Leevi Enontekiö

Model Distillation in Federated Learning for Human Action Recognition

Master's thesis

School of Technology and Innovations

Industrial Engineering and Management, Master of Science in Technology

Industrial Systems Analytics

Vaasa 2025

VAASAN YLIOPISTO
Tekniikan ja innovaatiojohtamisen akateeminen yksikkö
Tekijä: Leevi Enontekiö
Tutkielman nimi: Model Distillation in Federated Learning for Human Action Recognition Master's thesis
Tutkinto: Diplomi-insinööri
Ohjelma: Industrial Systems Analytics
Työn ohjaaja: Jani Boutellier, Bo Tan, Masud Fahim
Valmistumisvuosi: 2025 **Sivumäärä:** 57

TIIVISTELMÄ:
Tämä maisterintyö tutkii yksityisyyttä säilyttävän koneoppimisen ja ihmisen toiminnan tunnistamisen (Human action recognition - HAR) risteyskohtaa tarkastelemalla hajautetun koneoppimisen lähestymistapoja reunalaitteille. HAR-järjestelmien yleistyessä, terveydenhuollossa, turvallisuudessa ja älykkäissä ympäristöissä huoli yksityisyyden suojasta kasvaa samassa suhteessa. Työ käsittelee näitä huolenaiheita arvioimalla hajautetun koneoppimisen käyttökelpoisuutta yksityisyyttä säilyttävänä vaihtoehtona perinteiselle keskitetylle koneoppimiselle.
Työssä käytetään CheckMATE (Checking Mutually Average Temporal Encapsulation) -algoritmia videoiden distilloimiseksi kuviksi, muuttaen monimutkaisen videoluokittelun helpommin hallittaviksi kuvanluokittelutehtäviksi, jotka soveltuvat paremmin resursseiltaan rajoitetuille reunalaitteille. UCF101- ja HMDB51-tietoaineistoja käyttäen vertailukohtina, toteutamme ja vertailemme kolmea hajautettua optimointistrategiaa (FedAVG, FedProx ja FedYogi) keskitettyyn koneoppmiseen.
Tuloksemme osoittavat, että hajautettu oppiminen CheckMATE:n kanssa voi saavuttaa keskitettyihin koneoppmisen lähestymistapoihin verrattavan suorituskyvyn, vain 0,2%:n tarkkuuden laskulla UCF101:ssä (80,5% vs. 80,7%) ja 7,1%:n erolla haastavammassa HMDB51-tietoaineistossa (59,7% vs. 66,8%). Havaitsemme myös, että eri hajautetun oppimisen optimointistrategiat toimivat eroavasti riippuen tietoaineiston jakaumista, FedYogin suoriutuessa parhaiten UCF101:ssä ja FedProxin HMDB51:ssä.
Tämä työ edistää kasvavaa yksityisyyttä säilyttävän koneoppimisen alaa, tarjoamalla empiiristä näyttöä siitä, että hajautetun koneoppimisen HAR-järjestelmät voivat saavuttaa hyväksyttävän tarkkuuden säilyttäen samalla yksityisyydensuojan. Se korostaa myös CheckMATE-algoritmin tehokkuutta laskentavaatimusten vähentämisessä reunalaitteille. Näillä löydöksillä on merkittäviä vaikutuksia yksityisyydensuojaa kunnioittavien HAR-sovellusten kehittämiseen aloille, joille arkaluonteisten henkilötietojen suojaaminen on ensiarvoisen tärkeää, uhraamatta merkittävästi sovellusten tehokkuudessa.

Avainsanat: Hajautettu koneoppiminen, Ihmisen toiminnan tunnistaminen, Yksityisyyden säilyttävä koneoppiminen, Reunalaskenta, Konenäkö, CheckMATE

UNIVERSITY OF VAASA
School of Technology and Innovations
Author: Leevi Enontekiö
Thesis title: Model Distillation in Federated Learning for Human Action Recognition Master's thesis
Degree: Master of Science in Technology
Programme: Industrial Systems Analytics
Supervisor: Jani Boutellier, Bo Tan, Masud Fahim
Year of graduation: 2025 **Number of pages:** 57

ABSTRACT:
This thesis explores the intersection of privacy-preserving machine learning and human action recognition (HAR) by investigating federated learning approaches for edge devices. As HAR systems become increasingly prevalent in healthcare, security, and smart environments, concerns about data privacy have grown proportionally. This research addresses these concerns by evaluating the viability of federated learning as a privacy-preserving alternative to traditional centralized training methods.
The study employs the CheckMATE (Checking Mutually Average Temporal Encapsulation) algorithm to distill video sequences into representative frames, transforming complex video classification into more manageable image classification tasks suitable for resource-constrained edge devices. Using the UCF101 and HMDB51 datasets as benchmarks, we implement and compare three federated optimization strategies (FedAVG, FedProx, and FedYogi) against centralized training.
Our results demonstrate that federated learning with CheckMATE can achieve performance comparable to centralized approaches, with only a 0.2% accuracy drop on UCF101 (80.5% vs. 80.7%) and a 7.1% gap on the more challenging HMDB51 dataset (59.7% vs. 66.8%). We observe that different federated optimization strategies excel under different data distribution characteristics, with FedYogi performing best on UCF101 and FedProx on HMDB51.
This research contributes to the growing field of privacy-preserving machine learning by providing empirical evidence that federated HAR systems can maintain acceptable accuracy while preserving data privacy. It also highlights the effectiveness of video distillation techniques in reducing computational requirements for edge deployment. These findings have significant implications for developing privacy-respectful HAR applications in domains where protecting sensitive personal data is paramount.

Keywords: Federated Learning, Human Action Recognition, Privacy-Preserving Machine Learning, Edge Computing, Computer Vision, CheckMATE

Contents

Figures

Tables

Algorithms

Abbreviations

2D Two-dimensional

3D Three-dimensional

CheckMATE Checking Mutually Average Temporal Encapsulation

CNN Convolutional Neural Network

FedAvg Federated Averaging

FedProx Federated Proximal

FedYogi Federated Yogi

FL Federated Learning

GDPR General Data Protection Regulation

GPU Graphics Processing Unit

HAR Human Action Recognition

HMDB51 Human Motion Database (51 action categories)

IID Independent and Identically Distributed

IoT Internet of Things

LEAF Large-scale and Extensible Toolkit for Federated Learning

non-IID Non-Independent and Identically Distributed

RGB Red-Green-Blue

RNN Recurrent Neural Network

UCF101 University of Central Florida 101 action categories dataset

ViT Vision Transformer

VRAM Video Random Access Memory

1 Introduction

The ability to recognize and interpret human actions has been fundamental to human social interaction and decision-making throughout history. In the domain of artificial intelligence, translating this capability into computational systems has emerged as human action recognition (HAR), a field that bridges the gap between human behavior and machine understanding. As our world becomes increasingly automated, HAR has grown from a theoretical concept into a crucial technology with applications spanning healthcare monitoring, security systems, smart environments, and human-computer interaction.

In recent years, deep learning has revolutionized the field of computer vision, dramatically enhancing the performance of HAR systems. The shift from traditional handcrafted feature extraction methods to end-to-end deep neural networks has enabled unprecedented levels of accuracy in recognizing complex human behaviors from visual data. Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and more recently, transformer-based architectures have pushed the boundaries of what's possible in action recognition tasks. These advances have made it feasible to deploy HAR systems in increasingly complex real-world environments and applications.

However, this progress comes with significant privacy concerns (Yu et al., 2024). HAR systems typically require vast amounts of sensitive personal data for training and operation. Video recordings of individuals in their homes, workplaces, or public spaces raise serious questions about data protection, consent, and potential misuse. As organizations deploy these technologies more widely, balancing the benefits of HAR with respect for individual privacy has become an increasingly critical challenge requiring careful and meticulous handling of data as well as conveying trust to end user (Knowles, 2016).

Federated Learning (FL) has emerged as a promising approach to address these privacy concerns in machine learning applications. Unlike conventional training methods that

require centralizing all data in one location, FL enables model training across multiple decentralized devices while keeping the raw data local. This paradigm shift allows edge devices to collaboratively learn a shared model while preventing sensitive data from leaving the device (Mammen, 2021). The core principle of FL involves iterative model updates: a central server distributes the current model to participating devices, each device improves the model using only its local data, and then the server aggregates these local updates to enhance the global model. This process not only protects privacy but also reduces bandwidth requirements and enables learning from data that might otherwise be inaccessible due to privacy regulations or practical constraints.

While federated learning offers privacy benefits, processing video data on edge devices presents significant computational challenges. The CheckMATE (Checking Mutually Average Temporal Encapsulation) algorithm addresses this by efficiently distilling video sequences into representative frames that capture essential spatiotemporal information (Fahim & Boutellier, 2024). This technique transforms the complex task of video analysis into a more manageable image classification problem. By condensing the relevant information from multiple frames into a single representation, CheckMATE potentially reduces the memory footprint and computational requirements, making it particularly suitable for federated learning scenarios where edge devices may have limited resources.

This research aims to address key challenges at the intersection of human action recognition and privacy-preserving machine learning. It explores whether the combination of federated learning and efficient video processing techniques can enable privacy-respectful HAR systems without compromising performance. Most importantly, the study seeks to answer the following research questions:

- **How does a distributedly trained model perform on camera-based human action recognition tasks compared to traditionally trained models?**

  This question addresses a critical gap in current HAR research. Although federated learning has shown promise in various domains, its effectiveness for vision-based

HAR systems remains underexplored. Camera-based HAR systems present unique challenges due to their high-dimensional data and complex feature relationships. Understanding the performance implications of distributed training is crucial for determining whether privacy-preserving HAR systems can achieve comparable accuracy to traditional centralized approaches.

- **How does distilling datasets with the CheckMATE procedure affect federated learning and final model performance?**

  The computational and communication demands of processing video data pose significant challenges for edge devices in federated learning systems. CheckMATE offers a potential solution by condensing video information into representative frames, but its impact on federated learning performance has not been thoroughly investigated. This question explores whether the benefits of reduced computational overhead can be achieved without sacrificing model accuracy.

These questions are particularly relevant as organizations increasingly seek to deploy HAR systems while respecting user privacy and working within the constraints of edge devices. The answers will provide practical insights for implementing privacy-preserving HAR systems in real-world applications.

For clarity, it should be noted that human action recognition (HAR) is sometimes used interchangeably with terms such as human activity recognition or human gesture recognition, but for consistency, we will use the term human action recognition throughout this thesis.

2 Background

2.1 Human Action Recognition

Human action recognition (HAR) represents the intersection of machine perception and human behavior understanding. At its core, HAR is a field of technology that enables machines to identify, interpret, and classify human actions in real-time or from recorded data. This capability emerges from the integration of various sensing technologies with advanced computer vision and machine learning approaches that can process and categorize complex human movements into meaningful activities (Jegham et al., 2020).

The evolution of HAR over the past two decades mirrors broader technological advances in computing power, sensor technology, and artificial intelligence. This progression has transformed HAR from a theoretical concept into a practical technology with wide-ranging applications. In healthcare, HAR systems monitor patient movements and detect falls or irregular behavior patterns (Zhou et al., 2020). In security applications, they identify suspicious activities or potential threats (Shao et al., 2015). The technology has also become integral to human-computer interaction, enabling more natural interfaces in virtual and augmented reality environments, and advancing the capabilities of robotic systems that must interact safely with humans (Haria et al., 2017).

The continued interest in HAR development is driven by more than just technological capability - it addresses fundamental needs in our increasingly automated world. As the boundaries between human and machine interactions blur, the ability to accurately interpret human actions becomes crucial for creating responsive and adaptive systems. This importance is particularly evident in emerging ecosystems where seamless human-computer interaction can significantly enhance efficiency, safety, and user experience (Beddiar et al., 2020).

To understand how HAR systems achieve these capabilities, we must examine their fun-

damental components and implementation approaches. Modern HAR systems operate through a sophisticated pipeline: first capturing human actions through various sensing modalities, then processing this raw data into meaningful features, and finally employing classification algorithms to recognize specific actions. Each stage presents unique technical challenges and requires careful design decisions that ultimately determine the system's effectiveness in real-world applications.

2.1.1 Techniques for Human Action Recognition

Understanding human actions requires capturing complex spatial and temporal information that humans process instinctively but machines must break down into measurable components. HAR systems have evolved to use diverse sensing approaches, each offering unique insights into human movement and behavior. These sensing modalities broadly fall into two categories: visual and non-visual data collection.

Visual data collection, encompassing RGB video, depth sensing, and infrared imaging, attempts to mirror human visual perception. These methods can capture subtle details of movement, posture, and interaction with the environment that might be missed by other sensing approaches (Z. Sun et al., 2022). For instance, RGB cameras capture color and texture information similar to human vision, while depth sensors add crucial spatial understanding that helps distinguish between overlapping movements. The field of vision-based human action recognition has witnessed a clear evolutionary trajectory in its modeling approaches. Early systems have relied heavily on handcrafted features with manually designed descriptors (Gu et al., 2021), which were limited in their representational capacity. With the advent of deep learning, 2D convolutional networks emerged as powerful feature extractors, often complemented by optical flow to capture temporal dynamics. As computational resources expanded, researchers shifted toward 3D CNNs that could directly model spatiotemporal relationships (X. Wang et al., 2020).

Non-visual sensing takes a different approach, measuring physical variables like acceleration, audio signals, or even WiFi signal disturbances. These methods often provide complementary information that visual sensors might miss. For instance, accelerometers can precisely measure movement intensity and patterns, while audio sensors can capture associated sounds that might indicate specific activities.

The choice between single-modality and multi-modal approaches depends on the complexity of the actions being recognized. Simple actions like walking or running might be reliably detected using a single sensor type, but more complex activities often benefit from combining multiple modalities. When monitoring someone preparing a meal, for example, visual data might track body movements while acoustic sensors detect cooking sounds, providing a more complete understanding of the activity (Ramanujam et al., 2021).

The rising prominence of non-visual sensors, particularly in smart devices and Internet of Things (IoT) applications, reflects growing privacy concerns surrounding vision-based monitoring (Chen et al., 2021). However, given this thesis's focus on vision-based approaches and their unique capabilities in capturing complex human actions, we will concentrate our discussion on visual sensing methods while acknowledging the broader sensing landscape.

Modality	Pros	Cons
RGB	• Provide rich appearance information • Easy to obtain and operate • Wide range of applications	• Sensitive to viewpoint • Sensitive to background • Sensitive to illumination
3D Skeleton	• Provide 3D structural information of subject pose • Simple yet informative • Insensitive to viewpoint • Insensitive to background	• Lack of appearance information • Lack of detailed shape information
Depth	• Provide 3D structural information • Provide geometric shape information	• Lack of color and texture information • Limited workable distance
Infrared Sequence	• Workable in dark environments	• Lack of color and texture information • Susceptible to sunlight
Point Cloud	• Provide 3D information • Provide geometric shape information • Insensitive to viewpoint	• Lack of color and texture information • High computational complexity
Event Stream	• Avoid much visual redundancy • High dynamic range • No motion blur	• Asynchronous output • Spatio-temporally sparse • Capturing device is relatively expensive
Audio	• Easy to locate actions in temporal sequence	• Lack of appearance information
Acceleration	• Can be used for fine-grained HAR • Privacy protecting • Low cost	• Lack of appearance information • Capturing device needs to be carried by subject
Radar	• Can be used for through-wall HAR • Insensitive to illumination • Insensitive to weather • Privacy protecting	• Lack of appearance information • Capturing device is relatively expensive
WiFi	• Simple and convenient • Privacy protecting • Low cost	• Lack of appearance information • Sensitive to environments • Noisy

Table 1. Comparison of Different Modalities for Human Activity Recognition, adapted from Z. Sun et al. (2022).

2.1.2 Vision-based Techniques

Computer vision techniques can be applied to vision-based data to extract features that indicate certain activities. The main advantage is the ease of applying these techniques for good results, but the main caveat is privacy concerns (Hussain et al., 2020). Despite privacy concerns, vision-based approaches remain a large area of interest for research, as images and videos offer rich and useful data that other nonvision-based hardware cannot compete with (Jegham et al., 2020). Vision-based approaches can also be considered a societally more acceptable technique due to other methods that require ”intrusive” adoption of sensors that a person would need to wear for data collection (Beddiar et al., 2020).

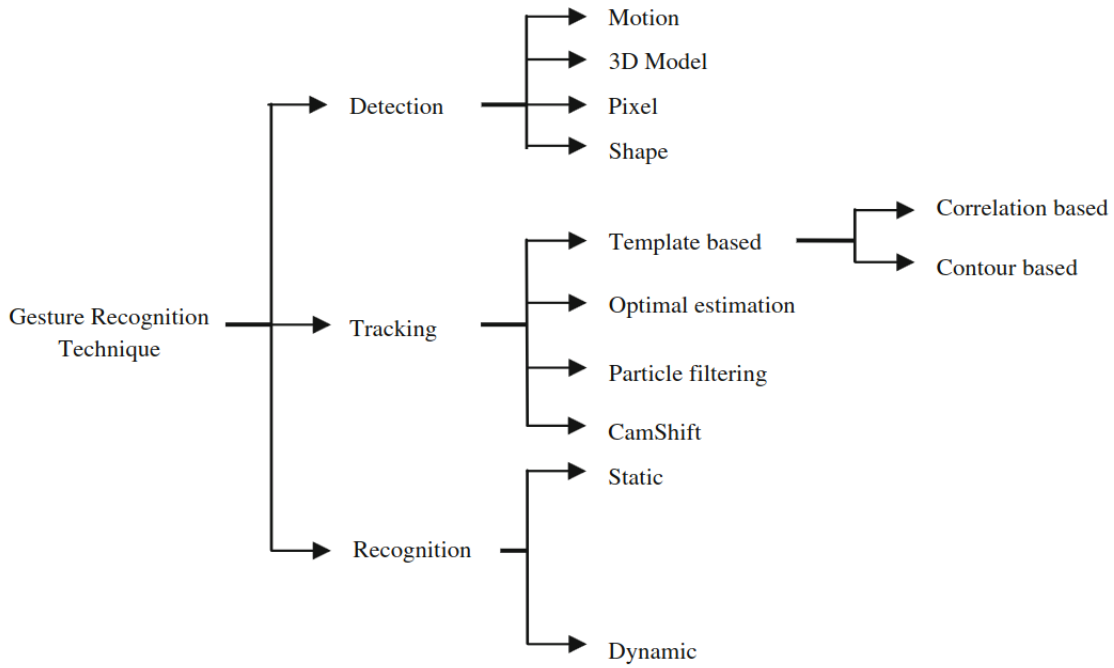

Figure 1. Human action recognition techniques, adapted from Rautaray and Agrawal (2015).

Vision-based action recognition consists of three distinct operations: detection, tracking, and classification, which flow in the respective order from start of capturing an action to determining the type of action (Rautaray & Agrawal, 2015). Detection phase seeks to segment the action from everything else going on in the video or image. From this, features can be extracted either through hand-crafted methods or utilizing deep learn-

ing approaches (Beddiar et al., 2020). Lastly, classification is made by an algorithmic approach of choice. One choice is optical flow that offers a displacement field between two images. Essentially, separating motion from appearance (Simonyan & Zisserman, 2014). The field of displacement maps and original images can then be passed through two stream network for classification.

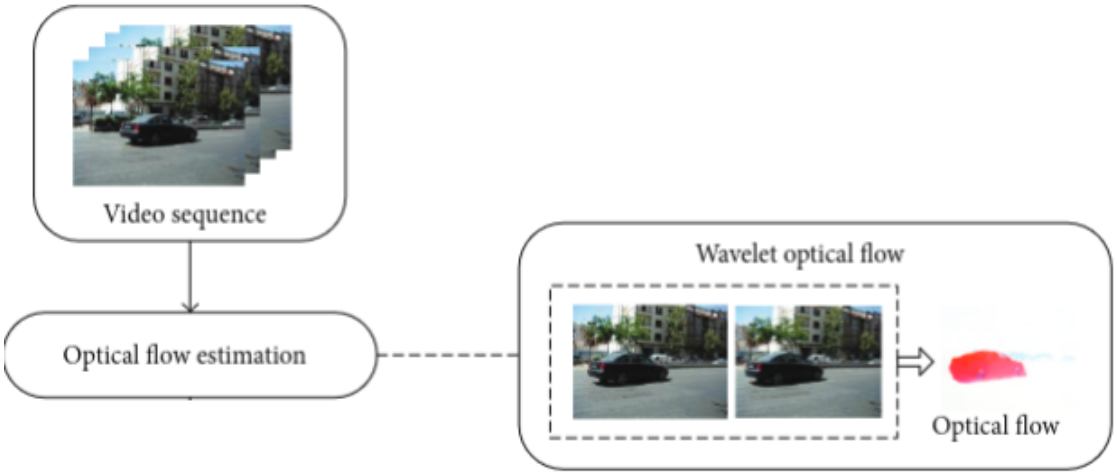

Figure 2. Motion captured and separated into horizontal and vertical components, adapted from Han et al. (2013).

While the preceding figure may depict HAR as achieved through separate techniques, a growing trend in contemporary research involves the application of deep learning methods. Deep learning offers a distinct advantage over conventional machine learning approaches in HAR and other domains. Unlike traditional methods, deep learning necessitates less manual intervention in the form of hand-crafted feature extraction and data preprocessing steps. This is particularly advantageous given the current advancements in data acquisition facilitated by the Internet of Things (IoT) and the exponential growth in computational power (LeCun et al., 2015).

Convolutional Neural Networks (CNNs) have proven particularly adept at action classification, object segmentation, and various other tasks within image and video stream data analysis. Initially as 2D CNNs and later 3D for capturing spatiotemporal features, recent years have witnessed the introduction of novel architectures like MobileNets, which achieve high accuracy levels (up to 75.2%) on the industry-standard ImageNet dataset

(Howard et al., 2019). This dataset encompasses 1,000 distinct object classes with annotations specifying object locations and classifications within each image (Russakovsky et al., 2015). Capitalizing on transfer learning, researchers have successfully adapted MobileNets trained on ImageNet to Human action Recognition (HAR) problems. This approach has yielded promising results, achieving F1-scores of 98.12% on image datasets containing sign language gestures (Liao et al., 2021).

Although CNNs have been effective at HAR tasks, the current state-of-the-art has been dominated by transformer-based architectures, which have demonstrated superior performance across benchmark datasets. Notable examples include VideoMaeV2 from Meta AI (L. Wang et al., 2023) and InternVideo2 (Y. Wang et al., 2024), both leveraging the self-attention mechanisms of Vision Transformers (ViT) (Kolesnikov et al., 2021) to achieve unprecedented accuracy on datasets such as HMDB51 and Kinetics-600 (Wensel et al., 2023). This progression from 2D to 3D CNN approaches, and subsequently to transformer-based models, reflects both the increasing computational capabilities available to researchers and the fundamental need to effectively capture spatiotemporal relationships in video understanding tasks.

A significant challenge with spatiotemporal models lies in the computational demands of video processing. With standard videos capturing 30 frames per second (fps), one minute of uncompressed high-definition content (1920×1080) requires approximately 10 GB of storage space. To address these resource constraints, researchers typically implement training protocols with reduced frame rates and spatial resolutions (Johnson, 2022). An effective sampling strategy involves selecting nonconsecutive frames at predetermined intervals, such as extracting 16 frames with a step size of six frames between selections. This process can be executed multiple times per video, with the model's final prediction calculated as the average across all sampling iterations, thereby balancing computational efficiency with model performance.

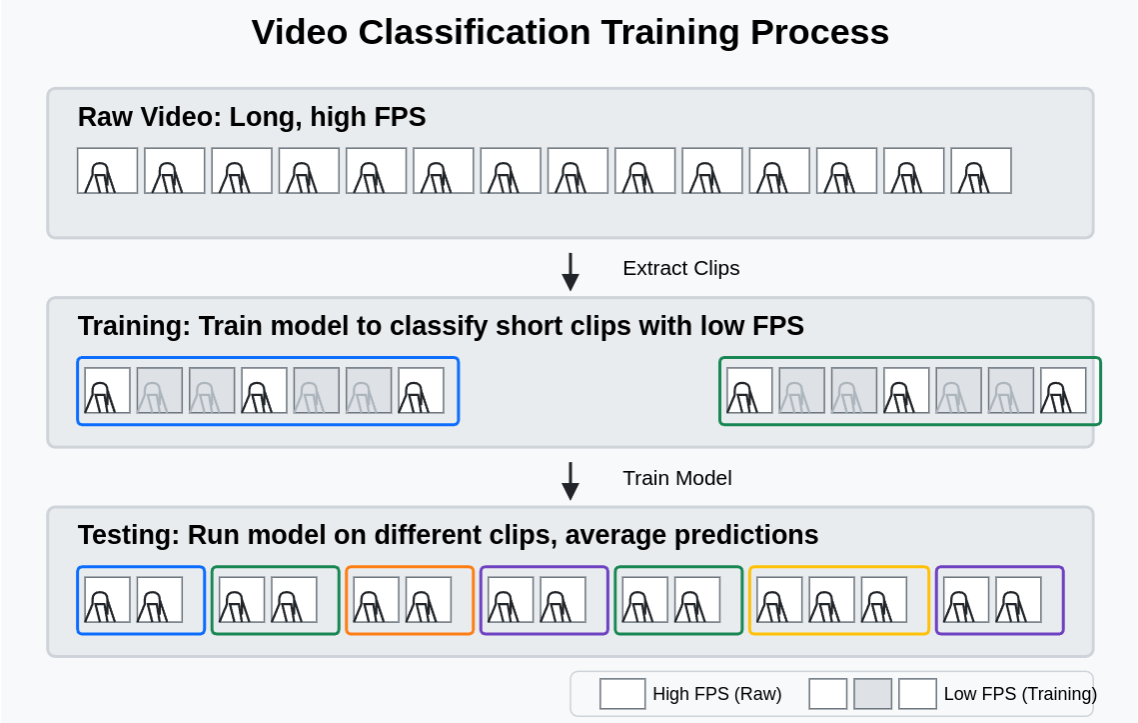

Figure 3. Low fps and spatial resolution sampling of video.

2.1.3 Human Action Categorization

HAR research requires systematic categorization frameworks to organize the diverse range of human activities. These frameworks serve as conceptual foundations that guide algorithm development, dataset design, and evaluation methodologies. This section examines the evolution of action categorization approaches and their practical implications for HAR systems.

Early HAR research predominantly employed hierarchical frameworks that decomposed complex activities into fundamental, low-level categories (Minh Dang et al., 2020). These frameworks typically organized human movements along a spectrum of increasing complexity, comprising four to six distinct levels. The progression moved from simple, atomic movements to complex social interactions, with recognition difficulty generally increasing at higher levels (Beddiar et al., 2020).

The field has subsequently evolved toward more data-driven methodologies. Contem-

porary large-scale datasets such as Kinetics (Zisserman et al., 2017), UCF101 (Soomro et al., 2012), and others have largely shifted to flat classification schemes without explicit hierarchical organization. This transition reflects both the increased capacity of deep learning models to directly classify diverse action categories and the practical challenges in maintaining consistent hierarchical relationships across heterogeneous real-world actions. Nevertheless, understanding the conceptual hierarchy remains valuable for system design and analysis.

A typical hierarchical framework organizes human actions into levels of increasing complexity and social dimension:

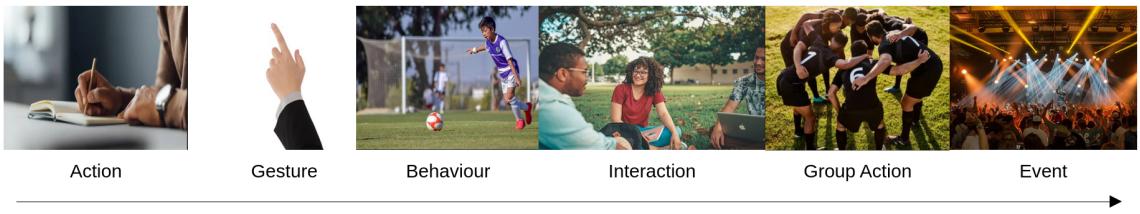

Figure 4. Human action types ranging by complexity, adapted from Beddiar et al. (2020).

- **Gestures or atomic events** constitute the fundamental building blocks of human action. They are characterized by brief durations and typically involve movement of a single body part. Examples include hand waving, nodding, or facial expressions such as smiling. These units are relatively simple to recognize in controlled environments but can become challenging in naturalistic settings due to variations in execution.

- **Behavior** represent coherent, purposeful activities that incorporate multiple gestures into a meaningful sequence. Examples include walking, running, or swimming. Actions typically involve whole-body movement patterns and exhibit greater temporal extent than gestures. The recognition complexity increases due to variations in execution style, speed, and contextual factors.

- **Interactions** introduce a relational dimension, encompassing either human-object or human-human scenarios. These involve coordinated sequences of actions and

gestures directed toward another entity. Examples include opening a door, drinking from a cup, or shaking hands. The recognition challenge expands to include both the actions of the primary subject and their relationship to the secondary entity.

- **Group activities** represent the highest level of complexity, involving multiple people performing coordinated or related actions. Examples include team sports, dance performances, or social gatherings. Recognition at this level requires analyzing multiple interactions simultaneously while accounting for spatial arrangements, temporal synchronization, and social dynamics among participants.

This hierarchical conceptualization provides analytical clarity, but implementation approaches often vary. Modern deep learning systems may implicitly learn these hierarchical relationships without explicit modeling, particularly when trained on large-scale datasets that span multiple levels of the hierarchy.

In practical HAR applications, categorization schemes are often tailored to specific use cases and available sensing modalities rather than adhering strictly to conceptual hierarchies. For instance, mobile phone-based HAR systems typically leverage built-in sensors (accelerometers, gyroscopes, GPS) to categorize activities into application-relevant domains (Incel et al., 2013), such as:

- **Locomotion states**: walking, running, climbing stairs, standing
- **Transportation modes**: driving, riding a bus, cycling, using an elevator
- **Daily activities**: eating, cooking, sleeping, working at a desk
- **Device interactions**: talking on the phone, texting, swiping, tapping

The sensor capabilities fundamentally constrain which action categories can be reliably distinguished. For example, wearable accelerometers excel at detecting rhythmic activi-

ties like walking but struggle with fine-grained manipulations. Vision-based systems offer broader coverage across the action hierarchy but face different challenges related to viewpoint, occlusion, and lighting conditions.

2.1.4 Challenges in Human Action Recognition

Human action recognition (HAR) systems face several categories of challenges that continue to drive research in the field (Chen et al., 2021; Jegham et al., 2020). These challenges can be broadly categorized into technical constraints, and privacy concerns.

Technical challenges arise primarily from environmental factors that affect vision-based systems. Issues such as varying lighting conditions, occlusion, and complex backgrounds can significantly impact system performance. Additionally, the detection and definition of abnormal activities, crucial for healthcare and security applications, presents unique computational and conceptual challenges (Jegham et al., 2020).

Privacy and security concerns have emerged as increasingly critical considerations in HAR system deployment. Despite growing research in privacy-preserving deep learning, these approaches have not been thoroughly explored in the context of HAR systems specifically (Hussain et al., 2020). Privacy considerations in machine learning typically address three key aspects: protecting training data, securing model outputs, and safeguarding the model itself. Recent research suggests that combining differential privacy methods with federated learning approaches may offer promising solutions (Shokri & Shmatikov, 2015).

2.2 Federated Learning

While HAR systems have demonstrated impressive capabilities in controlled environments, their deployment in real-world settings raises important concerns about privacy and data security. This challenge has led researchers to explore federated learning (FL), a dis-

tributed machine learning paradigm that enables model training without centralizing sensitive user data. The intersection of HAR and FL presents an opportunity to develop privacy-preserving systems that can learn from real-world human actions while protecting individual privacy.

Federated learning - also known as collaborative learning, was coined by a Google research paper called ”Communication-Efficient Learning of Deep Networks from Decentralized Data (McMahan et al., 2017). As the authors explain it, model training is done by a loose federation of clients and a server that facilitates communication. A key benefit of federated learning lies in its ability to separate model training from the requirement for direct access to raw training data. This significantly reduces privacy and security risks associated with traditional centralized learning approaches. While some level of trust is still necessary in the server coordinating the training process, the attack surface is minimized. In centralized learning, both the device and the cloud are vulnerable points. Federated learning, on the other hand, restricts potential attacks to the local device as the raw data remains there.

Ideal applications for federated learning are cases in which real-world data collected directly from pervasive environments such as smartphones offers a significant advantage over using proxy data typically available in centralized data centers. This real-world data captures the nuances and variations of actual user behavior. Often the collected data might be privacy-sensitive, containing information that users wouldn’t want shared. Additionally, the data volume might be substantial, making it impractical or even infeasible to transfer it entirely to a central server for training purposes (Ek et al., 2021). Federated learning addresses both concerns by keeping the data on the devices while still enabling collaborative model training. For supervised learning tasks, the labels associated with the data can be inferred directly from user interactions. For example, in an action recognition system, the user’s actions (walking, running, etc.) can be automatically labeled based on sensor readings. This eliminates the need for explicit labeling, which can be time-consuming and potentially error-prone (McMahan et al., 2017).

2.2.1 Federated Learning Workflow

Federated learning operates on a fixed set of K participating clients, each possessing a unique, local dataset. The training process unfolds across iterative rounds. At the commencement of each round, a predetermined fraction (C) of clients is randomly chosen to participate. This selection strategy balances efficiency with performance. While including all clients might appear ideal, experiments by (McMahan et al., 2017) have demonstrated diminishing returns in terms of performance improvement beyond a specific client participation threshold.

Following client selection, the server broadcasts the current global model state (e.g., the most recent model parameters) to all chosen clients. These clients then leverage the received global model state in conjunction with their own local datasets to perform localized computations. Essentially, each selected client trains a localized copy of the global model using its unique data.

Upon completion of local training, each participating client transmits an update (encapsulating the learned information) back to the server. The server then aggregates these updates received from all participating clients and utilizes them to update the global model state. This refined global model state serves as the foundation for the subsequent training round. This iterative process continues, with the global model progressively improving as it incorporates the collective learning gleaned from local datasets distributed across participating devices.

Algorithm 1 Federated Learning Server Process

Require: Number of global epochs E, number of clients K, fraction of clients per round C
Ensure: Global model parameters w_g

Initialize global model parameters w_g
for each global epoch $e = 1, 2, ..., E$ **do**
    $m \leftarrow \max(C \cdot K, 1)$ ▷ Select number of clients
    $S_t \leftarrow$ random set of m clients
    for each client $k \in S_t$ in parallel **do**
        $w_k^e \leftarrow$ ClientUpdate(k, w_g)
    end for
    $w_g \leftarrow \sum_{k=1}^{m} \frac{n_k}{n} w_k^e$ ▷ Weighted average
end for
return w_g

Algorithm 2 ClientUpdate(k, w_g): Federated Learning Client Process

Require: Local epochs E_{local}, batch size B, learning rate η
Require: Local dataset $\mathcal{D}k$
Ensure: Updated model parameters w_k

$w_k \leftarrow w_g$ ▷ Receive global model
for each local epoch $i = 1, 2, ..., E_{local}$ **do**
    for batch $b \in \mathcal{D}_k$ of size B **do**
        $g \leftarrow \nabla \mathcal{L}(w_k; b)$ ▷ Compute gradients
        $w_k \leftarrow w_k - \eta g$ ▷ Update local model
    end for
end for
return w_k to server

2.2.2 Challenges

Federated learning has many shortcomings currently such as non-independent and identically distributed random variables (non-IID), unbalanced datasets, massively distributed client space with average number of examples less than the clients sampled during training, and finally communication limitations with devices being offline or on slow, and communcation costs (McMahan et al., 2017).

From strictly model training point of view non-IID data is the greatest challenge that can impact model training by up to 55% difference in accuracy if training data is highly skewed (Zhao et al., 2018). Skewness can be related to differences in distribution of attributes, labels, or temporal data. What this means in practice is that clients participating in the training have vastly different data distributions, which might make model trained on a client level accurate but generalizing a global model difficult. For example, non-IID data can create significant problems when hospitals contribute patient data from different demographic populations, causing models trained locally at a children's hospital to perform poorly when aggregated with models from geriatric facilities due to the fundamentally different underlying data distributions they've learned from. To address this challenge, researchers have proposed various solutions across three main categories: data-level, algorithm-level, and system-level changes (Zhu et al., 2021).

- **Data-Level** solutions aim to directly modify the underlying data distributions to make them more homogenous. This can involve techniques like data sharing (Federated server having more balanced global dataset that it can distribute to each client along the clients own data) or data augmentation (artificially generating additional data points to address imbalances).
- **Algorithm-level** solutions focus on modifying how local model updates from clients are aggregated to create the global model. These solutions aim to achieve an optimal global model that performs well on all client datasets despite their differences. Examples of such algorithms include FedAvg (Federated Averaging) and Fed-

Per (Federated Learning with Personalization). Moreover, transfer learning can be used which leverages pre-trained models for basic vision tasks and then fine-tune them specifically for the task of HAR. This approach helps the model learn faster and adapt better to diverse datasets.

- **System-level** solutions explore broader system-wide modifications to tackle heterogenous data. Client clustering assumes that there are more than one global model and creates clusters of clients that are found to be similar. Based on which global models can be created that serve the clients the best.

Each solution have their advantages and disadvantages. For example FedPer algorithm was found to work better for heterogenous data compared to more homogenous one where FedAvg outperforms. Data sharing has been found to be effective measure but it can undermine the promise of FL being a privacy preserving method depending how the global dataset for federated server is curated (Zhao et al., 2018). More research is continuously done on the topic which bring novel algorithms for generalizing a better global model (Gao et al., 2022) (Sattler et al., 2019).

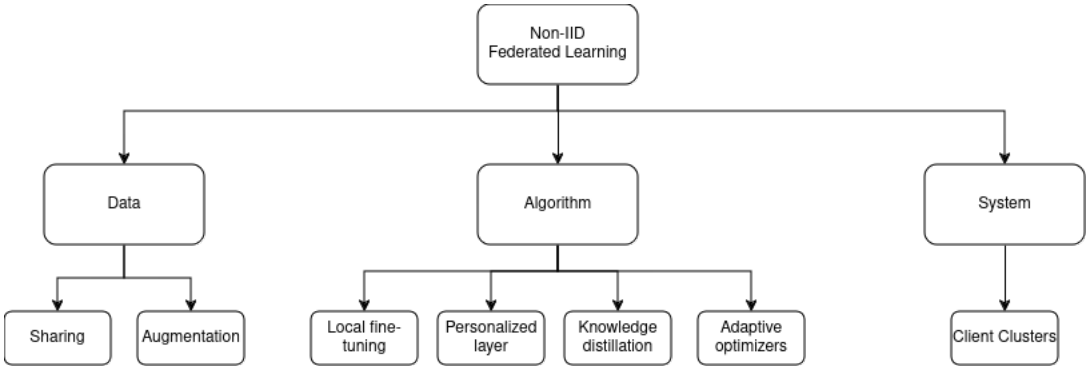

Figure 5. Summary of existing approaches for mitigating non-IID data in federated learning systems.

In the context of increasingly expansive federated learning architectures, communication overhead emerges as a significant constraint on system scalability. Attaining required model fidelity necessitates multiple training iterations, thereby imposing substantial bandwidth utilization when transmitting parameters bidirectionally between the central server and a potentially vast distributed network of edge devices (Mohammadi et al.,

2024). Contemporary advancements in edge computing capabilities have facilitated the deployment of increasingly sophisticated models characterized by parametric complexity several orders of magnitude greater than previous generations. Consequently, the bandwidth requirements for parameter synchronization constitute the predominant performance bottleneck within federated learning ecosystems (Lim et al., 2020), potentially impeding the practical implementation of such systems at global scale.

2.3 Related Work

2.3.1 CheckMATE

CheckMATE (Checking Mutually Average Temporal Encapsulation) is an algorithm to summarize video clips into representative single frames while preserving spatiotemporal information (Fahim & Boutellier, 2024). The algorithm enables the transformation of complex video classification tasks traditionally requiring 3D Convolutional Networks or video transformers into simpler 2D image classification problems. This approach is particularly beneficial for edge devices with limited computational resources, as it may significantly reduce memory requirements and allow for larger batch sizes during training by allowing the use of a 2D convolutional model.

The algorithm operates through an iterative process of temporal encapsulation and feature matching:

Algorithm 3 CheckMATE Video Summarization Loop

Require: Video sequence V with M clips, backbone network B_θ
Ensure: Summary frame V^s (e.g. average complete clip)
1: Segment video V into M clips: $\{v_n | n = 1, ..., M\}$
2: Initialize V^s by averaging random clip frames
3: **for** each iteration **do**
4: Extract temporal encapsulations from clips using B_θ
5: Extract encapsulation from V^s using B_θ
6: Compute similarity loss between clips and V^s
7: Update V^s using gradient descent
8: **end for**
9: **return** V^s

The temporal encapsulation process combines feature embeddings from multiple layers of the backbone network, capturing both high-level semantic information and low-level motion details. This multiscale feature extraction ensures that the final summary frame retains critical spatiotemporal characteristics of the original video sequence while significantly reducing computational overhead during inference.

2.3.2 FedProx: Federated Optimization in Heterogenous Networks

Federated learning presents unique challenges due to its heterogeneous nature, both in terms of systems characteristics (systems heterogeneity) and non-IID data across devices (statistical heterogeneity). Li et al. (2020) proposed FedProx, a generalized version of FedAvg that addresses these challenges.

FedProx has two distinct modifications to FedAvg: (1) it mitigates statistical heterogeneity by restricting local updates from deviating too far from the initial global model by introducing a proximal term, and (2) it allows for variable amounts of work to be performed across devices based on their system capabilities, thus addressing systems heterogeneity. Unlike FedAvg, which often discards stragglers (devices that fail to complete a fixed number of local epochs), FedProx incorporates partial solutions from these devices. Each device can perform different amounts of local computation based on its available resources, measured through a device-specific inexactness parameter γ_k^t. This flexible approach leverages more information from the network while accommodating device-level constraints (Li et al., 2020).

Proximal term used in FedProx to alleviate client drift:

$$h_k(w; w^t) = F_k(w) + \frac{\mu}{2}||w - w^t||^2 \tag{1}$$

The variables in this equation are:

- $h_k(w; w^t)$: The modified objective function for client k that includes the proximal term
- $F_k(w)$: The original loss function for client k
- w: The local model parameters being optimized by client k
- w^t: The global model parameters received from the server at communication round t
- μ: A positive regularization parameter that controls the strength of the proximal term

Empirical evaluations show that FedProx achieves more stable and accurate convergence than FedAvg across various datasets, with particularly significant improvements in highly heterogeneous settings. The algorithm's effectiveness stems from its ability to balance the tension between local and global objectives, especially in non-IID data environments where local models might otherwise diverge from the global optimization goal.

2.3.3 FedYogi: Adaptive Optimization in Federated Learning

FedYogi is an adaptive federated optimization algorithm introduced by Reddi et al. (2021) that addresses the convergence challenges associated with client heterogeneity and communication limitations in federated learning. FedYogi employs adaptive server optimization to improve the training convergence properties of federated learning systems.

Unlike the standard FedAvg algorithm that uses fixed learning rates, FedYogi incorporates an adaptive learning rate mechanism based on the Yogi optimizer Zaheer et al., 2018. This approach makes FedYogi particularly effective in settings with non-IID data across clients, which is a common characteristic in real-world federated learning scenarios.

The key innovation in FedYogi lies in its second-moment accumulator update rule Reddi et al., 2021:

$$v_t = v_{t-1} - (1 - \beta_2)\Delta_t^2 \text{sign}(v_{t-1} - \Delta_t^2) \tag{2}$$

where v_t represents the second-moment estimate, β_2 is a decay parameter, and Δ_t is the average model update from clients at round t. The sign-based correction term helps to address the problem of diminishing step sizes encountered in other adaptive methods like Adam. The server then updates the global model using:

$$x_{t+1} = x_t + \eta \frac{m_t}{\sqrt{v_t + \tau}} \tag{3}$$

where $m_t = \beta_1 m_{t-1} + (1 - \beta_1)\Delta_t$ is the first-moment estimate (incorporating momentum), η is the server learning rate, and τ is a small constant for numerical stability.

Reddi et al. (2021) claim that it is easier to tune than non-adaptive methods, shows robustness to hyperparameter choices, and remains compatible with cross-device federated learning scenarios where clients cannot maintain state between rounds. Empirical evaluations across diverse federated learning tasks show that FedYogi significantly outperforms non-adaptive methods, particularly on tasks with sparse gradients such as natural language processing.

2.3.4 Federated Learning Research

Early work exploring federated learning for HAR focused on evaluating basic FedAvg approaches. Sozinov et al. (2018) conducted one of the first studies to apply FedAvg to HAR using smartphone sensor data, demonstrating slightly worse but acceptable accu-

racy compared to centralized training. However, their work highlighted significant challenges with heterogeneity of the data between users that affect the convergence of the models. The work on communication protocols has since accelerated. Many protocols focus on reducing communication costs between clients and the server (Presotto et al., 2022), (X. Liu et al., 2023) as well as improving performance benchmark datasets such as LEAF (Caldas et al., 2018), but some focus on specific tasks such as prediction of human intention in human-robot assembly tasks (Cai et al., 2024) or defense against malicious poisoning attacks (L. Sun et al., 2024).

To address heterogeneity challenges, researchers have explored adaptations to basic federated learning. Ek et al. (2021) evaluated federated multitask learning approaches to better handle non-IID data distributions in HAR applications. Some studies have proposed clustering models that leverage the similarity between users to enable more effective collaborative learning through clustering, and altering FedAVG to account for data heterogeneity (Gao et al., 2022), (Ouyang et al., 2021), (Presotto et al., 2022).

Research into federated learning healthcare care applications has drastically accelerated after Covid-19 and the European General Data Protection Regulation (GDPR) (Pfitzner et al., 2021), (Rauniyar et al., 2023). FL has been proposed for a variety of tasks such as cancer detection (Jiang et al., 2022), (Jiménez-Sánchez et al., 2023) , Covid-19 and pneumonia detection (Lu et al., 2022) (Díaz & García, 2023), HAR (Zhang et al., 2022), and prediction of autism spectrum disorders (Peng et al., 2022).

Although these works have made important progress, several key limitations remain in current research. Most existing studies focus on simpler tasks, such as medical imaging instead of HAR with limited numbers of activities, users, and data modalities. Evaluation for HAR has largely been conducted using small-scale datasets collected in controlled environments, leaving questions about real-world applicability. There is also a notable gap in research that examines federated learning for more complex HAR tasks utilizing modern deep learning architectures. Especially image-based systems are underrepresented in current research. Although centralized deep learning approaches have shown impressive

results for HAR (Chen et al., 2021), extending these to the federated setting introduces additional challenges that have not been thoroughly explored.

3 Methodology

This chapter describes the systematic approach used to investigate federated learning for privacy-preserving human action recognition on edge devices. The research methodology is designed to address the previously described research questions through a comparative study of centralized and federated learning approaches.

The investigation consists of several key components. First, we establish a baseline using traditional centralized training on the UCF101 and HMDB51 datasets. These widely used action recognition datasets provide a robust foundation for evaluating model performance (“Papers with Code - Machine Learning Datasets”, 2025). Next, we implemented a federated learning system for CheckMATE using the Flower framework, distributing the training process across multiple simulated edge devices while maintaining data privacy. Throughout both approaches, we employ the CheckMATE algorithm to optimize video processing.

The methodology is designed to systematically evaluate and compare both training approaches while ensuring practical applicability on resource-constrained devices. The following sections detail each component of this research framework, beginning with dataset preparation and proceeding through the implementation of both training approaches.

3.1 Dataset

3.1.1 Dataset Selection

UCF101 and HMDB51 are two prominent datasets commonly employed in the domain of human action recognition research. UCF101, with its extensive collection of 101 action classes and diverse video sequences, presents a robust benchmark for evaluating the performance of action recognition models. In contrast, HMDB51, encompassing a broader spectrum of actions captured from various sources, is well suited for investigating the

robustness of models to different video conditions. Both datasets, characterized by their complexity and diversity, serve as valuable resources to compare the results of the study with other works in the field.

HMDB51 consists of 51 distinct action categories. In total, there are 6,766 video clips with each class containing at least 101 clips. The clips have been sourced from youtube, google videos, digitized movies, and each clips quality and annotation has been verified by at least two reviewers. The selected clips represent various lighting conditions, camera angles, camera types, camera movement, and more criteria to ensure diverse high-quality data set (Kuehne et al., 2011).

UCF101 consists of 13,320 video clips with 101 categories which can be further classified within 5 groups of actions. In total there is 27 hours of video data sourced from Youtube. The clip lengths vary from 1.06 seconds to 71.04 seconds with constant 25 frames per second framerate (Soomro et al., 2012). The authors note that the clips are varied with different camera motions, partial occlusions, lighting conditions, blurriness and low quality frames at times.

Other dataset considerations were something-something V2 (Goyal et al., 2017), Kinetics 400 and 600 (Zisserman et al., 2017) (Carreira et al., 2018). All the aforementioned datasets are newer and larger than UCF101 and HMDB51 but this also pose more demanding requirements in terms of storage, and computing power to prepare and train within adequate timespan. Kinetics 400 for example is over 400GB in size, Kinetics 600 being even larger. As federated learning is meant to be performed on edge devices these datasets would defeat the purpose by being too large. Fortunately, UCF101 and HMDB51 are appropriate sized and still relevant as Papers with Code – website featuring latest research and datasets in machine learning, shows that in 2024 UCF101 was featured in 225 papers and HMDB51 in 63 as of writing this (“Papers with Code - Machine Learning Datasets”, 2025). Kinetics 400 was featured in 105. Kinetics 600 and something-something v2 even less than this.

3.1.2 Dataset Preparation

UCF101 and HMDB51 datasets are distilled using CheckMATE. CheckMATE's description and implementation can be found on Github and read from the CheckMATE paper.

UCF101 comes with three predetermined train and test splits that were used to evaluate the 3-fold accuracy. With HMDB51, every fifth clip is chosen for testing and the rest are for training. HMDB51 also offers three predetermined splits by the authors, but CheckMATE follows every fifth sample testing split. To have fair comparison with CheckMATE's results, same split approach was taken.

3.1.3 Video Distillation Process

The video distillation process differs slightly between the UCF101 and HMDB51 datasets to accommodate their distinct characteristics. For UCF101, each video is processed by sampling eight frames with a frame step of size six to cover greater area of the clip. Sampled frames are normalized according to ImageNet dataset mean and standard deviation (Russakovsky et al., 2015) and then passed through the ConvNext backbone model. Backbone model output latents are compared with the latents of initial averaged frame of the clip. Initial averaged frame is then adjusted based on the mean squared error loss between the latents. In the case of UCF101, 15 samples of frames are taken and used to adjust the initial averaged frame three times per sample resulting in total 45 adjustments. The HMDB51 dataset follows a similar but modified approach, where 16 frames are sampled instead of eight. However, the optimization process differs, with averaged frames being compared to a single random clip and the optimization loop executing twice per sampled clip to avoid overfitting.

Table 2. CheckMATE Video Distillation Configuration.

Parameter	**UCF101**	**HMDB51**
Frame sampling	8 frames with step size 6	16 frames with step size 6
Frame normalization	ImageNet mean and std	
CheckMATE Parameters	15 random clips, 3 optimizations per clip	1 random clip, 2 optimizations per clip
Total frame optimizations	45	2
Rationale	Greater coverage needed for diverse actions	Limited adjustments to prevent overfitting

The data augmentation pipeline includes several transformations to enhance the robustness of the distillation and is applied 50% of the time:

- Random cropping of frames
- Horizontal flipping
- Brightness adjustments
- Saturation modifications
- Resolution scaling ($224 \rightarrow 32 \rightarrow 224$)

3.1.4 Training Methodology

The training phase employs custom classification heads specifically designed for both UCF101 and HMDB51 datasets. A key feature of the training methodology is the creation of four distinct versions of each distilled image to improve classification accuracy:

1. Original distilled image
2. Random flipped variant
3. Random rotated variant
4. 3D dropout variant

These four versions are concatenated and processed through the model to generate latent representations. The final prediction is obtained by averaging these latent representations in the custom classification head. This approach helps in capturing different perspectives of the same action and improves the model's robustness to variations in the input. The drawback of using heavy augmentation is the memory required to train the model. The reasoning for the use of this heavy data augmentation was to reproduce the results of CheckMATE as closely as possible.

The training process utilizes the Adam optimizer with a learning rate of $1e^{-4}$. Each model is trained for 20 epochs with a batch size of 16. The complete implementation, including the specific architecture of the custom heads and detailed data processing pipelines, is available in the project's GitHub repository.

3.2 Federated Learning Architecture

Federated learning is orchestrated using Flower framework (Beutel et al., 2022). Flower is a Python library created for building federated learning systems. Flower offers a highly customizable, machine learning framework agnostic, easily understandable baseline to start building systems for federated learning. Because the framework is machine learning framework agnostic, it only functions as a means to communicate between the server and clients. Flower simulations also offer quicker prototyping possibilities by mimicking federated learning system training with parallellized virtual clients with computing cluster resources. The federated learning implementation of this work can be found on Github [1].

During training, Flower tracks centralized and distributed metrics such as loss and the accuracy. For the distributed case, the metrics are aggregated by weighted averaging between each client. The test dataset exists on server side (centralized) and each client have their own smaller validation sets for tracking the distributed metrics. The separation

[1] https://github.com/LeeviEnontekio/CheckMATE-FL

of centralized and distributed metrics offers more holistic understanding of the model convergence globally and locally for each client.

The experiments are carried out with 8 clients of which all are used for fitting and evaluating the global model. The clients receive partitioned training sets from which 20% is allocated for local model evaluation. Training data is partitioned homogenously with random splits for all clients when training is commenced. Due to independent and identically distributed data between the clients, federated averaging (FedAvg) is chosen for global model aggregation strategy with Flower as baseline. FedAvg performs weighted averaging operation to aggregate each local models' parameters after local fit round to be used for the global model (McMahan et al., 2017).

3.3 Model Architecture

Based on CheckMATE's proposed method, ImageNet-1K pre-trained ConvNext-small is used as a backbone. ConvNext is a modernized ResNet model which takes inspiration from transfomer architectures (Z. Liu et al., 2022). The small variant of ConvNext offers good performance with 50 million parameters which is adequate for many edge devices with strict memory constraints. The classification model uses the same backbone, as the summary frames during distillation process are adjusted based on temporal encapsulations derived from the backbone. The specific layers chosen for feature extraction as well as custom classification head for the backbone can be found on Github.

ConvNeXt is a pure convolutional neural network architecture that demonstrates how modernizing traditional ConvNet design principles can achieve performance competitive with state-of-the-art Vision Transformers. ConvNeXt systematically adapts several key design elements of Vision Transformers while maintaining the simplicity and efficiency of standard ConvNets.

The architecture incorporates several key modifications to the traditional ResNet design:

Table 3. ConvNext small architecture specifications.

Stage	Output Size	Blocks	Channels
Stem	56×56	1	96
Stage 1	56×56	3	96
Stage 2	28×28	3	192
Stage 3	14×14	9	384
Stage 4	7×7	3	768

- Macro design: Adopts the "patchify" stem layer using non-overlapping convolutions and adjusts the stage compute ratio to follow Vision Transformer patterns
- Micro design: Replaces ReLU with GELU activation, reduces normalization layers, and substitutes BatchNorm with LayerNorm
- Block modifications: Uses depthwise convolution with increased kernel sizes (up to 7×7), inverted bottleneck design, and fewer activation functions per block

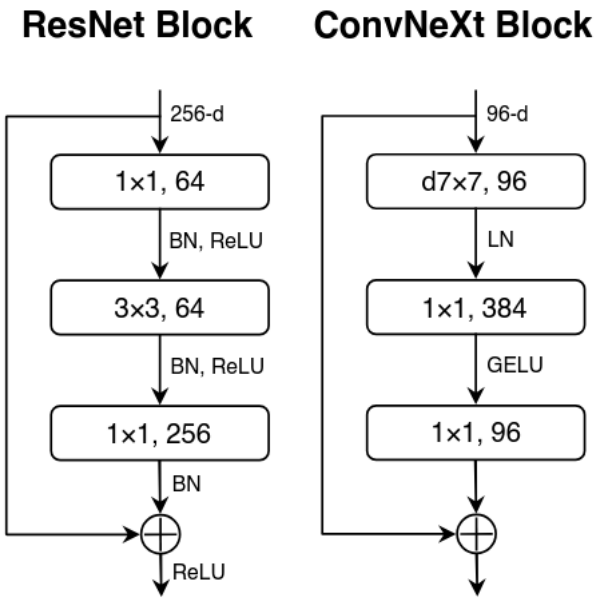

Figure 6. The ConvNeXt block design compared to ResNet block. Unlike Transformer-based designs, it maintains pure convolutional operations while achieving competitive performance.

These architectural changes enable ConvNeXt to achieve performance comparable to or superior to Swin Transformers across multiple computer vision tasks (Z. Liu et al.,

2022). The tiny model (ConvNeXt-T) achieves 82.1% accuracy on ImageNet-1K classification, while larger variants demonstrate strong scaling behavior, with ConvNeXt-XL reaching 87.8% accuracy when pre-trained on ImageNet-22K. Importantly, Z. Liu et al., 2022 note that ConvNeXt maintains higher throughput than equivalent Transformer models while achieving these results through standard convolutional operations.

3.4 Metrics

To quantitatively assess the performance of a human action recognition model, two evaluation metrics are used for monitoring training and evaluation. The metrics are carefully chosen to provide comprehensive insights into both the model's learning process and its final performance.

3.4.1 Loss Function

Two distinct loss measurements are utilized to monitor the training process, server-side and distributed loss. Cross-entropy loss is used as a loss function (Bishop & Bishop, 2023) as the model is performing classification task on video dataset.

$$-\frac{1}{N}\Sigma_{i=1}^{N}(y_i.log(p_i) + (1-y_i)log(1-p_i)) \tag{4}$$

1. *Server side Loss*: This metric quantifies the model's performance on server-side testing dataset, providing immediate feedback about the learning process after a round of training on distributed clients.

2. *Distributed Loss*: Given the distributed nature of the training setup, we measure the aggregated loss across all training clients. Distributed loss offers a holistic view of the model's convergence on clients. The distributed loss is a weighted average between all clients who participate in a training round.

3.4.2 Classification Accuracy

The primary metric for evaluating the model's performance is classification accuracy, which measures the proportion of correctly classified action sequences in the test set.

We compute accuracy using:

$$Accuracy = \frac{(Number\,of\,Correct\,Predictions)}{(Total\,Number\,of\,Predictions)} \cdot 100 \tag{5}$$

This metric is particularly suitable for balanced datasets and provides an intuitive measure of the model's performance. Accuracy scores are reported separately for both UCF101 and HMDB51 datasets to ensure comprehensive evaluation across different action recognition scenarios.

For UCF101, which contains 101 action categories, the accuracy metric effectively captures the model's ability to distinguish between a wide range of human actions. Similarly, for HMDB51, with its 51 action categories and emphasis on natural videos, the accuracy metric helps assess the model's robustness to real-world variations in human motion patterns.

To ensure robust evaluation, the standard protocol for evalution is followed by using all three train/test splits for UCF101 and reporting the average accuracy across splits. This approach helps mitigate any potential bias that might arise from a single split configuration.

3.5 Hardware

The experiments were conducted using a single NVIDIA RTX 4090 GPU with 24GB of VRAM. The system ran on Ubuntu 22.04 operating system. Both centralized and federated learning training setups were simulated on the same machine, with the federated learning environment being virtualized using Flower framework to simulate multiple clients while utilizing the single GPU's resources. This hardware configuration provided sufficient computational capacity for training the ConvNeXt model and processing the video datasets used in the experiments.

4 Results and Analysis

This chapter presents the experimental results comparing centralized and federated learning approaches for human action recognition, with particular focus on the effectiveness of CheckMATE in a distributed learning context.

4.1 Comparison of Federated Learning Strategies

Table 4. Classification accuracy (%) on action recognition datasets.

Dataset	Centralized	FedAVG	FedProx	FedYogi
UCF101	80.7%	79.9%	80.1%	80.5%
HMDB51	66.8%	58.4%	59.7%	59.3%

Table 3 presents the classification accuracy results for both the UCF101 and HMDB51 data sets in different training approaches. The centralized CheckMATE method establishes our baseline performance, achieving 80.7% accuracy on UCF101 and 66.8% on HMDB51.

When examining federated learning methods, we observe that all three approaches (FedAVG, FedProx, and FedYogi) maintain competitive performance in the UCF101 dataset. FedYogi demonstrates the strongest performance among the federated approaches with 80.5% accuracy, showing only a marginal decrease of 0.2 percentage points compared to the centralized baseline. This suggests that for datasets with relatively consistent action representations such as UCF101, federated learning can achieve nearly equivalent performance while preserving data privacy.

The results on HMDB51 show a different pattern, with a more pronounced gap between the centralized and federated approaches. The best performing federated method on this dataset is FedProx at 59.7%, which still represents a 7.1 percentage point decrease from the centralized baseline. This performance difference likely stems from HMDB51's greater variability in video quality, camera angles, and action executions, making it more challenging for federated models to generalize effectively across distributed data.

4.2 Analysis & Implications

The performance disparity between datasets highlights an important consideration for privacy-preserving HAR systems: while federated learning approaches can achieve near-centralized performance on well-structured datasets, they may require additional optimization techniques when dealing with more diverse, real-world action representations. However, the relatively small performance trade-off, particularly with FedYogi on UCF101, demonstrates the viability of federated learning for privacy-preserving human action recognition. Furthermore, the results demonstrate the effectiveness of CheckMATE in a federated learning environment.

For UCF101, FedYogi adaptive optimization offers highest accuracy, but it must noted that all strategies are within 0.6% accuracy. Due to the homogenous nature of UCF101 it could be argued that the stategies do not differentiate themselves from each other much as FedYogi and FedProx are meant to alleviate problems with heterogenous, non-IID data which UCF101 is a poor example of. Regardless, FL with CheckMATE demonstrates competitive performance.

The more substantial performance gap observed on HMDB51 (7.1% between centralized and the best federated approach) reflects the dataset's greater complexity and variability, presenting a more challenging scenario for federated optimization. This observation aligns with findings in prior research that client heterogeneity can significantly impact federated learning convergence (Karimireddy et al., 2020). FedProx's regularization approach, which explicitly constrains local updates to remain close to the global model (Li et al., 2020), may be more effective for highly heterogeneous data distributions. The proximal term in FedProx appears to mitigate the effects of client drift when learning from diverse action representations, providing more stable convergence than purely adaptive methods in such scenarios.

These findings have several important implications for designing privacy-preserving HAR systems:

First, the selection of federated optimization strategy should be data-dependent. For applications involving relatively uniform action distributions across users (similar to UCF101), adaptive methods like FedYogi may offer the performance. Conversely, applications with highly personalized or variable action patterns may benefit from proximal regularization techniques like FedProx.

Second, the reduced performance on HMDB51 indicates that federated HAR systems may require supplementary techniques to close the gap with centralized approaches on complex datasets. Potential strategies could include personalization layers (Arivazhagan et al., 2019) that can adapt to user-specific action patterns while maintaining a strong shared representation.

Third, our results validate CheckMATE's architecture amenable to federated learning, with its spatio-temporal modeling proving robust even when trained on distributed data. The multi-stage architecture allows effective extraction of spatial-temporal features across clients, with the performance gap primarily emerging from the challenges in aggregating diverse action representations rather than architectural limitations.

Finally, the trade-off between privacy preservation and recognition accuracy appears manageable in practical scenarios. Even the largest performance gap (7.1% on HMDB51) may be acceptable in many privacy-sensitive applications where protecting user data is a primary concern. This suggests that federated HAR systems based on CheckMATE can be deployed in real-world settings where both privacy and recognition performance are valued, particularly if appropriate federated optimization strategies are selected based on the expected data distribution characteristics.

5 Conclusions and Future Work

Our experimental results demonstrate that federated learning represents a viable approach for developing privacy-preserving HAR systems, though with varying degrees of performance across different datasets and federated optimization strategies. On the UCF101 dataset, which contains relatively consistent action representations, federated approaches achieved performance nearly equivalent to centralized training. FedYogi demonstrated the strongest results with 80.5% accuracy, showing only a marginal decrease of 0.2 percentage points compared to the centralized baseline of 80.7%. This minimal performance gap suggests that for well-structured HAR datasets, federated learning can effectively preserve data privacy without significant accuracy compromise.

However, the more challenging HMDB51 dataset revealed a more pronounced difference between centralized and federated approaches. The best-performing federated method on this dataset was FedProx at 59.7%, representing a 7.1 percentage point decrease from the centralized baseline of 66.8%. This larger performance gap highlights how heterogeneity in action representations, camera angles, and video quality—characteristics more prevalent in HMDB51—poses greater challenges for federated optimization.

5.1 Implications

The findings from this study have several important implications for the development and deployment of privacy-preserving HAR systems: First, the selection of federated optimization strategy should be tailored to the expected data distribution characteristics. Applications with relatively uniform action patterns across users may benefit from adaptive methods like FedYogi, while applications involving more diverse or personalized action patterns might achieve better results with regularization-based approaches like FedProx.

Second, the acceptable performance tradeoff observed, even in the most challenging scenario (7.1% on HMDB51), suggests that federated HAR systems can be practically deployed

in privacy-sensitive contexts where protecting user data is a primary concern. This trade-off represents a reasonable compromise between privacy preservation and recognition accuracy for many real-world applications.

Third, the successful integration of CheckMATE with federated learning demonstrates a promising path toward computationally efficient, privacy-preserving HAR systems suitable for edge deployment. By reducing video processing requirements while maintaining acceptable performance, this approach addresses both privacy and resource constraints simultaneously.

5.2 Addressing Research Questions

1. **How does a distributedly trained model perform on camera-based human action recognition tasks compared to traditionally trained models?**

   Our experiments demonstrate that distributedly trained models can approach the performance of traditionally trained models, with the gap varying based on dataset characteristics. For datasets with more uniform action distributions (like UCF101), the performance difference is minimal (less than 1%). For more heterogeneous datasets (like HMDB51), the gap widens but remains within an acceptable range for many practical applications (approximately 7%).

2. **How does distilling datasets with the CheckMATE procedure affect federated learning and final model performance?**

   The CheckMATE procedure has proven highly compatible with federated learning, effectively reducing the computational demands of video processing while preserving essential spatiotemporal information. This compatibility is evident in the competitive performance achieved by federated models trained on CheckMATE-distilled frames. Transformation from video sequences to representative frames enables efficient model training on edge devices with limited resources without substantial compromise to recognition accuracy.

In conclusion, this research demonstrates that federated learning, combined with efficient video processing techniques like CheckMATE, offers a promising approach to privacy-preserving human action recognition on edge devices. While performance tradeoffs exist, particularly for more heterogeneous datasets, the gap between centralized and federated approaches is manageable for many practical applications. Future work focused on model efficiency, real-world deployment, and optimization for heterogeneous data distributions will further enhance the viability and effectiveness of privacy-preserving HAR systems.

6 Bibliography

References

Arivazhagan, M. G., Aggarwal, V., Singh, A. K., & Choudhary, S. (2019). Federated learning with personalization layers. *arXiv preprint arXiv:1912.00818*.

Beddiar, D. R., Nini, B., Sabokrou, M., & Hadid, A. (2020). Vision-based human activity recognition: A survey. *Multimedia Tools and Applications*, *79*(41), 30509–30555. https://doi.org/10.1007/s11042-020-09004-3

Beutel, D. J., Topal, T., Mathur, A., Qiu, X., Fernandez-Marques, J., Gao, Y., Sani, L., Li, K. H., Parcollet, T., de Gusmão, P. P. B., & Lane, N. D. (2022, March). Flower: A Friendly Federated Learning Research Framework [arXiv:2007.14390 [cs, stat]]. https://doi.org/10.48550/arXiv.2007.14390

Bishop, C. M., & Bishop, H. (2023). *Deep learning: Foundations and concepts*. Springer Nature.

Cai, J., Gao, Z., Guo, Y., Wibranek, B., & Li, S. (2024). Fedhip: Federated learning for privacy-preserving human intention prediction in human-robot collaborative assembly tasks [ISBN: 1474-0346 Publisher: Elsevier]. *Advanced Engineering Informatics*, *60*, 102411.

Caldas, S., Duddu, S. M. K., Wu, P., Li, T., Konečný, J., McMahan, H. B., Smith, V., & Talwalkar, A. (2018). Leaf: A benchmark for federated settings. *arXiv preprint arXiv:1812.01097*.

Carreira, J., Noland, E., Banki-Horvath, A., Hillier, C., & Zisserman, A. (2018). A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*.

Chen, K., Zhang, D., Yao, L., Guo, B., Yu, Z., & Liu, Y. (2021). Deep learning for sensor-based human activity recognition: Overview, challenges, and opportunities [ISBN: 0360-0300 Publisher: ACM New York, NY, USA]. *ACM Computing Surveys (CSUR)*, *54*(4), 1–40.

Díaz, J. S.-P., & García, Á. L. (2023). Study of the performance and scalability of federated learning for medical imaging with intermittent clients [ISBN: 0925-2312 Publisher: Elsevier]. *Neurocomputing*, *518*, 142–154.

Ek, S., Portet, F., Lalanda, P., & Vega Baez, G. E. (2021). Evaluating Federated Learning for human activity recognition. *Workshop AI for Internet of Things, in conjunction with IJCAI-PRICAI 2020*. https://hal.science/hal-03102880

Fahim, M. A.-N. I., & Boutellier, J. (2024). CheckMATE: Efficient Video Summarization by Checking Mutually Averaged Temporal Encapsulation, 8343–8348. Retrieved April 3, 2025, from https://openaccess.thecvf.com/content/CVPR2024W/DDCV/html/Fahim_CheckMATE_Efficient_Video_Summarization_by_Checking_Mutually_Averaged_Temporal_Encapsulation_CVPRW_2024_paper.html

Gao, L., Fu, H., Li, L., Chen, Y., Xu, M., & Xu, C.-Z. (2022). Feddc: Federated learning with non-iid data via local drift decoupling and correction. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10112–10121.

Goyal, R., Ebrahimi Kahou, S., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., & Mueller-Freitag, M. (2017). The" something something" video database for learning and evaluating visual common sense. *Proceedings of the IEEE international conference on computer vision*, 5842–5850.

Gu, F., Chung, M.-H., Chignell, M., Valaee, S., Zhou, B., & Liu, X. (2021). A survey on deep learning for human activity recognition [ISBN: 0360-0300 Publisher: ACM New York, NY]. *ACM Computing Surveys (CSUR)*, *54*(8), 1–34.

Han, P., Du, J., Zhou, J., & Zhu, S. (2013). An Object Detection Method Using Wavelet Optical Flow and Hybrid Linear-Nonlinear Classifier [Publisher: John Wiley & Sons, Ltd]. *Mathematical Problems in Engineering*, *2013*(1), 965419. Retrieved April 24, 2025, from https://onlinelibrary.wiley.com/doi/abs/10.1155/2013/965419

Haria, A., Subramanian, A., Asokkumar, N., Poddar, S., & Nayak, J. S. (2017). Hand gesture recognition for human computer interaction [ISBN: 1877-0509 Publisher: Elsevier]. *Procedia computer science*, *115*, 367–374.

Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., & Vasudevan, V. (2019). Searching for mobilenetv3. *Proceedings of the IEEE/CVF international conference on computer vision*, 1314–1324.

Hussain, Z., Sheng, M., & Zhang, W. E. (2020). Different Approaches for Human Activity Recognition: A Survey [arXiv:1906.05074 [cs]]. *Journal of Network and Computer Applications*, *167*, 102738. https://doi.org/10.1016/j.jnca.2020.102738

Incel, O. D., Kose, M., & Ersoy, C. (2013). A review and taxonomy of activity recognition on mobile phones [ISBN: 2191-1630 Publisher: Springer]. *BioNanoScience*, *3*(2), 145–171.

Jegham, I., Ben Khalifa, A., Alouani, I., & Mahjoub, M. A. (2020). Vision-based human action recognition: An overview and real world challenges. *Forensic Science International: Digital Investigation*, *32*, 200901. https://doi.org/10.1016/j.fsidi.2019.200901

Jiang, M., Wang, Z., & Dou, Q. (2022). Harmofl: Harmonizing local and global drifts in federated learning on heterogeneous medical images [Issue: 1]. *Proceedings of the AAAI Conference on Artificial Intelligence*, *36*, 1087–1095.

Jiménez-Sánchez, A., Tardy, M., Ballester, M. A. G., Mateus, D., & Piella, G. (2023). Memory-aware curriculum federated learning for breast cancer classification [ISBN: 0169-2607 Publisher: Elsevier]. *Computer Methods and Programs in Biomedicine*, *229*, 107318.

Johnson, J. (2022, April). Videos. Retrieved February 28, 2025, from https://web.eecs.umich.edu/~justincj/slides/eecs498/WI2022/598_WI2022_lecture24.pdf

Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S., Stich, S., & Suresh, A. T. (2020). Scaffold: Stochastic controlled averaging for federated learning. *International conference on machine learning*, 5132–5143.

Knowles, B. (2016). Emerging trust implications of data-rich systems [ISBN: 1536-1268 Publisher: IEEE]. *IEEE Pervasive Computing*, *15*(4), 76–84.

Kolesnikov, A., Dosovitskiy, A., Weissenborn, D., Heigold, G., Uszkoreit, J., Beyer, L., Minderer, M., Dehghani, M., Houlsby, N., Gelly, S., Unterthiner, T., & Zhai, X. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.

Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., & Serre, T. (2011). HMDB: A Large Video Database for Human Motion Recognition.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning [ISBN: 0028-0836 Publisher: Nature Publishing Group UK London]. *nature*, *521*(7553), 436–444.

Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., & Smith, V. (2020). Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, *2*, 429–450.

Liao, S., Li, G., Wu, H., Jiang, D., Liu, Y., Yun, J., Liu, Y., & Zhou, D. (2021). Occlusion gesture recognition based on improved SSD [ISBN: 1532-0626 Publisher: Wiley Online Library]. *Concurrency and Computation: Practice and Experience*, *33*(6), e6063.

Lim, W. Y. B., Luong, N. C., Hoang, D. T., Jiao, Y., Liang, Y.-C., Yang, Q., Niyato, D., & Miao, C. (2020). Federated learning in mobile edge networks: A comprehensive survey [ISBN: 1553-877X Publisher: IEEE]. *IEEE communications surveys & tutorials*, *22*(3), 2031–2063.

Liu, X., Deng, Y., Nallanathan, A., & Bennis, M. (2023). Federated learning and meta learning: Approaches, applications, and directions [ISBN: 1553-877X Publisher: IEEE]. *IEEE Communications Surveys & Tutorials*, *26*(1), 571–618.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A convnet for the 2020s. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11976–11986.

Lu, W., Wang, J., Chen, Y., Qin, X., Xu, R., Dimitriadis, D., & Qin, T. (2022). Personalized federated learning with adaptive batchnorm for healthcare [ISBN: 2332-7790 Publisher: IEEE]. *IEEE Transactions on Big Data*.

Mammen, P. M. (2021). Federated learning: Opportunities and challenges. *arXiv preprint arXiv:2101.05428*.

McMahan, B., Moore, E., Ramage, D., Hampson, S., & Arcas, B. A. y. (2017, April). Communication-Efficient Learning of Deep Networks from Decentralized Data. In A. Singh & J. Zhu (Eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics* (pp. 1273–1282, Vol. 54). PMLR. https://proceedings.mlr.press/v54/mcmahan17a.html

Minh Dang, L., Min, K., Wang, H., Jalil Piran, M., Hee Lee, C., & Moon, H. (2020). Sensor-based and vision-based human activity recognition: A comprehensive survey. *Pattern Recognition*, *108*, 107561. https://doi.org/10.1016/j.patcog.2020.107561

Mohammadi, S., Balador, A., Sinaei, S., & Flammini, F. (2024). Balancing privacy and performance in federated learning: A systematic literature review on methods and metrics [ISBN: 0743-7315 Publisher: Elsevier]. *Journal of Parallel and Distributed Computing*, 104918.

Ouyang, X., Xie, Z., Zhou, J., Huang, J., & Xing, G. (2021). Clusterfl: A similarity-aware federated learning system for human activity recognition. *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*, 54–66.

Papers with Code - Machine Learning Datasets. (2025). Retrieved March 1, 2025, from https://paperswithcode.com/datasets

Peng, L., Wang, N., Dvornek, N., Zhu, X., & Li, X. (2022). Fedni: Federated graph learning with network inpainting for population-based disease prediction [ISBN: 0278-0062 Publisher: IEEE]. *IEEE Transactions on Medical Imaging*, *42*(7), 2032–2043.

Pfitzner, B., Steckhan, N., & Arnrich, B. (2021). Federated learning in a medical context: A systematic literature review [ISBN: 1533-5399 Publisher: ACM New York, NY]. *ACM Transactions on Internet Technology (TOIT)*, *21*(2), 1–31.

Presotto, R., Civitarese, G., & Bettini, C. (2022). Fedclar: Federated clustering for personalized sensor-based human activity recognition. *2022 IEEE international conference on pervasive computing and communications (PerCom)*, 227–236.

Ramanujam, E., Perumal, T., & Padmavathi, S. (2021). Human activity recognition with smartphone and wearable sensors using deep learning techniques: A review [ISBN: 1530-437X Publisher: IEEE]. *IEEE Sensors Journal*, *21*(12), 13029–13040.

Rauniyar, A., Hagos, D. H., Jha, D., Håkegård, J. E., Bagci, U., Rawat, D. B., & Vlassov, V. (2023). Federated learning for medical applications: A taxonomy, current trends, challenges, and future research directions [ISBN: 2327-4662 Publisher: IEEE]. *IEEE Internet of Things Journal*, *11*(5), 7374–7398.

Rautaray, S. S., & Agrawal, A. (2015). Vision based hand gesture recognition for human computer interaction: A survey [ISBN: 0269-2821 Publisher: Springer]. *Artificial intelligence review*, *43*, 1–54.

Reddi, S., Charles, Z. B., Zaheer, M., Garrett, Z., Rush, K., Konečný, J., Kumar, S., & McMahan, B. (Eds.). (2021). *Adaptive Federated Optimization* [Publication Title: International Conference on Learning Representations]. https://openreview.net/forum?id=LkFG3lB13U5

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet Large Scale Vi-

sual Recognition Challenge. *International Journal of Computer Vision*, *115*(3), 211–252. https://doi.org/10.1007/s11263-015-0816-y

Sattler, F., Wiedemann, S., Müller, K.-R., & Samek, W. (2019). Robust and communication-efficient federated learning from non-iid data [ISBN: 2162-237X Publisher: IEEE]. *IEEE transactions on neural networks and learning systems*, *31*(9), 3400–3413.

Shao, J., Kang, K., Change Loy, C., & Wang, X. (2015). Deeply learned attributes for crowded scene understanding. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4657–4666.

Shokri, R., & Shmatikov, V. (2015). Privacy-preserving deep learning. *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, 1310–1321.

Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, *27*.

Soomro, K., Zamir, A. R., & Shah, M. (2012). UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild [arXiv: 1212.0402]. *CoRR*, *abs/1212.0402*. http://arxiv.org/abs/1212.0402

Sozinov, K., Vlassov, V., & Girdzijauskas, S. (2018). Human Activity Recognition Using Federated Learning. *2018 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Ubiquitous Computing & Communications, Big Data & Cloud Computing, Social Computing & Networking, Sustainable Computing & Communications (ISPA/IUCC/BDCloud/SocialCom/SustainCom)*, 1103–1111. https://doi.org/10.1109/BDCloud.2018.00164

Sun, L., Tian, J., & Muhammad, G. (2024). FedKC: Personalized federated learning with robustness against model poisoning attacks in the metaverse for consumer health [ISBN: 0098-3063 Publisher: IEEE]. *IEEE Transactions on Consumer Electronics*.

Sun, Z., Ke, Q., Rahmani, H., Bennamoun, M., Wang, G., & Liu, J. (2022). Human Action Recognition from Various Data Modalities: A Review [arXiv:2012.11866 [cs]]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–20. https://doi.org/10.1109/TPAMI.2022.3183112

Wang, L., Huang, B., Zhao, Z., Tong, Z., He, Y., Wang, Y., Wang, Y., & Qiao, Y. (2023). Videomae v2: Scaling video masked autoencoders with dual masking. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14549–14560.

Wang, X., Wu, Y., Zhu, L., & Yang, Y. (2020). Symbiotic attention with privileged information for egocentric action recognition [Issue: 07]. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*, 12249–12256.

Wang, Y., Li, K., Li, X., Yu, J., He, Y., Chen, G., Pei, B., Zheng, R., Wang, Z., & Shi, Y. (2024). Internvideo2: Scaling foundation models for multimodal video understanding. *European Conference on Computer Vision*, 396–416.

Wensel, J., Ullah, H., & Munir, A. (2023). Vit-ret: Vision and recurrent transformer neural networks for human activity recognition in videos [ISBN: 2169-3536 Publisher: IEEE]. *IEEE Access*, *11*, 72227–72249.

Yu, S., Carroll, F., & Bentley, B. L. (2024). Trust and Trustworthiness: Privacy Protection in the ChatGPT Era. In *Data Protection: The Wake of AI and Machine Learning* (pp. 103–127). Springer.

Zaheer, M., Reddi, S., Sachan, D., Kale, S., & Kumar, S. (2018). Adaptive methods for nonconvex optimization. *Advances in neural information processing systems*, *31*.

Zhang, K., Liu, X., Xie, X., Zhang, J., Niu, B., & Li, K. (2022). A cross-domain federated learning framework for wireless human sensing [ISBN: 0890-8044 Publisher: IEEE]. *IEEE Network*, *36*(5), 122–128.

Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D., & Chandra, V. (2018). Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*.

Zhou, X., Liang, W., Kevin, I., Wang, K., Wang, H., Yang, L. T., & Jin, Q. (2020). Deep-learning-enhanced human activity recognition for internet of healthcare things [ISBN: 2327-4662 Publisher: IEEE]. *IEEE Internet of Things Journal*, *7*(7), 6429–6438.

Zhu, H., Xu, J., Liu, S., & Jin, Y. (2021). Federated learning on non-IID data: A survey. *Neurocomputing*, *465*, 371–390. https://doi.org/10.1016/j.neucom.2021.07.098

Zisserman, A., Carreira, J., Simonyan, K., Kay, W., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., & Suleyman, M. (2017). The Kinetics Human Action Video Dataset.