance dont 98.51%, 96.75% et 95.48% de précision ont été obtenues pour trois positions différentes.

Chowdhury *et al.* [63] ont proposé une technique de RAH en temps réel utilisant le balayage des signaux IR-UWB disposés sous forme de matrice de Distance-Temps. Différents classifieurs ont été entrainés et testés sur les caractéristiques extraites par l'algorithme d'Analyse Bidimensionnelle-Bidirectionnelle des composants principales (two-dimensional two-directional principal component analysis : 2D2D-PCA). Une précision de 86.4 $\pm$5.2% a était obtenue par l'algorithme kNN. Les auteurs envisagent de rajouter la détection de signes vitaux dans les travaux prochains.

Maitre *et al.* [116] ont proposé un système binaire en vue de la classification des incidents de chute et de non-chute en utilisant trois IR-UWB. Les données ont été recueillies à partir de dix volontaires dans trois zones distinctes au sein d'une résidence. Leur approche a été validée en utilisant la méthode Leave One Subject Out : LASO. Une précision de 90% a été obtenue en utilisant un modèle hybride à architecture CNN-LSTM.

Li *et al.* [66] ont fusionné différentes caractéristiques extraites à partir des images Fréquence-Temps et du DCV pour classifier les mouvements individuels et séquentiels à l'aide d'un radar FMCW et de trois IR-UWB. Quarante-sept caractéristiques du Fréquence-Temps et sept caractéristiques du DCV ont été extraites pour entrainer une mémoire à court terme bidirectionnelle (BidirectionalLSTM : Bi-LSTM). Une précision de 98,2% et de 93% a été obtenue respectivement pour les mouvements individuels et séquentiels respectivement.

Bordvik *et al.* [150] ont proposé un système qui assure la détection des comporte-

ment normaux et anormaux en explorant différentes méthodes de fusion de capteurs et d'algorithmes. A cette fin ils utilisent des images de profondeur RVB-D et des données de IR-UWB pour entrainer un CNN et un LSTM. La fusion au niveau décisionnel a permis d'obtenir la précision la plus élevée de 99.98% par le modèle CNN.

Imbeault-Nepton *et al.* [151] ont calculé dix-sept caractéristiques à partir des données de trois IR-UWB pour la classification de quinze AVQ. Bien constaté que ce nombre de caractéristiques est trop important, les auteurs ont proposé d'utiliser l'ACP afin de réduire le nombre de caractéristiques et éviter le sur-ajustement du classifieur RF. Le modèle n'a pu atteindre qu'une précision moyenne de 59%.

Yubin *et al.* [152] ont proposé un système de IR-UWB distribué composé de cinq nœuds afin d'assurer l'éclairage total du corps humain dans la zone de mesure. Neuf activités ont été impliquées à savoir marcher, se tenir debout de manière stationnaire, s'asseoir, se lever, se pencher depuis la position assise, se pencher depuis la position debout, chute en marchant, se relever après une chute et chute en position debout lorsque la cible est immobile. Ces activités ont été effectuées dans des directions arbitraires par rapport à la ligne de visée des capteurs et représentées par des images de Distance-Fréquence. Les résultats ont montré que la fusion de données permet de tirer parti d'une plus grande quantité de données pour former les modèles de classification et d'exploiter la diversité spatiale des mêmes actions vues simultanément par plusieurs radars.

Simin *et al.* [153] ont étudié la faisabilité d'extraire des caractéristiques spatio-temporelles à partir des images Fréquence-Temps de données provenant d'un réseau de IR-UWB distribués composé de cinq nœuds. A cette fin les auteurs ont proposé de combiner un CNN avec une GRU et tester différentes combinaisons de fusion à savoir : la fusion précoce, fusion tardive et fusion à mi-chemin. D'après les résultats obtenus, la précision la plus élevée de 90.8% a été constaté pour la fusion à mi-chemin. Cela indique que cette dernière permet un équilibre entre la combinaison de détails fins et de représentations de caractéristiques de haut niveau et qui est le meilleur choix de schéma de fusion.

Guendel *et al.* [154] ont exploité différentes représentations des données d'un réseau de IR-UWB à savoir Distance-Fréquence, Fréquence-Temps, Distance-Temps et Transformée de Fourier synchro-squeezed. Multiple classifieurs ont été appliqués aux données fusionnées à savoir SVM, kNN, DT, Naïve Bayes (NB) et testés pour la fusion précoce et la fusion tardive. Bien que la fusion des caractéristiques améliore les performances de classification, il a été constaté qu'elle fournit un large volume de données qui par la suite nécessite des ressources de traitement importantes.

Contrairement aux travaux antérieurs qui se sont concentrés sur la classification des activités à partir des données radars traités comme des matrices à valeurs réelles, Ximei *et al.* [155] ont étudié la faisabilité d'utiliser les données complexes des images Distance-Temps et Fréquence-Temps pour entrainer les modèles de classification. L'idée clé est de séparer les parties réelle et imaginaire des données d'entrée en deux canaux et les traiter comme deux nombres réels indépendants. Les résultats ont montré que les implémentations à valeurs complexes ne permettent d'améliorer la précision de la classification que pour certains formats de données et certaines architectures et n'est pas généralisable.

Lai *et al.* [156] ont proposé un système à double IR-UWB pour la surveillance de posture de sommeil en utilisant les techniques d'extraction de caractéristiques conventionnelles et automatiques. Les résultats obtenus ont montré que le système à double radar est plus performant qu'un système à radar unique. Les caractéristiques statistiques prédéterminées avec le classifieur RF ont donné la meilleure précision de 88.7% en surpassant les approches d'apprentissage profond.

2.6.2.4 Reconnaissance des actions humaines dans le cas des données limités

Bien que les approches d'apprentissage profond aient montré un grand potentiel d'atteindre les meilleurs résultats, leurs performances dans les applications de reconnaissance d'actions sont souvent limitées en dépit des petits ensembles de données disponibles pour l'apprentissage. Cela s'explique principalement par le fait que la collecte de données exige que la cible répète chaque essai séparément pour obtenir des échantillons indépendants, un processus coûteux en temps, en termes d'opérations et de ressources humaines. Afin de réduire de manière significative le coût et l'effort de mesure des données, plusieurs stratégies pour élargir le volume de données d'entrainement ont été proposées.

La première approche consiste à simuler les données radars en utilisant la capture de mouvement (Motion Capture : MOCAP) afin de saisir les mouvements de cible à l'aide de marqueurs ou de capteurs montés sur le corps. Les données MOCAP peuvent être combinées avec des calculs électromagnétiques (EM) pour simuler des échos radar basés sur des mouvements humains *et al.* [157]. Les données MOCAP fournissent les paramètres de mouvement telles que les angles d'articulation, les vitesses ou les trajectoires des membres et les calculs EM intègrent les caractéristiques de propagation radar pour générer des échos radar simulés [157]. Cette approche permet de générer des signatures radar réalistes associées à différentes activités humaines. Depuis la proposition des données de capture

de mouvement par Ram *et al.* [158], ils ont été utilisés pour une variété d'études pour l'extraction de caractéristiques pertinentes pour l'analyse la classification des mouvements [64,95,159,160]. Cependant, le principal inconvénient de la simulation est que les données générées sont trop propres et parfaites, alors que dans les scénarios réels, les données sont affectées par divers facteurs environnementaux, par les paramètres du capteur et ainsi que les caractéristiques de la cible. Comme exemple, les variations introduites par le milieu environnant, telles que les obstructions causées par les murs, les objets ou les mouvements qui ne sont pas liés à la cible. L'écart entre le monde réel et la simulation peut considérablement détériorer les performances des modèles. Pour résoudre cette lacune, le réseau adversarial génératif (Generative Adversial Network : GAN) a été proposé comme moyen de générer des images simulées très réalistes [90,91]. Le GAN utilise un réseau générateur alimenté par un bruit aléatoire ainsi que les données réelles et étiquetées afin de générer des données synthétiques dont la distribution ressemble à celle des données radars réels. Une première tentative d'application du GAN pour synthétiser les signatures micro-Doppler a été proposée par Shi *et al.* [93] pour les allures de marche à différentes vitesses. Une approche similaire a été proposée par Kiasari *et al.* [134] pour générer différentes actions humaines, étendues à d'autres mouvements que la simple marche. Gurbuz *et al.* [33] ont utilisé un GAN pour synthétiser les signatures micro-Doppler collectées à partir de trois radars pour l'apprentissage inter-fréquences. Les résultats montrent une augmentation de la précision globale de la classification. Erol *et al.* [92] ont utilisé des réseaux adversaires génératifs à classificateur auxiliaire (Auxiliary Classifier Generative Adversarial Networks : ACGAN) pour générer des signatures micro-Doppler synthétiques plus diverses et plus nettes. Les résultats ont montré l'efficacité des données synthétiques ACGAN adaptées à différents environnements et position de détection. Zhong *et al.* [65] ont proposé d'utiliser le Réseau adversarial génératif à convolution profonde (Deep Convolution Generative Adversarial Network : DCGAN) pour augmenter les données de l'ensemble d'échantillons micro-Doppler. Les résultats expérimentaux ont montré que la combinaison du GAN et du CNN permet d'obtenir une reconnaissance efficace. Toutefois, l'expérimentation a été réalisée dans un environnement idéal sans interférence. Bien que le GAN ait démontré sa capacité à générer des données synthétiques réalistes, la fidélité de ces derniers n'est pas garantie. Il a été démontré que les signatures radars générées par le GAN correspondent à des comportements cinématiquement impossibles ou à des classes de mouvements différentes de celles attendues [161,162].

L'apprentissage par transfert (Transfert Learning : TL) et les méthodes d'adaptation

au domaine (Domain Adaptation : DA) constituent une autre approche pour faire face au nombre faible des échantillons d'apprentissage. TL implique l'apprentissage de modèles sur de grands ensembles de données, puis le réglage fin de leurs poids sur de petits ensembles de données pour une nouvelle tâche de classification [163]. Cela réduit considérablement la dépendance des modèles à l'égard de large volume de données d'apprentissage, améliore la précision de la reconnaissance et la vitesse de convergence. Plusieurs modèles profond pré-entraînés sur l'ensemble de données ImageNet ont été utilisés pour classifier des signatures micro-Doppler d'activités humaines [84, 85]. Du *et al.* [84] ont présenté un réseau résiduel à apprentissage par transfert pour classer les activités humaines sur la base des spectrogrammes micro-Doppler. Les performances du modèle ont été évaluées sur l'ensemble de données Mocap. Du *et al.* [85] ont proposé d'utiliser le VGG-19 pré-entrainé sur l'ensemble de données imageNet pour la classification des images micro-Doppler. Les résultats expérimentaux démontrent que le VGG-19 entrainé par transfert est plus performant que le modèle formé à partir de zéro et permet de réduire le nombre de paramètres et d'opérations de calcul. Cependant, la précision requise ne peut être atteinte uniquement en transférant directement les caractéristiques. Les caractéristiques des signatures radar diffèrent largement de celles des images optiques. Les modèles peuvent présenter des performances imprévisibles en cas d'inadéquation entre le contenu de la formation de la source et celui de la cible. D'autre part, la DA [94, 95, 164], qui relève du domaine de recherche du TL, est déployée pour atténuer le changement de distribution. L'objectif principal est de parvenir à un alignement intermédiaire entre les domaines source et cible. Toutefois, les conditions à remplir en cas de changements importants dans les domaines ne sont pas clairement définies, ce qui rend l'alignement des domaines plus difficile.

2.7 Conclusion

Au cours de ce chapitre, nous avons examiné de manière approfondie les algorithmes d'intelligence artificielle prédominants ainsi que leurs concepts opérationnels. Par la suite, nous avons établie une revue de la littérature détaillée des solutions dédiées à la RGM et des AH exploitant les IR-UWBs. Le chapitre suivant sera consacré à la présentation de notre solution de réseau de capteurs pour la RGM et des AH, qui repose sur l'utilisation des IR-UWBs et des algorithmes d'IA.

Chapitre 3

Réseaux de capteurs pour la reconnaissances des gestes de la main et des actions humaines

3.1 Introduction

Notre principal objectif est de développer un système de RGM et des AH en utilisant un réseau de IR-UWB comme illustré sur la Figure 3.1. Cependant, nous avons rencontré un défi majeur : le manque d'une base de données complète qui regroupe à la fois les GM et les AH nécessaires pour entraîner notre modèle. Pour surmonter cette limitation, nous avons eu recours à la combinaison de deux ensembles de données distincts [33, 101]. L'idée maîtresse derrière notre approche est d'utiliser un modèle de classification commun pour les deux tâches requises. Nous aspirons à créer un système puissant et polyvalent, ouvrant ainsi la voie à de multiples applications.

La section 3.2 explique les détails de la conception du système de reconnaissance proposé, en commençant par le pré-traitement des données et en se prolongeant jusqu'au modèle de classification. Elle aborde également les critères d'évaluation et la stratégie d'entrainement appliquée au modèle. Les sections 3.3 et 3.4 décrivent respectivement les bases de données utilisées pour la RGM et des AH. Elles discutent également les résultats obtenus et établissent une comparaison avec les travaux antérieurs issus de la littérature. Enfin, la section 3.5 conclue le chapitre.

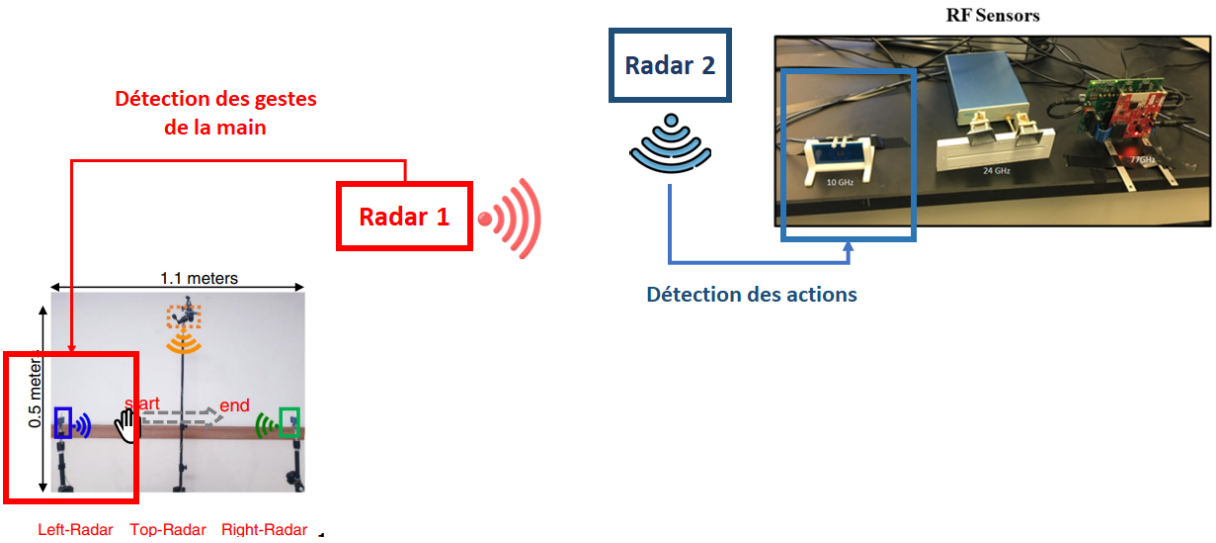


Figure 3.1 – Réseau de radars UWB pour la reconnaissance des gestes de la main et des actions humaines [33,101].

3.2 Système de reconnaissance des gestes de la main et des actions humaines

3.2.1 Description

Le système de RGM et des AH proposé comprend cinq éléments fondamentaux, à savoir:

1. Le pré-traitement des données :

   Le pré-traitement des données permet d'étendre les données radar sur différentes représentations en extrayant des caractéristiques de bas-niveau qui sont par la suite utilisées en données d'entrée pour le modèle de classification proposé. La procédure de pré-traitement comprend trois étapes : (1) Filtrage des données, (2) Renforcement des caractéristiques, (3) Normalisation et étiquetage des données. A la fin du processus de pré-traitement, il en résulte trois paires images-étiquette alignées pour chaque échantillon original. Le processus de pré-traitement permet non seulement d'augmenter la quantité d'informations utilisées pour décrire un geste/une action, mais aussi d'améliorer les caractéristiques de fond en réduisant le bruit dans les données et en fournissant des informations plus diversifiées.

2. Le bloc d'extraction des caractéristiques spatiales :

   Le CNN agit comme un extracteur de caractéristiques hiérarchiques spatiales.

Il prend comme données d'entrée les images étendues générées lors de l'étape de pré-traitement (voir Figure 3.2) . La sortie du bloc CNN comporte un ensemble de cartes de caractéristiques qui représentent les informations extraites des données d'entrée. L'objectif principal du bloc CNN dans l'architecture hybride CNN-LSTM est d'effectuer un processus d'extraction automatique des caractéristiques spatiales pour capturer les corrélations locales dans les données.

3. Module de concaténation :

   Le module de concaténation regroupe les cartes de caractéristiques générées par les trois branches CNN verticalement sous forme d'un cube. Ensuite la sortie du module de concaténation est remodelée de manière à correspondre à l'entrée du bloc suivant. Le module de concaténation des données vise à concaténer les meilleures caractéristiques de niveau intermédiaire qui peuvent conclure au geste/action de la cible.

4. Le bloc d'extraction des caractéristiques temporelles :

   Les données d'entrée du LSTM sont les cartes de caractéristiques générées par le module de concaténation de données précédent. Le bloc LSTM prend les données d'entrée comme un ensemble de séquences temporelles et capture les dépendances temporelles à long terme. La sortie du bloc LSTM est une carte de caractéristiques dans laquelle chaque valeur est une représentation de haut niveau des données d'entrée pour une région locale correspondante. La sortie du bloc LSTM est converti sous forme de vecteur et est fournie comme entrée au classifieur finale.

5. Le bloc de classification :

   Le bloc de classification consiste en l'utilisation du classifieur SVM, ajouté au sommet du modèle proposé. Il prend le vecteur de caractéristiques de haut niveau du bloc LSTM et génère les prédictions finales pour les données d'entrées. Pendant l'apprentissage du modèle, l'objectif principal du bloc SVM est d'apprendre diverses fonctions de mappage entre les différentes caractéristiques de haut niveau et la classe de geste/action finale.

3.2.1.1 Pré-traitement des données

La classification est un processus complexe qui nécessite la prise en compte de nombreux facteurs à savoir le pré-traitement des données ainsi qu'un nombre suffisant d'échantillons d'entrainement. La phase de pré-traitement dans notre système a pour trois objectifs

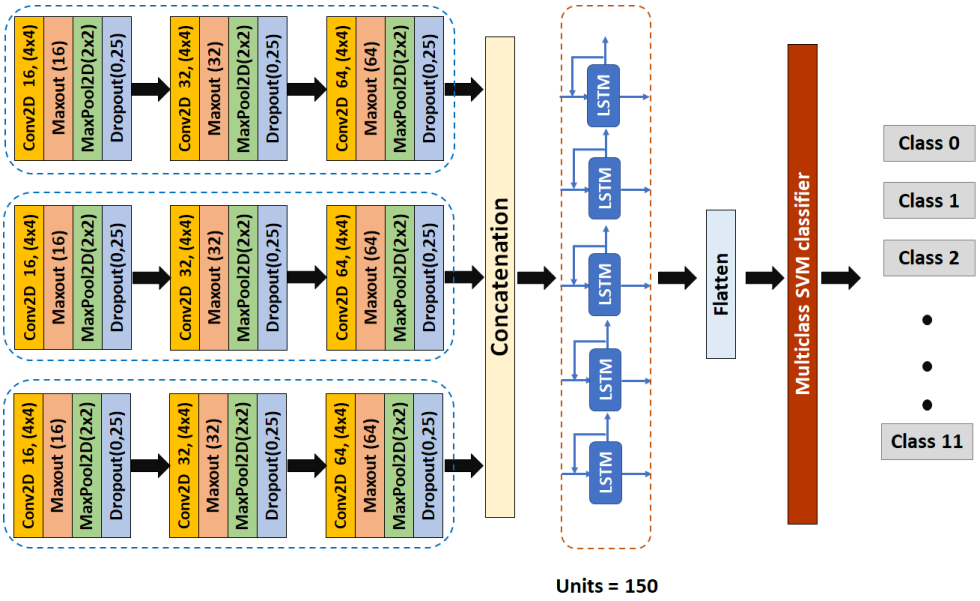


Figure 3.2 – Architecture du modèle de classification des GM et AH.

principaux : Premièrement, nous visons à réduire tout bruit ou effet indésirable présent dans les données afin d'obtenir une plus grande précision, ce qui peut être réalisé grâce à un filtrage de données. Deuxièmement, nous cherchons à renforcer et améliorer le contenu des échantillons qui peuvent être fournis comme entrée au modèle de classification. L'amélioration des images joue un rôle majeur dans les techniques de pré-traitement, car elle nous fournit de meilleures données pour un examen plus approfondi. Troisièmement, la normalisation des données nous permet de préparer un ensemble de données uniforme à des fins d'apprentissage. La normalisation des données ou la mise à l'échelle des caractéristiques réduit la variance globale des données de mesure, ce qui facilite la classification.

— **Filtrage de données**

La binarisation des images est l'une des étapes de pré-traitement les plus pertinentes qui permet de réduire considérablement la quantité d'informations soumises à une analyse ultérieure et d'en augmenter la vitesse de traitement. Cette opération est généralement appliquée dans de nombreux systèmes qui utilisent principalement des méthodes de reconnaissance des formes et ne nécessitent pas d'analyse de la couleur ou de la texture. Nous avons opté pour notre approche l'utilisation du seuillage global pour la binarisation des images (voir Figure 3.3). L'algorithme de seuillage global utilise une valeur de seuil d'intensité fixe T (de 0 à 255). Il compare l'intensité de chaque pixel de l'image source à la valeur seuil prédéfinie et réattribue des valeurs d'intensité différentes. Si la valeur d'intensité d'un pixel est supérieure à T, le pixel devient noir (1), sinon il devient blanc

(0).

$$g(x,y) = \begin{cases} 1, & f(x,y) > \text{seuil} \\ 0, & f(x,y) < \text{seuil} \end{cases} \tag{3.1}$$

où g(x,y) représente les pixels de l'image seuil et f(x,y)représente les pixels de l'image originale. L'utilisation du seuillage globale à valeur fixe est suffisant pour notre approche afin d'obtenir une qualité constante. Cela est dû au traitement d'un lot d'images où les distributions d'intensité des pixels du premier plan et de l'arrière-plan ne varient pas entre les images.

Cette étape de filtrage nous a permis l'élimination du bruit afin d'éviter les informations redondantes sur les pixels et facilite l'analyse ultérieure des images.

— **Renforcement des caractéristiques**

Les caractéristiques d'une image représentent un élément d'information sur son contenu. Il s'agit généralement de savoir si une certaine région de l'image possède certaines propriétés. Les caractéristiques peuvent être des structures spécifiques de l'image telles que des points, des bords ou des objets. Cependant, l'amélioration du contenu de l'image peut être définie comme la conversion de la qualité visuelle d'une image à un niveau meilleur et plus compréhensible pour l'extraction de caractéristiques à des fins de réalisation de taches de plus haut niveau, telle que la classification.

Pour ce faire, nous proposons d'utiliser l'opérateur Sobel pour l'extraction des caractéristiques de bas niveau. Cela permet d'accentuer certaines caractéristiques de l'image tout en préservant ses détails. L'opérateur Sobel est un détecteur de contour bien connu, qui calcule le gradient de l'intensité de l'image. Il a été utilisé pour extraire des caractéristiques locales dans de nombreux travaux [165–167]. La méthode du gradient détecte les bords en recherchant le maximum et le minimum à partir de la première dérivée de l'image. L'estimation du gradient d'intensité en un pixel dans les directions x et y pour une image f, est donnée par :

$$\frac{\delta f_x}{\delta x} = \Delta x = \frac{f(x+d_x, y) - f(x,y)}{d_x} \tag{3.2}$$

$$\frac{\delta f_x}{\delta y} = \Delta y = \frac{f(x, y+d_y) - f(x,y)}{d_y} \tag{3.3}$$

où dx et dy mesurent la distance le long des directions x et y respectivement. Dans les images discrètes, on peut considérer dx et dy en termes de nombre de pixels entre deux points. d_x= d_y= 1 (espacement des pixels) est le point où les coordonnées des pixels sont

(i, j) donc :

$$\Delta x = f(i+1, j) - f(i, j) \tag{3.4}$$

$$\Delta y = f(i, j+1) - f(i, j) \tag{3.5}$$

L'opérateur Sobel implique deux filtres H_x et H_y de dimension (3 x 3) pour chaque image $f(x, y)$. L'un pour l'estimation du gradient D_x dans le sens horizontal et l'autre pour l'estimation du gradient D_y dans le sens vertical. Les deux filtres sont convolués avec l'image originale. Le calcul du gradient (D_x, D_y) peut être exprimé comme suit :

$$D_x = \begin{pmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{pmatrix} * f(x, y) \quad D_y = \begin{pmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{pmatrix} * f(x, y)$$

A chaque pixel de l'image, les approximations du gradient données par D_x et D_y sont combinées pour donner l'amplitude du gradient, en utilisant :

$$D_{xy} = \sqrt{D_x^2 + D_y^2} \tag{3.6}$$

Les images obtenues après l'extraction des contours contiennent des valeurs nulles, positives et négatives du fait que l'opérateur Sobel utilise des filtres avec des valeurs nulles, positives et négatives. À des fins d'affichage, les valeurs absolues de la carte de gradient (entre 0 et 255) sont utilisées. Les valeurs très négatives et très positives apparaissent plus foncés et sont représentées par une valeur de 255, alors que les valeurs nulles apparaissent en couleur claire et sont représentées par un 0.

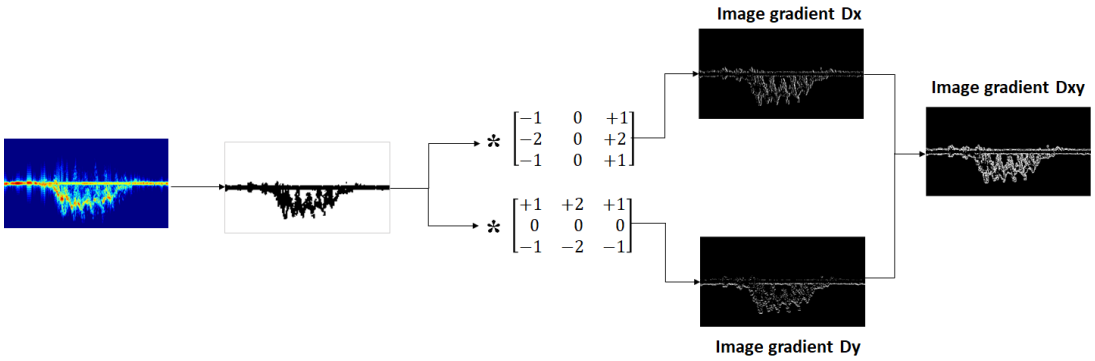


Figure 3.3 – Diagramme du traitement des données.

— **Normalisation et étiquetage des données**

La normalisation ou la mise à l'échelle des caractéristiques sert à préparer un ensemble uniforme de données à des fins d'apprentissage. La normalisation réduit la variance globale des données de mesure, ce qui facilite la classification. Pour la normalisation, nous avons utilisé l'équation suivante :

$$Z_{norm(i)} = (255 - 0)\frac{z_i - min(\vec{z})}{max(\vec{z}) - min(\vec{z})} \tag{3.7}$$

où Z_{norm} désigne la valeur normalisée obtenue pour le *ime* échantillon et Z_i est l'image pré-traitée. La normalisation permet d'obtenir des images où les valeurs des caractéristiques prennent les valeurs 0 et 1. Cette étape permet de réduire la complexité des calculs et améliorer la convergence du modèle. Après normalisation des données, l'étape d'annotation est réalisée. Il en résulte trois paires d'images associés à une étiquette pour chaque échantillon.

3.2.1.2 Extraction des caractéristiques spatiales

La principale utilisation de l'architecture CNN dans le modèle hybride CNN-LSTM est d'effectuer une extraction hiérarchique des caractéristiques spatiales sur les données d'entrée. Le CNN peut découvrir automatiquement des caractéristiques locales et invariantes par rapport aux translations des variations. Il s'agit donc d'une propriété cruciale pour les tâches de RGM et AH du fait que les données radar varient en fonction des différents participants et de leurs manières de les exécutés. La structure d'une branche CNN est détaillée dans le tableau 3.1.

Le modèle CNN se compose de trois branches avec des configurations de couches similaires. Chaque branche comporte une couche d'entrée de taille 75x75x1 et de trois blocs dont chacun consiste en quatre composants principaux, à savoir : (1) la couche de convolution, (2) la couche d'activation non-linéaire, (3) la couche de mise en commun et (4) la couche de régularisation.

Les couches de convolutions utilisent 16, 32 et 64 filtres de taille 4x4, respectivement. En outre, un pas de 2x2 est utilisé pour réduire davantage le coût de calcul. Chaque branche prend en entrée une représentation d'image différente. Les images sont traitées séparément en effectuant de multiples opérations de convolution pour extraire les carac-

Tableau 3.1 – Structure détaillée d'une branche CNN.

Type de couche	Filtre	Padding	Strides	Paramètres	La sortie
Input layer	-	-	-	-	(None, 75, 75, 1)
Conv2D	4x4x16	same	2x2	272	(None, 38, 38, 16)
Maxout	16	-	-	0	(None, 38, 38, 16)
MaxPooling2D	2x2	-	-	0	(None, 19, 19, 16)
Dropout	0.25	-	-	0	(None, 19, 19, 16)
Conv2D	4x4x32	same	2x2	8224	(None, 10, 10, 32)
Maxout	32	-	-	0	(None, 10, 10, 32)
MaxPooling2D	2x2	-	-	0	(None, 5, 5, 32)
Dropout	0.25	-	-	0	(None, 5, 5, 32)
Conv2D	4x4x64	same	2x2	32832	(None, 3, 3, 64)
Maxout	64	-	-	0	(None, 3, 3, 64)
MaxPooling2D	2x2	-	-	0	(None, 1, 1, 64)
Dropout	0.25	-	-	0	(None, 1, 1, 64)

téristiques spatiales. L'opération de convolution est exprimée comme suit :

$$F(i,j) = (I * K)(i,j) = \sum_i \sum_j I(i+m, j+n)K(m,n) \tag{3.8}$$

où I est l'image d'entrée, i et j sont respectivement la hauteur et la largeur, K est le filtre convolutif 2D de taille m x n, et F est la carte de caractéristiques 2D de sortie.

Après chaque couche de convolution, la non-linéarité est introduite dans le modèle par l'ajout de couches d'activation non-linéaires. La couche d'activation permet au modèle d'apprendre toute relation complexe entre l'entrée et la sortie. Elle convertit les correspondances linéaires apprises en formes non-linéaires pour les propagées aux couches supérieures. Cela permet au modèle de s'adapter aux données qui changent progressivement d'une couche à l'autre.

Étant donné que les fonctions d'activations peuvent avoir des performances différentes selon les ensembles de données, le choix de la fonction peut devenir un hyperparamètre du modèle lui-même. Dans ce travail, la fonction d'activation Maxout présentée par Ian Goodfellow *et al.* [168] est utilisée pour effectuer une transformation non-linéaire par éléments sur les données de la couche de convolution.

La fonction d'activation Maxout renvoie la valeur maximale au sein d'un groupe de

différentes cartes caractéristiques générées à partir des couches de convolution. L'approche mathématique du Maxout est définie comme suit :

$$f(x) = max(w_1x + b_1, w_2x + b_2, \ldots, w_kx + b_k) \tag{3.9}$$

où x est la matrice de données, w_1, w_2, ... w_k, sont les filtres de convolution et b_1, b_2, ..., b_k sont les biais.

L'hyperparamètre k doit également être défini avant l'apprentissage. Le choix de k est important pour l'architecture du modèle, car il détermine également sa complexité. Un modèle avec un k élevé est capable d'acquérir plus de caractéristiques de données d'entrée, mais il y a toujours un risque de sur-ajustement. Notre modèle applique la fonction Maxout sur des pairs de cartes caractéristiques ($k = 2$) pour obtenir une efficacité de calcul maximale pendant l'apprentissage. La sortie de chaque couche Maxout est égale au nombre d'entrée. Il en résulte 16, 32, et 64 matrices de données, respectivement, pour chaque couche Maxout. La Figure 3.4 montre un exemple de l'application de l'unité Maxout dans une architecture CNN, où x est l'image d'entrée. L'unité Maxout prend la valeur maximale des opérations de convolution y_1 et y_2. La non-linéarité du neurone Maxout est obtenue par le processus de sélection de la valeur maximale comme dans l'équation 3.9. Cette non-linéarité peut également être considérée comme un processus de sélection des caractéristiques.

Une couche supplémentaire est insérée au niveau de chaque bloc de convolution qui est la mise en commun. Chaque carte de caractéristiques obtenue à partir de la couche Maxout est réduite dimensionnellement. Nous avons utilisé la mise en commun maximale avec une taille de pool de 2x2 pour préserver les caractéristiques les plus pertinentes identifiées. Cette opération permet également de réduire les coûts de calcul et éviter les calculs inutiles.

Une dernière couche est insérée à la fin de chaque bloc de convolution qui est la couche d'exclusion (Dropout). Lorsque l'exclusion est appliquée pendant l'apprentissage, une fraction prédéfinie d'unités aléatoires est désactivée, ce qui réduit la quantité d'informations qui circule dans ces couches. Après l'application de l'exclusion, chaque unité de la région de mise en commun peut être exclue avec une certaine probabilité. Dans notre modèle, nous utilisons une exclusion avec une probabilité p =25% des unités sont désactivées, tandis que 75% restantes sont utilisées. Un exemple de l'opération d'exclusion est illustré sur

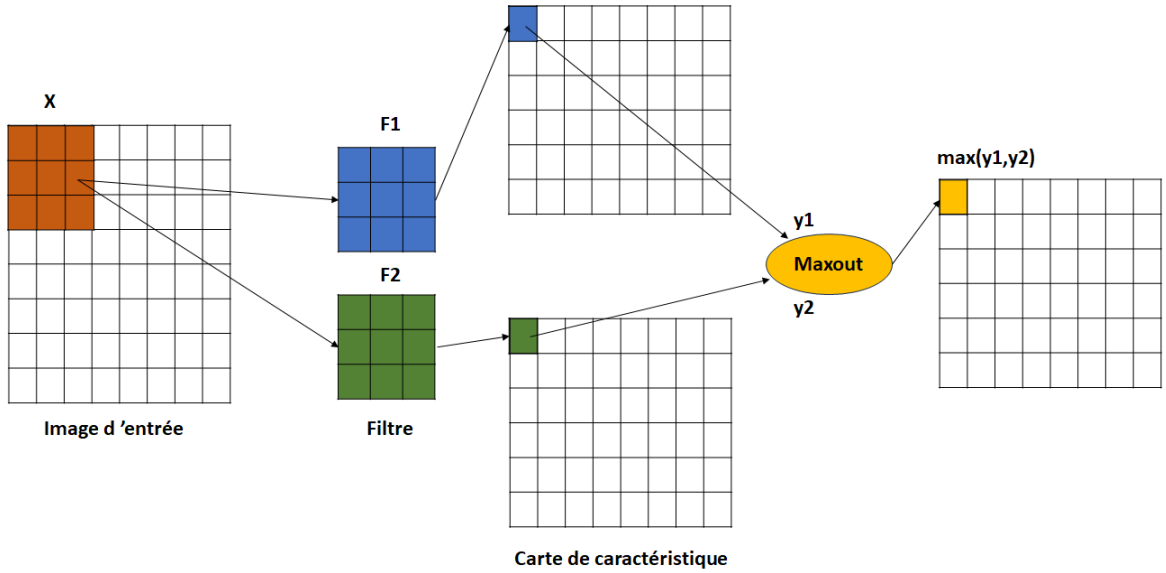


Figure 3.4 – Principe de la non-linéarité Maxout.

la Figure 3.5.

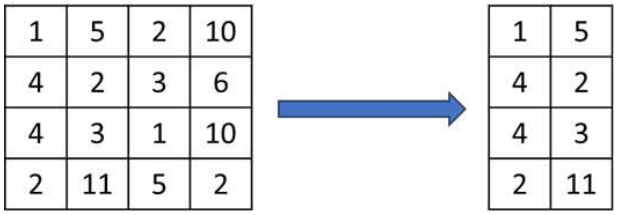


Figure 3.5 – Exemple d'exclusion avec une probabilité $p = 50\%$.

3.2.1.3 Bloc de concaténation

Notre approche consiste à étendre un échantillon sur trois représentations différentes comme mentionner dans l'étape pré-traitement. L'architecture de notre modèle CNN ne se contente pas uniquement d'exploiter la force de multiples représentations de bas niveau pour extraire des caractéristiques complémentaires de la même cible, elle introduit également le concept de concaténation de caractéristiques dans l'architecture pour obtenir des représentations plus holistiques.

La concaténation des caractéristiques peut être divisée en trois catégories selon le niveau sur lequel les données sont concaténées, à savoir la fusion précoce (ou fusion au niveau du signal), la fusion tardive (ou fusion au niveau de la décision) et la fusion à mi-chemin (ou fusion au niveau des caractéristiques) comme illustré sur la Figure 3.6.

Nous proposons dans notre modèle d'utiliser la concaténation à mi-chemin qui représente un compromis entre les schémas de fusion tardive et précoce. Elle combine les représentations des caractéristiques extraites à partir des couches intermédiaires de chaque branche. Cela permet au modèle d'extraire et de préserver les caractéristiques sans les altérer avec d'autres représentations de données. Les données concaténées sont par la suite envoyées au réseau de neurones LSTM pour un traitement ultérieur.

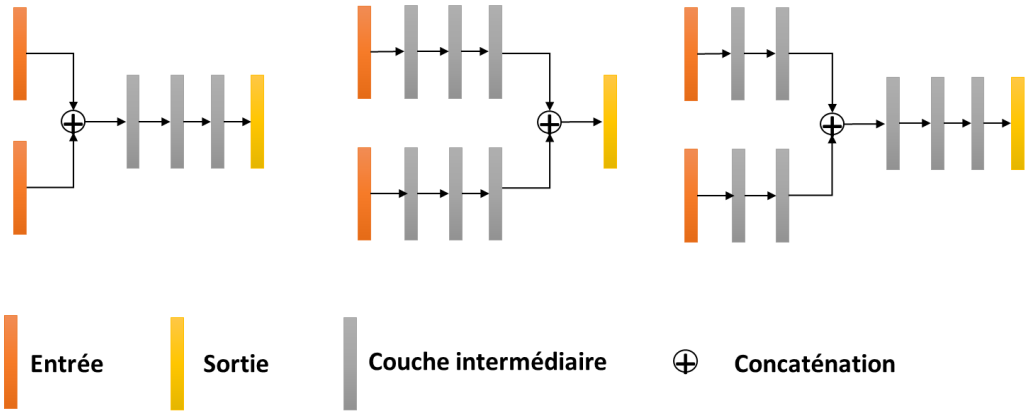


Figure 3.6 – Schéma de concaténation des données.

3.2.1.4 Extraction des caractéristiques temporelles

Le but de notre travail est de classifier correctement des GM et des AH dynamiques. Ces derniers sont par nature des données de séries temporelles. Elles représentent des variations de distance et de vitesse dans le temps des différentes parties du corps humain. Il est donc nécessaire d'impliquer un processus qui exploite les corrélations temporelles des caractéristiques spatiales plutôt que de détecter simplement les formes géométriques globales des signatures radars. C'est pourquoi l'architecture LSTM est introduite après le module de concaténation des données pour modéliser directement les signatures radars dans le temps.

Comme la couche LSTM prend en son entrée des données sous la forme (nombre de pas, nombre de caractéristiques), une couche reshape est d'abord insérée pour remodeler la sortie du bloc de concaténation. Le nombre de pas représente l'occurrence de la cellule LSTM pour la même séquence d'entrée.

Un autre point important à prendre en considération est le schéma d'entrée-sortie du modèle LSTM. Comme illustré sur la Figure 3.7, il existe principalement quatre types de

schémas d'entrée-sortie, à savoir : (a) un-à-plusieurs, (b) un-à-plusieurs, (c) plusieurs-à-un, (d) plusieurs-à-plusieurs-A.

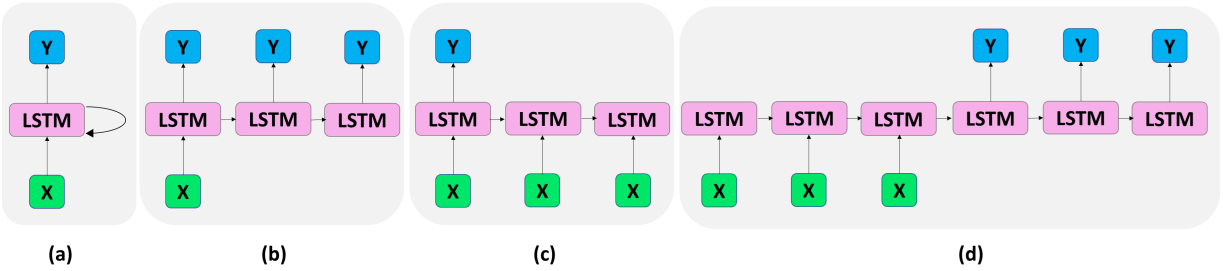


Figure 3.7 – Schéma d'entrée-sortie LSTM.

Comme notre système de reconnaissance proposé a pour objectif de classifier des gestes et des actions individuelles, le LSTM prend en entrée une seule séquence de donnée relative à un geste ou une action. Pour ce faire nous utilisons le schéma (a) un-à-plusieurs. Notre modèle LSTM consiste en une seule couche composée de plusieurs unités. Les unités possèdent la même structure comme détaillé à la section 2.5.2 du chapitre 2, mais initialisées avec des poids différents. Elles calculent alors différentes caractéristiques et leurs sorties ne sont pas identiques.

Afin de lutter contre l'ajustement excessif dans la couche LSTM, l'exclusion récurrente est utilisée (Recurrent dropout). Elle repose sur le même principe que l'exclusion utilisée au niveau des couches de convolutions à savoir cette fois-ci l'exclusion des états cachés dans le sens vertical comme illustré sur la Figure 3.8. Nous avons utilisé l'exclusion récurrente avec un taux égale à 20%.

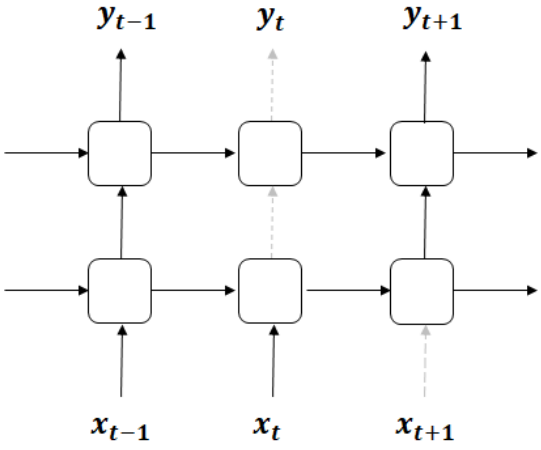


Figure 3.8 – Exclusion récurrente.

3.2.1.5 Classification

Le bloc de classification utilise l'algorithme SVM comme classifieur final pour effectuer les prédictions. Bien que l'algorithme SVM est conçu principalement pour traiter des cas binaires, il existe une multitude de technique pour étendre son application à des problèmes de classification multi-classes. Pour ce faire, nous proposons d'utiliser la stratégie "un contre tous" (one vs all). Considérons un problème à M classes, pour lequel nous disposons de N échantillons d'apprentissage :$\{x_1,y_1\}$,...,$\{x_N,y_N\}$. x_i représente le vecteur de caractéristiques et $y_i \in \{1, 2, ...M\}$ est l'étiquette de classe correspondante.

L'approche "un contre tous" construit M classifieur SVM binaires, chacun séparant une classe de toutes les autres. Le ième SVM est entrainé avec tous les exemples d'apprentissage de la ième classe avec des étiquettes positives, et tout le reste avec des étiquettes négatives. Les classifieurs SVM binaires prédisent la probabilité de correspondance pour chaque classe concernée. En analysant les scores de probabilité, nous prédisons le résultat comme étant l'indice de classe ayant un score de probabilité maximal. Le principe de la stratégie un contre tous est illustré à la Figure 3.9.

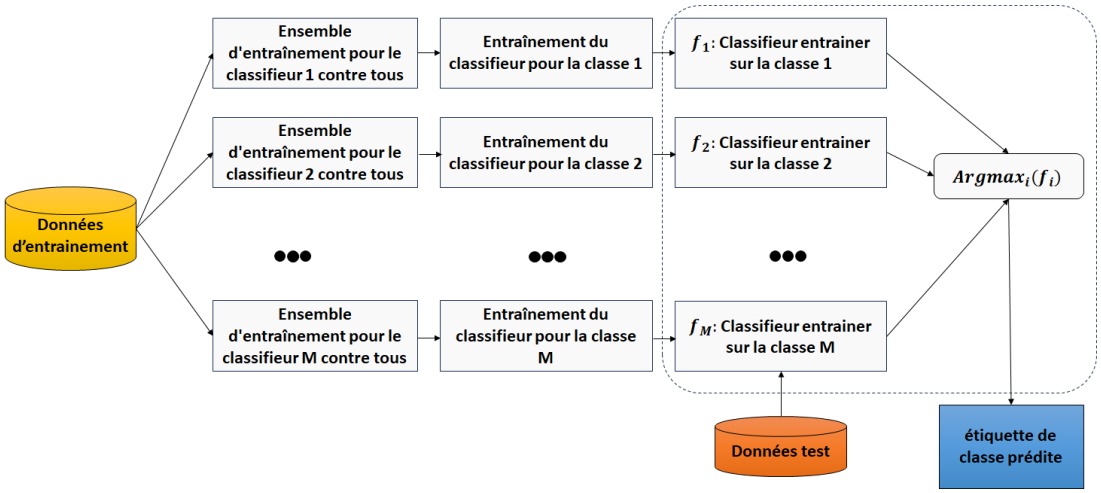


Figure 3.9 – Principe d'entrainement "un contre tous".

3.2.2 Critères d'évaluation

Afin d'analyser les performances de notre modèle, différentes métriques sont utilisées, notamment l'exactitude (Accuracy), la précision (Precision), le rappel (Recall), le score F1 (F1-score) et la matrice de confusion. Ces métriques sont basées sur les vrais positifs (T_P), les vrais négatifs (T_N), les faux positifs (F_P) et les faux négatifs (F_N).

— **La matrice de confusion:** La matrice de confusion résume les performances globales du modèle sous forme de structure tabulaire NxN, où N représente le nombre de classes. Les vraies étiquettes sont représentées par des lignes, tandis que les étiquettes prédites sont représentées par des colonnes. La Figure 3.10 illustre un exemple de matrice de confusion.

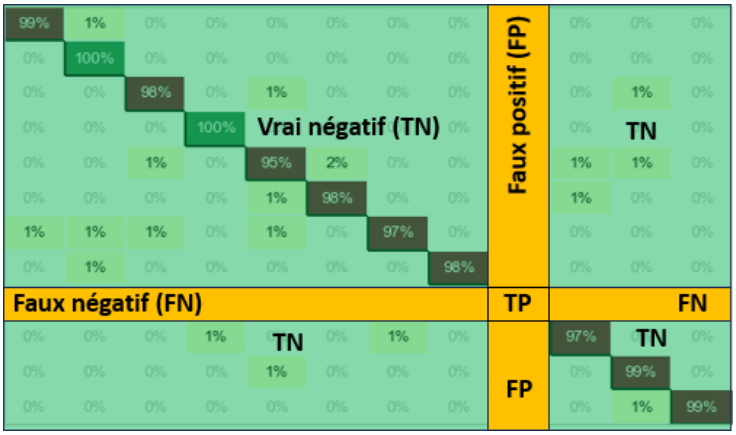


Figure 3.10 – Exemple de matrice de confusion.

Chaque cellule de la matrice de confusion représente un facteur d'évaluation.

— Vrai positif (True Positif : T_P) : résultat dans lequel le modèle prédit qu'une observation appartient à une classe et l'observation appartient effectivement à cette classe.

— Vrai négatif (True Negatif : T_N) : résultat dans lequel le modèle prédit qu'une observation n'appartient pas à une classe et qu'elle n'appartient effectivement pas à cette classe.

— Faux positif (False Positif : F_P) : résultat dans lequel le modèle prédit qu'une observation appartient à une classe alors qu'elle n'y appartient pas.

— Faux négatif (False Negatif : F_N) : résultat dans lequel le modèle prédit qu'une observation n'appartient pas à une classe alors qu'elle y appartient en réalité.

— **L'exactitude (Accuracy):** L'exactitude est le rapport entre le nombre de prédictions correctes et le nombre total de prédictions effectuées, définie comme suit :

$$Accuracy = \frac{T_P + T_N}{T_P + T_N + F_P + F_N} \tag{3.10}$$

— **La précision (Precision):** La précision est le rapport entre le nombre de prédictions

positives correctes et le nombre total de prédictions positives, définie comme suit :

$$Precision = \frac{T_P}{T_P + F_P} \tag{3.11}$$

— **Le rappel (Recall):** Le rappel est le rapport entre le nombre de prédictions positives correctes et le nombre total de prédictions dans la classe réelle.

$$Recall = \frac{T_P}{T_P + F_N} \tag{3.12}$$

— **Le F1-score:** Le F1-score est la moyenne harmonique du rappel et de la précision du modèle.

$$F1 - score = \frac{2T_P}{2T_P + F_P + F_N} \tag{3.13}$$

En outre, nous avons utilisé deux approches graphiques comprenant la courbe de Précision-Rappel (PR) et la courbe Receiver Operating Characteristic (ROC), car il s'agit d'indicateurs plus robustes. La courbe PR est obtenu en traçant la précision en fonction du rappel, tandis que la courbe ROC est obtenu en traçant le taux de vrais positifs (True Positif Rate : TPR) en fonction du taux de faux positifs (False Positif Rate : FPR) pour différents seuils. Le TPR est synonyme de rappel, et le FPR est défini comme suit :

$$FPR = \frac{F_P}{F_P + F_N} \tag{3.14}$$

3.2.3 Entrainement et évaluation

L'entrainement du modèle est une étape cruciale qui implique plusieurs processus clés. La première étape consiste à diviser l'ensemble de données : un ensemble d'entrainement (Train), dont les données sont utilisées pour l'entraînement du modèle, et un ensemble test, réservé uniquement à l'évaluation des performances du modèle. La répartition généralement préférée est la suivante : 20% de données de test, 80% de données d'entraînement. Pour ce faire, nous utilisons la fonction train_test_split() incluse dans le package python sklearn. Cette fonction mélange aléatoirement les données avant de le diviser en deux parties (voir Figure 3.11).

Certains modèles nécessitent un ajustement de leurs hyperparamètres afin d'obtenir les meilleurs résultats. Un troisième ensemble doit être déduit à partir de l'ensemble de données, appelé ensemble de validation. Ce dernier est utilisé pour évaluer la précision initiale du modèle, observer son processus d'apprentissage et affiner les hyperparamètres.

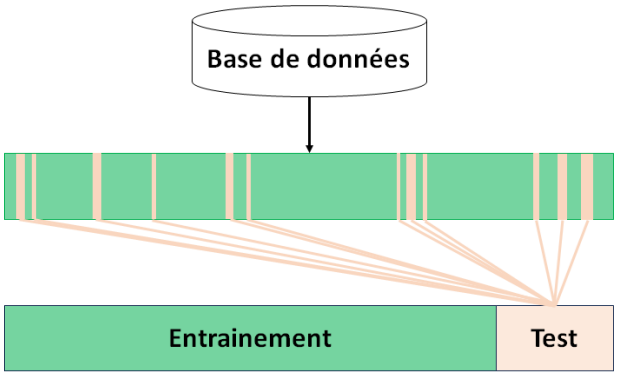


Figure 3.11 – Partitionnement de l'ensemble de données en deux parties ; apprentissage et teste.

— **Entrainement du CNN-LSTM à trois entrées**

L'entrainement de l'extracteur des caractéristiques CNN-LSTM est effectué par l'algorithme Back propagation [169]. Les paramètres du modèle sont initialisés de manière aléatoire au début et la fonction de perte est utilisée pour mesurer l'écart entre les probabilités de classe prédites et les vraies étiquettes par l'optimiseur Adam. La fonction de perte entropie catégorielle éparse (Sparse Categorical Crossentropy) est choisie compte tenu que les étiquettes d'entrées sont codées en nombres entiers plutôt que de vecteurs. Cela permet d'économiser du temps en mémoire et en calculs lorsqu'il s'agit d'un grand nombre de classes. La formule de la cross entropie catégorielle éparse est la suivante :

$$J(w) = -\sum_{i=0}^{M} y_{pred} log(y) \tag{3.15}$$

où $J(w)$ représente la valeur de la perte, M représente le nombre de classes, y_{pred} est la distribution de probabilité prédite et y représente les vraies étiquettes de classe.

— **Entrainement du classifieur SVM multi-classes**

L'un des principaux problèmes pratiques rencontrés lors de l'implémentation du modèle de classification est la découverte de la combinaison idéale d'hyperparamètres qui minimise ou maximise l'erreur. La création d'un espace de recherche statique peut représenter un défi important si l'on considère le grand nombre de paramètres combinés. Le processus d'apprentissage du classifieur SVM multi-classes est réalisé à l'aide de Optuna [170] pour tester différentes combinaisons d'hyperparamètres. Optuna est un framework d'optimisation open-source qui nous permet de réaliser des expériences complexes rapidement, efficacement, impérativement et dynamiquement. Il existe différents logiciels

d'optimisation, mais les aspects suivants ont été pris en compte lors du choix d'Optuna : (i) logiciel open-source ; (ii) langage Python ; (iii) espace de recherche dynamique ; (iv) suffisamment léger pour fonctionner sur des ordinateurs portables-Jupyter Notebook et (v) interface web (tableau de bord) pour visualiser les historiques et les résultats de l'optimisation.

Au cours du processus d'apprentissage, le réglage des hyperparamètres du classifieur SVM, y compris le coefficient de pénalité C, la fonction noyau, la variable de relâchement ξ_i et le paramètre γ, est réalisé sur la base de plusieurs essais répétés en utilisant la validation croisée k.

La validation croisée consiste à entraîner et à valider le modèle sur différentes portions de l'ensemble de données Train (voir Figure 3.12). En divisant l'ensemble Train en k parties, nous pouvons entraîner notre modèle sur les quatre premières et le valider sur la cinquième. Le processus sera ensuite répété pour toutes les configurations possibles. Ainsi, le modèle est entraîné et validé k fois. Ensuite, le réglage des hyperparamètres est mis en œuvre sur la base des résultats de validation calculés en moyenne sur les expériences. Le modèle avec la meilleure performance globale est choisi comme modèle final. Pour l'évaluation, le modèle affiné est évalué sur l'ensemble de test.

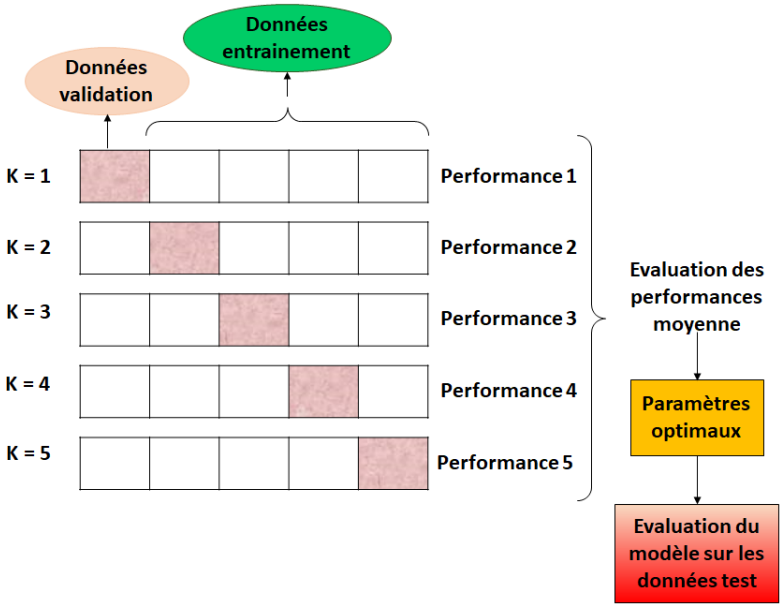


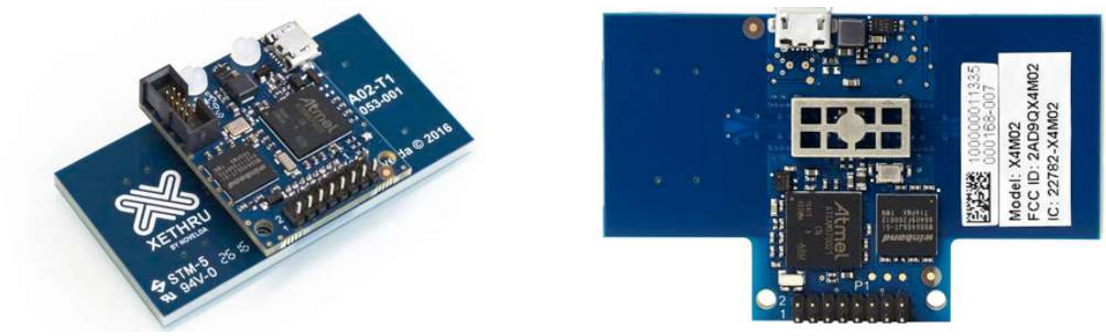
Figure 3.12 – Exemple de validation croisée avec k = 5.

3.3 Contribution 1 : Amélioration de la reconnaissance dynamique des gestes de la main en utilisant la concaténation de caractéristiques via un modèle hybride à entrées multiples

3.3.1 Base de données

Pour la RGM, nous avons utilisé le tout premier ensemble de données publiques (appelé UWB-Gestures) recueillies à l'aide de radars à impulsions ultra-large bande (UWB) [101]. L'aspect intéressant et unique de cet ensemble de données est qu'il contient la plupart des gestes précédemment utilisés dans les études de reconnaissance des gestes basées sur le radar. Pour construire l'ensemble de données UWB-Gestures les auteurs ont utilisé trois radars UWB XeThru X4 de Novelda (Norvège) disponible sur le marché, illustré à la Figure 3.13. Le XeThru X4 est fabriqué à partir de matériaux semi-conducteurs avancés et de composants électroniques. Le principal matériau utilisé pour sa construction est généralement une combinaison de silicium et d'autres matériaux semi-conducteurs, utilisés pour fabriquer les circuits intégrés et autres composants électroniques du capteur radar.

Les paramètres des radars utilisés pour la collecte des données sont indiqués sur le tableau 3.2. Les trois radars utilisés sont situés à trois positions, à gauche, à droite et en haut, comme le montre la Figure 3.15.



(a) **(b)**

Figure 3.13 – Modèle IR-UWB Xethru X4 (a) face avant, (b) face arrière.

Les auteurs ont commencé par effectuer un pré-traitement pour atténuer l'impact du bruit et des parasites sur les données. Plus précisément, ils ont utilisé le filtrage en boucle sur les données brutes pour supprimer les parasites. Cela a permis d'obtenir des matrices de

Tableau 3.2 – Spécifications techniques des radars utilisés.

Paramètre	**Spécification**
Exactitude (Accuracy)	$\sim$ 1 millimètre
Fréquence centrale	8.745 GHz
Fréquence d'échantillonnage	23 GHz
Fréquence de trame	20 trames/seconde
Bande passante (-10 dB)	1.5 GHz
Fréquence de répétition des impulsions	40.5 MHz
Largeur du faisceau d'antenne	65 degrés
Nombre d'antenne	1 paire d'émetteur Tx et récepteur Rx

temps-lent temps-rapide pour représenter les gestes. Les données ont été recueillies auprès de huit participants différents afin d'introduire davantage de variations intra-gestuelles. L'âge moyen des participants est de 25,75 ans et l'indice de masse corporelle moyen (IMC) était de 22,19 $\pm$ 5 kg/m2. Pour capturer les gestes de la main, chaque participant lui a été demandé d'effectuer douze gestes prédéterminés représentés sur la Figure 3.14, à savoir : glisser à gauche, à droite (LR), glisser de droite à gauche (RL), glisser de gauche à droite (RL), gauche-droite (LR), droite-gauche (RL), haut-bas (UD), bas-haut (DU), diagonale (diag)-LR-UD, (diag)-LR-UD swipe, diag-LR-DU swipe, diag-RL-UD swipe, diag-RL-DU dans le sens des aiguilles d'une montre, dans le sens inverse des aiguilles d'une montre, en poussant vers l'intérieur et en faisant un geste vide. Chaque geste est répété 100 fois. Chaque type de geste comporte alors 2400 données. Pendant le processus de collecte de données, les participants se tenaient dans une zone située à 1,5 mètre des trois radars qui leur faisaient face.

3.3.2 Objectifs

L'objectif de notre proposition sur la reconnaissance des gestes de la main est l'amélioration des résultats déjà publiés dans la littérature [101]. À des fins de comparaison adéquate, nous avons suivi la même procédure que les travaux antérieurs, et n'avons utiliser que les données du radar gauche.

3.3.3 Implémentation

Le schéma du système de RGM proposé est décrit dans la Figure 3.16.

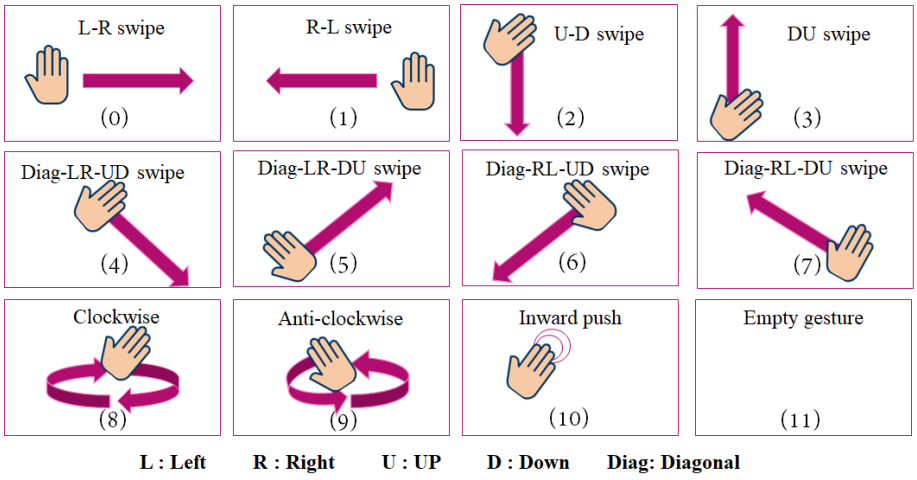


Figure 3.14 – Les douze gestes de la main de l'ensemble de données UWB-Gestures.

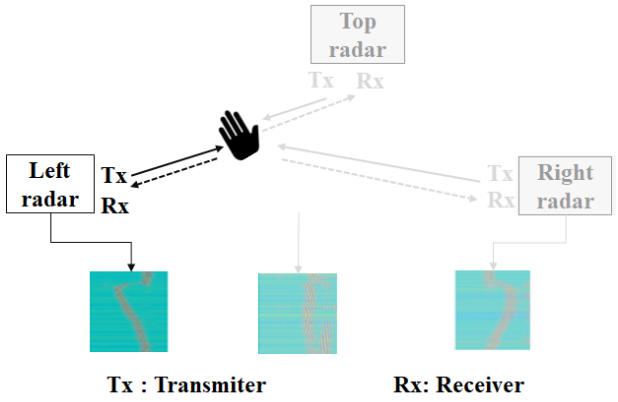


Figure 3.15 – L'emplacement des trois IR-UWB utilisés pour la collecte des données.

Le modèle CNN-LSTM-SVM à trois entrées est mis en œuvre en Python à l'aide du frameworkKeras avec Tensorflow sur une machine exécutant un environnement avec un CPU Intel (R) Core (TM) i5 2,40 GHz, 16 Go de RAM, 1 To de disque dur et Windows 10. Bien que le framework Optuna permet la création automatique d'un ensemble de validation, nous nous sommes contentés de diviser la base de données uniquement sur deux partie train et test. Les échantillons du jeu de données sont répartis aléatoirement en 80% pour l'apprentissage et 20% pour le test à l'aide de la fonction train_test_split incluse dans le package python sklearn. En utilisant la validation croisée sous Optuna avec k = 5, $\frac{100}{k}\%$ = 20% des données train sont utilisées pour la validation. Le paramètre "randomseed" est également utilisé pour s'assurer que les échantillons de test sont les mêmes dans chaque expérience réalisée. Pendant le processus d'apprentissage du CNN-LSTM à trois entrées,

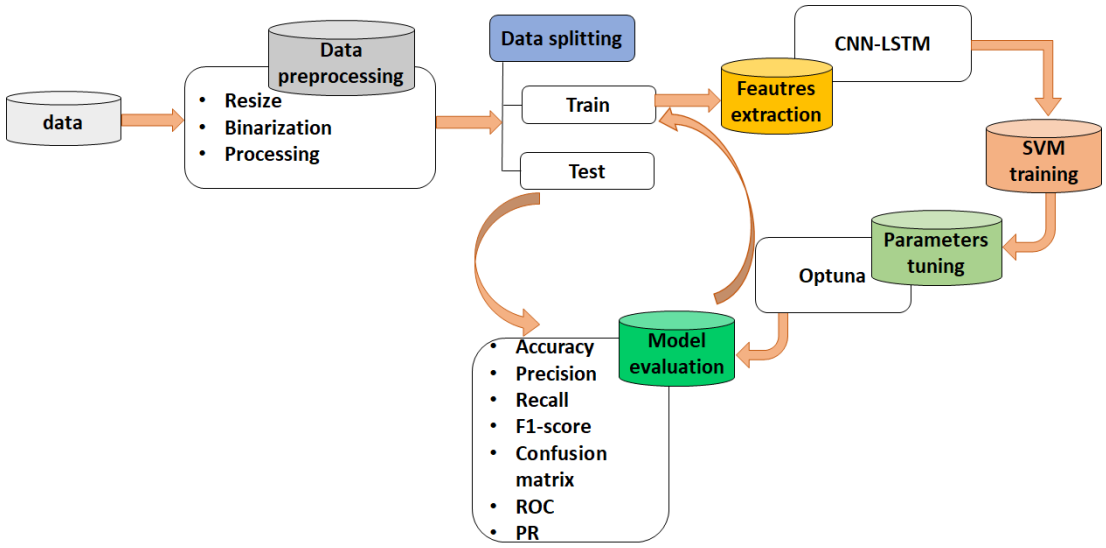


Figure 3.16 – Structure générale du système de reconnaissance des gestes de la main proposé.

l'optimiseur Adam est utilisé avec un taux d'apprentissage fixé à 0,001, une taille de lot (batch-size) de 16. Le processus d'entrainement est divisé en trois phases principales :

- Le CNN-LSTM est entraîné sur 25 itérations avec un classifieur SoftMax pour l'extraction des caractéristiques spatiotemporelles.
- Le SoftMax est remplacé par le SVM multi-classes. Ce dernier est entraîné et optimisé sous Optuna en utilisant la validation croisée k, avec k = 5 pour 100 essais.
- Le SVM multi-classes est ensuite entraîné une seconde fois avec les hyper-paramètres optimaux.

Après que l'apprentissage soit achevé, le processus d'évaluation réalisé a consisté à évaluer les performances du modèle sur les données de test.

3.3.4 Résultats

Dans cette section, nous présentons les résultats des études réalisées dans la section de mise en œuvre. Nous présentons d'abord les résultats de la phase d'apprentissage du modèle sur les données d'entrainement. Nous présentons ensuite les résultats de l'évaluation du modèle sur des données de test. Enfin, nous présentons les résultats des comparaisons sur différents tests afin de valider notre approche et de prouver sa supériorité par rapport aux approches existantes dans la littérature.

3.3.4.1 Apprentissage

Plusieurs tests ont été effectués pour obtenir les bons hyperparamètres pour le classifieur SVM. Ces paramètres comprennent le noyau, le degré du noyau et le paramètre de régularisation C. Le meilleur modèle est celui qui présente la précision la plus élevée et le taux d'erreur le plus faible sous différentes combinaisons des hyperparamètres. Le taux d'apprentissage le plus élevé pour le SVM multi-classes affiné est de 99,62% comme illustré sur la Figure 3.17. L'influence des différentes valeurs d'hyperparamètres sur les performances du modèle est présentée sur la Figure 3.18.

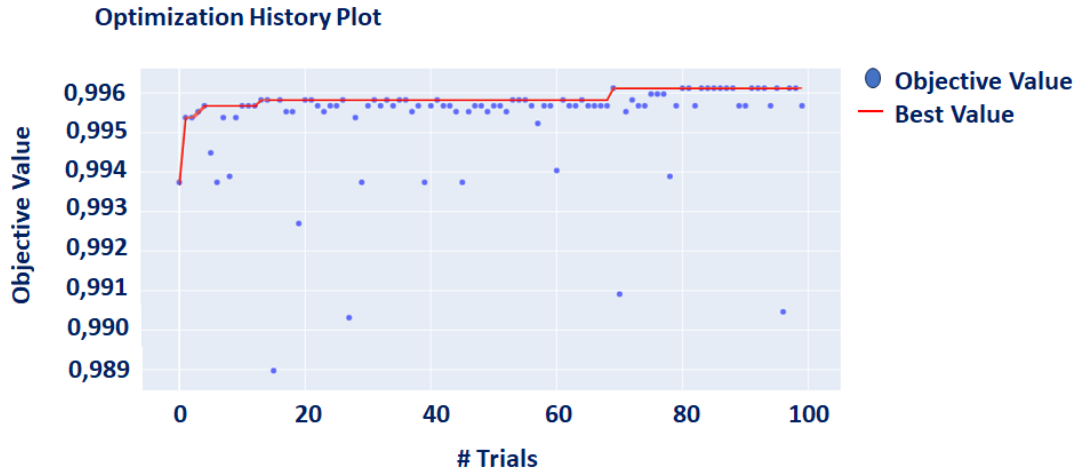


Figure 3.17 – Historique d'optimisation.

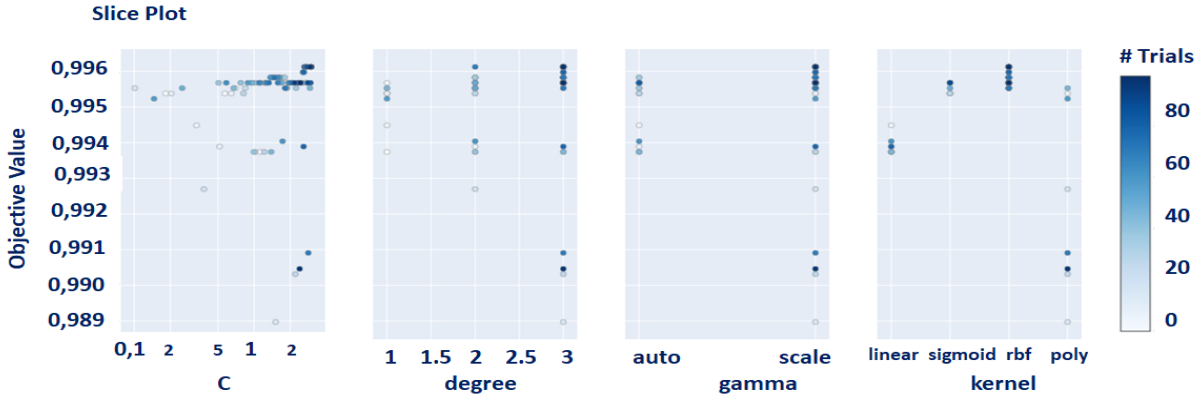
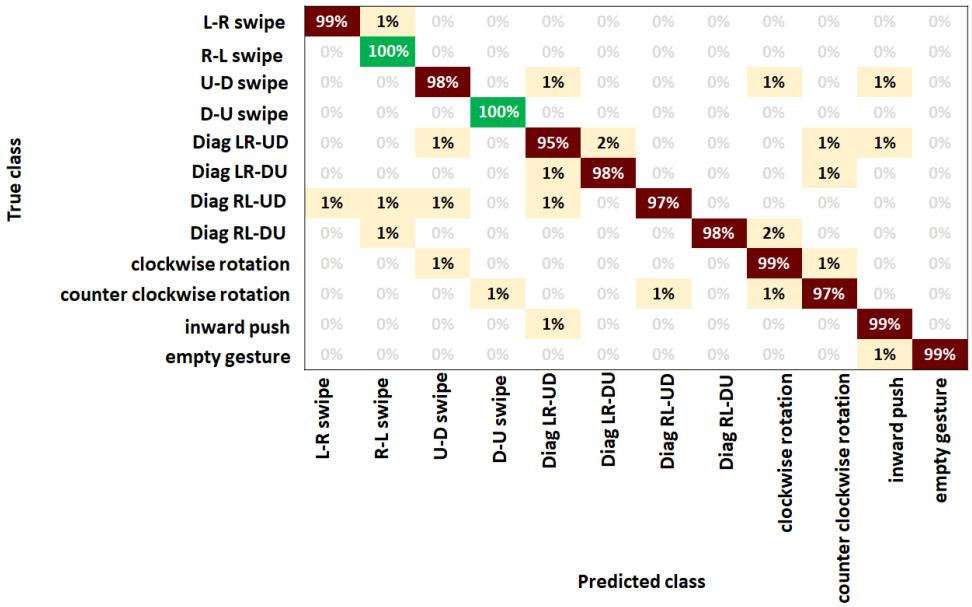


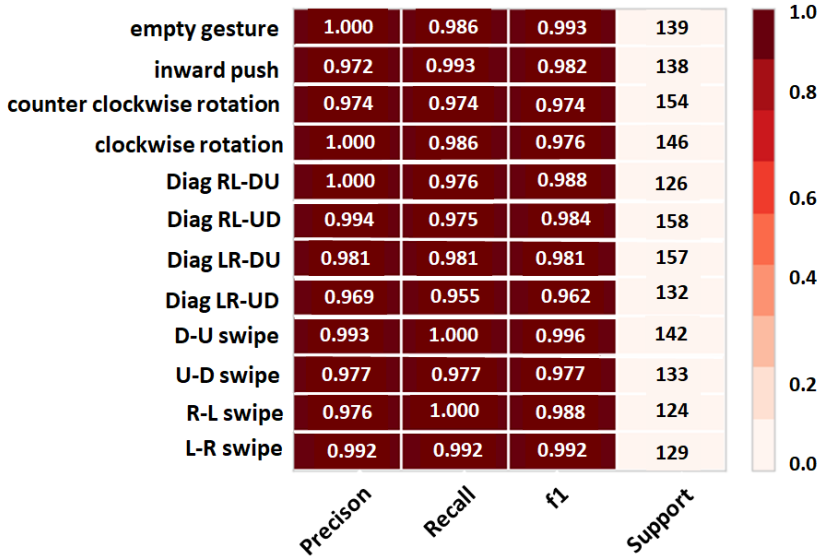
Figure 3.18 – Graphe des différentes valeurs d'hyperparamètres sur les performances du modèle.

3.3.4.2 Evaluation

Le modèle CNN-LSTM-SVM à trois entrées est évalué sur l'ensemble de test. La matrice de confusion et le rapport de classification obtenus à partir des données de test sont représentés ci-dessous (voir Figure 3.19). En outre, les courbes ROC et PR sont tracées pour comparer les performances globales (voir Figure 3.20).

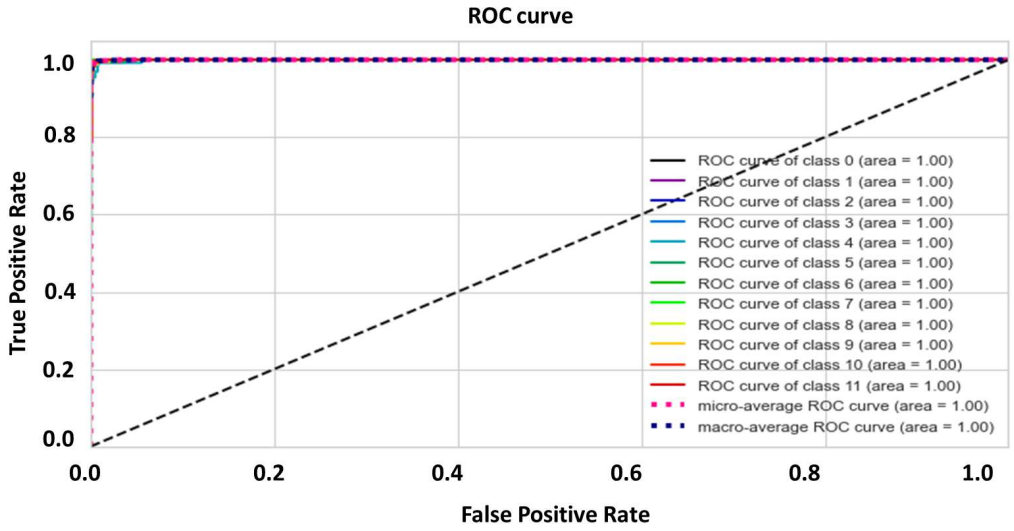


(a)

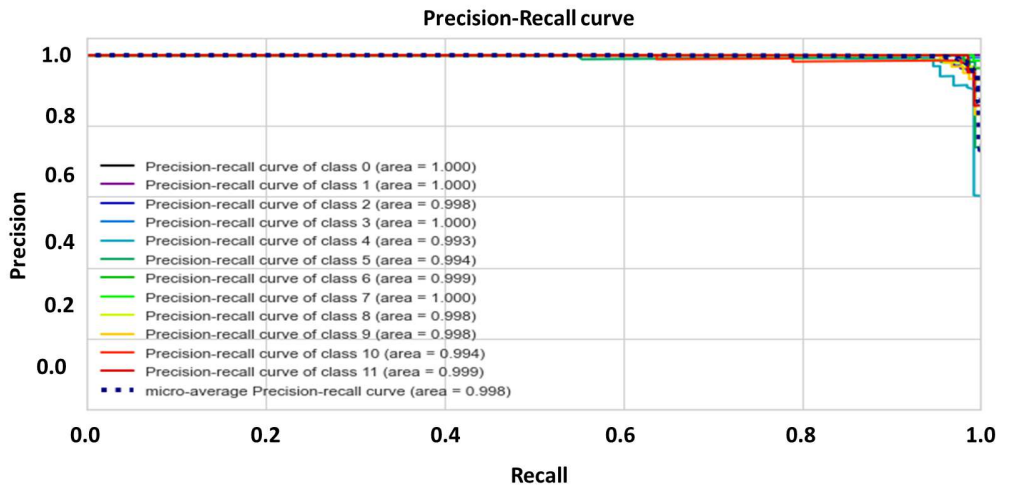


(b)

Figure 3.19 – (a) Matrice de confusion, (b) Rapport de classification.



(a)



(b)

Figure 3.20 – (a) Courbe ROC, (b) Courbe PR.

3.3.4.3 Comparaison

— **Vérification du traitement des données**

Pour démontrer l'efficacité de l'approche de traitement des données proposée, la première expérience est divisée en deux parties principales. Dans la première partie, le modèle CNN-LSTM-SVM à trois entrées est entraîné à l'aide d'images brutes (images originales sans traitement), où le même échantillon est fourni simultanément aux trois branches du CNN. Dans la seconde partie, le modèle est entraîné à l'aide d'images étendues, où chaque

branche du CNN est alimentée par une représentation différente des caractéristiques de bas niveau. Les résultats sont présentés dans le tableau 3.3.

Tableau 3.3 – Comparaison des performances de classification sur les images originales et prétraitées.

Métrique	**Image originales**	**Image pré-traitées**
Train Accuracy %	98.34	99.62
Test Accuracy %	92.49	98.27
Precision %	92.77	98.30
Recall %	92.40	98.29
F1-score %	92.44	98.27

— **Vérification de la structure du modèle**

La deuxième expérience vise à démontrer la supériorité de l'utilisation de plusieurs branches pour le traitement parallèle des données par rapport à une seule branche pour le traitement séquentiel des données, lors de l'utilisation de différentes représentations des caractéristiques. Nous avons alimenté le CNN-LSTM-SVM à entrée unique avec les images étendues et avons comparé ses performances aux résultats obtenus lors de l'expérience précédente. Notons que le CNN-LSTM-SVM à une entrée utilisée dans cette expérience consiste en les mêmes couches et la même configuration de paramètres qu'une seule branche CNN du CNN-LSTM-SVM à trois entrées. Les résultats sont présentés dans le tableau 3.4.

Tableau 3.4 – Comparaison des performances de classification du CNN-LSTM-SVM à entrée unique/trois entrées.

Métrique	**CNN-LSTM-SVM à entrée unique**	**CNN-LSTM-SVM à trois entrée**
Train Accuracy %	96.30	99.62
Test Accuracy %	93.09	98.27
Precision %	93.24	98.30
Recall %	93.10	98.29
F1-score %	93.44	98.27

— **Comparaison avec les approches de pointe**

La troisième expérience vise à comparer les performances de classification du modèle CNN-LSTM-SVM à trois entrées avec d'autres modèles, notamment les modèles CNN-SoftMax à trois entrées et CNN-LSTM-SoftMax à trois entrées, définis par modèle 1 et modèle 2. Ces derniers représentent des versions initiales avant d'aboutir à la version finale du modèle. Cette expérience est réalisée pour démontrer l'impact de l'utilisation de caractéristiques

spatiales uniquement par rapport à l'utilisation de caractéristiques spatiotemporelles sur le taux de reconnaissance. En outre, nous cherchons à trouver le classifieur optimal pour le modèle en comparant SoftMax et SVM. Finalement, les résultats du modèle que nous proposons peuvent être directement comparés à ceux de Ahmed *et al.* [101] et de Noori *et al.* [100], puisqu'ils ont également utilisé le même ensemble de données. Les résultats sont présentés dans le tableau 3.5.

Tableau 3.5 – Comparaison des performances de classification du CNN-LSTM-SVM à trois entrées avec les méthodes les plus récentes.

Référence	**Modèle**	**Test Accuracy (%)**	**N° de paramètres**
Ahmed *et al.* [101]	CNN	94	-
Notre approche : Modèle 1	Three input-CNN-SoftMax	95.41	126300
Noori *et al.* [100]	LSTM	97	847055
Notre approche : Modèle 2	Three input-CNN-LSTM-SoftMax	97.20	331596
Notre approche : Modèle final	Three input-CNN-LSTM-SVM	98.27	332578

3.3.5 Discussion

Ce travail présente un modèle hybride de bout en bout pour la classification des gestes dynamiques de la main à l'aide de données IR-UWB. Nous avons utilisé un CNN-LSTM à trois entrées pour l'apprentissage automatique des caractéristiques spatiotemporelles, combiné à un classifieur SVM multi-classes. Les performances du modèle proposé ont été évaluées à l'aide de diverses méthodes. Les résultats de la matrice de confusion et du rapport de classification présentés à la Figure 3.19 montrent que le modèle a atteint une exactitude de 98,27% et a identifié la moitié des classes avec une exactitude supérieure à 98 %, une précision de 98,30%, un rappel de 98,29% et un score-F1 de 98,27%. En outre, la courbe ROC de la Figure 3.20 (a) indique que le modèle proposé basé sur des caractéristiques combinées a produit d'excellents résultats, avec un taux de vrais positifs pour chaque classe de geste.

Le tableau 3.3 montre clairement que le modèle CNN-LSTM-SVM à trois entrées est surajusté lors de la duplication des entrées à l'aide d'images brutes. Le modèle a atteint une exactitude d'apprentissage de 98,34%, alors qu'une exactitude de test plus faible de

92,49%. Cela est dû à l'ensemble de données non représentatif, qui ne fournit pas suffisamment d'informations pour permettre au modèle de bien généraliser. Pour résoudre ce problème et éviter que le modèle ne soit surajusté, nous avons utilisé l'ensemble de données étendu qui comprend les images de gradients dans la direction x, dans la direction y et dans les directions x et y à la fois. Comme le montre le tableau 3.3, le modèle CNN-LSTM-SVM à trois entrées permet une meilleure reconnaissance des gestes lorsque différentes représentations de bas niveau sont fournies comme entrées, ce qui entraîne une augmentation de l'exactitude d'environ 6% par rapport à l'utilisation d'images originales dupliquées. Le modèle a atteint une exactitude d'apprentissage et de test respectivement de 99,62% et 98,27%, avec un sur-ajustement beaucoup plus faible et une meilleure capacité de généralisation. En effet, l'utilisation de représentations multiples pour un même geste permet d'extraire et de fusionner davantage de caractéristiques, ce qui fournit au classifieur final plus d'informations pour prendre une décision. Ces résultats soulignent l'importance d'utiliser différentes représentations de bas niveau ainsi qu'un modèle à entrées multiples dans les cas où les caractéristiques représentatives sont insuffisantes, et indiquent que la duplication inutile des entrées n'améliore pas significativement les performances mais augmente la complexité du modèle.

La supériorité de notre méthode est plus évidente lorsque nous comparons les performances des modèles à une entrée et à trois entrées. Les résultats du tableau 3.4 montrent une différence significative dans les performances du modèle lors du traitement séquentiel et indépendant des données en parallèle. Le modèle CNN-LSTM-SVM à une entrée a atteint une exactitude de 93,09%, tandis que le modèle CNN-LSTM-SVM à trois entrées a atteint une exactitude de 98,27%. Par rapport au CNN-LSTM-SVM à une entrée, l'architecture CNN-LSTM-SVM à trois entrées que nous proposons non seulement tire parti de la force de plusieurs représentations de bas niveau pour extraire des caractéristiques complémentaires de la même cible, mais introduit également le concept de concaténation des caractéristiques dans l'architecture pour obtenir des représentations plus holistiques. Le traitement séparé de chaque représentation de données permet d'extraire et de préserver ses caractéristiques sans les altérer avec d'autres représentations de données, ce qui permet au modèle d'apprendre efficacement des caractéristiques distinctes, discriminantes et complémentaires. Nous pouvons conclure que dans des conditions de données limitées, l'extension des données brutes à différentes représentations de caractéristiques et les fournitures en tant qu'entrées devraient être appliquées à des branches distinctes afin d'obtenir une plus grande précision. Nous émettons également l'hypothèse

que la raison de la faible performance du modèle CNN-LSTM-SVM à entrée unique est due à des caractéristiques dissemblables, où le même geste est représenté trois fois différemment, ce qui entraîne une confusion pour le modèle et le rend sujet à l'erreur. Nous pouvons raisonnablement conclure de cette expérience que, grâce à la concaténation efficace d'informations multiples, la prédiction est faite sur la base de l'utilisation complète des caractéristiques cibles, ce qui améliore la précision de la reconnaissance.

D'autre part, comme le montre le tableau 3.5, le CNN-LSTM-SVM à trois entrées proposé surpasse les travaux de recherche les plus récents [100, 101]. Par rapport au CNN proposé par Ahmed *et al.* [101], le CNN-SoftMax à trois entrées a obtenu une augmentation moyenne de l'exactitude de 1,41%. Ce résultat suggère que le CNN-SoftMax à trois entrées peut capturer plus d'informations sur le contexte spatial pour la classification en apprenant les détails lorsqu'un grand nombre d'échantillons d'entraînement est fourni. Bien que l'expansion des données contribue au modèle, la précision reste limitée par le manque d'informations supplémentaires. Bien que la structure CNN puisse apprendre des caractéristiques de plus haut niveau, elle ignore les dépendances temporelles des caractéristiques, ce qui signifie que les entrées et les sorties sont indépendantes, ce qui conduit à des performances de reconnaissance limitées.

Le tableau 3.5 montre que l'ajout d'unités LSTM peut améliorer les performances de classification. La combinaison CNN-LSTM-SoftMax à trois entrées a montré son efficacité en atteignant une exactitude de 97,20%, surpassant de 3,20% le CNN proposé par Ahmed *et al.* [101] et de 1,79% le CNN-SoftMax à trois entrées. Cette amélioration est due aux opérations de convolution et de concaténation, ainsi qu'à la structure sophistiquée du LSTM, le tout qui maintient les relations spatiales et temporelles. Le LSTM est cascadé pour apprendre et intégrer des caractéristiques temporelles, qui peuvent fournir des informations supplémentaires et améliorer les performances de classification. Le LSTM aide à capturer et à mémoriser la façon dont les caractéristiques extraites par les couches du CNN changent au fil du temps. En combinant les forces du CNN et du LSTM, on obtient les avantages de l'apprentissage spatial et temporel, ce qui est très efficace pour améliorer le taux de reconnaissance des gestes dynamiques de la main. En outre, le modèle CNN-LSTM-SoftMax à trois entrées a fourni des performances comparables à celles des travaux antérieurs de Noori *et al.* [100]. Cependant, le modèle CNN-LSTM-SoftMax à trois entrées a atteint une exactitude de 97,20% et comportait moins de paramètres entraînables (331596) que le modèle proposé par Noori *et al.* [100], qui a été maintenu avec 847055 paramètres et a atteint une exactitude de 97%. Afin de sélectionner le classifieur

optimal pour notre modèle, le CNN-LSTM à trois entrées a été entraîné avec une couche SoftMax et un classifieur SVM. Les résultats présentés dans le tableau 3.5 montrent une augmentation de 1,07% de la précision en utilisant le SVM comme classifieur final. Ce gain est principalement dû à l'utilisation des différents hyperparamètres optimaux sélectionnés à l'aide du framework Optuna. Par conséquent, la capacité de généralisation du SVM est supérieure à celle de SoftMax.

3.4 Contribution 2 : Reconnaissance avancée des actions humaines par augmentation de données et concaténation de caractéristiques des signatures micro-Doppler

3.4.1 Base de données

Pour la RAH nous avons utilisé la base de données d'AH du réseau de capteurs multifréquence proposée par Gurbuz *et al.* [33].La base de données se compose de trois ensembles de données acquises à partir de trois capteurs radar synchronisés fonctionnant en mode monostatique. Les capteurs comprennent le radar FMCW IWR1443 de 77 GHz de Texas Instruments réglé à une fréquence centrale de 77 GHz et à une largeur de bande de 750 MHz, le radar FMCW de 24 GHz d'Ancortek avec une fréquence centrale de 24 GHz et une largeur de bande de 1500 MHz et le radar à impulsions UWB XeThru X4 transmettant sur une bande d'environ 7 GHz-10 GHz. Lors de l'acquisition des données, les capteurs ont été placés côte à côte sur une hauteur de 1 mètre du sol comme illustré sur la Figure 3.1. Six participants d'âge, de taille et de poids différents ont participé à cette étude. Au total, onze activités et marches ambulatoires différentes ont été prises en compte à savoir : WLKT (marchant vers le radar), WALKA (s'éloignant du radar), PICK (ramasser un objet du sol), BEND (se pencher), SIT (s'asseoir sur une chaise), KNEEL (s'agenouiller), CRWL (ramper vers le radar), LIMP (boitant avec une jambe droite raide), WTOES (marchant sur les deux orteils), SHTEPS (marchant à petits pas), et SCSSR (marchant avec des ciseaux). Chaque participant se déplace de 0,5 à 3 mètres le long de la ligne de visée des radars et effectue 10 répétitions de chaque action. Cela donne un total de 60 échantillons par classe et par radar. La Figure 3.21 représente les spectrogrammes micro-Doppler des différentes actions pour chaque radar.

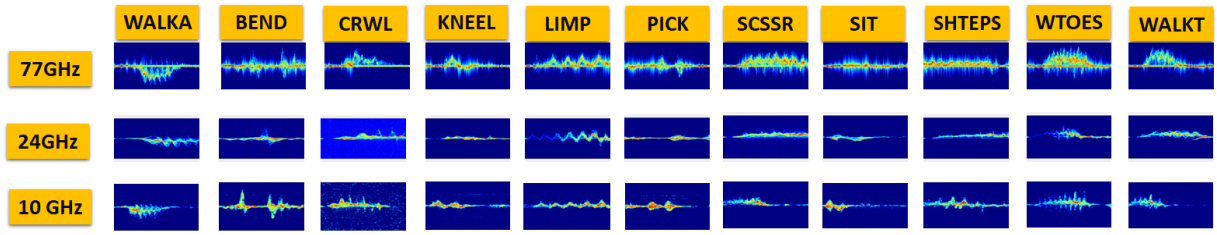


Figure 3.21 – Signature micro-Doppler pour chaque radar/action humaine [33].

3.4.2 Objectifs

Ce travail vise à développer un système de RAH qui permet d'obtenir à la fois des exigences de calcul réduites et une précision de classification élevée avec de petits ensembles de données micro-Doppler, avec deux objectifs principaux à l'esprit.

Le premier objectif est de proposer une stratégie simple et efficace pour générer des spectrogrammes micro-Doppler utilisés pour l'entraînement des algorithmes de classification. La principale contribution de cette recherche réside dans la création de nouveaux échantillons dérivés de l'ensemble de données original avec des caractéristiques qui ne se chevauchent pas. Nous nous concentrons sur le développement d'un processus d'augmentation qui garantit la préservation des caractéristiques essentielles des signatures micro-Doppler sans introduire de distorsions involontaires. En outre, nous nous efforçons d'établir une stratégie d'augmentation robuste et adaptable qui puisse être appliquée à différents ensembles de données, plutôt que de dépendre d'un ensemble spécifique. Contrairement aux études précédentes [33], notre approche se distingue par l'utilisation exclusive de techniques de manipulation d'images directement appliquées aux signatures micro-Doppler. Au lieu de recourir aux transformations d'images conventionnelles [171], nous proposons l'application de la transformée en ondelettes discrète (Discret Wavelet-Transform : DWT) comme nouvelle alternative. Cette approche unique distingue notre travail et contribue à l'avancement dans le contexte des stratégies d'augmentation des données. Les spectrogrammes micro-Doppler originaux sont projetés sur le sous-espace de la transformée en ondelettes discrète. À partir de cette projection, les spectrogrammes sont décomposés, ce qui permet d'obtenir différentes images de sous-bandes. Ensuite, les composantes de décomposition renvoyées par le processus DWT sont utilisées dans différentes configurations pour générer de nouveaux échantillons. À partir d'un seul spectrogramme, trois échantillons sont générés, ce qui augmente le nombre de données d'entraînement sans créer de doublons. Cette approche permet d'améliorer les caractéristiques des actions

humaines afin d'accroître la précision du modèle de classification sans avoir à acquérir à nouveau des données.

Le deuxième objectif est d'étendre notre travail précédent basé sur la RGM à l'application de RAH. L'idée est de réappliquer le même pré-traitement sur les données d'actions humains et réutiliser le même modèle CNN-LSTM-SVM à trois entrées pour la classification des actions.

3.4.2.1 Augmentation des données

— **Rappel sur la transformée en ondelette discrète**

La transformée en ondelettes discrète est un outil puissant utilisé dans le traitement des images pour diverses applications telles que la compression, la détection, la reconnaissance et la suppression du bruit [172]. Le principe de l'algorithme consiste à diviser une image à chaque itération en quatre blocs : trois blocs concernant les détails de l'image (coefficients de détails), et le quatrième correspondant aux informations les plus importantes pour l'œil (coefficient d'approximation).La procédure de décomposition DWT sur une image est illustrée à la Figure 3.22.

Cette décomposition implique une série d'opérations de filtrage et de mise à l'échelle (sous-échantillonnage). Ceux-ci comprennent deux filtres issus du choix d'ondelette : le filtre passe-bas utilisé pour extraire les informations approximatives et le filtre passe-haut pour extraire les détails de l'image. En effet une ondelette n'est qu'une fonction mathématique utilisée pour diviser une fonction donnée ou un signal en différentes composantes fréquentielle à différentes échelles. La DWT utilise différentes ondelettes telles que Haar, Symlet et Daubechies [172].

Dans le traitement des images, la 2D-DWT est utilisée pour effectuer des opérations sur les lignes et les colonnes des images originales. Cela équivaut à deux transformées 1D consécutives mises en œuvre sous la forme d'une transformée 1D de ligne suivie d'une transformée 1D de colonne. Dans la première étape, les filtres passe-bas (L)égale a ($\frac{1}{\sqrt{2}}$, $\frac{1}{\sqrt{2}}$) et passe-haut (H) égale a ($\frac{1}{\sqrt{2}}$, -$\frac{1}{\sqrt{2}}$), sont appliqués à chaque ligne de l'image et ensuite échantillonnée par un facteur de 2, ce qui donne deux demi-images. L'une avec des coefficients d'échelle (L) et l'autre avec des coefficients d'ondelettes (H). Les deux images correspondent à la moitié de la largeur des lignes de l'image originale. Dans la seconde étape, les filtres passe-bas (L) et passe-haut (H) sont appliqués dans le sens des colonnes des images générées par la première étape. Cela donne lieu à quatre quarts sous-bandes qui

représentent différentes gammes de fréquences et orientations spatiales au sein de l'image originale.

- Approximation (LL) : La sous-bande LL représente les composantes à basse fréquence ou l'approximation grossière de l'image (coefficients d'approximation : cA). Elle capture la structure globale ou le contenu des basses fréquences.
- Détail horizontal (LH) : La sous-bande LH représente les composantes à haute fréquence dans la direction horizontale (coefficients horizontaux cH).
- Détail vertical (HL) : La sous-bande HL représente les composantes à haute fréquence dans la direction verticale (coefficients verticaux cV).
- Détail diagonal (HH) : La sous-bande HH représente les composantes à haute fréquence dans la direction diagonale (coefficients diagonaux cD).

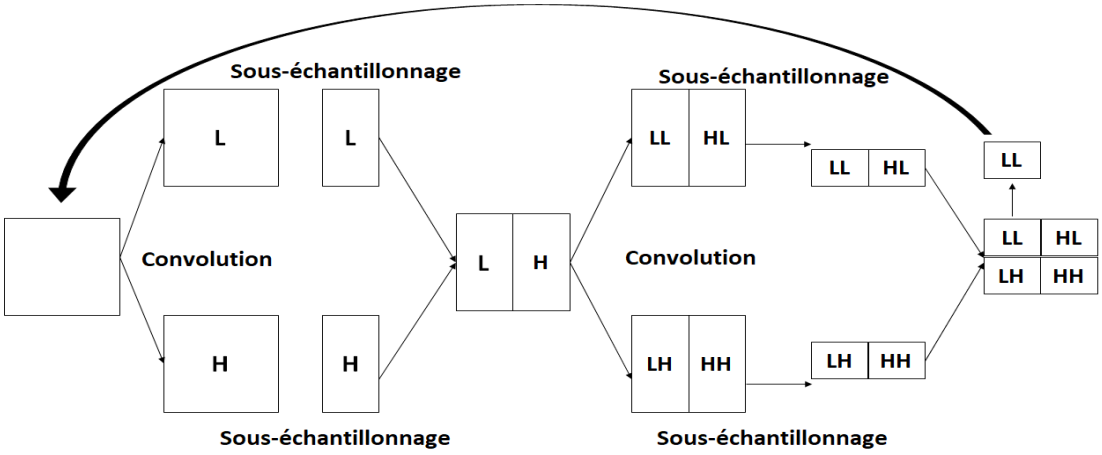


Figure 3.22 – Processus de décomposition en ondelettes discrètes 2D.

- **Processus d'augmentation des données**

Bien que les modèles d'apprentissage profond se sont avérés efficaces pour la RGM et des AH, leurs performances restent parfois insuffisantes en raison de la rareté des données [33,83]. Les modèles d'apprentissage profond sont connus pour être gourmands en données, nécessitant de grandes quantités d'échantillons d'entraînement étiquetés pour obtenir des résultats satisfaisants. Malheureusement, il est difficile de satisfaire à l'exigence d'une grande quantité de données radar. L'acquisition et l'annotation des données restent des tâches complexes, coûteuses et fastidieuses. En outre, étant donné que chaque donnée radar a ses propres caractéristiques de configuration, il n'est pas possible de combiner et d'utiliser différents ensembles de données.

L'une des meilleures façons d'y parvenir est d'utiliser l'augmentation des données qui

nous permet d'éviter la redondance des données, d'accroître la diversité des données et d'augmenter la précision de prédiction du modèle. Dans ce travail, nous nous concentrons sur la génération d'échantillons micro-Doppler 2D pour enrichir l'ensemble d'apprentissage. Nous proposons d'utiliser la DWT dans le but d'étudier sa caractéristique notable qui permet la décomposition d'une image en composantes. Nous proposons d'utiliser l'ondelette de Haar car elle offre une approche simple, plus rapide et efficace en termes de calcul. L'image originale est décomposée en quatre sous-bandes, à savoir les coefficients d'approximation (cA), les coefficients horizontaux (cH), les coefficients verticaux (cV) et les coefficients diagonaux (cD) à chaque niveau. Nous considérons qu'une décomposition à un seul niveau est une taille appropriée, car des niveaux de décomposition supplémentaires peuvent entraîner la perte d'informations utiles. À partir de l'image originale de taille NxN, le processus d'augmentation des signatures micro-Doppler est le suivant :

— Tout d'abord, nous effectuons une DWT sur l'ensemble des lignes et des colonnes des spectrogrammes. Quatre sous-images se référant aux coefficients de décomposition à un niveau sont générées, notamment cA1, cD1, cH1 et cV1 de taille $\frac{N}{2}$x$\frac{N}{2}$ chacune.

— Les informations importantes sont concentrées dans cA1. Les coefficients cH1, cV1 et cD1 peuvent facilement être perturbés par des bruits. C'est pourquoi nous proposons d'injecter chacun de ces coefficients dans cA1 et d'additionner les pixels entre les deux images. De cette façon, trois nouvelles images sont générées comme suit :

$$cA1 + cD1$$
$$cA1 + cV1$$
$$cA1 + cH1$$

En tirant parti des propriétés de décomposition de la DWT, de nouveaux échantillons dont les caractéristiques ne se chevauchent pas avec celles des images originales sont générées et utilisées pour entrainer le modèle de reconnaissance. La Figure 3.23 montre les signatures micro-Doppler générées à l'aide de la décomposition DWT.

3.4.3 Implémentation

Le schéma du système de reconnaissance des AH proposé est décrit dans la Figure 3.24.

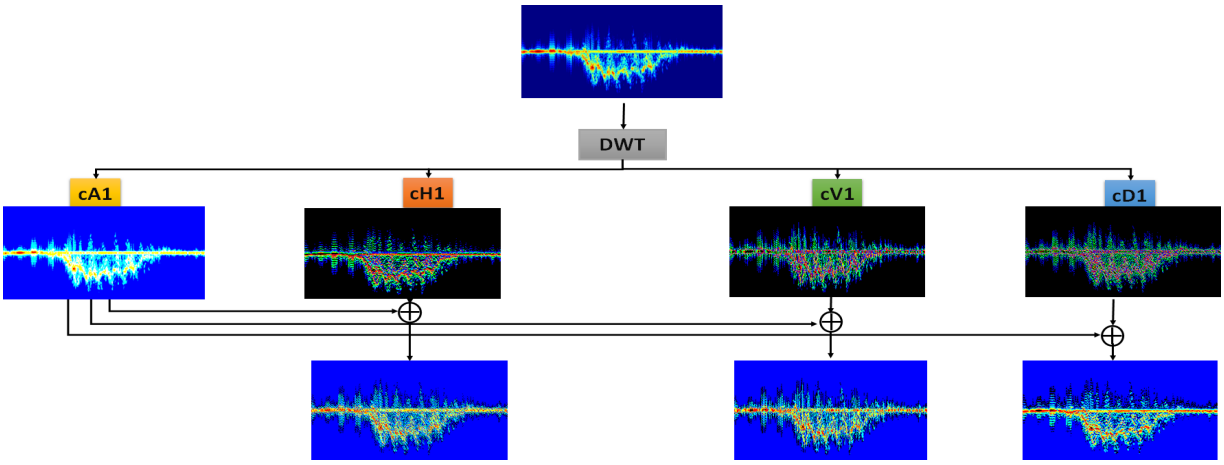


Figure 3.23 – Processus d'augmentation proposé.

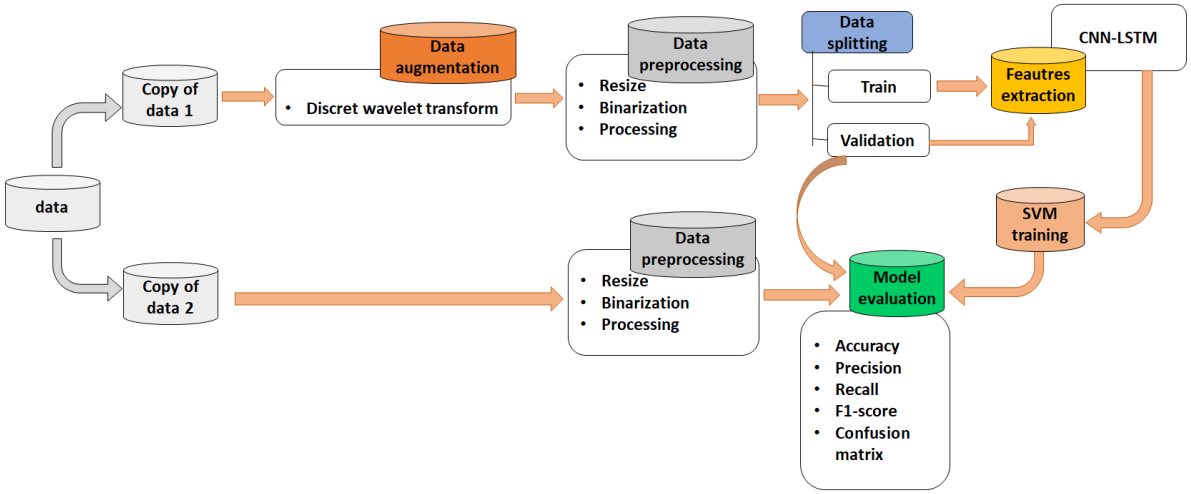


Figure 3.24 – Structure générale du système de reconnaissance des actions humaines proposé.

Bien que nous souhaitions réutiliser le même modèle de classification utilisé pour la reconnaissance des gestes de la main, tout en concevant les hyperparamètres du classifieur SVM (c.a.d ne pas ré-entraîner avec Optuna), nous avons choisis cette fois-ci de diviser les ensembles de données sur trois portions à savoir entrainement, validation et test. L'ensemble de validation est introduit cette fois-ci manuellement afin d'évaluer le comportement du modèle pendant la phase d'apprentissage. Lorsque l'on utilise les ensembles de données originaux avec un nombre restreint d'échantillons, l'ensemble de validation sert à identifier le sur-ajustement. En outre, lors de l'utilisation d'échantillons augmentés, l'ensemble de validation a un double objectif. Premièrement, il permet de s'assurer que le

modèle est capable d'apprendre efficacement les échantillons augmentés. Deuxièmement, il permet de s'assurer que les caractéristiques des échantillons augmentés s'alignent avec celles des données de test. Cela aide le modèle à gérer les variations dans les données en atteignant une grande précision et empêche l'ajustement excessif.

La mise en œuvre de la technique d'augmentation proposée est effectuée à l'aide de Matlab R2021 sur une machine fonctionnant dans un environnement doté d'un processeur Intel (R) Core (TM) i5 2,40 GHz, de 16 Go de RAM, d'un disque dur de 1To et de Windows 10. Le reste des opérations, y compris l'entraînement, la validation et le test du modèle de classification, a été exécuté sur Google Collaboratory.

Pour évaluer notre proposition dans l'analyse expérimentale, quatre expériences sont réalisées pour chaque ensemble de données :

- **Expérience 1:** utilisation des ensembles de données originaux sans augmentation ni pré-traitement. Chaque ensemble de données est divisé en 80% pour l'entrainement, 10% pour la validation et 10% pour le test. Afin d'effectuer une comparaison concise, nous utilisons le même ensemble de test dans toutes les expériences. À cette fin, nous utilisons le paramètre de random seed pour garantir la même partition.
- **Expérience 2:** utilisation des ensembles de données originaux pré-traités sans augmentation. Chaque ensemble de données est divisé en 80% pour l'entrainement, 10% pour la validation et 10% pour le test, comme dans l'expérience 1. Chaque branche du CNN est alimentée par Dx, Dy et Dxy, respectivement.
- **Expérience 3:** utilisation des échantillons augmentés pour l'entrainement et des échantillons originaux pour le test. Les échantillons augmentés sont binarisés et répartis en 90% pour l'entrainement et 10 % pour la validation. Deux tests sont réalisés. Test (a) : Utiliser uniquement la portion de 10% de l'ensemble de données original comme ensemble de test, comme indiqué dans l'expérience 1. Test (b) : pour éviter une évaluation biaisée, nous testons le modèle sur la totalité de l'ensemble de données original, y compris tous les échantillons.
- **Expérience 4:** utilisation des échantillons augmentés et pré-traités pour l'entrainement et des échantillons originaux pré-traités pour le test. Le modèle est entraîné à l'aide d'images augmentées étendues, où chaque branche du CNN est alimentée par Dx, Dy et Dxy, respectivement. Pour le test, nous suivons la même procédure que pour l'expérience 3.

Notez que le modèle pour toutes les expériences est formé à partir de zéro pour 100 itérations avec une taille de lot de 16 en utilisant l'optimiseur Adam avec un taux d'apprentissage

fixé à 0,001. Toutes les configurations par défaut du modèle sont laissées intactes, comme indiqué sur le tableau 3.1. À l'exception de la couche LSTM, le nombre d'unités a été augmenté à 300 pour obtenir de meilleurs résultats. Pour le classifieur SVM, le nombre de classifieurs binaires est fixé à 11, ce qui correspond au nombre de classe d'AH.

3.4.4 Résultats

Les résultats de chaque expérience dans le cadre proposé sont présentés dans cette section. Chaque sous-section présente les résultats séparément pour chaque ensemble de données.

3.4.4.1 L'ensemble de données IR-UWB 10 GHz

Les résultats sont présentés dans le tableau 3.6. La matrice de confusion et le rapport de classification obtenus grâce à la combinaison de l'augmentation et du pré traitement appliqués aux données de test sont décrits dans la Figure 3.25 (a) et (b), respectivement.

Tableau 3.6 – Performances comparatives de la classification sur l'ensemble de données de 10 GHz.

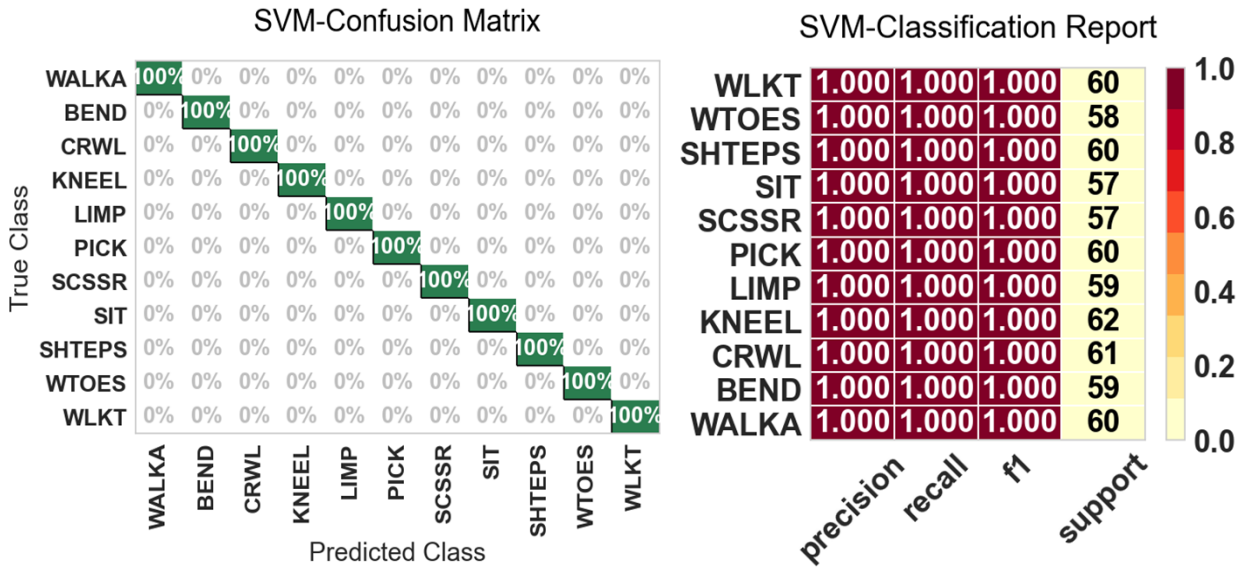
Expérience	**Train Acc** %	**Val Acc** %	**Test Acc** %	**Precision** %	**Recall** %	**F1-score** %
Exp1	100	88.29	81.33	86.58	81.82	80.33
Exp2	100	91.52	83.33	88.71	83.33	82.80
Exp 3(a)	100	99.48	100	100	100	100
Exp 3(b)	100	99.48	98.18	98.23	98.15	98.17
Exp 4(a)	100	100	100	100	100	100
Exp 4(b)	100	100	100	100	100	100

3.4.4.2 L'ensemble de données FMCW 77 GHz

Les résultats sont présentés dans le tableau 3.7. La matrice de confusion et le rapport de classification obtenus grâce à la combinaison de l'augmentation et du pré-traitement appliqués aux données de test sont décrits dans la Figure 3.26 (a) et (b), respectivement.

3.4.4.3 L'ensemble de données FMCW 24 GHz

Les résultats sont présentés dans le tableau 3.8. La matrice de confusion et le rapport de classification obtenus grâce à la combinaison de l'augmentation et du pré-traitement



(a) (b)

Figure 3.25 – (a) Matrice de confusion, (b) Rapport de classification.

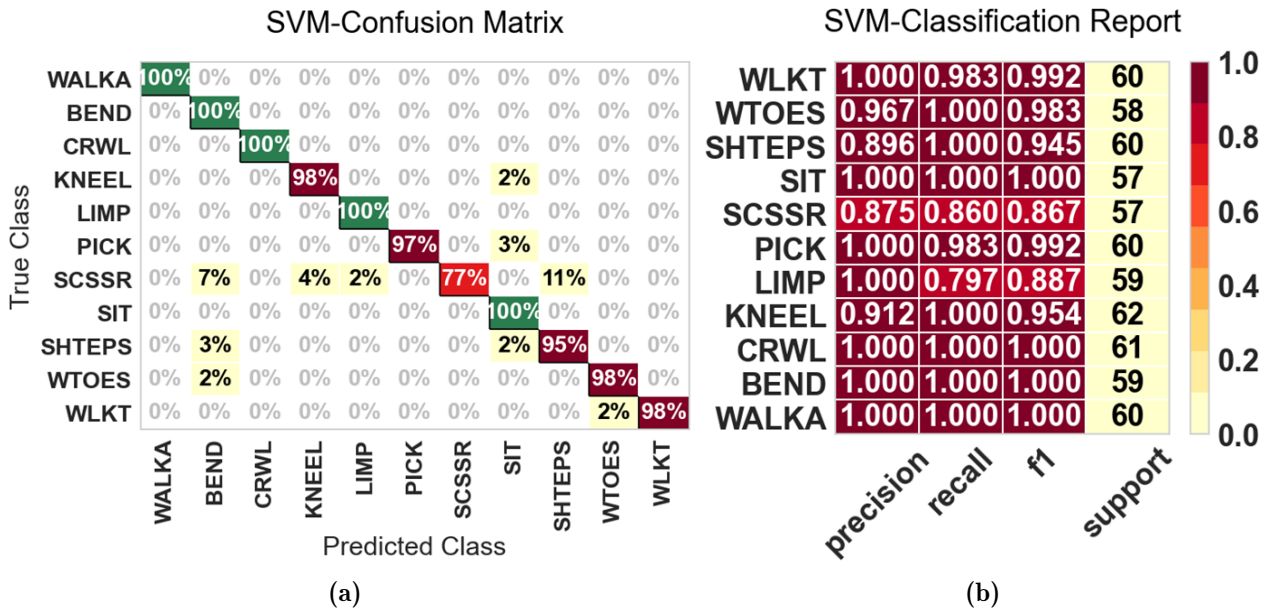
Tableau 3.7 – Performances comparatives de la classification sur l'ensemble de données de 77 GHz.

Expérience	**Train Acc** %	**Val Acc** %	**Test Acc** %	**Precision** %	**Recall** %	**F1-score** %
Exp1	100	93.32	92.36	92.58	92.33	92.30
Exp2	100	95.44	93.93	95.12	93.94	93.82
Exp 3(a)	100	98.01	95.45	97.47	95.45	94.73
Exp 3(b)	100	98.01	96.47	96.61	96.42	96.40
Exp 4(a)	100	98.21	98.48	98.99	98.18	98.45
Exp 4(b)	100	98.21	96.78	96.97	96.71	96.67

appliqués aux données de test sont décrits dans la Figure 3.27 (a) et (b), respectivement.

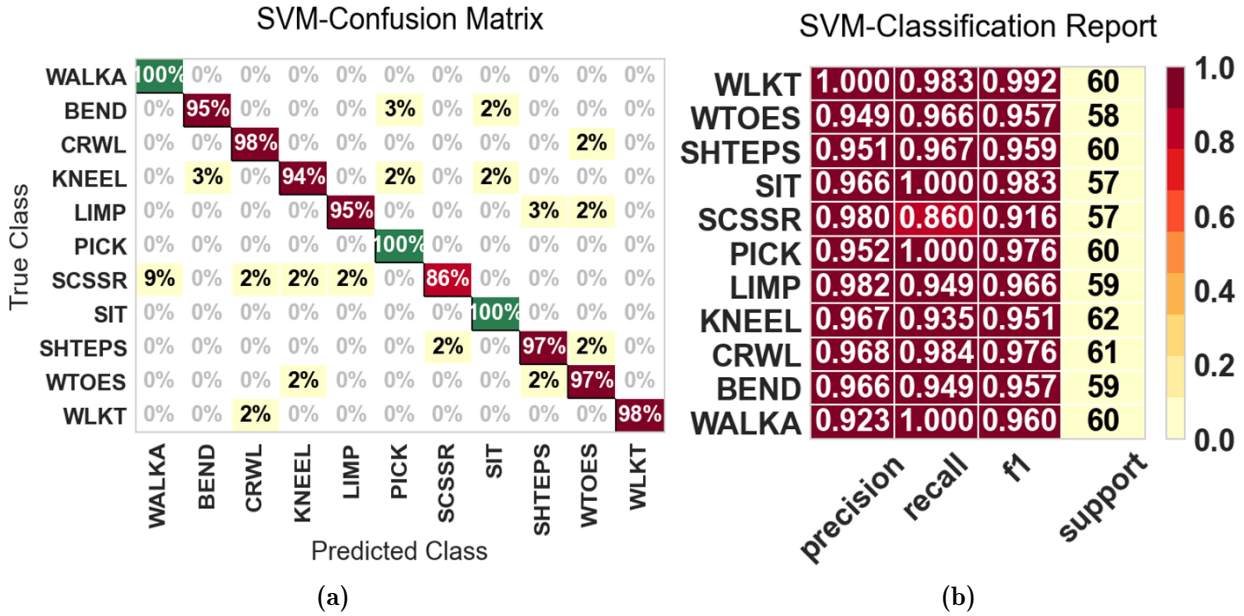
Tableau 3.8 – Performances comparatives de la classification sur l'ensemble de données de 24 GHz.

Expérience	**Train Acc** %	**Val Acc** %	**Test Acc** %	**Precision** %	**Recall** %	**F1-score** %
Exp1	100	86.44	84.84	85.12	87.86	85.25
Exp2	100	88.41	86.63	86.49	86.64	85.87
Exp 3(a)	100	95.91	92.42	92.10	94.60	92.81
Exp 3(b)	100	95.91	94.27	94.50	94.14	94.14
Exp 4(a)	100	98.09	98.18	98.18	98.18	97.98
Exp 4(b)	100	98.09	96.32	96.40	96.30	96.30



(a) (b)

Figure 3.26 – (a) Matrice de confusion, (b) Rapport de classification.



(a) (b)

Figure 3.27 – (a) Matrice de confusion, (b) Rapport de classification.

3.4.4.4 Comparaison avec les approches les plus récentes

Les résultats du modèle que nous proposons peuvent être directement comparés à ceux de Gurbuz *et al.* [33] et de Vishwakarma *et al.* [171], qui ont utilisé le même ensemble de

données de 77 GHz. Les résultats sont présentés dans le tableau 3.9.

Tableau 3.9 – Comparaison des performances de classification l'ensemble de données 77 GHz avec les méthodes les plus récentes.

Référence	**Modèle**	**Test Accuracy (%)**
Gurbuz *et al.* [33]	CAE	85.40
Vishwakarma *et al.* [171]	Modified Alexnet	96.44
Notre approche	Three input-CNN-LSTM-SVM	96.78

3.4.5 Discussion

Les résultats de l'expérience 1 présentés dans les tableaux 3.6, 3.7 et 3.8 montrent que le modèle souffre d'un sur-ajustement. Pendant la phase d'entrainement, le modèle fait preuve d'une exactitude exceptionnelle, atteignant 100 % sur les trois ensembles de données. Néanmoins, lorsqu'il est validé/testé sur de nouvelles données inédites, l'exactitude du modèle chute de manière significative et reste nettement inférieure à son exactitude d'entrainement. Cet écart entre l'exactitude de la formation et celle de la validation/test est dû à la taille limitée de l'ensemble de données d'entrainement, qui entraîne un manque de généralisation. Avec des données d'apprentissage restreintes, le modèle n'est pas en mesure de saisir tous les modèles et toutes les variations possibles dans les données. Par conséquent, il a tendance à mémoriser plutôt qu'à apprendre des modèles généralisables.

L'expérience 2 des tableaux 3.6, 3.7 et 3.8 montre que l'utilisation de données pré-traitées permet d'améliorer l'exactitude du modèle sur tous les ensembles de données. Un taux d'amélioration de la précision du test de 1,52%, 2% et 2,2% est constaté pour les ensembles de données 77 GHz, 10 GHz et 24 GHz, respectivement. L'étape de pré-traitement joue un rôle crucial en améliorant le contenu des images et en facilitant l'extraction et la concaténation de caractéristiques supplémentaires. Malgré ces avantages, la capacité de généralisation du modèle reste limitée. Cela est dû à une exposition insuffisante aux variations nécessaires à l'apprentissage de caractéristiques robustes pouvant être généralisées efficacement à de nouvelles données.

Comme l'a montré l'expérience 3, l'inclusion du processus d'augmentation pendant la phase d'entrainement a renforcé la robustesse et amélioré les performances du modèle. Après avoir effectué le test (a) de l'expérience 3, comme le montrent les tableaux 3.6, 3.7 et 3.8, on observe une amélioration significative de l'exactitude dans les trois ensembles de

données. Ce progrès notable est principalement attribué à l'introduction d'une plus grande variabilité parmi les échantillons d'entraînement. Cette variation permet au modèle de discerner efficacement les modèles et de surmonter les limites associées à la généralisation. Par conséquent, le modèle devient plus apte à faire des prédictions précises sur des données non vues, réduisant ainsi les effets néfastes du sur-ajustement qui étaient apparents dans les résultats des expériences 1 et 2. L'analyse des résultats du test (b) de l'expérience 3, tels que présentés dans les tableaux 3.6, 3.7 et 3.8, met en évidence la performance comparative avec les résultats du test (a). Le modèle a atteint une précision de 98,18%, 96,43% et 94,27% sur les ensembles de données de 10 GHz, 77 GHz et 24 GHz, respectivement. Cela indique que le modèle est bien généralisé et qu'il a le potentiel de fonctionner efficacement sur de nouvelles données inédites de grande taille. Cela confirme la fiabilité de l'approche d'augmentation proposée. Cependant, en comparant les performances du modèle sur le test (a) et le test (b) de l'expérience 3, l'utilisation d'un ensemble de test plus grand a donné une exactitude supérieure pour les ensembles de données de 77 GHz et de 24 GHz. Cette différence de précision suggère que l'utilisation d'une fraction seulement des échantillons, souvent obtenue par fractionnement aléatoire, pourrait introduire une plus grande variabilité. Ce problème découle de la possibilité que les instances sélectionnées ne représentent pas avec précision l'ensemble de l'ensemble de données, ce qui conduit à des évaluations biaisées. Pour répondre à cette préoccupation, il est plus approprié d'envisager d'évaluer le modèle sur un plus grand nombre d'échantillons de test, favorisant ainsi une évaluation équitable.

À la suite de l'expérience 4, une amélioration supplémentaire est obtenue en combinant l'augmentation des données et le pré-traitement. Les rapports de classification présentés aux Figures 3.25, 3.26 et 3.27, respectivement, montrent des résultats cohérents en termes de précision, de rappel et de score F1 pour les trois ensembles de données. Le modèle a obtenu des précisions de 100%, 96,78% et 96,32% pour les ensembles de données 10 GHz, 77 GHz et 24 Ghz, respectivement. L'analyse des matrices de confusion des Figures 3.26 et 3.27, respectivement, correspondant aux ensembles de données de 77 GHz et 24 GHz respectivement, révèle un léger sur-ajustement. Le modèle éprouve des difficultés à distinguer avec précision les différentes activités au sein des diverses classes. Cette difficulté peut être attribuée à la présence de caractéristiques très similaires partagées par ces classes. L'existence de caractéristiques communes entraîne des confusions et des erreurs de classification, ce qui contribue aux limites du modèle en matière de différenciation. Nous pensons qu'une révision de l'architecture et des paramètres du modèle pourrait contribuer

à améliorer sa capacité à distinguer les activités. Le rapport de classification (b) présenté à la Figure 3.25 indique que le modèle est plus performant sur l'ensemble de données de 10 GHz en termes de métriques de classification. En outre, l'examen de la matrice de confusion (a) présentée à la Figure 3.25 montre que le nombre et le pourcentage de faux positifs et de faux négatifs dans l'ensemble de test sont nuls. Ce résultat est d'autant plus significatif que l'ensemble de données comprend davantage de motifs discriminants, ce qui permet au modèle de distinguer efficacement les différentes activités. Cet avantage découle de la haute résolution de l'IR-UWB, qui permet de détecter des modèles de mouvement subtils. Par conséquent, le modèle atteint une exactitude supérieure et la capacité de capturer des modèles uniques associés à des actions et des individus distincts.

Pour justifier la pertinence de l'approche proposée, une analyse comparative des performances avec celles rapportées dans la littérature utilisant le même ensemble de données est effectuée. Les résultats sont présentés dans le tableau 3.9. Gurbuz *et al.* [99] ont proposé d'utiliser les données synthétisées par le GAN à partir de l'ensemble de données de 77 GHz pour entraîner CAE. Leur modèle a atteint une exactitude modeste de 85,40%. Les auteurs ont attribué cette performance sous-optimale à une inadéquation entre les distributions des données synthétiques et réelles. En revanche, notre modèle a fourni un taux d'amélioration de 11,38%, atteignant une exactitude de 96,78%. Grâce à notre approche, nous avons réussi à atténuer le problème d'inadéquation en introduisant une stratégie d'augmentation qui favorise une meilleure généralisation sur des données inédites. Vishwarkarma *et al.* [171] ont proposé d'ajouter artificiellement du bruit blanc gaussien additif (Additive white Gaussian noise : AWGN) aux échantillons de l'ensemble de données 77 GHz et de les utiliser pour former un AlexNet modifié avec un mécanisme d'attention. Par rapport à leur modèle, qui a atteint une précision de 96,44%, le nôtre a apporté une légère amélioration de 0,34%. Contrairement à notre approche, qui n'utilise que des échantillons augmentés, les auteurs de [171] ont utilisé une combinaison d'échantillons originaux et augmentés à des fins de formation. L'introduction d'échantillons originaux a probablement contribué à l'amélioration de leurs performances car ces échantillons partagent la même distribution que les échantillons de test. En outre, les auteurs ont utilisé une architecture complexe avec des millions de paramètres, alors que notre modèle consiste en une structure simple et légère maintenue avec 635397 paramètres entraînables. En outre, il est important de reconnaître que leur approche a été évaluée uniquement sur l'ensemble de données de 77 GHz et que, par conséquent, sa généralisation à d'autres ensembles de données reste incertaine.

En conclusion, il a été démontré que l'approche proposée améliore l'efficacité des données au cours du processus d'entrainement. Les échantillons augmentés générés présentent des attributs significatifs, comme le confirment les résultats de l'expérience 3. Ces résultats fournissent une validation solide de l'efficacité du mécanisme d'augmentation. En conséquence, nous pouvons déduire que la performance prédictive du modèle, entraîné avec des images augmentées à l'aide de la décomposition en ondelettes, présente des caractéristiques favorables relatives à la généralisation et à la pertinence.

3.5 Conclusion

Dans ce chapitre, une proposition a été faite pour un système de RGM et des AH. Le système est composé de deux radars, chacun dédié à l'une des tâches requises. Pour gérer les deux tâches de reconnaissance, une modalité de pré-traitement ainsi qu'un modèle de classification commun ont été proposés. Le système dans son ensemble offre de bonnes performances sous différentes métriques avec une architecture de taille raisonnable. Une des limitations de notre solution réside dans l'utilisation des données provenant d'un seul radar. Cela peut parfois entraîner des problèmes, notamment en cas de génération d'échantillons erronés, souvent dus à un angle défavorable entre la cible et le radar. Pour pallier cette faiblesse, le chapitre suivant propose une approche de RGM en intégrant un système multi-capteurs, afin d'améliorer à la fois les performances et la fiabilité.

Chapitre 4

Contribution3 : Système multi-capteurs pour la reconnaissance des gestes de la main

4.1 Introduction

Outre les nombreux avantages du IR-UWB cité dans les chapitres 1 et **??**, sa vulnérabilité aux variations d'orientation pose problème. Cette vulnérabilité devient évidente dans les situations où la perception du radar devient incertaine en raison d'angles d'aspect défavorables entre la trajectoire du mouvement de la cible et la ligne de visée du radar. Cette incertitude fait qu'il n'est pas pratique de s'appuyer uniquement sur un seul capteur.

Rappelons que nous visons à développer un système de RGM et AH spécifiquement dédié au suivi des patients post-AVC. Pour atteindre notre objectif, le système que nous développons doit opérer de manière continue et capable de fournir des prédictions fiables basées sur les données recueillies par les radars.

Pour différencier les GM et les AH plus efficacement, le système de reconnaissance doit intégrer plusieurs capteurs. En combinant les données de plusieurs IR-UWB, des informations plus concises sont ajoutées. En cas de données manquantes ou peu fiables fournies par un capteur, le modèle de fusion peut toujours prendre des décisions éclairées sur la base des données fournies par les autres capteurs. Cette intégration renforce la robustesse du système, le rendant plus résistant aux défaillances des capteurs. Elle réduit également l'ambiguïté et l'incertitude tout en renforçant la confiance.

4.2 Système de reconnaissance multi-capteurs

Pour remédier aux lacunes des IR-UWB susmentionnées ci-dessus, nous proposons une nouvelle architecture à entrées et sorties multiples (Multiple Input Multiple Output : MIMO) basée sur l'intégration d'un CNN avec l'algorithme de classification ETs. Le modèle proposé ne nécessite pas de pré-traitement lourd des données ni d'ingénierie manuelle des caractéristiques. Il élimine la dépendance à l'égard de l'expérience des experts et des connaissances préalables. En tirant parti des avantages offerts par le CNN et les ETs, nous avons mis au point un modèle complet et efficace capable d'extraire automatiquement des caractéristiques discriminantes et de produire des résultats plus précis. Contrairement aux travaux précédents où les données de plusieurs capteurs étaient traitées ensemble [61, 99, 127, 128],nos eXtra-arbres convolutifs à entrées et sorties multiples (Multiple Input Multiple Output eXtraTrees : MIMO-CxT) peuvent prendre les données de plusieurs capteurs en entrée et les traiter de manière indépendante. La sortie de chaque branche du CNN est combinée avant que la prédiction ne soit effectuée par les ETs. Ce processus améliore l'opération d'extraction des caractéristiques. En se concentrant sur les données de chaque capteur séparément, la capacité à détecter des caractéristiques plus profonds et plus utiles est accrue. Le traitement des données en parallèle peut effectivement aider à extraire des informations complémentaires sur la même cible à travers plusieurs capteurs, ce qui permet d'apprendre une représentation plus complète et d'obtenir un classifieur plus efficace. Outre ses performances encourageantes, notre modèle comporte moins de paramètres, ce qui réduit les problèmes de complexité et de coût de calcul. Il est donc tout à fait adapté au développement d'un système de reconnaissance embarqué qui peut être utilisé comme outil de télé-médecine conçu pour la rééducation à distance des patients victimes d'AVC.

4.2.1 MIMO-CxT pour les systèmes multicapteurs

Plusieurs études ont démontré l'efficacité de l'utilisation de plusieurs IR-UWB ou de leur combinaison avec d'autres capteurs pour augmenter les performances de détection et de classification [61, 79, 99]. Souvent, ces capteurs fonctionnent indépendamment les uns des autres, ce qui signifie qu'ils peuvent recueillir des données de valeurs, d'échelles ou même de natures différentes dans le cas de systèmes de capteurs hétérogènes. Il est donc raisonnable de les traiter séparément afin d'extraire des caractéristiques discriminantes des données de chaque capteur et de préserver leurs caractéristiques sans les altérer. Le

MIMO-CxT se compose de trois branches CNN parallèles utilisées dans une configuration hybride avec l'algorithme ETs. Contrairement au modèle CNN standard, qui mélange et traite l'ensemble des données en une seule fois, ce qui entraîne parfois la perte de caractéristiques spécifiques, notre modèle est conçu pour effectuer l'opération d'extraction des caractéristiques sur les données d'un certain capteur dans chaque branche de manière indépendante. Les sorties de toutes les branches CNN sont concaténées, mises sous forme de vecteur et transmises aux ETs, qui agissent comme le classifieur de l'architecture et fournissent le résultat de la prédiction. La structure du MIMO-CxT est décrite à la Figure 4.1. L'une des caractéristiques remarquables de l'architecture proposée est la possibilité d'extraire des caractéristiques indépendantes et appropriées correspondant à des capteurs fonctionnant indépendamment, ce qui est le cas dans notre travail. Grâce à leur architecture flexible, les branches CNN peuvent être facilement adaptées à de nouvelles configurations, comme l'ajout d'une nouvelle branche d'entrée lors de l'exploitation de données supplémentaires.

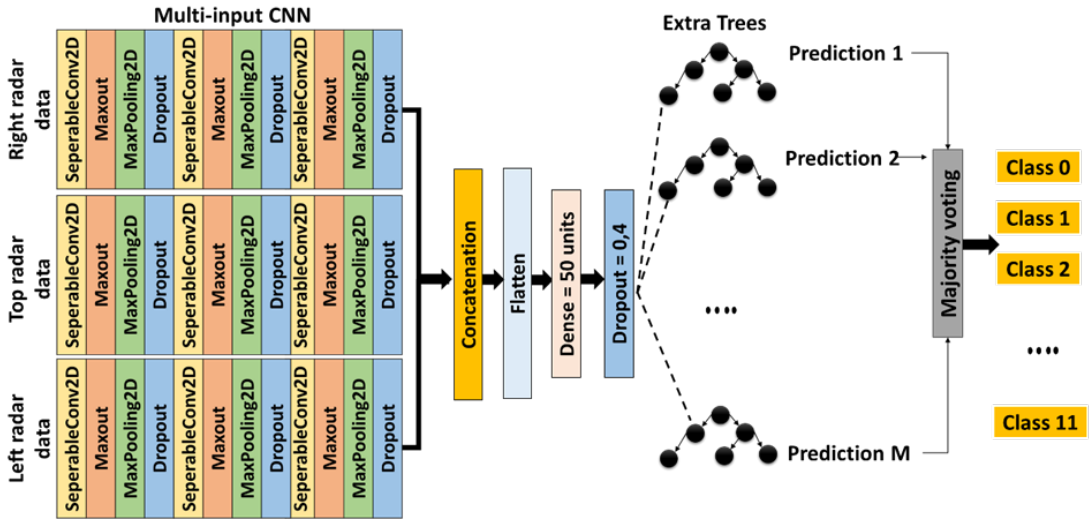


Figure 4.1 – Architecture du modèle MIMO-CxT.

4.2.1.1 Structure CNN à entrées multiples

Une configuration de couches similaire entre les trois branches du CNN a été adoptée, où chacune consiste en une couche d'entrée de taille 75×75×1, qui correspond à un seul canal de l'image binaire d'entrée 75×75 pixels. Le tableau 4.1 présente en détail la composition d'une seule branche du CNN. Étant donné que l'un de nos objectifs est de concevoir un modèle efficace qui réduit autant que possible la quantité de calculs et le nombre

de paramètres sans perdre en performance de classification. Nous proposons d'utiliser la convolution séparable en profondeur. Contrairement à la convolution conventionnelle qui exécute le calcul dans le sens du canal et dans le sens spatial en une seule étape, la convolution séparable en profondeur factorise le calcul en deux étapes : une convolution en profondeur suivie d'une convolution par point (voir Figure 4.2). La convolution en profondeur prend en entrée une carte de caractéristiques F dont la taille est de $W_{in} \times H_{in} \times M$ (largeur, hauteur et nombre de filtres) et génère une carte de caractéristiques de sortie O dont la taille est $W_{out} \times H_{out} \times N$. En effectuant une convolution spatiale indépendamment sur chaque canal de l'entrée à l'aide d'un noyau convolutif $\hat{\mathrm{K}}$ de taille $W_K \times H_K \times M \times N$, la convolution en profondeur peut être exprimée comme suit :

$$\hat{O}_{k,l,m} = \sum_{ij} \hat{K}_{i,j,m,n} \hat{O}_{K+i-1,l+j-1,m} \tag{4.1}$$

Ensuite, la convolution par point est effectuée à l'aide d'un filtre $\check{K}$ de taille 1x1 afin de combiner la sortie totale générée.

$$O_{k,l,n} = \sum_{m} \check{K}_{m,n} \hat{O}_{K-1,l-1,m} \tag{4.2}$$

La combinaison de la convolution par profondeur et de la convolution par point est plus efficace que la convolution classique en termes de complexité de calcul. Elle réduit considérablement le nombre de paramètres du réseau, en éliminant une grande partie de la multiplication, ce qui permet d'obtenir un modèle plus rapide à former et à exécuter. En outre, cette combinaison garantit la précision de la classification du modèle, ce qui le rend moins sujet au sur-ajustement. Après chaque couche de convolution séparable dans le sens de la profondeur, une couche Maxout est insérée. Cette dernière a été détaillé dans la section 3.2.1.2 du chapitre 3. Le modèle incorpore également une couche MaxPooling2D alimentée par les cartes de caractéristiques obtenues, afin de réduire leurs dimensions, de préserver les caractéristiques les plus pertinentes identifiées dans la couche précédente et d'éviter les calculs inutiles. Enfin, la couche Dropout, d'une valeur de 25%, est utilisée pour éviter l'effet de sur-ajustement du modèle. Les sorties des trois branches du CNN sont concaténées, puis aplaties pour former un vecteur de caractéristiques à traiter ultérieurement par le classifieur ETs.

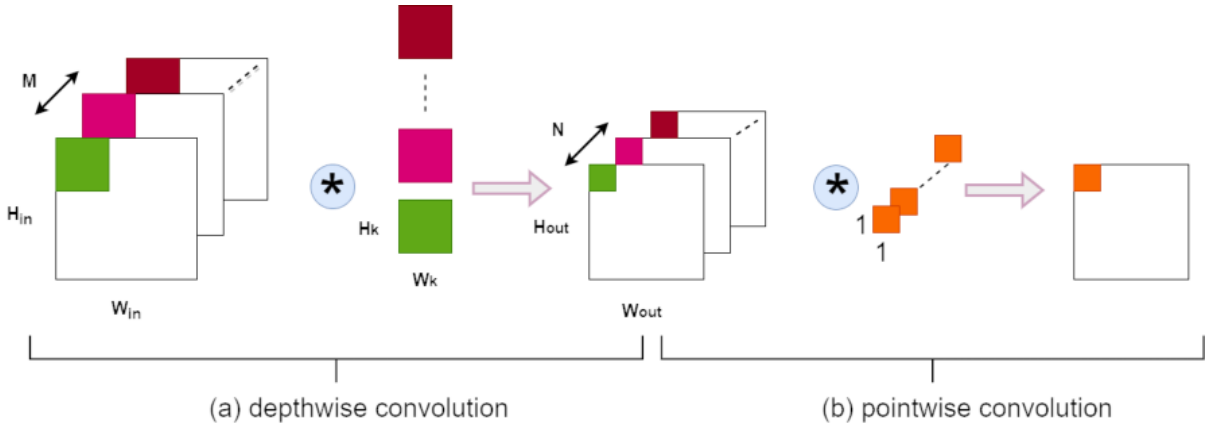


Figure 4.2 – Convolution séparable en profondeur.

Tableau 4.1 – Structure détaillée d'une branche CNN.

Type de couche	Filtre	Padding	Strides	Paramètres	La sortie
Input layer	-	-	-	-	(None, 75, 75, 1)
SeparableConv2D	4x4x64	same	2x2	144	(None, 38, 38, 64)
Maxout	64	-	-	0	(None, 38, 38, 64)
MaxPooling2D	2x2	-	-	0	(None, 19, 19, 64)
Dropout	0.25	-	-	0	(None, 19, 19, 64)
SeparableConv2D	4x4x64	same	2x2	5184	(None, 10, 10, 64)
Maxout	64	-	-	0	(None, 10, 10, 64)
MaxPooling2D	2x2	-	-	0	(None, 5, 5, 64)
Dropout	0.25	-	-	0	(None, 5, 5, 64)
Conv2D	4x4x64	same	2x2	5184	(None, 3, 3, 64)
Maxout	64	-	-	0	(None, 3, 3, 64)
MaxPooling2D	2x2	-	-	0	(None, 1, 1, 64)
Dropout	0.25	-	-	0	(None, 1, 1, 64)

4.2.1.2 Arbres extrêmement aléatoires (Extra-Trees : ETs)

Nous avons choisi d'utiliser les ETs comme classifieur final pour notre modèle. Cet algorithme a été élaboré au chapitre 2, section 2.3.3. Parmi les différentes méthodes de classification basées sur les arbres, les ETs sont le choix le plus approprié dans le cadre de notre approche, et ce pour plusieurs raisons.

— Le schéma de randomisation extrême des ETs les rend beaucoup plus rapides et efficaces en termes de temps d'apprentissage. En outre, il réduit considérablement la variance élevée et empêche ainsi le sur-ajustement excessif.

— L'utilisation de l'ensemble d'apprentissage complet d'origine pour l'apprentissage de chaque DT, permet une forte capacité de généralisation.
— La nature d'ensemble de l'algorithme ETs permet l'agrégation de divers arbres de décision, chacun formé sur un sous-ensemble différent des données. Cette diversité contribue à améliorer les performances globales et la robustesse de l'algorithme.
— La mise en œuvre de l'algorithme ETs ne nécessite pas une grande concentration sur la sélection des valeurs des hyperparamètres.

4.3 Implémentation

Pour valider l'efficacité de notre modèle MIMO-CxT, nous utilisons la base de données publique UWB Gestures [101] décrite dans le chapitre précédent (voir section 3.3.1). Cette fois-ci les trois ensembles de données radars Left, Top et Right sont utilisés simultanément pour entrainer le modèle. Le modèle est mis en œuvre en Python à l'aide du package Keras avec Tensorflow. Les autres bibliothèques utilisées sont sklearn, Pandas, Numpy et Yellowbrick. L'implémentation du modèle est réalisée sur la même machine avec les caractéristiques décrites sur la section 3.3.3 du chapitre 3. Tous les échantillons de la base de données ont été réduits à une taille de 75x75, puis convertis en images binaires comme montré sur la Figure 4.3. Les raisons d'utiliser des images binaires sont les suivantes : plus rapide lors de l'inférence, évite les pré-traitements inutiles et nécessite un faible stockage par rapport aux images RVB. Ensuite, les échantillons convertis ont été divisés aléatoirement en 80% pour l'entraînement et 20% pour le test. Pour obtenir des performances optimales, les hyperparamètres du classifieur ETs sont comparés et sélectionnés au cours du processus d'entrainement à l'aide de Optuna.

Afin d'analyser les performances du modèle proposé, nous avons utilisé les mêmes critères d'évaluation que le chapitre précédent mentionné dans la section 3.2.2, notamment l'exactitude, la précision, le rappel et le score-F1, la courbe PR et la courbe ROC.

4.4 Résultats

Dans cette section, nous présentons les résultats obtenus par le modèle proposé dans les phases d'entrainement et de test à l'aide des différentes mesures d'évaluation. Dans ce qui suit une comparaison avec d'autres modèles existants est établie.

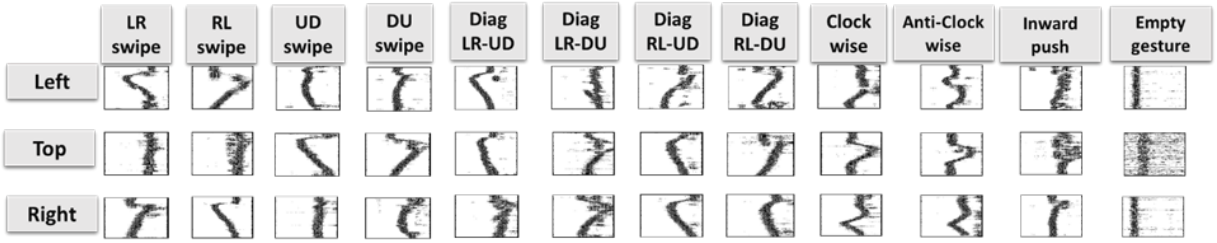


Figure 4.3 – Exemple d'échantillons binarisés de l'ensemble de données UWB Gestures.

4.4.1 Processus d'entrainement

Le processus d'apprentissage est entièrement supervisé et peut être divisé en trois étapes principales. Tout d'abord, le modèle CNN-Softmax est utilisé, où chaque branche du CNN est alimentée par des données provenant d'un radar différent. Ensuite, le CNN est entraîné sur 15 itérations à l'aide de l'algorithme de rétro-propagation avec différentes tailles de lots. Enfin, Softmax est remplacé par le classifieur ETs, qui est alimenté par les caractéristiques extraites et fusionnées des multiples branches CNN pour la classification. Les résultats ont montré que les performances du MIMO-CxT s'améliorent avec des tailles de lots plus petites. Les meilleures précisions d'apprentissage ont été obtenues avec des lots de 8 et 16. Cependant, une taille de lot de 16 nécessite relativement moins de temps d'apprentissage qu'une taille de lot de 8, et a donc été sélectionnée pour les expériences restantes. Le tableau 4.2 résume les résultats obtenus pour les différentes tailles de lots.

Tableau 4.2 – Exactitude d'entrainement pour différentes tailles de lot.

Métrique	**Taille de lot**				
	8	**16**	**32**	**64**	**128**
Train Accuracy %	99.67	99.55	99.30	99.21	98.98

Ensuite, le MIMO-CxT est ré-entraîné en utilisant une taille de lot de 16 avec l'optimiseur Optuna pour affiner ses hyperparamètres. Le meilleur modèle MIMO-CxT a atteint une précision de 99,7%, comme le montre l'historique de l'optimisation sur la Figure 4.4. Les valeurs optimales des hyperparamètres sont représentés sur la Figure 4.5.

4.4.2 Processus d'évaluation

Pour l'évaluation des performances, la configuration du modèle MIMO-CxT est utilisée, ainsi que les hyperparamètres illustrés sur la Figure 4.5. Les résultats obtenus sur les données test sont illustrés à la Figure 4.6, la matrice de confusion étant présentée à la

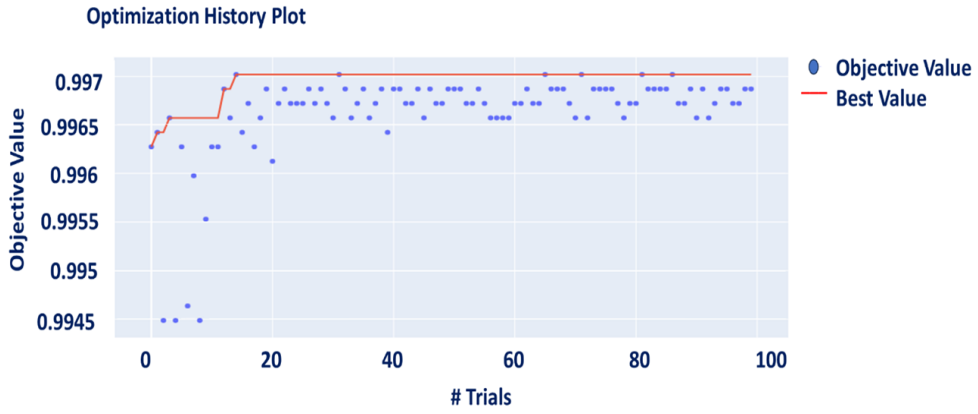


Figure 4.4 – Historique d'optimisation.

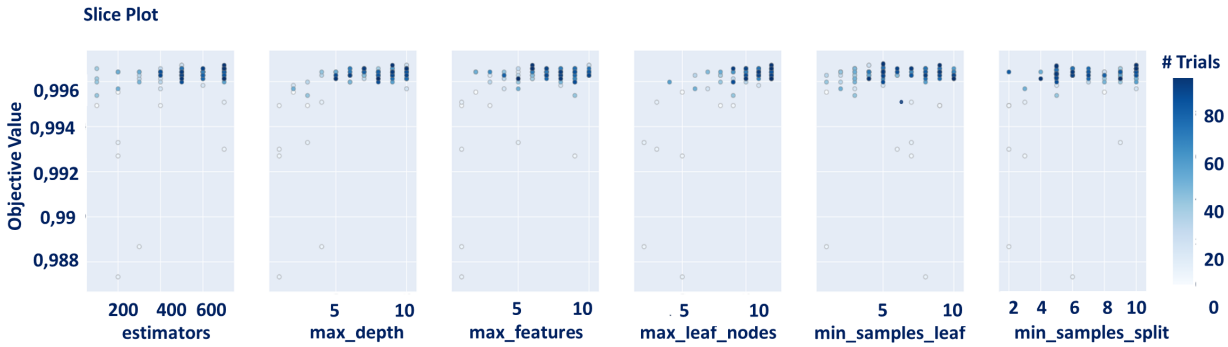


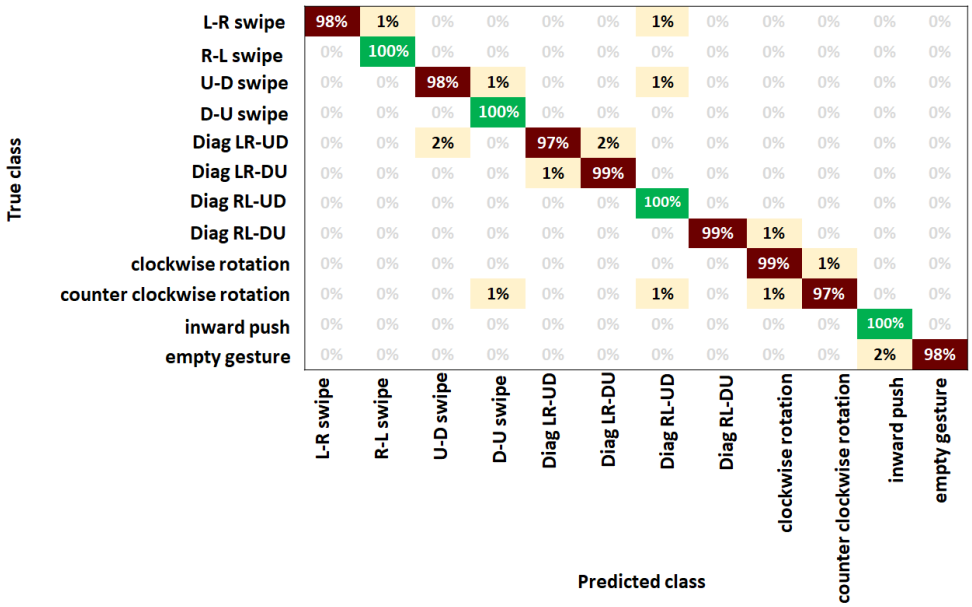
Figure 4.5 – Graphe des différentes valeurs d'hyperparamètres sur les performances du modèle.

Figure 4.6 (a) et le rapport de classification à la Figure 4.6 (b). Bien que les résultats obtenus soient excellents, nous examinons également les approches graphiques illustrées à la Figure 4.7, notamment la courbe ROC présentée à la Figure 4.7 (a) et la courbe PR présentée à la Figure 4.7 (b).

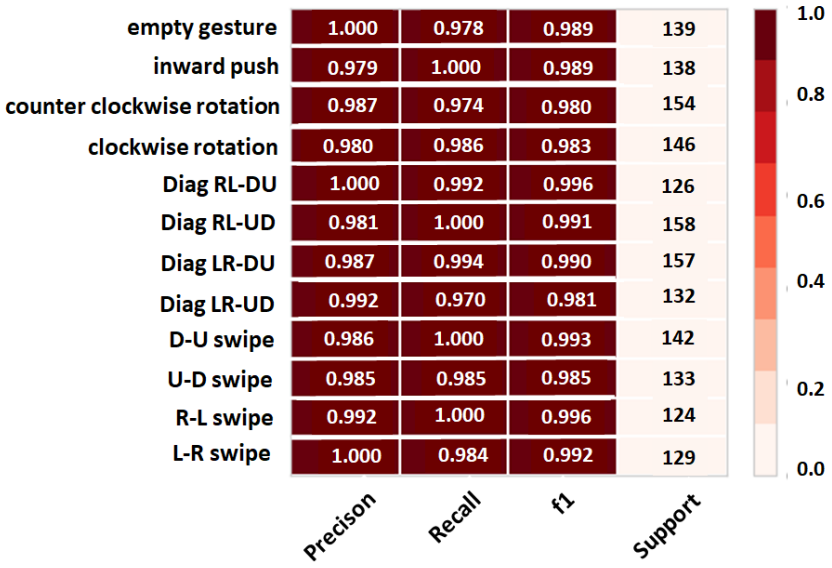
4.4.3 Comparaison

4.4.3.1 Vérification de la méthode de concaténation des données

Pour démontrer la faisabilité et la supériorité du modèle proposé, deux expériences différentes ont été menées. Premièrement, le modèle à entrée unique a été alimenté par les données provenant de chaque capteur séparément. Deuxièmement, le modèle à entrée



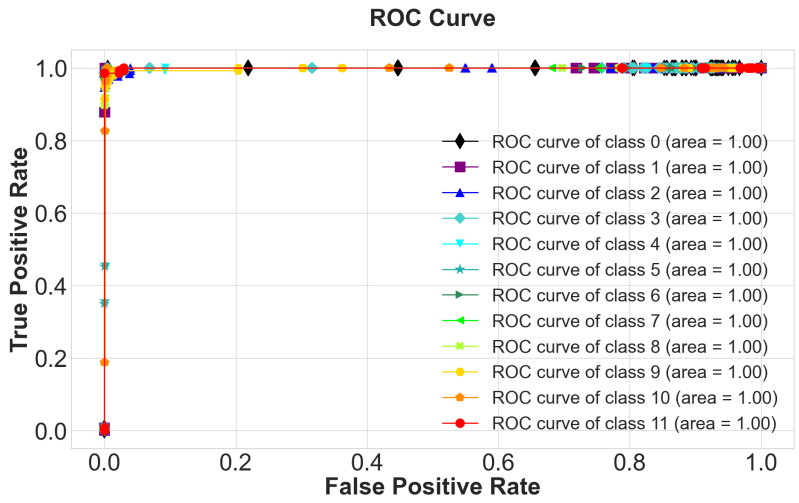
(a)



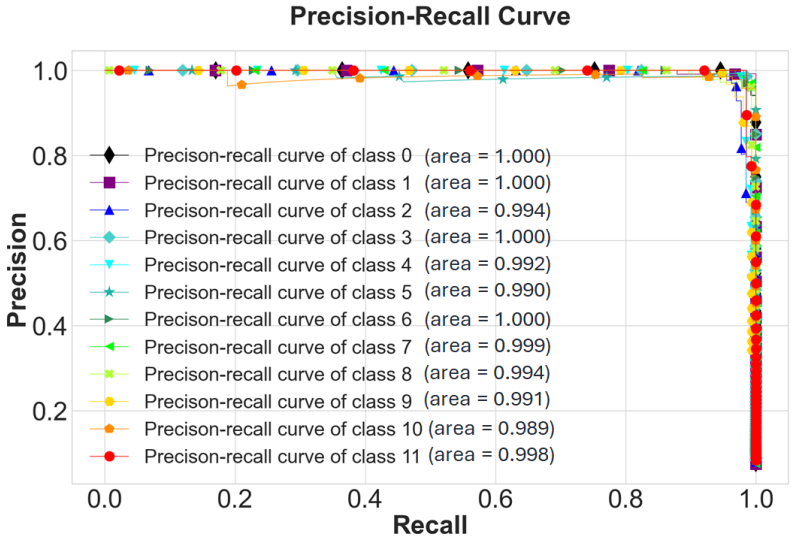
(b)

Figure 4.6 – (a) Matrice de confusion, (b) Rapport de classification.

unique a été alimenté par les données des trois capteurs simultanément. Enfin, les performances de ces approches ont été comparées à celles du modèle MIMO-CxT. Les résultats de ces expériences sont résumés dans le tableau 4.3.



(a)



(b)

Figure 4.7 – (a) Courbe ROC, (b) Courbe PR.

Tableau 4.3 – Comparaison des performances de classification modèle à entrée unique et multiples.

Métrique	**Single-Input Multi-Output CxT**				**MIMO-CxT**
	Left	**Top**	**Right**	**Left+Top+Right**	
Train Accuracy %	88.86	85.65	86.76	78.91	99.70
Test Accuracy %	81.52	75.26	80.95	72.70	98.86
Precision %	82.14	76.21	80.43	73.62	98.90
Recall %	81.52	75.26	80.09	73.70	98.86
F1-score %	81.68	75.31	80.07	72.66	98.85

4.4.3.2 Vérification du pré-traitement des données

Afin d'évaluer l'efficacité de l'utilisation des images binaires par rapport aux images RVB, le modèle MIMO-CxT a été entraîné et testé sur les deux types d'images. Une analyse complète a été réalisée, prenant en compte des facteurs tels que la performance de la classification, la complexité du modèle et les temps de calcul. Les résultats de cette évaluation sont présentés dans le tableau 4.4.

Tableau 4.4 – Comparaison des performances de classification sur les images RVB/binaires.

Métrique	Images RVB	Images binaires
Train Accuracy %	98.73	99.70
Test Accuracy %	97.19	98.86
Precision %	97.23	98.90
Recall %	97.15	98.86
F1-score %	97.21	98.85
Paramètres CNN	32016	31536
Temps d'entrainement (s)	9527.5141	450.3863
Temps de test (s)	2.7767	1.9582

4.4.3.3 Comparaison avec les modèles existants

La troisième expérience vise à comparer les performances de différents classifieurs par rapport aux ETs lorsqu'ils sont utilisés dans une configuration hybride avec un CNN à entrées multiples.Nous avons comparé notre modèle MIMO-CxT à des modèles à entrée unique utilisés dans le contexte de la reconnaissance des gestes de la main et à des modèles à entrées multiples proposés dans la littérature. Le tableau 4.5 résume les résultats obtenus par les différents modèles entraînés sur l'ensemble de données publiques UWB Gestures.

4.5 Discussion

L'utilisation de systèmes multi-capteurs pour la RGM offre de nombreux avantages tels qu'une précision accrue, la robustesse, la capacité de détecter des gestes complexes, la polyvalence et l'adaptabilité à différents besoins. Ces avantages rendent les systèmes multi-capteurs préférables aux solutions à capteur unique. Cependant, il est important d'envisager une solution capable de traiter et d'intégrer efficacement les informations

Tableau 4.5 – Comparaison des performances de classification de la MIMO-CxT par rapport aux approches existantes.

	Modèle	**Train Accuracy %**	**Test Accuracy %**
Etrée unique	Four-layer CNN [101]	90.47	78.90
	Six-layer CNN [60]	96.94	87.63
Etrées multiples	Three-input CNN [173]	98.94	97.37
	MIMO-CNNLSTM	96.11	97.79
	MIMO-CNNSoftmax	97.14	97.99
	Multi-stream CNN [174]	98.93	98.15
	MIMO-CNNSVM	98.94	98.18
Notre modèle	MIMO-CxT	99.70	98.86

provenant de diverses sources de capteurs. Pour y remédier, nous présentons MIMO-CxT, une approche d'apprentissage profond hybride de bout en bout pour la RGM basé sur des systèmes multi-capteurs.

La première expérience a été menée pour analyser les performances du modèle en fonction de la façon dont les caractéristiques des données de capteurs sont extraites : (i) séparément, (ii) ensemble, ou (iii) indépendamment en parallèle. Le tableau 4.3 montre que le modèle à entrées multiples est capable de reconnaître les gestes beaucoup mieux que le modèle à entrée unique dans tous les scénarios testés. L'une des principales raisons de cette amélioration est la quantité de caractéristiques extraites. Il est évident que le modèle à entrées multiples atteint un taux de reconnaissance élevé en raison de sa capacité à extraire et à fusionner des informations plus diverses concernant le même geste à travers plusieurs capteurs. En outre, nous avons observé une différence significative dans les performances du modèle lors du traitement simultané et indépendant de toutes les données en parallèle. Nous émettons l'hypothèse que la raison de cette faible performance est la présence de caractéristiques communes. Nous avons mené une deuxième expérience dans laquelle nous avons comparé les performances du modèle basé sur la technique de traitement d'image adoptée dans notre travail. Le tableau 4.4 montre que l'utilisation d'images binaires permet de réduire les temps d'inférence et de prédiction par rapport aux images RVB. L'utilisation d'images binaires permet de filtrer les informations inutiles tout en conservant les principales caractéristiques des gestes. En outre, elle réduit le nombre de paramètres entraînables tout en maintenant une performance élevée du modèle. Par conséquent, la mise en œuvre de notre modèle ne nécessite pas de matériel informatique

puissant.

La même expérience a été réalisée sur des modèles à entrée unique/multiples déjà proposés dans la littérature, et les résultats sont présentés dans le tableau 4.5. Si l'on compare les performances des modèles CNN à quatre couches à entrée unique [101] et CNN à six couches [60] par rapport à MIMO-CxT, il est évident que ce dernier est plus performant en termes de précision de test et de généralisabilité. Les CNN standard sont conçus pour extraire des caractéristiques à partir d'une seule entrée, ce qui peut se traduire par une capacité insuffisante pour saisir toute la gamme des variations de données entre différents capteurs. Toutefois, la mise en œuvre du MIMO-CxT avec une structure multi-branches permet de surmonter cette limitation. La structure multi-branches agit comme un régularisateur, permettant au modèle de capturer efficacement les diverses variations présentes dans les données. En utilisant différentes branches, le modèle peut apprendre des représentations plus robustes et plus discriminantes, ce qui permet d'améliorer les performances de classification. D'après le tableau 4.5, nous constatons que l'utilisation de modèles à entrées multiples a permis d'améliorer considérablement les résultats. Cependant, MIMO-CxT présente plusieurs avantages par rapport aux modèles proposés dans la littérature. MIMO-CxT a obtenu une augmentation de précision de 1,49% et de 0,71% par rapport au CNN à trois entrées [173] et au Multi-Stream CNN [174], respectivement. Les deux modèles [173] et [174] ont une architecture plus profonde et plus complexe que la nôtre. Ils utilisent tous deux un grand nombre de filtres convolutifs et plusieurs couches entièrement connectées avec un grand nombre d'unités. Cela augmente considérablement le nombre de paramètres entraînables, ce qui accroît le temps d'entraînement. Nous pouvons également remarquer que la combinaison du CNN avec le LSTM augmente légèrement la précision, mais reste inférieure à celle obtenue par le MIMO-CxT. Compte tenu de l'architecture complexe et de la nature séquentielle du LSTM, les ETs sont faciles à mettre en œuvre et plus efficaces. Les temps d'entraînement et d'inférence des ETs sont plus rapides, en particulier pour les grands ensembles de données. Cette efficacité de calcul est bénéfique lorsque l'on travaille avec des prédictions en temps réel ou quasi-réel. Bien que le Soft-Max soit un classifieur puissant, l'utilisation des ETs a permis d'améliorer la précision de 0,86%. Cette amélioration peut être principalement attribuée à la mise en œuvre de techniques d'apprentissage d'ensemble par les ETs. L'utilisation d'arbres individuels permet de capturer et d'apprendre des informations distinctes, tandis que l'agrégation des prédictions de plusieurs arbres permet d'obtenir des prédictions plus fiables. L'utilisation des ETs comme classifieur final a permis d'obtenir des performances similaires à celles du

SVM. Cependant, les ETs sont intrinsèquement bien adaptés aux tâches de classification multi-classes. Contrairement aux SVM, les ETs peuvent les traiter directement sans avoir à les réduire en plusieurs problèmes de classification binaire. En outre, les ETs tendent à être plus évolutives que les SVM, en particulier lorsqu'elles traitent un grand nombre de points de données. La complexité des SVMs croît rapidement avec l'augmentation du nombre de points de données.

4.6 Conclusion

Dans ce chapitre un nouveau modèle hybride profond, nommé MIMO-CxT, pour la RGM à partir d'un réseau de capteurs est proposé. L'architecture du modèle tire parti de la robustesse du CNN pour capturer les caractéristiques de bas niveau et profondes, ainsi que de la simplicité et de la forte capacité de généralisation des ETs pour la classification des données. Les avantages du modèle MIMO-CxT comprennent l'entrée multi-sources, la vitesse de calcul rapide, le faible coût et les performances de généralisation élevées. Les résultats ont été comparés à ceux des approches conventionnelles et il s'est avéré que notre modèle est nettement plus performant. Sur la base des performances du MIMO-CxT, il peut être considéré comme une solution prometteuse pour aider les professionnels de la santé en tant qu'outil de surveillance en ligne à domicile pour les patients victimes d'AVC.

Conclusion générale

Dans le contexte de cette thèse, notre étude s'est concentrée sur la mise en œuvre d'un système visant à reconnaître les gestes de la main ainsi que les actions humaines, en utilisant la technologie de radar ultra large bande. Ce système a été spécifiquement proposé pour le suivi des patients post AVC lors des séances de rééducation à domicile. Pour répondre à cet objectif, nous avons élaboré deux architectures de modèle d'apprentissage profond, considérées comme des solutions performantes pour la classification des gestes et actions.

En ce qui concerne l'application de notre étude, une limitation importante que nous avons rencontrée est l'absence d'une base de données complète qui regroupe à la fois les gestes de la main et les actions humaines. Pour résoudre ce problème, nous avons pris l'initiative de combiner deux bases de données distinctes pour former un réseau de radars. Par la suite, nous avons conçu un modèle hybride qui combine deux aspects essentiels du traitement de données : une partie convolutive et une partie récurrente. Cette approche nous a permis d'exploiter de manière directe et efficace les caractéristiques spatiales et temporelles présentes dans nos données radar. En utilisant cette architecture hybride, nous avons pu saisir plus efficacement les motifs complexes inhérents aux gestes et aux actions, améliorant ainsi la performance de notre système de reconnaissance. Un autre défi que nous avons abordé est le nombre limité d'échantillons disponibles dans la base de données des actions humaines. Pour résoudre cette problématique, nous avons introduit une nouvelle méthode d'augmentation de données, visant à générer des variations synthétiques de nos données d'entraînement. Cette approche a permis d'élargir le nombre d'exemples d'entraînement disponibles, améliorant ainsi la généralisation et la robustesse de notre modèle. L'architecture du modèle que nous avons développée présente plusieurs avantages. Elle offre une solution simple et efficace, avec des performances qui se situent au-dessus ou au moins à la hauteur des méthodes de pointe actuelles. De plus, ce modèle se caractérise par sa légèreté, avec un nombre réduit de paramètres entraînables à gérer. Son fonction-

nement repose sur une structure à couches parallèles, permettant l'extraction simultanée de caractéristiques à partir de différentes représentations de bas niveau. Cette conception a ouvert de meilleures perspectives d'apprentissage, conduisant à une performance améliorée dans la reconnaissance des gestes et des actions. Dans notre démarche, une nouvelle limitation s'est manifestée, soulignant la non fiabilité de se reposer exclusivement sur un seul capteur pour mener à bien la reconnaissance. Face à ce défi, nous avons conçu une solution novatrice : le modèle MIMO-CxT (Multiple-Input Multiple-Output Convolutional Extra Trees), spécialement adapté pour les systèmes multi-capteurs. L'élaboration du MIMO-CxT a été entreprise dans le but de surmonter cette limitation en offrant une approche différente de celles qui ont été employées jusqu'ici dans la littérature concernant les IR-UWB. Contrairement aux méthodes préexistantes, le MIMO-CxT adopte une approche entrées-sorties multiples, ce qui signifie qu'il est capable de traiter les données provenant de plusieurs capteurs de manière autonome et indépendante. Cet aspect permet au modèle de mieux saisir les nuances et les particularités des informations recueillies par chaque capteur, contribuant ainsi à une reconnaissance plus précise, fiable et robuste.

Le principal travail à poursuivre dans cette recherche consiste à développer une nouvelle base de données spécifiquement conçue pour le suivi des patients post AVC. Cette base de données devrait être conçue pour enregistrer et suivre les activités des patients post-AVC, permettant ainsi une évaluation continue de leur rétablissement et de leur autonomie. Elle serait particulièrement utile pour fournir des données précieuses aux professionnels de la santé, aux aidants et aux patients eux-mêmes.

Nous visons à créer une base de données complète regroupant à la fois les gestes de la main et les actions humaines. Cela nous permettra d'évaluer simultanément le modèle sur les deux tâches de classification, afin de vérifier sa capacité à gérer ces deux aspects de manière cohérente et précise.

De plus, nous souhaitons étendre la diversité des activités évaluées, en incluant des variations telles que le changement d'orientation et la variation de distance entre les capteurs et la cible. Ces ajouts aideront à tester la flexibilité et l'adaptabilité du modèle dans des conditions variées et réalistes.

Ces améliorations visent à renforcer la performance globale et la fiabilité du système de reconnaissance, particulièrement dans des contextes cliniques et de rééducation.

Production Scientifique

Publications internationales

— D. S. Korti,Z. Slimane and K. Lakhdari. "Enhancing Dynamic Hand Gesture Recognition using Feature Concatenation via Multi-Input Hybrid Model". International journal of electrical and computer engineering systems,2023, Vol. 14, No. 5, 535-546.

— D. S. Korti and Z. Slimane. "Advanced Human Activity Recognition through Data Augmentation and Feature Concatenation of Micro-Doppler Signatures". International journal of electrical and computer engineering systems,2023, Vol. 14, No. 8, 893-902.

— D. S. Korti and Z. Slimane. "Unobtrusive hand gesture recognition using ultra-wide band radar and deep learning". International journal of electrical and computer engineering ,2023, Vol. 13, No. 6, 6872-6881.

Communications internationales

— D. S. Korti, F.Derraz, Z.Slimane and K. Lakhdari. "Enhanced Approach for On-road Obstacles Detection using Level-Set-YOLOv3 Combination". The 1st International Conference on Electronics, Artificial ntelligence and New Technologies, 2021, Oum El Boughi, Algeria.

— D. S. Korti and Z.Slimane. "Improving Dynamic Hand Gesture Recognition based IR-UWB using Offline Data Augmentation and Deep Learning." IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET), 2022, Kota Kinabalu, Malaysia, p. 1-6.

— D. S. Korti and Z.Slimane. "Multimodal Radar Data Fusion for Human Activity Recognition." The 2nd International Conference on Innovative Solutions in Software

Engineering (ICISSE), 2023, Ivano-Frankivsk, Ukraine.

— D. S. Korti and Z.Slimane. "Exploring the Potential of UWB Radar in Human Activity Recognition: A Brief Survey." The 2nd International Conference on Innovative Solutions in Software Engineering (ICISSE), 2023, Ivano-Frankivsk, Ukraine.

— D. S. Korti and Z.Slimane. "A Comparative Analysis of Multi-Sensor Radar Integration for Human Activity Recognition." The 2nd International Conference on Information Technologies and Educational Engineering (ICITEE23), 2023, Tirana, Albania.

Communications nationales

— D. S. Korti, Z.Slimane and K. Lakhdari. "Classification des obstacles pour véhicule autonome en utilisant l'apprentissage profond". Journée d'études sur les télécommunications, 2021, Université Belhadj Bouchaib Ain Témouchent.

Bibliographie

[1] A. Podury, S. M. Raefsky, L. Dodakian, L. McCafferty, V. Le, A. McKenzie, J. See, R. J. Zhou, T. Nguyen, B. Vanderschelden, *et al.*, "Social network structure is related to functional improvement from home-based telerehabilitation after stroke," *Frontiers in neurology*, vol. 12, p. 603767, 2021.

[2] C. Grefkes and G. R. Fink, "Recovery from stroke: current concepts and future perspectives," *Neurological research and practice*, vol. 2, no. 1, pp. 1–10, 2020.

[3] Y. Béjot, H. Bailly, M. Graber, L. Garnier, A. Laville, L. Dubourget, N. Mielle, C. Chevalier, J. Durier, and M. Giroud, "Impact of the ageing population on the burden of stroke: the dijon stroke registry," *Neuroepidemiology*, vol. 52, no. 1-2, pp. 78–85, 2019.

[4] O. O. Seminog, P. Scarborough, F. L. Wright, M. Rayner, and M. J. Goldacre, "Determinants of the decline in mortality from acute stroke in england: linked national database study of 795 869 adults," *Bmj*, vol. 365, 2019.

[5] S. Strilciuc, D. A. Grad, C. Radu, D. Chira, A. Stan, M. Ungureanu, A. Gheorghe, and F.-D. Muresanu, "The economic burden of stroke: a systematic review of cost of illness studies," *Journal of medicine and life*, vol. 14, no. 5, p. 606, 2021.

[6] J. R. Villar, S. González, J. Sedano, C. Chira, and J. M. Trejo-Gabriel-Galan, "Improving human activity recognition and its application in early stroke diagnosis," *International journal of neural systems*, vol. 25, no. 04, p. 1450036, 2015.

[7] H. U. Nam, J. S. Huh, J. N. Yoo, J. M. Hwang, B. J. Lee, Y.-S. Min, C.-H. Kim, and T.-D. Jung, "Effect of dominant hand paralysis on quality of life in patients with subacute stroke," *Annals of rehabilitation medicine*, vol. 38, no. 4, pp. 450–457, 2014.

[8] S. Anwer, A. Waris, S. O. Gilani, J. Iqbal, N. Shaikh, A. N. Pujari, and I. K. Niazi, "Reply to morone, g.; giansanti, d. comment on "anwer et al. rehabilitation of upper limb motor impairment in stroke: A narrative review on the prevalence, risk factors, and economic statistics of stroke and state of the art therapies. healthcare 2022, 10, 190"," in *Healthcare*, vol. 10, p. 847, MDPI, 2022.

[9] J. C. Martins, L. T. Aguiar, S. Nadeau, A. A. Scianni, L. F. Teixeira-Salmela, and C. D. C. de Morais Faria, "Measurement properties of self-report physical activity assessment tools in stroke: a protocol for a systematic review," *BMJ open*, vol. 7, no. 2, p. e012655, 2017.

[10] M. H. Lee, D. P. Siewiorek, A. Smailagic, A. Bernardino, and S. B. i. Badia, "Enabling ai and robotic coaches for physical rehabilitation therapy: iterative design and evaluation with therapists and post-stroke survivors," *International Journal of Social Robotics*, pp. 1–22, 2022.

[11] A. McCluskey, M. Lannin, K. Schurr, S. Dorsch, M. Curtin, J. Adams, and M. Egan, "Optimizing motor performance and sensation after brain impairment," in *:*, Elsevier, 2017.

[12] C. S. Mang and S. Peters, "Advancing motor rehabilitation for adults with chronic neurological conditions through increased involvement of kinesiologists: a perspective review," *BMC Sports Science, Medicine and Rehabilitation*, vol. 13, no. 1, pp. 1–11, 2021.

[13] C. E. Lang, J. R. MacDonald, D. S. Reisman, L. Boyd, T. J. Kimberley, S. M. Schindler-Ivens, T. G. Hornby, S. A. Ross, and P. L. Scheets, "Observation of amounts of movement practice provided during stroke rehabilitation," *Archives of physical medicine and rehabilitation*, vol. 90, no. 10, pp. 1692–1698, 2009.

[14] J. F. Gallagher, M. Sivan, and M. Levesley, "Making best use of home-based rehabilitation robots," *Applied Sciences*, vol. 12, no. 4, p. 1996, 2022.

[15] G. Burdea, N. Kim, K. Polistico, A. Kadaru, N. Grampurohit, D. Roll, and F. Damiani, "Assistive game controller for artificial intelligence-enhanced telerehabilitation post-stroke," *Assistive technology*, vol. 33, no. 3, pp. 117–128, 2021.

[16] I. Boukhennoufa, X. Zhai, V. Utti, J. Jackson, and K. D. McDonald-Maier, "Wearable sensors and machine learning in post-stroke rehabilitation assessment: A systematic review," *Biomedical Signal Processing and Control*, vol. 71, p. 103197, 2022.

[17] G. Burdea, N. Kim, K. Polistico, A. Kadaru, N. Grampurohit, D. Roll, and F. Damiani, "Assistive game controller for artificial intelligence-enhanced telerehabilitation post-stroke," *Assistive technology*, vol. 33, no. 3, pp. 117–128, 2021.

[18] I. Boukhennoufa, X. Zhai, V. Utti, J. Jackson, and K. D. McDonald-Maier, "Wearable sensors and machine learning in post-stroke rehabilitation assessment: A systematic review," *Biomedical Signal Processing and Control*, vol. 71, p. 103197, 2022.

[19] F. Porciuncula, A. V. Roto, D. Kumar, I. Davis, S. Roy, C. J. Walsh, and L. N. Awad, "Wearable movement sensors for rehabilitation: a focused review of technological and clinical advances," *Pm&r*, vol. 10, no. 9, pp. S220–S232, 2018.

[20] J. C. Augusto, V. Callaghan, D. Cook, A. Kameas, and I. Satoh, "Intelligent environments: a manifesto," *Human-centric Computing and Information Sciences*, vol. 3, pp. 1–18, 2013.

[21] A. Aztiria, A. Izaguirre, and J. C. Augusto, "Learning patterns in ambient intelligence environments: a survey," *Artificial Intelligence Review*, vol. 34, pp. 35–51, 2010.

[22] M. Mohammadi, H. Arts, and R. d. Pagter, "Contribution of ambient intelligence to the subjective wellbeing: an overview of the advantages and disadvantages," in *Proceedings of the IADIS International Conference (e-Health'11)*, 2012.

[23] G. Adamson, "Safeguards in a world of ambient intelligence. series: The international library of ethics, law and technology," 2009.

[24] R. Godwin-Jones, "Emerging spaces for language learning: Ai bots, ambient intelligence, and the metaverse," 2023.

[25] A. Scemama, "Guide des connaissances sur l'activité physique et la sédentarité," *Haute autorité de santé*, 2022.

[26] C. J. Caspersen, K. E. Powell, and G. M. Christenson, "Physical activity, exercise, and physical fitness: definitions and distinctions for health-related research.," *Public health reports*, vol. 100, no. 2, p. 126, 1985.

[27] S. J. Strath, L. A. Kaminsky, B. E. Ainsworth, U. Ekelund, P. S. Freedson, R. A. Gary, C. R. Richardson, D. T. Smith, and A. M. Swartz, "Guide to the assessment of physical activity: clinical and research applications: a scientific statement from the american heart association," *Circulation*, vol. 128, no. 20, pp. 2259–2279, 2013.

[28] A. Benmansour, A. Bouchachia, and M. Feham, "Multioccupant activity recognition in pervasive smart home environments," *ACM Computing Surveys (CSUR)*, vol. 48, no. 3, pp. 1–36, 2015.

[29] G. Saleem, U. I. Bajwa, and R. H. Raza, "Toward human activity recognition: a survey," *Neural Computing and Applications*, vol. 35, no. 5, pp. 4145–4182, 2023.

[30] M. G. Morshed, T. Sultana, A. Alam, and Y.-K. Lee, "Human action recognition: A taxonomy-based survey, updates, and opportunities," *Sensors*, vol. 23, no. 4, p. 2182, 2023.

[31] Y. Li, G. Yang, Z. Su, S. Li, and Y. Wang, "Human activity recognition based on multienvironment sensor data," *Information Fusion*, vol. 91, pp. 47–63, 2023.

[32] H. Du, T. Jin, Y. He, Y. Song, and Y. Dai, "Segmented convolutional gated recurrent neural networks for human activity recognition in ultra-wideband radar," *Neurocomputing*, vol. 396, pp. 451–464, 2020.

[33] S. Z. Gurbuz, M. M. Rahman, E. Kurtoglu, T. Macks, and F. Fioranelli, "Cross-frequency training with adversarial learning for radar micro-doppler signature classification (rising researcher)," in *Radar Sensor Technology XXIV*, vol. 11408, pp. 58–68, SPIE, 2020.

[34] K. Zuo, X. Li, Y. Dong, J. Dezert, and S. S. Ge, "Combination of different-granularity beliefs for sensor-based human activity recognition," *IEEE Sensors Journal*, 2023.

[35] A. Chakraborty and N. Mukherjee, "A deep-cnn based low-cost, multi-modal sensing system for efficient walking activity identification," *Multimedia Tools and Applications*, vol. 82, no. 11, pp. 16741–16766, 2023.

[36] B. Jokanovic, M. Amin, and B. Erol, "Multiple joint-variable domains recognition of human motion," in *2017 IEEE Radar Conference (RadarConf)*, pp. 0948–0952, IEEE, 2017.

[37] G. Diraco, G. Rescio, P. Siciliano, and A. Leone, "Review on human action recognition in smart living: Sensing technology, multimodality, real-time processing, interoperability, and resource-constrained processing," *Sensors*, vol. 23, no. 11, p. 5281, 2023.

[38] R. Elbasiony and W. Gomaa, "A survey on human activity recognition based on temporal signals of portable inertial sensors," in *The International Conference on Advanced Machine Learning Technologies and Applications (AMLTA2019) 4*, pp. 734–745, Springer, 2020.

[39] X. Wang, H. Yu, S. Kold, O. Rahbek, and S. Bai, "Wearable sensors for activity monitoring and motion control: A review," *Biomimetic Intelligence and Robotics*, p. 100089, 2023.

[40] S. Jindal, M. Sachdeva, and A. K. S. Kushwaha, "Deep learning for video based human activity recognition: Review and recent developments," in *Proceedings of International Conference on Computational Intelligence and Emerging Power System: ICCIPS 2021*, pp. 71–83, Springer, 2022.

[41] A. Sánchez-Caballero, D. Fuentes-Jiménez, and C. Losada-Gutiérrez, "Real-time human action recognition using raw depth video-based recurrent neural networks," *Multimedia Tools and Applications*, vol. 82, no. 11, pp. 16213–16235, 2023.

[42] A. B. Sargano, P. Angelov, and Z. Habib, "A comprehensive review on handcrafted and learning-based action representation approaches for human activity recognition," *applied sciences*, vol. 7, no. 1, p. 110, 2017.

[43] K. Chaccour, R. Darazi, A. H. el Hassans, and E. Andres, "Smart carpet using differential piezoresistive pressure sensors for elderly fall detection," in *2015 IEEE 11th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*, pp. 225–229, IEEE, 2015.

[44] S. Zhou, G. Lin, Q. Qian, and C. Xu, "Binary classification of floor vibrations for human activity detection based on dynamic mode decomposition," *Neurocomputing*, vol. 432, pp. 227–239, 2021.

[45] A. Zhuravchak, O. Kapshii, and E. Pournaras, "Human activity recognition based on wi-fi csi data-a deep neural network approach," *Procedia Computer Science*, vol. 198, pp. 59–66, 2022.

[46] H. Sarieddeen, M.-S. Alouini, and T. Y. Al-Naffouri, "An overview of signal processing techniques for terahertz communications," *Proceedings of the IEEE*, vol. 109, no. 10, pp. 1628–1665, 2021.

[47] W. Jiang, C. Miao, F. Ma, S. Yao, Y. Wang, Y. Yuan, H. Xue, C. Song, X. Ma, D. Koutsonikolas, *et al.*, "Towards environment independent device free human activity recognition," in *Proceedings of the 24th annual international conference on mobile computing and networking*, pp. 289–304, 2018.

[48] B. Fu, N. Damer, F. Kirchbuchner, and A. Kuijper, "Sensing technology for human activity recognition: A comprehensive survey," *Ieee Access*, vol. 8, pp. 83791–83820, 2020.

[49] N. Gupta, S. K. Gupta, R. K. Pathak, V. Jain, P. Rashidi, and J. S. Suri, "Human activity recognition in artificial intelligence framework: A narrative review," *Artificial intelligence review*, vol. 55, no. 6, pp. 4755–4808, 2022.

[50] R. Banstola, R. Bera, and D. Bhaskar, "Review and design of uwb transmitter and receiver," *International Journal of Computer Applications*, vol. 69, no. 13, 2013.

[51] L.-T. Wang, Y. Xiong, M. He, and M. Kheir, "Review on uwb bandpass filters," in *UWB Technology-Circuits and Systems*, pp. 1–24, IntechOpen London, United Kingdom, 2019.

[52] X. Li, Y. He, and X. Jing, "A survey of deep learning-based human activity recognition in radar," *Remote Sensing*, vol. 11, no. 9, p. 1068, 2019.

[53] S. Ahmed, K. D. Kallu, S. Ahmed, and S. H. Cho, "Hand gestures recognition using radar sensors for human-computer-interaction: A review," *Remote Sensing*, vol. 13, no. 3, p. 527, 2021.

[54] D. Yim, W. H. Lee, J. I. Kim, K. Kim, D. H. Ahn, Y.-H. Lim, S. H. Cho, H.-K. Park, and S. H. Cho, "Quantified activity measurement for medical use in movement disorders through ir-uwb radar sensor," *Sensors*, vol. 19, no. 3, p. 688, 2019.

[55] Y. Wang, J. Zhou, J. Tong, and X. Wu, "Uwb-radar-based synchronous motion recognition using time-varying range–doppler images," *IET Radar, Sonar & Navigation*, vol. 13, no. 12, pp. 2131–2139, 2019.

[56] M. Piriyajitakonkij, P. Warin, P. Lakhan, P. Leelaarporn, N. Kumchaiseemak, S. Suwajanakorn, T. Pianpanit, N. Niparnan, S. C. Mukhopadhyay, and T. Wilaiprasitporn, "Sleepposenet: Multi-view learning for sleep postural transition recognition using uwb," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 4, pp. 1305–1314, 2020.

[57] H. Sadreazami, M. Bolic, and S. Rajan, "Tl-fall: Contactless indoor fall detection using transfer learning from a pretrained model," in *2019 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, pp. 1–5, IEEE, 2019.

[58] A. H. Victoria, V. Gayathri, and A. Vasudevan, "A deep convolutional neural network for remote life activities detection using fmcw radar under realistic environments," in *International Conference on Information Systems and Management Science*, pp. 400–412, Springer, 2022.

[59] C. Ding, L. Zhang, C. Gu, L. Bai, Z. Liao, H. Hong, Y. Li, and X. Zhu, "Non-contact human motion recognition based on uwb radar," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 8, no. 2, pp. 306–315, 2018.

[60] S. Ahmed, F. Khan, A. Ghaffar, F. Hussain, and S. H. Cho, "Finger-counting-based gesture recognition within cars using impulse radar with convolutional neural network," *Sensors*, vol. 19, no. 6, p. 1429, 2019.

[61] S. Ahmed and S. H. Cho, "Hand gesture recognition using an ir-uwb radar with an inception module-based classifier," *Sensors*, vol. 20, no. 2, p. 564, 2020.

[62] J. Park and S. H. Cho, "Ir-uwb radar sensor for human gesture recognition by using machine learning," in *2016 IEEE 18th International Conference on High Performance Computing and Communications; IEEE 14th International Conference on Smart City; IEEE 2nd International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, pp. 1246–1249, IEEE, 2016.

[63] A. Chowdhury, T. Das, S. Rani, A. Khasnobish, and T. Chakravarty, "Activity recognition using ultra wide band range-time scan," in *2020 28th European Signal Processing Conference (EUSIPCO)*, pp. 1338–1342, IEEE, 2021.

[64] M. S. Seyfioglu, B. Erol, S. Z. Gurbuz, and M. G. Amin, "Diversified radar micro-doppler simulations as training data for deep residual neural networks," in *2018 IEEE radar Conference (radarConf18)*, pp. 0612–0617, IEEE, 2018.

[65] Y.-J. Zhong and Q.-S. Li, "Human motion recognition in small sample scenarios based on gan and cnn models," *Prog. Electromagn. Res. M*, vol. 113, pp. 101–113, 2022.

[66] H. Li, A. Mehul, J. Le Kernec, S. Z. Gurbuz, and F. Fioranelli, "Sequential human gait classification with distributed radar sensor fusion," *IEEE Sensors Journal*, vol. 21, no. 6, pp. 7590–7603, 2020.

[67] X. Li, Z. Li, F. Fioranelli, S. Yang, O. Romain, and J. L. Kernec, "Hierarchical radar data analysis for activity and personnel recognition," *Remote Sensing*, vol. 12, no. 14, p. 2237, 2020.

[68] S. Wang, J. Song, J. Lien, I. Poupyrev, and O. Hilliges, "Interacting with soli: Exploring fine-grained dynamic gesture recognition in the radio-frequency spectrum," in *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, pp. 851–860, 2016.

[69] S. Z. Gurbuz and M. G. Amin, "Radar-based human-motion recognition with deep learning: Promising applications for indoor monitoring," *IEEE Signal Processing Magazine*, vol. 36, no. 4, pp. 16–28, 2019.

[70] Z. Zhang, Z. Tian, and M. Zhou, "Latern: Dynamic continuous hand gesture recognition using fmcw radar sensor," *IEEE Sensors Journal*, vol. 18, no. 8, pp. 3278–3289, 2018.

[71] N. Ren, X. Quan, and S. H. Cho, "Algorithm for gesture recognition using an ir-uwb radar sensor," *Journal of Computer and Communications*, vol. 4, no. 3, 2016.

[72] H. Sadreazami, M. Bolic, and S. Rajan, "On the use of ultra wideband radar and stacked lstm-rnn for at home fall detection," in *2018 IEEE Life Sciences Conference (LSC)*, pp. 255–258, IEEE, 2018.

[73] H. Sadreazami, M. Bolic, and S. Rajan, "Residual network-based supervised learning of remotely sensed fall incidents using ultra-wideband radar," in *2019 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 1–4, IEEE, 2019.

[74] S. Y. Kim, H. G. Han, J. W. Kim, S. Lee, and T. W. Kim, "A hand gesture recognition sensor using reflected impulses," *IEEE Sensors Journal*, vol. 17, no. 10, pp. 2975–2976, 2017.

[75] F. Khan and S. H. Cho, "Hand based gesture recognition inside a car through ir-uwb radar," , pp. 154–157, 2017.

[76] F. Khan, S. K. Leem, and S. H. Cho, "Hand-based gesture recognition for vehicular applications using ir-uwb radar," *Sensors*, vol. 17, no. 4, p. 833, 2017.

[77] A. Ghaffar, F. Khan, and S. H. Cho, "Hand pointing gestures based digital menu board implementation using ir-uwb transceivers," *IEEE Access*, vol. 7, pp. 58148–58157, 2019.

[78] S. K. Leem, F. Khan, and S. H. Cho, "Detecting mid-air gestures for digit writing with radio sensors and a cnn," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 4, pp. 1066–1081, 2019.

[79] F. Khan, S. K. Leem, and S. H. Cho, "In-air continuous writing using uwb impulse radar sensors," *IEEE Access*, vol. 8, pp. 99302–99311, 2020.

[80] Y. Shao, S. Guo, L. Sun, and W. Chen, "Human motion classification based on range information with deep convolutional neural network," in *2017 4th International Conference on Information Science and Control Engineering (ICISCE)*, pp. 1519–1523, IEEE, 2017.

[81] D. Kumar, A. Sarkar, S. R. Kerketta, and D. Ghosh, "Human activity classification based on breathing patterns using ir-uwb radar," in *2019 IEEE 16th India Council International Conference (INDICON)*, pp. 1–4, IEEE, 2019.

[82] H. Sadreazami, M. Bolic, and S. Rajan, "Capsfall: Fall detection using ultra-wideband radar and capsule network," *IEEE Access*, vol. 7, pp. 55336–55343, 2019.

[83] M. M. Rahman and S. Z. Gurbuz, "Multi-frequency rf sensor data adaptation for motion recognition with multi-modal deep learning," in *2021 IEEE Radar Conference (RadarConf21)*, pp. 1–6, IEEE, 2021.

[84] H. Du, Y. He, and T. Jin, "Transfer learning for human activities classification using micro-doppler spectrograms," in *2018 IEEE International Conference on Computational Electromagnetics (ICCEM)*, pp. 1–3, IEEE, 2018.

[85] H. Du, T. Jin, Y. Song, Y. Dai, and M. Li, "Efficient human activity classification via sparsity-driven transfer learning," *IET Radar, Sonar & Navigation*, vol. 13, no. 10, pp. 1741–1746, 2019.

[86] S. Yang, J. Le Kernec, O. Romain, F. Fioranelli, P. Cadart, J. Fix, C. Ren, G. Manfredi, T. Letertre, I. D. H. Sáenz, *et al.*, "The human activity radar challenge: Benchmarking based on the 'radar signatures of human activities' dataset from glasgow university," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 4, pp. 1813–1824, 2023.

[87] R. Bravin, L. Nanni, A. Loreggia, S. Brahnam, and M. Paci, "Varied image data augmentation methods for building ensemble," *IEEE Access*, vol. 11, pp. 8810–8823, 2023.

[88] A. Oubara, F. Wu, A. Amamra, and G. Yang, "Survey on remote sensing data augmentation: Advances, challenges, and future perspectives," in *International Conference on Computing Systems and Applications*, pp. 95–104, Springer, 2022.

[89] K. Devanand Bathe and N. S. Patil, "Leveraging potential of deep learning for remote sensing data: A review," *Intelligent Systems and Human Machine Collaboration: Select Proceedings of ICISHMC 2022*, pp. 129–145, 2023.

[90] Y. Lang, C. Hou, H. Ji, and Y. Yang, "A dual generation adversarial network for human motion detection using micro-doppler signatures," *IEEE Sensors Journal*, vol. 21, no. 16, pp. 17995–18003, 2021.

[91] L. Qu, Y. Wang, T. Yang, and Y. Sun, "Human activity recognition based on wrgan-gp-synthesized micro-doppler spectrograms," *IEEE Sensors Journal*, vol. 22, no. 9, pp. 8960–8973, 2022.

[92] B. Erol, S. Z. Gurbuz, and M. G. Amin, "Motion classification using kinematically sifted acgan-synthesized radar micro-doppler signatures," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 56, no. 4, pp. 3197–3213, 2020.

[93] X. Shi, Y. Li, F. Zhou, and L. Liu, "Human activity recognition based on deep learning method," in *2018 International Conference on Radar (RADAR)*, pp. 1–5, IEEE, 2018.

[94] Y. Lang, Q. Wang, Y. Yang, C. Hou, D. Huang, and W. Xiang, "Unsupervised domain adaptation for micro-doppler human motion classification via feature fusion," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 3, pp. 392–396, 2018.

[95] H. Du, T. Jin, Y. Song, and Y. Dai, "Unsupervised adversarial domain adaptation for micro-doppler based human activity classification," *IEEE geoscience and remote sensing letters*, vol. 17, no. 1, pp. 62–66, 2019.

[96] B. Erol and M. G. Amin, "Fall motion detection using combined range and doppler features," in *2016 24th European Signal Processing Conference (EUSIPCO)*, pp. 2075–2080, IEEE, 2016.

[97] B. Erol and M. G. Amin, "Radar data cube analysis for fall detection," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2446–2450, IEEE, 2018.

[98] W. Chen, C. Ding, Y. Zou, L. Zhang, C. Gu, H. Hong, and X. Zhu, "Non-contact human activity classification using dcnn based on uwb radar," in *2019 IEEE MTT-S International Microwave Biomedical Conference (IMBioC)*, vol. 1, pp. 1–4, IEEE, 2019.

[99] S. Skaria, D. Huang, A. Al-Hourani, R. J. Evans, and M. Lech, "Deep-learning for hand-gesture recognition with simultaneous thermal and radar sensors," in *2020 IEEE SENSORS*, pp. 1–4, IEEE, 2020.

[100] F. M. Noori, M. Z. Uddin, and J. Torresen, "Ultra-wideband radar-based activity recognition using deep learning," *IEEE Access*, vol. 9, pp. 138132–138143, 2021.

[101] S. Ahmed, D. Wang, J. Park, and S. H. Cho, "Uwb-gestures, a public dataset of dynamic hand gestures acquired using impulse radar sensors," *Scientific Data*, vol. 8, no. 1, p. 102, 2021.

[102] K. Bouchard, J. Maitre, C. Bertuglia, and S. Gaboury, "Activity recognition in smart homes using uwb radars," *Procedia Computer Science*, vol. 170, pp. 10–17, 2020.

[103] M. Piriyajitakonkij, P. Warin, P. Lakhan, P. Leelaarporn, N. Kumchaiseemak, S. Suwajanakorn, T. Pianpanit, N. Niparnan, S. C. Mukhopadhyay, and T. Wilaiprasitporn, "Sleepposenet: Multi-view learning for sleep postural transition recognition

using uwb," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 4, pp. 1305–1314, 2020.

[104] M. Mostafa and S. Chamaani, "Unobtrusive human activity classification based on combined time-range and time-frequency domain signatures using ultrawideband radar," *IET Signal Processing*, vol. 15, no. 8, pp. 543–561, 2021.

[105] F. Khan, S. K. Leem, and S. H. Cho, "Human–computer interaction using radio sensor for people with severe disability," *Sensors and Actuators A: Physical*, vol. 282, pp. 39–54, 2018.

[106] G. Mokhtari, Q. Zhang, and A. Fazlollahi, "Non-wearable uwb sensor to detect falls in smart home environment," in *2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, pp. 274–278, IEEE, 2017.

[107] J. Bryan and Y. Kim, "Classification of human activities on uwb radar using a support vector machine," in *2010 IEEE Antennas and Propagation Society International Symposium*, pp. 1–4, IEEE, 2010.

[108] G. Diraco, A. Leone, and P. Siciliano, "A fall detector based on ultra-wideband radar sensing," in *Sensors: Proceedings of the Third National Conference on Sensors, February 23-25, 2016, Rome, Italy 3*, pp. 373–382, Springer, 2018.

[109] L. Ma, M. Liu, N. Wang, L. Wang, Y. Yang, and H. Wang, "Room-level fall detection based on ultra-wideband (uwb) monostatic radar and convolutional long short-term memory (lstm)," *Sensors*, vol. 20, no. 4, p. 1105, 2020.

[110] S. Sharma, H. Mohammadmoradi, M. Heydariaan, and O. Gnawali, "Device-free activity recognition using ultra-wideband radios," in *2019 International Conference on Computing, Networking and Communications (ICNC)*, pp. 1029–1033, IEEE, 2019.

[111] M. Bocus, R. Piechocki, and K. Chetty, "A comparison of uwb cir and wifi csi for human activity recognition," IEEE Radar Conference (RadarCon), 2021.

[112] A. Li, E. Bodanese, S. Poslad, T. Hou, K. Wu, and F. Luo, "A trajectory-based gesture recognition in smart homes based on the ultrawideband communication system," *IEEE Internet of Things Journal*, vol. 9, no. 22, pp. 22861–22873, 2022.

[113] G. Park, V. K. Chandrasegar, and J. Koh, "Accuracy enhancement of hand gesture recognition using cnn," *IEEE Access*, vol. 11, pp. 26496–26501, 2023.

[114] S. Sani, S. Massie, N. Wiratunga, and K. Cooper, "Learning deep and shallow features for human activity recognition," in *International conference on knowledge science, engineering and management*, pp. 469–482, Springer, 2017.

[115] L. K. Hansen and P. Salamon, "Neural network ensembles," *IEEE transactions on pattern analysis and machine intelligence*, vol. 12, no. 10, pp. 993–1001, 1990.

[116] J. Maitre, K. Bouchard, C. Bertuglia, and S. Gaboury, "Recognizing activities of daily living from uwb radars and deep learning," *Expert Systems with Applications*, vol. 164, p. 113994, 2021.

[117] N. Hendy, H. M. Fayek, and A. Al-Hourani, "Deep learning approaches for air-writing using single uwb radar," *IEEE Sensors Journal*, vol. 22, no. 12, pp. 11989–12001, 2022.

[118] X. Yang, P. Chen, M. Wang, S. Guo, C. Jia, and G. Cui, "Human motion serialization recognition with through-the-wall radar," *IEEE Access*, vol. 8, pp. 186879–186889, 2020.

[119] S. Sadatmir and N. K. Maryamabadi, "Hesitation strategies engaged by foreign language (fl) learners during an interview," *LANGUAGE & COMMUNICATION*, vol. 1, no. 2, pp. 188–201, 2014.

[120] J. Galván-Ruiz, C. M. Travieso-González, A. Tejera-Fettmilch, A. Pinan-Roescher, L. Esteban-Hernández, and L. Domínguez-Quintana, "Perspective and evolution of gesture recognition for sign language: A review," *Sensors*, vol. 20, no. 12, p. 3571, 2020.

[121] K. Faheem, L. Seong Kyu, and S. H. CHO, "Algorithm for fingers counting gestures using ir-uwb radar sensor," in *2018 IEEE Sensors Applications Symposium*, pp. 144–146, IEEE, 2018.

[122] A. Ghaffar, F. Khan, and S. H. Cho, "Hand pointing gestures based digital menu board implementation using ir-uwb transceivers," *IEEE Access*, vol. 7, pp. 58148–58157, 2019.

[123] B. Li, J. Yang, Y. Yang, C. Li, and Y. Zhang, "Sign language/gesture recognition based on cumulative distribution density features using uwb radar," *IEEE transactions on instrumentation and measurement*, vol. 70, pp. 1–13, 2021.

[124] Y. Li, X. Wang, B. Shi, and M. Zhu, "Hand gesture recognition using ir-uwb radar with shufflenet v2," in *Proceedings of the 5th International Conference on Control Engineering and Artificial Intelligence*, pp. 126–131, 2021.

[125] J. Park, J. Jang, G. Lee, H. Koh, C. Kim, and T. W. Kim, "A time domain artificial intelligence radar system using 33-ghz direct sampling for hand gesture recognition," *IEEE Journal of Solid-State Circuits*, vol. 55, no. 4, pp. 879–888, 2020.

[126] S. Skaria, A. Al-Hourani, and R. J. Evans, "Deep-learning methods for hand-gesture recognition using ultra-wideband radar," *IEEE Access*, vol. 8, pp. 203580–203590, 2020.

[127] L. Qiao, Z. Li, B. Xiao, Y. Shu, W. Li, and X. Gao, "Gesture-proxylessnas: A lightweight network for mid-air gesture recognition based on uwb radar," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2023.

[128] R. R. Sharma, K. A. Kumar, and S. H. Cho, "Novel time-distance parameters based hand gesture recognition system using multi-uwb radars," *IEEE Sensors Letters*, vol. 7, no. 5, pp. 1–4, 2023.

[129] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea, "Machine recognition of human activities: A survey," *IEEE Transactions on Circuits and Systems for Video technology*, vol. 18, no. 11, pp. 1473–1488, 2008.

[130] C. Jobanputra, J. Bavishi, and N. Doshi, "Human activity recognition: A survey," *Procedia Computer Science*, vol. 155, pp. 698–703, 2019.

[131] J. Bryan, J. Kwon, N. Lee, and Y. Kim, "Application of ultra-wide band radar for classification of human activities," *IET Radar, Sonar & Navigation*, vol. 6, no. 3, pp. 172–179, 2012.

[132] K. Ota, Y. Ota, M. Otsu, and A. Kajiwara, "Elderly-care motion sensor using uwb-ir," in *2011 IEEE Sensors Applications Symposium*, pp. 159–162, IEEE, 2011.

[133] K. Saho, T. Sakamoto, T. Sato, K. Inoue, and T. Fukuda, "Accurate and real-time pedestrian classification based on uwb doppler radar images and their radial velocity features," *IEICE transactions on communications*, vol. 96, no. 10, pp. 2563–2572, 2013.

[134] M. A. Kiasari, S. Y. Na, and J. Y. Kim, "Classification of human postures using ultra-wide band radar based on neural networks," in *2014 International Conference on IT Convergence and Security (ICITCS)*, pp. 1–4, IEEE, 2014.

[135] B. Erol, M. Amin, Z. Zhou, and J. Zhang, "Range information for reducing fall false alarms in assisted living," in *2016 IEEE Radar Conference (RadarConf)*, pp. 1–6, IEEE, 2016.

[136] G. Mokhtari, S. Aminikhanghahi, Q. Zhang, and D. J. Cook, "Fall detection in smart home environments using uwb sensors and unsupervised change detection," *Journal of Reliable Intelligent Environments*, vol. 4, pp. 131–139, 2018.

[137] A.-K. Seifert, A. M. Zoubir, and M. G. Amin, "Radar-based human gait recognition in cane-assisted walks," in *2017 IEEE Radar Conference (RadarConf)*, pp. 1428–1433, IEEE, 2017.

[138] Z. Baird, S. Rajan, and M. Bolic, "Classification of human posture from radar returns using ultra-wideband radar," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 3268–3271, IEEE, 2018.

[139] A.-K. Seifert, A. M. Zoubir, and M. G. Amin, "Radar classification of human gait abnormality based on sum-of-harmonics analysis," in *2018 IEEE Radar Conference (RadarConf18)*, pp. 0940–0945, IEEE, 2018.

[140] G. Wang and Z. Zhu, "Classification of human motion status using uwb radar based on decision tree algorithm," in *Communications, Signal Processing, and Systems: Proceedings of the 8th International Conference on Communications, Signal Processing, and Systems 8th*, pp. 985–992, Springer, 2020.

[141] K. Tsuchiyama and A. Kajiwara, "Accident detection and health-monitoring uwb sensor in toilet," in *2019 IEEE Topical Conference on Wireless Sensors and Sensor Networks (WiSNet)*, pp. 1–4, IEEE, 2019.

[142] M. Hämäläinen, L. Mucchi, S. Caputo, L. Biotti, L. Ciani, D. Marabissi, and G. Patrizi, "Ultra-wideband radar-based indoor activity monitoring for elderly care," *Sensors*, vol. 21, no. 9, p. 3158, 2021.

[143] F. Qi, Z. Li, Y. Ma, F. Liang, H. Lv, J. Wang, and A. E. Fathy, "Generalization of channel micro-doppler capacity evaluation for improved finer-grained human activity classification using mimo uwb radar," *IEEE Transactions on Microwave Theory and Techniques*, vol. 69, no. 11, pp. 4748–4761, 2021.

[144] T. Han, W. Kang, and G. Choi, "Ir-uwb sensor based fall detection method using cnn algorithm," *Sensors*, vol. 20, no. 20, p. 5948, 2020.

[145] H. Sadreazami, M. Bolic, and S. Rajan, "Contactless fall detection using time-frequency analysis and convolutional neural networks," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 10, pp. 6842–6851, 2021.

[146] I. Brishtel, S. Krauss, M. Chamseddine, J. R. Rambach, and D. Stricker, "Driving activity recognition using uwb radar and deep neural networks," *Sensors*, vol. 23, no. 2, p. 818, 2023.

[147] X. Jiang, L. Zhang, and L. Li, "Multi-task learning radar transformer (mlrt): A personal identification and fall detection network based on ir-uwb radar," *Sensors*, vol. 23, no. 12, p. 5632, 2023.

[148] R. Qi, X. Li, Y. Zhang, and Y. Li, "Multi-classification algorithm for human motion recognition based on ir-uwb radar," *IEEE Sensors Journal*, vol. 20, no. 21, pp. 12848–12858, 2020.

[149] J. Maitre, K. Bouchard, and S. Gaboury, "Fall detection with uwb radars and cnn-lstm architecture," *IEEE journal of biomedical and health informatics*, vol. 25, no. 4, pp. 1273–1283, 2020.

[150] D. A. Bordvik, J. Hou, F. M. Noori, M. Z. Uddin, and J. Torresen, "Monitoring in-home emergency situation and preserve privacy using multi-modal sensing and deep learning," in *2022 International Conference on Electronics, Information, and Communication (ICEIC)*, pp. 1–6, IEEE, 2022.

[151] T. Imbeault-Nepton, J. Maitre, K. Bouchard, and S. Gaboury, "Filtering data bins of uwb radars for activity recognition with random forest," *Procedia Computer Science*, vol. 201, pp. 48–55, 2022.

[152] Y. Zhao, R. G. Guendel, A. Yarovoy, and F. Fioranelli, "Distributed radar-based human activity recognition using vision transformer and cnns," in *2021 18th European Radar Conference (EuRAD)*, pp. 301–304, IEEE, 2022.

[153] S. Zhu, R. G. Guendel, A. Yarovoy, and F. Fioranelli, "Continuous human activity recognition with distributed radar sensor networks and cnn–rnn architectures," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.

[154] R. Guendel, N. C. Kruse, F. Fioranelli, and A. Yarovoy, "Exploiting radar data domains for classification with spatially distributed nodes," in *SET-312 Research Specialists' Meeting over" Distributed Multi-Spectral/Statics Sensing*, 2022.

[155] X. Yang, R. G. Guendel, A. Yarovoy, and F. Fioranelli, "Radar-based human activities classification with complex-valued neural networks," in *2022 IEEE Radar Conference (RadarConf22)*, pp. 1–6, IEEE, 2022.

[156] D. K.-H. Lai, L.-W. Zha, T. Y.-N. Leung, A. Y.-C. Tam, B. P.-H. So, H.-J. Lim, D. S. K. Cheung, D. W.-C. Wong, and J. C.-W. Cheung, "Dual ultra-wideband (uwb) radar-based sleep posture recognition system: Towards ubiquitous sleep monitoring," *Engineered Regeneration*, vol. 4, no. 1, pp. 36–43, 2023.

[157] X. Shi, X. Yao, X. Bai, F. Zhou, Y. Li, and L. Liu, "Radar echoes simulation of human movements based on mocap data and em calculation," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 6, pp. 859–863, 2019.

[158] S. S. Ram and H. Ling, "Simulation of human microdopplers using computer animation data," in *2008 IEEE Radar Conference*, pp. 1–6, IEEE, 2008.

[159] S. S. Ram, C. Christianson, Y. Kim, and H. Ling, "Simulation and analysis of human micro-dopplers in through-wall environments," *IEEE Transactions on Geoscience and remote sensing*, vol. 48, no. 4, pp. 2015–2023, 2010.

[160] B. Erol, C. Karabacak, S. Z. Gürbüz, and A. C. Gürbüz, "Simulation of human micro-doppler signatures with kinect sensor," in *2014 IEEE Radar Conference*, pp. 0863–0868, IEEE, 2014.

[161] B. Erol, S. Z. Gurbuz, and M. G. Amin, "Synthesis of micro-doppler signatures for abnormal gait using multi-branch discriminator with embedded kinematics," in *2020 IEEE International Radar Conference (RADAR)*, pp. 175–179, IEEE, 2020.

[162] M. M. Rahman, E. A. Malaia, A. C. Gurbuz, D. J. Griffin, C. Crawford, and S. Z. Gurbuz, "Effect of kinematics and fluency in adversarial synthetic data generation for asl recognition with rf sensors," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 58, no. 4, pp. 2732–2745, 2022.

[163] J. Fix, I. Hinostroza, C. Ren, G. Manfredi, and T. Letertre, "Transfer learning for human activity classification in multiple radar setups," in *2022 30th European Signal Processing Conference (EUSIPCO)*, pp. 1576–1580, IEEE, 2022.

[164] X. Li, X. Jing, and Y. He, "Unsupervised domain adaptation for human activity recognition in radar," in *2020 IEEE Radar Conference (RadarConf20)*, pp. 1–5, IEEE, 2020.

[165] S. P. Sahoo, R. Silambarasi, and S. Ari, "Fusion of histogram based features for human action recognition," in *2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS)*, pp. 1012–1016, IEEE, 2019.

[166] D. Sharifrazi, R. Alizadehsani, M. Roshanzamir, J. H. Joloudari, A. Shoeibi, M. Jafari, S. Hussain, Z. A. Sani, F. Hasanzadeh, F. Khozeimeh, *et al.*, "Fusion of convolution neural network, support vector machine and sobel filter for accurate detection of covid-19 patients using x-ray images," *Biomedical Signal Processing and Control*, vol. 68, p. 102622, 2021.

[167] H. Shahverdi, M. Nabati, P. Fard Moshiri, R. Asvadi, and S. A. Ghorashi, "Enhancing csi-based human activity recognition by edge detection techniques," *Information*, vol. 14, no. 7, p. 404, 2023.

[168] I. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio, "Maxout networks," in *International conference on machine learning*, pp. 1319–1327, PMLR, 2013.

[169] T. P. Lillicrap, A. Santoro, L. Marris, C. J. Akerman, and G. Hinton, "Backpropagation and the brain," *Nature Reviews Neuroscience*, vol. 21, no. 6, pp. 335–346, 2020.

[170] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 2623–2631, 2019.

[171] S. Vishwakarma, W. Li, C. Tang, K. Woodbridge, R. R. Adve, and K. Chetty, "Attention-enhanced alexnet for improved radar micro-doppler signature classification," *IET Radar, Sonar & Navigation*, vol. 17, no. 4, pp. 652–664, 2023.

[172] E. Guariglia, R. C. Guido, and G. J. Dalalana, "From wavelet analysis to fractional calculus: A review," *Mathematics*, vol. 11, no. 7, p. 1606, 2023.

[173] Y. Sun, L. Zhu, G. Wang, F. Zhao, *et al.*, "Multi-input convolutional neural network for flower grading," *Journal of Electrical and Computer Engineering*, vol. 2017, 2017.

[174] J. A. Aghamaleki and V. Ashkani Chenarlogh, "Multi-stream cnn for facial expression recognition in limited training data," *Multimedia Tools and Applications*, vol. 78, no. 16, pp. 22861–22882, 2019.