Apprentissage profond pour l'authentification continue et la reconnaissance d'activités humaines à partir de données comportementales multimodales

par

Khalil SNOUSSI

MÉMOIRE PRÉSENTÉ À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
COMME EXIGENCE PARTIELLE À L'OBTENTION DE LA MAÎTRISE
AVEC MÉMOIRE
M. Sc. A.

MONTRÉAL, LE 28 AOÛT 2025

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

PRÉSENTATION DU JURY

CE MÉMOIRE A ÉTÉ ÉVALUÉ

PAR UN JURY COMPOSÉ DE:

M. Wael Jaafar, directeur de mémoire
Département de génie logiciel et TI à l'École de technologie supérieure

M. Rami Langar, codirecteur
Département de génie logiciel et TI à l'École de technologie supérieure

Mme. Bassant Selim, présidente du jury
Département de génie des systèmes à l'École de technologie supérieure

M. Ulrich Aïvodji, membre du jury
Département de génie logiciel et TI à l'École de technologie supérieure

IL A FAIT L'OBJET D'UNE SOUTENANCE DEVANT JURY ET PUBLIC

LE 27 AOÛT 2025

À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

Apprentissage profond pour l'authentification continue et la reconnaissance d'activités humaines à partir de données comportementales multimodales

Khalil SNOUSSI

RÉSUMÉ

Les smartphones modernes embarquent une variété de capteurs, tels que l'accéléromètre, le gyroscope et l'écran tactile capacitif, offrant une opportunité unique pour l'authentification continue et la reconnaissance d'activités humaines. Ce mémoire propose deux architectures basées sur l'apprentissage profond et les modèles à espaces d'états structurés (SSM) : S4HI, pour l'identification robuste à partir de données comportementales inertielles, et MEDUSAA, un encodeur-décodeur multimodal combinant capteurs inertiels et interactions tactiles. Ces approches exploitent la capacité des SSM à capturer efficacement les dépendances temporelles longues, tout en réduisant la complexité et le temps d'inférence par rapport aux architectures de type Transformer. Les modèles sont évalués sur plusieurs jeux de données publics (HMOG, OU-ISIR, WhuGait, PAMAP2, Capture-24) et démontrent des performances supérieures en précision, évolutivité et déploiement en ligne. Les résultats obtenus confirment le potentiel de ces méthodes pour des applications embarquées sécurisées et à faible consommation énergétique

Mots-clés: Apprentissage profond, biométrie comportementale

Deep learning for continuous authentication and human activity recognition from multimodal behavioral data

Khalil SNOUSSI

ABSTRACT

Modern smartphones are equipped with a variety of sensors, such as the accelerometer, gyroscope, and capacitive touchscreen, which offer a unique opportunity for continuous authentication and human activity recognition. This thesis proposes two deep learning and structured state-space model (SSM) based architectures : S4HI, for robust identification using inertial behavioral data, and MEDUSAA, a multimodal encoder-decoder that combines inertial sensors and tactile interactions. These approaches leverage the ability of SSMs to effectively capture long-term temporal dependencies while reducing complexity and inference time compared to Transformer-type architectures. The models are evaluated on several public datasets (HMOG, OU-ISIR, WhuGait, PAMAP2, Capture-24) and demonstrate superior performance in terms of precision, scalability, and online deployment. The results confirm the potential of these methods for secure, low-energy embedded applications.

Keywords: Deep Learning, behavioral biometrics

TABLE DES MATIÈRES

Page

LISTE DES TABLEAUX

LISTE DES FIGURES

Page

LISTE DES ABRÉVIATIONS, SIGLES ET ACRONYMES

SSM	State Space Model
IoT	Internet of Things
MEMS	Micro-Electro-Mechanical System
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
GRU	Gated Recurrent Unit
HAR	Human Activity Recognition
EER	Equal Error Rate
FFT	Fast Fourier Transform
iFFT	inverse Fast Fourier Transform
SSL	Self-Supervised Learning
IMU	Inertial Measurement Unit
MB	MegaBytes
ms	milliseconds
GFLOPS	Giga Floating Point Operations Per Second
HSS	Hidden State Size

LISTE DES SYMBOLES ET UNITÉS DE MESURE

$\mathbf{A_x}, \mathbf{A_y}, \mathbf{A_z}$	Accélération selon les axes X, Y, Z
$\omega_\mathbf{x}, \omega_\mathbf{y}, \omega_\mathbf{z}$	Vitesse angulaire autour des axes X, Y, Z
$\mathbf{S}$	Surface de contact sur l'écran tactile
$\mathbf{P}$	Pression exercée sur l'écran tactile
$\mathbf{x}, \mathbf{y}$	Coordonnées du point de contact sur l'écran tactile
$\mathbf{t}$	Temps
$\mathbf{f}$	Fonction d'activation
η	Taux d'apprentissage
$\mathbf{E}$	Fonction de coût (erreur)
$\mathbf{\Delta}$	Intervalle d'échantillonnage
$\mathbf{K(t)}$	Noyau de convolution dans un SSM continu
$\mathbf{K_d[j]}$	Noyau de convolution discret

INTRODUCTION

Les téléphones portables jouent un rôle très important dans notre vie quotidienne. Nous payons nos factures et accédons à nos comptes bancaires rapidement grâce à nos téléphones. C'est pour cette raison que la sécurité de ces appareils est au cœur de ce travail de recherche.

Bien que de nombreuses méthodes d'authentification dites « traditionnelles » comme les mots de passe, les codes PIN et la reconnaissance faciale existent, elles partagent toutes une faiblesse majeure : une fois l'authentification réussie, l'appareil reste ouvert jusqu'à ce qu'il soit verrouillé, créant ainsi une possibilité d'attaques.

L'authentification continue vise à corriger cette faille en remplaçant la vérification statique et périodique par une évaluation constante et en continu qui se fonde sur la surveillance des comportements de l'utilisateur actuel, en tenant compte de son historique d'utilisation précédent.

La majorité de tous les récents téléphones portables contiennent des capteurs exploitables pour mesurer le comportement de l'utilisateur tenant son téléphone. Parmi ces capteurs, on peut citer les capteurs inertiels comme l'accéléromètre et le gyroscope ainsi que l'écran tactile, qui contient des capteurs capacitifs capables de mesurer la surface de contact, la pression et les coordonnées du toucher. Ces capteurs génèrent une quantité énorme de données qu'on peut exploiter pour analyser les comportements et les interactions des utilisateurs avec leurs appareils dans le but de trouver des patrons subtils et difficiles à reproduire par un autre utilisateur dit imposteur.

Malgré son potentiel L'authentification en continue fait face à beaucoup de défis techniques, notamment la variabilité inter-utilisateur, puisqu'un utilisateur peut faire différentes activités. Ainsi, les données capturées lors d'une session de jogging diffèrent beaucoup de celles lorsqu'il est allongé, même si elles proviennent d'un même utilisateur. De plus, il faut donc penser à construire des systèmes d'authentification qui sont légers, puisque ces systèmes doivent

s'exécuter sur des appareils à ressources limitées. Aussi, la méthode de collecte de données ne doit pas drainer les batteries.

Face à ces défis, ce travail de recherche répond à la problématique suivante : comment concevoir un système d'authentification continue robuste, léger et qui emploie conjointement la fusion des données issues des capteurs inertiels et des capteurs capacitifs de l'écran tactile ?

Cette problématique se compose de plusieurs sous-questions :

- Quelles caractéristiques extraites des données inertielles et des données d'interaction avec l'écran tactile discriminent le mieux entre les utilisateurs ?
- Quelles architectures de réseaux de neurones sont les plus adaptées pour ces types de données ?
- Comment fusionner des données de nature différente afin de les utiliser comme entrée à nos réseaux de neurones implémentés ?
- Comment assurer la légèreté des réseaux de neurones en termes de consommation énergétique sur les téléphones portables ?

Ce mémoire est structuré en cinq chapitres. Le premier chapitre présente les notions préliminaires essentielles à la compréhension du travail. Le deuxième chapitre est consacré à une revue de littérature, mettant en perspective les travaux existants dans le domaine. Les troisième et quatrième chapitres correspondent à deux articles scientifiques : le premier, soumis et accepté à la conférence CIoT 2024, et le second, soumis dans le journal IEEE TBIOM. Enfin, le cinquième chapitre expose la conclusion générale ainsi que les perspectives de recherche futures.

CHAPITRE 1

NOTIONS PRÉLIMINAIRES

Dans ce chapitre, nous allons explorer des notions essentielles à ce travail de recherche et à la compréhension de notre approche d'authentification continue comportementale. Plus précisément, nous présentons les différents capteurs utilisés, à savoir l'accéléromètre, le gyroscope et les capteurs capacitifs des écrans tactiles, puis nous décrivons le fonctionnement des réseaux de neurones employés pour l'analyse et la classification.

1.1 Capteurs utilisés

1.1.1 L'écran tactile

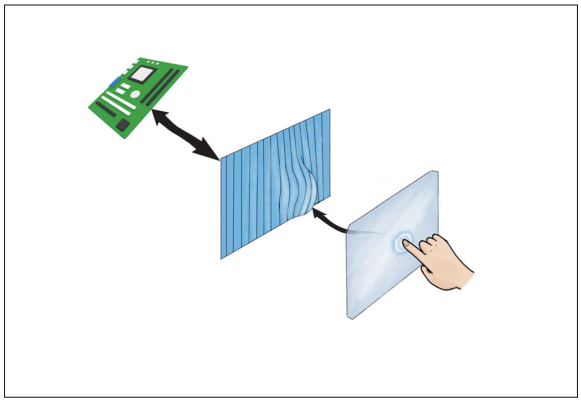


Figure 1.1 Principe du fonctionnement de l'écran tactile

L'écran tactile est la composante principale qui permet à l'utilisateur d'interagir avec l'interface utilisateur. Cette technologie repose sur des capteurs capacitifs et les propriétés électriques du corps humain. En effet, l'écran tactile des smartphones modernes contient une couche d'oxyde d'indium-étain qui est formée par une grille d'électrodes de telle façon qu'à ce que, lorsqu'un utilisateur touche un endroit sur l'écran, il déforme le champ électrique de cet endroit. Cette déformation permet aux contrôleurs de calculer :

- Les coordonnées (x, y) du point qui indiquent la position exacte du toucher.
- La surface de contact

- La pression appliquée
- Horodatage précis de chaque événement tactile.

1.1.2 L'accéléromètre

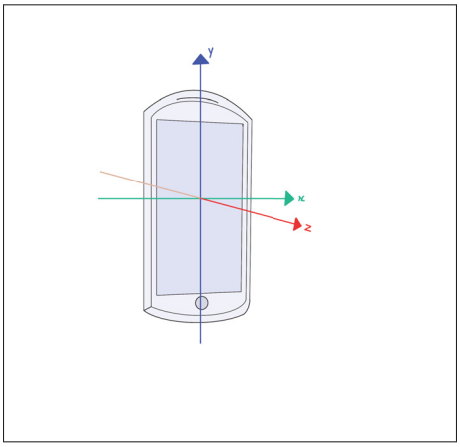


Figure 1.2 Orientation des axes de l'accéléromètre

L'accéléromètre est un capteur qui mesure les forces auxquelles est soumis le téléphone portable selon les trois axes orthogonaux (x, y, z). Ces axes sont généralement placés comme suit :

- axe X : horizontal, parallèle au petit côté de l'écran
- axe Y : vertical, parallèle au grand côté de l'écran
- axe Z : perpendiculaire à l'écran

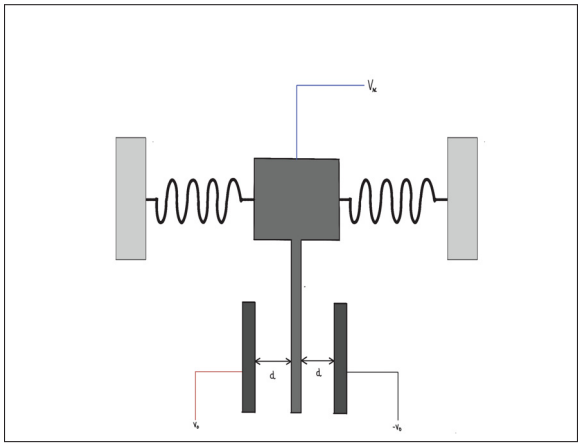


Figure 1.3 Modélisation du principe de l'accéléromètre

Les accéléromètres trouvés dans les smartphones modernes sont généralement de type MEMS (micro-electro-mechanical systems) et se composent de petites masses dont les déplacements causés par des forces extérieures sont mesurés par la variation de la capacité électrique.

En effet, le déplacement de la masse dans la figure 1.3 fait varier la capacité électrique entre les deux électrodes, et c'est cette variation de la capacité qui est convertie en un signal électronique qui peut être ensuite traité pour calculer la force exercée sur la masse et finalement trouver l'accélération par la loi de Newton : $F = m.a$

Dans le contexte de l'authentification comportementale, l'accéléromètre joue un rôle majeur, car il permet aux réseaux de neurones de capturer les mouvements caractéristiques de chaque utilisateur. Un grand avantage de l'utilisation de ce capteur est le fait qu'il fournit des mesures d'accélération sur les trois axes X, Y, Z en continu et à un taux d'échantillonnage très élevé, entre $30Hz$ et $100Hz$, même pour les petites secousses involontaires des utilisateurs lors de la manipulation de leur dispositif. Ces secousses involontaires peuvent s'avérer cruciales dans la discrimination entre les utilisateurs.

1.1.3 Le gyroscope

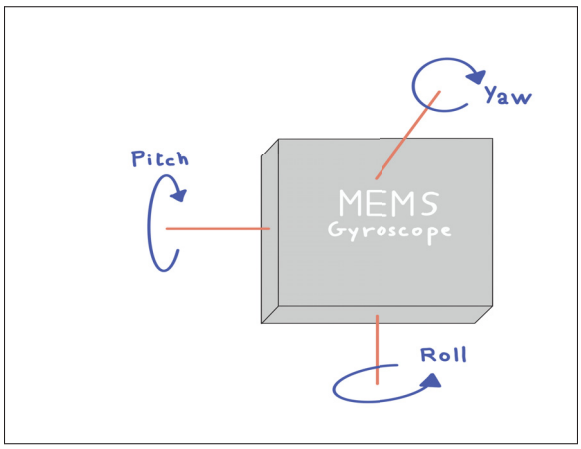


Figure 1.4 Modélisation du principe de l'accéléromètre

Le gyroscope est un capteur essentiel pour l'authentification comportementale, car, contrairement à l'accéléromètre qui calcule les changements d'accélération linéaire, il permet de mesurer la vitesse angulaire du dispositif, ce qui permet de calculer l'orientation, l'inclinaison et les rotations du dispositif par rapport aux trois axes dimensionnels :

- axe x (pitch) : rotation autour de l'axe longitudinal
- axe y (roll) : rotation autour de l'axe transversal
- axe z (yaw) : rotation autour de l'axe vertical

Les gyroscopes qu'on trouve généralement dans les smartphones sont également de type MEMS. Couplés à l'accéléromètre, ces deux capteurs peuvent être utilisés pour calculer une signature comportementale unique à chaque utilisateur. Le tableau suivant récapitule les données utilisées dans ce travail de recherche :

Tableau 1.1 Résumé des données collectées à partir des capteurs inertiels et de l'écran tactile pour l'authentification comportementale.

Capteur	**Données mesurées**	**Unité**
Accéléromètre	Accélération $(\mathbf{A_x}, \mathbf{A_y}, \mathbf{A_z})$	m/s^2
Gyroscope	Vitesse angulaire $(\omega_\mathbf{x}, \omega_\mathbf{y}, \omega_\mathbf{z})$	$/s$
Écran tactile	Coordonnées $(\mathbf{x}, \mathbf{y})$, Surface $\mathbf{S}$, Pression $\mathbf{P}$	Pixels, mm^2, N

L'un des avantages de l'utilisation des capteurs inertiels est leur capacité à fournir des informations détaillées et en continu à grand taux d'échantillonnage, $100Hz$ ce qui permet une très haute résolution de données contrairement aux méthodes statiques comme la reconnaissance vocale et la reconnaissance faciale, ce qui veut dire que, même lorsque l'utilisateur n'utilise pas son téléphone portable, cela permet une authentification passive qui se fait en arrière-plan sans demander à l'utilisateur de fournir un effort pour l'authentifier. D'autre part, il est difficile pour les imposteurs de reproduire des mouvements identiques à l'utilisateur authentique.

Pour conclure, l'utilisation des capteurs inertiels dans l'authentification continue ouvre la voie à des solutions d'authentification beaucoup plus sécurisées, robustes et surtout transparentes, ne nécessitant presque aucun effort de la part des utilisateurs.

1.2 Les réseaux de neurones

Les réseaux de neurones sont une famille de concepts dans l'apprentissage machine qui, à la lumière de la constitution du cerveau humain formé de plusieurs millions de neurones connectés, consistent en plusieurs couches de neurones qui transforment progressivement les données afin d'en apprendre une représentation plus compacte dans le but d'apprendre des patrons pertinents et propices à l'accomplissement des tâches complexes comme la reconnaissance de la voix, la classification des images, le traitement du signal et la compréhension du langage naturel.

1.2.1 Aperçu historique

Les premiers travaux sur les réseaux de neurones remontent à 1943, lorsque des mathématiciens et des neuroscientifiques ont proposé un premier modèle théorique permettant de comprendre comment le cerveau traite l'information McCulloch & Pitts (1943). Dans ce travail, les auteurs ont proposé le tout premier modèle théorique d'un neurone capable de faire des opérations logiques, posant ainsi une base solide pour tous les réseaux de neurones modernes.

Certes, Rosenblatt (1958) a introduit le perceptron en 1958. Il consiste en un seul neurone dont les entrées sont pondérées puis sommées pour finalement être transmises vers une fonction d'activation qui est souvent une fonction seuil afin de produire une classification binaire.

Soit y la sortie du perceptron, elle peut s'écrire mathématiquement comme suit :

$$y = f\left(\sum_{i=1}^{n} w_i x_i + b\right) \tag{1.1}$$

où x_i sont les entrées, w_i les poids associés à chaque entrée et b représente le biais. La fonction d'activation seuil f est définie par :

$$f(u) = \begin{cases} 1 & \text{si } u \geq 0 \\ 0 & \text{sinon} \end{cases} \tag{1.2}$$

Ainsi, la sortie y prend une valeur binaire indiquant l'appartenance à l'une des deux classes considérées.

$$w_i \leftarrow w_i + \eta(y_{\text{réel}} - y_{\text{prédit}})x_i \tag{1.3}$$

$$b \leftarrow b + \eta(y_{\text{réel}} - y_{\text{prédit}}) \tag{1.4}$$

où :

- η est le taux d'apprentissage, un paramètre positif contrôlant la vitesse d'apprentissage ;
- $y_{\text{réel}}$ est la valeur cible attendue pour l'entrée considérée ;
- $y_{\text{prédit}}$ est la sortie obtenue par le perceptron pour cette entrée.

Toutefois, ce type de réseaux de neurones a vite montré ses faiblesses, car en 1969, Minsky & Papert (1969) a démontré dans *Perceptrons* que le perceptron ne peut pas résoudre des problèmes où les données sont non-linéairement séparables. Un exemple classique de telles données est la fonction logique XOR, car il n'existe aucune frontière de décision linéaire qui peut séparer les classes.

$$\text{XOR} = \begin{cases} 0 & \text{si } (x_1 = 0 \text{ et } x_2 = 0) \text{ ou } (x_1 = 1 \text{ et } x_2 = 1) \\ 1 & \text{si } (x_1 = 0 \text{ et } x_2 = 1) \text{ ou } (x_1 = 1 \text{ et } x_2 = 0) \end{cases} \tag{1.5}$$

Cette limitation a gelé temporairement les recherches sur les réseaux de neurones jusqu'en 1986, lorsque Rumelhart, Hinton & Williams (1986) ont développé l'apprentissage par rétropropagation et ont par conséquent révolutionné l'apprentissage machine, car cet algorithme a permis l'entrainement de plusieurs couches de neurones de façon efficace grâce à la règle de la chaine qui, couplée à la descente du gradient, permet de propager l'erreur depuis la couche de sortie vers les couches cachées intermédiaires. En effet, la règle de la chaîne permet de calculer la

dérivée d'une fonction composée, par exemple lorsqu'une fonction f dépend d'une variable g, qui dépend elle-même d'une variable x.

$$\frac{df}{dx} = \frac{df}{dg} \cdot \frac{dg}{dx}. \tag{1.6}$$

Dans le cadre de l'apprentissage profond, la mise à jour des poids d'un réseau de neurones consiste à minimiser une fonction de coût E en ajustant les poids. w_{ij}

$$w_{ij} \leftarrow w_{ij} - \eta \frac{\partial E}{\partial w_{ij}} \tag{1.7}$$

Où η est le taux d'apprentissage et $\frac{\partial E}{\partial w_{ij}}$ est obtenu en appliquant la règle de la chaîne :

Grâce à cette avancée, les perceptrons multicouches ont pu être entraînés efficacement, ouvrant la voie à une nouvelle ère dans l'intelligence artificielle. L'algorithme de rétropropagation a permis aux modèles de reconnaître des structures complexes dans les données, surpassant les limites du perceptron simple qui ne pouvait apprendre que des fonctions linéairement séparables.

Cette percée a inspiré le développement de nombreuses architectures plus sophistiquées, comme les réseaux convolutionnels (CNN) pour la vision par ordinateur et les réseaux récurrents (RNN, LSTM, GRU) pour le traitement du langage naturel et les séries temporelles.

1.2.2 Le perceptron multicouche

Le perceptron multicouche est un réseau de neurones composé d'une succession de couches, chacune contient plusieurs neurones dont chaque neurone est connecté à tous les neurones de la couche suivante.

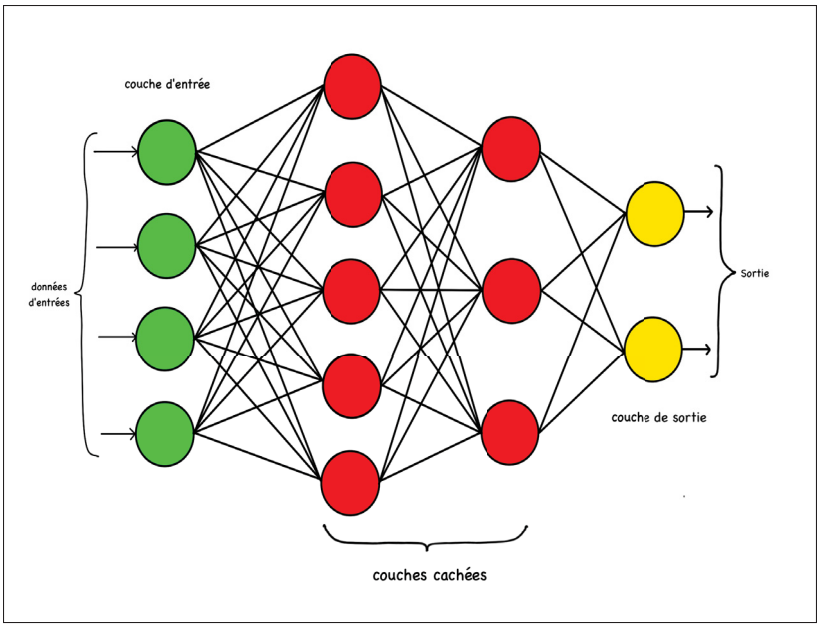


Figure 1.5 Architecture d'un perceptron multicouches

En general l'architecture MLP comprends trois composantes :

- une couche d'entree notee $l = 0$
- plusieurs couches cachees notees $l = 1$ jusqu'a $l = L - 1$
- une couche de sortie $l = L$

En effet, le perceptron multicouche transforme les donnees d'entree a travers les L couches successives pour produire une prediction ou une classification selon les equations suivantes : Pour chaque couche $l = 1 \cdots L$

$$\mathbf{z}^{(l)} = \mathbf{W}^{(l)}\mathbf{a}^{(l-1)} + \mathbf{b}^{(l)}, \tag{1.8}$$

$$\mathbf{a}^{(l)} = f^{(l)}(\mathbf{z}^{(l)}). \tag{1.9}$$

- $\mathbf{x} = \mathbf{a}^{(0)}$: Vecteur d'entrée.
- $\mathbf{W}^{(l)} \in \mathbb{R}^{n_l \times n_{l-1}}$: Matrice des poids.
- $\mathbf{b}^{(l)} \in \mathbb{R}^{n_l}$: Vecteur de biais.

- $\mathbf{z}^{(l)} = \mathbf{W}^{(l)}\mathbf{a}^{(l-1)} + \mathbf{b}^{(l)}$: Combinaison linéaire.
- $\mathbf{a}^{(l)} = f^{(l)}(\mathbf{z}^{(l)})$: Activation.
- $\hat{y} = \mathbf{a}^{(L)}$: Prédiction finale.

D'autre part, l'entrainement du reseau se fait a l'aide des equations de la retropropagation pour minimiser une fonction objective E :

L'erreur de la dernière couche peut être calculée à l'aide de la règle de la chaîne :

$$\delta^{(L)} = \frac{\partial E}{\partial \mathbf{a}^{(L)}} \odot f'^{(L)}(\mathbf{z}^{(L)}). \tag{1.10}$$

$$\delta^{(l)} = (\mathbf{W}^{(l+1)})^T \delta^{(l+1)} \odot f'^{(l)}(\mathbf{z}^{(l)}). \tag{1.11}$$

où $\odot$ représente le produit élément par élément (produit d'Hadamard).

Les gradients des poids et des biais sont donnés par :

$$\frac{\partial E}{\partial \mathbf{W}^{(l)}} = \delta^{(l)}(\mathbf{a}^{(l-1)})^T, \quad \frac{\partial E}{\partial \mathbf{b}^{(l)}} = \delta^{(l)}. \tag{1.12}$$

tous ces gradients sont calculés pour mettre à jour les paramètres du réseau pour minimiser l'erreur, où η est le taux d'apprentissage.

$$\mathbf{W}^{(l)} \leftarrow \mathbf{W}^{(l)} - \eta \frac{\partial E}{\partial \mathbf{W}^{(l)}}, \quad \mathbf{b}^{(l)} \leftarrow \mathbf{b}^{(l)} - \eta \frac{\partial E}{\partial \mathbf{b}^{(l)}}. \tag{1.13}$$

1.2.3 Les réseaux de neurones à convolution

Le besoin d'un autre type de réseau de neurones a été rapidement ressenti, surtout pour des données structurées à plusieurs dimensions, comme les images.

En effet, les MLP ne sont pas efficaces pour le traitement des données multidimensionnelles, car chaque neurone est connecté à tous les neurones de la couche précédente. Cela entraîne à la fois une explosion du nombre de paramètres à entraîner et la destruction de la structure spatiale des données, comme les images. En effet, chaque pixel de l'image est traité par un MLP comme une donnée indépendante, alors qu'il devrait être analysé en fonction de ses pixels voisins.

Les réseaux de neurones à convolution Lecun, Bottou, Bengio & Haffner (1998) répondent aux limitations des réseaux MLP en utilisant l'opération de convolution mathématique et en introduisant la notion de champs réceptifs locaux. Cela signifie que chaque neurone est connecté uniquement à une petite région de l'image, au lieu d'être relié à tous les pixels, comme c'est le cas dans les MLP.

En mathématiques et en traitement du signal, la convolution est une opération fondamentale qui décrit comment une fonction modifie l'autre. Sachant deux fonctions f et g, la convolution de f par g est définie comme :

$$(f * g)(t) = \int_{-\infty}^{+\infty} f(\tau)g(t-\tau)d\tau \tag{1.14}$$

Il s'agit d'un filtre qui modifie la fonction f en tout point de son domaine de définition. Cette équation est le fondement même des filtres électroniques passe-bas et passe-haut.

Dans le domaine du traitement d'image, nous utilisons la version discrète de l'équation :

$$Z_{ij} = \sum_{m=0}^{k} \sum_{n=0}^{k} K_{mn} \cdot X_{i-m,j-n} \tag{1.15}$$

Le noyau K est une matrice de poids de taille $k \times k$ qui définit la transformation appliquée à l'image. La taille du filtre détermine la zone de l'image considérée pour chaque calcul de convolution, et c'est cette zone qui est appelée le champ réceptif. Plus le noyau est grand, plus le champ réceptif couvre une large portion de l'image, tandis que lorsque la taille du noyau est petite,

la convolution se concentre sur des détails plus locaux. Ces filtres sont appris automatiquement par l'algorithme de la descente du gradient et la retropropagation.

1.2.4 Les réseaux de neurones récurrents

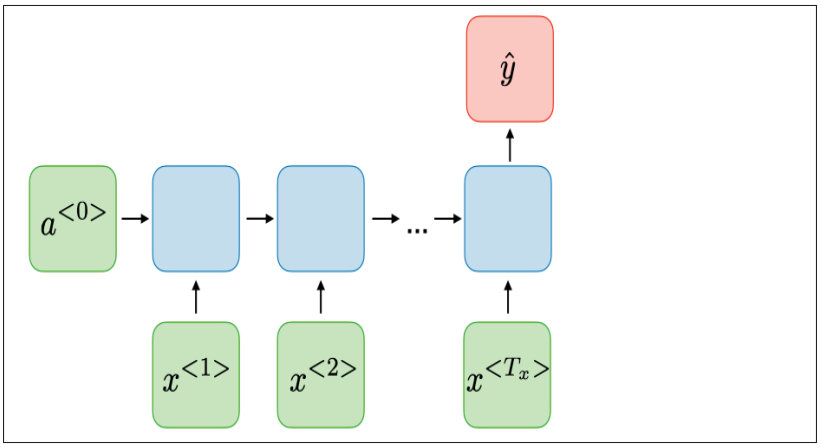


Figure 1.6 Architecture d'un réseau de neurones

Les réseaux de neurones récurrents viennent combler le besoin d'outils adaptés au traitement des données dont la nature dépend du temps, également connues sous le nom de séries temporelles. En effet, la nature séquentielle de certains types de données, comme les données audio, le langage naturel ou encore les données séquentielles des capteurs inertiels, tels que les accéléromètres, nécessite des outils capables de prendre en compte les dépendances temporelles. Contrairement aux réseaux de neurones MLP ou CNN, qui traitent chaque donnée indépendamment du temps, les réseaux récurrents intègrent l'ensemble du contexte passé pour prendre une décision dans le présent. Cela est possible grâce aux équations de rétroaction (feedback).

$$\mathbf{h}_t = \sigma(\mathbf{W_x X}_t + \mathbf{W_h h}_{t-1} + \mathbf{b_h}) \tag{1.16}$$

ou W_x et W_h sont des matrices de poids, b est le biais, σ est une fonction d'activation non-linéaire.

Cependant, les réseaux de neurones récurrents rencontrent des difficultés à exploiter pleinement leur objectif, à savoir apprendre les dépendances à long terme, en raison du phénomène d'explosion et de disparition du gradient, principalement causé par les équations de rétroaction. En effet, les phénomènes de disparition ou d'explosion du gradient dans le cas des réseaux récurrents sont causés par les propriétés spectrales des matrices de poids. Plus précisément, dans les équations de la rétro-propagation temporelle (backpropagation through time), le calcul de la fonction de perte L par rapport aux états cachés h_t pour chaque instant $t < T$:

$$\frac{\partial E}{\partial \mathbf{h}_t} = \frac{\partial E}{\partial \mathbf{h}_T} \frac{\partial \mathbf{h}_T}{\partial \mathbf{h}_t} \tag{1.17}$$

En décomposant la dérivée partielle $\frac{\partial \mathbf{h}_T}{\partial \mathbf{h}_t}$ en produit de jacobiens

$$\frac{\partial \mathbf{h}_T}{\partial \mathbf{h}t} = \prod_{i=t}^{T-1} \mathbf{J}_i \tag{1.18}$$

À chaque étape, la matrice jacobienne est donnée par :

$$\frac{\partial \mathbf{h}i}{\partial \mathbf{h}i-1} = \text{diag}(\sigma'(\mathbf{W_h}\mathbf{h}_{i-1} + \mathbf{W_x}\mathbf{x}_i + \mathbf{b})) \cdot \mathbf{W_h} \tag{1.19}$$

L'analyse asymptotique de cette expression nécessite l'étude des valeurs propres de la matrice des poids W_h. En effet, si le spectre de W_h est supérieur à 1, alors la norme du produit des jacobiens croît exponentiellement avec la longueur de $(T - t)$ ce qui entraîne une explosion du gradient. À l'inverse, si la norme du spectre est inférieure à 1, le gradient décroît exponentiellement jusqu'à sa disparition. Ainsi, ces limitations entraînent une difficulté à faire apprendre aux réseaux de neurones récurrents des dépendances à long terme, car des gradients trop grands ou trop petits empêchent une mise à jour adéquate des poids.

1.2.5 Transformeurs

L'architecture du transformeur Vaswani *et al.* (2017) constitue un paradigme de modèle pour le traitement de séquences, se distinguant par son abandon des mécanismes de récurrence au profit exclusif de l'attention. La structure est bipartite, comprenant un module d'encodage et un module de décodage, chacun composé d'une pile de N couches identiques. Le bloc encodeur a pour fonction de transformer une séquence de symboles d'entrée en un ensemble de représentations vectorielles continues dans un espace latent. Chaque couche de l'encodeur intègre deux sous-modules principaux : un mécanisme d'auto-attention multi-têtes 1.9 et un réseau de neurones à propagation avant entièrement connecté. Le décodeur, quant à lui, opère de manière auto-régressive pour générer la séquence de sortie, symbole par symbole. En plus des deux sous-modules présents dans l'encodeur, le décodeur incorpore une troisième sous-couche d'attention qui effectue une attention inter-sources sur les représentations de sortie de l'encodeur, permettant ainsi au modèle de pondérer la pertinence des éléments de la séquence d'entrée lors de la génération de chaque élément de la séquence de sortie. Des connexions résiduelles suivies d'une normalisation de couche sont appliquées autour de chaque sous-module pour faciliter l'entraînement de ce modèle profond

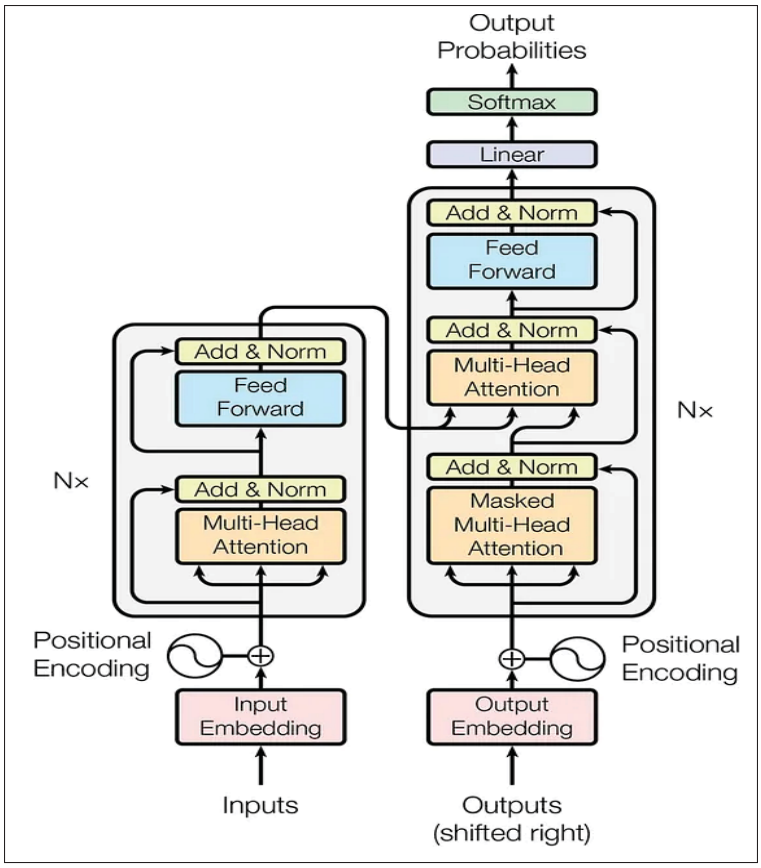


Figure 1.7 Architecture transformers tirée de Vaswani *et al.* (2017)

Le mécanisme d'attention 1.8 représente l'unité de calcul fondamentale de l'attention au sein du transformeur. Sa fonction est de mapper un vecteur de requête (Query) et un ensemble de paires clé-valeur (Key-Value) à un vecteur de sortie. La sortie est calculée comme une somme pondérée des vecteurs de valeur, où le poids attribué à chaque valeur est déterminé par une fonction de compatibilité entre la requête et la clé correspondante. Cette compatibilité est évaluée par un produit scalaire. L'algorithme calcule les produits scalaires de la requête avec toutes les clés, puis chaque résultat est divisé par un facteur d'échelle, $\sqrt{d_k}$, où d_k est la dimension des vecteurs de clé. Cette mise à l'échelle a pour objectif de contrer les effets de saturation de la fonction softmax lorsque les dimensions sont grandes, stabilisant ainsi le processus d'apprentissage. Une fonction softmax est ensuite appliquée pour obtenir une distribution de probabilités sur les

vecteurs de valeur, représentant les poids d'attention. Le résultat final est la somme des vecteurs de valeur, pondérée par ces poids.

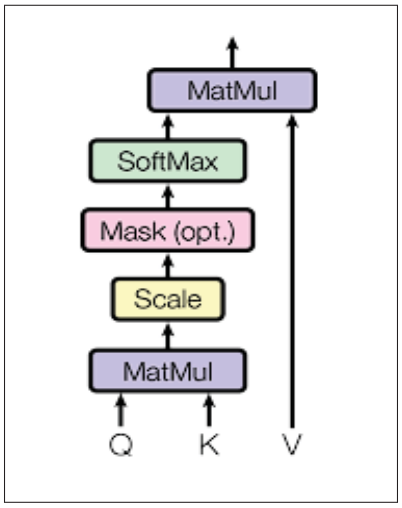


Figure 1.8 Mécanisme d'attention tirée de Vaswani *et al.* (2017)

Le mécanisme d'attention multi-têtes 1.9 est une évolution de l'attention à produit scalaire simple, conçu pour permettre au modèle de capturer conjointement des informations provenant de différents sous-espaces de représentation à diverses positions. Au lieu d'exécuter un seul mécanisme d'attention, les entrées requêtes, clés et valeurs sont d'abord projetées linéairement h fois à l'aide de différentes projections paramétrées. Chacune de ces versions projetées est ensuite traitée en parallèle par un mécanisme d'attention à produit scalaire distinct, formant ainsi h « têtes » d'attention. Chaque tête peut ainsi se spécialiser dans la détection de types de relations différents au sein de la séquence. Les sorties des h têtes sont ensuite concaténées, puis soumises à une projection linéaire finale pour produire le vecteur de sortie de la dimension attendue. Cette parallélisation permet au modèle d'intégrer une richesse d'informations contextuelles plus grande à chaque position de la séquence.

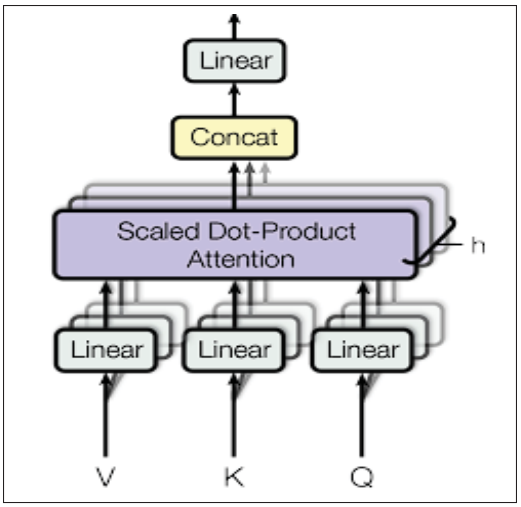


Figure 1.9 Mécanisme d'attention multi-têtes tirée de Vaswani *et al.* (2017)

1.2.6 Notre approche proposée : les modèles espace-états

Les modèles espace-états, plus connus sous le nom de SSM (State Space Models), sont des équations très utilisées dans la théorie du contrôle ainsi que dans les systèmes linéaires dynamiques. Ces modèles offrent à la fois une formulation continue et une autre discrète, proposant ainsi une alternative parallélisable aux réseaux récurrents pour la modélisation des séries temporelles.

$$\begin{aligned} \frac{dh(t)}{dt} &= Ah(t) + Bx(t), \\ y(t) &= Ch(t) + Dx(t), \end{aligned} \tag{1.20}$$

où $h(t)$ représente l'état caché, $x(t)$ est le signal d'entrée, et $y(t)$ est la sortie. Le système est régi par les matrices suivantes :

$$A \in \mathbb{R}^{N\times N}, \quad B \in \mathbb{R}^{N\times 1}, \quad C \in \mathbb{R}^{1\times N}, \quad D \in \mathbb{R} \tag{1.21}$$

Ces matrices définissent respectivement les transitions d'état (A), la projection de l'entrée (B), la projection de la sortie (C) et les connexions directes (D). Ce qui rend ce système particulièrement intéressant, c'est qu'il peut être exprimé aussi bien sous la forme d'une opération de convolution que sous celle d'une relation récurrente. La solution de l'équation différentielle 1.20 est connue et prend la forme :

$$h(t) = e^{At}h(0) + \int_0^t e^{A(t-\tau)}Bx(\tau)\,d\tau \tag{1.22}$$

Avec e^{At} est l'exponentielle de la matrice At.

En posant $h(0) = 0$:

$$h(t) = B\int_0^t e^{A(t-\tau)}x(\tau)\,d\tau. \tag{1.23}$$

En remplaçant dans 1.20 nous trouvons :

$$y(t) = \int_0^t \left(Ce^{A(t-\tau)}B\right)x(\tau)\,d\tau + Dx(t). \tag{1.24}$$

Nous retrouvons l'équation de la convolution présentée dans l'équation 4.9 qu'on peut réécrire :

$$y(t) = \int_0^t x(\tau)K(t-\tau)\,d\tau + Dx(t). \tag{1.25}$$

Avec : $K(t) = Ce^{A(t-\tau)}B$, la sortie donc de notre système est une convolution globale entre l'entrée $x(t)$ et le noyau apprenable$K(t)$.

$$y(t) = (K * x)(t) + Dx(t) \tag{1.26}$$

Pour intégrer ce système dans un réseau de neurones, il est nécessaire de le discrétiser en utilisant la transformée bilinéaire.

D'autre part, on peut écrire l'équation 1.20 sous sa forme récurrente. Nous faisons appel à la théorie du contrôle et, plus précisément, nous transformons l'équation-état en 1.20 dans le

domaine de Laplace :

$$sH(s) - h(0) = AH(s) + BX(s) \tag{1.27}$$

$$H(s) = (sI - A)^{-1}BX(s). \tag{1.28}$$

Nous trouvons la fonction de transfert $G(s)$ donnée donc par :

$$G(s) = \frac{H(s)}{X(s)} = (sI - A)^{-1}B \tag{1.29}$$

en substituant par :

$$s = \frac{2}{\Delta}\frac{z-1}{z+1} \tag{1.30}$$

Cette substitution s'appelle la transformée de Laplace et sert a faire la correspondance entre le domaine continu et le domaine discret avec :

- s est la variable de Laplace ,
- z est la transformée en z,
- Δ l'intervalle d'échantillonnage.

après manipulations algébriques nous trouvons :

$$A_d = \left(I - \frac{\Delta}{2}A\right)^{-1}\left(I + \frac{\Delta}{2}A\right), \tag{1.31}$$

$$B_d = \left(I - \frac{\Delta}{2}A\right)^{-1}\Delta B. \tag{1.32}$$

Ces matrices forment la version discrète des équation-états :

$$\begin{aligned} h_k &= A_d h_{k-1} + B_d x_k \\ y_k &= Ch_k + Dx_k \end{aligned} \tag{1.33}$$

a partir de l'équation discrète 4.14 on peut retrouver la forme de la convolution du domaine continue , en développant récursivement l'équation d'état discrète et en réitérant :

$$h_{k-1} = A_d h_{k-2} + B_d x_{k-1}, \tag{1.34}$$

nous obtenons, par substitution :

$$h_k = A_d^k h_0 + \sum_{j=0}^{k-1} A_d^j B_d x_{k-j}. \tag{1.35}$$

Avec :

$$K_d[j] = C A_d^j B_d. \tag{1.36}$$

Cette équation montre que la sortie est obtenue via une convolution discrète avec un noyau $K_d[j]$ dépendant des paramètres du système. En effet, Pour une séquence de longueur L, nous pouvons écrire explicitement la forme du noyau K_d :

$$K_d = \begin{bmatrix} CB_d \\ CA_dB_d \\ CA_d^2B_d \\ \vdots \\ CA_d^{L-1}B_d \end{bmatrix}. \tag{1.37}$$

Donc, la convolution entre l'entrée x et le noyau K_d s'écrit :

$$y_k = \sum_{j=0}^{L-1} K_d[j] x_{k-j} + D x_k. \tag{1.38}$$

La formule 1.38 montre que la sortie y_k est obtenue via une convolution discrète avec un noyau $K_d[j]$, qui encode l'évolution temporelle des états en fonction des paramètres A_d, B_d et C.

Les modèles espace-états sont une alternative prometteuse et flexible pour la modélisation des séries temporelles, en combinant les avantages des systèmes dynamiques linéaires et des réseaux de neurones modernes. Contrairement aux réseaux récurrents (RNNs) et aux Transformers,

les SSMs sont fondés sur une formulation mathématique rigoureuse, ce qui leur confère des propriétés uniques en termes de stabilité, d'efficacité computationnelle et de capacité à capturer des dépendances à long terme.

D'une part, les réseaux récurrents souffrent de l'instabilité de l'entraînement à cause du problème d'explosion ou de disparition du gradient, principalement dû aux non-linéarités répétitives. En revanche, les SSM garantissent la stabilité du gradient en procédant par des transitions linéaires entre les états cachés, même pour des séquences très longues. Mais aussi, la nature séquentielle des réseaux récurrents les rend très coûteux en termes de temps d'inférence et de mémoire lorsqu'il s'agit de traiter de longues séquences d'entrée. En revanche, les SSM offrent la possibilité de paralléliser les calculs grâce à leur formulation convolutionnelle des équations d'état. La parallélisation devient possible grâce à la transformée de Fourier rapide (FFT), ce qui permet de réduire la complexité de $O(L)$ à $O(L \log L)$. D'autre part, en ce qui concerne les transformers, ceux-ci présentent une complexité quadratique $O(L^2)$ en raison de leur mécanisme d'attention, ce qui les rend coûteux pour les longues séquences.

Tableau 1.2 Comparaison des avantages et limitations des SSMs, RNNs et Transformers.

Aspect	SSM	RNN	Transformeurs
Complexité Temporelle	$O(L \log L)$	$O(L)$	$O(L^2)$
Complexité en Mémoire	$O(L)$	$O(L)$	$O(L^2)$
Stabilité des Gradients	Stable (linéarité)	Instable	Stable (attention normalisée)
Longues Séquences	Excellente (noyau global)	Limitée	Bonne (attention globale)
Calcul Parallèle	Oui	Non (séquentiel)	Oui (attention parallèle)
Interprétabilité	Élevée	Faible	Faible
Non Linéairité	Limitée	Bonne (non-linéarités)	Excellente
Initialisation	**Délicate** (HiPPO requis)	Simple	Simple

.

CHAPITRE 2

REVUE DE LA LITTÉRATURE

2.0.1 Authentification Continue Basée sur la Biométrie Comportementale

Les travaux récents en authentification continue s'orientent de plus en plus vers des approches multimodales et des architectures avancées capables d'exploiter efficacement les dépendances temporelles et spatiales dans les données biométriques comportementales. Par exemple, Senerath *et al.* (2023) introduit *Behaveformer*, un cadre exploitant une architecture Transformer à double attention spatio-temporelle (STDAT) pour l'authentification basée sur la dynamique de frappe et les données issues de capteurs inertiels. La branche d'attention spatiale modélise les relations inter-caractéristiques à chaque instant (par exemple entre les axes d'un accéléromètre), tandis que la branche temporelle capture les dépendances séquentielles du comportement. En séparant explicitement les domaines spatial et temporel, cette architecture parvient à isoler des motifs subtils propres à chaque individu, atteignant un EER de 2,95% sur *HuMIdb* Acien, Morales, Fierrez, Vera-Rodriguez & Delgado-Mohatar (2020) et 1,80% sur *Aalto DB* Palin, Feit, Kim, Kristensson & Oulasvirta (2019), surpassant les approches uni- et multimodales existantes.

D'autres travaux se sont focalisés sur l'exploitation des signaux de marche via des architectures spécialisées. M-GaitFormer Delgado-Santos, Tolosana, Guest, Deravi & Vera-Rodriguez (2023a) adopte une approche à deux flux Transformer pour le traitement conjoint des données d'accéléromètre et de gyroscope. Le flux temporel capte la dynamique de la démarche, tandis que le flux canal analyse les interdépendances entre capteurs. L'utilisation d'un mécanisme d'auto-corrélation basé sur la FFT exploite la nature périodique de la marche humaine. L'architecture intègre également un bloc CNN multi-échelle et un module Transformer récurrent pour préserver les dépendances temporelles longues, obtenant un EER de 3,42% sur *whuGAIT* Zou, Wang, Wang, Zhao & Li (2020) et 2,90% sur *OU-ISIR* Iwama, Okumura, Makihara & Yagi (2012), dépassant nettement les méthodes CNN, LSTM et hybrides CNN-LSTM.

Historiquement, les approches d'authentification continue reposaient sur deux grandes catégories : les méthodes de *template matching* et celles basées sur l'apprentissage automatique classique. Les premières comparent les nouvelles données à un gabarit stocké à l'aide de mesures de similarité comme le *dynamic time warping* Muller (2007), la corrélation croisée Wren, Wren, Do, Rethlefsen & Healy (2006) ou le coefficient de corrélation de Pearson Berman (2016). Les secondes, souvent fondées sur des SVM ou kNN Nickel, Wirtl & Busch (2012), nécessitent un important travail d'ingénierie de caractéristiques, incluant par exemple des transformations de Fourier ou d'ondelettes Watanabe (2014).

L'émergence de l'apprentissage profond (DL) a marqué un tournant, en permettant l'extraction automatique de représentations complexes à partir de données brutes. Les travaux de Zeng *et al.* (2014) ont été pionniers dans l'utilisation de réseaux de neurones convolutionnels (CNN) pour l'extraction de caractéristiques de la démarche humaine, surpassant les méthodes traditionnelles. Néanmoins, les techniques de segmentation par *template matching* utilisées dans cette étude sont coûteuses en mémoire et peu adaptées au calcul parallèle. En réponse à cela, Ronneberger, Fischer & Brox (2015) propose une extraction inspirée de l'architecture U-Net avec convolutions 1D, offrant un traitement parallélisable et une précision supérieure à 93,5

Enfin, pour réduire la dépendance aux jeux de données annotés, particulièrement coûteux dans le domaine des signaux de capteurs, plusieurs études se sont tournées vers l'apprentissage auto-supervisé. Dans Stragapede *et al.* (2022), HuMINet exploite le jeu de données *HuMIdb* Acien *et al.* (2020) et entraîne des RNN séparés pour chaque modalité biométrique à l'aide d'une perte *triplet*. La fusion des modalités (frappes fixes, gestes, données inertielle) au niveau des scores améliore nettement les performances, atteignant un EER compris entre 4% et 9% sur des fenêtres de 3 secondes.

2.0.2 Reconnaissance d'Activités Humaines

Les premières approches en HAR reposaient principalement sur des méthodes classiques telles que les machines à vecteurs de support (SVM), les forêts aléatoires (RF) ou les k-plus proches

voisins, combinées à des caractéristiques conçues manuellement à partir des signaux capteurs Rojanavasu, Jantawong, Jitpattanakul & Mekruksavanich (2023); Deep & Zheng (2019). Les données d'accéléromètre étaient découpées en fenêtres temporelles (3 à 10 secondes) pour extraire des mesures statistiques et fréquentielles (moyenne, variance, énergie, pics spectraux, etc.). Bien que légères, interprétables et performantes sur de petits ensembles de données, ces méthodes présentaient des limites de généralisation vers de nouveaux utilisateurs ou contextes, notamment dans des environnements où plusieurs activités pouvaient se chevaucher Zhou, Wang, Su & Xu (2022).

Avec l'essor de l'apprentissage profond, la tendance s'est orientée vers des architectures capables de modéliser automatiquement la complexité temporelle et la variabilité des activités. Par exemple, Li *et al.* (2021) propose THAT, une architecture double flux séparant le traitement temporel et celui des canaux à travers des tours MCAT (*Multi-scale Convolution Augmented Transformer*). Ces modules combinent auto-attention multi-tête et convolutions multi-échelles adaptatives, tout en intégrant un encodage gaussien de plage temporelle pour distinguer des actions réversibles comme "s'asseoir" et "se lever". THAT atteint des performances supérieures à celles des modèles classiques et DL, avec une précision moyenne de 98,6% sur des environnements intérieurs variés.

Parallèlement, les réseaux convolutionnels (CNN) se sont imposés comme référence pour le traitement de séquences locales. Des optimisations comme la normalisation par lot, les connexions résiduelles Mekruksavanich & Jitpattanakul (2023) ou les convolutions dilatées Hamad, Kimura, Yang, Woo & Wei (2021) ont permis d'élargir le champ réceptif et de stabiliser l'entraînement. Toutefois, les CNN analysent souvent des fenêtres courtes et peuvent perdre des détails fins nécessaires pour distinguer des activités proches ("s'asseoir" vs. "parler assis") Park, Kim & Lee (2024).

Certaines approches ont cherché à combiner l'apprentissage local et global. Dans Mahmud *et al.* (2020), une architecture basée sur l'auto-attention intègre un mécanisme spécifique à

chaque modalité, un encodage positionnel pour préserver l'ordre temporel, ainsi qu'une attention temporelle globale pour capter les dépendances à long terme, atteignant 96

L'exploitation multi-échelle a également été explorée. L'architecture DKInception Yu & Alqaness (2025) applique des convolutions parallèles de tailles variées (1×3 à 1×9) pour extraire simultanément des motifs temporels fins et étendus, avant de fusionner les cartes de caractéristiques via un module d'auto-attention. Cette approche obtient 89% de précision sur PAMAP2 Reiss (2012).

Enfin, l'apprentissage auto-supervisé (SSL) a récemment ouvert de nouvelles perspectives. Dans Yuan *et al.* (2024), les auteurs exploitent de larges ensembles de données comme *UK Biobank* pour préentraîner un modèle CNN multitâche avec des tâches auxiliaires (permutation, flèche du temps, déformation temporelle). Cette stratégie réduit la dépendance aux données annotées et améliore les performances sur différents jeux de données, avec des gains en F1-score allant jusqu'à 100% selon le contexte, y compris sur Capture-24 Chan *et al.* (2024).

En résumé, les approches existantes en HAR oscillent entre CNN locaux à faible coût mais limités en contexte temporel, et architectures à attention coûteuses en calcul. Notre approche, basée sur des modèles à espaces d'états structurés, conserve la capacité à modéliser des dépendances temporelles longues avec une complexité linéaire et une empreinte mémoire réduite, offrant ainsi une solution adaptée à l'inférence embarquée tout en assurant une généralisation robuste sur des environnements contrôlés (PAMAP2) et réels (Capture-24).

CHAPITRE 3

S4HI : A NOVEL LEARNING-BASED HUMAN IDENTIFICATION METHOD FROM BEHAVIOURAL DATA

Khalil Snoussi[1] , Wael Jaafar[1] , Rami Langar[1 2]

[1] Département de Génie logiciel et TI, École de Technologie Supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3
[2] LIGM-CNRS UMR 8049, University Gustave Eiffel,
F-77420 Marne-la-Vallée, France

Article soumis et accepté à la Conférence « CIOT » en octobre 2024.

3.1 Abstract

Behavioral biometrics have recently gained attention as a complementary key enabler of security and authentication, owing to the recent success of different deep learning architectures. State-of-the-art works explored the spatial information of gait signals using temporal features with convolutional or recurrent neural networks, or through combinations of both. At the same time, the recent success of transformers led to their use for gait-based identity recognition. However, their efficient use is hindered by the structure of data, their features, and their long training times. Alternatively, in this paper, we propose a novel identity recognition solution that relies on structural state space models to handle the long range dependencies of the inertial measurement unit (IMU) data and leverages the strengths of both convolutional and recurrent architectures. This duality provides a robust mechanism for handling sequential data by utilizing global convolutional views during training and recurrent views during inference realizing fast training, while the recurrent structure ensures fast inference. Through experiments, we demonstrate the superiority of our method in terms of accuracy (best 95.6%), complexity (up to 7 times less complex), and training time (up to 50 times faster), compared to the Transformer and other benchmarks, and for two different IMU datasets.

3.2 Introduction

Various biometric characteristics, including retinal scans, fingerprints, voice, and face features, are frequently employed for authentication purposes. Nevertheless, these methods have some limitations, such as serious privacy concerns with how data is stored and used, improper use of personal data, and identity theft Salama *et al.* (2023). Due to the increasing concerns over duplication and hacking, there has been a rising interest in behavioral biometrics within the cybersecurity area. The modalities of behavioral biometrics include those based on phone movements, touchscreen interaction behavior, and keyboard usage behavior. These methods utilize the distinct individual habits of the user, making it more difficult for counterfeiters or hackers to imitate or breach the system and offering a relevant response to authentication systems security. In addition, the gathering of traditional biometric data (non-behavioral), such as facial images, retinal scans, voice, and fingerprints, necessitates a significant commitment from users, thus undermining users' privacy. Alternatively, authentication based on behavioral biometrics requires the acquisition of data that is non-intrusive and easily understood. In this context, several researchers started investigating the use of phone movement-based biometrics, with a particular emphasis on gait motions. Indeed, authors of Park, Lee & Koo (2021) suggested that gait has the potential to be effective given its uniqueness. Prior studies on gait identification utilizing inertial measurement units (IMUs) can be classified into two distinct categories : 1) Approaches centered on *template matching* and, 2) machine learning *(ML)*-based approaches. The template matching-based methods identify users by quantifying the similarity between the current time series obtained from IMU sensors and previously recorded ones. If the similarity level is above a predefined threshold, then the current user is identified as genuine. The main manners to quantify similarity encompass dynamic time warping (DTW) Muller (2007), cross-correlation Wren *et al.* (2006), and Pearson correlation coefficient Berman (2016). In contrast, ML-based techniques rely mainly on unsupervised learning such as support vector machines (SVM) and k-nearest neighbors (K-NN) Nickel *et al.* (2012), with a strong emphasis on feature engineering. Specifically, these methods calculate additional characteristics based on raw IMU data, such as velocity and orientation. Although template matching methods are more robust than SVM

and K-NN, they still depend on the quality of determined features and data collection technique Watanabe (2014).

On the other hand, recent research indicated that deep learning (DL) is also suitable for gait recognition, given that neural networks (NNs) have an excellent capacity to capture nonlinear information. Indeed, NNs can independently and without feature engineering learn complex features from raw IMU data. The ability to learn complex non-linear representations of data provides precise and robust identification of walking patterns, even when the data may differ due to factors like walking speed, surface conditions, or device orientation. Authors of Zeng *et al.* (2014) pioneered the utilization of NNs to extract features from human gait. They proved the superiority of NNs over traditional approaches, in terms of classification rate (i.e., data corresponding to which individual). The study also presented a gait cycle extractor that uses template matching to identify and extract gait cycles. The process entails employing a low-pass filter with a cutoff frequency of 3 Hz to eradicate noise in accelerometer results and finally, a similarity measure is utilized to detect matches between distinct cycles, choosing the one that maximizes the similarity coefficient. Despite the efficiency of the above technique, it consumes a significant amount of memory proportional to the size of time series-based data. This makes the inference speed of models particularly slow, especially on smartphones. In addition, its iterative nature prevents parallelized computing, which could otherwise speed up the process. The method in Ronneberger *et al.* (2015) extracted gait cycles with segmentation similar to the U-net architecture and reduced the complexity using one-dimensional convolutions. This parallelizable approach accurately identified and authenticated individuals based on their gait patterns in natural conditions, with an accuracy above 93.5%. Behavioral biometrics primarily employs a two-stage method, consisting of *activity recognition* and then *identity recognition*. Although effective, especially when applied to gait data from sources like the WhuGait dataset, this technology frequently depends on computationally demanding segmentation architectures like U-nets for activity detection Zeng *et al.* (2014); Ronneberger *et al.* (2015). These requirements impede the ability to make immediate deductions on commonly used mobile devices. To overcome this constraint, the current study investigates a streamlined framework that can achieve

quicker inference on mobile/IoT devices and a cohesive single-step method for behavior-based identification. Specifically, a machine learning model architecture that shows promise in this regard is the S4 model, which is based on structural state space modeling Gu, Goel & Re (2022a).

Despite these results, additional gains in behavioral biometrics, particularly in gait recognition with phone movement data, can be further achieved using more advanced techniques. However, we need to address the following challenges. First, given that IMU data should be collected continuously, processing these data with the current state-of-the-art algorithms and techniques can be very challenging even when they suggest satisfactory results, since processing very long sequences of IMU data may take a substantial amount of energy and processing power on mobile phones. Also, IMU data contains a lot of noisy measurements due to different factors (magnetic interference, temperature change, mechanical vibrations, etc.), and finally IMU data contains temporal dependencies as the state of the user motion depends on their past states. To overcome these limitations, we propose in this paper a new approach based on Structural State Space Models (SSMs) Gu *et al.* (2022a) to further improve human identification performance using behavioral data. To the best of our knowledge, this is the first work to propose the use of SSMs for IMU data analysis, aiming to identify individuals, using only phone motion data. Also, to make it suitable for limited-capacity devices such as mobile and Internet-of-things (IoT) gadgets, we proposed the use of a lightweight SSM-based method, called S4HI, which can accelerate training without sacrificing human identification accuracy. Through extensive results, we demonstrate the superiority of our approach in terms of accuracy, complexity, and execution time, compared to other benchmarks.

The remainder of the paper is organized as follows. Section II provides the background information on the SSM-based model. In section III, we present the results of the experiments conducted, in terms of accuracy, training speed, and inference speed. Finally, section IV concludes the paper.

3.3 Proposed SSM-based method : S4HI

3.3.1 Background of State Spaces and S4 Model

Behavioral biometrics are based on the analysis of patterns in sensor data over a period of time. The history of an input function denoted as $f(x)|_{x \leq t}$ $(\forall t \geq 0)$, is of crucial significance when considering sensor functions such as accelerometer readings. As the quantity of collected data increases, storing the complete history of the data becomes difficult. The Hippo framework tackles this issue by transforming the signal history into an N-dimensional space using orthogonal polynomials Gu, Dao, Ermon, Rudra & Ré (2020b). This technique compresses the signal's past data into a vector with N dimension and analyzes its variations over time within a linear time-invariant system.

Specifically, assuming that $f(x)$ is the input measure function, we aim to approximate it on a fixed values interval $[a, b]$. Let $\phi_n(x)_{n=0}^{\infty}$ be a set of orthogonal polynomials defined on the same interval, such that :

$$< \phi_b, \phi_m >= \int_a^b \phi_m(x)\phi_n(x)\mathrm{d}x = \begin{cases} 0, & \text{if } m \neq n \\ h_n, & \text{if } m = n, \end{cases} \tag{3.1}$$

where $h_n = \int_a^b \phi_n(x)^2 \, dx$ is the normalization constant that ensures the orthogonality condition. Hippo objective is to determine c_n coefficients $(\forall n = 0, \ldots, N)$ such that $f(x) \approx f_N(x) = \sum_{n=1}^{N} c_n \phi_n(x)$. To obtain the optimal solution, we minimize the mean square error (MSE) defined as $\int_a^b \left[f(x) - f_N(x)\right]^2 \mathrm{d}x$. Indeed, the c_n coefficients are obtained by projecting $f(x)$ onto the orthogonal polynomials $\phi_n(x)$ dimensions as follows :

$$c_n = \frac{1}{h_n} \int_a^b f(x)\phi_n(x)\mathrm{d}x, \ \forall n = 1, \ldots, N. \tag{3.2}$$

The accuracy of the approximation depends on the number of terms N used in the series expansion and the properties of the selected orthogonal polynomials. Increasing N would

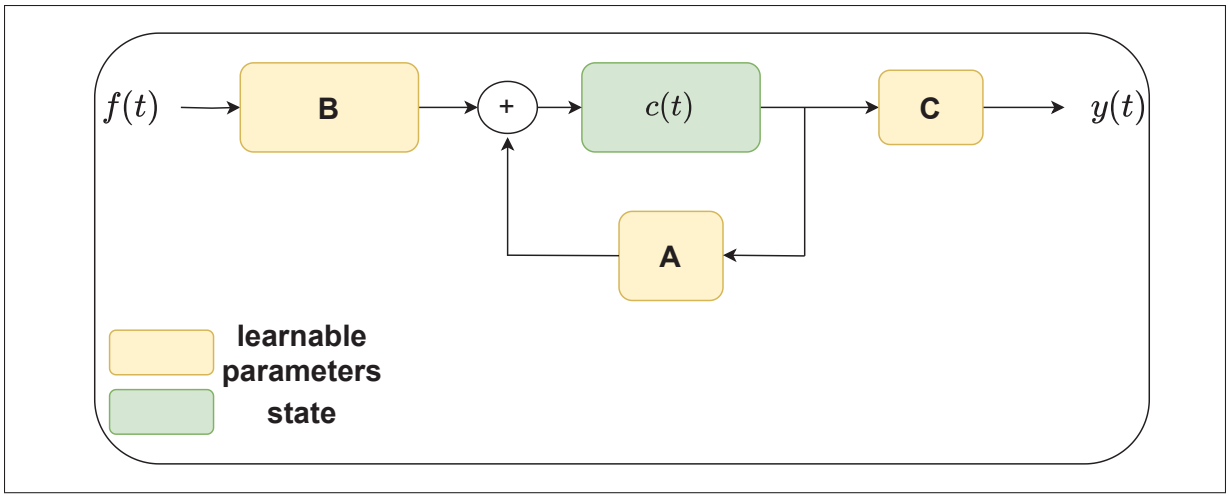


Figure 3.1 Diagram of an SSM model.

generally improve the approximation's accuracy but at the cost of increased computational complexity.

The state space model can be defined by the following equations :

$$\mathbf{c}'(t) = \frac{d\mathbf{c}(t)}{dt} = \mathbf{A}\mathbf{c}(t) + \mathbf{B}f(t), \tag{3.3}$$

$$y(t) = \mathbf{C}\,\mathbf{c}(t), \tag{3.4}$$

where the 1-dimensional (1D) input signal $f(t)$ is mapped to an N-dimensional (ND) latent state $\mathbf{c}(t)$, prior to projecting it to a 1D output signal $y(t)$. $\mathbf{A} \in \mathbb{C}^{N\times N}$, $\mathbf{B} \in \mathbb{C}^{N\times 1}$, and $\mathbf{C} \in \mathbb{C}^{1\times N}$. The SSM system in eqs.(3.3)–(3.4) is also equivalent to a convolution with kernel function $K(t)$, such that :

$$\mathbf{K}(t) = \mathbf{C}e^{t\mathbf{A}}\mathbf{B}, \tag{3.5}$$

$$y(t) = \mathbf{K}(t) * f(t), \tag{3.6}$$

where $(*)$ is the convolution operation.

When dealing with discrete time series data, i.e., a sample data series $\mathbf{f} = (f_0, f_2, \ldots, f_{L-1})$ rather than $f(t)$, the above equations should be discretized at a step Δ, which represents the measuring granularity or periodicity of the function $f(t)$. Consequently, the corresponding discrete SSM can be written as follows :

$$c_k = \overline{\mathbf{A}} c_{k-1} + \overline{\mathbf{B}} f_k, \tag{3.7}$$

$$y_k = \overline{\mathbf{C}} c_k, \tag{3.8}$$

where $\overline{\mathbf{A}} = (\mathbf{I} - \Delta/2\mathbf{A})^{-1}(\mathbf{I} - \Delta/2\mathbf{A})$, $\overline{\mathbf{B}} = \Delta(\mathbf{I} - \Delta\mathbf{A})^{-1}\mathbf{B}$, and $\overline{\mathbf{C}} = \mathbf{C}$ are the bilinear transform of $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$, respectively. $\mathbf{I}$ is the identity matrix and $(\cdot)^{-1}$ is the inverse matrix operation.

Eqs. (3.7)–(3.8) represent a sequence-to-sequence mapping, where (3.7) is a recurrence of c_k, thus enabling the discrete SSM to be processed as a recurrent neural network (RNN) and learn the matrices $\overline{\mathbf{A}}$, $\overline{\mathbf{B}}$, and $\overline{\mathbf{C}}$, as illustrated in Fig. 4.1. SSM equations are powerful because they can be viewed as both convolution Eqs. (5)-(6) and recurrent Eqs. (7)-(8) taking the best of both worlds. Indeed, the convolution view is favored for accelerated training, while the recurrent view achieves faster inference.

In the particular case of the S4 model, matrices $\overline{\mathbf{A}}$, $\overline{\mathbf{B}}$, and $\overline{\mathbf{C}}$ can be simply set up using the HiPPO scaled Legendre measure (HiPPO-LegS) approach such that :

$$\bar{A}_{ij} = \begin{cases} (2i+1)^{1/2}(2j+1)^{1/2} & \text{if } i > j, \\ i+1 & \text{if } i = j, \\ 0 & \text{if } i < j \end{cases} \tag{3.9}$$

$$\bar{B}_i = (2i+1)^{1/2}, \ \forall (i,j) \in \{1, \ldots, N\}^2. \tag{3.10}$$

The S4 model can capture long-range dependecies in sequences due to the initialization of its state transition matrix $\mathbf{A}$ derived from the highly accurate orthogonal polynomials approximators. It imposes a structure on its parameters hence the name structured state space models (S4)Gu *et al.* (2020b).

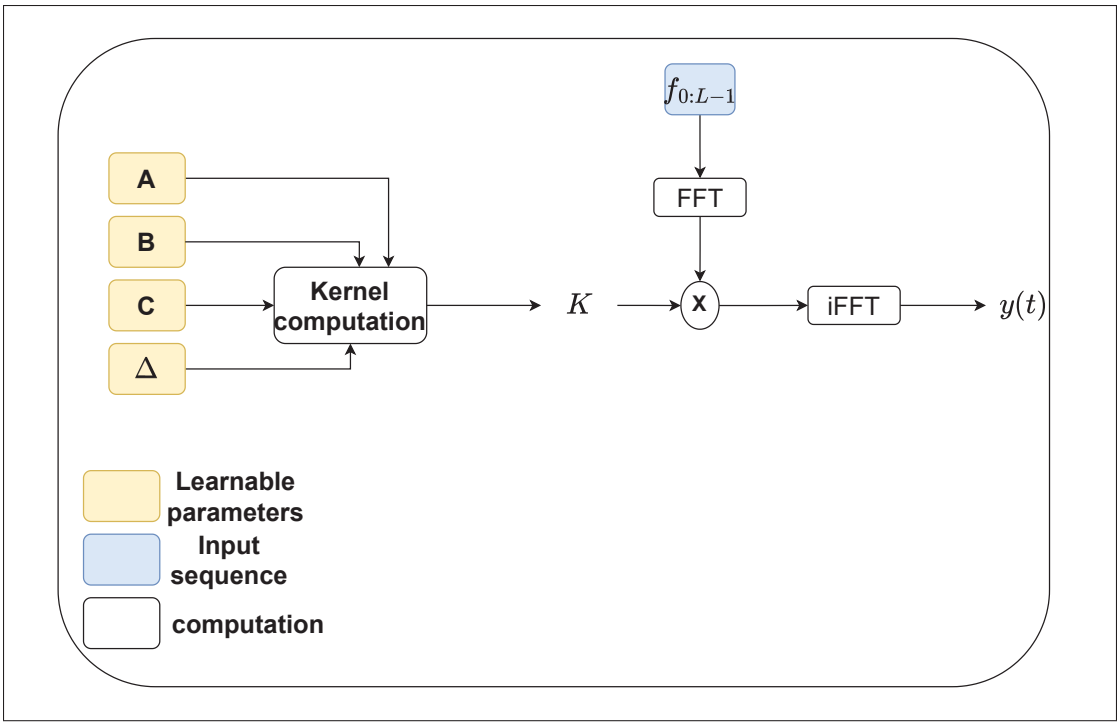


Figure 3.2 Diagram of the S4 model.

However, this structure comes at a cost. Indeed, calculating the kernel in (3.5) using the transition matrix in (3.9) would consume a lot of memory resources and computation power since it involves repeated matrix power. In this regard, the kernel computation block, shown in Fig. 3.2, calculates the kernel from (3.5) in its discrete form as follows :

$$\mathbf{K} = (\overline{\mathbf{CB}}, \overline{\mathbf{CAB}}, \overline{\mathbf{CA}}^2\overline{\mathbf{B}}, ..., \overline{\mathbf{CA}}^{L-1}\overline{\mathbf{B}}). \tag{3.11}$$

In Gu *et al.* (2022a), a solution based on calculating the kernel in the frequency domain has been proposed. It transforms the kernel $\mathbf{K}$ into a transfer function and the convolution operation into a multiplication operation using the Z-transform operator $\mathcal{Z}$, as shown below :

$$\mathcal{Z}(y(t)) = \mathcal{Z}(\mathbf{K}(t))\mathcal{Z}(f(t)), \tag{3.12}$$

$$Y(z) = H(z)F(z), \tag{3.13}$$

where $Y(z)$ and $F(z)$ are, respectively, the output and input in the frequency domain, while $H(z)$ is the transfer function of the SSM obtained by applying the Z-transform on the state equations (3.3)–(3.4) and is expressed by :

$$H(z) = \frac{Y(z)}{F(z)} = \mathbf{C}\left(z\mathbf{I} - \mathbf{A}\right)^{-1}\mathbf{B}. \tag{3.14}$$

Eq. (3.14) is convenient since it turns multiple powers of the transition matrix $\mathbf{A}$ into a single inverse operation. By applying the Woodbury identity and the Cauchy kernel Gu *et al.* (2022a), we can further simplify $H(z)$ since its second term $(z\mathbf{I} - \mathbf{A})^{-1}$ can be rewritten as :

$$\begin{aligned}
(z\mathbf{I} - \mathbf{A})^{-1} &= (z\mathbf{I} + \mathbf{D} - \mathbf{U}\mathbf{V})^{-1} \\
&= (z\mathbf{I} + \mathbf{D})^{-1} \\
&- \left(z\mathbf{I} + \mathbf{D})^{-1}\mathbf{U}\left(\mathbf{I} + \mathbf{V}\left(z\mathbf{I} + \mathbf{D}\right)^{-1}\mathbf{U}\right)^{-1}\right. \\
&\times \mathbf{V}\left(z\mathbf{I} + \mathbf{D}\right)^{-1}
\end{aligned} \tag{3.15}$$

where $\mathbf{A} = -\mathbf{D} + \mathbf{U}\mathbf{V}$, $\mathbf{D} = diag\left[\lambda_1, \ldots, \lambda_N\right]$, and $(\lambda_j)_{1 \le j \le N}$ are the eigenvalues of the diagonal matrix $\mathbf{D}$. The Fourier transform can be regarded as a particular instance of the Z-transform, obtained when the Z-transform is evaluated on the unit circle in the complex plane $z = e^{-j2k\pi/N}$ for $k = 1 \ldots N$. Consequently, to compute (3.13), we transform our input sequence $\mathbf{f} = (f_0, \ldots, f_{L-1})$ into the frequency domain using the fast Fourier transform algorithm (FFT) and multiply it by the computed transfer function $H(z = e^{-j2k\pi/N)}$ yielding the output $Y(z)$. Then, we apply the inverse fast Fourier Transform (iFFT) to obtain the output in the time-domain $y(t)$, as illustrated in Fig. 3.2.

3.3.2 Proposed S4HI Approach

In this work, we propose to set a diagonal state space model and prove its suitability for lightweight edge training. Our motivation stems from the fact that diagonal state space models can be as

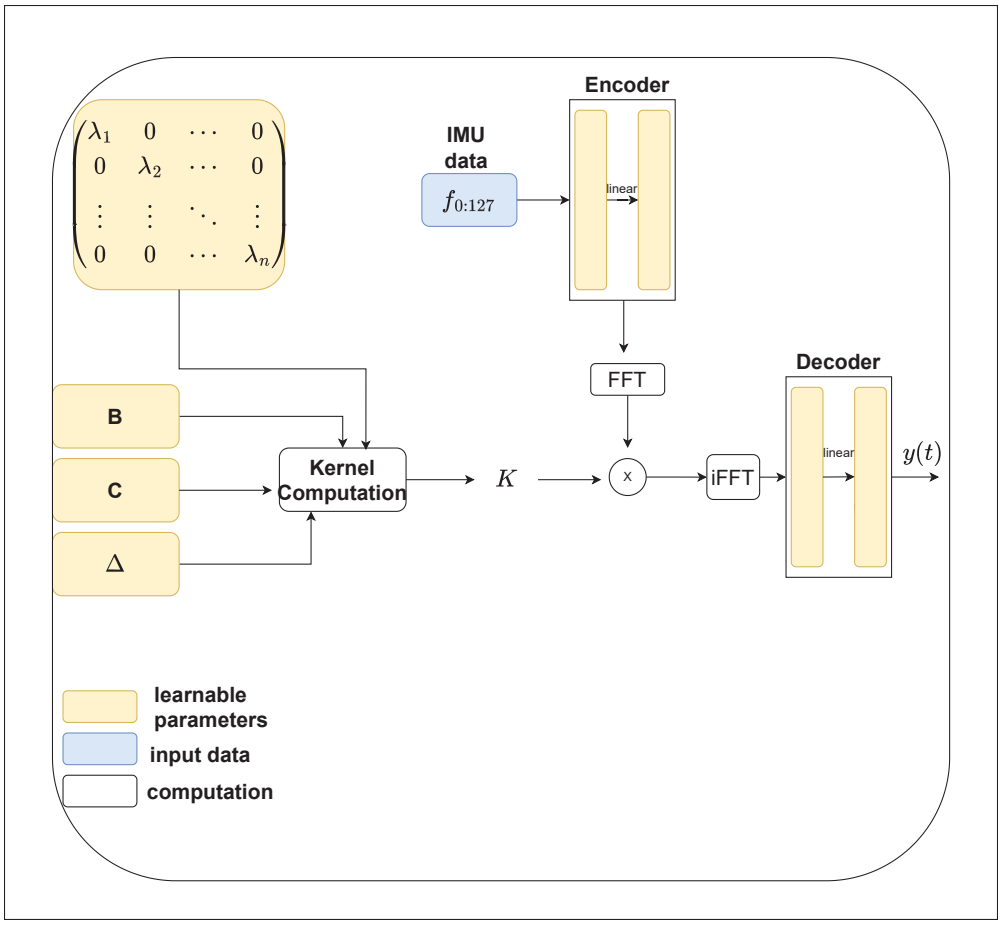


Figure 3.3 Diagram of the proposed S4HI model.

effective as structured SSMs, as demonstrated in Gupta, Gu & Berant (2022). On the other hand, kernel computation can be simplified using a diagonal state transition matrix. Indeed, the term $(z\mathbf{I} - \mathbf{A})^{-1}$ in the transfer function $H(z)$ is calculated straightforwardly without any optimization, given that it is simply equal to the Cauchy kernel $(z\mathbf{I} - \mathbf{A})^{-1} = diag(\frac{1}{z-\lambda_1}, \cdots, \frac{1}{z-\lambda_N})$. We initialize a diagonal state transition matrix, as follows :

$$\mathbf{A} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_N \end{bmatrix}, \tag{3.16}$$

where the diagonal values λ_i ($\forall i = 1, \ldots, N$) are complex. We use a 2-layer linear encoder, as shown in Fig. 3.3, to project the input sequences into an N-dimensional state space. Then, we use FFT to transform the input sequence into the frequency domain. The transformed expression is multiplied by the transfer function $H(z) = \mathbf{C}(z\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}$, which yields the output $Y(z)$. Subsequently, we apply iFFT and a 2-layer decoder to project the output back into its original dimension and get the time-domain output $y(t)$.

On the other hand, the solution to the differential equations in the diagonal case (3.5)–(3.6) is given by :

$$\mathbf{c}(t) = e^{\mathbf{A}t}\left(\mathbf{c}(0) + \int_0^t e^{-\mathbf{A}\tau}\mathbf{B}f(\tau)\,d\tau\right). \tag{3.17}$$

Given that $\mathbf{A}$ is diagonal and assuming the initial state $\mathbf{c}(0) = [0, \ldots, 0]$, after some manipulations, each element of $\mathbf{c}(t)$ can be written as :

$$c_j(t) = b_j e^{\lambda_j t}\int_0^t e^{-\lambda_j \tau} f(\tau)d\tau, \forall j = 1, \ldots, N, \tag{3.18}$$

where $\mathbf{B} = [b_1, \ldots, b_N]^t$ is a learnable vector which we initialize randomly from a normal distribution , and $(\cdot)^t$ is the transpose operator. We initialize the learnable parameters λ_j ($j = 1, \ldots, N$) in such a way that the state vector persists for a long time to capture long-range dependencies. Indeed, if we want our model to capture long-range dependencies in observed data, λ_j should be initialized such that the state vector $\mathbf{c}(t)$ stably persists over time. Specifically, a very slow decay on $e^{\lambda_j t}$ (Eq.17) should be achieved by selecting the real part of λ_j to be negative but very close to -1.

3.4 Experimental results

In this section, we present the results of our experiments conducted for different machine learning (ML)-based human identification methods. We compare our proposed S4HI model (HSS=64 or

128) [1] with four benchmarks including (i) the Informer model Zhou *et al.* (2020), (ii) THAT Transformer model Li *et al.* (2021), (iii) the MGaitFormer model Delgado-Santos *et al.* (2023a), and (iv) the conventional S4 model Gu, Goel & Ré (2022b). Experiments have been conducted on a laptop equipped with AMD Ryzen 7 5800 Hz CPU and an Nvidia GeForce RTX 3060 GPU.

To evaluate the effectiveness of the aforementioned approaches, we use data extracted from two publicly available walking time-series datasets, namely the OU-ISIR dataset from Osaka University Iwama *et al.* (2012), and the whuGait dataset from Wuhan University Zou *et al.* (2020). The former dataset is the largest from which we extracted data of 744 participants, each having 6 dimensions and a length of 128. In the latter dataset, we collected the inertial data of 20 individuals, gathered via smartphones in free-living conditions, including triaxial accelerometer and gyroscope data with a sampling frequency of 50 Hz.

Using the OU-ISIR dataset, we conduct experiments in Fig. 3.4 to determine the most suitable model structures for S4 and S4HI, to be used later. Specifically, we evaluated the accuracy performance as a function of the hidden state size (HSS) within the S4 and S4HI models. The accuracy is defined as follows :

$$Accuracy = \frac{TP + TN}{Total\ Sample\ Size}\ (\text{in } \%), \tag{3.19}$$

where TP and TN are the numbers of true positives and true negatives, respectively. Results show that, for a given HSS, the S4HI model outperforms the S4 one. This result highlights the potential of the proposed S4HI model compared to S4. For S4 (resp. S4HI), as the hidden state size increases from 64 to 256, the accuracy significantly improves from 91% (resp. 93%) to 94.1% (resp. 95.6%). However, a further increase of the hidden state size above 256 results in an accuracy degradation. Moreover, we can notice that, as the hidden state size grows, the computation complexity of the model (illustrated with colored “X” and the number above it) rises, for instance, from 4 to 24 Giga Floating Point Operations Per Second (GFLOPS) for S4,

[1] The selection of a hidden state size (HSS) of only 64 or 128 is motivated by the objective of designing a lightweight model that is suitable for deployment in IoT and smartphone devices with limited computing power.

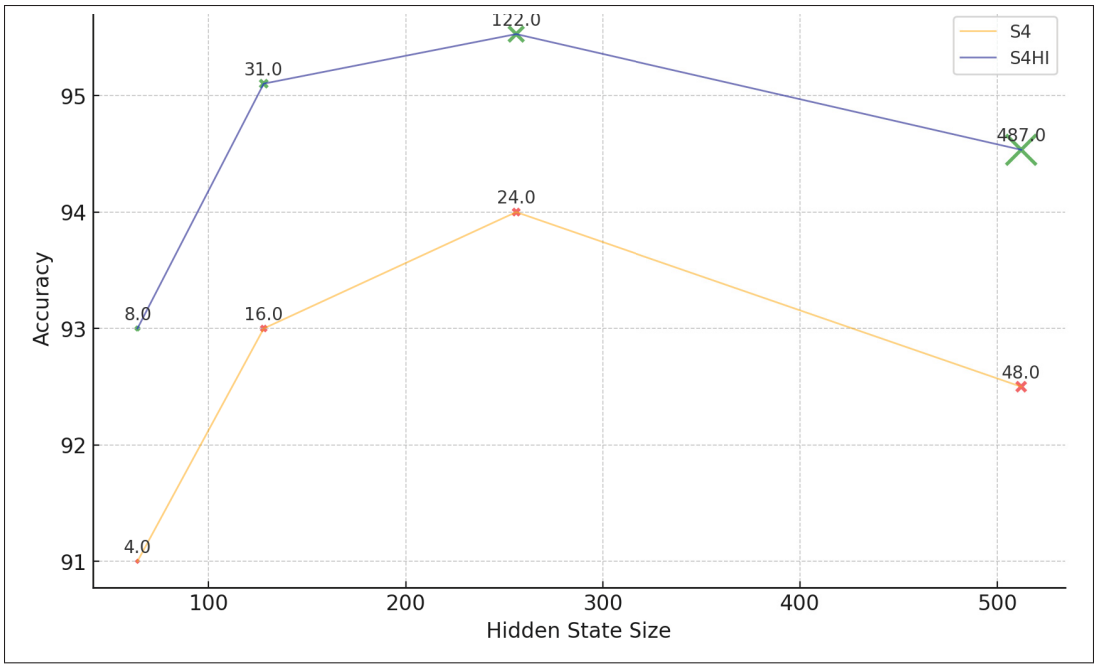


Figure 3.4 Accuracy vs. hidden state size and computational complexity using OU-ISIR dataset.

Tableau 3.1 Performance comparison of proposed and benchmark methods

Approach	**Accuracy**	**Complexity (GFLOPS)**	**Peak GPU Memory Usage**	**Duration of a training step**
S4HI (HSS = 128)	**95.6 %**	31	200 MB	**50 ms**
S4HI (HSS = 64)	93 %	**8**	**154 MB**	**50 ms**
S4 (HSS=256)	94.1 %	48	307 MB	320 ms
S4 (HSS=128)	93	24	210 MB	165 ms
MGaitFormer	93 %	52	400 MB	2700 ms
THAT Transformer	85.73 %	24	380 MB	2450 ms
Informer	59.4 %	16.83	410 MB	2500 ms

and from 8 to 122 GFLOPS for S4HI. Selecting a hidden state size of 256 for S4 ensures a good balance between accuracy (94.1%) and complexity (24 GFLOPS), while increasing the HSS beyond this value is not advantageous. Meanwhile, selecting HSS = 128 or HSS = 64 for the S4HI model is sufficient to achieve similar or better accuracy to S4 while maintaining the computation complexity at an acceptable or better level.

Tableau 3.2 Performances of S4HI (HSS = 64) using different datasets

Dataset	Accuracy	Precision	Recall	F1-score
WhuGait	97 %	98 %	97 %	97%
OU-ISIR	93 %	94 %	95 %	94 %

Table 3.1 compares the performances of different approaches in terms of accuracy, defined in Eq. (15), complexity (in GFLOPS) [2], peak GPU memory usage (in MegaBytes -MB), and the average duration of a training step (in milliseconds -ms), using the OU-ISIR dataset. From this table, we can see that the S4HI (HSS=128) approach outperforms all other methods in terms of accuracy (95.6%) while keeping a relatively equivalent complexity (31 GFLOPS) to that of THAT Transformer and S4 (HSS=128) but lower complexity than MGaitFormer. Moreover, it achieves a maximum GPU memory usage (200 MB) and an average training step time (50 ms) better than the other benchmarks. In contrast, the Informer method demonstrates the worst accuracy performance (59.4%) while being less complex (16 GFLOPS), consuming a high GPU memory (410 MB), and spending a long training time (2450 ms). Even though the THAT Transformer and MGaitFormer approaches do better in terms of accuracy (resp. 85.73% and 93%), they either have a relatively high complexity (e.g., 52 GFLOPS for MGaitFormer), a high GPU peak usage (around 400 MB), or long training times (above 2500 ms per training step. Finally, our proposed S4HI (HSS=64) method demonstrates a tradeoff between FL accuracy and cost. Indeed, it achieves an accuracy of 93%, which is 1.1% and 2.6% lower than S4 (HSS=256) and S4HI (HSS=128), with a low complexity of 8 GFLOPS, a low peak GPU usage of 154 MB, and with the fastest training time of 50 ms per step. The efficacy of our S4HI method makes it a suitable tool for deployment in edge devices equipped with sensors, such as IoT equipment, smartphones, or smartwatches.

In Table 3.2, we extend our evaluations of S4HI (HSS=64) using both OU-ISIR and WhuGait datasets, and evaluating, in addition to accuracy, the related precision, recall, and F1-score

[2] Complexity computation is conducted on a sampled tensor of shape $(1449, 128, 6)$.

performance metrics, where

$$Precision = \frac{TP}{TP + FP},\ Recall = \frac{TP}{TP + FN}, (\text{in } \%), \tag{3.20}$$

and

$$F1\ score = 2 * \frac{Precision\ *\ Recall}{Precision\ +\ Recall}, \tag{3.21}$$

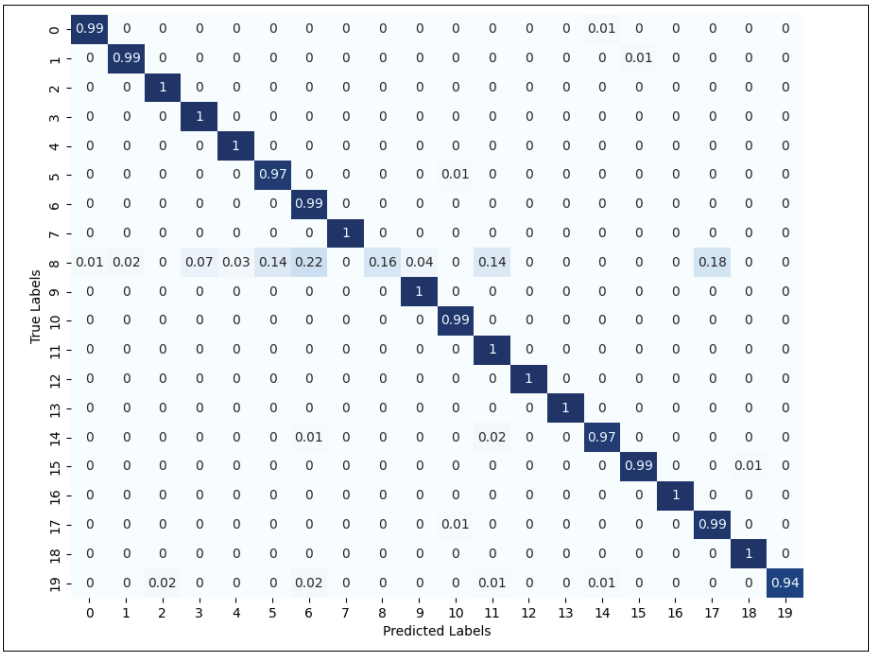
where FP and FN are the numbers of false positives and false negatives, respectively.

As shown in Table 3.2, S4HI shows high precision, precision, and F1 score in both datasets, suggesting that S4HI is capable of correctly classifying instances with a high positive rate, while minimizing false positives and false negatives. Nevertheless, we notice a slightly higher performance with the WhuGait compared to the OU-ISIR dataset. This is due to the small test dataset per individual of OU-ISIR, which deflates the performance results. Overall, these results suggest that the architecture of our model is highly suitable for gait recognition and authentication.

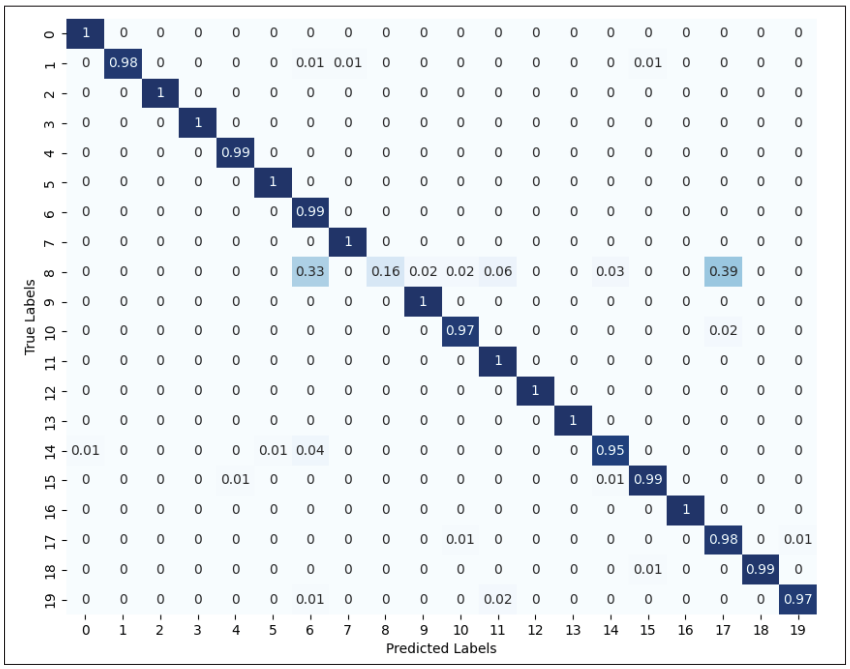
Finally, we present in Fig. 3.5 the confusion matrices for the S4 (HSS=256) and S4HI (HSS=64) methods when used with the WhuGait dataset. As it can be seen, most individuals have been correctly identified using S4 and S4HI. Also, we notice that both approaches performed badly to identify the individual with index 8. This is due to the low quality of collected data for this particular individual.

3.5 Conclusion

In this paper, we proposed a new model to identify humans from their behavioral data based on SSMs. Using two well-known datasets, we showed that our approach outperforms several benchmarks in classifying individuals based on their walking patterns, in terms of accuracy (95.6%), complexity (up to 7 times less complex), and training time (up to 20 times faster). Our approach represents therefore a promising solution for authentication systems that can be deployed on mobile devices since it uses very little memory during training. Further research



a) S4 (HSS=256) method



b) S4HI method (HSS=64)

Figure 3.5 Confusion matrices using WhuGait dataset.

would include testing our model on various gait patterns that capture a wide range of human activities beyond walking gait. In addition, we aim to combine our approach with other behavioral biometric identification to yield more remarkable results, thereby opening up new possibilities for innovative applications in behavioral identification, medical diagnostics, and other fields.

CHAPITRE 4

MEDUSAA : A MULTIMODAL ENCODER-DECODER USING STATE SPACES FOR CONTINUOUS HUMAN AUTHENTICATION AND ACTIVITY RECOGNITION

Khalil Snoussi[1] , Wael Jaafar[1] , Rami Langar[1 2]

[1] Département de Génie logiciel et TI, École de Technologie Supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3
[2] LIGM-CNRS UMR 8049, University Gustave Eiffel,
F-77420 Marne-la-Vallée, France

Article soumis au journal « IEEE TBIOM » en aout 2025.

4.1 Abstract

Smartphones have revolutionized personal information management by integrating powerful sensors and connectivity into everyday life. However, this convenience comes with growing security risks, as sensitive data is often protected by static, one-time authentication methods that are vulnerable to post-login attacks. While deep learning (DL) models, particularly those based on Transformer architectures, have shown promise for continuous authentication, their high computational demands make them impractical for deployment on smartphones or other edge devices. To address this gap, we propose MEDUSAA : a lightweight DL framework for continuous user authentication and human activity recognition (HAR) using inertial measurement unit (IMU) data from embedded smartphone sensors such as accelerometers and gyroscopes. Our approach continuously monitors user behavior to enhance security beyond traditional methods like personal identification numbers (PINs) and pattern locks. At the core of our model lies a structured state-space architecture featuring a diagonal state transition matrix that effectively captures temporal dependencies in sensor data and acts as a compact memory mechanism. The proposed framework provides a local, efficient, and privacy-preserving security layer for mobile devices. Experimental results demonstrate that it achieves state-of-the-art performance, with equal error rates of 5.43% and 3.26% on the multimodal HMOG and FETA datasets, respectively, and 2.83% and 2.4% on the unimodal WHUGAIT and OU-ISIR datasets, respectively. Furthermore, it achieves over a 40-fold improvement in inference speed compared

to Transformer-based approaches. This work represents a significant step toward secure, adaptive, and resource-efficient authentication systems tailored to the dynamic nature of mobile user behavior.

4.2 Introduction

Smartphones have induced many changes in the way users manage personal information. We now store significant amounts of sensitive data, such as banking credentials, private pictures, and conversations, necessitating strong security measures. This highlights the need for continuous security research in smartphone services to protect user data. As the first layer of security, traditional user authentication methods, such as passwords, personal identification numbers (PINs), and patterns, have attracted interest due to their ease of use and implementation. However, they are unreliable and present significant vulnerabilities given their time-discrete nature and their operation as single-point verification systems, thus creating post-login risks. For instance, in Aviv, Gibson, Mossop, Blaze & Smith (2010), researchers highlighted the security vulnerabilities in touchscreens, particularly for Android's pattern lock system, and were able to recover the entire password pattern 68% of the time under optimal conditions. Also, authors of Yan, Liu, Zhou, Guo & Zhang (2020) discovered a new attack, called the surfing attack, which exploits ultrasonic waves propagating through solid surfaces to activate and control voice assistants like Siri and Google Assistant, thus bypassing voice recognition authentication. Nevertheless, modern smartphones are equipped with several sensors that can be leveraged to feed machine-learning (ML)-based approaches for continuous user behavior analysis and identity verification.

Continuous authentication is a paradigm shift from single-point verification to a persistent identity confirmation throughout the user's interaction with a device. It typically operates in two phases : enrollment and monitoring. During enrollment, a user generates a biometric template by interacting with the device, where features are extracted from touch events and sensor data, e.g., accelerometer and gyroscope. Monitoring involves comparing real-time interactions against this template using ML models. For instance, Biotouch employed in Estrela, Albuquerque,

Amaral, Giozza & Júnior (2021) random forest (RF) algorithms to achieve 96% accuracy in distinguishing legitimate users from impostors during banking transactions.

Indeed, the rise in popularity of behavioral biometrics as a research topic is largely attributable to the advancements in deep learning (DL) Sepas-Moghaddam & Etemad (2021). The end-to-end learning capability of DL models enables the automatic discovery of discriminative features, which stands in contrast to conventional machine learning methods that depend on handcrafted feature engineering. The inherent time-series nature of behavioral biometric data, such as keystroke dynamics or gait patterns, makes recurrent neural networks (RNNs) a highly suitable architecture due to their native ability to process sequential information Peinado-Contreras & Munoz-Organero (2020). In contrast, Convolutional Neural Networks (CNNs) are employed as hierarchical feature extractors. By applying convolutional filters, they automatically learn localized and translation-invariant patterns, enabling the identification of discriminative characteristics within the raw signal Nithyakani, Shanthini & Ponsam (2019). Nevertheless, there are a number of drawbacks to both recurrent and convolutional neural networks. CNNs are fundamentally limited by their focus on local spatial correlations, restricting their ability to capture long-range dependencies essential for complex pattern recognition. This shortcoming is particularly critical for real-world sensor data, which often contains missing values and overlapping activity patterns that demand broader contextual modeling Abdellatef, Al-Makhlasawy & Shalaby (2025). Furthermore, as noted in Saha, Saha, Kabir, Fattah & Saquib (2024), shallow CNN architectures struggle to achieve high classification accuracy in the presence of sensor noise and overlapping patterns. Although deeper CNNs can improve performance to some degree Yuan *et al.* (2024), their computational demands significantly hinder real-time deployment on resource-constrained devices such as smartphones. On the other hand, the inherent sequential nature of RNNs makes them non-parallelizable, which results in slow inference. This presents a major bottleneck for real-time deployment on edge systems. Additionally, it is well known that RNNs suffer from the vanishing gradient problem, making them difficult to train and unstable Pascanu, Mikolov & Bengio (2012), especially when dealing with very long sequences. Transformers have been studied more recently as a potential solution for the problems

in continuous authentication and activity recognition that recurrent and convolutional networks face Delgado-Santos *et al.* (2023a) Delgado-Santos, Tolosana, Guest, Vera-Rodriguez & Fierrez (2023b) thanks to their success in natural language. While cross-attention blocks excel with multimodal data (accelerations, orientation, scrolling, and keystrokes), their quadratic complexity makes them unsuitable for long-range dependencies that we can find in Inertial Measurement Unit (IMU) time series data, especially for resource-constrained systems like Raspberry Pi and smartphones. Multiple works have adapted transformers to time series ; however, they showed acceptable results, often at the expense of high computation costs, particularly in multimodal datasets. In addition, sensor data is always noisy data that is inappropriate for the attention mechanism. IMU data is frequently irregularly sampled, which can impair the attention mechanism an result in poor generalization. Despite significant progress, continuous authentication and activity recognition systems still face several challenges, including training instabilities such as vanishing or exploding gradients in long sequences, limited parallelization that results in slow inference, and substantial computational overhead stemming from the quadratic complexity $O(L^2)$ of attention mechanisms, where $\mathcal{L}$ refers to the sequence length. In addition, the inability to capture global context from local operations limits performance, particularly in the presence of noisy or overlapping sensor data. Shallow models often underperform, while deeper ones improve accuracy at the expense of efficiency, making real-time deployment on resource-constrained devices difficult.

To address these challenges, we propose in this paper MEDUSAA : a Multimodal Encoder-Decoder model Using State spaces for Authentication and Activity recognition. The main contributions of our paper can be summarized as follows :

- First, we propose a novel DL framework based on structured state-space models with diagonal transition matrices, enabling efficient modeling of long-range temporal dependencies in IMU sensor data for both continuous authentication and human activity recognition.
- Second, we design a multimodal Siamese encoder-decoder architecture that compresses raw sensor signals into compact embedding vectors and optimizes their separability using a

triplet loss function, enhancing discriminative feature learning across both unimodal and multimodal modalities.

- Third, we demonstrate the effectiveness of our model through extensive experiments on six benchmark datasets, where it achieves state-of-the-art performance in Equal Error Rate (EER), F1-score, and inference time, outperforming existing transformer and CNN based architectures, while remaining scalable and lightweight for mobile deployment.

4.3 Related Work

In the following, we first survey the existing literature on human authentication methods, followed by an overview of related studies on human activity recognition.

4.3.1 Human Authentication

Prior studies on continuous authentication can be classified into two distinct categories : Template matching and ML-based approaches. In template matching methods, users are authenticated by measuring the similarity between their newly recorded data and previously stored data. If the similarity measure exceeds a fixed threshold, the users are classified as genuine. Otherwise, they are classified as imposters. Such similarity scores include dynamic time warping Muller (2007), cross-correlation Wren *et al.* (2006), and the Pearson correlation coefficient Berman (2016). On the other hand, ML-based continuous authentication relys mainly on support vector machines (SVM) and K-nearest neighbors (kNN) Nickel *et al.* (2012) with a strong emphasis on feature engineering. However, these methods did not extract features directly from raw data, and needed extensive feature engineering like Fourier transforms and wavelet transforms Watanabe (2014).

Deep learning (DL) pushed continuous authentication further thanks to neural networks' (NNs) capacity to capture and represent non-linear information, without relying on feature engineering. The ability to learn complex non-linear representations of data provides precise and robust identification of walking patterns, even when the data may differ due to factors like walking speed, surface conditions, or device orientation. For instance, authors of Zeng *et al.* (2014)

pioneered the utilization of convolutional neural networks (CNNs) to extract features from human gait. They proved the superiority of CNNs over traditional methods in terms of classification rate, i.e., matching data to individuals. The study also presented a gait cycle extractor that uses template matching to identify and extract gait cycles. The process entails employing a low-pass filter with a cutoff frequency of 3 Hz to eradicate noise in accelerometer results, and finally, a similarity measure is utilized to detect matches between distinct cycles and select the one that maximizes the similarity coefficient. Despite the efficiency of this technique, it is memory-greedy, proportional to the size of time series-based data. This makes the inference speed particularly slow, especially on smartphones. In addition, its iterative nature prevents the leverage of parallelized computing. Alternatively, the proposed method in Ronneberger *et al.* (2015) extracted gait cycles with segmentation, similarly to the U-net architecture, and reduced the complexity using one-dimensional convolutions. This parallelizable approach accurately identified and authenticated individuals based on their gait patterns in natural conditions, with an accuracy above 93.5%. All these methods require large datasets that are often very expensive to annotate. This challenge comes from the fact that, unlike image datasets, human beings cannot easily recognize patterns in accelerometer signals. This has led recent literature to focus heavily on self-supervised learning methods as a solution to the need for manual annotations, especially given their outstanding results in other areas such as speech recognition Baevski, Zhou, Mohamed & Auli (2020). In this context, authors of HuMINet Stragapede *et al.* (2022) used the large and publicly available dataset of real-world interaction and sensor data, HuMIdb Acien *et al.* (2020). The authors trained separate recurrent neural networks (RNNs) with triplet loss for each biometric modality, including touch-based inputs like keystrokes and gestures, and motion sensor data. Results showed that while individual modalities like magnetometer signals and fixed-text keystrokes are highly discriminative, the fusion of modalities at the score level significantly enhanced human authentication performances, achieving Equal Error Rates (EER) between 4% and 9%, on a 3-second window.

In Senerath *et al.* (2023), the authors introduce the Behaveformer framework, which leverages a novel Spatio-Temporal Dual Attention Transformer (STDAT) architecture for continuous

authentication based on multimodal behavioral biometrics, including keystroke dynamics and inertial sensor data. The Behaveformer framework is designed to effectively capture both intra-modality and cross-temporal dependencies through a two-branch attention mechanism. The spatial attention component models inter-feature relationships at each timestep, such as between accelerometer axes or between different keystroke timing metrics. In contrast, temporal attention captures sequential behavioral dependencies over time, which is crucial for identifying users based on patterns like typing rhythm or device motion flow. Unlike conventional transformer architectures, STDAT explicitly separates and learns from spatial and temporal domains, enhancing its ability to extract subtle, identity-specific behavioral patterns. This architectural design proves particularly effective in passive authentication, where behavioral patterns are often noisy and overlapping across users. Empirically, BehaveFormer achieves impressive results : an Equal Error Rate (EER) of 2.95% on the HuMIdb dataset Acien *et al.* (2020) using combined keystroke and IMU data and an EER of 1.80% on the Aalto DB dataset Palin *et al.* (2019) using only keystroke dynamics, both outperforming several strong unimodal and multimodal baselines. These findings underscore the strength of STDAT in modeling the fine-grained and multimodal nature of behavioral biometrics and position BehaveFormer as a state-of-the-art solution for mobile continuous authentication.

M-GaitFormer Delgado-Santos *et al.* (2023a) introduces a dedicated two-stream Transformer-based architecture for mobile gait biometric verification using inertial sensor data. The model processes six-axis accelerometer and gyroscope signals through two distinct branches : a temporal stream to model gait dynamics over time and a channel stream to capture cross-sensor dependencies. M-GaitFormer integrates an auto-correlation attention mechanism based on the Fast Fourier Transform (FFT) to harness the periodic nature of gait signals, enabling the model to detect temporal repetitions inherent in human walking patterns. This is complemented by a multi-scale Biometric Gait Verification (BGV) CNN block, which extracts features at varying temporal resolutions, and a block-recurrent Transformer module that maintains long-range temporal dependencies crucial for characterizing individual gait cycles. Furthermore, the model employs Gaussian Range Encoding to preserve positional information across input sequences,

ensuring sensitivity to temporal ordering without overfitting to single time points. Unlike classification-based architectures, M-GaitFormer generates discriminative embeddings evaluated through similarity-based metrics, such as Euclidean distance, OC-SVM, and B-SVM, within a triplet loss training framework to optimize for verification tasks. Experimentally, M-GaitFormer achieves state-of-the-art performance on public gait datasets, reporting an Equal Error Rate (EER) of 3.42% on the whuGAIT database Zou *et al.* (2020) and 2.90% on the OU-ISIR database Iwama *et al.* (2012), markedly surpassing prior methods, including CNN, LSTM, and hybrid CNN-LSTM approaches, which reported EERs in the range of 4.49% to 10.43%.

In summary, while prior studies have advanced multimodal learning for behavioral biometrics through transformers and attention mechanisms, such methods remain computationally demanding and vulnerable to noisy, irregularly sampled IMU signals. To address these limitations, our proposed MEDUSAA framework integrates a structured state-space encoder with a lightweight Siamese encoder–decoder trained using triplet loss, as will be detailed in Section III. This architecture retains long-range dependencies with linear computational complexity and exhibits robustness to sensor noise and irregular sampling. Additionally, it inherently supports multimodal fusion, delivers high verification accuracy, and is optimized for practical, on-device continuous authentication on smartphones.

4.3.2 Human Activity Recognition

The first Human Activity Recognition (HAR) systems used Support Vector Machines (SVM), Random forests (RF) and k-nearest neighbors and relied on hand-crafted features and extensive domain knowledge Rojanavasu *et al.* (2023) Deep & Zheng (2019). These models classify activities based on human expertise in signal processing theory by typically windowing accelerometer data into sequences of 3 to 10 seconds in order to calculate features like mean, variance, energy, frequency peaks, etc. While these models are lightweight and interpretable and perform well on small datasets, they cannot capture complex data produced in the real world, where multiple activities could overlap in time and occur simultaneously Zhou *et al.* (2022). The

reliance on hand-crafted features often limits the generalizability of these models to new users and new environments, and they often lag behind deep learning (DL) models in performance.

To overcome these limitations and better model the temporal complexity and variability of human activities, recent research has shifted towards DL approaches. For instance, Yuan *et al.* leveraged large-scale datasets in Yuan *et al.* (2024), particularly the UK Biobank, to implement multi-task self-supervised learning (SSL) for human activity recognition. Their approach incorporated auxiliary tasks, including permutation, arrow-of-time, and time-warping, to develop pre-trained foundational CNN models that exhibited enhanced generalization capabilities across diverse datasets, thus diminishing reliance on labeled data. This methodological framework effectively addressed domain and task shift challenges, yielding significant improvements in F1-scores across multiple gait datasets. The magnitude of these improvements ranged from 2.5% to 100%, with the smallest gain observed on the Capture-24 dataset Chan *et al.* (2024). However, a CNN classifier looks only at local windows and loses fine-grained information needed to distinguish similar activities ("sit" vs. "talk-sit") Park *et al.* (2024).

Recent CNN-based architectures for HAR often incorporate techniques such as batch normalization, residual connections Mekruksavanich & Jitpattanakul (2023), and dilated convolutions Hamad *et al.* (2021) o enhance training stability and expand the receptive field. Meanwhile, research interest in recurrent neural networks (RNNs) has declined, as newer architectures address key limitations such as vanishing or exploding gradients and the inherently sequential, non-parallelizable nature of RNNs. For instance, authors in Li & Wang (2022) mitigates the vanishing gradient problem by integrating residual connections into RNNs, achieving a notable F1-score of 97.35%. However, this approach incurs higher computational costs and relies on access to both past and future context, limiting its applicability in real-time scenarios where future input data is unavailable.

Authors in Li *et al.* (2021) introduces THAT, a novel dual-stream architecture for human activity recognition using Channel State Information (CSI). It separates the input into temporal and channel streams, each processed through parallel Multi-scale Convolution Augmented

Transformer (MCAT) towers. These towers employ multi-head self-attention and adaptive multi-scale CNNs to capture long-range dependencies and local patterns across time and channel dimensions. THAT further enhances temporal semantics by incorporating Gaussian Range Encoding, which encodes order over time ranges instead of discrete points, an approach shown to distinguish better reversible actions such as "sit down" versus "stand up." Experimental results on various indoor environments, including office rooms, activity rooms, meeting rooms, and their combination, demonstrate the model's superior performance. THAT consistently outperforms classical and DL baselines (e.g., LSTM, CNN, HMM), achieving the highest average accuracy across all settings : 98.4% in the office room, 98.4% in the activity room, 99.0% in the meeting room, and 98.6% in the merged dataset, highlighting its effectiveness and generalizability in complex multi-environment HAR tasks.

In Mahmud *et al.* (2020), the authors proposed a self-attention-based architecture that incorporates a modality-specific attention mechanism to weigh the contribution of each sensor modality, positional encoding to retain temporal order, multiple intra-sequence self-attention layers to model local dependencies, and a global temporal attention module to capture long-term dependencies, achieving 96% accuracy.

In a more recent study Yu & Al-qaness (2025), the authors introduced an architecture featuring the DKInception module, which leverages parallel convolutions with kernel sizes of 1×3, 1×5, 1×5, 1×7, and 1×9 to capture temporal features at multiple scales. Smaller kernels emphasize fine-grained temporal details, while larger ones are designed to model longer-range dependencies. The resulting feature maps are then aggregated into a unified representation using a self-attention mechanism, enhancing contextual understanding. This approach achieved an accuracy of 89% on the PAMAP2 dataset Reiss (2012).

Overall, existing HAR approaches either operate on short temporal windows using CNN variants or depend on computationally intensive attention mechanisms, which hinder efficient on-device deployment. Our proposed structured state-space architecture offers the best of both paradigms, capturing long-term temporal dependencies comparable to attention and RNNs, while

maintaining linear time complexity and a compact memory state that mitigates the vanishing gradient problem. As will be demonstrated in our experiments, this design achieves strong generalization across both controlled environments (using the PAMAP2 dataset) and real-world settings (using the Capture-24 dataset).

Table 4.1 summarizes the capabilities of prior methods and underscores the practical advantages of our proposed MEDUSAA model, which uniquely supports multimodal inputs, enables online continuous operation, captures long-range dependencies, and remains lightweight enough for deployment on edge devices.

Tableau 4.1 Comparison of related works across key capabilities. Only our model supports all features including multimodal input, scalability, and online deployment.

Model	**Multi-modal**	**HAR**	**Conti-nuous Auth**	**Long -Term Dep.**	**Online Operation Support**	**Light-weight**	**Self-Super-vised**	**Scalable**
Biotouch (RF)			✓			✓		
CNN			✓					
U-Net 1D			✓	✓	✓	✓		
HuMINet	✓		✓				✓	
BehaveFormer	✓		✓	✓				
M-GaitFormer			✓	✓				
THAT		✓		✓				
DKInception		✓		✓				
Res-BiLSTM		✓		✓				
Attention CNN		✓		✓				
Informer		✓	✓	✓		✓		
DuoNet	✓		✓					
Dilated conv			✓					
ResNet		✓					✓	✓
MEDUSAA	✓	✓	✓	✓	✓	✓	✓	✓

4.4 Proposed DL Model for Human Authentication and Activity Recognition

4.4.1 System Model : Structured state spaces

Behavioral biometrics and human activity recognition using wearable devices both rely on time series pattern recognition, where the core challenge lies in effectively leveraging the historical context of sensor data. Formally, if f denotes a function representing sensor measurements, then $f(x)|_{x \leq t}$ (for all $t \geq 0$) corresponds to the history of sensor readings up to time t. A typical machine learning classification model learns patterns from this history and their corresponding annotations y_i. In continuous authentication, y represents the user's identity, while in human activity recognition, it denotes the current activity. Despite differences in labels, both tasks share the same underlying objective : mapping temporal sensor patterns to their associated classes.

However, as the volume of sensor readings increases, especially with high-frequency sensors such as accelerometers and gyroscopes, storing the complete history becomes impractical on resource-constrained edge devices like wearables, smartphones, and IoT systems. Furthermore, extracting meaningful patterns from long sequences poses additional computational challenges. The HiPPO (High-order Polynomial Projection Operators) framework Gu, Johnson, Timalsina, Rudra & Ré (2022c) offers a solution by leveraging orthogonal polynomials to incrementally compress the entire history of sensor data into a fixed-size embedding vector of size N . This approach enables efficient representation, storage, and processing of temporal information without sacrificing essential contextual information.

Assuming that the function $f(x)|_{x \leq t}$ (for all $t \geq 0$) represents the history of sensor readings up to time t, we aim to approximate f over fixed intervals $[a, b]$ using a basis of orthogonal polynomials. Let $\{\phi_n(x)\}_{n=0}^{\infty}$ be a set of orthogonal polynomials defined on the interval $[a, b]$, such that the inner product satisfies :

$$\langle \phi_m, \phi_n \rangle = \int_a^b \phi_m(x)\phi_n(x)\, dx = \begin{cases} 0, & \text{if } m \neq n \\ h_n, & \text{if } m = n, \end{cases} \tag{4.1}$$

where $h_n = \int_a^b \phi_n(x)^2\,dx$ is the normalization constant associated with ϕ_n, ensuring the orthogonality condition.

To approximate the function $f(x)$ on the interval $[a, b]$, the objective is to determine the coefficients c_n (for all $n = 0, \ldots, N$) such that :

$$f(x) \approx f_N(x) = \sum_{n=0}^{N} c_n \phi_n(x).$$

To obtain the optimal coefficients, we minimize the mean square error (MSE) between $f(x)$ and its approximation $f_N(x)$, defined as :

$$\text{MSE} = \int_a^b [f(x) - f_N(x)]^2\ dx.$$

Thanks to the orthogonality of the basis functions, the coefficients c_n can be obtained by projecting $f(x)$ onto the corresponding polynomial $\phi_n(x)$ as follows :

$$c_n = \frac{1}{h_n} \int_a^b f(x)\phi_n(x)\,dx, \quad \forall n = 0, \ldots, N. \tag{4.2}$$

The accuracy of this approximation depends on the number of terms N used in the series expansion and the choice of orthogonal polynomials. Increasing N typically improves the fidelity of the approximation but also increases computational complexity.

4.4.2 MEDUSAA : Multimodal Encoder-Decoder Using State spaces for Authentication and Activity recognition

In this section, we present our proposed MEDUSAA model, which is based on a DL architecture built upon a state-space framework featuring a diagonal state transition matrix. This matrix acts as an implicit memory, regulating the retention of historical information from the input time series.

In our study, the evolution of the hidden state $\mathbf{h}(t)$ is governed by the following state-space representation :

$$\begin{aligned} \frac{\partial \mathbf{h}(t)}{\partial t} &= \mathbf{A}h(t) + \mathbf{B}u(t), \\ \frac{\partial \mathbf{y}(t)}{\partial t} &= \mathbf{C}h(t) + \mathbf{D}u(t), \end{aligned} \tag{4.3}$$

where $\mathbf{A} \in \mathbb{R}^{N\times N}$ is the diagonal state transition matrix, $\mathbf{B} \in \mathbb{R}^{N\times d}$ projects the input $\mathbf{u}(t)$ into the hidden state space, $\mathbf{C} \in \mathbb{R}^{m\times N}$ maps the hidden states to the output, and $\mathbf{D}$ is an optional direct feedthrough term.

The input function $\mathbf{u}(t)$ represents the raw sensor readings from the Inertial Measurement Unit (IMU), typically comprising accelerometer and gyroscope measurements. Specifically,

$$\mathbf{u}(t) = \begin{bmatrix} a_x(t) \\ a_y(t) \\ a_z(t) \\ \omega_x(t) \\ \omega_y(t) \\ \omega_z(t) \end{bmatrix} \in \mathbb{R}^6 \tag{4.4}$$

where $a_x(t), a_y(t), a_z(t)$ correspond to the accelerometer readings in three spatial dimensions, and $\omega_x(t), \omega_y(t), \omega_z(t)$ represent the angular velocity measured by the gyroscope. These signals capture user movement patterns, enabling the model to extract behavioral biometrics for continuous authentication.

The architecture, as illustrated in Figs. 4.1 and 4.2 follows a Siamese structure where input signals (anchor, positive, and negative samples), are first passed through a shared encoder. This encoder transforms raw input sequences into a latent representation suitable for temporal modeling followed by multiple stacked diagonal SSM perfect for capturing long-range dependencies in

the time domain with linear computational cost $O(L)$. The decoder then reconstructs this representation into a lower-dimensional embedding vector space.

These embeddings denoted as A, P, and N for the anchor, positive, and negative inputs, respectively, are optimized using the *triplet loss* function, as shown in Fig. 4.2. The objective is to bring the anchor closer to the positive sample (same user) and push it away from the negative sample (impostor) in the learned embedding space. The triplet loss is defined as follows :

$$\mathcal{L}_{\text{triplet}}(A, P, N) = \max\left(d(A, P) - d(A, N) + \alpha, 0\right), \tag{4.5}$$

where $d(\cdot, \cdot)$ denotes a distance metric such as cosine similarity, and α is a margin that enforces a minimum separation between genuine and impostor pairs. This loss formulation encourages the model to learn discriminative embeddings that preserve inter-user differences while minimizing intra-user variability.

The structured state-space model effectively captures temporal dependencies within sequential IMU data, enabling users to differentiate based on their unique motion patterns. The diagonal structure of the state transition matrix $\mathbf{A}$ ensures computational efficiency, allowing the model to maintain long-range dependencies while remaining scalable for real-time applications, facilitating seamless and continuous authentication.

What makes the model in equations (4.3) particularly compelling is that it corresponds to an ordinary differential equation for which the analytical solution is known. The solution takes the following form :

$$h(t) = e^{\mathbf{A}t}h(0) + \int_0^t e^{\mathbf{A}(t-\tau)}\mathbf{B}x(\tau)\, d\tau \tag{4.6}$$

Assuming the initial hidden state $h(0) = 0$, the closed form of the hidden state will be :

$$h(t) = B\int_0^t e^{A(t-\tau)}x(\tau)\, d\tau. \tag{4.7}$$

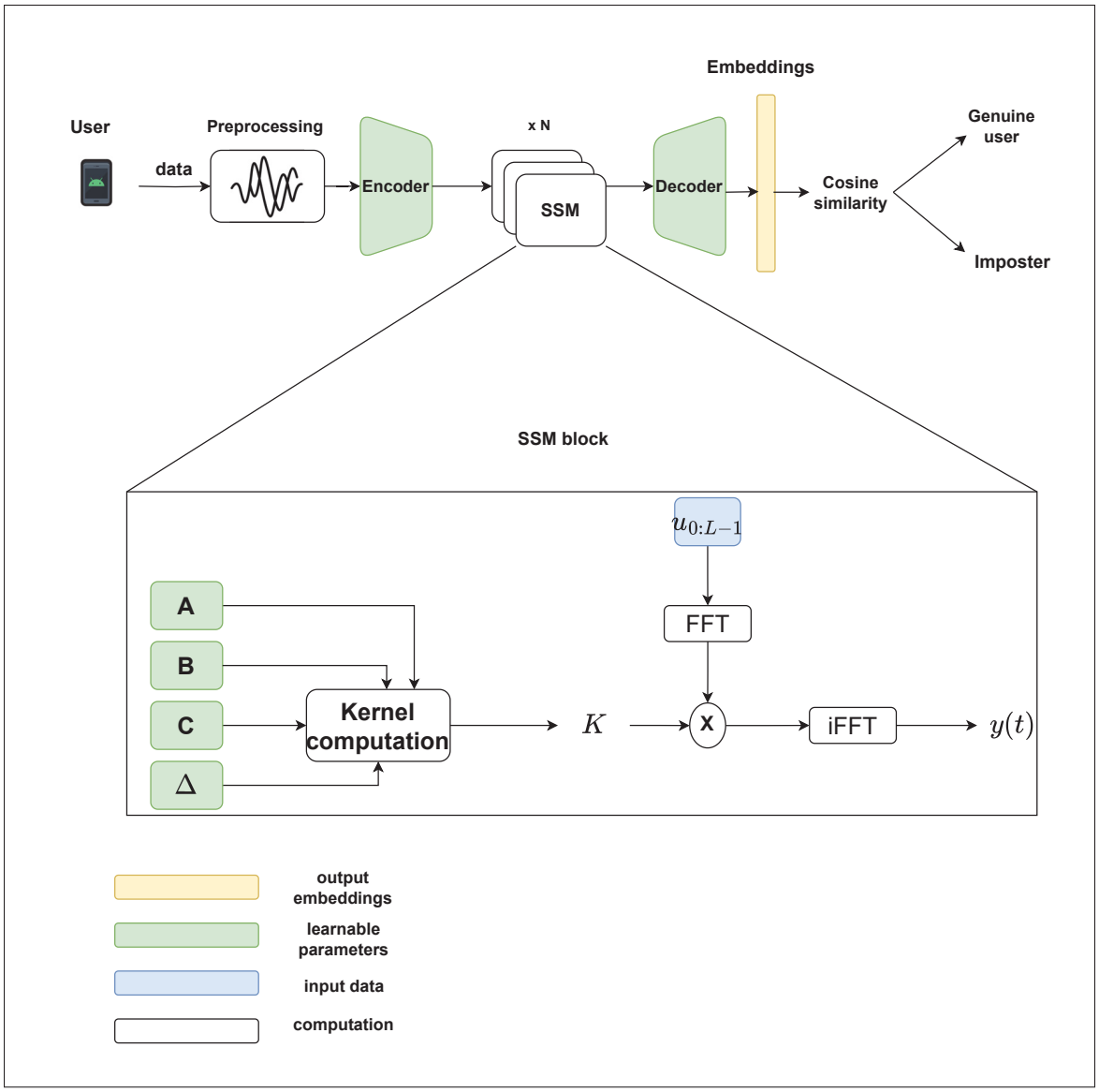


Figure 4.1 Our proposed MEDUSAA model

By substituting into the output expression $y(t)$, we observe that when the kernel $K(t) = \mathbf{C}e^{\mathbf{A}(t-\tau)}\mathbf{B}$ is used, which is fully defined by the matrices **A**, **B** and **C**, the state-space model becomes equivalent to a convolution operation.

$$y(t) = \int_0^t x(\tau)K(t-\tau)\,d\tau + Dx(t). \tag{4.8}$$

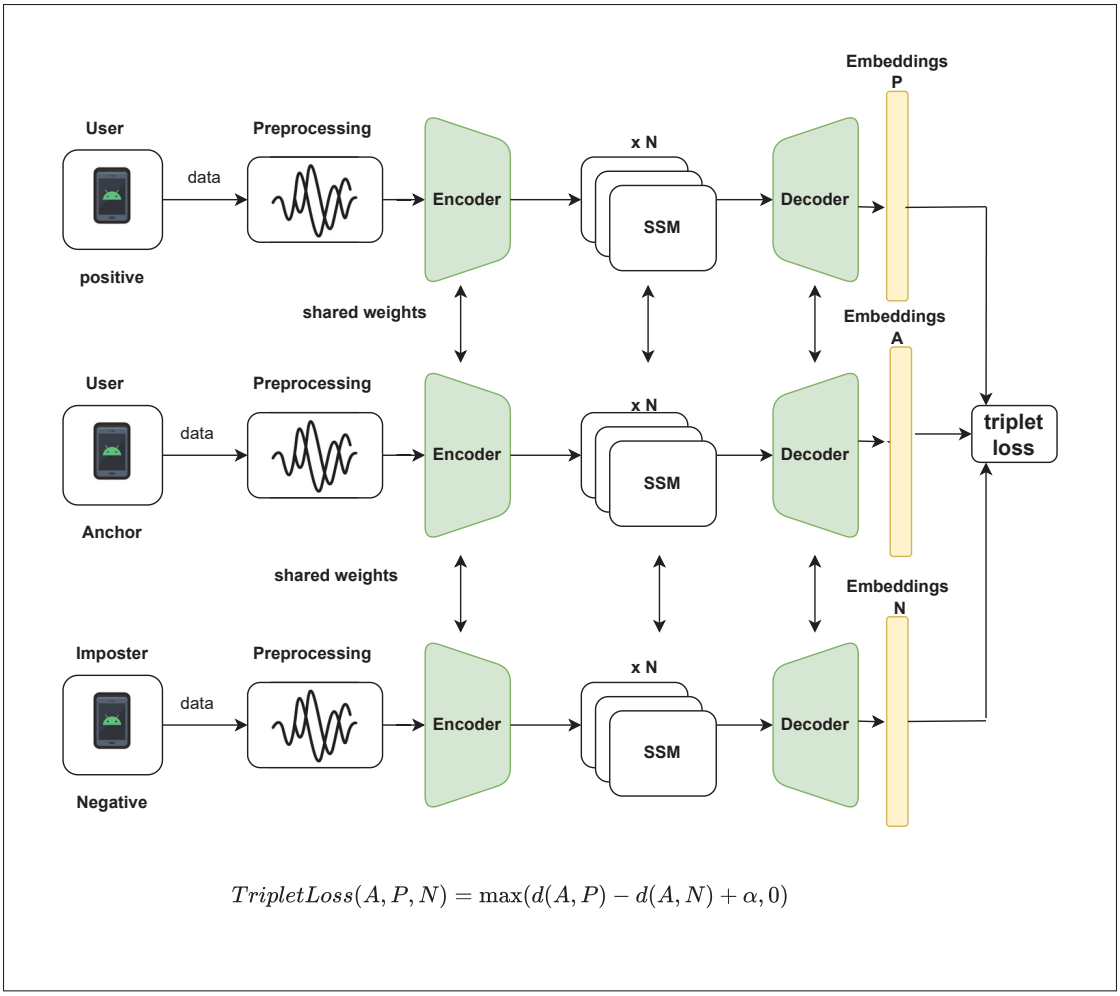


Figure 4.2 Overview of the proposed multimodal architecture for continuous authentication and human activity recognition. The model leverages shared-weight encoders, state space model (SSM) blocks, and decoders to produce embeddings for anchor, positive, and negative samples, optimized using the triplet loss function.

In fact, the mathematical convolution operation is defined as follows, given two functions g and f :

$$(f * g)(t) = \int_{-\infty}^{+\infty} f(\tau)g(t-\tau)d\tau \tag{4.9}$$

The state-space model in (4.3) can be thus written as follows :

$$y(t) = (K * x)(t) + Dx(t) \tag{4.10}$$

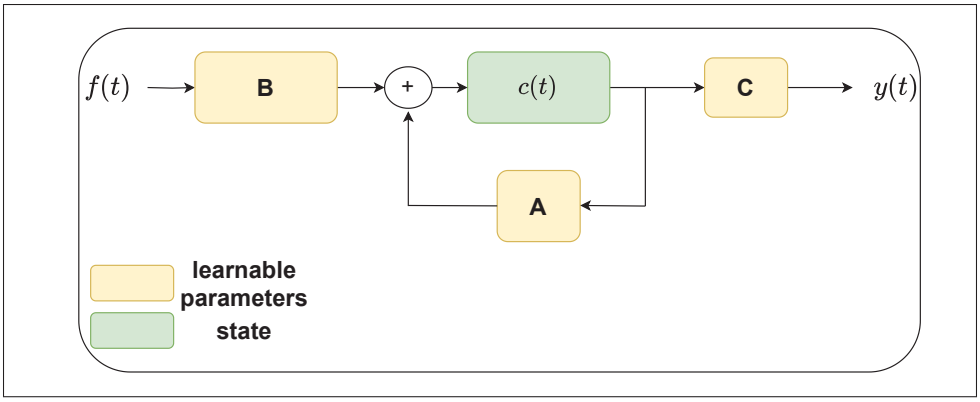


Figure 4.3 Recurrent view of our proposed model

On the other hand, continuous state-space models can be discretized in the Laplace domain. Specifically, the continuous ordinary differential equations presented in (4.3) can be transformed using the Laplace transform as follows :

$$\begin{aligned} sH(s) - h(0) &= AH(s) + BX(s) \\ H(s) &= (sI - A)^{-1}BX(s) \end{aligned} \tag{4.11}$$

and using the bilinear transform, we can find that :

$$A_d = \left(I - \frac{\Delta}{2}A\right)^{-1}\left(I + \frac{\Delta}{2}A\right), \tag{4.12}$$

$$B_d = \left(I - \frac{\Delta}{2}A\right)^{-1}\Delta B. \tag{4.13}$$

Consequently, the discrete form of the state equations highlights a key strength of state-space models (SSMs) : their dual representation as both a global convolution and a recurrence, as illustrated in Fig. 4.3. This duality offers the best of both worlds, leveraging the parallelism of convolutions for efficient training while preserving the advantages of recurrence, such as memory efficiency and suitability for online processing. Indeed, the recurrent formulation maintains a fixed memory footprint, independent of sequence length, since only the current state needs to be stored. Moreover, it supports real-time, incremental updates as new inputs arrive, eliminating the

need to recompute the entire sequence. This capability is particularly valuable for applications that require continuous and adaptive processing of streaming data. Formally, the state-space models can be written as follows :

$$\begin{aligned} h_k &= \mathbf{A_d} h_{k-1} + \mathbf{B_d} x_k \\ y_k &= \mathbf{C} h_k + \mathbf{D} x_k \end{aligned} \tag{4.14}$$

4.5 Experimental results

To evaluate the effectiveness of the proposed MEDUSAA framework for continuous user authentication, we conducted experiments using several datasets : the HMOG (Hand Movement, Orientation, and Grasp) dataset Sitová *et al.* (2015), OU-ISIR Iwama *et al.* (2012), WHU-GAIT Zou *et al.* (2020), FETA Georgiev, Eberz, Turner, Lovisotto & Martinovic (2022), PAMAP2 Reiss (2012) and CAPTURE-24 Yuan *et al.* (2024) . These publicly available dataset capture sensor data from mobile devices, including accelerometer, gyroscope, magnetometer readings, and touch interaction features. With data collected from multiple users across various sessions, these datasets provide a robust benchmark for assessing continuous authentication models.

We begin by describing the datasets, followed by an outline of the experimental setup. We then present the performance results, first for human authentication and subsequently for human activity recognition.

4.5.1 Datasets description

Building upon the datasets described earlier, we evaluate the proposed MEDUSAA model on two behavioral biometric tasks : scrolling dynamics and gait recognition. We conduct both unimodal and multimodal experiments. For the unimodal setting, the model is trained exclusively on IMU sensor data. The WHU-GAIT Zou *et al.* (2020), OU-ISIR Iwama *et al.* (2012) and PAMAP2 Reiss (2012) datasets serve as critical validation platforms, representing pure motion-based authentication scenarios without the additional context of touch interactions. These datasets

present unique challenges : WHU-GAIT contains accelerometer and gyroscope readings from waist-mounted smartphones during natural walking, while OU-ISIR provides inertial signals from a larger cohort of subjects under controlled conditions. On the other hand In the multimodal configuration, we combine IMU signals with scrolling interaction features to assess the benefits of sensor fusion for continuous user authentication.

For the IMU features, we extract raw data from both the accelerometer and the gyroscope. Each sensor generates a three-dimensional vector for every timestamp at a sampling rate of 100 Hz. Consequently, the IMU feature vector for both the gyroscope and the accelerometer is six-dimensional, represented as follows :

$$\mathbf{f}_{\text{IMU}} = \begin{bmatrix} a_x & a_y & a_z & g_x & g_y & g_z \end{bmatrix}$$

where a_x, a_y, a_z are the accelerometer readings along the three axes, and g_x, g_y, g_z are the gyroscope readings, we normalize this feature vector within the range $[0, 10]$. In the multimodal model, it is logical to synchronize IMU readings with timestamps from other modalities, such as scrolling, and discard readings that fall outside the range of scrolling events.

For the scrolling data, we extract 11 features that describe the touch interaction. These include the initial touch coordinates (*Start_X*, *Start_Y*), the pressure value at initial touch (*Start_pressure*), and the contact size at initial touch (*Start_size*). Additionally, we capture the type of action being performed (*Current_action_type*, along with the movement coordinates (*Current_X*, *Current_Y*), the pressure value during movement (*Current_pressure*), and the contact size during movement (*Current_size*). Finally, we compute the distance moved in both the X and Y directions (*Distance_X*, *Distance_Y*). These features provide a detailed representation of the user's scrolling behavior, which is crucial for multimodal analysis. Table 4.2 presents the parameter specifications for the described datasets.

Tableau 4.2 Overview of Datasets Used in Experiments

Dataset	Activities	Sensors Used
OU-ISIR Iwama *et al.* (2012)	6	Accelerometer, Gyroscope
WHU-GAIT Zou *et al.* (2020)	6	Accelerometer, Gyroscope
HMOG Sitová *et al.* (2015)	5	IMU, Touchscreen
FETA Georgiev *et al.* (2022)	Daily Activities	Accelerometer
PAMAP2 Reiss (2012)	12	IMU, Heart Rate Monitor
Capture-24 Yuan *et al.* (2024)	Daily Activities	Accelerometer, Gyroscope

4.5.2 Experimental settings

All training and inference procedures were performed on a workstation with an NVIDIA GeForce RTX 3060 GPU (6 GB VRAM). This setup balanced computational power and resource constraints, allowing us to efficiently process the sensor data and expedite model convergence. By leveraging the RTX 3060's parallel processing capabilities, we could train and optimize our deep neural network models within a reasonable time frame. Our proposed model is implemented in Pytorch with Adam optimizer, a learning rate of 0.001, and a batch size of 64.

For the loss function, we employ the Triplet Loss illustrated in Fig. 4.3, which is well-suited for learning an embedding space where similar data points are positioned closer together, while dissimilar ones are pushed farther apart. Given an anchor sample A, a **positive** sample P (same class as the anchor), and a **negative** sample N (different class), the loss function is defined as :

$$TripletLoss(A, P, N) = \max(d(A, P) - d(A, N) + \alpha, 0)$$

- $d(A, P)$ is the distance between the anchor and the positive sample.
- $d(A, N)$ is the distance between the anchor and the negative sample.
- α is a margin that ensures the negative sample is sufficiently far from the anchor.

For the distance between samples, we use the cosine similarity.

4.5.3 Continuous Authentication Performance Results

This section evaluates our proposed model's discriminative capabilities across two benchmark multimodal datasets, HMOG Sitová *et al.* (2015) and FETA Georgiev *et al.* (2022), as well as three unimodal benchmark datasets, WHU-GAIT Zou *et al.* (2020), OU-ISIR Iwama *et al.* (2012), and PAMAP2 Reiss (2012), using metrics critical to security applications : precision, recall, F1-score, and Equal Error Rate (EER). These metrics quantify the model's capacity to distinguish genuine users from impostors while minimizing false acceptances and rejections. Obtained results are supported by the t-distributed Stochastic Neighbor Embedding (t-SNE) visualizations (see Figs. 4 and 5) and comparative analyses in Tables 4.3 and 4.4.

In the following, we present a detailed analysis of the results, focusing on models' performance and scalability as well as computational complexity and inference time.

4.5.3.1 Models' performance and Scalability

Table 4.3 presents a comprehensive performance comparison of all evaluated authentication models across the aforementioned datasets. We can see that our proposed approach outperforms Transformer-based and recurrent models across key metrics, including F1 score, Equal Error Rate (EER), and Accuracy, particularly when enhanced with triplet loss function for optimized embedding learning. By combining behavioral robustness with computational efficiency, our framework significantly advances the state of the art in seamless and secure user authentication.

Specifically, as shown in Table 4.3, our proposed multimodal model (while using both IMU and scrolling patterns) performs better on both the HMOG and FETA datasets. On HMOG, it attains an F1-score of 99%, outperforming the IMU-only model (F1-score : 96.4%), demonstrating the value of integrating touch dynamics with motion data. On FETA, MEDUSAA achieves near-perfect precision 100% and recall 99.55%, with an F1-score of 0.9978, highlighting its robustness to sensor noise and behavioral variability. These results emphasize the complementary roles of inertial signals that capture gait and device handling and scrolling dynamics that encode

Tableau 4.3 Comprehensive Performance Comparison of Authentication Models

Dataset	Type	Model	Accuracy	Precision	Recall	F1-Score	EER
HMOG	Multimodal	MEDUSAA[3]	98%	99%	93.93%	96.4%	7.06
		MEDUSAA[4]	**99%**	**99.23%**	**98.88%**	**99.06%**	**5.43**
FETA	Multimodal	MEDUSAA[4]	**99.28%**	**100%**	**99.55%**	**99.78%**	**3.26**
		HumiNet	95%	94%	94%	94%	27.02
		DuoNet	92%	89.1%	90.9%	90%	28.31
		BehaveFormer	99%	97%	99%	98%	17.44
PAMAP2	Unimodal	MEDUSAA[3]	**94.2%**	**93.8%**	**94.1%**	**93.9%**	7.2
		MGaitFormer	91.1%	89%	91%	90%	12.09
		THAT	85.25%	83.7%	86.34%	85%	14.1
		Informer	57%	56%	55%	56%	28.22
WHU-GAIT	Unimodal	MEDUSAA[3]	**97%**	**98%**	**97%**	**97%**	**2.83**
		MGaitFormer	94.25%	92%	94%	93%	3.42
		THAT	92.9%	91%	93%	92%	7.21
		Informer	54.5%	53%	55%	54%	27.33
OU-ISIR	Unimodal	MEDUSAA[3]	**95.6%**	**94%**	**95%**	**94%**	**2.4**
		MGaitFormer	93.33%	89.5%	93%	91.2%	2.9
		THAT	85.7%	84.4%	86%	85.17%	15.1
		Informer	59.4%	56.1%	59%	55%	26.1

touch pressure and spatial patterns, which reduce intra-user variability while amplifying inter-user distinctions.

In addition, our proposed model achieves a 3.26% equal error rate on the FETA dataset, significantly outperforming baseline methods such as BehaveFormer Senerath *et al.* (2023) (17.44% EER) and HuMINet Stragapede *et al.* (2022) (27.02% EER). This 14.18% absolute reduction in EER underscores the model's ability to balance false acceptances and false rejections, a critical metric for security-sensitive applications. The structured state-space architecture's capacity to model long-range temporal dependencies in IMU sequences, combined with triplet loss optimization, enables precise separation of user embeddings in the latent space. The low

[3] Unimodal : IMU

[4] Multimodal : IMU and scrolling pattern

EER positions the model as a benchmark for continuous authentication systems, particularly in scenarios requiring high inference speed and usability.

To further validate the effectiveness of the proposed MEDUSAA model, Fig. 4.4 illustrates a t-SNE visualization of user embeddings from the multimodal configuration combining IMU signals with scrolling patterns on the HMOG dataset. The projection reveals ten clearly separated clusters, each corresponding to a unique user, with strong intra-class cohesion and substantial inter-class separation. This structure aligns with the quantitative results in Table 4.3 (precision = 99.23%, recall = 98.88%, F1-score = 99.06%). The compact clusters indicate high session-to-session consistency in extracted behavioral features, while the distinct boundaries reflect robustness to impersonation attempts. Notably, the IMU + Scrolling model exhibits more pronounced cluster separation than the IMU-only variant, confirming the added discriminative power provided by fusing complementary inertial and touch modalities, thereby explaining the observed performance gains.

While the multimodal results demonstrate the efficiency of our approach in fusing complementary data streams, we further evaluate the model's performance on established unimodal benchmark datasets to assess its generalizability. This evaluation aims to determine whether the structured state-space architecture that excelled in multimodal contexts can maintain its discriminative power when limited to a single modality.

Indeed, as reported in Table 4.3, MEDUSAA consistently achieves high performance across all three unimodal benchmarks, with particularly notable results on the WHU-GAIT dataset (97.0% accuracy, 98.0% precision, 97.0% recall, and 97.0% F1-score). These results underscore the model's ability to capture subtle gait-specific patterns from smartphone accelerometer and gyroscope signals. Similarly, the strong performance on PAMAP2 (94.2% accuracy, 93.8% precision, 94.1% recall, and 93.9% F1-score) demonstrates the model's capability to generalize across diverse human activities beyond walking.

Fig. 4.5 provides further qualitative evidence of the model's discriminative power through a t-SNE visualization of user embeddings from the PAMAP2 dataset, where eight distinct

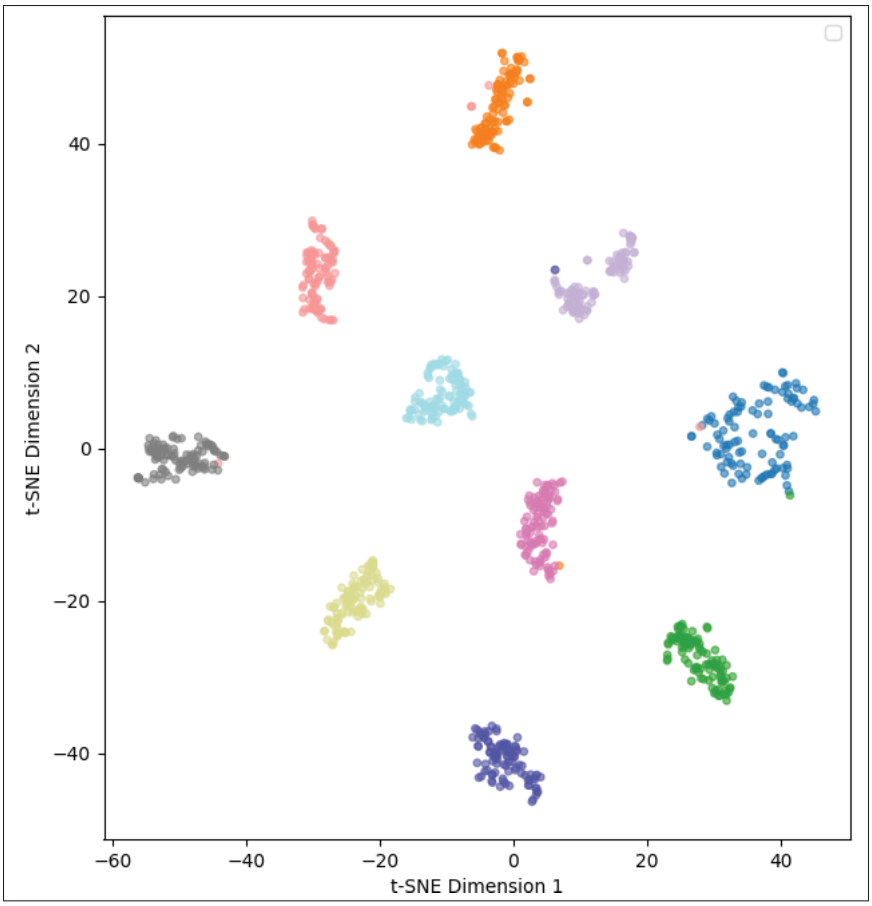


Figure 4.4 t-SNE visualization of user embeddings from IMU+Scrolling model (HMOG dataset, 10 users). Clear separation between genuine users (colored clusters).

clusters emerge with minimal overlap, indicating effective user separation even in a complex activity recognition scenario. The confusion matrix of our model, depicted in Fig. 4.6, further substantiates these findings, showing strong diagonal dominance with minimal misclassifications between users.

In addition, the strong performance on OU-ISIR (93.0% accuracy, 94.0% precision, 93.0% recall, and 94.0% F1-score), a challenging and large-scale gait dataset, highlights the scalability of our approach.

All these results show that the structured state-space architecture effectively models temporal dependencies in motion data alone, preserving its discriminative power even without the

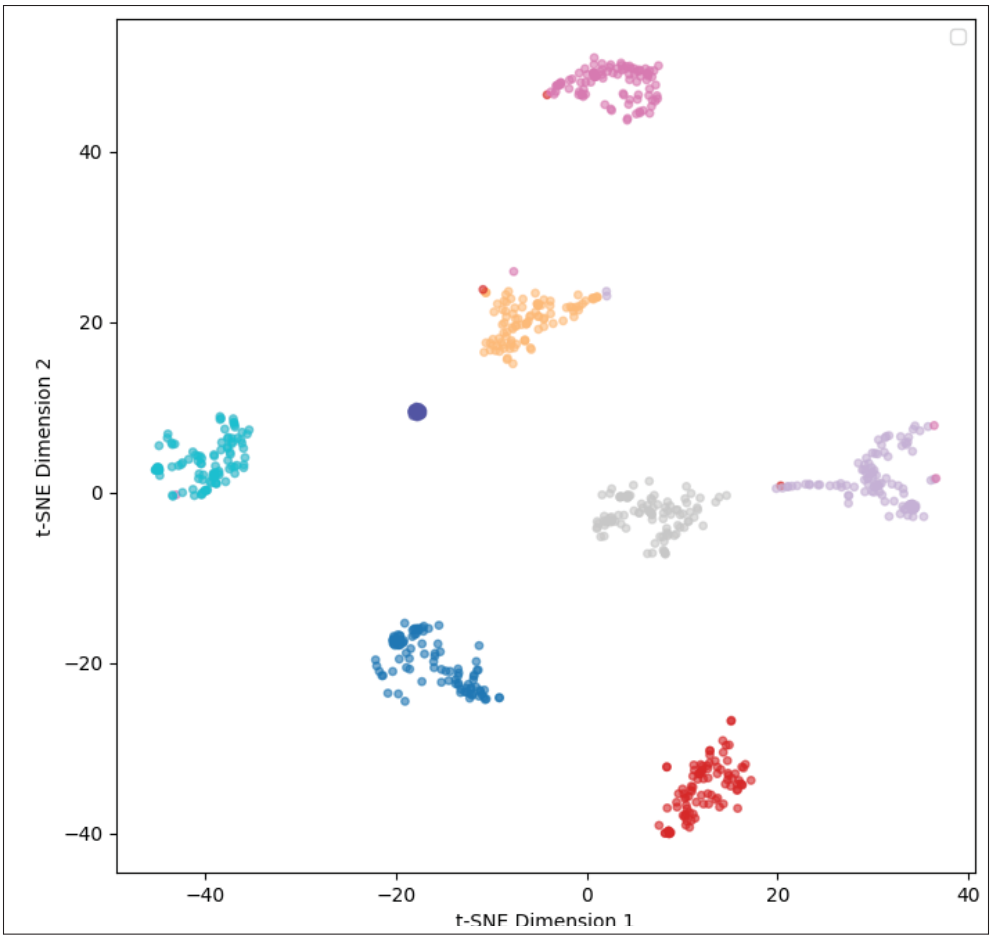


Figure 4.5 t-SNE visualization users embeddings (PAMAP2 dataset 8 users). Distinct clusters emerge.

contextual patterns provided by touch interactions in multimodal settings. The consistent performance across diverse unimodal datasets validates the robustness of our temporal modeling strategy and underscores its applicability to a wide range of authentication scenarios, from handheld devices to wearable sensors.

4.5.3.2 Complexity Analysis and Inference Time

Table 4.4 reports the complexity[5] (in terms of GFLOPS) and the inference time of all methods. From that table, we can see that MEDUSAA operates at an efficient 31 GFLOPS, striking an

[5] Complexity computation is conducted on a sampled tensor of shape $(1449, 128, 6)$.

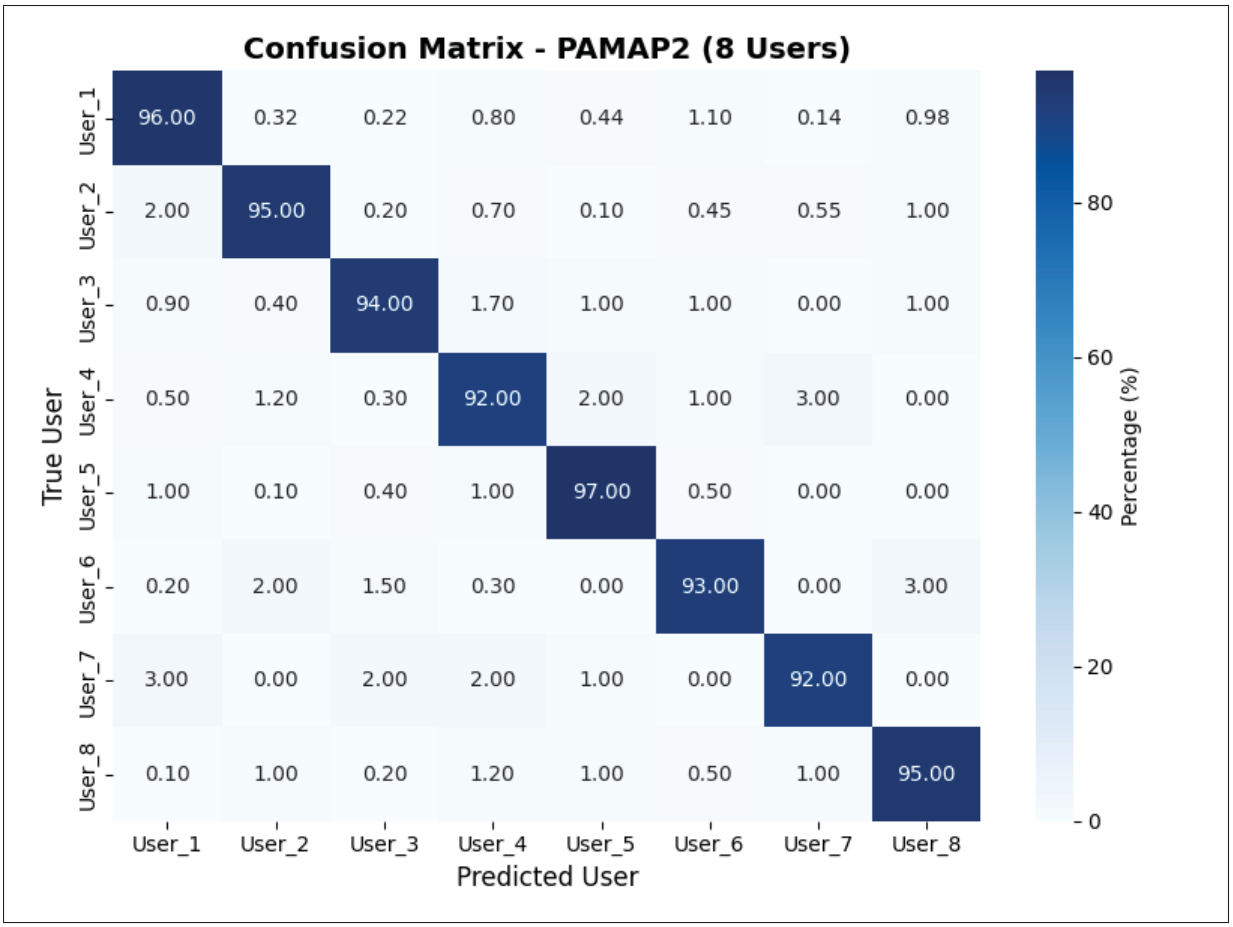


Figure 4.6 Confusion Matrix for IMU-Based PAMAP2 Multi-User Classification (8 users)

Tableau 4.4 Complexity (in GFLOPS), step time during training, and peak GPU memory usage for different authentication models.

Method	**GFLOPS**	**Step Time (ms)**	**GPU Mem. (MB)**
MEDUSAA	38.0	54	200
BehaveFormer Senerath *et al.* (2023)	58.0	3000	420
MGaitFormer Delgado-Santos *et al.* (2023a)	52.0	2700	400
THAT Li *et al.* (2021)	24.0	2450	380
Informer Zhou *et al.* (2020)	16.8	2500	410

effective balance between high performance and manageable computational demand. This is considerably more efficient than larger contemporary models such as BehaveFormer Senerath *et al.* (2023) (58 GFLOPS) and MGaitFormer Delgado-Santos *et al.* (2023a) (52 GFLOPS),

while outperforming leaner architectures like THAT Li *et al.* (2021) (24 GFLOPS) and Informer Zhou *et al.* (2020) (16.8 GFLOPS) on key accuracy metrics.

In addition, our model achieves a training step time of only 50 ms, over 40× faster than THAT Li *et al.* (2021), Informer Zhou *et al.* (2020), and BehaveFormer Senerath *et al.* (2023), while maintaining the lowest peak GPU memory footprint (200 MB). This combination of speed, efficiency, and minimal resource usage underscores its practicality for real-world security applications. The gains arise from the structured state-space architecture's ability to model long-range temporal dependencies without the quadratic complexity of transformer self-attention, delivering state-of-the-art authentication accuracy with dramatically reduced computational overhead.

4.5.4 Human Activity Recognition Performance Results

Tableau 4.5 Performance of each HAR model on Capture-24 and PAMAP2 datasets across multiple metrics

Model	Dataset	Accuracy (%)	F1-score (%)	Precision (%)	Recall (%)
attention CNN	Capture-24	61.3	61.3	61.3	61.3
	PAMAP2	96.45	96.4	96.4	96.4
Residual-BiLSTM	Capture-24	63.5	63.5	63.5	63.5
	PAMAP2	97.93	97.9	97.9	97.9
THAT	Capture-24	65.4	63.8	66.0	61.8
	PAMAP2	90.21	89.65	91.00	88.30
S4	Capture-24	66.1	65.0	67.3	62.9
	PAMAP2	91.75	91.22	92.60	90.00
DKInception	Capture-24	64.9	64.0	66.4	61.8
	PAMAP2	89.3	91.22	92.60	90.00
12-layers ResNet	Capture-24	70.9	70.0	71.8	68.3
	PAMAP2	91.75	91.22	92.60	90.00
MEDUSAA	Capture-24	**71.2**	**70.0**	**68.8**	**71.24**
	PAMAP2	**98.27**	**97.0**	**96.3**	**97.6**

Having demonstrated the effectiveness of our structured state space architecture for user authentication, we extend its evaluation to the domain of Human Activity Recognition (HAR) to assess its generalizability. While authentication focuses on identifying user-specific behavioral

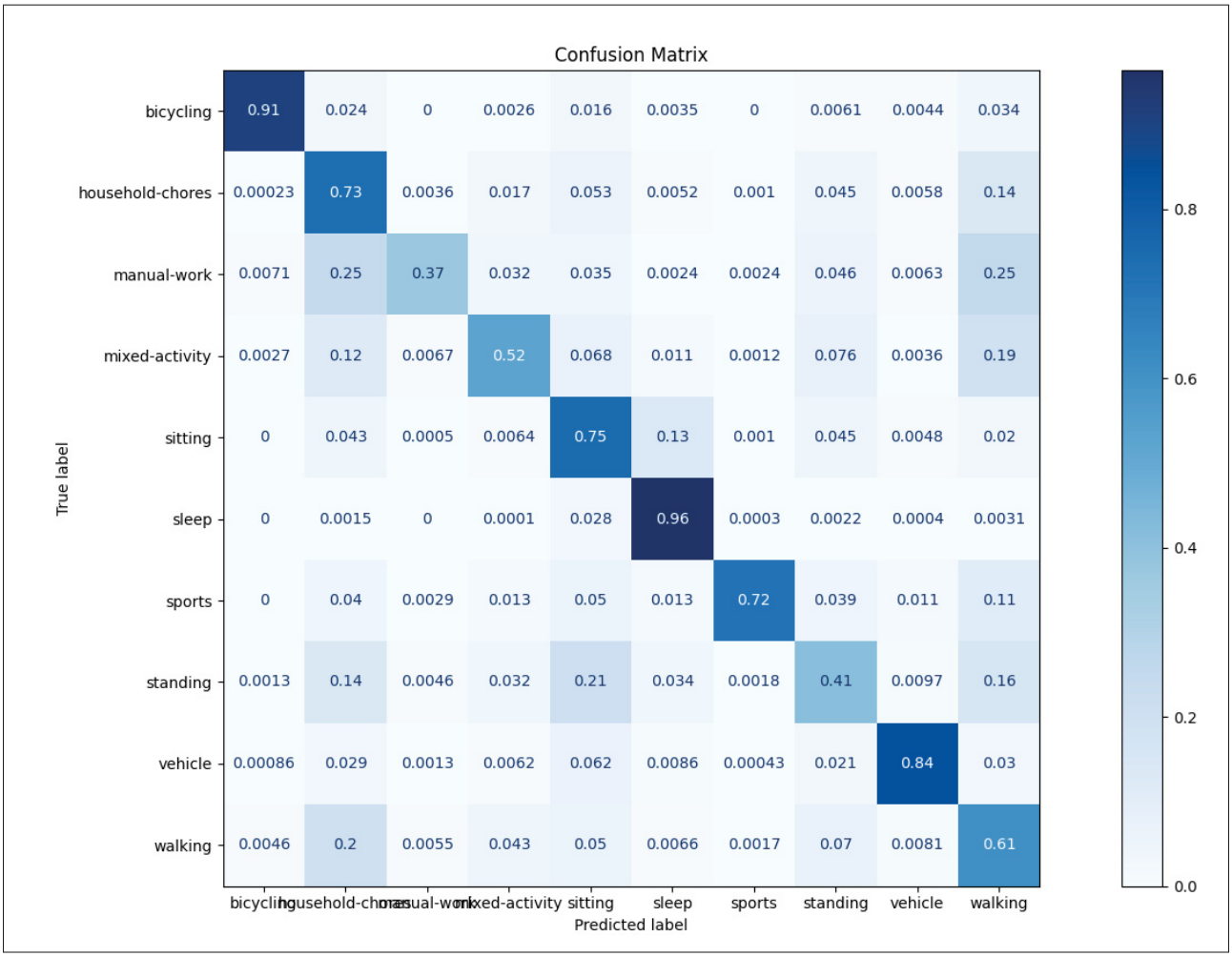


Figure 4.7 Confusion matrix of MEDUSAA using the capture-24 dataset

Tableau 4.6 Complexity (in GFLOPS), step time during training, and peak GPU memory usage for different HAR models.

Method	**GFLOPS**	**Step Time (ms)**	**GPU Mem.(MB)**
MEDUSAA	38.0	54	200
Res-BiLSTM Li & Wang (2022)	36.0	300	420
DkInception Yu & Al-qaness (2025)	42.0	1650	350
THAT Li *et al.* (2021)	24.0	2450	380
attention-CNN Mahmud *et al.* (2020)	29.8	2000	210
S4 Gu *et al.* (2022c)	48.0	320	307

patterns, HAR requires the accurate classification of physical activities from sensor data, a task that similarly demands strong temporal modeling capabilities.

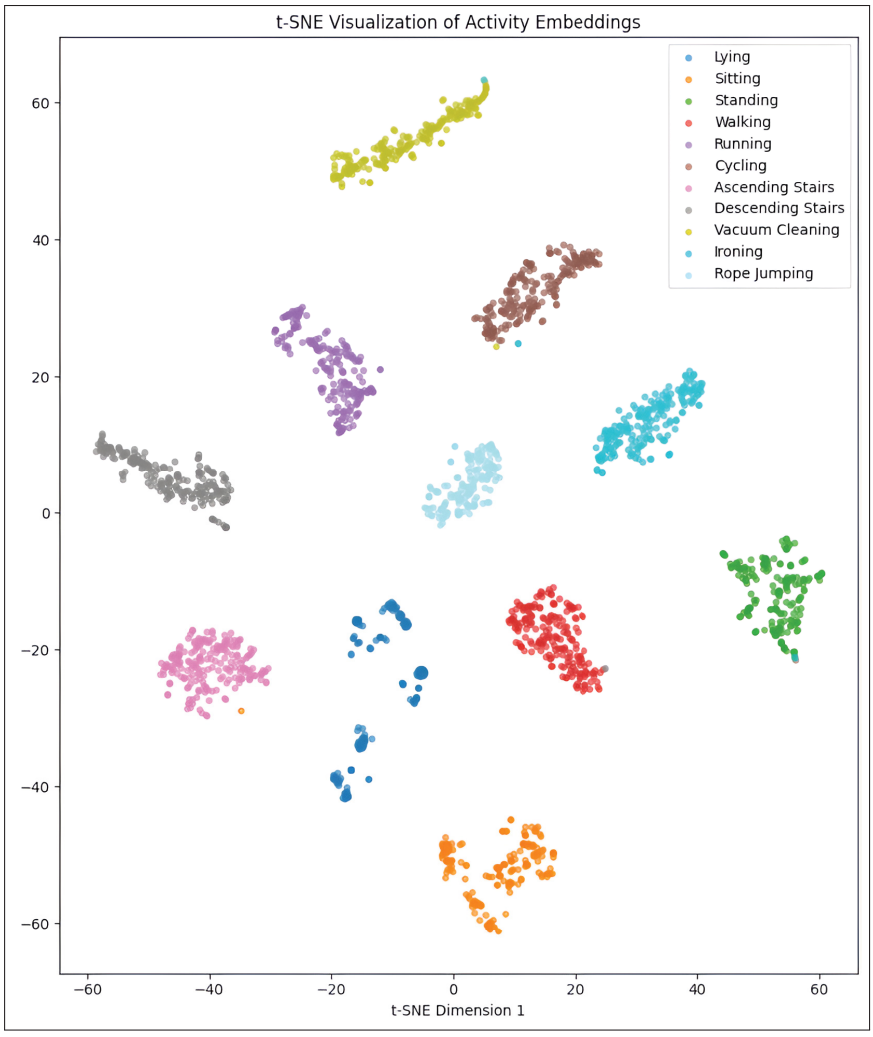


Figure 4.8 t-SNE visualization of activity embeddings (PAMAP2 dataset, 12 activities). Distinct clusters emerge for : 1) Stationary activities (lying, sitting, standing), 2) Locomotion (walking, running), and 3) Household activities (vacuuming, ironing).

To this end, we evaluate our model on two widely used HAR benchmarks : PAMAP2 and Capture-24. These datasets differ significantly in their acquisition conditions. PAMAP2 is collected in controlled laboratory environments, which ensures clean and consistent data but lacks the diversity and unpredictability of real-world activity execution. In contrast, Capture-24 is gathered in the wild during free living conditions, presenting a more realistic but significantly more challenging scenario due to noisy signals, varied user behaviors, and class ambiguities.

Table 4.5 presents the performance comparison between our proposed method and several state-of-the-art HAR approaches on the two aforementioned datasets, in terms of accuracy, precision, F1-score and recall. As shown in this table, our model outperforms state-of-the-art performance on the Capture-24 dataset. This underscores its robustness and strong generalization capability in uncontrolled and real-world conditions, a notable achievement given the difficulties most models encounter outside laboratory settings. On the PAMAP2 dataset, our model also delivers competitive results. However, the more structured nature of this environment allows most architectures and state of the art models to achieve high scores.

In Fig. 4.7, we show the confusion matrix of our model when using the Capture-24 dataset. We can see that high accuracy is achieved for well-defined classes such as sleep (96%) and bicycling (91%), whereas ambiguous activities like manual work and mixed activity show higher misclassification due to overlapping sensor patterns with similar activities (e.g., walking, household chores). This highlights the inherent challenges of performing HAR in real-world, unconstrained settings, while demonstrating our model's ability to consistently deliver high accuracy across diverse activity classes.

On PAMAP2, the t-SNE visualization, depicted in Fig. 4.8, exhibits clearly separated activity clusters, evidencing the architecture's ability to learn compact, discriminative representations from raw sensor data crucial for accurate classification and downstream applications.

Finally, Table 4.6 presents the computational overhead of the evaluated HAR models. Consistent with earlier observations, MEDUSAA achieves the fastest training step time (54 ms, nearly 6× faster than Residual-BiLSTM Li & Wang (2022) and S4 Gu *et al.* (2022c)) and the lowest peak GPU memory usage (200 MB), while maintaining a balanced computational complexity of 38.0 GFLOPS. These results confirm its suitability for deployment on resource-constrained devices without compromising performance.

Overall, these findings demonstrate that our MEDUSAA framework transfers effectively from authentication to HAR, outperforming competing models on challenging real-world datasets. Its combination of state-of-the-art accuracy, computational efficiency, and architectural versatility

positions it as a strong candidate for unified behavioral monitoring systems, capable of supporting both continuous authentication and contextual activity recognition, even under constrained on-device deployment scenarios.

4.6 Conclusion

In this paper, we have presented MEDUSAA, a novel deep learning framework based on a structured state-space architecture, for the dual tasks of continuous user authentication and Human Activity Recognition (HAR) on resource-constrained devices. Our proposed model addresses the prohibitive computational cost of Transformer-based alternatives by employing a diagonal state transition matrix within a Siamese encoder-decoder, optimized with a triplet loss function. This approach efficiently captures long-range temporal dependencies in inertial sensor data to learn highly discriminative embeddings for unique user behaviors and physical activities.

The effectiveness of our proposed framework is demonstrated through extensive evaluations on diverse benchmark datasets. It achieves state-of-the-art Equal Error Rates (EER) for authentication on multimodal (HMOG, FETA) and unimodal (WHUGAIT, PAMAP2, Capture-24) datasets. MEDUSAA also demonstrates strong generalizability, outperforming existing architectures in HAR on the challenging "in-the-wild" Capture-24 dataset. These quantitative achievements, which include a 49x faster inference time than leading alternatives, are qualitatively supported by t-SNE visualizations revealing distinct, compact clusters for different users and activities.

These findings underscore the potential of structured state-space architectures as a robust, scalable, and computationally efficient foundation for unified behavioral monitoring systems, enabling reliable operation in real-world mobile and edge computing environments.

CONCLUSION ET RECOMMANDATIONS

Ce mémoire a introduit des architectures S4HI et MEDUSAA basées sur les modèles à espaces d'états structurés SSM pour l'authentification continue et la reconnaissance d'activités humaines HAR sur des appareils mobiles. En exploitant les données de capteurs multimodaux, nos modèles capturent efficacement les dépendances temporelles à longue portée avec une complexité de calcul linéaire, une alternative légère aux architectures de type Transformer.

Les résultats expérimentaux démontrent une performance supérieure sur plusieurs jeux de données de référence (HMOG, FETA, OU-ISIR, PAMAP2, Capture-24). Nos approches atteignent des taux d'erreur égaux EER plus faibles en authentification et une meilleure précision en HAR, tout en étant jusqu'à 40 fois plus rapides en inférence que les modèles de pointe.

Cette recherche valide ainsi le potentiel des SSM pour le déploiement de systèmes de surveillance comportementale sécurisés, évolutifs et efficaces sur des appareils à ressources limitées, ouvrant la voie à des applications robustes en conditions réelles.

BIBLIOGRAPHIE

Abdellatef, E., Al-Makhlasawy, R. M. & Shalaby, W. A. (2025). Detection of human activities using multi-layer convolutional neural network. *Scientific Reports*, 15(1), 7004. doi : 10.1038/s41598-025-90307-6.

Acien, A., Morales, A., Fierrez, J., Vera-Rodriguez, R. & Delgado-Mohatar, O. (2020). BeCAPTCHA : Behavioral Bot Detection using Touchscreen and Mobile Sensors benchmarked on HuMIdb. Repéré à https://arxiv.org/abs/2005.13655.

Aviv, A. J., Gibson, K., Mossop, E., Blaze, M. & Smith, J. M. (2010). Smudge attacks on smartphone touch screens. *Proceedings of the 4th USENIX Conference on Offensive Technologies*, (WOOT'10), 1–7.

Baevski, A., Zhou, H., Mohamed, A. & Auli, M. (2020). wav2vec 2.0 : A Framework for Self-Supervised Learning of Speech Representations. Repéré à https://arxiv.org/abs/2006.11477.

Berman, J. J. (2016). Chapter 4 - Understanding Your Data. Dans Berman, J. J. (Éd.), *Book - Data Simplification* (pp. 135-187). Boston : Morgan Kaufmann. doi : https://doi.org/10.1016/B978-0-12-803781-2.00004-7.

Chan, S., Yuan, H., Tong, C., Acquah, A., Schonfeldt, A., Gershuny, J. & Doherty, A. (2024). CAPTURE-24 : A large dataset of wrist-worn activity tracker data collected in the wild for human activity recognition. Repéré à https://arxiv.org/abs/2402.19229.

Chow, C. K. & Liu, C. N. (1968). Approximating discrete probability distributions with dependence trees. *IEEE Trans. on Info. Theo.*, IT-14(3), 462–467.

Deep, S. & Zheng, X. (2019). Leveraging CNN and Transfer Learning for Vision-based Human Activity Recognition. *2019 29th International Telecommunication Networks and Applications Conference (ITNAC)*, pp. 1-4. doi : 10.1109/ITNAC46935.2019.9078016.

Delgado-Santos, P., Tolosana, R., Guest, R., Deravi, F. & Vera-Rodriguez, R. (2023a). Exploring transformers for behavioural biometrics : A case study in gait recognition. *Patt. Recogn.*, 143, 109798.

Delgado-Santos, P., Tolosana, R., Guest, R., Vera-Rodriguez, R. & Fierrez, J. (2023b). M-GaitFormer : Mobile biometric gait verification using Transformers. *Engineering Applications of Artificial Intelligence*, 125, 106682. doi : https://doi.org/10.1016/j.engappai.2023.106682.

Estrela, P. M. A. B., Albuquerque, R. O., Amaral, D. M., Giozza, W. F. & Júnior, R. T. S. (2021). A Framework for Continuous Authentication Based on Touch Dynamics Biometrics for Mobile Banking Applications. *Sensors*, 21(12), 4212. doi : 10.3390/s21124212.

Gadaleta, M. & Rossi, M. (2018). IDNet : Smartphone-based gait recognition with convolutional neural networks. *Patt. Recogn.*, 74, 25-37. doi : https://doi.org/10.1016/j.patcog.2017.09.005.

Georgiev, M., Eberz, S., Turner, H., Lovisotto, G. & Martinovic, I. (2022). FETA : Fair Evaluation of Touch-based Authentication. *arXiv preprint*, arXiv :2201.10606. Repéré à https://arxiv.org/abs/2201.10606.

Gu, A., Dao, T., Ermon, S., Rudra, A. & Ré, C. (2020a). HiPPO : Recurrent Memory with Optimal Polynomial Projections. *Proc. Adv. Neural Info. Process. Syst.* doi : https://doi.org/10.48550/arXiv.2008.07669.

Gu, A., Dao, T., Ermon, S., Rudra, A. & Ré, C. (2020b). HiPPO : Recurrent Memory with Optimal Polynomial Projections. *Proc. Adv. Neural Info. Process. Syst.*, 33, 1474–1487.

Gu, A., Goel, K. & Re, C. (2022a). Efficiently Modeling Long Sequences with Structured State Spaces. *Proc. Int. Conf. Learn. Represent. (ICLR)*.

Gu, A., Goel, K. & Ré, C. (2022b). Efficiently Modeling Long Sequences with Structured State Spaces. *Proc. Int. Conf. Learn. Represent. (ICLR)*.

Gu, A., Johnson, I., Timalsina, A., Rudra, A. & Ré, C. (2022c). How to Train Your HiPPO : State Space Models with Generalized Orthogonal Basis Projections. Repéré à https://arxiv.org/abs/2206.12037.

Gupta, A., Gu, A. & Berant, J. (2022). Diagonal state spaces are as effective as structured state spaces. *Proc. Int. Conf. Neural Info. Process. Syst. (NeurIPS)*.

Hamad, R. A., Kimura, M., Yang, L., Woo, W. L. & Wei, B. (2021). Dilated causal convolution with multi-head self attention for sensor human activity recognition. *Neural Comput. Appl.*, 33(20), 13705–13722. doi : 10.1007/s00521-021-06007-5.

Iwama, H., Okumura, M., Makihara, Y. & Yagi, Y. (2012). The OU-ISIR Gait Database Comprising the Large Population Dataset and Performance Evaluation of Gait Recognition. *IEEE Trans. Info. Forens. Sec.*, 7(5), 1511-1521. doi : 10.1109/TIFS.2012.2204253.

Lecun, Y., Bottou, L., Bengio, Y. & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278-2324. doi : 10.1109/5.726791.

Li, B., Cui, W., Wang, W., Zhang, L., Chen, Z. & Wu, M. (2021). Two-Stream Convolution Augmented Transformer for Human Activity Recognition. *Proc. Conf. on AI (AAAI)*, 35(1), 286-293. doi : 10.1609/aaai.v35i1.16103.

Li, Y. & Wang, L. (2022). Human Activity Recognition Based on Residual Network and BiLSTM. *Sensors*, 22(2). doi : 10.3390/s22020635.

Mahmud, S., Tonmoy, M. T. H., Bhaumik, K. K., Rahman, A. M., Amin, M. A., Shoyaib, M., Khan, M. A. R. & Ali, A. (2020). Human Activity Recognition from Wearable Sensor Data Using Self-Attention. *arXiv preprint arXiv :2003.09018*.

McCulloch, W. S. & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5, 115–133. doi : 10.1007/BF02478259.

Mekruksavanich, S. & Jitpattanakul, A. (2023). A Deep Learning Network with Aggregation Residual Transformation for Human Activity Recognition Using Inertial and Stretch Sensors. *Computers*, 12(7). doi : 10.3390/computers12070141.

Minsky, M. & Papert, S. (1969). *Perceptrons : An Introduction to Computational Geometry*. Cambridge, MA : MIT Press.

Muller, M. (2007). Dynamic Time Warping. Dans *Book - Information Retrieval for Music and Motion* (pp. 69–84). Berlin, Heidelberg : Springer Berlin Heidelberg. doi : 10.1007/978-3-540-74048-3_4.

Nickel, C., Wirtl, T. & Busch, C. (2012). Authentication of Smartphone Users Based on the Way They Walk Using k-NN Algorithm. *Proc. Int. Conf. Intelli. Info. Hiding Multimed. Sig. Process.*, pp. 16-20. doi : 10.1109/IIH-MSP.2012.11.

Nithyakani, P., Shanthini, A. & Ponsam, G. (2019). Human Gait Recognition using Deep Convolutional Neural Network. *2019 3rd International Conference on Computing and Communications Technologies (ICCCT)*, pp. 208-211. doi : 10.1109/ICCCT2.2019.8824836.

Palin, K., Feit, A. M., Kim, S., Kristensson, P. O. & Oulasvirta, A. (2019). How do People Type on Mobile Devices ? Observations from a Study with 37,000 Volunteers. *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services*, (MobileHCI '19). doi : 10.1145/3338286.3340120.

Park, G., Lee, K. M. & Koo, S. (2021). Uniqueness of gait kinematics in a cohort study. *Sci. Rep.*, 11.

Park, J., Kim, D.-W. & Lee, J. (2024). CALANet : Cheap All-Layer Aggregation for Human Activity Recognition. *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. Repéré à https://openreview.net/forum?id=ouoBW2PXFQ.

Pascanu, R., Mikolov, T. & Bengio, Y. (2012). Understanding the exploding gradient problem. *CoRR*, abs/1211.5063. Repéré à http://arxiv.org/abs/1211.5063.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems : Networks of Plausible Inference*. San Mateo, CA : Morgan Kaufman Publishers.

Pei, Y., Biswas, S., Fussell, D. S. & Pingali, K. (2019). An elementary introduction to Kalman filtering. *Commun. ACM*, 62(11), 122–133. doi : 10.1145/3363294.

Peinado-Contreras, A. & Munoz-Organero, M. (2020). Gait-Based Identification Using Deep Recurrent Neural Networks and Acceleration Patterns. *Sensors*, 20(23). doi : 10.3390/s20236900.

Reiss, A. [Accessed : 2025-08-08]. (2012). PAMAP2 Physical Activity Monitoring. UCI Machine Learning Repository. Repéré à https://doi.org/10.24432/C5NW2H.

Rojanavasu, P., Jantawong, P., Jitpattanakul, A. & Mekruksavanich, S. (2023). Improving Inertial Sensor-based Human Activity Recognition using Ensemble Deep Learning. *Proc. Joint Int. Conf. Dig. Arts Med. Technol. with ECTI Northern Sec. Conf. Elec. Electron. Compu. Telecom. Engineer. (ECTI DAMT & NCON)*, pp. 488-492. doi : 10.1109/ECTIDAMTNCON57770.2023.10139689.

Ronneberger, O., Fischer, P. & Brox, T. (2015). U-Net : Convolutional Networks for Biomedical Image Segmentation. *Proc. Med. Image Comput. Computer-Assist. Intervention (MICCAI)*, pp. 234–241.

Rosenblatt, F. (1958). The perceptron : A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6), 386–408. doi : 10.1037/h0042519.

Rumelhart, D. E., Hinton, G. E. & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088), 533–536. doi : 10.1038/323533a0.

Saha, U., Saha, S., Kabir, T., Fattah, S. A. & Saquib, M. (2024). Decoding Human Activities : Analyzing Wearable Accelerometer and Gyroscope Data for Activity Recognition. Repéré à https://arxiv.org/abs/2310.02011.

Salama, R., Al-Turjman, S., Altrjman, C., Al-Turjman, F., Prakash, R., Yadav, S. P. & Vats, S. (2023). Authentication using Biometric Data from Mobile Cloud Computing in Smart Cities. *Proc. Int. Conf. Adv. Electron. & Commun. Eng. (AECE)*, pp. 445-448. doi : 10.1109/AECE59614.2023.10428426.

Senerath, D., Tharinda, S., Vishwajith, M., Rasnayaka, S., Wickramanayake, S. & Meedeniya, D. (2023). BehaveFormer : A Framework with Spatio-Temporal Dual Attention Transformers for IMU enhanced Keystroke Dynamics. Repéré à https://arxiv.org/abs/2307.11000.

Sepas-Moghaddam, A. & Etemad, A. (2021). Deep Gait Recognition : A Survey. *CoRR*, abs/2102.09546. Repéré à https://arxiv.org/abs/2102.09546.

Sitová, Z., Sedenka, J., Yang, Q., Peng, G., Zhou, G., Gasti, P. & Balagani, K. (2015). HMOG : New Behavioral Biometric Features for Continuous Authentication of Smartphone Users. *IEEE Transactions on Information Forensics and Security*, 11, 1-1. doi : 10.1109/TIFS.2015.2506542.

Stragapede, G., Vera-Rodriguez, R., Tolosana, R., Morales, A., Acien, A. & Lan, G. L. (2022). Mobile Passive Authentication through Touchscreen and Background Sensor Data. *2022 International Workshop on Biometrics and Forensics (IWBF)*, pp. 1-6. doi : 10.1109/IWBF55382.2022.9794524.

Tay, Y., Dehghani, M., Abnar, S., Shen, Y., Bahri, D., Pham, P., Rao, J., Yang, L., Ruder, S. & Metzler, D. (2021). Long Range Arena : A Benchmark for Efficient Transformers. *Proc. Int. Conf. Learn. Represent. (ICLR)*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. & Polosukhin, I. (2017). Attention Is All You Need. *CoRR*, abs/1706.03762. Repéré à http://arxiv.org/abs/1706.03762.

Watanabe, Y. (2014). Influence of Holding Smart Phone for Acceleration-Based Gait Authentication. *Proc. Int. Conf. Emerg. Sec. Technol.*, pp. 30-33. doi : 10.1109/EST.2014.24.

Wren, T. A. L., Wren, T. A. L., Do, K. P., Rethlefsen, S. A. & Healy, B. S. (2006). Cross-correlation as a method for comparing dynamic electromyography signals during gait. *J. of Biomech.*, 39 14, 2714-2718.

Yan, Q., Liu, K., Zhou, Q., Guo, H. & Zhang, N. (2020, 01). *SurfingAttack : Interactive Hidden Attack on Voice Assistants Using Ultrasonic Guided Waves*. Proc. Netw. and Dist. Syst. Sec. Symp.

Yu, X. & Al-qaness, M. A. A. (2025). Human Activity Recognition Using Deep Residual Convolutional Network Based on Wearable Sensors. *IEEE Journal of Biomedical and Health Informatics*, 29(3), 1950-1958. doi : 10.1109/JBHI.2024.3510860.

Yuan, H., Chan, S., Creagh, A. P., Tong, C., Acquah, A., Clifton, D. A. & Doherty, A. (2024). Self-supervised learning for human activity recognition using 700,000 person-days of wearable data. *npj Digital Medicine*, 7(1). doi : 10.1038/s41746-024-01062-3.

Zeng, M., Nguyen, L. T., Yu, B., Mengshoel, O. J., Zhu, J., Wu, P. & Zhang, J. (2014). Convolutional Neural Networks for human activity recognition using mobile sensors. *Proc. Int. Conf. Mob. Comput. Appl. Serv.*, pp. 197-205. doi : 10.4108/icst.mobicase.2014.257786.

Zhou, F., Wang, R., Su, H. & Xu, S. (2022). A Human Activity Recognition Model Based on Wearable Sensor. *2022 9th International Conference on Digital Home (ICDH)*, pp. 169-174. doi : 10.1109/ICDH57206.2022.00033.

Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H. & Zhang, W. (2020). Informer : Beyond Efficient Transformer for Long Sequence Time-Series Forecasting. *Proc. Conf. on AI (AAAI 2020)*.

Zou, Q., Wang, Y., Wang, Q., Zhao, Y. & Li, Q. (2020). Deep Learning-Based Gait Recognition Using Smartphones in the Wild. *IEEE Trans. Info. Forens. Sec.*, 15, 3197-3212. doi : 10.1109/TIFS.2020.2985628.