جامعة بجاية
Tasdawit n Bgayet
Université de Béjaïa

République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université A. Mira de Béjaia
Faculté des Sciences Exacts
Département d'Informatique

MÉMOIRE DE MASTER RECHERCHE
En Informatique

Option : *Systèmes d'informations avancés*

Thème

Reconnaissance des activités humaines à base de réseaux de neurones profonds

Présenté par :

Merzouk Lyza et Kasdi Nesrine

Soutenu le 01 juillet 2024 devant le jury composé de :

Présidente	Mme Nassima BOUADEM	U. A/Mira Béjaia.
Encadrant	Mr Achour ACHROUFENE	U. A/Mira Béjaia.
Co-Encadrant	Mme Salima SABRI	U. A/Mira Béjaia.
Examinatrice	Mme Amel HOUHA	U. A/Mira Béjaia.

Année universitaire 2023/2024

Dédicace

À ma mère qui a toujours travaillé sans relâche pour assurer notre confort et notre succès. Pour mon père sans qui je n'aurais jamais atteint où j'en suis aujourd'hui.

À mes sœurs Taous et Lylia qui n'ont jamais cessé de me soutenir, de m'encourager et de croire en moi.

À mon seul frère Aghilas, pour son appui inconditionnel et sa confiance en moi.

À mes chères petites nièces Miral et Jana... je vous aime énormément.

À mes chères amies : Hadjer, Tina, Thilleli, Lynda, Saida pour la meilleure compagnie qui puisse être et qui m'ont aidé dans plusieurs situations.

Je souhaite remercier certaines personnes pour leur soutien et leurs encouragements discrètement tout au long de mon parcours. Votre aide a été inestimable. Merci infiniment.

Pour finir, une dédicace particulière à mon binôme « Nesrine ».

Lyza.

Dédicace

À mes chers parents, qui ont tout sacrifié pour ma réussite, que ce soit par leur travail acharné ou leurs prières incessantes.

À mes frères Salleh, Youcef, Athmane et Saadi, mes piliers de courage et de force, grâce à qui j'ai surpassé de nombreux défis et accompli tant de choses grâce à leur soutien constant.

À mes sœurs Randa et Lynda, ainsi qu'à mes belles-sœurs Houda et Sara, qui ont toujours été à mes côtés, m'aidant et m'encourageant à croire en moi-même.

Pour mes chers petits Djihane, Raouane, Aline, Rinade, Lina, Ghaithe,Iyad que j'aime énormément.

À mes amies pour leur compagnie précieuse et leur soutien inébranlable.

Pour terminer, une dédicace spéciale à ma binôme « Lyza ».

Nesrine.

Remerciements

Ce mémoire n'aurait pas pu voir le jour sans le soutien précieux et l'aide généreuse de nombreuses personnes.

Nous tenons à exprimer notre profonde gratitude à notre encadrant, M. Achroufene, pour son soutien indéfectible, ses précieux conseils et son engagement tout au long de ce projet. Votre expertise, votre disponibilité et votre patience ont largement contribué à la réussite de ce mémoire..

Nous remercions également notre co-encadrant, M. Sabri, pour son aide et ses conseils.

Nous remercions notre jury pour la lecture et l'évaluation sincère et constructive de ce modeste travail.

Nous souhaitons également remercier nos professeurs et collègues pour leurs précieux conseils et leurs encouragements tout au long de ce parcours.

Un grand merci à nos amis et à nos familles pour leur soutien moral et leur compréhension durant cette période intense. Vos encouragements et votre présence nous ont été d'un grand réconfort et ont rendu ce voyage beaucoup plus supportable.

Enfin, nous remercions tous ceux qui, de près ou de loin, ont contribué à la réalisation de ce mémoire.

Merci à tous.

Lyza et Nesrine.

Table des matières

Table des figures

Liste des tableaux

Liste des abréviations

AE	Autoencoder Neural Networks
Bi-LSTM	Bidirectionnel Long Short-Term Memory
BSN	Sensor Networks
BSTMs	Binary Space-Time Maps
CNN	Convolutional Neural Networks
Conv-LSTM	Convolution Long Short-Term Memory
DL	Deep Learning
DMI	Dynamic Motion Image
GAP	Global Average Pooling
GRU	Gated Recurrent Unit
GAN	Generative Adversarial Networks
HAR	Human Activity Recognition
IA	Intelligence Artificielle
ML	Machine Learning
MLFF	Multi-Level Fusion Framework
PAMAP2	Physical Activity Monitoring and Assessment Dataset
RNN	Recurrent Neural Networks
SVM	Support Vector Machines
WISDM	Wireless Sensor Data Mining

Introduction générale

La reconnaissance des activités humaines (HAR) se définit comme l'ensemble des techniques et méthodes visant à identifier et à classer les actions réalisées par une personne à partir de données de capteurs ou de vision. Ce domaine trouve des applications dans divers secteurs tels que la santé, le sport, la surveillance et les interactions homme-machine, en jouant un rôle essentiel dans l'amélioration de la qualité de vie et la sécurité.

La reconnaissance de l'activité humaine est un domaine de recherche important qui utilise diverses techniques d'apprentissage automatique pour identifier et classer les mouvements humains. Bien que ces méthodes traditionnelles se soient révélées utiles, elles présentent des limites en termes de précision et de capacité à généraliser à différents situations. Dans ce contexte, des techniques d'apprentissage profond ont émergé qui offrent de nouvelles perspectives pour relever ces défis.

Les techniques de deep learning, en particulier les réseaux de neurones profonds, ont montré un potentiel significatif pour améliorer les performances des systèmes HAR. En exploitant des architectures complexes comme les réseaux de neurones convolutifs (CNN) et les réseaux de neurones récurrents (RNN), ces modèles peuvent extraire et apprendre des caractéristiques riches et variées des données. Cependant, malgré leurs avantages, les approches basées sur le deep learning rencontrent également des défis, notamment dans la détection précise de certaines activités et la réduction des confusions entre différentes actions similaires.

Pour surmonter ces limitations, l'intégration de modèles hybrides s'est avérée une solution prometteuse. C'est dans cette direction que s'insère le présent travail, qui a pour but la mise en œuvre d'un système combinant les points forts des réseaux CNN et des GRU (Gated Recurrent Units). En exploitant cette combinaison, il est possible d'améliorer la précision et la robustesse des systèmes HAR. Les réseaux CNN sont efficaces pour extraire des caractéristiques spatiales des données, tandis que les GRU sont particulièrement performants pour capturer les dépendances temporelles. Cette synergie permet de mieux reconnaître les activités humaines et

de réduire les erreurs de classification.

À cette fin, ce document est organisé en quatre chapitres :

- Le premier a pour object d'introduire quelques aspects généraux et quelques définitions sur la reconnaissance de l'activité humaine et l'apprentissage profond.
- Le deuxième chapitre est un état de l'art des travaux récents portant sur la HAR basée sur les méthodes de deep learning.
- Ensuite, le chapitre trois vise à expliquer le processus HAR en utilisant le réseau neuronal profond du système proposé.
- Le dernier chapitre présente les détails de mise en œuvre et les résultats des différents tests effectués.

Ce présent mémoire se termine par une conclusion qui résume le processus suivi pour réalisé le système HAR proposé et les résultats obtenus lors de sa validation.

Chapitre 1

Généralités sur les activités humaines

1.1 Introduction

L'activité humaine est un élément essentiel de notre existence et façonne le monde dans lequel nous vivons. Depuis les débuts de l'humanité, nous avons été constamment engagés dans une multitude d'activités qui ont évolué et se sont transformées au fil du temps. Ces activités sont le reflet de notre capacité à penser, à créer, à innover et à interagir avec notre environnement.

Dans ce chapitre, nous plongerons dans les fondements de la reconnaissance d'activité humaine. Nous explorerons comment ses avancées technologiques ont permis de saisir, d'interpréter et de comprendre les mouvements, gestes et actions physiques de l'individu. De plus, nous introduirons quelques techniques de machine learning utilisées en reconnaissance d'activité et les mesures de performance qui servent à les évaluer. En examinant ses concepts, nous enrichirons notre compréhension des applications et des ramifications de la reconnaissance d'activité humaine dans divers aspects de notre quotidien et de la recherche contemporaine.

1.2 Définition d'activité

L'activité humaine fait référence aux mouvements, gestes ou actions physiques effectués impliquant une sortie d'énergie. Il existe deux catégories d'activités humaines : simples et complexes. Selon [34] les activités humaines simples considèrent la posture du corps et le mouvement pour définir les différentes activités (marcher, courir, ...) et les activités humaines complexes consistent en des activités simples accompagnées d'une fonction spécifique (manger par exemple).

Thomas et al. [35] ont proposé une catégorisation selon la complexité des mouvements réalisés :

- Le geste : c'est le mouvement élémentaire engendré par le déplacement d'un membre humain tel que, lever la main, tourner la tête, etc.
- L'action : c'est un mouvement complexe composé de plusieurs mouvements élémentaires (plusieurs gestes), par exemple courir, tirer un ballon, etc.
- L'activité : c'est un mouvement de plus en plus complexe et composé de plusieurs actions, par exemple joué au football, boire du thé, etc.

Dans la littérature [32], on distingue sept groupes d'activités. Ces groupes et les activités individuelles qui leur sont associées sont résumés dans la Table 2.3.

Groupe	Activités
Ambulation	Marcher, courir, s'asseoir, être debout, être allongé, monter et descendre un escalier, un escalier mécanique ou un ascenseur
Transportation	Monter un bus, faire du vélo et conduire
Utilisation du téléphone	Envoyer un message, passer un appel
Activités quotidiennes	Manger, boire, travailler sur PC, regarder la TV, lire, se brosser les dents, s'étirer, frotter, passer l'aspirateur
Exercice physique	Faire de l'aviron, soulever des poids, tourner, marche nordique, faire des pompes.
Militaire	Ramper, ouvrir une porte. évaluer la situation
Haut du corps	Mâcher, parler, avaler, bouger la tête

TABLE 1.1 – Types d'activités reconnues dans la littérature [32]

L'un des objectifs de l'étude des activités humaines est leur reconnaissance. Le reste de ce chapitre est consacré à la présentation de la reconnaissance d'activités.

1.3 Reconnaissance d'activité humaine

La reconnaissance d'activité humaine (En anglais : Human Activity Recognition (HAR)) est devenue un sujet très populaire dans le domaine de la vision par ordinateur et du traitement du signal. Un grand nombre d'applications de reconnaissance des actions humaines à partir des vidéos peuvent être trouvées : la vidéo-surveillance, l'interaction homme-machine et l'indexation des vidéos. Le but d'un système de reconnaissance d'activité humaine est d'identifier les actions simples de la vie quotidienne (comme marcher, courir, sauter ...). Chacune de ces actions, réalisées par une seule personne dans un laps de temps précis, doit être représentée par un modèle de mouvement simple.

Au cours de ces dernières années, de nombreuses méthodes ont été proposées pour la reconnaissance et la compréhension des actions humaines [36]. Cependant, le domaine de la reconnaissance d'activités humaines est particulièrement difficile vue le nombre de contraintes à surmonter dans les deux domaines [37] suivants :

Domaine de la vision, on peut mentionner :
- le changement du point de vue de la caméra,
- le changement du fond,
- la variation de la luminosité,
- l'énorme quantité de données vidéo, etc.

Domaine du signal, on peut mentionner :

- la présence du bruit,
- la perte du signal,
- la variations des activités humaines, etc.

Les principaux objectifs de la Human Activity Recognition (HAR) mentionnés dans [38] sont :

- Décrire, analyser, reconnaître, comprendre et suivre les activités et les mouvements de personnes.
- Elle peut être utile pour détecter tôt les comportements anormaux de certaines personnes : difficultés dues à l'âge ou à une maladie.
- Aider les humains dans leurs tâches après avoir fourni des informations.

Les avancées dans les domaines du deep learning ont considérablement amplifié les capacités des systèmes de reconnaissance d'activité humaine, ouvrant ainsi la voie à une multitude d'applications dans divers secteurs. Nous abordons ça dans la section à venir.

1.4 Domaines d'applications

Avec les avancées rapides dans les technologies de l'informatique et de la miniaturisation, les systèmes de reconnaissance d'activité humaine sont devenus une partie importante de notre quotidien et sont largement utilisés dans de nombreuses applications telles que la gestion de la santé, la surveillance médicale, l'interaction homme-machine, la robotique, la surveillance, les sciences du sport, la réadaptation, le contrôle à distance [39]. On va détailler quelques secteurs tels que :

• **Secteur de la santé :** La HAR basée sur le DL offre une solution de pointe pour les professionnels de la santé afin de fournir des interventions précises et plus proactives, réduisant la charge sur les systèmes de santé et améliorant le bien-être des patients tout en augmentant la qualité globale des soins. Dans ce secteur, plusieurs systèmes similaires ont été étudiés, en prenant l'étude de [40] qui présente une approche améliorée de l'algorithme d'optimisation du coyote avec une HAR assistée par l'apprentissage profond (ICOADL-HAR) pour la surveillance de santé. L'objectif de la technique ICOADL-HAR est d'analyser les informations des capteurs des patients pour déterminer les différents types d'activités.

• **Secteur du sport :** Appliquer la HAR dans domaine du sport offre de nombreuses opportunités pour aider les athlètes, les entraîneurs et les équipes à at-

teindre leurs objectifs en permettant une analyse et un suivi plus efficaces et détaillés des performances, par exemple les chercheurs de cet article [41] explorent les solutions basées sur la vision par ordinateur pour reconnaître les actions des joueurs pouvant être appliquées dans des scènes de handball non contraintes, sans capteurs supplémentaires et avec des exigences modestes, permettant une adoption plus large des applications de vision par ordinateur.

• **Secteur militaire :** Il est devenu une tendance de développer des dispositifs portables avec des fonctions diverses pour l'armée, par exemple dans cet article [42] les auteurs proposent un cadre de fusion multi-niveaux Multi-Level Fusion Framework (MLFF) basé sur les Réseaux de Capteurs Corporels Sensor Networks (BSN) des soldats, son objectif principal n'est pas uniquement de reconnaître l'activité humaine au sens traditionnel, comme pour des applications de santé ou de fitness. Au lieu de cela, il est spécifiquement conçu pour soutenir les opérations militaires en fournissant des informations multidimensionnelles qui incluent non seulement le mouvement mais aussi d'autres facteurs physiologiques et environnementaux pour aider les soldats à prendre des décisions éclairées en cas de danger ou de situation critique sur le terrain.

1.5 Approches de reconnaissance d'activités humaines

Pour saisir les activités, plusieurs recherches en HAR ([43],[44]) sont classées les approches en deux catégories principales : la recherche basée sur la vision et la recherche basée sur les capteurs comme le montre la figure 1.1.

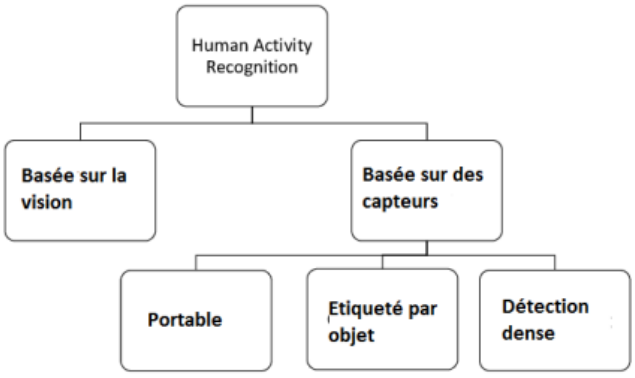


FIGURE 1.1 – Classification des approches de reconnaissance de l'activité humaine

1.5.1 Vision par ordinateur

Cette méthode se présente sous forme des caméras statiques installées à divers endroits à des fins de surveillance enregistrent des vidéos et les stockent sur des serveurs. Ces flux de caméra ou vidéos enregistrés sont ensuite utilisés à des fins de surveillance. Ce type de reconnaissance d'activité humaine est utilisé pour la sécurité routière, la sécurité publique, la gestion du trafic, la surveillance de foule, etc. La Figure 1.2 montre les étapes typiques d'un système de reconnaissance d'activités humaines basé sur la vision [45].

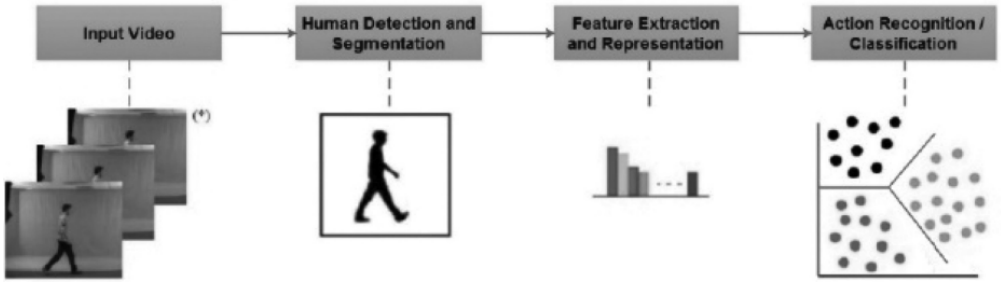


FIGURE 1.2 – Un système typique de reconnaissance d'activité humaine

1.5.2 Utilisation de capteurs

Selon [33], la reconnaissance de l'activité humaine basée sur l'utilisation des signaux provenant de différents types de capteurs est un problème de classification de série chronologique. La figure 1.3 donne une illustration du fonctionnement de ce genre de reconnaissance d'activités.

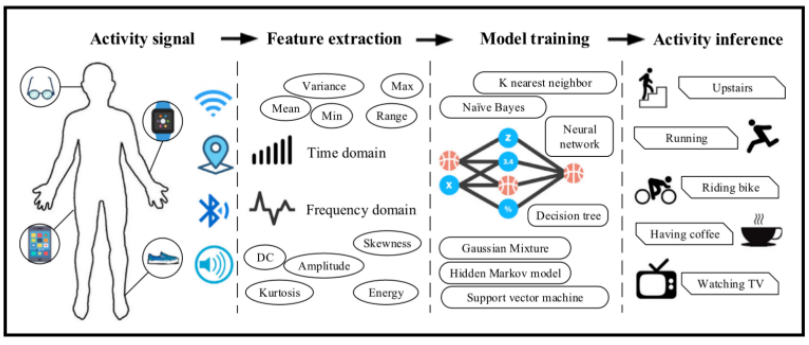


FIGURE 1.3 – Illustration du processus de la reconnaissance de l'activité humaine basée capteurs

Ces capteurs peuvent être répartis en trois grandes classes :

Capteurs portés : L'utilisateur doit porter les capteurs sur lui lorsqu'il effectue une activité. De nombreuses études ont été faites sur la reconnaissance des activités en utilisant des capteurs portables, mais le problème majeur avec ce type d'approche est qu'il n'est parfois pas possible de porter un capteur sur soi. Par exemple, dans le cas des personnes âgées ou des patients, ils peuvent oublier de les porter où ils résistent à porter les étiquettes du tout [46].

Capteurs Objet : Selon [46] les solutions qui utilisent une approche par objet, des capteurs sont attachées à des objets d'usage quotidien. Sur la base de l'interaction d'un utilisateur avec ces objets, différentes activités sont reconnues. Il s'agit d'une approche liée au périphérique, c'est-à-dire que les utilisateurs doivent utiliser des objets spécifiques (objets marqués) uniquement. Comme l'approche portable, cette approche peut également ne pas être réalisable tout le temps car elle limite les utilisateurs à utiliser seulement des objets marqués

Capteurs Ambiants : Au cours des dernières années, les chercheurs se sont concentrés sur une approche sans appareil dans laquelle les utilisateurs ne transportent pas d'appareil avec eux. L'idée est de déployer capteurs dans l'environnement (l'installation dans laquelle l'activité est en cours) et lorsqu'une personne effectue une activité, les données seront capturées par ces capteurs, qui peuvent ensuite être utilisé pour la reconnaissance d'activité. L'approche sans appareil est plus pratique car elle ne nécessite pas que l'utilisateur transporter avec lui un appareil spécifique lorsqu'il effectue ses activités [46]. Le principal inconvénient de cette approche c'est seules certaines activités peuvent être reconnue avec certitude [33].

Les différents types de capteurs utilisés sont récapitulés dans la Table 1.2 en spécifiant leurs avantages et leurs inconvénients.

Type de capteur	Type d'événement	Avantages	Inconvénients
portés	Mouvements du corps	déploiement facile	
Objets	Mouvement des objets	Informations détaillées	déploiement difficile
Ambiants	changements dans l'environnement		Déploiement difficile, Facilement affecté par l'environnement, Seules certaines activités peuvent être reconnues avec certitude

TABLE 1.2 – récapitulatif sur les différents types de capteurs[33].

En tenant compte de la portabilité et du confort du port des capteurs, certaines études ont utilisé un seul capteur pour la reconnaissance des actions [47],[48]. Cependant, un seul capteur ne peut acquérir que des informations de mouvement locales et présente une faible précision de reconnaissance pour des mouvements complexes. Ces dernières années, les chercheurs ont essayé différentes stratégies de fusion pour combiner plusieurs capteurs afin de tirer pleinement partie des données, des caractéristiques et des classificateurs pour une surveillance efficace des activités [49]. La figure 1.4 montre le processus de fusion d'informations sur différents niveaux.

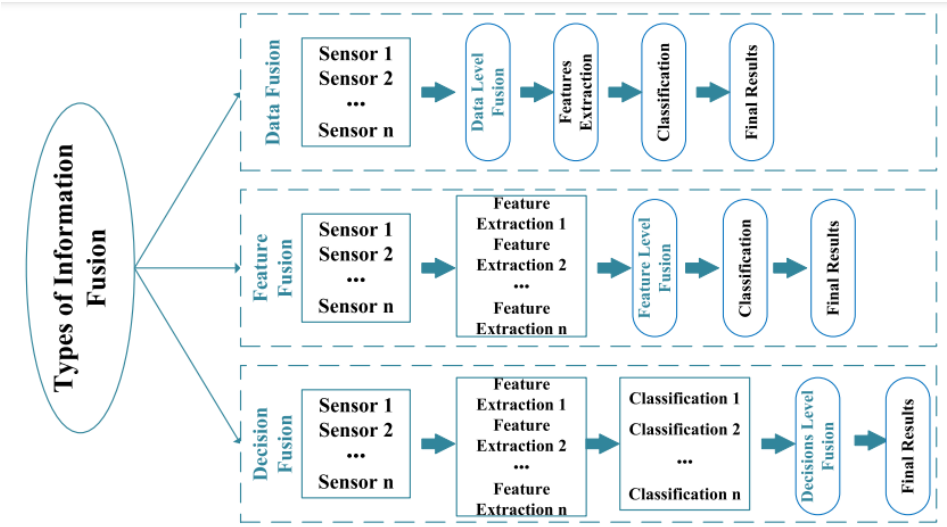


FIGURE 1.4 – Classification de la Fusion multi-sensor[1]

Les niveaux de fusion d'informations couramment adoptés sont :

Fusion au niveau des données : La fusion de données implique l'intégration de données collectées par plusieurs appareils mobiles et capteurs portables pour améliorer la fiabilité, la robustesse et la généralisation du système de reconnaissance. Lorsque le système implique plusieurs capteurs homogènes mesurant le même phénomène physique, les données des capteurs peuvent être fusionnés directement.

Fusion au niveau des caractéristiques : Cette méthode implique la combinaison des ensembles de caractéristiques extraits de multiples sources de données, qui peuvent provenir de différents capteurs ou d'un même dispositif équipé de plusieurs capteurs physiques. L'objectif est de créer un vecteur de caractéristiques de haute dimension qui résume l'information provenant de toutes ces sources, optimisant ainsi l'entrée pour l'étape de reconnaissance de motifs ou de classification.

Fusion au niveau des décisions : La sortie de fusion au niveau de la décision est une décision unique obtenue à partir des décisions locales de plusieurs capteurs homogènes ou hétérogènes, et la décision de haut niveau est réalisée en utilisant les informations extraites jusqu'à un certain niveau à travers le traitement préliminaire des données des capteurs ou au niveau des caractéristique.

1.6 Processus de reconnaissance d'activité

L'objectif principal de l'étude de la reconnaissance des activités humaines est de créer un système intelligent capable de reconnaître l'activité (comme cuisiner) [50]. Ce système fonctionne selon le processus de reconnaissance d'activité présenté par la figure 1.5.

Nous décrivons à présent en détail les différentes étapes de ce processus :

- **Définition des activités cibles (Target Activities) :** Définition et analyse des caractéristiques réelles des activités cibles à reconnaître. Cela peut inclure, par exemple, leur durée, leur distribution, leur similarité avec d'autres activités, etc.
- **Configuration de l'appareil (Device Setup) :** Identification et étude des exigences et détermination des appareils à utiliser dans la phase de collecte de données, basée sur les activités humaines cibles.
- **Collecte de données (Data collection) :** Durant cette phase, les données sont collectées à partir de capteurs, de dispositifs portables ou d'autres dispositifs qui capturent des informations sur les mouvements et les actions de la personne.
- **Annotation des donnéess (Data Annotation) :** Processus d'attribution de labels aux activités humaines en cours d'exécution.En associant les

données d'entrée avec les labels correspondants, le modèle peut apprendre à faire des prédictions précises et généraliser ses connaissances à des exemples non vus.

- **Prétraitement des données (Data Preprocessing) :** Les données collectées sont ensuite prétraitées pour éliminer le bruit, les informations non pertinentes sont filtrées, et les données sont préparées pour l'analyse. Dans le cadre de cela, les analyses suivantes sont réalisées :
  - **Extraction de caractéristiques :** Les données prétraitées sont analysées pour extraire des caractéristiques pertinentes qui peuvent être utilisées pour classer les activités humaines. Ces caractéristiques peuvent inclure des modèles de mouvement, la position du corps, ou d'autres caractéristiques.
  - **Sélection de caractéristiques :** Une fois les caractéristiques extraites, un sous-ensemble de caractéristiques peut être sélectionné pour être utilisé dans le modèle de classification. Cela aide à réduire la dimensionnalité (par exemple, le nombre de caractéristiques) des données et à améliorer la précision du modèle.
- **Génération et test du modèle(Model generation and testing) :** Un modèle HAR (Machine Learning (ML) ou Deep Learning (DL)) est développé pour classer les activités humaines basées sur les caractéristiques sélectionnées lors de cette phase. Le modèle peut être entraîné en utilisant un jeu de données étiqueté ou des techniques d'apprentissage non supervisé. Après que le modèle a été généré, les étapes suivantes sont réalisées avant que le modèle soit prêt à être utilisé :
  - **Évaluation du modèle :** Le modèle développé est ensuite évalué en utilisant un jeu de données de test pour évaluer son exactitude et sa performance. Cette phase aide à identifier tout problème ou domaine à améliorer dans le modèle.
  - **Déploiement :** Enfin, le modèle développé est déployé dans un environnement réel, où il est utilisé pour classifier les activités humaines.

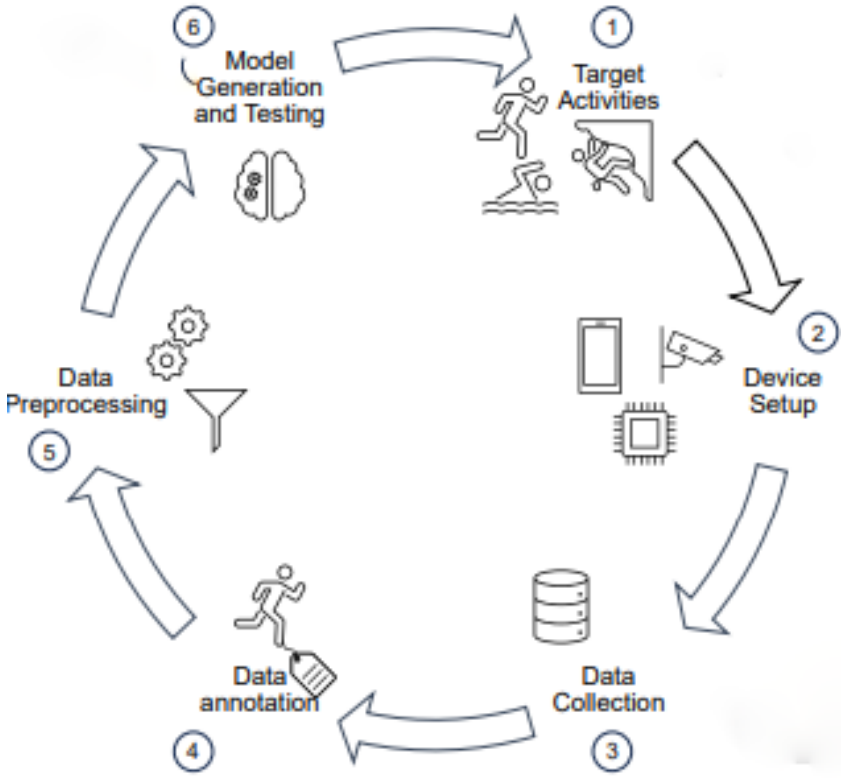


FIGURE 1.5 – Processus de reconnaissance d'activités [2]

Pour mettre en oeuvre des systèmes HAR, différentes techniques de l'Intelligence Artificielle (IA), en particulier machine learning et deep learning, ont été utilisées. Dans la section suivante, nous abordons leurs concepts.

1.7 Quelques techniques de machine learning et deep learning

D'après Steeven JANNY [51], l'Intelligence artificiel (IA) peut être décrite comme un domaine de l'informatique et des mathématiques rassemblant un ensemble de techniques algorithmiques et de théories permettant de réaliser des machines imitant l'intelligence humaine. Son but est de reproduire l'intelligence afin d'être capable de résoudre des problèmes complexes.

1.7.1 Apprentissage automatique

Conformément à la description fournie dans [52], ML peut être défini comme étant une technologie d'intelligence artificielle permettant aux machines d'apprendre sans avoir été au préalablement programmées spécifiquement à cet effet.

Généralement, les techniques d'apprentissage automatique sont réparties en trois grand types [3] comme le montre la figure 1.6.

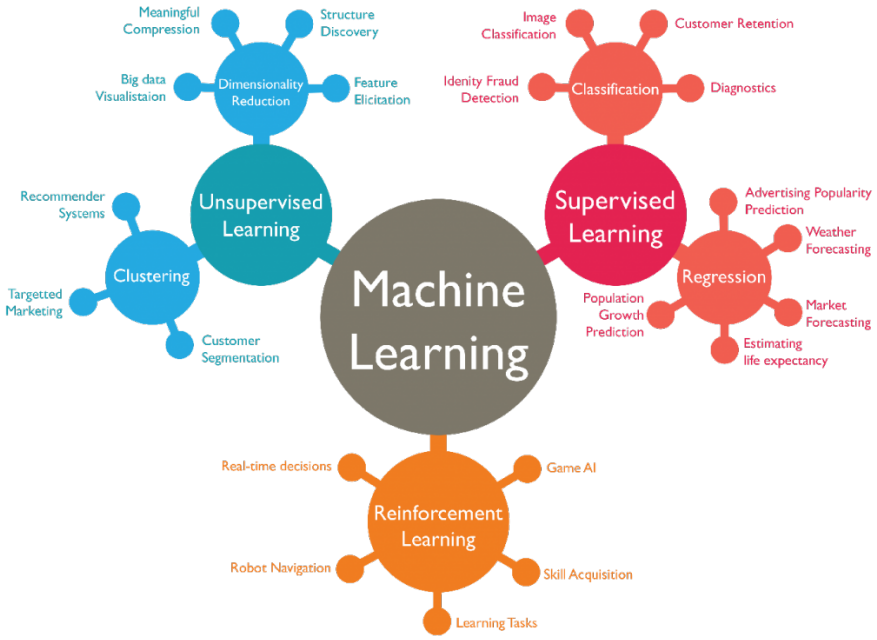


FIGURE 1.6 – Représentation des types de ML [3].

1.7.1.1 Apprentissage supervisé

Pour Velibor Božić [53], l'apprentissage supervisé est le type le plus courant d'apprentissage automatique, où l'algorithme est entraîné sur des données étiquetées. Les algorithmes les plus utilisés en apprentissage supervisé comprennent la régression linéaire, la régression logistique, les arbres de décision et les machines à vecteurs de support.

1.7.1.2 Apprentissage non-supervisé

En accord avec les informations disponibles sur [54], l'apprentissage non supervisé concerne les données non étiquetées, où l'objectif est de découvrir des motifs

ou des structures cachées au sein des données. Des algorithmes tels que le regroupement K-means, le regroupement hiérarchique et l'analyse en composantes principales (ACP) sont largement utilisés dans cette catégorie.

1.7.1.3 Apprentissage par renforcement

En suivant les définitions proposées sur [55], on peut comprendre l'apprentissage par renforcement comme une technique d'apprentissage automatique qui entraîne les logiciels à prendre des décisions en vue d'obtenir les meilleurs résultats. Elle imite le processus d'apprentissage par tâtonnements employé par les êtres humains pour atteindre leurs objectifs.

1.7.1.4 Apprentissage semi-supervisé

Un quatrième type d'apprentissage peut être obtenu en combinant les apprentissages supervisé et non-supervisé. Selon les explications fournies sur [56], l'apprentissage semi-supervisé est une approche en apprentissage automatique qui utilise une combinaison de données étiquetées et non étiquetées pour entraîner un modèle.

Nous allons nous intéresser aux techniques de deep learning qui sont largement adoptées dans les systèmes HAR et qui seront également utilisées dans le présent travail.

1.7.2 Deep learning

D'après la définition [57], le DL peut être défini comme un sous-domaine de l'apprentissage automatique caractérisé par l'utilisation de réseaux de neurones artificielles pour traiter et apprendre des données (voir la figure 1.7). Il se distingue par sa capacité à extraire automatiquement les caractéristiques hiérarchiques des données brutes, ce qui lui permet de s'attaquer à des problèmes complexes auparavant considérées comme insurmontables par les algorithmes traditionnels (Machines à vecteurs de support (SVM),k-Plus proches voisins (k-NN), Arbres de décision et Régression linéaire...).

À l'ère du deep learning, une vaste gamme de méthodes et d'architectures a été développée.[58]

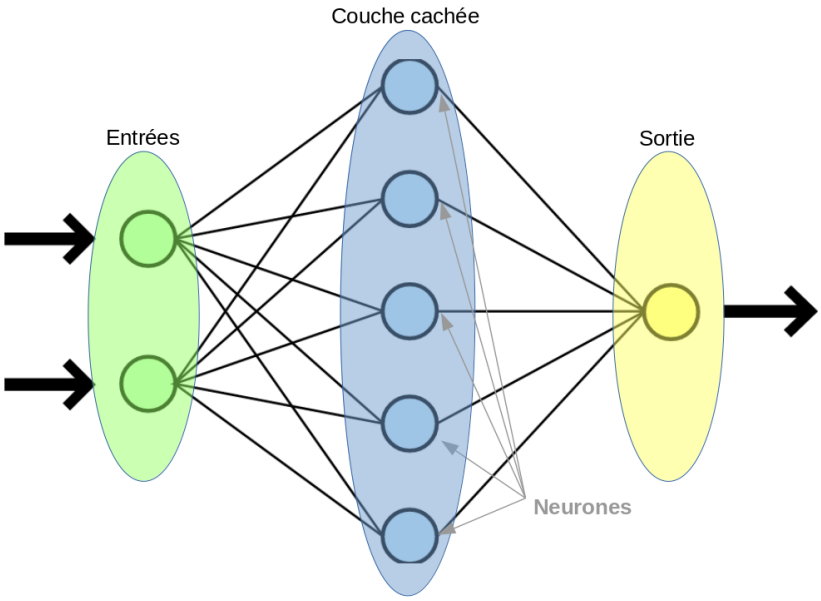


FIGURE 1.7 – Architecture de DL [4].

1.7.2.1 Réseaux de neurone convolutifs

Selon [59], on peut comprendre les Convolutional Neural Networks (CNN) comme étant un sous-ensemble des modèles de réseaux de neurones artificiels basés sur l'apprentissage profond, principalement développés pour l'extraction et l'évaluation de données visuelles, y compris les images et les vidéos. L'architecture de CNN est illustrée dans la figure (1.8).

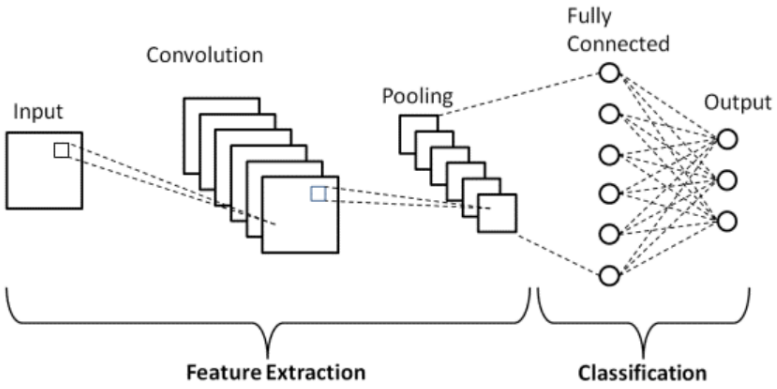


FIGURE 1.8 – Architecture de CNN [5]

Caractéristiques Clés des Réseaux de Neurones Convolutionnels :

— **Couche Convolutionnelle :**
Le cœur qui construit un CNN est la couche convolutionnelle. Les couches convolutionnelles extraient des caractéristiques des images en appliquant un ensemble de filtres ou de noyaux à l'image d'entrée. Ces filtres glissent sur l'image, effectuant des produits scalaires entre leurs poids et les pixels correspondants de l'image d'entrée.
— **Couche de Pooling :**
Les couches de pooling sont connues pour réduire les dimensions spatiales des cartes de caractéristiques, rendant le réseau plus efficace sur le plan computationnel et réduisant le risque de surapprentissage. Les opérations de pooling courantes incluent le max pooling, l'average pooling et le pooling L2-norm.
— **Couche Complètement Connectée :**
Les couches complètement connectées sont typiquement utilisées dans les étapes finales d'un CNN pour classifier les images ou prédire les emplacements des objets. Ces couches prennent la sortie aplatie des couches convolutionnelles et de pooling et connectent chaque neurone de la couche précédente à chaque neurone de la couche actuelle [60].

1.7.2.2 Réseaux de neurone récurrents

D'après [61], les Recurrent Neural Networks (RNN) peut être définie comme l'une des approches d'apprentissage profond séquentielles. Ont des connexions récurrentes entre les unités cachées de leur structure, ce qui permet de relier les informations passées aux informations actuelles.

Différents types de modèles RNN, tels que les LSTM (Long Short-Term Memory), GRU (Gated Recurrent Unit), ont été développés pour répondre à des défis spécifiques dans diverses applications [58] (voir la figure 1.9).

Long Short-Term Memory (LSTM) : Selon [62], la mémoire à long terme (LSTM) est un type de réseau RNN. La sortie de l'étape précédente est utilisée comme entrée dans l'étape actuelle dans les RNN. Elle peut traiter des séquences en utilisant sa "mémoire d'états internes" qui est supérieure. Dans LSTM, le niveau intermédiaire est une porte cachée. La porte d'oubli est utilisée pour déterminer les données qui doivent être sauvegardées et celles qui doivent être oubliées pour un apprentissage productif à long terme. La couche intermédiaire prend les données de la couche d'entrée et l'effet est affiché par la couche de sortie.

Gated Recurrent Unit (GRU) : Selon [63], il s'agit d'un algorithme basé sur RNN qui est comparable aux LSTM mais qui possède une architecture plus simple.

Le problème fondamental des RNN est la question des gradients qui disparaissent et explosent, ce qui se produit en raison des multiplications continues lors de la rétropropagation temporelle. La GRU résout ce problème en utilisant des portes, à savoir la porte de mise à jour et la porte de réinitialisation.

La première étape dans le développement d'un modèle GRU consiste à calculer la porte de mise à jour (Z_t) en utilisant la formule de l'équation (1.1), qui définit la quantité d'informations passées devant être conservées.

$$z_t = \sigma(w^{(z)}x_t + u^{(z)}h_{t-1} + b) \tag{1.1}$$

où w et u sont des poids, x_t est l'entrée, h_{t-1} est l'état caché, et b est le biais.
La porte de réinitialisation (r_t) est ensuite calculée en utilisant l'équation (1.2), qui définit combien d'informations antérieures doivent être supprimées et comment combiner l'entrée entrante avec les anciennes informations. La formule de la porte de réinitialisation est la suivante :

$$r_t = \sigma(w^{(r)}x_t + u^{(r)}h_{t-1} + b) \tag{1.2}$$

Ensuite, on calcule l'état caché candidat (h'_t) qui sera utilisé par la porte de réinitialisation pour conserver les informations importantes du passé. L'équation (1.3) exprime l'état caché candidat :

$$h'_t = \tanh(wx_t + r_t \odot uh_{t-1}) \tag{1.3}$$

où $\odot$ est le produit de Hadamard.
La dernière étape consiste à calculer l'état caché (h_t) en utilisant l'équation (1.4). Cet état caché sert de sortie (y_t) :

$$h_t = z_t \odot h_{t-1} + (1 - z_t) \odot h'_t \tag{1.4}$$

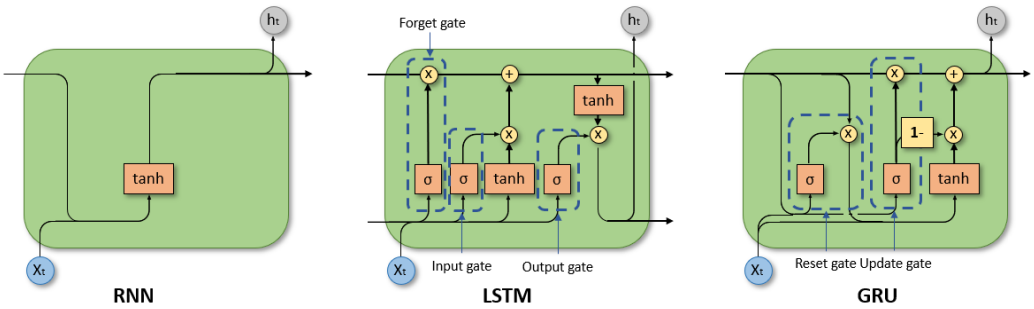


FIGURE 1.9 – Architecture de RNN, LSTM, GRU [6]

1.7.2.3 Réseaux de neurone auto-encodeurs

En se référant aux définitions fournies sur [64], Les Autoencoder Neural Networks (AE) peut être définir comme un type de réseau neuronal qui excelle dans la réduction de dimensionnalité. Ils se composent de deux parties : un encodeur et un décodeur. L'encodeur compresse les données d'entrée dans une représentation de dimension inférieure à l'aide de techniques non linéaires. Le décodeur tente ensuite de reconstruire les données d'entrée originales à partir de cette représentation compressée, minimisant ainsi les erreurs de reconstruction (voir la figure(1.10)).

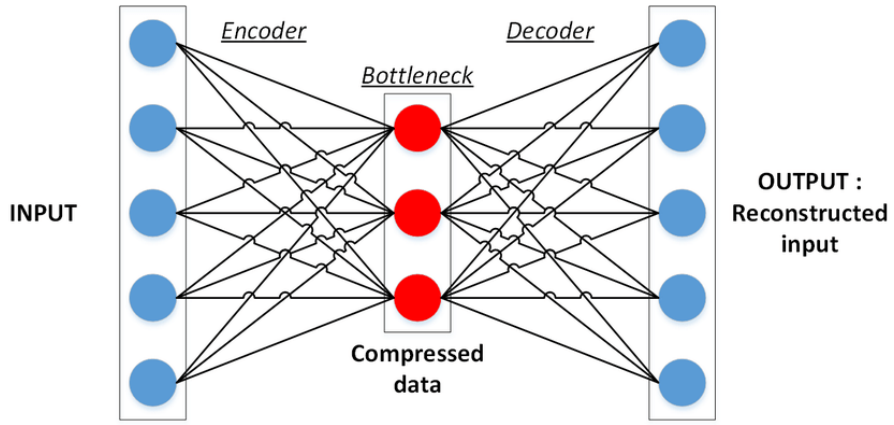


FIGURE 1.10 – Architacture de auto-encodeurs [7]

1.7.2.4 Réseaux de neurone générateurs adverses

En suivant les descriptions données sur [65], Un Generative Adversarial Networks (GAN) est composé de deux réseaux de neurones distincts, le générateur et le discriminateur. Le générateur crée de nouvelles données, tandis que le discriminateur évalue la qualité de ces données. Les deux réseaux s'entraînent en boucle, améliorant ainsi leurs performances respectives (voir la figure(1.11)).

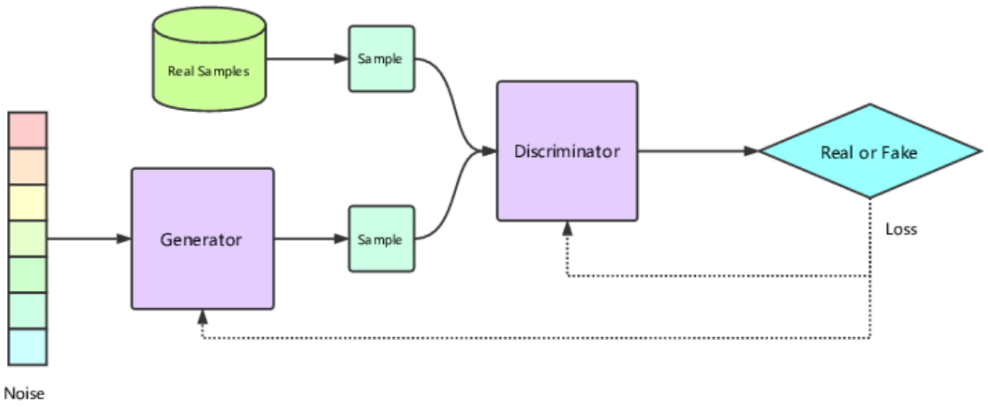


FIGURE 1.11 – Architacture de GAN [8]

1.7.2.5 Réseaux de neurone transformateurs

En se basant sur les informations fournies sur [66], on peut comprendre le transformer comme étant un réseau de neurones profonds conçu pour ingérer des données d'apprentissage séquentielles en utilisant des mécanismes d'attention (voir la figure(1.12)).

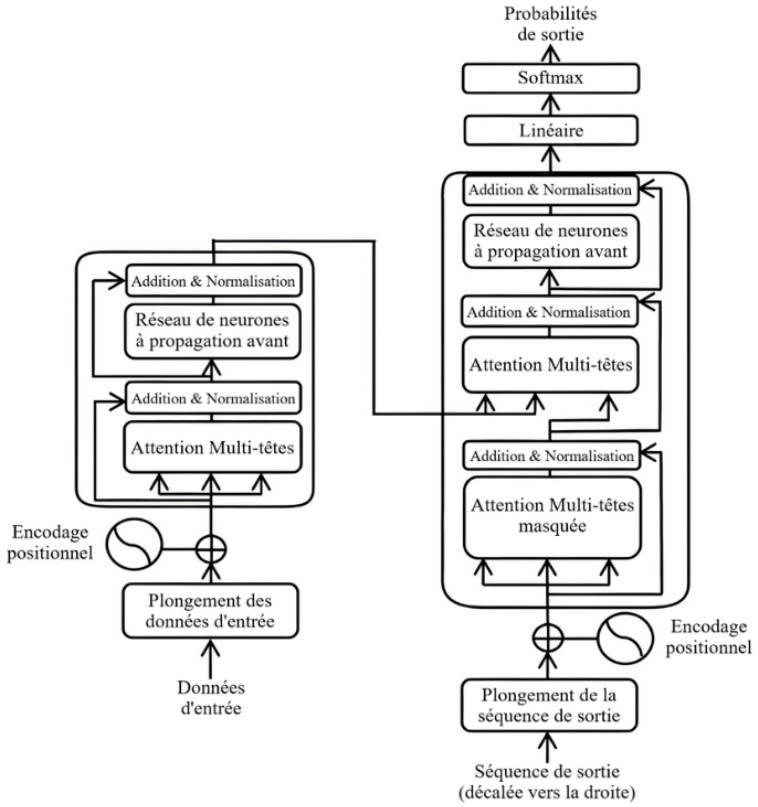


FIGURE 1.12 – Architacture de transformateurs [9]

Les avancées technologiques ont intégré les systèmes de reconnaissance d'activité humaine dans divers secteurs, cependant, pour garantir leur efficacité, ces systèmes doivent être évalués à l'aide de certains mesures.

1.8 Mesures de performances d'un système HAR

Les systèmes HAR sont essentiels pour divers applications, pour cela il faut bien assurer que le système fonctionne correctement et répond aux attentes en termes de précision et de fiabilité. Selon [10] pour évaluer la performance de chaque modèle basé sur les résultats expérimentaux, nous devons examiner leur performance à l'aide de quatre indicateurs d'évaluation : l'exactitude, la précision, le rappel et le score F1 :

— L'**exactitude** sert d'indice d'évaluation qui indique le plus intuitivement les performances du modèle. Elle est calculée en divisant le nombre de données

correctement prédites par le nombre total de données (équation (1.5)).

$$\text{Exactitude} = \frac{TP + TN}{2TP + FN + FP + TN} \tag{1.5}$$

avec :

- TP : le nombre de vraies acceptations (le modèle prédit correctement la classe positive) ;
- TN : le nombre de vrais rejets (le modèle prédit correctement la classe négative) ;
- FP : le nombre de fausses acceptations (le modèle prédit incorrectement la classe positive) ;
- FN : le nombre de faux rejets (le modèle prédit incorrectement la classe négative).

— La **précision** est la proportion d'éléments que le modèle classe comme vrais qui sont effectivement vrais. Elle peut être exprimée par l'équation (1.6).

$$\text{Précision} = \frac{TP}{TP + FP} \tag{1.6}$$

— Le **rappel** est un indicateur qui représente la proportion d'observations positives réelles parmi celles prédites par le modèle. Elle peut être exprimée par l'équation (1.7).

$$\text{Rappel} = \frac{TP}{TP + FN} \tag{1.7}$$

— Le **score F1**, exprimé par l'équation (1.8), sert d'indicateur représentant la moyenne harmonique de la précision et du rappel. Un score F1 faible implique que la précision et le rappel sont tous les deux faibles, indiquant un scénario dans lequel les deux métriques ont diminué. Un score de précision faible suggère que le modèle d'apprentissage automatique fait de nombreuses prédictions incorrectes, tandis qu'un rappel faible indique que le modèle échoue à capturer certains événements qui auraient dû être détectés.

$$\text{Score F1} = 2 \times \frac{\text{Précision} \times \text{Rappel}}{\text{Précision} + \text{Rappel}} \tag{1.8}$$

Le calcul de ces quatre indicateurs d'évaluation de performance implique l'utilisation d'une matrice de confusion montrée par la figure 1.13.

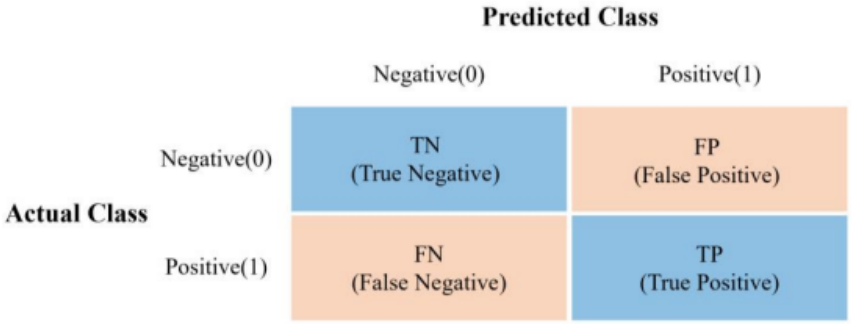


FIGURE 1.13 – Matrice de confusion [10]

1.9 Conclusion

Dans ce chapitre, nous avons défini les notions les plus importantes de l'activité humaine, nécessaires à la suite du projet. Une vue globale sur les différentes les techniques, systèmes utilisées pour la fusion a été détaillée, nous permettant ainsi d'entamer un état de l'art sur les travaux récemment réalisés dans le deuxième chapitre de ce travail.

Chapitre 2

État de l'art sur les systèmes de reconnaissance d'activités humaines

2.1 Introduction

Au fil des années, de nombreuses techniques ont été développées pour aborder le problème de la reconnaissance d'activités humaines. Durant cette dernière décennie, les méthodes HAR basées sur les techniques deep learning ont démontré des performances remarquables.

Dans ce chapitre, nous examinerons les travaux de chercheurs qui ont contribué à l'avancement de la HAR en utilisant des techniques de Deep Learning. Les travaux présentés sont catégorisés en trois classes : basés sur la vision par ordinateur, utilisant des signaux provenant de capteurs et hybrides. Nous terminons le chapitre par une comparaison des différents travaux selon un certain nombre de critères.

2.2 Approches HAR : Capteurs, Vision et hybrides

selon [67], La reconnaissance de l'activité humaine (HAR) peut être abordée à travers deux approches principales : basée sur les capteurs et basée sur la vision. Les capteurs portés sur le corps, tels que les accéléromètres et les unités de mesure inertielle (IMU), collectent des données de mouvement et identifient les activités allant des gestes simples aux mouvements complexes. Le HAR basé sur la vision, quant à lui, utilise des données visuelles provenant de vidéos ou d'images capturées par des caméras de surveillance, des smartphones ou d'autres appareils d'enregistrement.

De plus, pour capturer les aspects spatiaux et temporels de l'activité humaine, les chercheurs associent fréquemment les CNN aux réseaux neuronaux récurrents (RNN). L'utilisation de cette combinaison combine les compétences de reconnaissance spatiale des CNN avec la compréhension des séquences des RNN, ce qui permet d'avoir une vision globale des aspects spatiaux et temporels des manipulations humaines.

2.2.1 HAR basé sur la vision par ordinateur

Nous allons présenté quelques travaux récents de la littérature exploitant la vision par ordinateur en utilisant deep learning.

2.2.1.1 Travail de Khelalef et al. [11]

L'article de recherche de Khelalef et al. [11] met en avant une nouvelle méthode pour le système de reconnaissance d'activité humaine basé sur l'apprentissage pro-

fond à l'aide de Binary Space-Time Maps (BSTMs) pour l'extraction des caractéristiques qui offre la capacité de reconnaître plusieurs sujets dans la même image car ils calculent les BSTMs uniquement à partir des corps humains extraits et non à partir de l'ensemble des images comme les techniques proposées dans [35] [68]. Les étapes de cette approche sont détaillées dans le schéma explicatif de la Figure 2.1.

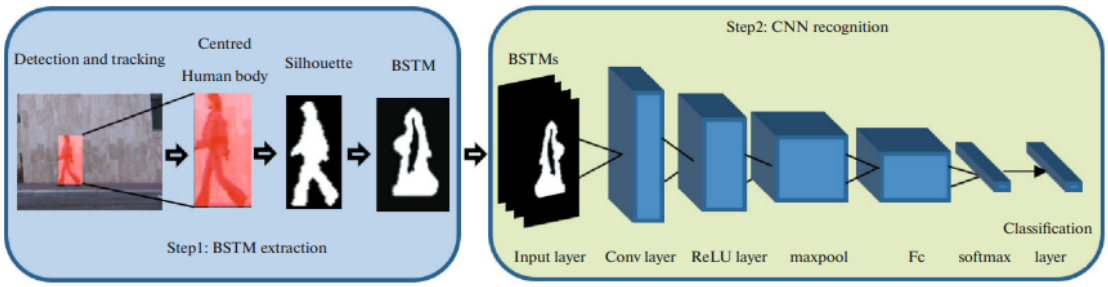


FIGURE 2.1 – Vue d'ensemble de la technique basée sur DL proposée [11]

Tous d'abord pour l'extraction efficace des caractéristiques la première étape consiste à détecter et suivre les corps humains dans les vidéos en utilisant un algorithme d'extraction de l'arrière-plan, après pour chaque image le corps humain est segmenté, ensuite une représentation visuelles des contours des personnes (silhouettes) sont extraites grâce à l'algorithme de segmentation d'image d'Otsu [69]. Ces silhouettes humaines segmentées de chaque image sont utilisées pour calculer les BSTMs qui capturent à la fois où se trouvent les personnes dans l'espace (représentation spatiale) et ce qu'elles font (actions) à différents moments et le changement d'état (représentation temporelle).

La deuxième étape ici utilise un CNN qui contient sept couches : une couche d'entrée a la même dimension que les BSTMs d'entrée, une couche de convolution, une couche ReLU, une couche de max-pooling, une couche complètement connectée, une couche softmax et une couche de classification qui renvoie la classe finale du BSTM construit en utilisant les résultats de la couche softmax.

Pour tester leur méthode les chercheurs ont utilisé trois bases de données différentes : Weizmann [70], Keck Gesture [71] et KTH [72] et ont atteint un taux de reconnaissance de 100% pour chaque une.

2.2.1.2 Travail de Karpathy et al. [12]

Les CNN sont imposés comme une classe de modèles puissants pour les problèmes de reconnaissance d'images. Karpathy et al. [12] sont encouragés par ces résultats à proposer un système de reconnaissance d'activités humaines pour la

classification vidéo à grande échelle en utilisant CNN. Leur but vise à améliorer la performance des CNN pour la classification vidéo en utilisant un ensemble de données massif de 1 million de vidéos YouTube. Pour cela, ils proposent une architecture multi-résolution avec l'utilisation de deux flux de traitement distincts (flux de contexte et un flux de fovéa) et chaque flux est un réseau de neurones à convolution qui traite les images de même résolution comme montre la Figure 2.2.

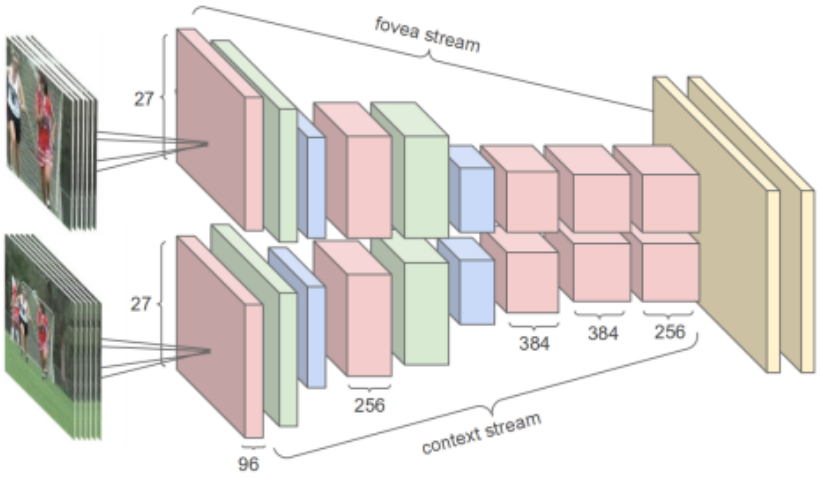


FIGURE 2.2 – Architecture de CNN à résolution multiple proposée en [12]

Afin de produire un modèle rapide, les auteurs proposent de travailler sur des images à basse résolution. Tout d'abord, ils utilisent en entrée un clip vidéo de taille 178 × 178 pixels, ensuite le flux de contexte reçoit des images échantillonnées à la moitié de la résolution spatiale d'origine (89 × 89 pixels), tandis que le flux de fovéa reçoit image centrée où se trouve le sujet d'intérêt de taille 89x89 à la résolution d'origine et les deux flux convergent vers deux couches entièrement connectées.

Les tests sont faits à l'aide de la base de donnée Sports-1M [12]. Après plusieurs expériences sur les réseaux CNN, ils choisissent le modèle Slow Fusion [12] comme représentant des réseaux conscients du mouvement car il donne les meilleurs résultats à 80%.

2.2.1.3 Travail de Simonyan et al. [13]

Dans [13], Simonyan et al. proposent une nouvelle architecture pour la reconnaissance d'actions dans les vidéos, elle est basée sur l'utilisation de réseaux Convolutifs (ConvNets). Un ConvNet de flux spatial pour capturer l'apparence spatiale Et transmettre des informations sur les objets dans les scènes à partir des

images individuelles de la séquence vidéo, ainsi qu'un ConvNet de flux temporel pour capturer le mouvement temporel du sujet à partir d'un ensemble d'images successives de la séquence vidéo, dont l'entrée est l'empilement du flux optique calculé entre plusieurs images consécutives de la séquence vidéo. La Figure 2.3 montre un exemple de calcul du flux optique et le système proposé est schématisé dans la Figure 2.4.

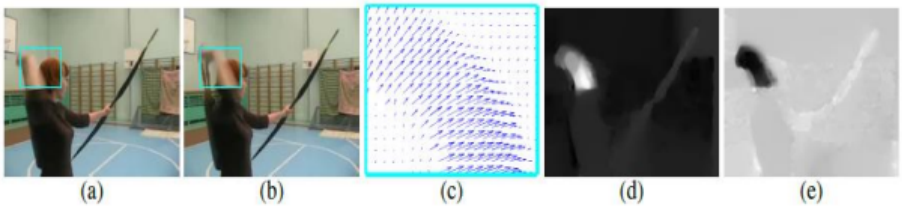


FIGURE 2.3 – Exemple du flux optique : (a et b) Deux images successives, (c) Le flux optique dans la zone en bleu, (d) La composante horizontale du flux optique et (e) La composante verticale du flux optique [13]

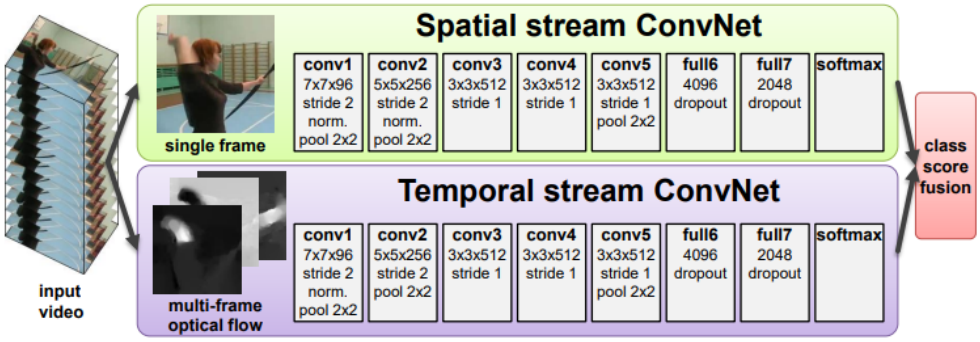


FIGURE 2.4 – Architecture à deux flux pour la classification vidéo [13]

Ensuite les résultats de classification obtenus à partir de ces deux flux sont combinés par une fusion tardive pour produire la prédiction finale de l'action dans la vidéo en utilisant deux méthodes de fusion : la moyenne des scores des deux flux ou en utilisant un Support Vector Machines (SVM) multiclasse sur les scores softmax. Pour évaluer la performance d'architecture de réseaux neuronaux, les auteurs utilisent les ensembles de données UCF-101 [73] et HMDB-51 [74] avec la fusion des flux temporels et spatiaux, la précision atteint 88.0% pour la base UCF-101 et 59.4% pour la base HMDB-51.

2.2.1.4 Travail de Gupta et al. [14]

L'approche proposée dans [14] consiste à identifier l'activité humaine dans les images vidéo basé sur un réseau Convolution Long Short-Term Memory (Conv-LSTM) et une nouvelle étape de prétraitement est appliquée pour produire des masques de segmentation humaine pour mettre en évidence le sujet humain par un modèle d'encodeur-décodeur basé sur DenseNet [75]. Après le prétraitement, un encodeur convolutionnel pré-entrainé Efficient-Net [76] est utilisé qui applique une mise à l'échelle composite construisant des vecteurs de caractéristiques qui sont ensuite alimentées à des LSTM bidirectionnels pour comprendre les relations temporelles entre les images vidéo. Les résultats de LSTM bidirectionnel sont ensuite fournis à la couche entièrement connectée pour l'apprentissage. Enfin, une couche softmax est utilisée pour prédire l'activité. Le système proposé est schématisé dans la Figure 2.5.

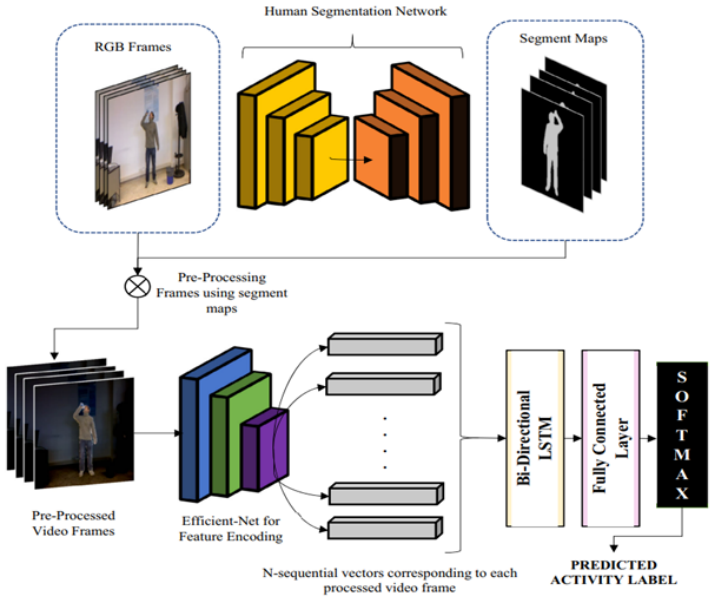


FIGURE 2.5 – Architecture guidée par segmentation proposée pour la reconnaissance de l'activité humaine [14].

L'approche proposée est soumise à des tests rigoureux sur trois ensembles de données publiques KARD [77], MSR Daily Activity [78] et SBU-interactions [79]. Les résultats soulignent l'efficacité et la robustesse de l'approche proposée.

2.2.1.5 Travail de Tej Singh et al. [15]

Dans cette étude [15], les chercheurs ont proposé un modèle ConvNet (voir la figure 2.6) qui utilise deux flux de données, les images RGB au niveau supé-

rieur pour l'extraction des caractéristiques spatiales avec un réseau Bidirectionnel Long Short-Term Memory (Bi-LSTM) pour l'acquisition d'informations séquentielles supplémentaire, et des images appelées Dynamic Motion Image (DMI) au niveau inférieur pour l'extraction des caractéristiques temporelles. Les deux flux sont entraînées en utilisant l'architecture profonde Inception-v3 [80] pré-entraînée. Les scores obtenus par ces deux flux sont fusionnés avec des techniques de fusion tardive au niveau de la décision.

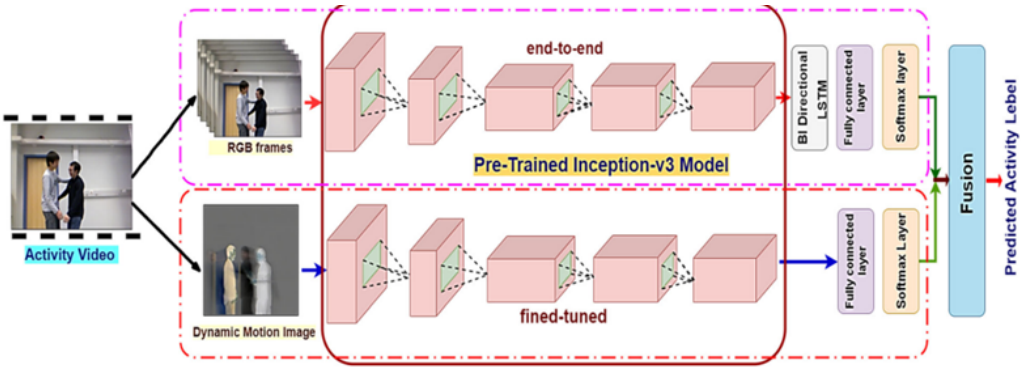


FIGURE 2.6 – L'architecture ConvNet proposée pour la reconnaissance de l'activité humaine [15].

Pour l'évaluation de modèle, ils ont utilisé quatre ensembles de données SBU-Interactions [79], MIVIA Action [81], MSR Action Pair [82], MSR Daily Activity [78]. Les résultats montre l'efficacité de la méthode proposée.

2.2.1.6 Travail de Samundra Deep et al. [16]

Dans cette étude [16], les chercheurs ont expérimenté avec trois modèles CNN à savoir VGG-16, VGG-19, InceptionNet-v3 pour la reconnaissance d'activités humaines sur l'ensemble de données largement utilisé en vision par ordinateur Weizmann [70]. Ils ont appliqué la technique de transfert d'apprentissage pour exploiter les connaissances acquises à partir de l'ensemble de données ImageNet [83] pour tous les modèles de CNN pour réduire le temps d'entraînement et améliorer la précision de modèle. Le schéma du système proposé est représenté dans la Figure 2.7.

La précision est améliorée de 1 à 6% en appliquant le transfert d'apprentissage. Ils ont obtenu des précisions pour VGG-16, VGG-19 et InceptionNet-v3, respectivement, de 96.95%, 96.54% et 95.63%.

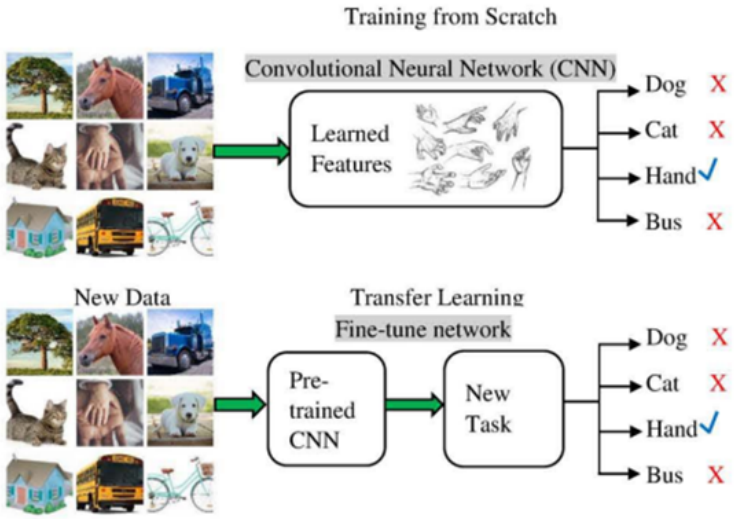


FIGURE 2.7 – Schéma de l'apprentissage par transfert proposé par [16].

2.2.1.7 Travail de Kushwaha et al. [17]

Dans cet article [17], les chercheurs ont utilisé une technique de résumé vidéo pour extraire les trames du clip vidéo (trames clés). Ensuite, ces trames sont utilisées comme entrée du modèle CNN proposé qui empile deux modules de convolution et d'identité. Les modules de convolution pour l'extraction des caractéristiques, tandis que les modules d'identité pour préserver les informations importantes sans les altérer. Deux types de couches de mise en commun sont utilisés pour réduire la dimensionnalité des données et la complexité de computationnelle. Ils ont utilisé aussi des connexions de raccourci pour préserver les gradients et les informations des couches précédentes. Après chaque couche de convolution une fonction d'activation ReLU est utilisée, la normalisation par lots est appliquée après les activations ReLU pour stabiliser et accélérer l'apprentissage. A la fin, le classificateur softmax est utilisé pour attribuer des probabilités afin de prédire l'activité. Le schéma du système proposé est représenté dans la Figure 2.8.

L'efficacité de l'approche proposée a été démontré par la réalisation de plusieurs expériences sur deux ensemble de données anciens : IXMAS[84], HMD51[85], ainsi que sur trois ensemble de données récents : Breakfast[86], YouTube-8M [87] et Kinetics-600 [88].

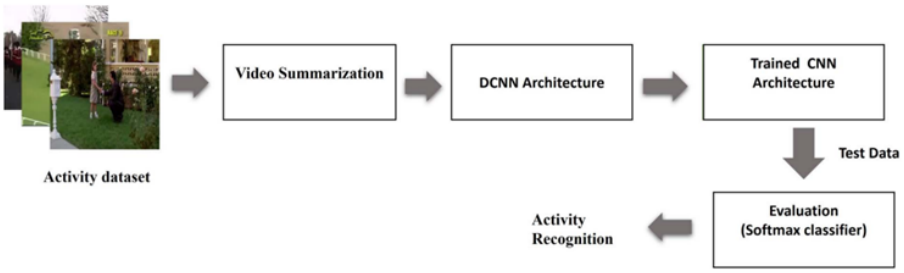


FIGURE 2.8 – Le cadre général de la méthode proposée pour HAR.[17]

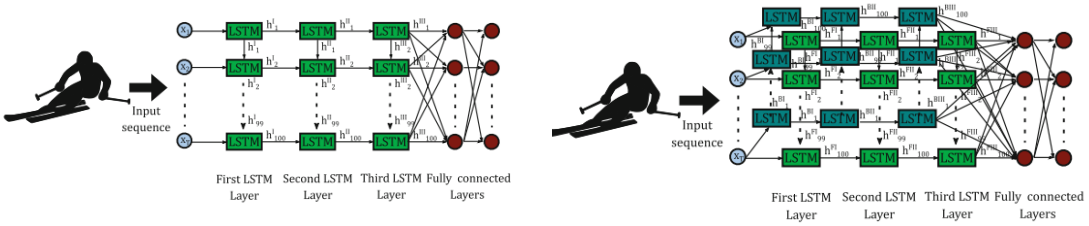
2.2.2 Traitement des signaux provenant des capteurs

Parmi les nombreux travaux HAR basé sur l'utilisation des signaux provenant des capteurs, dans cette sous-section, nous allons présenter certains travaux utilisant les techniques deep learning.

2.2.2.1 Travail de Pawlyta et al. [18]

Pawlyta et al. [18] ont présenté une nouvelle approche pour la reconnaissance d'activités humaines spécifiques, en l'occurrence le ski, utilisant le deep learning. Dans cette étude, les auteurs utilisent les capteurs placés sur un skieur (un accéléromètre, un gyroscope, et un magnétomètre) pour recueillir des données synchrones. Étant donné la rapidité et la variabilité du mouvement des skieurs, il est difficile de reconnaître les activités spécifiques réalisées pendant le ski, donc pour mieux gérer cette complexité les auteurs proposent l'implémentation d'un RNN basé sur deux modèles d'architecture LSTM (Unidirectionnel,bidirectional Long Short-Term Memory). Les signaux d'entrée sont segmentés en fenêtres de longueur T et introduits dans le modèle. Chaque fenêtre contient une séquence d'échantillons individuels observés par le capteur au temps t $(x_1, x_2, ..., x_T)$. Les meilleurs résultats ont été obtenus pour une fenêtre d'une longueur $T = 32$ et trois couches cachées comportant chacune 100 unités cachées. Chaque couche cachée est suivie d'une couche de dropout pour prévenir le surajustement. Dans les deux modèles, la dernière couche est une couche pleinement connectée et elle est utilisée pour effectuer l'interprétation des caractéristiques de haut niveau. Le système proposé est schématisé dans la Figure 2.9.

Ensuite, les auteurs évaluent le modèle à partir de leur propre base de données et les performances des modèles proposés pour chaque activité sont résumées dans la TABLE 2.1.



(a) LSTM-Unidirectionnel

(b) LSTM-Bidirectionnel

FIGURE 2.9 – Le réseau LSTM à trois couches proposé [18].

Task	F1-score (%)		Accuracy (%)	
	uni-LSTM	bi-LSTM	uni-LSTM	bi-LSTM
Turn	99.30	99.57	99.33	99.59
Leg lift	92.55	96.41	99.61	99.78
ski2ski orientation	99.21	99.43	99.21	99.43
Body position	99.03	99.10	99.13	99.19

TABLE 2.1 – Moyenne macro du score F1 et de l'exactitude des modèles proposés pour chaque événement

2.2.2.2 Travail de Murad and Pyun [19]

Murad and Jae-Young Pyun ont mis en place une nouvelle architecture HAR reposant sur l'utilisation d'un RNN basée LSTM pour le traitement des séquences temporelles de données capturées par un accéléromètre et un gyroscope. Le système proposé, illustré par la Figure 2.10, fait en premier lieu l'entrée d'une séquence discrète d'échantillons régulièrement espacés $(x_1, x_2, ..., x_T)$, où chaque point de données x_t est un vecteur d'échantillons individuels observés par les capteurs au temps t. Ces échantillons sont segmentés en fenêtres d'un indice de temps maximal T et alimentés dans un modèle RNN. Pour chaque fenêtre de données traitée le modèle produit une séquence de vecteurs de scores où chaque vecteur représente les prédictions des classes d'activités à l'instant t. Ensuite, les scores de prédiction pour chaque classe d'activité sont additionnés sur tous les instants T en utilisant une technique de fusion tardive et puis calculer une moyenne on divisant la somme des scores par le nombre total d'instants T. En dernière étape, une couche softmax est appliquée à ce résultat pour convertir les scores en probabilités de classe qui facilite la classification et la prise de décision.

Les auteurs ont développé des architectures pour trois modèles de DRNN qui sont les suivants : (Unidirectional LSTM, Bidirectional LSTM, Cascaded Bidirectional and Unidirectional LSTM). Pour évaluer leur modèle, ils ont utilisé plusieurs datasets (UCI-HAD [89], USC-HAD [90], Opportunity [91], Daphnet FOG [92],

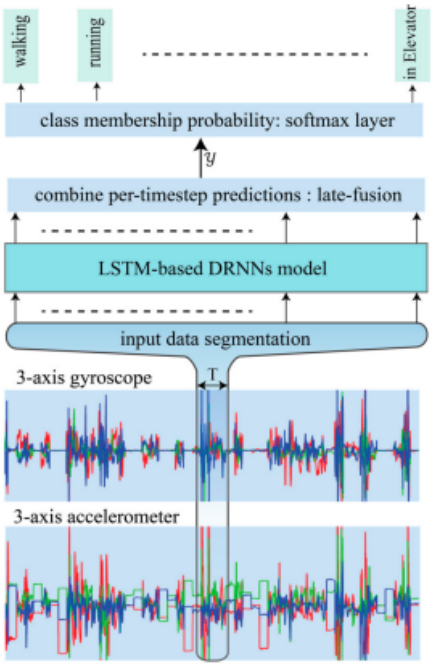


FIGURE 2.10 – Architecture HAR proposée à base d LSTM [5] [19]

Skoda [93]) et ont obtenu les performances résumés dans la TABLE 2.2.

Model	Dataset	Accuracy	Average Precision	F1 Score
Unidirectional DRNN	UCI, USC-HAD	96.7%	96.8%	96.7%
Bidirectional DRNN	Opportunity	92.5%	86.7%	83.5%
Cascaded DRNN	Daphnet FOG	94.1%	84.7%	78.9%
Cascaded DRNN	Skoda	92.6%	93.0%	92.6%

TABLE 2.2 – Evaluation metrics for different DRNN models on various datasets.

2.2.2.3 Travail de Mohsen [20]

Dans ce travail [20], Mohsen met en place un système pour classifier les activités humaines avec l'utilisation de RNN basé sur un algorithme Gated Recurrent Unit (GRU). Cet algorithme est appliqué à l'ensemble de données Wireless Sensor Data Mining WISDM [94] qui comprend six activités humaines différentes dans le but d'obtenir une précision élevée. L'architecture de l'algorithme GRU comprend une couche d'entrée, deux couches GRU et une couche de sortie (voir la Figure 2.11). La couche d'entrée est composée de trois caractéristiques : ax, ay et az qui sont utilisées pour capturer les informations temporelles des données d'activité

humaine. Deux couches GRU sont empilées pour améliorer la stabilité et la précision de l'algorithme avec l'utilisation d'une fonction d'activation linéaire rectifiée (ReLU) pour renforcer la robustesse de l'algorithme et une couche de sortie comprend six neurones avec une fonction d'activation softmax et l'optimiseur Adam qui ajuste les poids de l'algorithme GRU. En outre, une technique de régularisation est mise en œuvre pour prévenir le surajustement.

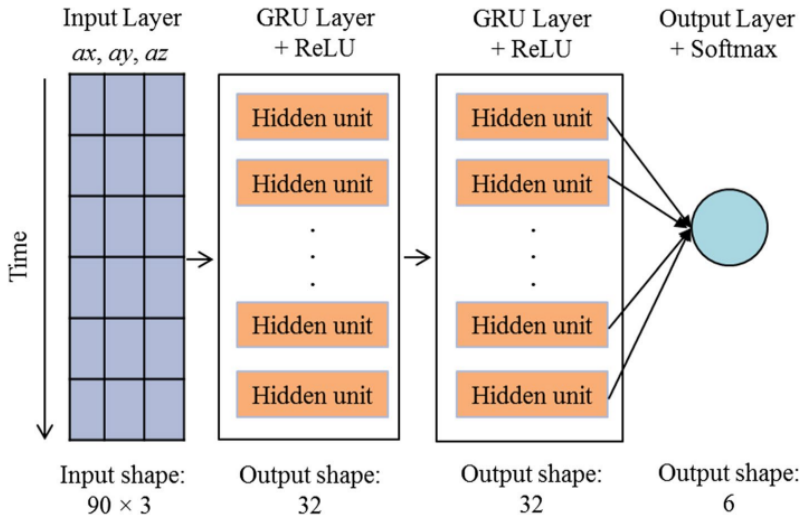


FIGURE 2.11 – L'architecture de GRU proposée en [20]

Les résultats démontrent des bonnes performances, dont 97,08% de précision, ce qui répond avec succès à la problématique posée au départ.

2.2.2.4 Travail de Ali et al.[21]

Les auteurs proposent un modèle d'apprentissage profond léger pour la reconnaissance des activités humaines qui utilise les données des capteurs (accéléromètres, gyroscopes et magnétomètres) directement avec très peu de prétraitement, afin de surmonter les difficultés d'extraction de caractéristiques à partir de données bruyantes ou inexactes, ainsi que la création de modèles volumineux qui ne peuvent pas être déployés sur des appareils à ressources limitées.

L'architecture proposée (voir la Figure 2.12) débute par une couche de convolution comprenant 64 filtres, destinée à extraire les caractéristiques des lectures des capteurs et à identifier les différentes relations entre ces données. Ensuite, les sorties de cette couche sont soumises à des couches de max-pooling pour réduire la taille de la carte des caractéristiques et la rendre plus gérable pour une analyse ultérieure, après ils vont réappliquer une couche de convolution similaire à la pre-

mière afin d'extraire des caractéristiques encore plus détaillées des données. Pour calculer la moyennes des caractéristiques et produire une valeur unique qui capture les informations les plus importantes, ils utilisent la technique Global Average Pooling (GAP). Enfin, le modèle utilise une couche entièrement connectée pour prédire le type d'activité en cours.

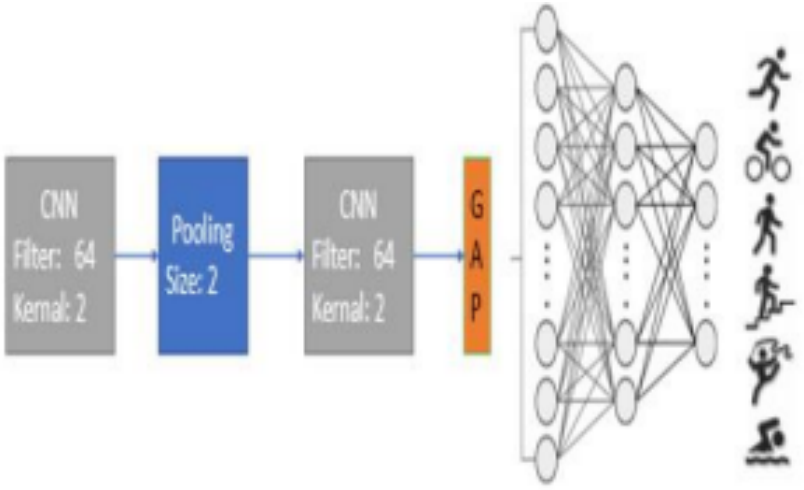


FIGURE 2.12 – Architecture du système proposé par Ali et al. [21]

Pour prouver la robustesse de modèle proposé, ils ont utilisé deux ensembles de données de référence WISDM [94] et PAMAP2 [95] en atteignant une précision de 99,71% et 94,31% respectivement.

2.2.2.5 Travail de Ryu et al. [22]

Dans ce travail [22], les chercheurs ont proposé une approche basée sur des signaux photopléthysmographie (PPG) et un réseau de neurones convolutif (1D CNN) (Figure 2.13). PPG est une technique qui mesure les variations de volume sanguin dans des vaisseaux sanguins à l'aide de capteurs optiques pour surveiller la fréquence cardiaque et évaluer l'activité physique. Les mesures brutes PPG de chaque participant sont sous-échantillonnées et segmentées, et redimensionnées pour être utilisées comme représentation d'entrée de modèle 1D CNN proposé. Le modèle classe ensuite les données d'entrés en cinq activités.

Le modèle comprenait dix couches convolutives et quatre couches de pooling maximum, et une couche de regroupement globale a été appliquée pour convertir la carte de caractéristiques extraite de couches convolutives en un vecteur 1D. Ce vecteur a traversé cinq couches entièrement connectées, ensuite il a été activé par softmax pour générer une classification, un dropout a été appliqué après les couches de pooling pour prévenir le surapprentissage. Les chercheurs ont mené deux

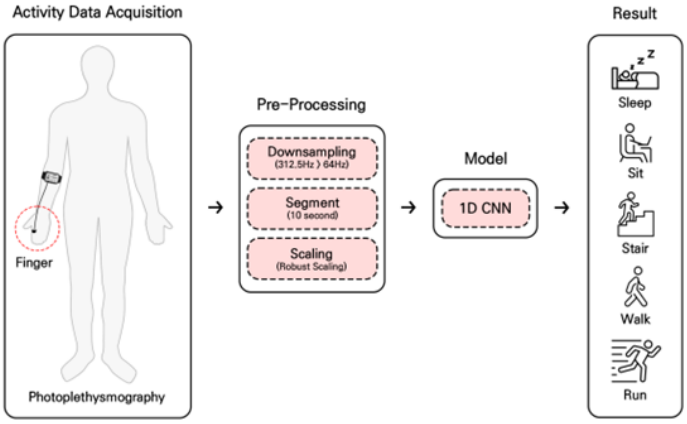


FIGURE 2.13 – Aperçu du cadre de reconnaissance de l'activité humaine proposé basé sur les signaux de photopléthysmogramme. en [22]

expériences, dans l'expérience 1, une validation croisée a été utilisé pour évaluer les performances de généralisation de l'approche proposée. Dans l'expérience 2, l'effet de la longueur du signal d'entrée sur les performances du système HAR a été étudié pour déterminer la taille de la fenêtre optimale utilisant une validation croisée intra-sujets. Les résultats montrent que la méthode proposée permet de distinguer avec succès les cinq activités considérées, avec une précision moyenne de test de 95,14 %.

2.2.2.6 Travail de Raj et Kos. [23]

Cette étude [23] vise à collecter des données à l'aide de capteurs mobiles, à les prétraiter, à les segmenter en segments appropriés, puis les utiliser pour entraîner un modèle de reconnaissance basé sur un CNN 2D. Les étapes comprennent également la validation et les tests du modèle, ainsi que l'adaptation des hyperparamètres pour améliorer les performances. Les chercheurs ont utilisé l'ensemble de données WISDM [94] en raison de sa richesse en échantillons. Le système décrit est illustré dans la Figure 2.14.

Une analyse comparative a été faite entre l'approche proposée et les techniques HAR basées sur le modèle CNN utilisant l'ensemble de données WISDM. Le taux de précision de la méthodologie proposée est supérieur à celui des autres. Le modèle atteint un taux de précision de 97,20%.

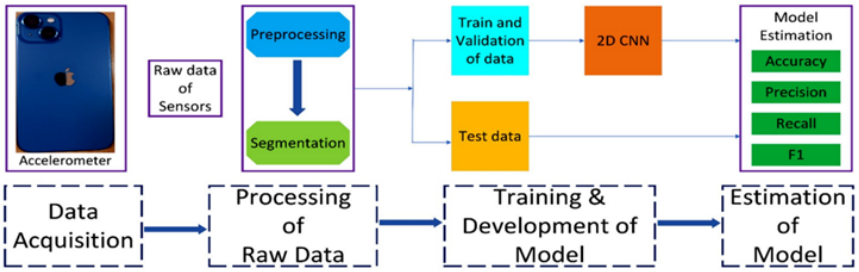


FIGURE 2.14 – Le cadre de reconnaissance de l'activité humaine proposé en. [23]

2.2.2.7 Travail de Parka et al. [24]

Dans l'article [24], les chercheurs ont présenté un travail de reconnaissance d'activités humaines avec les caractéristiques des angles des articulaire du corps basés sur RNN (voir la Figure 2.15). Ils ont extrait 28 caractéristiques d'angles articulaires à partir de 14 parties du corps clés et ils ont défini le nombre d'images pour une activité comme étant 50 pour créer une matrice de caractéristiques d'entrée des angles articulaires calculés à partir de l'ensemble de données d'activités MSRC-12 [96]. Cet ensemble de données se compose de séquences d'activités humaines contenant 12 activités capturées à l'aide d'une caméra de profondeur. Ensuit, ils ont entraîné les RNN avec la matrice des caractéristiques.

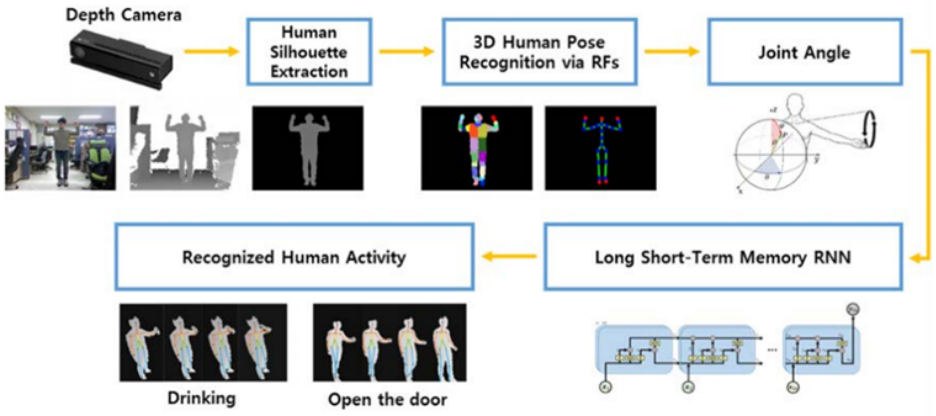


FIGURE 2.15 – Le système HAR basé sur RNN proposé en. [24]

Les performances de ce système basé sur RNN sont comparées à d'autres systèmes de reconnaissance tels que les modèles de Markov cachés (HMM) [96] et les réseaux de croyance profonde (DBN) [97] utilisant la base de données MSRC-12. Les résultats montrent que l'approche proposée surpasse les méthodes HMM

et DBN, avec une précision de reconnaissance moyenne de 99,55%, surpassant de 7,06% la méthode HMM et de 2,01% la méthode DBN.

2.2.2.8 Travail de Wang et Liu [25]

L'article[25] propose une nouvelle structure LSTM hiérarchique (H-LSTM) avec deux couches cachées basées sur des capteurs portables (la Figure 2.16). Premièrement, les données brutes sont prétraitées à l'aide de techniques de lissage et de débruitage, ensuite, les caractéristiques sont extraites par la méthode de domaine temps-fréquence. A la fin le modèle H-LSTM est utilisé pour classer les activités.

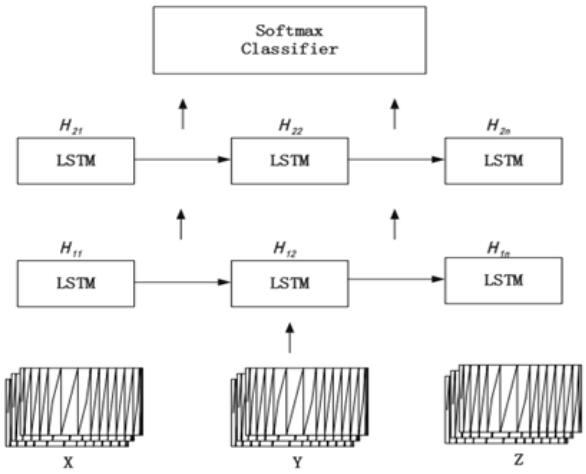


FIGURE 2.16 – Processus de reconnaissance de l'activité humaine basé sur H-LSTM. [25]

Les chercheurs ont utilisé trois ensembles de données UCI pour évaluer l'approche. Les résultats expérimentaux montrent que le système H-LSTM offre de meilleures performances que d'autres algorithmes d'apprentissage profond. L'exactitude de la reconnaissance des activités humaines atteint 99,15% avec l'approche H-LSTM.

2.2.3 Méthodes hybrides d'apprentissage profond

Divers efforts de recherche ont été orientés vers l'obtention de caractéristiques robustes et efficaces pour la reconnaissance des activités humaines en combinant des méthodes discriminatives. À partir de la littérature disponible sur la mise en œuvre hybride, le réseau neuronal convolutif semble être le meilleur choix pour

de nombreuses études afin d'être hybridé avec d'autres modèles pour la reconnaissance des activités humaines. Par exemple, le réseau neuronal convolutif et le réseau neuronal récurrent. Dans cette section, nous examinons quelques articles qui utilisent des méthodes hybrides.

2.2.3.1 Travail de Ghosh et al.[26]

Akash et al. [26] ont présenté une avancée récente dans le domaine de la reconnaissance des activités humaines et cela en proposant une approche novatrice qui combine les avantages des dispositifs de détection multimodaux et d'un modèle d'apprentissage profond hybride. Le modèle proposé utilise une architecture multi-canal, intégrant un réseau CNN et un réseau RNN bidirectionnel à mémoire court-terme (BLSTM). Chaque canal comprend trois couches Conv1D (convolutions unidimensionnelles) empilées pour extraire des caractéristiques des données de capteurs et représenter les informations spatiales des données, ensuite la sortie de ces couches est dirigées vers des couches de max-pooling et enfin les caractéristiques des différents canaux sont concaténées. Par la suite, les caractéristiques concaténées sont fournies à des couches BLSTM afin de capturer les relations séquentielles entre ces séquences de données et de prédire l'activité humaine correspondante en ajoutant une couche de dropout entre les couches LSTM pour éviter le sur-apprentissage. La sortie des couches BLSTM est fournie à une couche entièrement connectée, suivie d'une couche de sortie avec une activation softmax. La Figure 2.17 est une représentation globale de l'architecture du système proposé.

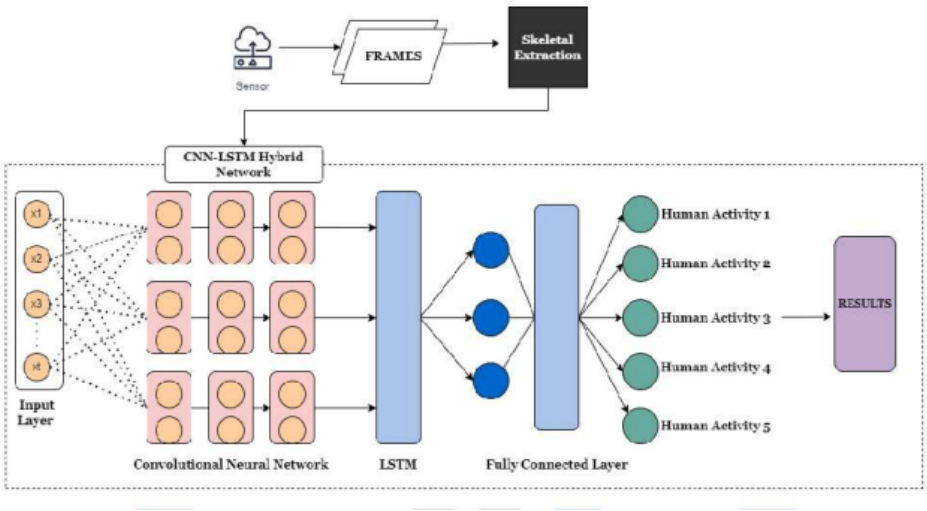


FIGURE 2.17 – Vue sur l'architecture multi-canal proposée dans [26].

D'après les résultats des tests sur deux ensembles de données disponibles publiquement, la technique CNN-LSTM atteint une précision de 92,32%.

2.2.3.2 Travail de Challa et al. [98]

Les auteurs Challa et al. [98] proposent dans cette étude un modèle de deep learning pour HAR basé sur les données des capteurs IMU. Le cadre proposé comprend deux modèles de réseau (CNN et Bi-LSTM) qui utilisent les mêmes données d'entrée pour extraire simultanément des caractéristiques spatiales et temporelles des données brutes des capteurs,ensuite les caractéristiques concaténées sont transmises à une couche dense, activée par une fonction ReLU, puis normalisées à l'aide de techniques de normalisation par lots et de dropout pour accélérer l'apprentissage et prévenir le surapprentissage. Enfin, les caractéristiques obtenues sont alimentées dans une couche de classification qui utilise la fonction d'activation softmax pour classifier les différentes activités. La particularité de ce modèle réside dans l'utilisation de l'algorithme d'optimisation métaheuristique Rao-3 pour déterminer les valeurs des hyperparamètres idéales afin d'améliorer les performances de reconnaissance du modèle DL proposé.

Les performances du modèle DL proposé ont été validées sur les ensembles de données PAMAP2 [95], UCI-HAR [89] et MHEALTH [99], et ont atteint respectivement des précisions de 94,91%, 97,16% et 99,25%. Les résultats indiquent que le modèle DL proposé surpasse les modèles de l'art existants.

2.2.3.3 Travail de Hassan et al. [27]

Hassan et al. [27] ont mené une étude décrivant en détail l'architecture de création du modèle spécialisé pour la reconnaissance dynamique des actions humaines en mettant l'accent sur l'utilisation de MobileNetV2 et du framework Deep BiLSTM. Ce système, présenté par la Figure 2.18, utilise MobileNetV2 pour extraire des caractéristiques à partir de cadres vidéo, ensuite elles sont affinées par le modèle Deep BiLSTM pour capturer les dépendances temporelles à long terme et classifier les activités. Plusieurs technique telles que la convolution, le MaxPooling sont utilisées pour optimiser les caractéristiques extraites, ainsi que les activations ReLU et softmax sont choisies pour améliorer l'efficacité computationnelle et la précision des prédictions. Enfin, le modèle est entraîné en utilisant une technique de réglage robuste, où il est entraîné plusieurs fois sur des batches de données.

La performance du modèle proposé a été évaluée de manière rigoureuse à l'aide de trois ensembles de données de référence, à savoir UCF11 [100] et UCF Sport [101], obtenant des précisions remarquables de 99,20%, 93,3 %, respectivement.

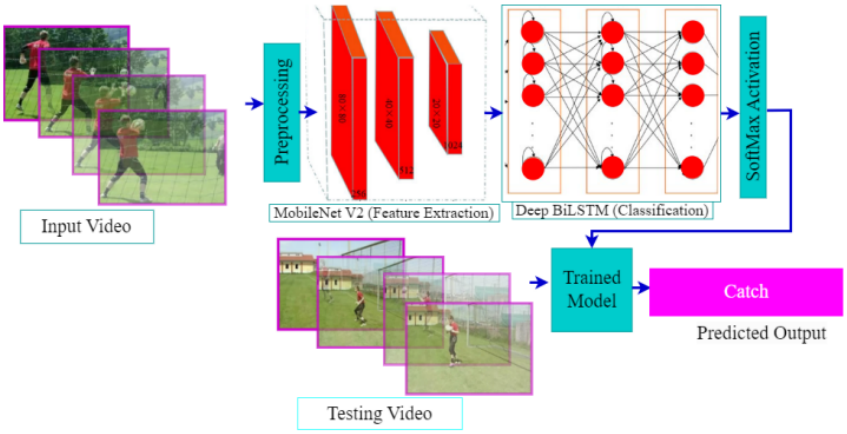


FIGURE 2.18 – Architecture de la méthode proposée dans [27].

2.2.3.4 Travail de Kumar et al. [102]

L'étude de Kumar et al. [102] propose un système pour la HAR utilisant trois modèles d'apprentissage profond (LSTM, CNN, MLP) sur le jeu de données UCF-101 [103], disponible publiquement pour la HAR. Le but de l'étude étant de faire la comparaison entre ces modèles pour la classification des activités. Après prétraitement des données et extraction de caractéristiques avec InceptionV3, elles sont utilisées comme entrée pour les différents modèles d'apprentissage profond. Pour les tests, le modèle CNN est d'abord testé individuellement, puis en combinaison avec les autres. D'après les résultats présentés, le modèle CNN-LSTM a atteint une précision de 79,21%. De plus, une précision catégorielle des 5 meilleures prédictions de 92,92% a été observée. Ces résultats indiquent que le modèle CNN-LSTM a surpassé les autres modèles évalués sur la base de données UCF-101.

2.2.3.5 Travail de Ding et al. [28]

Les chercheurs de [28] ont proposé un nouveau modèle de deep learnig pour la reconnaissance d'activités humaines appelé HAR-DeepConvLG. Le modèle se compose de trois couches de convolution CNN et un bloc de compression et d'excitation (SE) qui sont utilisés pour extraire précisément les caractéristiques spatiale des données brutes collectées par des capteurs (voir la Figure 2.19). Les caractéristiques extraites sont utilisées comme entrée de trois chemins parallèles, chacun contenant une couche LSTM connectée en séquence à une couche GRU pour apprendre la représentation temporelle. Les sorties des chemins sont concaténées et renvoyées au bloc de reconnaissance (RB) pour reconnaitre l'activité humaine finale.

Pour évaluer le modèle des expériences ont été menées sur quatre ensembles

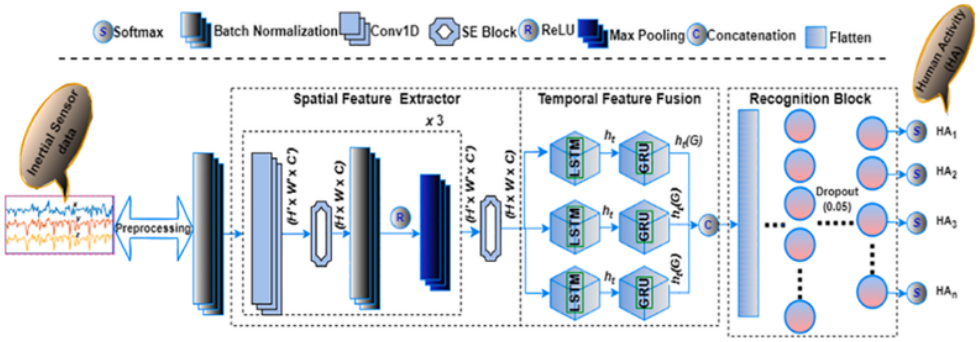


FIGURE 2.19 – Structure du modèle HAR-DeepConvLG proposé en.[28]

de données : UCI-HAR [89], WISDM [94], PAMP2 [95] et USC-HAR [90]. Les résultats expérimentaux indiquent que le modèle HAR-DeepConvLG atteint une précision de classification de 97,52 %, 98,48 %, 97,85 % et 98,55 % sur les ensembles de données donnés dans l'ordre précédent.

2.2.3.6 Travail de Rijayanti et al. [29]

Dans cette étude [29], les chercheurs ont comme objectif de déterminer les mouvements des travailleurs, les postures et les interactions avec les objets environnants. Après l'extraction des caractéristiques de la vidéo, les données sont divisées en données d'entraînement et de test 75 % et 25 %, ils ont utilisé la méthodologie Mask R-CNN [104], pour la détection des objets. Ensuite, ils ont crée des boîtes englobantes (bbox) et des masques de segmentation. Les poses sont identifiées et classifiées pour reconnaître les mouvements corporels des travailleurs à partir du framework MediaPipe Holistic. A la fin, le modèle génère un texte descriptif basé sur la pose identifiée du travailleur et son interaction avec les objets environnants et classifie le comportement comme normal ou anormal. Le système décrit est illustré dans la Figure 2.20.

Le modèle proposé a détecté avec succès les comportements anormaux dans les vidéos testées avec une précision de détection d'objet acceptable d'environ 97% et une précision de reconnaissance de pose d'environ 96%.

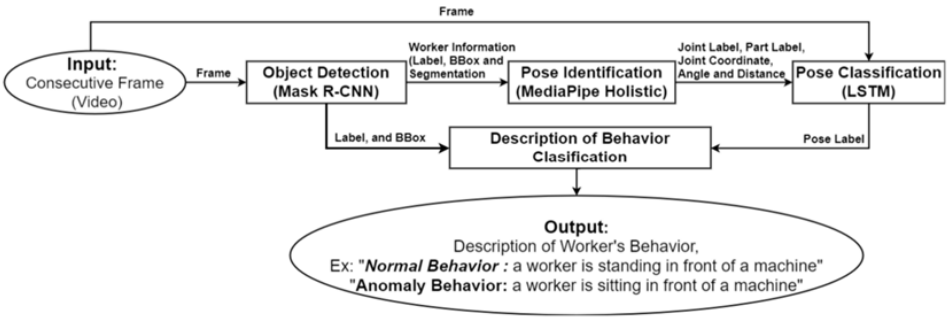


FIGURE 2.20 – Architecture proposée dans [29].

2.3 Comparaison des travaux

Dans le domaine de HAR, il y a des chercheurs qui continuent de proposer de nouvelles approches pour les systèmes de reconnaissance d'activités humaines, d'autres se consacrent à la réalisation de revues ou de comparaisons. Ces travaux rassemblent généralement les approches récemment proposées et les classifient selon une taxonomie spécifique. Ils les comparent ensuite en utilisant des mesures de performance standard dans le but d'aider la recherche dans ce domaine. Un exemple de cela, est la revue de Sharma et al. [45] qui fait un état de l'art sur les méthodes utilisées dans les systèmes HAR. De la même façon, nous mettons en œuvre, dans cette section, un tableau comparatif (Table 2.3) des travaux synthétisés afin d'avoir une vision plus claire et de pouvoir comparer les approches et leurs points distinctifs selon des mesures de performance fixes. Les critères de comparaison sont le matériel utilisé dans l'expérimentation, le dataset,le type de Deep learning, les performances atteintes et limites constatées.

TABLE 2.3 – Tableau comparatif des travaux

Auteur & Année	Matériaux	Dataset	Deep	Limites	Performance
Khelalef et al. [11] 2019	Silhouettes	Wiezman, Keck Esture, KTH	CNN	L'efficacité du système dépend de la qualité des silhouettes extraites	Acc= 100%
Karpathy et al. [12] 2014	Caméra vidéos	1M Sports	CNN (Slowfusion)	Le modelé n'est pas sensible aux détails spécifique de la connective temporelle	Average Acc =80%

Continued on next page

Table 2.3 – continued from previous page

Auteur & Année	Matériaux	Dataset	Deep	Limites	Performance
simonyan et al. [13] 2021	Caméras	ECF-101, HMDB-51	Conv-Nets	l'absence de la méthode de regroupement de caractéristique qui capteur les information spatiale et temporelle importante autour de trajectoire dans les vidéos	UCF-101 Average Acc=86.9%, SVM=88% HMDB-51 Average Acc=58%, SVM=59.4%
Pawlita et al. [18] 2019	Capteurs (accéléromètre, gyroscope et magnétomètre)	Données collectées	LSTM (bi-LSTM, Uni-LSTM)	Leur ensemble de données propre présente un déséquilibre, et comme le modèle est uniquement entraîné sur les données provenant d'une seule piste de ski, cela pourrait limiter leur capacité à généraliser à d'autres environnements ou conditions.	Uni-LSTM Acc=99.13, F1-Score=99.10% Bi-LSTM Acc=99.19%, F1-Score=99.03
Murad et al. [19] 2017	Capteurs (Accelerométre et gyroscope)	UCI-HAD,USC-HAD, Opportunity	LSTM	Bien que les CNN puissent capturer les dépendances locales entre les échantillons d'entrée, le partage de paramètres dans le temps et la connectivité locale ne sont pas suffisants pour capturer toutes les corrélations entre les échantillons.	UCi-HAD Average Acc=96.7%, USC-HAD Average Acc=97.4% Opportunity Average Acc=86.7%
Saeed et al. [20] 2023	Capteurs (accélérométre)	WISDM	RNN (GRU)	La contrainte liée à l'utilisation des données provenant d'un seul capteur représente un défi pour maintenir la précision du modèle en présence de données sensorielles manquantes dues à des pertes de données	Acc= 97.28%, Average Acc= 97.11%,F1-Score= 97.10%

Continued on next page

Table 2.3 – continued from previous page

Auteur & Année	Matériaux	Dataset	Deep	Limites	Performance
Ali et al. [21] 2023	les capteurs (accéléromètre, gyroscope et magnétomètre)	WISDM, PAMAP2	CNN	Le modèle léger proposé par les auteurs, conçu pour les appareils avec des ressources limitées, pourrait manquer de profondeur pour traiter efficacement des activités complexes. En conséquence, sa performance pourrait être moins satisfaisante dans des environnements complexes.	WISDM Acc =99.71%, PAMAP2 Acc=94.31%
Akash et al. [26] 2023	Les capteurs	Disponible publiquemnt	CNN, Bi-Lstm	l'article incluent la complexité de l'intégration multimodale, le besoin de données étiquetées en grande quantité pour les modèles d'apprentissage profond hybrides	Average Acc=92.32%
kumar et al. [98] 2023	Capteurs IMU (accéléromètre, gyroscope et magnétomètre)	PAMAP2, UCI-HAR, M-HEALTH	CNN, Bi-Lstm	l'utilisation de l'algorithme d'optimisation métaheuristique Rao-3 pour déterminer les valeurs des hyperparamètresUne peuvent ne pas produire une solution optimale pour tous les cas en fonction d'une fonction objective.	PAMAP2 Acc=94.91, F1-Score=94.89% M-HEALTH Acc=99.25%, F1-Score=99.28% UCI-HAR Acc=97.16%, F1-Score=97.14
Hassan et al. [27] 2024	Caméras	UCF-11, UCF SPORT, JHMDB	MobileNet, Bilstm	la limite de modelé propose lié aux préoccupations de confidentialité des individus surveillés donc aux futurs pourraient aborder ces préoccupations en explorant des techniques préservant la vie privée.	UCF-11 Acc= 99.2%, UCF SPORT Acc= 93.3%,JHMDB ACC= 76.30%

Continued on next page

Table 2.3 – continued from previous page

Auteur & Année	Matériaux	Dataset	Deep	Limites	Performance
Kumar et al. [102]2023	Caméras	UCF-101	CNN-lstm, CNN-MLP	Les auteurs n'ont pas correctement ajusté les valeurs des hyperparamètres, notamment la taille du noyau dans les CNN, ce qui a eu un impact négatif sur la capacité du modèle à extraire des caractéristiques pertinentes.	CNN-lstm Acc =79.21%, CNN-MLP Acc=70.04%
Gupta et al. [14] 2024	caméra-vidéos	KARD, MSR Daily Activity, SBU-interaction	conv-LSTM	la nécissité d'améliorer la cohérence temporelle pour les séquences vidéos de plus longue durée.	KARD Acc = 99% , MSR Daily Activity Acc = 98% , SBU-interaction Acc = 99% .
Tej et al. [15] 2020	images RGB, images dynamiques (caméra vidéos)	SBU-interaction, MIVIA Action, MSR Action pair, MSR Daily Activity	CNN- BiLSTM	confusion entre les activités similaires	SBU-interaction Acc = 98.70%, MIVIA Action Acc = 99.44%, MSR Action pair Acc = 98.30%, MSR Daily Activity Acc = 94.37%.
Samundra Deep et al. [16] 2019	caméras	Weizmann	VGG-16, VGG-19,Inception-V3	les connaissances transférées à partir de ImageNet peut être compromise car ImageNet contient des images de plusieurs catégories différentes	VGG-16 Acc= 96.95%, Recall = 97%, F1-score = 97%. VGG-19 Acc = 96.54%, F1-score = 96%, Recall = 97%. Inception-V3 Acc = 95.63%,Recall = 96%.
Kushwaha et al. [17] 2024	caméra vidéos	IXMAS, HMDS1, Breakfast, youTube-8M, Kinetics-600	CNN	la technique de résumé vidéo pourrait entraîner une perte d'informations importantes ce qui pourrait affecter la capacité du modèle à capturer des caractéristiques subtiles	Acc = 94.196% Recall = 0.981 F-Measure = 0.98

Continued on next page

Table 2.3 – continued from previous page

Auteur & Année	Matériaux	Dataset	Deep	Limites	Performance
Semin Ryu et al. [22] 2024	PPG	données collectées	1D CNN	le modèle est entraîné uniquement avec des données de certain groupe d'âge, ce qui pourrait produire des résultats biaisés.	Acc = 95.14%
Raj et Kos. [23] 2023	capteurs (accélemèter)	WISDM	2D CNN	l'utilisation des données collectées à partir d'un seul dispositif ne fonctionne pas parfaitement et ne pas être généralisables.	Acc = 97.20%, Loss = 2.80%
Parka et al. [24] 2016	caméra de profondeur	MSRC-12	LSTM	l'utilisation de caméra de profondeur peut nécessiter des ajustements spécifiques de l'environnement.	Acc = 99.55%
Wang et Liu. [25] 2020	capteurs	UCI	H-LSTM	le modèle n'a pas été suffisament validée dans des scénarions en temps réel.	Acc = 91.65%
Ding et al. [28] 2023	capteurs	UCI-HAR, WISDN, PAMP2,USC-HAR	CNN, LSRM, GRU	coût computationnel légèrement élevé en raison de l'utilisation de quelques paramétre hautement entraînables.	UCI-HAR Acc = 97.52%, WISDN Acc = 98.48%, PAMP2 Acc = 97.85%, USC-HAR Acc = 98.55%.
Rijayanti et al. [29] 2023	caméras	Données collectées	CNN, LSTM	Des limites de précision pour certain types de comportement anormaux des travailleurs.	Acc = 97%

2.4 Conclusion

Les nombreux travaux examinés dans ce chapitre témoignent de la richesse et de la diversité des approches utilisées pour aborder les défis des systèmes HAR. Nous avons vu que les premières approches sont basées sur la vision par ordinateur, alors que les secondes utilisent différents types de capteurs fournissant des signaux sur l'activité. De plus, certains travaux ont combiné ces deux approches pour améliorer les performances de leurs systèmes.
Bien que les avancées récentes soient prometteuses, il reste encore beaucoup à faire pour faire progresser davantage le domaine de la HAR. Cela nécessitera un engagement continu de la part des chercheurs et l'exploration de nouvelles approches innovantes pour surmonter les défis et rendre la reconnaissance d'activité humaine plus précise, adaptable et généralisable dans divers contextes d'application. C'est dans sens que nous allons proposer une amélioration des systèmes HAR utilisant Deep learning, en l'occurrence CNN, dans le chapitre suivant.

Chapitre 3

Système de reconnaissance d'activité humaine proposé

3.1 Introduction

Comme vu dans les chapitres un et deux, la reconnaissance d'activités humaines peut bénéficier de l'utilisation de réseaux de neurones profonds pour surmonter les limitations des approches traditionnelles. L'intégration de différentes méthodes de traitement des données, telles que les réseaux de neurones CNN et RNN, rend le domaine de la reconnaissance d'activités humaines très vaste et riche en possibilités. Plusieurs propositions de systèmes, comme celles présentées précédemment dans la synthèse de documents, ont émergé selon divers objectifs de recherche. Ce chapitre proposera donc un système de reconnaissance d'activités humaines basé sur les réseaux de neurones. Il abordera tout d'abord la problématique étudiée, puis présentera la solution proposée en justifiant le choix des réseaux de neurones récurrents. Enfin, il détaillera les différentes phases du système proposé.

3.2 Problématique

La reconnaissance dynamique des activités humaines est un domaine d'étude qui attire actuellement une attention considérable dans les domaines de la vision par ordinateur et le traitement de signal. Le besoin croissant de systèmes basés sur l'IA pour évaluer le comportement humain et renforcer la sécurité souligne la pertinence de cette recherche [27]. Malgré le succès des modèles DL dans la reconnaissance des activités humaines, l'extraction de caractéristiques à la fois des détails spatiaux complexes et les dynamiques temporelles reste un défi en raison du déséquilibre des classes et des données bruyantes [98]. Pour aborder les problèmes mentionnés ci-dessus, l'utilisation des modelés hybride qui combine des couches de réseau de neurones est une alternative intéressante qui commence déjà à être adoptée dans plusieurs propositions différentes de systèmes, notamment celles présentées précédemment dans la synthèse de documents. L'analyse des travaux clés sur le DL appliqué à la reconnaissance d'activités humaines a mis en évidence divers défis et limitations inhérents à ces modèles. Parmi eux, l'interprétabilité et l'explicabilité se démarquent comme des problèmes majeurs, étant donné la complexité des architectures profondes qui rendent difficile la compréhension des décisions du modèle. Le surajustement est également une préoccupation significative, avec des modèles ayant tendance à mal se généraliser sur de nouvelles données [105].

3.3 Solution proposée

Dans ce travail, nous proposons de développer un système de reconnaissance d'activités humaines non coûteux basé sur des modèles hybrides de réseaux de

neurones convolutionnels (CNN) et de réseaux de neurones récurrents (RNN) compréhensibles, fiables et facile d'utilisation, ce qui est particulièrement important dans la reconnaissance d'activités humaines où les décisions du modèle peuvent avoir des impacts significatifs. Cependant, la proposition d'un tel système soulève plusieurs questions concernant le choix des architectures à combiner, les traitements spécifiques à appliquer pour exploiter au mieux ces modèles, ainsi que la méthode d'extraction des informations afin d'améliorer les performances globales du système. Dans ce qui suit, nous justifions notre choix, tout en expliquant les étapes de la solution proposée.

3.3.1 Choix des Capteurs : Approches pour HAR

La reconnaissance des activités humaines consiste à détecter automatiquement les différentes activités physiques quotidiennes que les individus réalisent. Ces activités peuvent être capturées à l'aide d'une gamme d'appareils, tels que des caméras ou des capteurs de mouvement, physiologiques, acoustiques et ambiants.

En fonction de la méthode de détection utilisée, la HAR peut être largement catégorisée en approches basées sur des capteurs externes et internes. Les méthodes externes englobent les signaux optiques (vidéo), les signaux Wi-Fi, les signaux environnementaux et même les ondes sismiques. Notamment, les approches basées sur les caméras ont démontré des performances remarquables en HAR, en particulier avec les avancées dans les réseaux neuronaux artificiels. Cependant, en raison des préoccupations en matière de confidentialité associées aux systèmes basés sur les caméras, des approches alternatives utilisant différents types de capteurs ont émergé. Avec la prolifération des appareils intelligents, les capteurs portables ont attiré une attention significative pour répondre aux préoccupations en matière de confidentialité et de sécurité [106].
Nous avons décidé d'opter pour des capteurs portés dans notre système, en prenant en considération les préoccupations grandissantes la confidentialité des données personnelles.

3.3.2 Choix de type de deep

La reconnaissance d'activité basée sur capteurs consiste à acquérir une connaissance de haut niveau sur les activités humaines à partir de lectures de nombreux capteurs de bas niveau. Ces dernières années, bien que les méthodes existantes d'apprentissage profond aient été largement utilisées pour la HAR basée sur capteurs avec des performances satisfaisantes, elles font toujours face à des défis tels que l'extraction et la caractérisation des caractéristiques, la segmentation des actions continues dans le traitement des problèmes de séries temporelles [107].

Dans le domaine de la Reconnaissance d'Activités Humaines (HAR) basée sur le deep learning, plusieurs méthodes d'extraction de caractéristiques sont largement utilisées. Par exemple, les architectures couramment employées incluent des couches de convolution , principalement utilisées pour analyser des séquences spatiales telles que celles provenant de capteurs comme décrit dans[21]. D'autres approches utilisant des couches LSTM [19] ou des couches GRU [20] sont également largement adoptées pour capturer les dépendances séquentielles dans les données temporelles. Cependant, un défi majeur de ces méthodes est leur capacité limitée à extraire des caractéristiques robustes et riches, ce qui restreint leur capacité à traiter une large gamme de données. L'utilisation séparée de ces méthodes ne permet pas de capturer de manière intégrée les aspects spatiaux et temporels des données, qui sont essentiels pour le suivi des activités humaines, car chacune de ces approches se spécialise davantage dans un type spécifique de caractéristiques à extraire.

Dans notre architecture, nous avons opté pour une méthode hybride afin de relever ces défis. Nous utilisons un réseau de neurones convolutionnels multi-canaux pour améliorer l'extraction des caractéristiques à différentes échelles spatiales. Par la suite, ces caractéristiques extraites sont combinées et introduites dans un RNN. Cette approche permet de capturer efficacement les informations spatiales avec les CNN et d'apprendre les relations temporelles cruciales avec les RNN, offrant ainsi une représentation enrichie et robuste des données.

3.3.3 Explication du système proposé

Dans ce travail, nous proposons un système de Reconnaissance d'Activités Humaines (HAR) basé sur l'utilisation de caractéristiques spécifiques, telles que les signaux provenant d'accéléromètres, de gyroscopes et de capteurs de magnétomètres. Ces signaux sont essentiels pour capturer les mouvements et les orientations corporelles qui permettent d'identifier et de classifier différentes activités humaines.

L'architecture proposé combine un CNN et du GRU pour réaliser l'extraction automatique des caractéristiques spatio-temporelles des données (voir la figure 3.1).

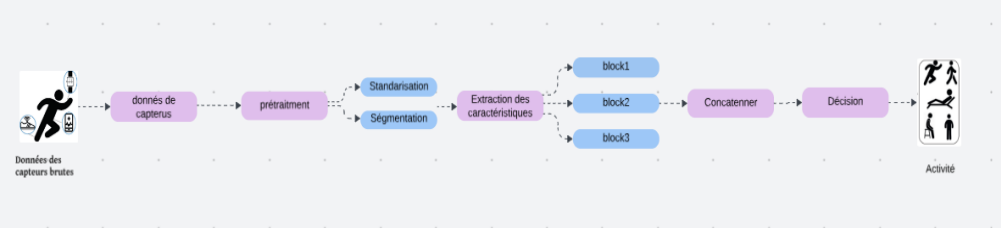


FIGURE 3.1 – Étapes du système HAR proposé

Le système proposée traite les données des capteurs prenant en compte que les données d'entrées sont brutes. Ensuite ces données passent par des prétraitements qui comprend la préparation, le nettoyage,la normalisation vers une échelle de [0, 1] et une segmentation qui divise les données en segments temporels pertinents dans le but de leur préparation à la phase d'extraction de caractéristiques. Cette phase traite les caractéristiques par plusieurs blocs de traitement, représentés par block1, block2 et block3 avec des tailles de filtres différentes (3, 5, et 7) pour capturer des caractéristiques à différentes échelles temporelles. La sortie de ces block sont ensuite concaténer pour combiner les caractéristiques extraites des données de capteurs, améliorer leur représentation et faciliter la prise de décision ultérieure pour la reconnaissance d'activités. Chaque étape de ce modèle sera détaillée dans les sections à venir.

3.4 Phase d'acquisition

Comme abordé précédemment dans l'architecture des systèmes HAR,la phase d'acquisition englobe le processus de définition des activités cibles à reconnaître ainsi la configuration des l'appareils à utiliser dans la phase de collecte de données, basée sur les activités humaines cibles, ensuite la collection de données à partir de capteurs, de dispositifs portables ou d'autres dispositifs qui capturent des informations sur les mouvements et les actions de la personne. Enfin, l'annotation des données où se fait l'attribution de labels aux activités humaines en cours d'exécution.
Toutefois, dans notre travail, ce processus n'est pas concrètement mis en œuvre, car nous avons utilisé des données provenant des différents capteurs déjà disponibles publiquement.

3.5 Phase de prétraitement

Les données anormales des capteurs telles que le bruit et les valeurs manquantes seront inévitablement générées pendant l'acquisition des données en raison de l'environnement complexe ou du taux d'échantillonnage instable [2].

Cette phase critique vise à préparer les données de manière à maximiser l'efficacité des analyses subséquentes et la précision des modèles développés. Il existe diverses techniques de pré-traitement dépendamment du type de données brutes des capteurs.

Dans notre cas,les pré-traitement auxquels se restreint notre étude sont : la normalisation, le calcul des magnitudes, le Ré-échantillonnage et la segmentation

pour adapter notre approche aux caractéristiques uniques de nos datasets à présenter dans le prochain chapitre, avec l'objectif de tirer le meilleur parti des données disponibles. Voici un résumé des étapes spécifiques du pré-traitement :

• **Normalisation :** Cette étape a impliqué trois procédures principales : la conversion des valeurs catégorielles, gérer les valeurs manquantes et la standardisation des caractéristiques numériques :

— la conversion des données catégorielles, notamment la transformation des états des capteurs tels que "ON" et "OFF" en valeurs numériques. Nous avons utilisé une approche directe et explicite dans le code de pré-traitement. Cette conversion manuelle a permis de passer de valeurs textuelles à des représentations numériques.

— dans le traitement des valeurs manquantes, plusieurs techniques ont été évaluées pour gérer les valeurs manquantes. Ici nous avons appliqué l'interpolation linéaire pour estimer les valeurs manquantes (NaN - Not a Number). L'interpolation linéaire est une technique précieuse dans le domaine du traitement des données séquentielles, car elle permet de remplir les valeurs manquantes de manière efficace et de maintenir la cohérence des séries de données au fil du temps ou de l'espace. Cela facilite une analyse plus complète et précise des données pour diverses applications.

— une fois le nettoyage effectué, nous pouvons passer à la standardisation de nos données. Nous avons employé la méthode StandardScaler de la bibliothèque scikitlearn qui réajuste les données pour qu'elles présentent une moyenne de zéro et un écart-type de un.

• **Calcul des magnitudes :** La magnitude est calculée en combinant les valeurs des axes x, y et z d'un capteur à trois dimensions dans le but d'améliorer la représentation des données issues des capteurs. La magnitude fournit une mesure supplémentaire qui peut capturer des informations pertinentes sur l'intensité ou la variation des mouvements enregistrés, ce qui n'est pas toujours évident à partir des caractéristiques individuelles sur chaque axe de capteurs. La magnitude est calculée en utilisant la formule suivante :

$$M = \sqrt{x^2 + y^2 + z^2} \tag{3.1}$$

où :

— **M** : Représente la magnitude du vecteur tridimensionnel. La magnitude est une mesure de la longueur ou de l'intensité globale du vecteur dans l'espace

3D

— **x, y, z** : Représentent les composantes du vecteur le long des trois axes cartésiens (axes x,y et z). Ces composantes sont les valeurs mesurées par un capteur dans chacune des trois directions.

• **Ré-échantillonnage :** En augmentant, ou en diminuant le nombre d'échantillons, les résultats peuvent être améliorés. La Figure 3.2 montre la distribution du nombre d'échantillons en fonction des activités. On note un déséquilibre significatif entre les différentes classes. Le déséquilibre évident entre les différentes classes peut entraîner un biais dans les modèles de DL. Dans cette example, les activités très représentées comme "Walking" et "Jogging" pourraient dominer les prédictions au détriment des activités moins fréquentes comme "Sitting" et "Standing".

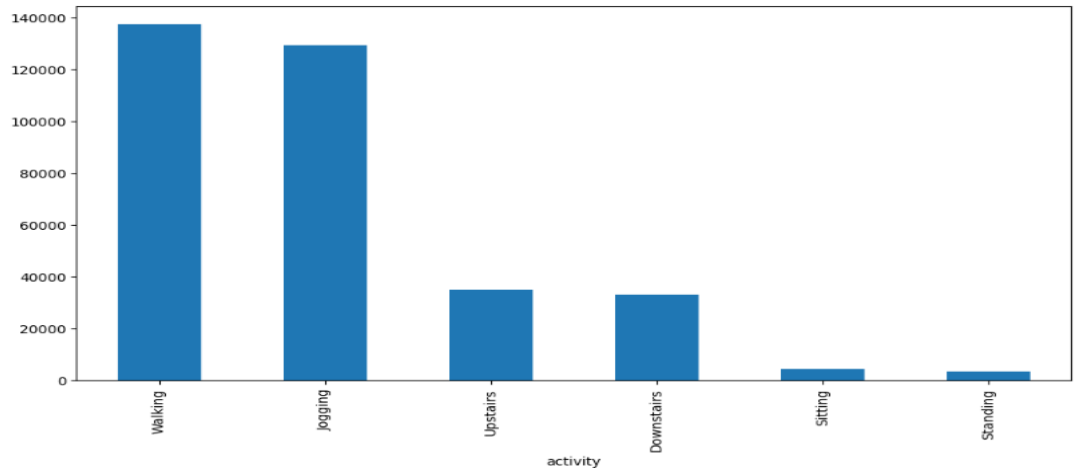


FIGURE 3.2 – Distribution du nombre d'échantillons en fonction des activité dans dataset WISDM.

Dans notre cas, l'augmentation artificielle du nombre d'exemples d'apprentissage est appliqué pour éviter que le modèle ne soit biaisé en faveur de la classe majoritaire, afin d'améliorer les performances globales du modèle.

• **Segmentation :** La segmentation fait référence à une technique qui consiste à diviser un ensemble de données en segments plus petits, afin de faciliter l'analyse ou l'application de techniques spécifiques sur chaque segment par exemple, pour appliquer des algorithmes ML et DL sur des segments de données. Dans notre cas, les données sont segmentés en fenêtres de taille fixe (voir la figure 3.3) et récupérer l'étiquette la plus fréquente dans chaque fenêtre. Elle prépare également les données pour l'entraînement ou le test de modèles de reconnaissance d'activité.

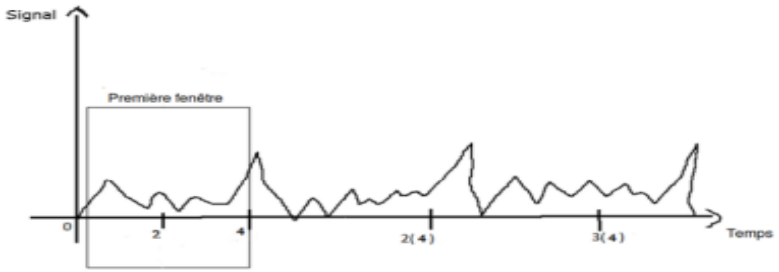


FIGURE 3.3 – Segmentation en fenêtres de taille fixe.

Par la suite, l'ensemble des données obtenues a été divisé en trois ensembles, où 70% ont été sélectionnés pour générer les données de formation destinée à l'entraînement du modèle, 10% pour les données de validation pour évaluer le modèle pendant son entraînement, ajuster les hyperparamètres du modèle et aide à surveiller le phénomène de surapprentissage (overfitting). Enfin, 20% des données sont utilisées pour constituer l'ensemble de test. Cet ensemble permettra de tester et d'évaluer les performances du modèle, offrant ainsi une estimation réaliste de son efficacité sur des données nouvelles et non vues. Cela détermine l'efficacité du modèle pour une utilisation pratique.

3.5.1 Phase d'extraction des caractéristiques

L'extraction de caractéristiques (feature extraction) fait référence au processus par lequel des informations significatives et discriminantes sont extraites des données d'entrée brutes. Cette phase est cruciale car elle prépare les données de manière à ce que le modèle puisse les utiliser efficacement pour la tâche spécifique à laquelle il est destiné, comme la classification. Comme nous l'avons déjà mentionné précédemment, dans cette étude nous avons utilisé un CNN et un GRU. L'extraction de caractéristiques implique l'utilisation d'une méthode de convolution multi-échelle, qui comprend trois blocs principaux d'extraction de caractéristiques. Dans ce qui suit, nous représentons les éléments utiles et étapes nécessaires dans l'approche donnée :

- **Structure des Blocs d'Extraction de Caractéristiques :** Il y a trois blocs d'extraction de caractéristiques principaux, chaque bloc est similaire en structure mais diffère par la taille du noyau de convolution utilisé : 3, 5 et 7.

- **Composition de chaque Bloc :** Chaque bloc d'extraction de caractéristiques comprend : deux couches convolutionnelles, avec des filtres de tailles spécifiques (3, 5, ou 7), une couche de pooling max pour réduire la dimensionnalité après la

première couche convolutionnelle, une couche de normalisation par lot (Batch Normalization) pour accélérer l'entraînement et réduire le surajustement, une couche GRU pour capturer les caractéristiques spatio-temporelles dans les données et une couche Dropout pour appliquer une régularisation en désactivant aléatoirement un pourcentage de neurones pour éviter le surapprentissage.

3.5.2 Phase Concaténation des Caractéristiques et de décision

Une fois que chaque canal a produit sa propre représentation des caractéristiques, les sorties finales de ces canaux sont concaténées (voir la figure 3.4). La concaténation combine les caractéristiques des trois canaux : canal avec les filtres de taille 3 pour extraire des caractéristiques locales à court terme, canal avec les filtres de taille 5 permettant de capturer des motifs légèrement plus étendus dans la séquence temporelle et un filtre de taille 7 idéal pour extraire des caractéristiques qui nécessitent une analyse à plus long terme et pour capturer des tendances sur une échelle de temps plus étendue pour former une représentation globale et riche des données. Après la concaténation, les caractéristiques combinées passent par une couche de normalisation par lots (Batch Normalization, BN) pour accélérer l'entraînement et réduire le surajustement. Enfin, les caractéristiques réduites sont envoyées à une couche dense avec une activation Softmax pour produire la sortie de classification.

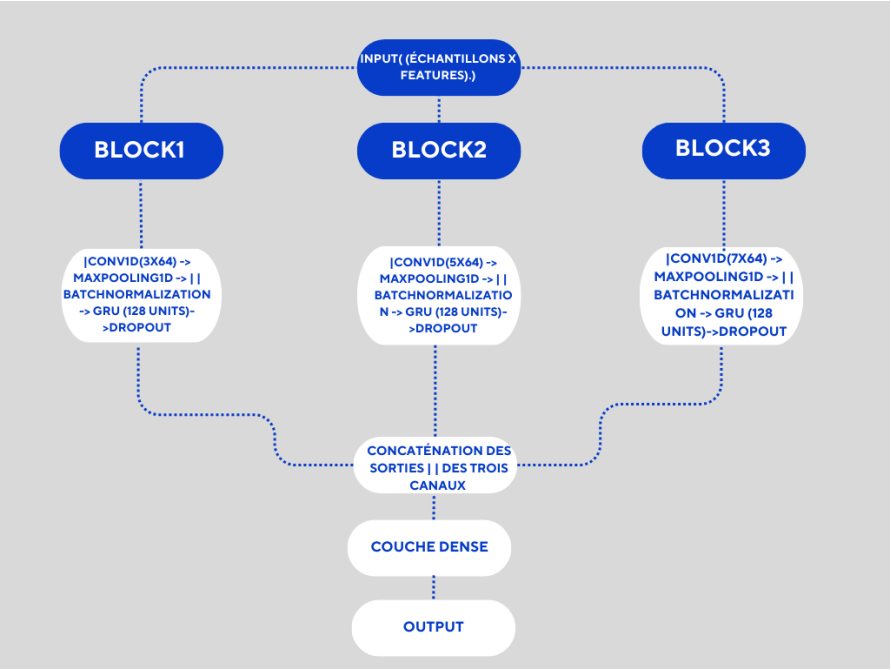


FIGURE 3.4 – Schéma de la fusion des caractéristiques extraites.

3.6 Conclusion

Dans ce chapitre, nous avons présenter d'une manière détaillée le système d'extraction de caractéristiques proposé pour notre modèle de reconnaissance d'activités humaines. Chaque étape du processus a été explicitée, notamment l'utilisation de convolutions unidimensionnelles dans les différents canaux du modèle. Nous avons mis en lumière l'approche de fusion des caractéristiques extraites à partir des filtres de tailles variées, ce qui est crucial pour la classification précise des activités humaines dans notre modèle HAR.

Le chapitre suivant porte sur l'expérimentation et la validation du système proposé sur deux datasets.

Chapitre 4

Expérimentation et validation du système HAR proposé

4.1 Introduction

Ce chapitre est dédié à la mise en œuvre, l'expérimentation et la validation d'un système HAR basé sur CNN et GRU. Tout d'abord, nous fournissons un aperçu de l'environnement de développement, y compris le matériel et les logiciels utilisés pour mettre en œuvre le système. Après, nous détaillons les ensembles de données utilisés pour entraîner et évaluer le modèle. Nous présentons ensuite les résultats obtenus, les comparons avec les approches existantes et discutons des avantages de notre approche hybride. Enfin, nous concluons ce chapitre par une analyse des performances du modèle et des perspectives pour les travaux futurs.

4.2 Environnement de développement

Pour implémenter le système HAR, nous avons utilisé plusieurs outils dont le langage Python adopté dans la plateforme Google Colab. Cette section présente brièvement les bibliothèques de l'environnement Python et le matériel utilisés dans ce travail.

Langage Python

Python est un langage de programmation interprété, interactif et orienté objet créé par Guido Van Rossum. Il compte plusieurs modules et permet la manipulation des types de données de façon dynamique. Autre que l'orienté objet, ce langage de programmation permet l'exécution de code procédural ou fonctionnel. Python est portable sur de multiples systèmes d'exploitation, cela inclut Linux, Windows, et MacOS [108].

Différentes bibliothèques Python ont été employées afin de mettre en place le système proposé. Nous en citons les plus sollicitées :

- **Tenserflow :** TensorFlow est l'une des bibliothèques les plus importantes ; il s'agit d'une bibliothèque de deep learning open-source créée par l'équipe "Google Brain". Elle fonctionne avec les CNN, les RNN et les GAN, qui sont tous des modèles de deep learning [109].

- **Scikit-learn :** également connue sous le nom de sklearn, est une bibliothèque open-source de modélisation de données et d'apprentissage automatique pour Python. Elle propose divers algorithmes de classification, de régression et de regroupement. Scikit-learn est conçu pour interagir avec les bibliothèques Python NumPy et SciPy [110].

- **Numpy :** NumPy est une bibliothèque pour langage de programmation Python, destinée à manipuler des matrices ou tableaux multidimensionnels ainsi que

des fonctions mathématiques opérant sur ces tableaux [111].

- **Keras :** une API de deep learning basée sur Python, a été créée par François Chollet, ingénieur logiciel chez Google et chercheur en intelligence artificielle. Elle utilise JAX, TensorFlow ou PyTorch. Keras gère le traitement des données, l'ajustement des hyperparamètres et le déploiement dans le flux de travail de machine learning pour accélérer les expérimentations et simplifier la construction et l'entraînement des modèles de deep learning [109].

Google Colab

Google Colab ou Colaboratory est un service cloud, offert par Google (gratuit), basé sur Jupyter Notebook et destiné à la formation et à la recherche dans l'apprentissage automatique. Cette plate-forme permet d'entraîner des modèles de Machine Learning directement dans le cloud. Sans donc avoir besoin d'installer quoi que ce soit sur notre ordinateur à l'exception d'un navigateur [112]. La figure 4.1 illustre les caractéristiques matérielles de Google Colab. Pour notre situation, nous faisons appel à un processeur CPU.

Parameter	Specification
GPU Model Name	Nvidia K80
GPU Memory	12 GB
GPU Memory Clock	0.82 GHz
GPU Performance	4.1 TFLOPS
CPU Model Name	Intel(R) Xeon(R)
CPU Frequency	2.30 GHz
Number of CPU Cores	2
Available RAM	12 GB
Disk Space	25 GB

FIGURE 4.1 – Spécifications matérielles de Google Colab [30]

4.3 Ensembles de données

Dans ce travail, nous avons choisi d'expérimenter le système implémenté sur deux ensembles de données disponibles publiquement et connus dans la communauté de HAR à savoir WISDM et PAMAP2. L'utilisation de deux datasets permet de consolider les résultast obrenus.

Dataset WISDM

Le dataset Wireless Sensor Data Mining (WISDM) est un ensemble de données de référence pour la reconnaissance d'activités humaines, dérivé du laboratoire Wireless Sensor Data Mining, et contient un total de 1 098 207 échantillons. Il s'agit d'un ensemble d'activités collectées auprès de 36 utilisateurs effectuant des activités quotidiennes, y compris les six comportements suivants : marche, position assise, jogging, descente d'escaliers, montée d'escaliers et position debout. Ces données ont été obtenues par les utilisateurs expérimentaux avec un téléphone Android dans la poche avant de leur pantalon, en utilisant le capteur accéléromètre intégré du téléphone avec une fréquence d'échantillonnage de 20 Hz [113]. La figure 4.2 est une illustration des captures des signaux d'accéléromètre sur les axes x, y, z de quelques activités tirées de WISDM.

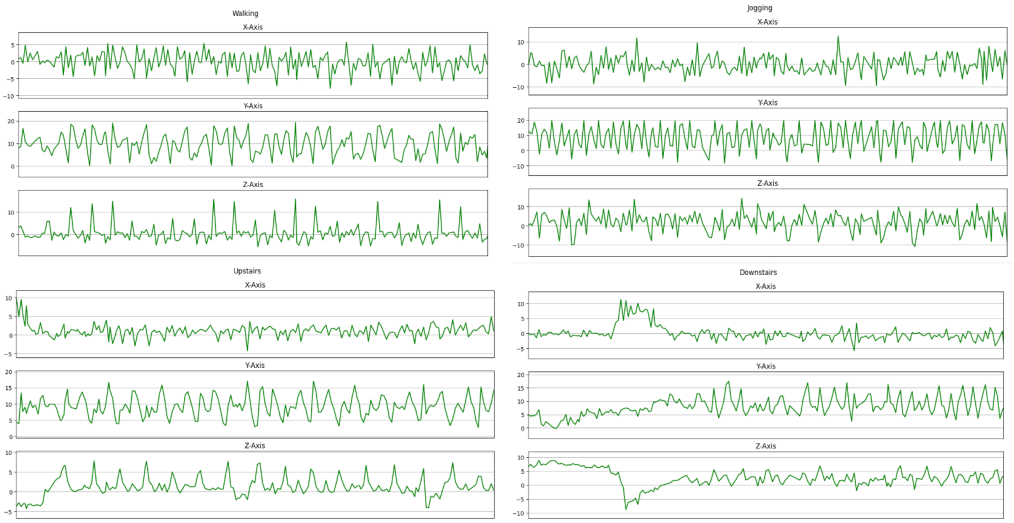


FIGURE 4.2 – Signaux d'accéléromètre sur les axes x, y, z pris de WISDM

Dataset PAMAP2

Le Physical Activity Monitoring and Assessment Dataset (PAMAP2) recueille diverses activités auprès de 9 volontaires (1 femme, 8 hommes), comprenant 12 activités protocolaires (allongé, assis, debout, marche, course à pied, cyclisme, marche nordique, repassage, nettoyage à l'aspirateur, saut à la corde, montée et descente des escaliers) et 6 activités facultatives (regarder la télévision, travailler sur ordinateur, conduire une voiture, plier le linge, nettoyer la maison, jouer au football). Les données d'activité ont été enregistrées par des capteurs IMU (unités de mesure inertielle) installés à différentes positions du corps humain (main,

poitrine et cheville). Un total de 52 caractéristiques ont été capturées à une fréquence d'échantillonnage de 100 Hz [114]. La disposition des IMU sur le dispositif PAMAP2 est illustrée dans la figure 4.3.

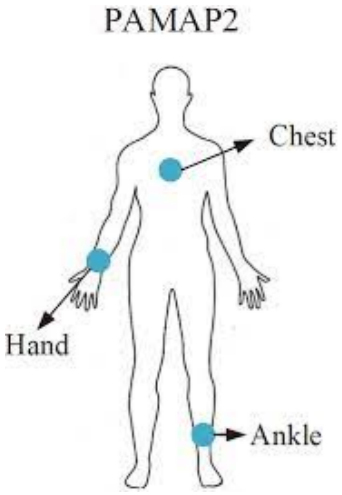


FIGURE 4.3 – Disposition des IMU sur le dispositif PAMAP2 [31]

4.4 Tests et discussion des résultats

Dans cette sous-section, nous examinons en détail la performance des modèles d'apprentissage automatique que nous avons appliqués. Nos objectifs initiaux consistaient à évaluer l'efficacité des modèles classiques d'apprentissage automatique dans divers contextes et à comparer leurs performances avec celles des techniques de deep learning. Nous avons commencé par développer un programme de référence et, par la suite, nous avons amélioré ce programme en effectuant divers tests et en explorant d'autres approches. En particulier, nous avons cherché à comprendre le comportement de ces modèles dans des scénarios de reconnaissance d'activités humaines.
Nous discuterons également les points forts et les limitations observés, ainsi que les anomalies ou résultats inattendus et leur impact potentiel sur notre solution. Cette analyse nous permettra de tirer des conclusions éclairées sur l'efficacité des modèles d'apprentissage automatique dans divers contextes et de proposer des recommandations pour les recherches futures.

4.4.1 Critères d'évaluation

Afin d'évaluer notre système, nous avons utilisé divers critères de performance d'apprentissage mentionnés dans le chapitre 2, notamment l'exactitude (accuracy), la précision (precision) et le score F1 (F1 score).

- **Accuracy (ACC) :** L'exactitude est le rapport du nombre de prédictions correctes au nombre total de prédictions effectuées.
- **Précision :** La précision est le rapport du nombre de vraies prédictions positives au nombre total de prédictions positives (vraies positives et fausses positives).
- **F1-score :** Le score F1 est la moyenne harmonique de la précision et du rappel (recall). Il offre un équilibre entre les deux métriques, surtout dans les situations où les classes sont déséquilibrées.

4.4.2 Présentation des tests sur le jeu de données WISDM

Nous allons expliquer graduellement les améliorations apportées soit au dataset soit à la méthode proposée.

Modèle Référentiel

En suivant la méthode décrite dans l'article [115], nous avons implémenté plusieurs étapes clés pour le traitement des données avant de les soumettre à l'algorithme CNN-GRU. Les données des capteurs continus ont été segmentées en fenêtres de taille 128 avec un chevauchement de 50%. Toutes les caractéristiques fournies par ces capteurs ont été extraites. Ensuite, les auteurs ont divisé les données en deux ensembles (ensemble d'apprentissage, ensemble test), en tenant compte des utilisateurs, ils ont sélectionné les données des 30 premiers utilisateurs comme ensemble d'apprentissage et celles des 6 utilisateurs suivants comme ensemble de test. L'application de cet méthode a donné une exactitude de 78% sur l'ensemble d'apprentissage et 53% sur l'ensemble de test.

Ajout des magnitudes des accélérations et celles des vitesses angulaires

En contraste avec l'approche précédente, nous avons d'abord enrichi le processus en intégrant le CNN-GRU standard avec l'ajout de la magnitude comme caractéristique supplémentaire. Cette modification a considérablement amélioré l'exactitude, la portant à 97% sur l'ensemble d'apprentissage et 80% sur l'ensemble de test.

Équilibrage des classes

Ensuite, nous avons appliqué la méthode de rééchantillonnage sur le CNN-GRU standard. Cette technique, qui concerne le dataset et non la méthode elle-même, consiste à sous-échantillonner les données pour équilibrer les classes par rapport à la classe majoritaire, maximisant ainsi la pertinence des données pour l'apprentissage. Cette adaptation a significativement augmenté le taux de précision, atteignant 99% sur l'ensemble d'apprentissage et 95% sur l'ensemble de test.

Équilibrage des classes et ajout des magnitudes

Bien que les résultats obtenus sont intéressants, nous souhaitons améliorer la capacité du modèle à reconnaître efficacement les activités humaines sur des données non vues. En vue de cela, nous allons créer un nouveau modèle intégrant ces deux techniques, c'est-à-dire le calcul de la magnitude comme caractéristique et l'application du rééchantillonnage. Cette approche vise à élever encore davantage la précision sur les données de test, renforçant ainsi la robustesse du système dans le contexte HAR, atteignant 99% sur l'ensemble d'apprentissage et 98% sur l'ensemble de test.

Le tableau 4.1 montre tous les résultats de toutes les métriques obtenues

Étapes	**Entraînement**			**ACC**
	ACC	Précision	F1-Score	**Test**
Référentiel	0.78	0.50	0.40	0.52
Référentiel + Magnitude	0.97	0.92	0.88	0.80
Référentiel + Augmentation	0.99	0.90	0.96	0.95
CNN+GRU(Magnitude+Augmentation)	0.99	0.99	0.99	0.98

TABLE 4.1 – Récapitulation des résultats obtenus sur le dataset WISDM

Les résultats obtenus révèlent l'impact significatif des ajustements techniques, tels que l'incorporation de la magnitude comme caractéristique supplémentaire et l'application du rééchantillonnage, sur l'amélioration des performances du modèle. Ces modifications ont non seulement augmenté l'exactitude et la précision du modèle lors de tests sur des ensembles d'apprentissage et de test, mais ont également démontré sa capacité à maintenir une haute précision sur des données non vues (voir la figure 4.4).

Étant donné l'exactitude élevée de modèle CNN-GRU avec équilibrage des classes et ajout des magnitudes des accélérations et celles des vitesses angulaires , avec une précision de 98% sur l'ensemble d'entraînement et de 97% sur l'ensemble de test, nous pouvons conclure que le modèle ne présente pas de sous-apprentissage

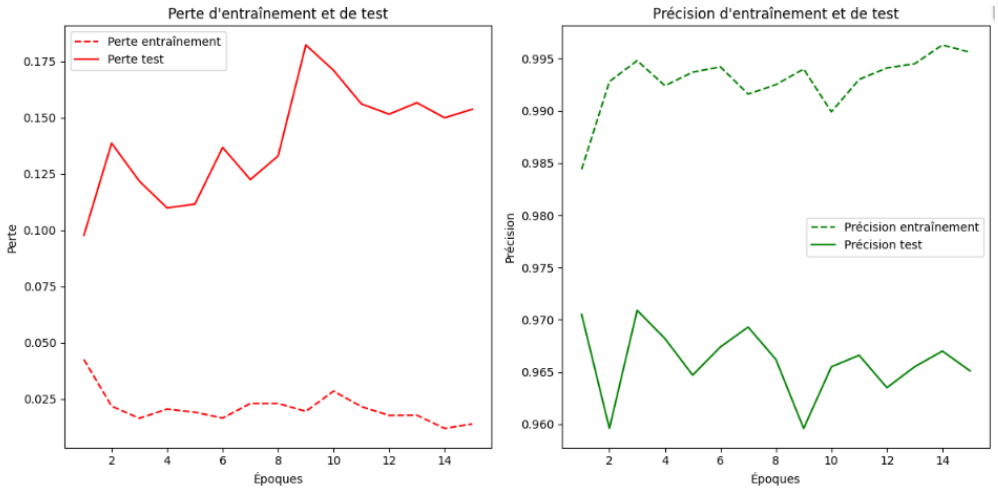


FIGURE 4.4 – Graphe représentant le taux d'exactitude

(underfitting) ou de surapprentissage (overfitting). Ces deux phénomènes seront expliqués ci-dessous.

En observant le graphe de la figure 4.4, nous remarquons que les courbes de précision pour l'entraînement et le test sont proches, de même pour les pertes d'entraînement et de test. Cela indique que l'apprentissage s'est bien déroulé. En cas de surapprentissage, nous aurions observé une grande différence entre les performances sur les ensembles d'entraînement et de test, avec une performance très élevée sur l'entraînement et beaucoup plus faible sur le test. En cas de sous-apprentissage, le modèle aurait eu des performances médiocres sur les deux ensembles, indiquant qu'il n'a pas bien appris les motifs des données.

Afin d'approfondir notre analyse des résultats obtenus, nous avons utilisé un histogramme (voir la figure 4.5) pour représenter la distribution des données. Un histogramme utilise des barres verticales pour montrer la fréquence des valeurs dans différents intervalles (bins). Chaque barre représente un intervalle de valeurs, et la hauteur de la barre indique le nombre de valeurs de l'ensemble de données qui tombent dans cet intervalle .

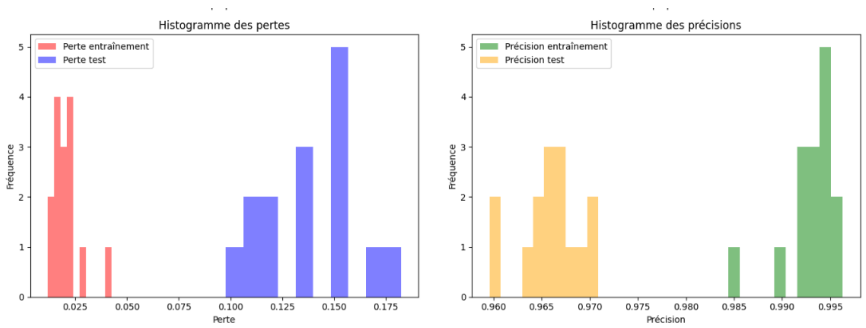


FIGURE 4.5 – Histogramme représentant les résultats obtenue

La matrice de confusion obtenue pour le modèle CNN-GRU appliqué au jeu de données WISDM, illustrée à la figure 4.6, indique une prédominance de classifications correctes par le modèle. Par exemple, l'activité 'UPSTAIRS' présente 608 instances correctement identifiées (vrais positifs) contre 30 instances incorrectement classées dans d'autres activités.

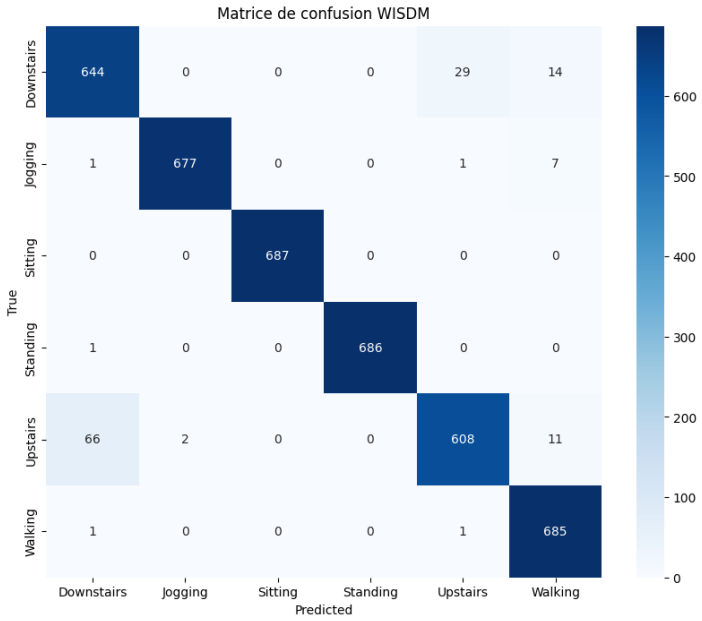


FIGURE 4.6 – Matrice de confusion du meilleur résultat de CNN-GRU sur WISDM

Les résultats significatifs obtenus avec le modèle sur le jeu de données WISDM, qui contient 6 activités à classer, nous ont incités à approfondir notre étude. Nous avons ainsi décidé de tester le modèle sur un autre jeu de données, PAMAP2, qui comporte 18 activités à classer, afin d'évaluer sa robustesse et sa capacité à généraliser à des ensembles de données plus complexes et variés.

4.4.3 Présentation des tests sur le jeu de données PAMAP2

Pour le dataset PAMAP2, nous allons suivre la même démarche que pour le dataset WISDM.

Modèle Référentiel

Dans l'évaluation des performances sur le jeu de données PAMAP2, le modèle CNN-GRU a été déployé avec ses paramètres par défaut, sans modifications ni optimisations particulières en suivant la méthode décrite dans l'article [115]. Les données des capteurs continus ont été segmentées en fenêtres de taille 128 avec un chevauchement de 50%. Toutes les caractéristiques fournies par ces capteurs ont été extraites. Ensuite, les auteurs ont été divisées les données en deux ensembles (ensemble d'apprentissage, ensemble test), en tenant compte des utilisateurs, ils ont sélectionné les utilisateurs 1, 2, 3, 4, 5, 8 et 9 comme ensemble d'apprentissage, et 6 et 7 comme ensemble de test. Cette approche initiale a produit une exactitude de 99% sur l'ensemble d'apprentissage, indiquant une performance élevée dès le départ sur cet ensemble, et 84% sur l'ensemble de test.

Cependant, malgré la précision élevée sur les données d'apprentissage, ce résultat révèle un problème potentiel. La précision est beaucoup plus faible sur les données de test (84%) suggère que le modèle a bien appris sur l'ensemble d'apprentissage mais manque de généralisation sur les données non vues, ce qui est crucial dans le domaine de la reconnaissance d'activités humaines.

Ajout de magnitude

Pour remédier à cela, nous avons décidé d'appliquer les mêmes étapes d'optimisation que celles utilisées pour le jeu de données WISDM. Ces étapes incluent l'ajout des magnitudes comme caractéristiques supplémentaires. Cette modification a apporté une légère amélioration, portant l'exactitude sur les données de test à 85%.

Équilibrage des classes

Nous avons appliqué la méthode de rééchantillonnage sur le CNN-GRU standard. Cette technique consiste à sous-échantillonner les données pour équilibrer les classes par rapport à la classe majoritaire, maximisant ainsi la pertinence des données pour l'apprentissage. Cette adaptation a significativement augmenté le taux de précision, atteignant 94% sur l'ensemble d'apprentissage et 94% sur l'ensemble de test.

Équilibrage des classes et ajout des magnitudes

Bien que les résultats obtenus soient intéressants, nous cherchons à améliorer la capacité du modèle à reconnaître efficacement les activités humaines sur des données inédites. Pour ce faire, nous allons développer un nouveau modèle intégrant deux techniques : le calcul de la magnitude comme caractéristique et l'application du rééchantillonnage. Cette approche vise à augmenter encore davantage la précision sur les données de test, renforçant ainsi la robustesse du système dans le contexte HAR, avec des taux de précision de 98% sur l'ensemble d'apprentissage et de 96% sur l'ensemble de test.

Les résultats obtenus sont présentés dans le tableau 4.2 :

Étapes	**Référentiel**	**Référentiel+ Magnitude**	**Référentiel+ Augmentation**	**CNN+GRU (Magnitude+Augmentation)**
ACC-Entraînement	0.99	0.99	0.94	0.98
présicion-Entraînement	0.88	0.87	0.94	0.96
F1-score-Entraînement	0.87	0.83	0.94	0.96
Perte sur donnes test	0.64	0.54	0.21	0.12
ACC-Test	0.84	0.85	0.94	0.96

TABLE 4.2 – Récapitulation des résultats appliqués sur le dataset PAMAP2

Ces résultats montrent que les ajustements apportés, notamment l'ajout de la magnitude et l'application du rééchantillonnage, ont considérablement amélioré la capacité de notre modèle à généraliser sur des données non vues (voir la figure 4.7). Le modèle optimisé a montré une performance robuste et fiable sur le dataset PAMAP, qui est plus complexe avec ses 18 activités à classer.

Pour approfondir notre analyse des résultats obtenus avec PAMAP, nous avons tracé un histogramme présenté à la figure 4.8 pour illustrer la distribution des données.

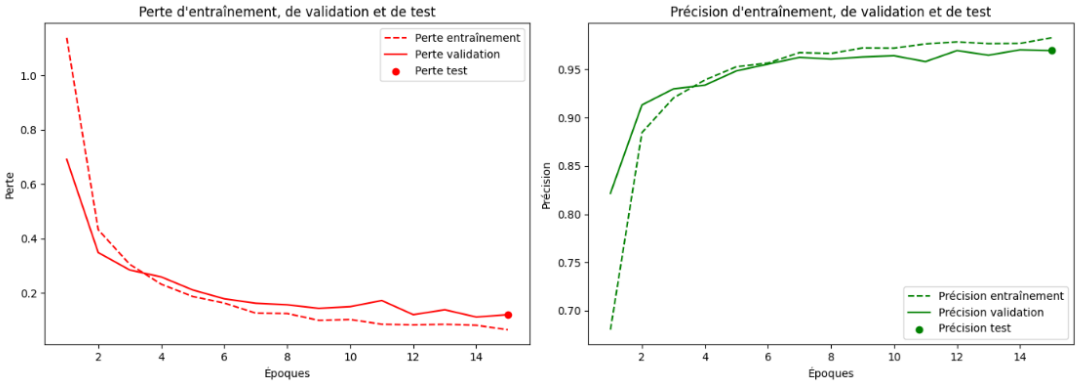


FIGURE 4.7 – graphe représentant les résultats obtenue

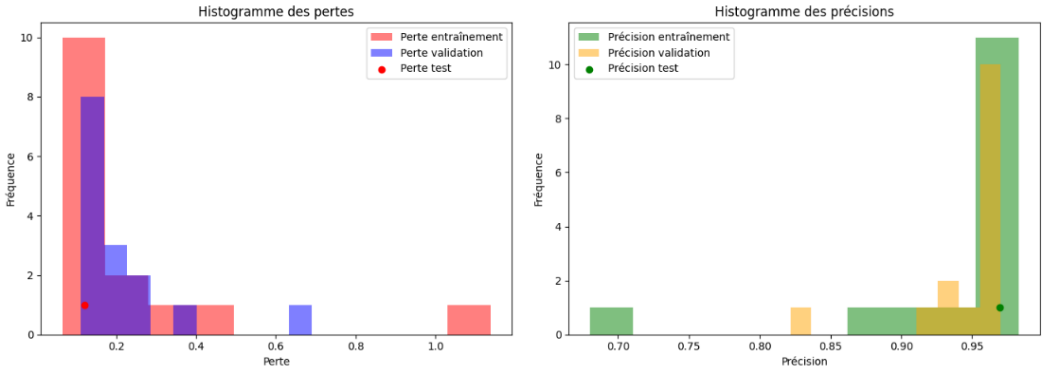


FIGURE 4.8 – graphe représentant les résultats obtenue

La matrice de confusion obtenue pour le modèle CNN-GRU appliqué au jeu de données PAMAP2, illustrée à la figure 4.9 , montre une forte proportion de classifications correctes par le modèle. Par exemple, l'activité 'STANDING' compte 181 instances correctement identifiées (vrais positifs) contre 5 instances mal classées dans d'autres activités.

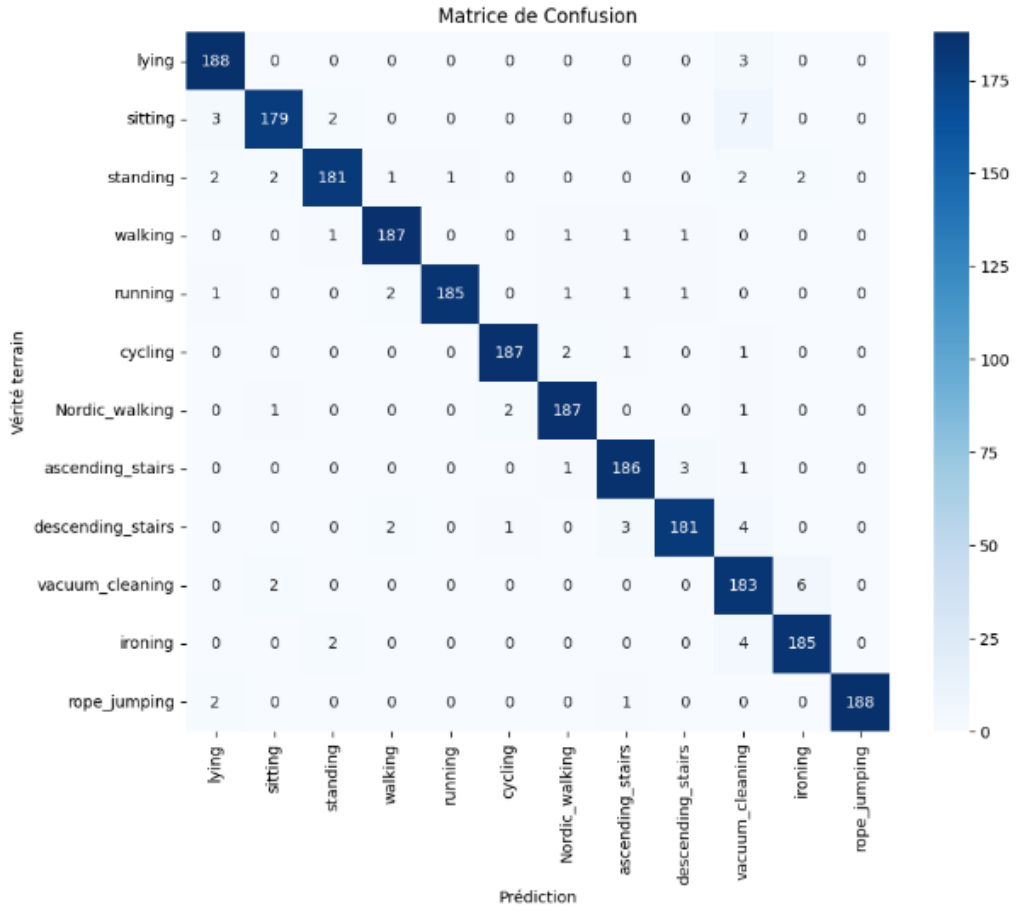


FIGURE 4.9 – Matrice de confusion du meilleur résultat de CNN-GRU sur PAMAP2

4.5 Perspectives futures

Dans le domaine de la reconnaissance des activités humaines, nous nous orientons vers des défis stimulants et des opportunités d'innovation cruciales. Jusqu'à présent, nos recherches se sont concentrées sur des activités spécifiques et bien définies. Cependant, pour répondre aux exigences des applications réelles et variées, il est essentiel de renforcer nos efforts vers des activités plus complexes et d'améliorer notre capacité de généralisation.
Une avenue prometteuse est l'exploration de la reconnaissance des activités humaines impliquant des interactions précises avec des objets. Par exemple, il devient crucial de détecter les transitions subtiles, telles que passer de la position assise à une interaction active avec une machine, comme le fait de taper sur un

clavier. Cela permettrait une classification plus précise des activités en identifiant non seulement ce que fait une personne, mais aussi comment elle interagit avec son environnement.
Pour relever ce défi, il sera nécessaire de développer des modèles HAR capables de détecter et de comprendre une diversité étendue de comportements humains dans des contextes dynamiques et variés. Cela inclut les interactions sociales complexes, les environnements non structurés, et les mouvements non conventionnels. Ces modèles devront maintenir des performances élevées en termes de précision et de résilience, adaptées à des scénarios réels où les conditions peuvent varier considérablement.

4.6 Conclusion

Ce chapitre est dédié à la mise en œuvre et à la validation d'un système hybride de reconnaissance de l'activité humaine basé sur CNN et GRU.
Il est crucial d'utiliser des méthodes avancées comme l'utilisation de caractéristiques de la magnitude et l'augmentation des données afin d'améliorer les performances des systèmes de reconnaissance de l'activité humaine.
Selon les résultats, l'emploi de ces méthodes permet d'améliorer considérablement les performances, en particulier sur des ensembles de données complexes comme PAMAP2. Ces résultats peuvent servir de guide pour le développement futur de systèmes HAR plus robustes et plus précis.

Conclusion générale

Le but de cette étude est de mettre en œuvre un système de reconnaissance de l'activité humaine basé sur des capteurs utilisant un modèle hybride de réseaux neuronaux CNN et GRU.

Nous avons commencé ce travail en discutant des aspects généraux de la reconnaissance de l'activité humaine et de diverses techniques d'apprentissage profond. L'architecture globale du système HAR comprend plusieurs modules tels que le module d'acquisition , le prétraitement, l'extraction de caractéristiques, concaténation des caractéristiques, et de décision. Nous avons également exploré l'utilisation de modèles hybrides de réseaux de neurones CNN et GRU pour améliorer la précision et l'efficacité du système HAR.

Après avoir réalisé un état de l'art, nous avons constaté que les approches de HAR peuvent être basées sur la vision ou sur des capteurs. Après une analyse comparative approfondie, nous avons opté pour une approche par capteurs. L'une des principales raisons de ce choix est la confidentialité, étant donné que les capteurs sont moins intrusifs que les systèmes basés sur la vision.

Lors de la mise en œuvre du système, différentes techniques ont été envisagées à différentes étapes du processus HAR, telles que la magnitude et l'augmentation des données. Les résultats des tests expérimentaux effectués sur les bases de données WISDM et PAMAP2 ont montré que la magnitude et l'augmentation des données peuvent être utilisées en combinaison pour améliorer considérablement les performances par rapport à une utilisation individuelle.

De plus, afin de généraliser nos résultats et d'observer un éventail d'activités plus large, nous avons étendu notre analyse à l'aide de deux bases de données distinctes. La base de données WISDM ne contient en réalité que 6 activités, tandis que la base de données PAMAP2 contient 18 activités, nous permettant d'explorer des scénarios d'activités plus diversifiés.

Nous prévoyons l'exploration de la reconnaissance des activités humaines impliquant des interactions précises avec des objets dans nos futurs travaux.

Bibliographie

[1] J. Llinas M. Liggins II, D. Hall. Handbook of multisensor data fusion : Theory and practice. 2017.

[2] Florenc Demrozi, Cristian Turetta, Fadi Al Machot, Graziano Pravadelli, and Philipp H Kindt. A comprehensive review of automated data annotation techniques in human activity recognition. *arXiv preprint arXiv :2307.05988*, 2023.

[3] Abdul Wahid. Big data and machine learning for Businesses. https://www.wordstream.com/wp-content/uploads/2021/07/machine-learning.png. . Consulté le : 08 mars 2024.

[4] Histoire du deep learning. https://www.natural-solutions.eu/blog/histoire-du-deep-learning. Consulté :Décembre 2023.

[5] Sai Balaji. Binary Image classifier CNN using TensorFlow. https://medium.com/techiepedia/binary-image-classifier-cnn-using-tensorflow-a3f5d6746697. Consulté le : 20 novembre 2023.

[6] Jonte Dancker. A brief introduction to recurrent neural networks. https://towardsdatascience.com/a-brief-introduction-to-recurrent-neural-networks-638f64a61ff4. Consulté le : 05 juillet 2024.

[7] « Auto-encodeurs en Deep Learning : tout savoir ». Formation Tech et Data en ligne | Blent.ai. https://blent.ai/blog/a/auto-encodeurs-deep-learning. Consulté le : 05 juillet 2024.

[8] Yuxuan Gu, Qixin Chen, Kai Liu, Le Xie, and Chongqing Kang. Gan-based model for residential load generation considering typical consumption patterns. In *2019 IEEE power & energy society innovative smart grid technologies conference (ISGT)*, pages 1–5. IEEE, 2019.

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[10] Yuhwan Kim, Chang-Ho Choi, Chang-Young Park, and Seonghyun Park. Examining recognition of occupants' cooking activity based on sound data using deep learning models. *Buildings*, 14(2) :515, 2024.

[11] Aziz Khelalef, Fakhreddine Ababsa, and Nabil Benoudjit. An efficient human activity recognition technique based on deep learning. *Pattern Recognition and Image Analysis*, 29 :702–715, 2019.

[12] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.

[13] Andrew Zisserman Karen Simonyan. Two-stream convolutional networks for action recognition in videos. *Nom du Journal*, 2021/2022.

[14] Shaurya Gupta, Dinesh Kumar Vishwakarma, and Nitin Kumar Puri. A human activity recognition framework in videos using segmented human subject focus. *The Visual Computer*, pages 1–17, 2024.

[15] Tej Singh and Dinesh Kumar Vishwakarma. A deeply coupled convnet for human activity recognition using dynamic and rgb images. *Neural Computing and Applications*, 33(1) :469–485, 2021.

[16] Samundra Deep and Xi Zheng. Leveraging cnn and transfer learning for vision-based human activity recognition. In *2019 29th international telecommunication networks and applications conference (ITNAC)*, pages 1–4. IEEE, 2019.

[17] Arati Kushwaha, Manish Khare, Reddy Mounika Bommisetty, and Ashish Khare. Human activity recognition based on video summarization and deep convolutional neural network. *The Computer Journal*, page bxae028, 2024.

[18] Magdalena Pawlyta, Marek Hermansa, Agnieszka Szczęsna, Mateusz Janiak, and Konrad Wojciechowski. Deep recurrent neural networks for human activity recognition during skiing. In *Man-Machine Interactions 6 : 6th International Conference on Man-Machine Interactions, ICMMI 2019, Cracow, Poland, October 2-3, 2019*, pages 136–145. Springer, 2020.

[19] Abdulmajid Murad and Jae-Young Pyun. Deep recurrent neural networks for human activity recognition. *international journal on the science and technology of sensors*, 2017.

[20] Saeed Mohsen. Recognition of human activity using gru deep learning algorithm. *Multimedia Tools and Applications*, 82(30) :47733–47749, 2023.

[21] BOUDJEMA ALI, TITOUNA FAIZA, HATTAB ABDESSALAM, and DJOUZI KHEYREDDINE. A light cnn architecture for human activity recognition based wearable sensors. 2023.

[22] Semin Ryu, Suyeon Yun, Sunghan Lee, and In cheol Jeong. Exploring the possibility of photoplethysmography-based human activity recognition using convolutional neural networks. *Sensors*, 24(5) :1610, 2024.

[23] Ravi Raj and Andrzej Kos. An improved human activity recognition technique based on convolutional neural network. *Scientific Reports*, 13(1) :22581, 2023.

[24] SU Park, JH Park, Mohammed A Al-Masni, Mugahed A Al-Antari, Md Z Uddin, and T-S Kim. A depth camera-based human activity recognition via deep learning recurrent neural network for health and social care services. *Procedia Computer Science*, 100 :78–84, 2016.

[25] LuKun Wang and RuYue Liu. Human activity recognition based on wearable sensor using hierarchical deep lstm networks. *Circuits, Systems, and Signal Processing*, 39(2) :837–856, 2020.

[26] Sakshi Ganesh Dighe et Ahay Gupta et Pranjal Gupta Akash Ghosh, Rahul Shakar. Integrating multimodal sensing and hybrid deep learning for enhanced human activity recognition. *International Journal of Recent Technology and Engineering (IJRTE)*, 2023.

[27] Abu Saleh Musa Miah Najmul Hassan and Jungpil Shin. A deep bidirectional lstm model enhanced by transfer-learning-based feature extraction for dynamic human activity recognition. *https ://www.mdpi.com/journal/applsci*, 2024.

[28] Abdel-Basset et Mohamed Reda Ding, Weiping. Har-deepconvlg : Hybrid deep learning-based model for human activity recognition in iot applications. *Information Sciences*, 646 :119394, 2023.

[29] Rita Rijayanti, Mintae Hwang, and Kyohong Jin. Detection of anomalous behavior of manufacturing workers using deep learning-based recognition of human–object interaction. *Applied Sciences*, 13(15) :8584, 2023.

[30] Zhandos Kegenbekov and Ilya Jackson. Adaptive supply chain : Demand–supply synchronization using deep reinforcement learning. *Algorithms*, 14(8) :240, 2021.

[31] Haojie Ma, Wenzhong Li, Xiao Zhang, Songcheng Gao, and Sanglu Lu. Attnsense : Multi-level attention mechanism for multimodal human activity recognition. In *IJCAI*, pages 3109–3115, 2019.

[32] Miguel A Labrador Oscar D Lara. A survey on human activity recognition using wearable sensors. *IEEE Communications Surveys and Tutorials, vol. 15, no. 3, pages 1192–1209*, 2013.

[33] Nassim Mokhtari. *Reconnaissance d'activités quotidiennes à partir de données capteurs pour l'assistance de vie dans un environnement connecté*. 2020.

[34] Saurabh Gupta. Deep learning based human activity recognition (har) using wearable sensor data. *www.elsevier.com/locate/jjime*, 2021.

[35] Adrian Hilton Thomas B. Moeslund and Volker K. A survey of advances in vision-based human motion capture and analysis. *in book :Computer Vision and Image Understanding, Vol. 104, no. 2-3, pp 90–126*, 2006.

[36] Hedi Tabia, Michèle Gouiffes, Lionel Lacassagne, and Bures sur Yvette. Reconnaissance des activités humaines à partir des vecteurs de mouvement quantifiés. *Institut d'Electronique Fondamentale d'Orsay, France*, 16(1), 2012.

[37] Mouna Selmi. *Reconnaissance d'activités humaines à partir de séquences vidéo*. PhD thesis, Institut National des Télécommunications, 2014.

[38] Oughlis Nadia. *Conception et réalisation d'un simulateur à base de SMA pour activités humaines*. PhD thesis, Université Abderrahmane Mira de Bejaia, 2020-2021.

[39] Yu-Liang Hsu, Shih-Chin Yang, Hsing-Cheng Chang, and Hung-Che Lai. Human daily and sport activity recognition using a wearable inertial sensor network. *IEEE Access*, 6 :31715–31728, 2018.

[40] Sana Alazwari, Majdy M Eltahir, Nabil Sharaf Almalki, Abdulrahman Alzahrani, Mrim M Alnfiai, and Ahmed S Salama. Improved coyote optimization algorithm and deep learning driven activity recognition in healthcare. *IEEE Access*, 2024.

[41] Kristina Host, Miran Pobar, and Marina Ivasic-Kos. Analysis of movement and activities of handball players using deep neural networks. *Journal of imaging*, 9(4) :80, 2023.

[42] Han Shi, Hai Zhao, Yang Liu, Wei Gao, and Sheng-Chang Dou. Systematic analysis of a military wearable device based on a multi-level fusion framework : research directions. *Sensors*, 19(12) :2651, 2019.

[43] Saurabh Gupta. Deep learning based human activity recognition (har) using wearable sensor data. *International Journal of Information Management Data Insights*, 1(2) :100046, 2021.

[44] L Minh Dang, Kyungbok Min, Hanxiang Wang, Md Jalil Piran, Cheol Hee Lee, and Hyeonjoon Moon. Sensor-based and vision-based human activity recognition : A comprehensive survey. *Pattern Recognition*, 108 :107561, 2020.

[45] Vijeta Sharma, Manjari Gupta, Anil Kumar Pandey, Deepti Mishra, and Ajai Kumar. A review of deep learning-based human activity recognition on benchmark video datasets. *Applied Artificial Intelligence*, 36(1) :2093705, 2022.

[46] Thanina BOULTACHE et LAMZAOUI Amar. *Reconnaissance d'activités humaines à l'aide de capteurs de smartphone*. PhD thesis, Université Mouloud MAMMERI de TIZI-OUZOU, 2019-2020.

[47] Alvin Raj, Amarnag Subramanya, Dieter Fox, and Jeff Bilmes. Rao-blackwellized particle filters for recognizing activities and spatial context from wearable sensors. In *Experimental Robotics : The 10th International Symposium on Experimental Robotics*, pages 211–221. Springer, 2008.

[48] Yi-Liang Kuo, Karen M Culhane, Pamela Thomason, Oren Tirosh, and Richard Baker. Measuring distance walked and step count in children with cerebral palsy : an evaluation of two portable activity monitors. *Gait & posture*, 29(2) :304–310, 2009.

[49] Sen Qiu, Hongkai Zhao, Nan Jiang, Zhelong Wang, Long Liu, Yi An, Hongyu Zhao, Xin Miao, Ruichen Liu, and Giancarlo Fortino. Multi-sensor information fusion based on machine learning for real applications in human activity recognition : State-of-the-art and research challenges. *Information Fusion*, 80 :241–265, 2022.

[50] D Roggen, S Magnenat, M Waibel, and G Tröster. Designing and sharing activity recognition systems across platforms : methods from wearable computing. *IEEE Robotics and Automation Magazine*, 12 :83–95, 2011.

[51] Ludovic DE MATTEIS, S Janny, S Nathan, and W Shu-Quartier. Introduction à l'apprentissage automatique, 2022.

[52] Le machine learning. https://ia-data-analytics.fr/machine-learning/. Consulté le 4 juin 2024.

[53] Velibor Božić. Machine learning vs deep learning. *March*, 2024.

[54] A comprehensive guide to supervised learning in machine learning. https://www.linkedin.com/pulse/comprehensive-guide-supervised-learning-machine-shobha-sharma-qzrec. Consulté : novembre 2023.

[55] Qu'est-ce que l'apprentissage par renforcement ? ,. https://aws.amazon.com/fr/what-is/reinforcement-learning/. Consulté : novembre 2023.

[56] Semi-supervised learning. https://maddevs.io/blog/what-is-semi-supervised-learning/. Consulté : novembre 2023.

[57] Abu Rayhan and Robert Kinzler. The fundamental concepts behind deep learning.

[58] Farhad Mortezapour Shiri, Thinagaran Perumal, Norwati Mustapha, and Raihani Mohamed. A comprehensive overview and comparative analysis on deep learning models : Cnn, rnn, lstm, gru. *arXiv e-prints*, pages arXiv–2305, 2023.

[59] Ravi Raj and Andrzej Kos. An improved human activity recognition technique based on convolutional neural network. *Scientific Reports*, 13(1) :22581, 2023.

[60] Shrishti Bisht, Sunita Joshi, Urvi Rana, et al. Comprehensive review of r-cnn and its variant architectures. *International Research Journal on Advanced Engineering Hub (IRJAEH)*, 2(04) :959–966, 2024.

[61] SU Park, JH Park, Mohammed A Al-Masni, Mugahed A Al-Antari, Md Z Uddin, and T-S Kim. A depth camera-based human activity recognition via deep learning recurrent neural network for health and social care services. *Procedia Computer Science*, 100 :78–84, 2016.

[62] Harsh Arora and Mamta Bansal. Feature extraction through sentiment analysis of tourist sentiments using deep learning techniques like cnn, rnn and lstm. *International Journal of Recent Technology and Engineering (IJRTE)*, 9(1), 2020.

[63] Mochamad Ridwan, Kusman Sadik, and Farit Mochamad Afendi. Comparison of arima and gru models for high-frequency time series forecasting. *Scientific Journal of Informatics*, 10(3) :389–400, 2023.

[64] Hermawan Nugroho, Gee Yang Tay, and Swaraj Dube. Perceptual autoencoder and exemplar selection for lifelong learning in convolutional neural networks (cnns). 2024.

[65] Réseaux antagonistes génératifs ou gan. https://datascientest.com/generative-adversarial-network-tout-savoir. Consulté : décembre 2023.

[66] Transformer : un réseau de neurones taillé pour le nlp. https://www.journaldunet.fr/intelligence-artificielle/guide-de-l-intelligence-artificielle/1508983-transformer-deep-learning/. Consulté : 6 juillet 2024.

[67] Michail Kaseris, Ioannis Kostavelis, and Sotiris Malassiotis. A comprehensive survey on deep learning methods in human activity recognition. *Machine Learning and Knowledge Extraction*, 6(2) :842–876, 2024.

[68] Zhuolin Jiang, Zhe Lin, and Larry Davis. Recognizing human actions by learning and matching shape-motion prototype trees. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(3) :533–547, 2012.

[69] Nobuyuki Otsu et al. A threshold selection method from gray-level histograms. *Automatica*, 11(285-296) :23–27, 1975.

[70] Lena Gorelick, Moshe Blank, Eli Shechtman, Michal Irani, and Ronen Basri. Actions as space-time shapes. *Transactions on Pattern Analysis and Machine Intelligence*, 29(12) :2247–2253, December 2007.

[71] Zhuolin Jiang, Zhe Lin, and Larry Davis. Recognizing human actions by learning and matching shape-motion prototype trees. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(3) :533–547, 2012.

[72] Christian Schuldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions : a local svm approach. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 3, pages 32–36. IEEE, 2004.

[73] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. A dataset of 101 human action classes from videos in the wild. *Center for Research in Computer Vision*, 2(11) :1–7, 2012.

[74] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb : a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE, 2011.

[75] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *Proc. 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017*, volume 2017, pages 2261–2269, 2017.

[76] Mingxing Tan and Quoc V. Le. Efficientnet : Rethinking model scaling for convolutional neural networks. In *36th Int. Conf. Mach. Learn. ICML 2019*, volume 2019-June, pages 10691–10700, 2019.

[77] Salvatore Gaglio, Giuseppe Lo Re, and Marco Morana. Human activity recognition process using 3-d posture data. *IEEE Trans. Human-Mach. Syst.*, 45(5) :586–597, 2015.

[78] Jingen Wang, Zicheng Liu, Ying Wu, and Junliang Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1290–1297, 2012.

[79] Kiwon Yun, Jean Honorio, Debdoot Chattopadhyay, Tamara L. Berg, and Dimitris Samaras. Two-person interaction detection using body-pose features and multiple instance learning. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 28–35, 2012.

[80] Tej Singh and Dinesh Kumar Vishwakarma. A deeply coupled convnet for human activity recognition using dynamic and rgb images. 2020.

[81] Vincenzo Carletti, Pasquale Foggia, Gennaro Percannella, Alessia Saggese, and Mario Vento. Recognition of human actions from rgb-d videos using a reject option. In *International workshop on social behaviour analysis*, 2013.

[82] Omar Oreifej and Zicheng Liu. Hon4d : histogram of oriented 4d normals for activity recognition from depth sequences. In *IEEE international conference on computer vision and pattern recognition (CVPR)*, Portland, OR, 2013.

[83] Jia Deng, Wei Dong, Richard Socher, Li Li, Li Kai, and Li Fei-Fei. Imagenet : A large-scale hierarchical image database. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, June 2009.

[84] Daniel Weinland, Remi Ronfard, and Edmond Boyer. Free viewpoint action recognition using motion history volumes. *Comput. Vision Image Understanding*, 104 :249–257, 2006.

[85] Hilde Kuehne, Hueihan Jhuang, Esti Garrote, Tomaso Poggio, and Thomas Serre. Hmdb : a large video database for human motion recognition. In *Proc. of the Int. Conference on Computer Vision (ICCV)*, pages 2556–2563, Barcelona, 2011. IEEE.

[86] Hilde Kuehne, Jurgen Gall, and Thomas Serre. An end-to-end generative framework for video segmentation and recognition. In *Proc. of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–8, Lake Placid, NY, USA, 2016. IEEE.

[87] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m : a large-scale video classification benchmark. *Computer Vision and Pattern Recognition*, 1 :1–10, 2016.

[88] Will Kay and et al. The kinetics human action video dataset. *Computer Vision and Pattern Recognition*, 1 :1–22, 2017.

[89] Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra, Jorge Luis Reyes-Ortiz, et al. A public domain dataset for human activity recognition using smartphones. In *Esann*, volume 3, page 3, 2013.

[90] Mi Zhang and Alexander A Sawchuk. Usc-had : A daily activity dataset for ubiquitous activity recognition using wearable sensors. In *Proceedings of the 2012 ACM conference on ubiquitous computing*, pages 1036–1043, 2012.

[91] Ricardo Chavarriaga, Hesam Sagha, Alberto Calatroni, Sundara Tejaswi Digumarti, Gerhard Tröster, José del R Millán, and Daniel Roggen. The opportunity challenge : A benchmark database for on-body sensor-based activity recognition. *Pattern Recognition Letters*, 34(15) :2033–2042, 2013.

[92] Marc Bachlin, Meir Plotnik, Daniel Roggen, Inbal Maidan, Jeffrey M Hausdorff, Nir Giladi, and Gerhard Troster. Wearable assistant for parkinson's disease patients with the freezing of gait symptom. *IEEE Transactions on Information Technology in Biomedicine*, 14(2) :436–446, 2009.

[93] Piero Zappi, Clemens Lombriser, Thomas Stiefmeier, Elisabetta Farella, Daniel Roggen, Luca Benini, and Gerhard Tröster. Activity recognition from on-body sensors : accuracy-power trade-off by dynamic sensor selection. In *Wireless Sensor Networks : 5th European Conference, EWSN 2008, Bologna, Italy, January 30-February 1, 2008. Proceedings*, pages 17–33. Springer, 2008.

[94] Yuwen Chen, Kunhua Zhong, Ju Zhang, Qilong Sun, and Xueliang Zhao. Lstm networks for mobile human activity recognition. In *2016 International conference on artificial intelligence : technologies and applications*, pages 50–53. Atlantis Press, 2016.

[95] Attila Reiss and Didier Stricker. Introducing a new benchmarked dataset for activity monitoring. In *2012 16th international symposium on wearable computers*, pages 108–109. IEEE, 2012.

[96] S. U. Park, J. H. Park, M. A. Al-masni, M. A. Al-antari, Md. Z. Uddin, and T. S. Kim. A depth camera-based human activity recognition via deep learning recurrent neural network for health and social care services. 2016.

[97] JH Park, SU Park, Md Zia Uddin, MA Al-Antari, MA Al-Masni, and T-S Kim. A single depth sensor based human activity recognition via convolutional neural network. In *6th International Conference on the Development of Biomedical Engineering in Vietnam (BME6) 6*, pages 541–545. Springer, 2018.

[98] Sravan Kumar Challa, Akhilesh Kumar, Vijay Bhaskar Semwal, and Nidhi Dua. An optimized deep learning model for human activity recognition using inertial measurement units. *Expert Systems*, 40(10) :e13457, 2023.

[99] Oresti Banos, Rafael Garcia, Juan A Holgado-Terriza, Miguel Damas, Hector Pomares, Ignacio Rojas, Alejandro Saez, and Claudia Villalonga. mhealthdroid : a novel framework for agile development of mobile health applications. In *Ambient Assisted Living and Daily Activities : 6th International Work-Conference, IWAAL 2014, Belfast, UK, December 2-5, 2014. Proceedings 6*, pages 91–98. Springer, 2014.

[100] Jingen Liu, Jiebo Luo, and Mubarak Shah. Recognizing realistic actions fromvideos. *the wild", IEEE Computer Vision and Pattern Recognition*, 2009.

[101] Seyed Morteza Safdarnejad, Xiaoming Liu, Lalita Udpa, Brooks Andrus, John Wood, and Dean Craven. Sports videos in the wild (svw) : A video dataset for sports analysis. In *2015 11th IEEE international conference and*

workshops on automatic face and gesture recognition (FG), volume 1, pages 1–7. IEEE, 2015.

[102] Mohit Kumar, Adarsh Rana, Ankita, Arun Kumar Yadav, and Divakar Yadav. Human activity recognition in videos using deep learning. In *International Conference on Soft Computing and its Engineering Applications*, pages 288–299. Springer, 2022.

[103] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. A dataset of 101 human action classes from videos in the wild. *Center for Research in Computer Vision*, 2(11) :1–7, 2012.

[104] Kaiming He Georgia Gkioxari Piotr Dollár, Ross Girshick, et al. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[105] Brondon Styve Tchamba Kuinze. *Comparaison de l'efficacité du deep learning et de l'apprentissage automatique classique pour la reconnaissance d'activités humaine dans les habitats intelligents*. PhD thesis, Université du Québec à Chicoutimi, 2024.

[106] Semin Ryu, Suyeon Yun, Sunghan Lee, and In cheol Jeong. Exploring the possibility of photoplethysmography-based human activity recognition using convolutional neural networks. *Sensors*, 24(5) :1610, 2024.

[107] Limeng Lu, Chuanlin Zhang, Kai Cao, Tao Deng, and Qianqian Yang. A multichannel cnn-gru model for human activity recognition. *IEEE Access*, 10 :66797–66810, 2022.

[108] « General Python FAQ ». Python Documentation. https://docs.python.org/3/faq/general.html#general-information. Consulté le : 20 novembre 2023.

[109] Yasmin Makki Mohialden, Raed Waheed Kadhim, Nadia Mahmood Hussien, and Samira Abdul Kader Hussain. Top python-based deep learning packages : A comprehensive review. *International Journal Papier Advance and Scientific Review*, 5(1) :1–9, 2024.

[110] sklearn. https://domino.ai/data-science-dictionary/sklearn. Consulté le : 20 novembre 2023.

[111] « NumPy ». DATAROCKSTARS. https://www.datarockstars.ai/glossary/numpy/. Consulté le : 20 novembre 2023.

[112] Henri Michel. « Google Colab : Le guide Ultime. https://ledatascientist.com/google-colab-le-guide-ultime/. Consulté le : 20 novembre 2023.

[113] Yuwen Chen, Kunhua Zhong, Ju Zhang, Qilong Sun, and Xueliang Zhao. Lstm networks for mobile human activity recognition. In *2016 International conference on artificial intelligence : technologies and applications*, pages 50–53. Atlantis Press, 2016.

[114] Attila Reiss and Didier Stricker. Introducing a new benchmarked dataset for activity monitoring. In *2012 16th international symposium on wearable computers*, pages 108–109. IEEE, 2012.

[115] Chuanlin Zhang, Kai Cao, Limeng Lu, and Tao Deng. A multi-scale feature extraction fusion model for human activity recognition. *Scientific Reports*, 12(1) :20620, 2022.

Abstract

Human activity recognition (HAR) is essential in various fields such as healthcare, sports, surveillance, and human-machine interactions, significantly contributing to improved quality of life and security. This work focuses on developing a sensor-based HAR system using a hybrid model that combines Convolutional Neural Networks (CNN) and Gated Recurrent Units (GRU). This combination leverages the effectiveness of CNNs for extracting spatial features and GRUs for capturing temporal dependencies, thereby enhancing the system's accuracy and robustness. Experimental results, conducted on the WISDM and PAMAP2 datasets, demonstrate improved performance through the use of magnitude and data augmentation techniques. In conclusion, this study presents an advanced approach to human activity recognition, with continuous improvement prospects and expansion to more complex activities involving precise interactions with objects.

Keywords : Human Activity Recognition, Deep Learning, Convolutional Neural Networks (CNN), Gated Recurrent Units (GRU), Hybrid Model, Sensor Data.

Résumé

La reconnaissance des activités humaines (HAR) est essentielle dans divers domaines tels que la santé, le sport, la surveillance et les interactions homme-machine, et elle contribue significativement à l'amélioration de la qualité de vie et de la sécurité. Ce travail se concentre sur le développement d'un système HAR basé sur des capteurs, utilisant un modèle hybride combinant les réseaux de neurones convolutifs (CNN) et les Gated Recurrent Units (GRU). Cette combinaison permet de bénéficier de l'efficacité des CNN pour extraire des caractéristiques spatiales et des GRU pour capturer les dépendances temporelles, améliorant ainsi la précision et la robustesse du système. Les résultats des tests expérimentaux, réalisés sur les bases de données WISDM et PAMAP2, démontrent une performance accrue grâce à l'utilisation de techniques de magnitude et d'augmentation des données. En conclusion, cette étude propose une approche avancée pour la reconnaissance des activités humaines, avec des perspectives d'amélioration continue et d'élargissement à des activités plus complexes impliquant des interactions précises avec des objets.

Mots clés : Reconnaissance des activités humaines,Apprentissage profond, Réseaux de neurones convolutifs (CNN),Gated Recurrent Units (GRU),Modèle hybride, Données de capteurs.