Human-Computer Interaction & Emerging Technologies, Vol. 233, 2026, 317–324
https://doi.org/10.54941/ahfe1007531

A Framework for Lightweight, Edge-Based Recognition of Dynamic American Sign Language Using Temporal Learning Models

Owasu Brown and Amir Schur

National University, San Diego, CA 92123, USA

ABSTRACT

Automated recognition of dynamic American Sign Language (ASL) gestures remains a significant challenge for real-time deployment on resource-constrained edge devices. Although recent advances in deep learning have achieved high accuracy in sign language recognition systems, such approaches typically rely on GPU acceleration and substantial computational resources, limiting their feasibility for accessible, real-world applications. This study proposes a theoretical and methodological framework for evaluating lightweight machine learning models for dynamic ASL recognition under CPU-dependent constraints. Grounded in Human-Computer Interaction Theory, Multimodal CommunicationTheory, and Computational LearningTheory, the framework formalizes the relationship between temporal representation, model complexity, and computational feasibility in edge-based environments. The proposed framework outlines a comparative evaluation strategy using pose-based time-series data extracted from glossed-annotated ASL videos, examining both sequence-preserving models (e.g., Canonical Interval Forest and InceptionTime) and aggregated-feature classifiers (e.g., Random Forest and Logistic Regression). Rather than reporting empirical findings, this paper establishes the conceptual foundations, modeling assumptions, and evaluation criteria necessary to determine whether lightweight classifiers can approximate the performance of deep learning approaches while remaining suitable for edge deployment. By explicitly linking theoretical principles to methodological design choices, this work provides a foundation for future empirical studies and contributes a structured approach for developing accessible, efficient, and scalable sign language recognition systems.

Keywords: American sign language, Edge computing, Lightweight machine learning, Time-series classification, Human-computer interaction, Assistive AI

INTRODUCTION

Dynamic American Sign Language (ASL) recognition is a central problem in accessible human-computer interaction, with direct implications for communication equity and assistive technologies (Angelini et al., 2025; Prietch et al., 2022). While recent deep learning approaches have demonstrated strong performance on sign language recognition tasks, these systems are typically developed for high-resource computing environments. They often rely on GPU acceleration, large parameter counts, and substantial computational overhead (Gan et al., 2023; Shams et al., 2024). Such

Received February 3, 2026; Revised April 1, 2026; Accepted April 17, 2026; Available online July 20, 2026

requirements significantly limit their applicability in real-world contexts that require low latency, low power consumption, and device-level inference.

From an accessibility standpoint, effective ASL recognition systems must operate in real time and function reliably on mobile or edge devices (Bekeš et al., 2024). However, much of the existing literature prioritizes accuracy under laboratory conditions, often overlooking deployment feasibility and usability constraints (Sihan et al., 2024). This disconnect has tangible consequences, as evidenced by documented failures of sign language interpretation in public communication settings (Alter, 2014; MacFarlane, 2017).

This paper addresses this gap by proposing a theory-driven framework for evaluating lightweight machine learning models for dynamic ASL recognition under realistic computational constraints. Rather than presenting empirical results, the study defines the conceptual, representational, and methodological foundations for assessing trade-offs among temporal fidelity, model complexity, and deployability.

THEORETICAL FOUNDATIONS

Table 1 shows the overview of the relevance and design implications of the underlying theories.

Table 1: Theoretical perspectives and design implications.

Theory	Relevance	Design Implication
HCI	Usability & latency	Edge deployment required
MCT	Temporal structure	Sequence modeling necessary
CLT	Efficiency constraints	Lightweight models preferred

Human-Computer Interaction and Accessibility Constraints

Human-Computer Interaction (HCI) research emphasizes usability, responsiveness, and system efficiency as essential properties of interactive technologies (Prietch et al., 2022). In the context of sign language recognition, latency and reliability directly affect communication flow and user experience, particularly for deaf and hard-of-hearing users (Angelini et al., 2025). Systems that introduce perceptible delays or require specialized hardware risk undermining accessibility rather than supporting it (Bekeš et al., 2024).

HCI scholarship consistently demonstrates that high computational overhead can degrade usability, especially in mobile or edge-based environments where processing resources are limited (Sihan et al., 2024). From this perspective, recognition accuracy alone is an insufficient metric for evaluating ASL recognition systems. Practical ASL recognition systems must instead balance performance with responsiveness and deployability.

Multimodal Communication Theory and Temporal Structure

Multimodal Communication Theory conceptualizes meaning as emerging from the coordinated temporal and spatial cues across multiple channels,

including hand motion, body posture, and facial expression (Guo, 2024). Within ASL, semantic content is inherently dynamic and sequence-dependent, making temporal modeling a fundamental requirement rather than an optional enhancement to recognition systems (Núñez-Marcos et al., 2023).

Prior work in sign language recognition consistently demonstrates that approaches preserving temporal structure outperform static or frame-based representations, particularly for continuous or dynamic gestures (Li et al., 2019; Shams et al., 2024). These findings underscore the necessity of modeling temporal dependencies when designing ASL recognition systems. Consequently, any evaluation framework for ASL recognition models must explicitly account for how temporal information is represented, preserved, and processed under computational constraints.

Computational Learning Theory and Efficiency Trade-offs

Computational Learning Theory (CLT) provides a formal foundation for analyzing trade-offs among model expressiveness, generalization capability, and computational feasibility (Clark & Lappin, 2012). Although highly complex models may achieve strong performance under ideal conditions, they frequently exhibit diminishing returns when deployed in constrained environments (Damdoo & Kumar, 2025).

From a CLT perspective, simpler or more structured models may generalize more efficiently when computational resources are constrained, particularly when feature representations are well aligned with the underlying task structure (Goehry et al., 2023). This theoretical lens motivates systematic exploration of lightweight alternatives to deep learning architectures for real-time ASL recognition, especially in scenarios where latency, power efficiency, and deployability are critical design factors.

From Theory to Task Framing for ASL Recognition

Scholarship on MCT informs the decision to treat automated, real-time ASL recognition as a supervised time-series classification problem. However classifications for Time Series have attributes that are ordered and violate typical assumptions that violate the typical machine learning assumption of independent and identically distributed samples (Bagnall et al., 2017; Löning et al., 2019). This distinction is critical for dynamic signs, in which meaning emerges from the sequence of hand configurations rather than any single frame. Classical Machine Learning models cannot directly capture sequential dependencies.

For example, the ASL sign for book consists of at least three ordered hand states; open palms, closed palms, and reopened palms, which collectively convey meaning only when interpreted as a sequence (Figure 1). Treating these frames as independent observations or collapsing them into a single aggregated representation (e.g., average hand position), risks erasing the very temporal structure that encodes linguistic content. As dataset size and class overlap increase, such aggregation can further blur inter-class distinctions and degrade classification reliability (Bagnall et al., 2017). Another task that

is common with time series classification is annotation. Pose detection such as those that can be captured by static signs, can be considered a supervised, time-series annotation task (Löning et al., 2019). However, it is not suitable for dynamic signs.

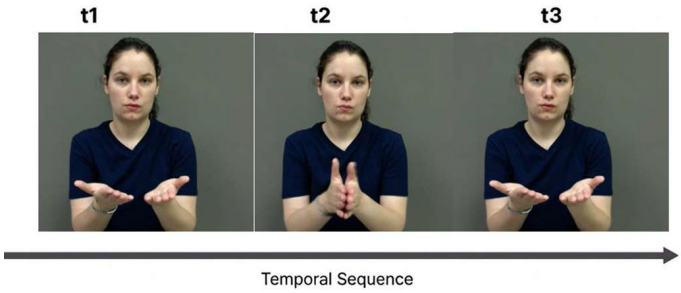

Figure 1: Still frames of ASL sign for "book" from WLASL file 69241.
Note. Adapted from ASL Sign for "Book" (Video No. 69241) [Video], by D. Li, C. Rodriguez, X. Yu, and H. Li, 2020, WLASL. Licensed under C-UDA 1.0. The images show the sequential opening movement of the hands from a closed position.

Scholarship in CLT motivates training all models on the same curated feature set prior to performance comparison. Although deep neural networks are often favored for modeling complex multimodal signals such as sign language, Guo (2024) warns that high model complexity and excessive parameters can increase the risk of overfitting. To avoid confounding effects from differing feature pipelines, this framework standardizes pose-based features into a common tabular representation used across all models, enabling fair comparison of representational strategies rather than feature engineering differences. Consistent with HCI principles, the framework further constrains hardware choices to camera-only inputs and requires models to operate with low latency on edge devices.

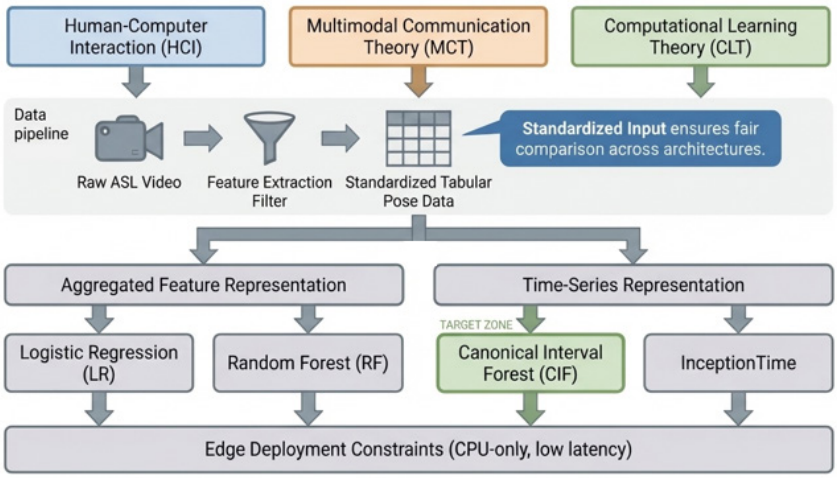

Figure 2: Conceptual framework for lightweight, edge-based ASL recognition.
Note: The framework integrates Human-Computer Interaction principles, Multimodal Communication Theory, and Computational Learning Theory to relate temporal representation strategies and model classes to computational constraints relevant to edge-based American Sign Language recognition.

Model Classes Relevant to Lightweight ASL Recognition

Table 2 shows the overview of the model's structure, computational cost, and suitability for edge deployment.

Table 2: Model classes and their temporal and computational properties.

Model	Type	Temporal Modeling	Computational Cost	Edge Suitability
Logistic Regression	Linear	None	Very Low	High
Random Forest	Ensemble	Aggregated	Low	High
CIF	Time-Series Ensemble	Yes	Moderate	Moderate
InceptionTime	Deep Learning	Yes	High	Low

To operationalize the theoretical considerations outlined above, this framework focuses on four representative model classes that span a range of computational paradigms and levels of temporal modeling complexity.

Logistic Regression (LR) serves as a linear baseline emphasizing simplicity, interpretability, and minimal computational overhead. Although LR does not model temporal dependencies directly, it provides a lower-bound reference for assessing the added value of more complex approaches (Goehry et al., 2023).

Random Forest (RF) is an ensemble-based classifier that operates on fixed-length feature representations. RF models are computationally efficient and robust to noise, but they rely on aggregated descriptors rather than native temporal modeling. As such, they are well-suited to resource-constrained environments (Bagnall et al., 2017).

The Canonical Interval Forest (CIF) is a time-series ensemble classifier that preserves temporal structure through interval-based feature extraction across multiple resolutions (Middlehurst et al., 2020). CIF occupies a middle ground between classical machine learning and deep learning, offering temporal modeling capabilities at a moderate computational cost.

InceptionTime represents a deep learning baseline for time-series classification, employing parallel convolutional filters to capture multi-scale temporal patterns (Middlehurst et al., 2021). Despite its effectiveness, its architectural complexity and resource demands limit suitability for CPU-based edge deployment (Gan et al., 2023).

Methodological Framework

This study adopts a quantitative, comparative methodological framework designed to evaluate lightweight machine learning models under controlled, resource-constrained conditions. The framework assumes a supervised learning paradigm in which models learn mappings between pose-based input features and labeled ASL gestures (Li et al., 2019).

Pose landmarks extracted from video sequences support two complementary representational strategies: sequential time-series inputs for temporal models and aggregated statistical descriptors for fixed-length classifiers. Evaluating

these representations under identical preprocessing conditions isolates the effect of model architecture on performance and computational feasibility (Middlehurst et al., 2020; Goehry et al., 2023).

Importantly, this paper defines an evaluation strategy without reporting empirical outcomes. Performance metrics, statistical comparisons, and deployment benchmarks are intentionally deferred to a subsequent empirical study informed by the framework presented here.

Implications and Research Continuity

By integrating HCI, Multimodal Communication Theory, and Computational Learning Theory, this framework reframes dynamic ASL recognition as a constrained optimization problem involving temporal fidelity, computational efficiency, and usability (Clark & Lappin, 2012; Prietch et al., 2022). Emphasis is placed on evaluation under realistic deployment conditions rather than idealized laboratory settings.

The framework is intended to inform a follow-on empirical investigation examining how different model architectures perform under CPU-based constraints typical of edge and mobile environments. The associated research questions and hypotheses are addressed in a companion empirical study that operationalizes the concepts developed here.

CONCLUSION

This paper presented a theoretical and methodological framework for evaluating lightweight machine learning models for dynamic American Sign Language recognition under resource-constrained, edge-based conditions. Grounded in established theory and prior empirical work, the framework formalizes key trade-offs among temporal modeling, computational efficiency, and usability, contributing to ongoing efforts in accessible, human-centered, and edge-based AI (Angelini et al., 2025; Gan et al., 2023).

The framework motivates a comparison of models that share the same curated pose-based features but differ in how they represent temporal structure, contrasting time-series–preserving approaches with aggregated-feature approaches. This conceptual foundation directly informs a subsequent empirical study conducted under CPU-only conditions.

REFERENCES

Angelini, R., Spiel, K., & De Meulder, M. (2025). Speculating Deaf Tech: Reimagining Technologies Centering Deaf People. Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, 1–18. https://doi.org/10.1145/3706598.3713238

Alter, C. (2014). Officials Linked to Bogus Mandela Interpreter Resign. Time. https://time.com

Bagnall, A., Lines, J., Bostrom, A., Large, J., & Keogh, E. (2017). The Great Time Series Classification Bake Off: A Review and Experimental Evaluation of Recent Algorithmic advances. Data Mining and Knowledge Discovery, 31(3), 606–660. https://doi.org/10.1007/s10618-016-0483-9

Bekeš, E. R., Galzina, V., & Kolar, E. B. (2024). Using human-computer interaction (HCI) and artificial intelligence (AI) in education to improve the literacy of deaf and hearing-impaired children. Proceedings of the 47th MIPRO ICT and Electronics Convention, 1375–1380. https://doi.org/10.1109/MIPRO60963.2024.10569417

Clark, A., & Lappin, S. (2012). Computational learning theory and language acquisition. In Philosophy of Linguistics (pp. 445–475). Elsevier. https://doi.org/10.1016/B978-0-444-51747-0.50013-5

Damdoo, R., & Kumar, P. (2025). SignEdgeLVM transformer model for enhanced sign language translation on edge devices. Discover Computing, 28(1), 15. https://doi.org/10.1007/s10791-025-09509-1

Gan, S., Yin, Y., Jiang, Z., Xie, L., & Lu, S. (2023). Towards real-time sign language recognition and translation on edge devices. Proceedings of the 31st ACM International Conference on Multimedia, 4502–4512. https://doi.org/10.1145/3581783.3611820

Goehry, B., Yan, H., Goude, Y., Massart, P., & Poggi, J.-M. (2023). Random forests for time series. REVSTAT - Statistical Journal, 21(2), 283–302. https://doi.org/10.57805/REVSTAT.V21I2.400

Guo, Y. (2024). Multimodal multilabel classification by CLIP. arXiv. https://doi.org/10.48550/arXiv.2406.16141

Li, D., Opazo, C. R., Yu, X., & Li, H. (2019). Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. arXiv. https://doi.org/10.48550/arXiv.1910.11006

Löning, M., Bagnall, A., Ganesh, S., Kazakov, V., Lines, J., & Király, F. J. (2019). Sktime: A unified interface for machine learning with time series. arXiv. https://doi.org/10.48550/arXiv.1909.07872

Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.-L., Yong, M. G., Lee, J., Chang, W.-T., Hua, W., Georg, M., & Grundmann, M. (2019). MediaPipe: A framework for building perception pipelines. arXiv. https://doi.org/10.48550/arXiv.1906.08172

MacFarlane, D. (2017). Sign language interpreter warns of "pizza" and "bear monster" in Irma briefing. The Weather Channel. https://weather.com

Middlehurst, M., Large, J., & Bagnall, A. (2020). The canonical interval forest (CIF) classifier for time series classification. IEEE International Conference on Big Data, 188–195. https://doi.org/10.1109/BigData50022.2020.9378424

Middlehurst, M., Large, J., Flynn, M., Lines, J., Bostrom, A., & Bagnall, A. (2021). HIVE-COTE 2.0: A new meta ensemble for time series classification. Machine Learning, 110(11-12), 3211–3243. https://doi.org/10.1007/s10994-021-06057-9

Mishra, A., Hannig, F., Teich, J., & Sabih, M. (2022). MOSP: Multi-objective sensitivity pruning of deep neural networks. IEEE International Green and Sustainable Computing Conference, 1–8. https://doi.org/10.1109/IGSC55832.2022.9969363

Núñez-Marcos, A., Perez-de-Viñaspre, O., & Labaka, G. (2023). A survey on sign language machine translation. Expert Systems with Applications, 213, 118993. https://doi.org/10.1016/j.eswa.2022.118993

Prietch, S., Sánchez, J. A., & Guerrero, J. (2022). A systematic review of user studies as a basis for the design of systems for automatic sign language processing. ACM Transactions on Accessible Computing, 15(4). https://doi.org/10.1145/3563395

Shams, K. A., Reaz, M. R., Rafi, M. R. U., Islam, S., Rahman, M. S., Rahman, R., Reza, M. T., Parvez, M. Z., Chakraborty, S., Pradhan, B., & Alamri, A. (2024). Multimodal ensemble approach leveraging spatial, skeletal, and edge features for sign language recognition. IEEE Access, 12, 83638–83657. https://doi.org/10.1109/ACCESS.2024.3410837

Shao, F., Zhang, T., Gao, S., Sun, Q., & Yang, L. (2024). Computer vision-driven gesture recognition: Toward natural and intuitive human-computer interfaces. Proceedings of the 4th International Conference on Electronic Information Engineering and Computer Communication, 1313–1317. https://doi.org/10.1109/EIECC64539.2024.10929472

Sihan, T., Itoyama, K., & Nakadai, K. (2024). Advancing human-computer interaction: End-to-end sign language translation. Transactions of the Human Interface Society, 26(4), 391–398. https://doi.org/10.11184/his.26.4_391